PROGRAM STATE DISTRIBUTION REGULARIZATION FOR LEARNING ACROSS SINGLE CELL TECHNOLOGIES: BENEFITS AND LIMITATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Single-cell integration aims to enable joint analysis of datasets by reconciling technical distribution shifts while preserving biological variation. Most methods rely on correspondences or invariances across datasets that can be ambiguous, unstable, or absent when biological support set differs across datasets. Methods based on gene programs offer a more structured unit of alignment, but matching program identity alone does not ensure comparable program states or their combinations. We posit the joint distribution of shared program states as the cross technology invariant, and explicitly decompose it into program specific state distributions and interprogram dependence. We implement this decomposition using a shared program bottleneck with inference and observation modules specific to each technology, aligning program marginals through quantiles and interprogram dependence in rank space. Controlled simulations with five independent generators and five optimization repeats show higher correlation between matched factors than the same model without alignment at all four tested levels of observation distortion. On a development cohort, adding dependence alignment improves mean clustering and transfer scores over marginal alignment across five paired fits; for donors excluded during training, prediction of protein from RNA raises median protein R^2 from 0.187 without alignment to 0.250 with the combined objective. These results show that aligning program state distributions can improve recovery of shared factors and prediction across technologies.

1 INTRODUCTION

Single-cell RNA sequencing technologies differ in capture, amplification, and quantification. Learning across unpaired technologies therefore requires a representation that recovers shared biological variation and supports downstream comparisons (Luecken et al., 2022). This raises a modelling question: which distributions should be aligned to make the shared representation useful?

Gene-program, topic, and structured-factor models connect cell coordinates to gene responses (Svensson et al., 2020; Zhao et al., 2021; Moinfar & Theis, 2024; Zhang et al., 2026). A coordinate's marginal distribution describes its range and prevalence; dependence describes how coordinates vary together. Matching program marginals alone does not determine their joint distribution: two representations can have identical coordinate-wise distributions but different inter-program dependence. We therefore regularize both objects separately, incorporating the marginal–dependence approach from domain adaptation (Ma et al., 2024) into single-cell count reconstruction.

Our method, *program-state distribution regularization*, uses a technology-specific encoder followed by a shared program bottleneck, a shared gene-space decoder, and a technology-specific observation decoder. Quantile matching aligns the full marginal distributions of individual coordinates to a reference-free Wasserstein barycenter. A copula-based loss compares joint rank vectors without fitting a parametric copula density. Both losses act on the coordinates used for reconstruction and downstream analysis, allowing the same backbone to test their contributions separately. The target assumes a shared biological state distribution *after capture and filtering*, formalized as captured-population comparability in Section 3.1.

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.

Our contributions are threefold. First, we formulate marginal and dependence alignment directly on a shared program representation used for count reconstruction. Second, we characterize the population assumptions under which these targets are biologically meaningful and analyze failures under selective capture. Third, experiments evaluate latent-factor recovery, integration, and held-out protein prediction, with same-backbone controls separating the alignment objectives. The evidence distinguishes whole-representation utility from individual-factor interpretability: improved recovery or prediction does not identify the learned coordinates as stable biological programs.

2 Related Work

Program representations. Program models differ in how they define coordinates and connect them to genes. Linear loadings, topics, and consensus expression programs make this connection explicit (Svensson et al., 2020; Zhao et al., 2021; Kotliar et al., 2019); structured nonlinear factors extend it beyond linear responses (Moinfar & Theis, 2024). Shared-variation models and program-level correspondences address relationships across datasets (Welch et al., 2019; Argelaguet et al., 2020; Zhang et al., 2026), while knowledge-guided models incorporate external biological structure (Kunes et al., 2024; Lotfollahi et al., 2023; Stark et al., 2021; Chen et al., 2026). Our focus is the distributional target imposed on shared program coordinates across unpaired technologies. We assess coordinate interpretation separately through full-decoder responses and independent biological evidence.

Integration and distribution matching. Integration methods address cross-dataset comparability through batch-aware generative models (Lopez et al., 2018; Hrovatin et al., 2025) and complementary representation or correction approaches (Zhu et al., 2026; He et al., 2025; Korsunsky et al., 2019; Hao et al., 2021). A separate design choice is what distribution to match: coordinate marginals, the joint latent distribution, or inter-coordinate dependence. Distribution-regularized autoencoders, differentiable transport, MMD, and unbalanced transport provide tools for these choices (Tolstikhin et al., 2018; Kolouri et al., 2018; Gretton et al., 2012; Chizat et al., 2017). We compare alignment targets within one reconstruction backbone to isolate their effects from architectural differences.

Dependence regularization in transfer learning. Ma et al. (2024) directly precede our use of separately weighted marginal and dependence discrepancies. Their supervised transfer setting contrasts with our count-reconstruction setting across unpaired single-cell technologies, where both constraints act on the program coordinates used downstream. Their copula-based domain adaptation network combines a supervised source-task loss with marginal discrepancies and Gaussian-copula differences in pairwise dependence strength. We use one-dimensional Wasserstein barycenters for marginals and MMD on joint smooth-rank vectors for dependence, without fitting a parametric copula density. The resulting formulation studies program-coordinate recovery and prediction under captured-population comparability.

Variational copula models for multimodal learning. Wu et al. (2025) introduce CM^2, combining Gaussian-mixture modality distributions with a copula model trained by variational inference for supervised multimodal fusion. They also sample missing-modality representations under a missing-at-random assumption. Their dependence object links modalities of an observation; ours links program coordinates within a cell and compares their population distributions across technologies without cell pairing. Semantic identification requires additional conditional-distribution assumptions beyond marginal–dependence decomposition and reconstruction (Shrestha & Fu, 2024).

3 Methods

We jointly fit a count-reconstruction model and a regularizer on full program marginal distributions and their dependence. The population target and architecture are defined below.

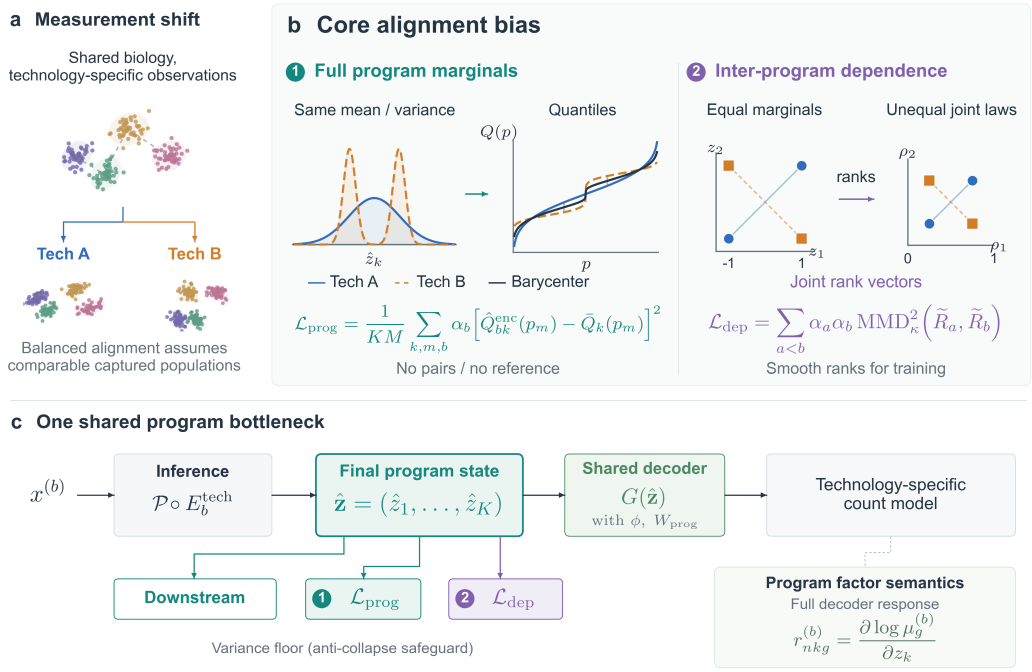

Figure 1: Program-state distribution regularization. (a) Technology-specific measurements under captured-population comparability. (b) Quantiles describe full marginal distributions; joint rank vectors describe dependence. (c) Technology-specific encoder, shared program bottleneck, shared gene-space decoder, and technology-specific observation decoder. The bottleneck supports both alignment losses and downstream analysis, with a soft total-variance penalty. Full-decoder responses support a separate semantic audit.

3.1 Problem setup and population target

We observe raw counts $\mathbf{X}^{(b)} \in \mathbb{N}_0^{N_b \times G}$ from B batches, with cell vector $\mathbf{x}_n$, batch b_n, and library size $\ell_n = \sum_g x_{ng}$. The model learns K scalar program coordinates $\widehat{\mathbf{z}}_n \in \mathbb{R}^K$, whose gene responses are defined through the decoder. Write Z_{nk} for a latent biological state and F_k for its CDF.

We define *captured-population comparability* as a shared joint biological state distribution across technologies after capture and filtering. The following assumptions separate its marginal and dependence components.

Assumption 1 (Shared captured-state marginals)**.** *Among cells actually captured and retained, $P(Z_{nk} \le t \mid b_n = b) = F_k(t)$ for every b, k, t. Cells are independent conditional on the population and experimental unit used in the sampling analysis; programs need not be independent.*

Assumption 2 (Shared inter-program dependence)**.** *The captured biological state vectors share a joint distribution Π. For continuous marginals, this adds a common copula to Assumption 1.*

The marginal-only control uses the first assumption; the primary model uses both. State-dependent capture can violate captured-population comparability even for aliquots of one source (Appendix B). Technologies may also differ in their observation kernels. Consequently, comparable captured populations can still yield different encoder-output distributions, and alignment acts as a regularizer on learned coordinates.

3.2 Encoding and decoding through a shared program bottleneck

The architecture follows four stages: a technology-specific encoder, a shared program bottleneck, a shared gene-space decoder, and a technology-specific observation decoder. The first two produce

the final coordinates:

$$\mathbf{y}_n = \log(1 + 10^4 \mathbf{x}_n/\ell_n), \qquad \widehat{\mathbf{z}}_n = C\tanh\big(\mathcal{P}(\mathcal{E}_{b_n}^{\mathrm{tech}}(\mathbf{y}_n))/C\big)\,. \tag{1}$$

The technology-specific encoder $\mathcal{E}_b^{\mathrm{tech}}$ combines a shared nonlinear encoder with a batch-specific low-rank residual. The shared program bottleneck applies the affine projection $\mathcal{P}$ and bounds its K output coordinates by $C = 6$. For decoding,

$$\mathbf{v}_{nk} = \phi([\widehat{z}_{nk}, \mathbf{e}_k]), \quad \mathbf{u}_n = \mathbf{c} + \mathbf{C}^{\mathrm{gene}}\left[\frac{1}{K}\sum_k \mathbf{v}_{nk} + \mathbf{W}^{\mathrm{prog}}\widehat{\mathbf{z}}_n\right], \quad \boldsymbol{\eta}_n = \mathcal{D}_{b_n}^{\mathrm{tech}}(\mathbf{u}_n). \tag{2}$$

Within the shared gene-space decoder, each nonlinear branch ϕ takes one coordinate and its learned program embedding $\mathbf{e}_k$. The shared linear basis $\mathbf{W}^{\mathrm{prog}}$ complements these branches; both use the same bottleneck, with no auxiliary latent bypass. $\mathbf{C}^{\mathrm{gene}}$ produces the gene-space feature vector $\mathbf{u}_n$. The technology-specific observation decoder $\mathcal{D}_b^{\mathrm{tech}}$ then maps this vector to expression logits.

Count reconstruction. Raw counts follow the working likelihood $x_{ng} \sim \mathrm{NB}(\mu_{ng}, \nu_{b_n g})$, where $\boldsymbol{\mu}_n = \ell_n \,\mathrm{softmax}(\boldsymbol{\eta}_n)$ and ν_{bg} is inverse dispersion (Lopez et al., 2018). The deterministic encoder minimizes batch-balanced negative log likelihood $\mathcal{L}_{\mathrm{rec}}$ with weights $\alpha_b = 1/B$ (Appendix C). Full-length reads remain read counts rather than UMI counts.

3.3 Full program marginal distributions and their dependence

For the encoded program-k values in batch b, let $\widehat{Q}_{bk}^{\mathrm{enc}}$ be the empirical quantile function. On midpoint levels $p_m = (m - 1/2)/M$, define

$$\overline{Q}_k(p_m) = \sum_b \alpha_b \widehat{Q}_{bk}^{\mathrm{enc}}(p_m), \qquad \mathcal{L}_{\mathrm{prog}} = \frac{1}{KM}\sum_{k,m,b} \alpha_b [\widehat{Q}_{bk}^{\mathrm{enc}}(p_m) - \overline{Q}_k(p_m)]^2. \tag{3}$$

This approximates squared one-dimensional Wasserstein distances to a reference-free quantile barycenter. The target is equality of full marginal shapes, including state proportions, so unmatched state regions also incur a penalty. Quantiles are recomputed within each minibatch. Sorting is differentiable almost everywhere and provides gradients except at ties (Kolouri et al., 2018). The loss requires no persistent cell correspondences.

Dependence alignment. For continuous marginals, equal marginal CDFs and equal copulas imply equal joint distributions. We approximate joint rank vectors with differentiable within-batch smooth ranks, whose empirical joint distribution is $\widetilde{R}_b$, and minimize

$$\mathcal{L}_{\mathrm{dep}} = \sum_{a<b} \alpha_a \alpha_b \,\mathrm{MMD}_\kappa^2(\widetilde{R}_a, \widetilde{R}_b). \tag{4}$$

The fixed Gaussian-kernel mixture compares joint rank vectors (Gretton et al., 2012). We use smooth ranks during training to obtain gradients, whereas evaluation uses exact midranks. The evaluation distribution is therefore discrete and depends on how ties are handled. The loss serves as a finite-sample surrogate for population copula matching. Appendix C specifies the transform and kernels; RQ2 compares its empirical effect with marginal alignment alone.

Scale control and joint optimization. Global coordinate shrinkage can reduce $\mathcal{L}_{\mathrm{prog}}$ while decoder compensation preserves reconstruction. We therefore use the soft total-variance floor

$$V = \mathrm{tr}[\mathrm{Cov}(\widehat{\mathbf{z}})], \qquad \mathcal{L}_{\mathrm{scale}} = [\max(0, \tau_V - V)]^2, \tag{5}$$

under the batch-balanced pooled distribution. This soft constraint penalizes small total variance while allowing individual coordinates to be inactive; collapse and redundant factors remain possible (Appendix B). The primary objective is

$$\mathcal{L} = \mathcal{L}_{\mathrm{rec}} + \lambda_{\mathrm{prog}}\mathcal{L}_{\mathrm{prog}} + \lambda_{\mathrm{dep}}\mathcal{L}_{\mathrm{dep}} + \lambda_{\mathrm{scale}}\mathcal{L}_{\mathrm{scale}} + \lambda_{\mathrm{adapt}}\mathcal{L}_{\mathrm{adapt}}, \tag{6}$$

where $\mathcal{L}_{\mathrm{adapt}}$ penalizes batch-specific residual parameters. Reconstruction and scale control warm up before both alignment weights ramp; all modules train jointly. Inference returns $\widehat{\mathbf{z}}_n$ directly. Frozen settings and numerical details are in Appendix C.

Mass relaxation under selection shift. We evaluate an unbalanced-transport variant when captured-population comparability fails, replacing balanced marginal alignment and disabling dependence matching. A matched balanced control uses the same numerical settings to isolate mass relaxation (Appendix C). Both objectives compare observed coordinate distributions; capture probabilities and absent states are outside the model.

Interpretation and sampling scope. A factor's gene response is the derivative of the log count mean through the full technology-specific decoder at fixed reference states and library size. Reconstruction and distribution matching leave coordinate semantics ambiguous (Appendix B); we assess a gene-program interpretation through response and activity stability plus independent biological evidence (Appendix G). The sampling result in Appendix B bounds marginal-distribution estimation for independent held-out cells under frozen encoders.

4 Experimental Evaluation

We evaluate factor recovery, cross-technology integration, and protein prediction, then use alignment-object and population-shift controls to explain the measured effects. GSE133549 informed configuration and is exploratory; Ding PBMC and the raw-count-compatible pancreas subset use the frozen model. Same-backbone controls compare no alignment, marginals alone, and marginals plus dependence; external methods provide context.

Unless stated otherwise, error bars and standard deviations over optimizer seeds quantify fitting variability on fixed cohorts rather than biological uncertainty. Simulations use independent generators as replicate populations and average optimizer repeats within each generator when available. Data, budgets, full numerical profiles, and probe sensitivities are in Appendices D–F.

4.1 Known-factor recovery and real integration (RQ1)

E1 crosses five independent generators with five optimizer seeds across four observation-distortion severities and a separate branching setting. Combined alignment improves matched-factor recovery over no alignment at all four severities, with paired generator intervals excluding zero (Figure 2). At severity 2, mean factor correlation increases from 0.671 to 0.714.

Improved recovery comes at the cost of lower true-state neighborhood recall and slightly worse reconstruction. No correlation gain is observed in the branching setting. External methods show different trade-offs across factor recovery, neighborhood preservation, and transfer (Appendix F).

On GSE133549, the frozen model's mean label-transfer F1 exceeds CONCORD and scCRAFT, while scCRAFT leads several integration measures. On Ding and pancreas, combined alignment improves mean ARI over no alignment with nearly unchanged F1; external baselines retain advantages on individual endpoints (Figure 3). Ding holds out an experimental round with unproven donor independence; pancreas uses within-technology cell holdout in a cross-study task.

4.2 Contributions of marginal and dependence alignment (RQ2)

E3 changes only the alignment target on one frozen backbone and the GSE133549 development split. Five paired optimizer seeds use the primary $H = 512$ minibatch setting (Table 1). Adding dependence alignment increases mean ARI from 0.258 to 0.301 and macro-F1 from 0.886 to 0.900 relative to marginal alignment alone, with improvements in four of five paired seeds for both metrics. It also reduces exact-rank MMD in all five seeds, but decreases biological ASW.

Other objectives retain advantages on different endpoints. Full-latent MMD has slightly higher mean batch ASW and iLISI than combined alignment, while joint Wasserstein has higher biological ASW and lower test NB NLL in all five pairs. Appendix F gives the full metric table, paired differences, and their directions.

A frozen-model diagnostic on GSE133549 finds smaller coordinate variance and greater decoder sensitivity with alignment. Held-out marginal discrepancy still decreases after normalization by pooled training variance in all five paired runs (Appendix F, Table 15); pure rescaling is therefore insufficient to account for the difference. Synthetic controls show different scale changes and

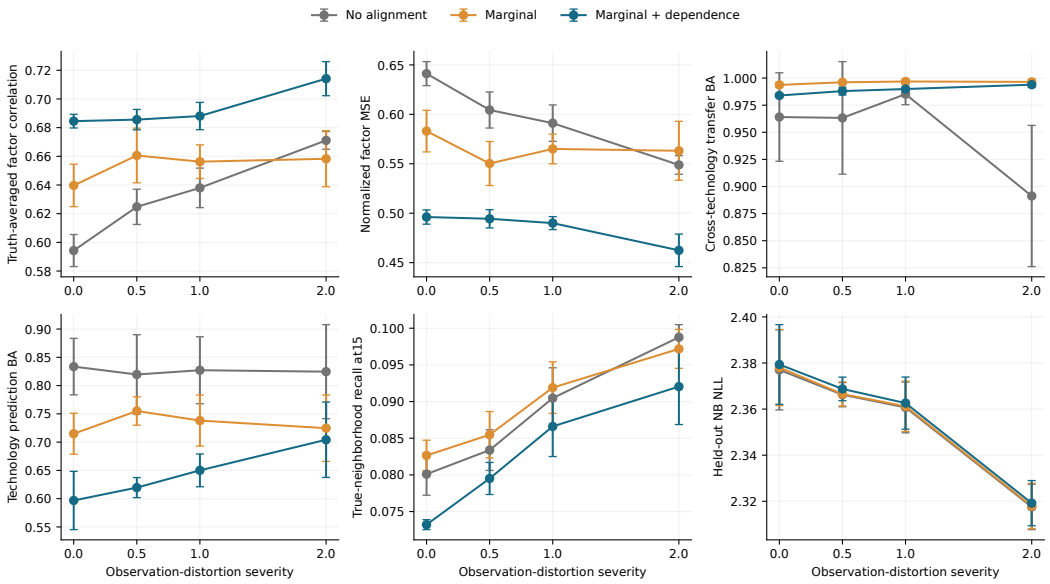

Figure 2: Known-factor recovery across four observation-distortion severities. Points and error bars show the mean and sample SD of five independent generator means, each averaging five optimizer seeds (300 fits). Severity zero retains depth and dispersion differences. The separate branching setting and seed identities are reported in Appendix F.

Table 1: Six alignment objects on the GSE133549 development split and frozen backbone, with $H = 512$. Entries are mean ± sample SD across five paired optimizer seeds (7/19/37/73/101). Full metrics are in Table 13.

Objective	ARI ↑	Bio ASW ↑	Rank MMD ↓	F1 ↑	Test NB ↓
No alignment	0.275 ± 0.008	0.550 ± 0.008	0.0161 ± 0.0020	0.895 ± 0.007	0.897 ± 0.008
Means/variances	0.259 ± 0.018	0.595 ± 0.026	0.0164 ± 0.0011	0.878 ± 0.012	0.890 ± 0.007
Full-latent MMD	0.288 ± 0.017	0.613 ± 0.011	0.0079 ± 0.0007	0.889 ± 0.010	0.890 ± 0.005
Joint Wasserstein	0.277 ± 0.018	0.624 ± 0.009	0.0117 ± 0.0006	0.879 ± 0.009	0.882 ± 0.003
Program marginals	0.258 ± 0.021	0.608 ± 0.016	0.0150 ± 0.0022	0.886 ± 0.009	0.893 ± 0.005
Marginals + dependence	0.301 ± 0.023	0.592 ± 0.010	0.0060 ± 0.0003	0.900 ± 0.006	0.888 ± 0.004

mixed normalized-discrepancy results. The diagnostic complements the capacity experiment's activity counts by examining decoder sensitivity explicitly.

4.3 Protein prediction and factor interpretation (RQ3)

Using the aligned RNA representation yields higher held-out protein-prediction R^2 for the task covering 134 BMMC ADTs (Table 2). Mean median-protein R^2 across five fits increases from 0.187 without alignment to 0.250 with the combined objective, compared with 0.209 for training-fitted RNA PCA. Spearman correlation is close to PCA.

Training donors 1–6 fit the representation and probes, donor 7 selects probe penalties, and donors 8–9 provide test measurements. Both test donors share an acquisition site with part of the training set.

This improvement applies to the representation as a whole. The separate semantic audit finds that 0/64 combined-model factors meet the joint response, activity, and stable ADT-evidence rule. Appendix F reports the complete audit, including the response/activity trade-off on BMMC and the limited top-gene stability on GSE133549 (Figure 5).

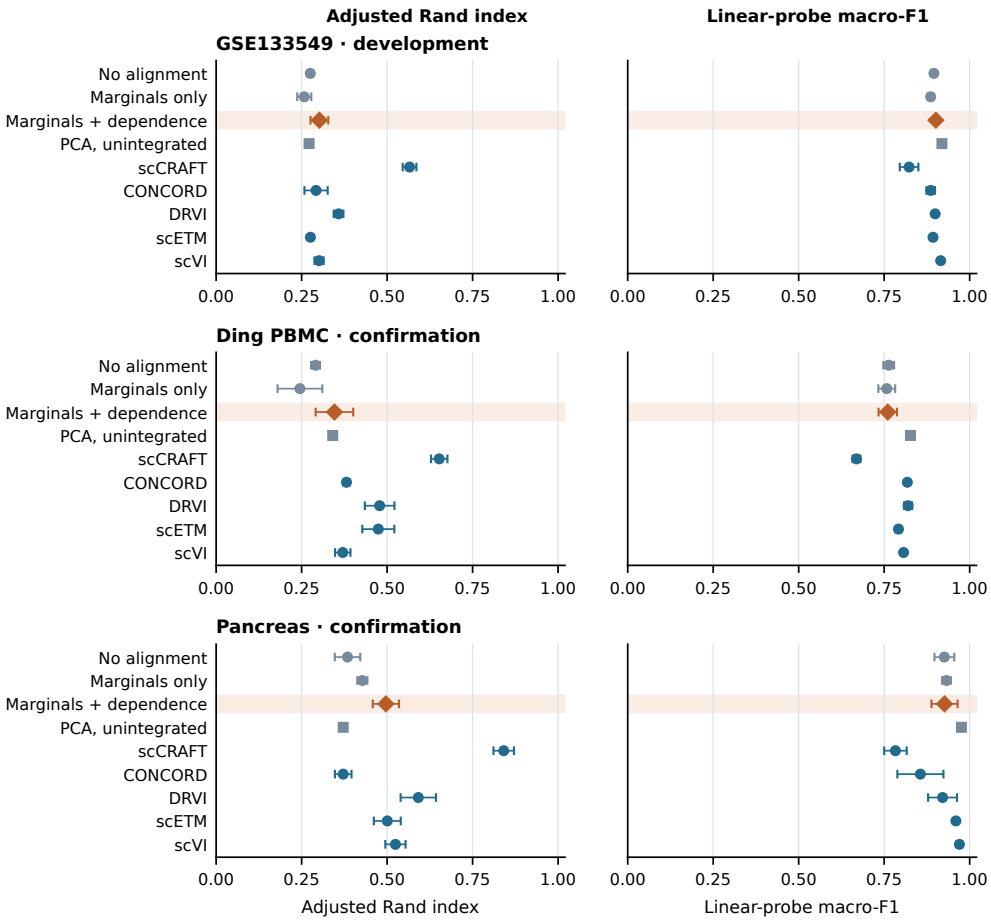

Figure 3: Clustering and prediction on GSE133549 (1,057 test cells), Ding (10,448 labelled PBMC2 query cells), and pancreas (1,650 within-technology held-out cells). Points and bars show means and sample SD across five caller seeds per task; training-fitted PCA is one fixed control. The highlighted marginal/dependence configuration is frozen. Full profiles and comparison budgets are in the appendices.

Table 2: E5 donor-held-out prediction of 134 ADTs. Entries are mean ± sample SD across five optimizer seeds of the median metric across proteins; fixed controls are shown without a seed standard deviation. Training used RNA only. Donors 8–9, the test units, both belong to acquisition site 4.

Representation or control	Median protein R^2	Median protein Spearman
Combined program alignment	0.250 ± 0.006	0.440 ± 0.012
No distribution alignment	0.187 ± 0.018	0.425 ± 0.023
Training-fitted RNA PCA	0.209	0.439
Library size and batch	-0.176	0.133
Cell type (evaluation only)	0.235	0.401

4.4 Sensitivity to captured-population comparability (RQ4)

E6 varies selective capture, missing states, and genuine dependence on three independent generators with one fixed optimizer seed. Increasing capture strength from $\gamma_c = 0$ to 1 changes the mean combined-minus-no-alignment factor-correlation difference from $+0.204$ to -0.156. Missing states also reduce recovery, and dependence alignment attenuates genuine biological dependence differences.

These results show why captured-population comparability matters for the regularizer. Mass relaxation provides no reliable rescue in the matched balanced/unbalanced controls. The full strength curve, three-scenario plot, and numerical controls are in Appendix F (Figure 6).

5 Discussion and Limitations

Program-state distribution regularization connects population-level biological structure to count-based representation learning. Full marginal distributions and their dependence constrain one shared program bottleneck used for reconstruction, alignment, and prediction. Building on marginal–dependence regularization (Ma et al., 2024) and related copula modelling (Wu et al., 2025), this formulation makes the alignment object an explicit modelling choice whose effects can be studied within one backbone.

Across simulations, integration, and protein prediction, distribution alignment can improve task-relevant recovery or prediction without uniformly preserving biological geometry. Factor recovery, local neighborhood structure, and predictive utility are different objectives: a representation can support one while losing information relevant to another. The simulation and alignment-object controls make these trade-offs explicit (Figure 2; Table 1). They also distinguish evidence for the combined regularizer against no alignment from evidence for the dependence term, which comes from the direct comparison with marginal alignment alone on the development cohort.

Distribution alignment can improve the utility of a representation without identifying its individual coordinates as stable biological programs. A predictor can use information distributed across several coordinates, whereas a program interpretation requires reproducible coordinate-level responses, activities, and independent biological evidence. The protein-prediction result supports the former use; the semantic audit leaves the latter unresolved (Table 2; Appendix F). Evaluating these outcomes separately makes clear what downstream performance establishes about the learned representation.

Application requires evidence for captured-population comparability. Selective capture can reverse the recovery gain, and dependence matching can suppress a genuine biological difference (Figure 6). Shared source material alone is insufficient because capture and filtering can change the retained population. The current unpaired model does not provide a data-only test of comparability or recover capture probabilities and absent states. The stress tests and mass-relaxation controls define the tested scope; choosing when to enable alignment still requires information about the populations being compared.

The evidence covers one model family and a finite set of cohorts and generators. The protein task retains site overlap between training and test donors. The soft variance floor controls total scale, while coordinate scale and the effective number of factors remain unresolved. The frozen-model diagnostic adds evidence that pure rescaling cannot fully explain the GSE133549 discrepancy reduction. The fixed-encoder CDF bound addresses distribution estimation, with rare-state coverage depending on the available sample. Broader donor/site evaluation and additional architecture and scale controls would clarify how far the observed benefits extend.

A next step is calibration based on supported state correspondences. Information from other genes could help locate a biologically silent state when one technology captures only active states; silence here denotes biological inactivity rather than an observed zero count. Developing this idea requires a constrained shared calibration map that allows unmatched states and different abundances, followed by known-state experiments designed for unequal support.

6 Conclusion

Program-state distribution regularization couples full program marginal distributions and dependence to single-cell count reconstruction through a shared bottleneck. Experiments provide evidence for improved recovery in simulations with captured-population comparability, mean clustering and transfer gains from dependence alignment on the development cohort, and higher protein-prediction R^2 on held-out donors. These findings support program-state distributions as a target for cross-technology representation learning, with trade-offs in local geometry and sensitivity to population differences. Improved representation utility does not identify individual coordinates as stable biological programs.

AI USE STATEMENT

We used generative AI tools to aid and polish manuscript writing, search for and summarize relevant literature, and support research ideation and execution. This support included method development, mathematical exposition, experiment planning, implementation and execution, and interpretation of results. AI agents generated the synthetic datasets used in the simulation experiments. The authors are responsible for the methods, results, interpretations, and content of the paper.

ETHICS STATEMENT

This work analyzes previously collected single-cell datasets described in Appendix D. Its outputs are statistical representations and model-derived gene responses for research use. Capture bias and donor/site confounding can affect biological interpretation, as illustrated by the population-shift experiments and the cohort-specific limitations. Clinical use or causal claims about gene regulation would require independent evidence beyond the evaluations reported here.

REPRODUCIBILITY STATEMENT

Appendix C specifies the architecture, objectives, and optimization; Appendix B gives the mathematical details; and Appendices D–G describe the data, comparisons, and evaluation protocols. The experimental records retain software versions, input and split provenance, configurations, failed runs, and result artifacts. The reproducibility package is not yet available. We will release an anonymized package of code, frozen configurations, and data-preparation, split, and evaluation instructions before the author-response (rebuttal) period, and add the access link to OpenReview when submission updates reopen.

REFERENCES

Ricard Argelaguet, Damien Arnol, Danila Bredikhin, Yonatan Deloro, Britta Velten, John C. Marioni, and Oliver Stegle. MOFA+: a statistical framework for comprehensive integration of multimodal single-cell data. *Genome Biology*, 2020. doi: 10.1186/s13059-020-02015-1.

Hegang Chen, Yuyin Lu, Yifan Zhao, Zhiming Dai, Fu Lee Wang, Qing Li, Yanghui Rao, and Yue Li. scE2TM improves single-cell embedding interpretability and reveals cellular perturbation signatures. *Nature Communications*, 2026. doi: 10.1038/s41467-026-76825-5.

Lénaïc Chizat, Gabriel Peyré, Bernhard Schmitzer, and François-Xavier Vialard. Scaling Algorithms for Unbalanced Optimal Transport Problems. *arXiv preprint arXiv:1607.05816*, 2017. URL https://arxiv.org/abs/1607.05816v2. Version 2, 22 May 2017; title as printed in the downloaded manuscript.

Chao Gao, Jialin Liu, April R. Kriebel, Sebastian Preissl, Chongyuan Luo, Rosa Castanon, Justin Sandoval, Angeline Rivkin, Joseph R. Nery, Margarita M. Behrens, Joseph R. Ecker, Bing Ren, and Joshua D. Welch. Iterative single-cell multi-omic integration using online learning. *Nature Biotechnology*, 39(8):1000–1007, 2021. doi: 10.1038/s41587-021-00867-x. URL https://www.nature.com/articles/s41587-021-00867-x.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A Kernel Two-Sample Test. *Journal of Machine Learning Research*, 13(25):723–773, 2012. URL https://www.jmlr.org/papers/v13/gretton12a.html.

Yuhan Hao, Stephanie Hao, Erica Andersen-Nissen, et al. Integrated analysis of multimodal single-cell data. *Cell*, 2021. doi: 10.1016/j.cell.2021.04.048.

Chuan He, Panagiotis Filippidis, Steven H. Kleinstein, and Leying Guan. Partially characterized topology guides reliable anchor-free scRNA-integration. *Communications Biology*, 2025. doi: 10.1038/s42003-025-07988-y.

Karin Hrovatin, Amir Ali Moinfar, Luke Zappia, Alejandro Tejada-Lapuerta, Ben Lengerich, Manolis Kellis, and Fabian J. Theis. Integrating single-cell RNA-seq datasets with substantial batch effects. *BMC Genomics*, 2025. doi: 10.1186/s12864-025-12126-3. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC12577435/.

Soheil Kolouri, Phillip E. Pope, Charles E. Martin, and Gustavo K. Rohde. Sliced-Wasserstein Autoencoder: An Embarrassingly Simple Generative Model. *arXiv preprint arXiv:1804.01947*, 2018. URL https://arxiv.org/abs/1804.01947.

Ilya Korsunsky, Nathan Millard, Jean Fan, et al. Fast, sensitive and accurate integration of single-cell data with Harmony. *Nature Methods*, 2019. doi: 10.1038/s41592-019-0619-0.

Dylan Kotliar, Adrian Veres, M. A. Nagy, Shervin Tabrizi, Eran Hodis, Douglas A. Melton, and Pardis C. Sabeti. Identifying gene expression programs of cell-type identity and cellular activity with single-cell RNA-Seq. *eLife*, 2019. doi: 10.7554/eLife.43803.

Robert C. Kunes et al. Supervised discovery of interpretable gene programs from single-cell data. *Nature Biotechnology*, 2024. doi: 10.1038/s41587-023-01940-3. First published online in 2023.

Romain Lopez, Jeffrey Regier, Michael B. Cole, Michael I. Jordan, and Nir Yosef. Deep generative modeling for single-cell transcriptomics. *Nature Methods*, 2018. doi: 10.1038/s41592-018-0229-2.

Mohammad Lotfollahi, Sergei Rybakov, Karin Hrovatin, et al. Biologically informed deep learning to query gene programs in single-cell atlases. *Nature Cell Biology*, 25:337–350, 2023. doi: 10.1038/s41556-022-01072-x.

Malte D. Luecken, Maren B"uttner, Kridsadakorn Chaichoompu, Anna Danese, Marius Interlandi, Michaela F. Mueller, Daniel C. Strobl, Luke Zappia, Martin Dugas, Maria Colom'e-Tatch'e, and Fabian J. Theis. Benchmarking atlas-level data integration in single-cell genomics. *Nature Methods*, 19:41–50, 2022. doi: 10.1038/s41592-021-01336-8.

Shumin Ma, Zhiri Yuan, Qi Wu, Yiyan Huang, Xixu Hu, Cheuk Hang Leung, Dongdong Wang, and Zhixiang Huang. Deep Into the Domain Shift: Transfer Learning Through Dependence Regularization. *IEEE Transactions on Neural Networks and Learning Systems*, 35(10):14409–14423, 2024. doi: 10.1109/TNNLS.2023.3279099. URL https://doi.org/10.1109/TNNLS.2023.3279099.

Pascal Massart. The Tight Constant in the Dvoretzky–Kiefer–Wolfowitz Inequality. *The Annals of Probability*, 18(3):1269–1283, 1990. doi: 10.1214/aop/1176990746.

Amir Ali Moinfar and Fabian J. Theis. Disentangling cellular heterogeneity into interpretable biological factors through structured latent representations. *bioRxiv*, 2024. doi: 10.1101/2024.11.06.622266. URL https://www.biorxiv.org/content/10.1101/2024.11.06.622266v4. Preprint; version 4 posted June 8, 2026; implementation documented in scvi-tools.

Sagar Shrestha and Xiao Fu. Towards Identifiable Unsupervised Domain Translation: A Diversified Distribution Matching Approach. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2401.09671.

Stefan G. Stark, Gilles Gut, Gunnar Rätsch, and Natalie R. Davidson. pmVAE: Learning Interpretable Single-Cell Representations with Pathway Modules. In *ICML Workshop on Computational Biology*, 2021. URL https://icml-compbio.github.io/2021/papers/WCBICML2021_paper_24.pdf.

Valentine Svensson, Adam Gayoso, Nir Yosef, and Lior Pachter. Interpretable factor models of single-cell RNA-seq via variational autoencoders. *Bioinformatics*, 36(11):3418–3421, 2020. doi: 10.1093/bioinformatics/btaa169.

Ilya Tolstikhin, Olivier Bousquet, Sylvain Gelly, and Bernhard Sch"olkopf. Wasserstein Auto-Encoders. In *International Conference on Learning Representations*, 2018. URL https://arxiv.org/abs/1711.01558.

A. W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, 1998.

Joshua D. Welch, Velina Kozareva, Ashley Ferreira, Charles Vanderburg, Carly Martin, and Evan Z. Macosko. Single-Cell Multi-omic Integration Compares and Contrasts Features of Brain Cell Identity. *Cell*, 2019. doi: 10.1016/j.cell.2019.05.006.

Feng Wu, Tsai Hor Chan, Fuying Wang, Guosheng Yin, and Lequan Yu. Cross-Modal Alignment via Variational Copula Modelling. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 67128–67146. PMLR, 2025. URL https://proceedings.mlr.press/v267/wu25d.html.

Yuelei Zhang, Wenxuan Ming, et al. Robust integration and annotation of single-cell and spatial omics data using interpretable gene programs. *Cell Genomics*, 2026. doi: 10.1016/j.xgen.2025.101105.

Yifan Zhao, Huiyu Cai, Zuobai Zhang, Qianxi Tang, and Yue Li. Learning interpretable cellular and gene signature embeddings from single-cell transcriptomic data. *Nature Communications*, 2021. doi: 10.1038/s41467-021-25534-2.

Qin Zhu, Zuzhi Jiang, et al. Revealing a coherent cell-state landscape across single-cell datasets with CONCORD. *Nature Biotechnology*, 2026. doi: 10.1038/s41587-025-02950-z.

A NOTATION

Bold symbols denote vectors or matrices; hats denote learned estimates. Superscripts obs and enc distinguish captured biological distributions from encoder-output distributions. Outputs are program coordinates or factors; biological interpretation requires the evidence in Appendix G.

Symbol	Meaning
B, N_b, G, K	Batches, cells per batch, genes, and fitted coordinate capacity.
$b_n, \mathcal{I}_b, \mathbf{x}_n, \ell_n$	Cell batch, batch index set, count vector, and library size.
Z_{nk}, F_k; $\widehat{\mathbf{z}}_n$	Unknown biological state and CDF; final learned coordinates.
$\mathcal{E}_b^{\text{tech}}$	Technology-specific encoder.
$\mathcal{P}$	Shared affine projection within the program bottleneck.
$\mathcal{D}_b^{\text{tech}}$	Technology-specific observation decoder.
$\phi, \mathbf{e}_k, \mathbf{W}^{\text{prog}}$	Nonlinear branch of the shared gene-space decoder, program embedding, and shared linear basis.
$\mathbf{u}_n, \boldsymbol{\eta}_n, \mu_{ng}, \nu_{bg}$	Shared gene-space feature vector, logits, count mean, and NB inverse dispersion.
$P_b^{\text{enc}}, P_{bk}^{\text{enc}}$	Joint encoded distribution and its coordinate marginal.
$\widehat{F}_{bk}^{\text{enc}}, \widehat{Q}_{bk}^{\text{enc}}, \overline{Q}_k$	Empirical CDF, quantile, and batch-weighted quantile barycenter.
α_b, M, H, C	Batch weight, quantile levels, sampled cells per batch, coordinate bound.
$\mathcal{C}_b, \rho, \widetilde{\rho}, \widetilde{R}_b$	Copula, exact midrank, smooth rank, and smooth-rank vector distribution.
v_k, V, τ_V	Coordinate variance, total variance, and its floor threshold.
$\Pi_b^{\text{obs}}, \mathsf{M}_b, c_b(s)$	Captured-state distribution, observation kernel, and capture probability.
$\pi, \tau_U, \varepsilon_U$	Transport measure, mass penalty, and entropic regularization.
$r_{nbkg}, \pi_{\text{match}}$	Full-decoder response and global cross-fit factor assignment.

Remaining dimensions and numerical constants are defined with their formulas. K denotes capacity, not a recovered biological program count.

B MATHEMATICAL DETAILS AND BOUNDARIES

B.1 OBSERVATION, CAPTURE, AND SEMANTIC AMBIGUITY

For a captured-state distribution Π_b^{obs} and observation kernel M_b,

$$P(\mathbf{x} \in A \mid b) = \int \mathsf{M}_b(A \mid \mathbf{z}) \Pi_b^{\text{obs}}(d\mathbf{z}). \tag{7}$$

Captured-population comparability and a common entire observation kernel imply batch-independent observations. Technology-dependent kernels permit different count distributions. Technical covariates belong to the kernel. With source density $p(s)$ and capture/retention probability $c_b(s)$, conditioning gives

$$p_b^{\text{obs}}(s) = \frac{c_b(s)p(s)}{\int c_b(t)p(t)\,dt}. \tag{8}$$

Thus proportional capture functions preserve the normalized population, whereas state-dependent differences need not. Selection and measurement are not identified from unrestricted unpaired counts: for $S \sim \text{Bernoulli}(1/2)$, $c_b(1) = q_b$, $c_b(0) = 1 - q_b$, and noiseless $X = S$, captured observations follow $\text{Bernoulli}(q_b)$. State-independent capture with an observation kernel returning that same Bernoulli law independently of S produces identical data.

Proposition 1 (Distribution-preserving ambiguity)**.** *Suppose encoders $\mathcal{E}_b$ and decoders $\mathcal{D}_b$ have a common encoded joint law Π. For invertible measurable maps $\mathcal{H}_b$ with $(\mathcal{H}_b)_{\#}\Pi = \Pi$, set $\mathcal{E}'_b = \mathcal{H}_b \circ \mathcal{E}_b$ and $\mathcal{D}'_b = \mathcal{D}_b \circ \mathcal{H}_b^{-1}$. Reconstruction and encoded distribution matching are unchanged.*

Proof. $\mathcal{D}'_b \circ \mathcal{E}'_b = \mathcal{D}_b \circ \mathcal{E}_b$, and for the observed law P_b, $(\mathcal{E}'_b)_{\#}P_b = (\mathcal{H}_b)_{\#}\Pi = \Pi$. □

These transformations need not preserve coordinate semantics (Shrestha & Fu, 2024). Shared modules restrict the parameterization but do not prove that every ambiguity is removed.

B.2 MARGINAL TRANSPORT AND DEPENDENCE

For joint laws P, Q with finite second moments, every joint coupling projects to feasible marginal couplings. Hence

$$W_2^2(P, Q) \geq \sum_k W_2^2(P_k, Q_k). \tag{9}$$

Indeed, the joint squared Euclidean cost is the sum of projected coordinate costs, each bounded below by its marginal optimum. For product laws, the product of optimal marginal couplings attains equality; independence is not assumed by our model. Equal mass on $(-1,-1),(1,1)$ versus $(-1,1),(1,-1)$ gives zero marginal discrepancy but joint squared-W_2 equal to four.

For continuous encoded marginals, the joint CDF decomposes as

$$F_b^{\text{joint}}(\mathbf{t}) = \mathcal{C}_b(F_{b1}^{\text{enc}}(t_1), \ldots, F_{bK}^{\text{enc}}(t_K)). \tag{10}$$

The marginal CDF transform has uniform marginals. Equal marginal CDFs and equal transformed joint laws recover the same encoded law through common generalized inverses. A characteristic-kernel population MMD distinguishes these transformed laws (Gretton et al., 2012). Atoms make copulas nonunique outside the marginal CDF ranges; empirical midranks are discrete and tie-dependent. Finite-bandwidth smooth ranks add approximation and scale dependence. This population decomposition does not provide either a finite-sample training guarantee or semantic identification.

B.3 QUANTILE BARYCENTER AND NORMALIZATION

For $q_b = \widehat{Q}_{bk}^{\text{enc}}(p)$ and $\bar{q} = \sum_b \alpha_b q_b$, expansion yields

$$\sum_b \alpha_b (q_b - q)^2 = \sum_b \alpha_b (q_b - \bar{q})^2 + (q - \bar{q})^2, \tag{11}$$

$$\sum_b \alpha_b (q_b - \bar{q})^2 = \sum_{a<b} \alpha_a \alpha_b (q_a - q_b)^2. \tag{12}$$

Thus $\bar{q}$ minimizes the weighted squared error and is nondecreasing in p. Integrating gives the one-dimensional squared-W_2 barycenter loss; the grid in Equation (3) approximates that integral. Its population version satisfies

$$\mathcal{L}_{\text{prog}}^{\text{pop}} = \frac{1}{K} \sum_{a<b} \alpha_a \alpha_b \sum_k W_2^2(P_{ak}^{\text{enc}}, P_{bk}^{\text{enc}}) \leq \frac{1}{K} \sum_{a<b} \alpha_a \alpha_b W_2^2(P_a^{\text{enc}}, P_b^{\text{enc}}). \tag{13}$$

A finite quantile grid is not necessarily an exact population lower bound. Uniform sampling of distinct batch pairs requires multiplying their mean squared quantile discrepancy by $(B-1)/(2B)$ to obtain the full uniform-weight normalization.

B.4 SCALE AND MASS-RELAXATION LIMITS

Rescaling coordinates by $c_0 > 0$ and compensating in the decoder preserves reconstruction whenever the function family permits it, but scales $\mathcal{L}_{\text{prog}}$ and V by c_0^2. Exact ranks remain unchanged; finite-bandwidth smooth ranks need not. The soft floor tends to the finite value τ_V^2 at collapse, which can be stationary because variance gradients vanish at identical coordinates. It does not force deactivation, remove redundancy, or select model order. In particular,

$$V = \sum_b \alpha_b \operatorname{tr}(\mathbf{\Sigma}_b) + \sum_b \alpha_b \|\mathbf{m}_b - \overline{\mathbf{m}}\|_2^2, \tag{14}$$

where $\mathbf{m}_b, \mathbf{\Sigma}_b$ are within-batch moments and $\overline{\mathbf{m}} = \sum_b \alpha_b \mathbf{m}_b$. Batch offsets or noise can satisfy the floor.

Unbalanced transport changes the matching preference without resolving Equation (8)'s ambiguity. Its unsmoothed cost admits the zero plan with value $2\tau_U$ for probability inputs, so small cost need not mean preserved biology. Per-factor plans need not select common cells. Neither source frequencies nor missing states are recovered by allowing unmatched mass.

B.5 HELD-OUT ESTIMATION FOR FROZEN ENCODERS

Proposition 2 (Fixed-encoder sampling bound)**.** *Fix encoders independently of the evaluation sample. For N_b independent evaluation cells per batch and $0 < \delta < 1$, simultaneously over all BK combinations with probability at least $1 - \delta$,*

$$\sup_t |\widehat{F}^{\mathrm{enc}}_{bk}(t) - F^{\mathrm{enc}}_{bk}(t)| \le \epsilon_b = \sqrt{\frac{\log(2BK/\delta)}{2N_b}}. \tag{15}$$

If coordinates lie in $[-C, C]$, the same event implies

$$W_1(\widehat{F}^{\mathrm{enc}}_{bk}, F^{\mathrm{enc}}_{bk}) \le 2C\epsilon_b, \qquad W_2(\widehat{F}^{\mathrm{enc}}_{bk}, F^{\mathrm{enc}}_{bk}) \le 2C\sqrt{\epsilon_b}. \tag{16}$$

Proof. The Dvoretzky–Kiefer–Wolfowitz bound is $2\exp(-2N_b\epsilon^2)$ (Massart, 1990). Set it to $\delta/(BK)$ and take a union bound. Integrating the CDF difference over $[-C, C]$ gives $W_1 \le 2C\epsilon_b$; applying $|x - y|^2 \le 2C|x - y|$ to an optimal W_1 coupling gives $W_2^2 \le 2CW_1$. □

Pairwise W_2 estimation error is consequently at most $2C(\sqrt{\epsilon_a} + \sqrt{\epsilon_b})$. Glivenko–Cantelli gives corresponding fixed-distribution consistency (van der Vaart, 1998). These statements concern frozen encoder-output populations, conditional on a sampling unit that justifies independent cells. They do not cover refitted training losses, donor-level uncertainty, dependence estimation, density-mode recovery, or biological calibration.

C IMPLEMENTATION AND OPTIMIZATION

C.1 ARCHITECTURE AND RECONSTRUCTION

The technology-specific encoder combines a shared nonlinear encoder $\mathcal{E}_0 : G \to 256 \to d_h$ with a batch-specific low-rank residual. The shared encoder uses GELU and training dropout 0.1. An affine projection $\mathcal{P} : d_h \to K$ followed by bounded activation forms the shared program bottleneck. The technology-specific encoder and observation decoder are

$$\mathcal{E}^{\mathrm{tech}}_b(\mathbf{y}) = \mathcal{E}_0(\mathbf{y}) + \mathbf{C}^{E,b}\,\mathrm{GELU}(\mathbf{L}^{E,b}\mathbf{y}), \qquad \mathcal{D}^{\mathrm{tech}}_b(\mathbf{u}) = \mathbf{u} + \mathbf{d}_b + \mathbf{C}^{D,b}\,\mathrm{GELU}(\mathbf{L}^{D,b}\mathbf{u}). \tag{17}$$

The input adapter matrices have shape $q \times G$; output matrices have shapes $d_h \times q$ and $G \times q$, respectively. Residual output weights start small and nonzero. Each nonlinear branch of the shared gene-space decoder has widths $(1 + d_e) \to 128 \to d_h$; $\mathbf{W}^{\mathrm{prog}} \in \mathbb{R}^{d_h \times K}$ and $\mathbf{C}^{\mathrm{gene}} \in \mathbb{R}^{G \times d_h}$. The technology-specific observation decoder has a residual that can mix genes and need not be invertible or monotone.

With $\nu_{bg} = \mathrm{softplus}(\omega^{\nu}_{bg}) + 10^{-4}$, the count model and loss are

$$\mu_{ng} = \ell_n\,\mathrm{softmax}(\boldsymbol{\eta}_n)_g, \quad x_{ng} \sim \mathrm{NB}(\mu_{ng}, \nu_{b_n g}), \quad \mathrm{Var}(x_{ng}) = \mu_{ng} + \mu^2_{ng}/\nu_{b_n g}, \tag{18}$$

$$\mathcal{L}_{\mathrm{rec}} = -\sum_b \frac{\alpha_b}{N_b G} \sum_{n \in \mathcal{I}_b, g} \log \mathrm{NB}(x_{ng}; \mu_{ng}, \nu_{bg}). \tag{19}$$

Log-softmax, log-gamma, and log-sum-exp provide stable likelihood computation. $\mathcal{L}_{\mathrm{adapt}} = \sum_{\omega \in \mathcal{S}} \|\omega\|^2_F/|\omega|$ averages squared entries within each of the four adapter matrices and $\mathbf{d}_b$ per batch; dispersion is excluded. Pooled moments use weights α_b/N_b. The variance floor uses current sampled cells, with $\tau_V = 1$ fixed across K; its two-batch nonlinear hinge is a stochastic surrogate, not an unbiased full-population loss.

C.2 QUANTILES AND RANK DEPENDENCE

Each update samples two distinct batches uniformly and H cells per batch, using replacement only when fewer than H training cells are available. Define

$$\widehat{F}^{\mathrm{enc}}_{bk}(t) = N_b^{-1} \sum_{n \in \mathcal{I}_b} \mathbf{1}\{\widehat{z}_{nk} \le t\}, \qquad \widehat{Q}^{\mathrm{enc}}_{bk}(p) = \inf\{t : \widehat{F}^{\mathrm{enc}}_{bk}(t) \ge p\}. \tag{20}$$

Sorted minibatch columns provide midpoint quantiles by the first order statistic reaching each level. Pairwise quantile and dependence losses both receive the factor $(B-1)/(2B)$; mini-batch quantiles remain biased population surrogates. Exact evaluation ranks and the training surrogate are

$$\rho_{nk}^{(b)} = N_b^{-1} \sum_{j \in \mathcal{I}_b} [\mathbf{1}\{\widehat{z}_{jk} < \widehat{z}_{nk}\} + \tfrac{1}{2}\mathbf{1}\{\widehat{z}_{jk} = \widehat{z}_{nk}\}], \tag{21}$$

$$\widetilde{\rho}_{ik}^{(b)} = H^{-1} \sum_{j=1}^{H} \sigma((\widehat{z}_{ik}^{(b)} - \widehat{z}_{jk}^{(b)})/h_r), \quad \sigma(x) = (1+e^{-x})^{-1}. \tag{22}$$

Self-pairs contribute $1/2$; gradients pass through both arguments. For smooth-rank samples $\mathbf{a}_i, \mathbf{b}_j$, the biased empirical MMD includes diagonals:

$$\widehat{\mathrm{MMD}}_\kappa^2 = H^{-2} \sum_{i,j} [\kappa(\mathbf{a}_i, \mathbf{a}_j) + \kappa(\mathbf{b}_i, \mathbf{b}_j) - 2\kappa(\mathbf{a}_i, \mathbf{b}_j)],$$

$$\kappa(\mathbf{a}, \mathbf{b}) = \tfrac{1}{3} \sum_{\sigma_0 \in \{0.1, 0.3, 1\}} \exp[-\|\mathbf{a}-\mathbf{b}\|_2^2/(2K\sigma_0^2)]. \tag{23}$$

Full-latent MMD instead uses raw coordinates and a bandwidth scale frozen from pooled training warm-up distances. Joint Wasserstein uses uniform raw sample masses with squared cost $\|\mathbf{z}_i - \mathbf{z}_j\|_2^2/K$. Each replaces both alignment terms without batchwise standardization.

C.3 Mass-relaxed computation

For finite nonnegative measures, generalized KL is $D_{\mathrm{KL}}(\mu\|\nu) = \int \log(d\mu/d\nu)d\mu - \mu(\Omega) + \nu(\Omega)$ when absolutely continuous, and infinity otherwise. The unsmoothed target and its marginal loss are

$$\mathcal{U}_{\tau_U}(P, Q) = \inf_{\pi \geq 0} \left\{ \int (x-y)^2 d\pi + \tau_U [D_{\mathrm{KL}}(\pi_1\|P) + D_{\mathrm{KL}}(\pi_2\|Q)] \right\}, \tag{24}$$

$$\mathcal{L}_{\mathrm{prog}}^{\mathrm{unbal}} = K^{-1} \sum_k \sum_{a<b} \alpha_a \alpha_b \mathcal{U}_{\tau_U}(\widehat{P}_{ak}^{\mathrm{enc}}, \widehat{P}_{bk}^{\mathrm{enc}}). \tag{25}$$

E6 uses a smoothed target (Chizat et al., 2017) on $M_U = 32$ quantile supports x_i, y_j with masses $w_i = w_j = 1/M_U$:

$$\min_{T \geq 0} \sum_{ij} T_{ij}(x_i - y_j)^2 + a u_U \left[\sum_i d(\sum_j T_{ij}, w_i) + \sum_j d(\sum_i T_{ij}, w_j) \right] + \varepsilon_U \sum_{ij} d(T_{ij}, w_i w_j), \tag{26}$$

where $d(a, w) = a\log(a/w) - a + w$, $\tau_U = 1$, and $\varepsilon_U = 0.05$. The differentiable returned loss includes the full KL term without self-cost subtraction. Log-domain generalized Sinkhorn uses exactly 100 updates with exponent $\tau_U/(\tau_U + \varepsilon_U)$ and no adaptive stopping. The logged last-update residual is a fixed-point residual, not a primal–dual gap. A matched balanced control fixes both marginals, uses exponent one, and retains the same supports, smoothing, and updates. Both disable dependence. Transported mass and solver diagnostics accompany E6; per-factor plans are not common cell couplings.

C.4 Frozen configuration

An epoch approximates one expected pass over the training-cell count under equal batch sampling. All modules train jointly; validation reconstruction selects checkpoints. No-alignment and marginal-only controls use weights $(0, 0)$ and $(1, 0)$. Population mode is fixed by task design, not query mixing. Inference freezes all parameters and returns final coordinates.

D Data and Evaluation Protocols

D.1 Cohorts and partitions

GSE133549 includes Chromium, CEL-Seq2, Drop-seq, Quartz-Seq2, and Smart-seq2. It uses split seed 7 and a 70/15/15 cell split within technology, without donor-level generalization. Its role is

Table 3: Primary settings selected using GSE133549 and frozen before Ding and pancreas evaluation. Simulation-specific H and seed sets are stated with their results.

Setting	Value
K, d_h, d_e, q, C	$64, 128, 16, 16, 6$
H, M, h_r	$512, 128, 0.05$
Optimizer	Adam, learning rate 10^{-3}; gradient-norm cap 10
Training/checkpoint	At most 300 epochs; validation-reconstruction patience 30
Warm-up/ramp	10 epochs of reconstruction/scale/adapters; 20-epoch alignment ramp
$\lambda_{\text{prog}}, \lambda_{\text{dep}}, \lambda_{\text{scale}}, \lambda_{\text{adapt}}$	$1, 3, 0.1, 10^{-3}$; $\tau_V = 1$
Optimizer seeds	7, 19, 37, 73, 101

Table 4: Real-data tasks. All gene counts refer to task-specific retained RNA genes. Fitting receives counts and technical labels only; biological labels and proteins are evaluator-only.

Dataset	Cells	Genes	Partition and role
GSE133549 cellular5	7,013	28,742	Development; 4,907/1,049/1,057 train/validation/test cells.
Ding PBMC	40,582	20,008	PBMC1 fitting/validation; 15,793 PBMC2 query cells.
Pancreas subset	10,963	16,016	Cross-study robustness; within-technology split, 1,650 query cells.
GSE194122 BMMC	90,261	13,953	Donors 1–6: 67,494 training; donor 7: 11,473 validation; donors 8–9: 11,294 test.

exploratory because it informed model configuration. Smart-seq2 full-length reads are not treated as UMI observations.

Ding uses six technologies in fitting and inference. Biological scoring covers the 10,448 externally labelled PBMC2 cells (66.16%) across five technologies; Smart-seq2 has no available test labels. This experimental-round holdout does not establish donor independence. Small training groups remain eligible under the sampling rule in Appendix C.

The pancreas task uses raw-integer-compatible inDrop1–4 and Smart-seq2 batches from the scIB cross-study input. Fractional-count batches are excluded without rounding. Captured-population comparability is unestablished across studies, and evaluation uses within-technology cell holdout.

BMMC supplies 134 ADTs for evaluation. Its technical label is acquisition site 1–4, rather than assay technology. Validation donor 7 is at site 3; test donors 8–9 are both at site 4, also represented by donor 1 in training. Thus this RNA-to-protein task holds out donors while retaining partial donor/site confounding. Its partition differs from the source benchmark's original split.

D.2 Preprocessing and information access

Each task uses the intersection of genuinely assayed genes and detection in at least ten training cells, without HVG selection. Zero-library cells are excluded; missing assays are not encoded as biological zeros. Count-based models receive integer counts, separately from their encoder normalization. Every learned preprocessing transform uses training cells. All methods within a track share gene order, partitions, and labelled evaluation masks. Validation selects declared checkpoints and probe settings; test labels never update representations. Native LIGER's joint unlabelled query preprocessing defines a separate access track (Appendix E).

D.3 Captured-population comparability

An independent donor or preparation is the biological replication unit. Shared source material alone does not establish captured-population comparability, which is also unidentifiable from unrestricted unpaired counts. Table 5 compares author-labelled proportions using $\mathrm{TV}(p_a, p_b) = \frac{1}{2}\sum_c |p_a(c) - p_b(c)|$ over unchanged within-cohort labels. Missing labels are excluded from composition estimates. Cellular5 coverage is complete; labelled Ding technologies cover 52.1–97.7% of PBMC1 and 52.4–100% of PBMC2. Smart-seq2 is excluded from this label-based comparison.

Table 5: Cross-technology composition differences based on author annotations. TV is the total-variation distance between labelled cell-type proportions; zero denotes identical proportions. Each range includes all ten labelled technology pairs. Bounds for the largest observed pair are conservative simultaneous 95% intervals, conditional on independent cell sampling and fixed annotations. The comparison concerns labelled composition within these cohorts; latent program-state distributions and donor-level uncertainty are separate questions.

Cohort	Cells / labelled	Pairwise TV range	Largest-TV pair (95% bounds)
GSE133549	7,013 / 7,013	0.056–0.389	Quartz-Seq2 / Smart-seq2 [0.199, 0.637]
Ding PBMC1	24,789 / 15,748	0.076–0.371	CEL-Seq2 / Seq-Well [0.133, 0.675]
Ding PBMC2	15,793 / 10,448	0.057–0.312	CEL-Seq2 / inDrops [0.084, 0.584]

The simultaneous bounds propagate Bonferroni-adjusted Clopper–Pearson intervals for technology/category probabilities through TV. They condition on independent sampled cells and fixed annotations. The observed composition differences may reflect selection, annotation sensitivity, or biological sampling; even equal coarse proportions would not establish captured-population comparability.

D.4 Uncertainty units

Following Section 4, real-data means and sample SDs summarize five optimizer runs on fixed splits, with scCRAFT's partial RNG reset noted below. PCA, cNMF, and LIGER single fits receive no invented seed uncertainty. Simulations use independent generators as replicate populations, averaging optimizer repeats within generator when available. Their paired Student t intervals are descriptive and unadjusted for multiple endpoints. Protein prediction retains its fixed donor split; dependent factor, assay, and protein rows are not independent biological replicates.

E Baseline Configurations and Access

All common-input comparisons fit RNA-only representations on the declared training cells. Table 6 records versions, native transformations, output, and budgets; complete frozen configurations accompany the artifact. Results are consolidated by dataset in Appendix F. Native schedules differ, so external comparisons are fixed-configuration context rather than equal-budget searches.

Table 6: Baseline settings. Output dimension is 64 except scCRAFT's native 256-dimensional sample projected by training-fitted PCA. "Inductive" means queries do not update the representation or its fitted preprocessing.

Method	Version/commit	Input	Output	Training budget	Query access
PCA	Fixed task control	Train-fitted normalized RNA	64	One fit	Inductive
CONCORD	1.0.8 / 7c4f3c7	Native log normalization	64	15 epochs	Inductive
scCRAFT	02ed693	Native common-input workflow	256→64	150 epochs	Inductive
scETM	0.5.0	Raw counts	64 topics	500 epochs	Inductive
scVI	scvi-tools 1.3.3	Raw counts	64 mean	Native early stopping	Inductive
DRVI	0.2.7	Native count workflow	64	Native early stopping	Inductive
LDVAE	scvi-tools 1.4.3	Raw counts	64 factors	Native early stopping	Inductive
pmVAE	e7f6335	Log-CPM; Reactome prior	64 mean	40 epochs, CPU	Inductive
cNMF	1.7.1	Native count/TPM workflow	64 usages	One consensus, 20 restarts	Frozen NNLS
LIGER	rliger 2.2.1	Native online iNMF	64 factors	At most 5 epochs	Joint query preprocessing

CONCORD uses batch size 256, learning rate 0.01, hidden width 1,000, element/feature masking 0.4/0.3, temperature 0.4, and contrastive weight 1. scVI and DRVI export native coordinates; LDVAE also exports linear loadings. Early-stopping validation remains within the designated fitting pool. scETM exports its native `delta`. cNMF uses all eligible genes without batch labels; query usages are nonnegative least-squares projections onto frozen training consensus spectra, then row-normalized. Its 20 internal restarts form one consensus fit rather than independent repeats.

Randomness and dimension. scCRAFT exports sampled latent coordinates rather than a posterior mean. Its implementation resets RNG before BatchNorm initialization and within training epochs; caller seeds still affect linear initialization and export. Five caller-seed runs therefore have partially shared random streams. The native-256 versus PCA64 sensitivity remains in Appendix F.

Extra pathway information. pmVAE uses 15 Reactome v4 modules chosen by coverage before inspecting results. Gencode v50 maps 28,084 of the 28,742 GSE133549 genes to symbols. Four dimensions per module plus four auxiliary dimensions yield 64 coordinates, not 64 pathways. It uses MSE training and posterior-mean export. Reactome is a fitting prior, so enrichment against the same resource is not independent evidence.

Query access. LIGER uses native online iNMF (Gao et al., 2021), with Matrix 1.7.6, $K = 64$, $\lambda = 5$, and minibatches of 500. Validation/test counts enter `runOnlineINMF` with `projection=TRUE` as jointly preprocessed unlabelled datasets. Its results form a separate track; neither native settings nor output selection use query labels. The same-backbone E3 controls, rather than cross-family rankings, isolate the alignment-object claim.

F Supplementary Experiments

This appendix provides protocols and numerical evidence for the known-state mechanism, real integration, alignment-object controls, semantic audit, and captured-population comparability. The statistical convention in Section 4 applies throughout; Appendix D specifies task-level units and exceptions.

F.1 E1: known-state mechanism

Each generator independently samples cells across technologies. A three-component mixture with frozen probabilities generates eight states via

$$z^\star_{nk} = 2\tanh(m^\star_{c_n k} + \epsilon^\star_{nk}), \quad m^\star_{ck} \in \{-1, 0, 1\}, \quad \epsilon^\star_{nk} \sim \mathcal{N}(0, 0.3^2). \tag{27}$$

A sparse signed loading matrix $\mathbf{A}^\star$ has 50 nonzero genes per factor, 20% overlap, and unit column norms. For $\mathbf{s}_n = \mathbf{A}^\star \mathbf{z}^\star_n$, technology-specific logits are

$$\boldsymbol{\eta}^\star_n = \mathbf{s}_n + \gamma \mathbf{d}^\star_b + \gamma \mathbf{B}^\star_b \tanh(\mathbf{a}^\star_b \odot \mathbf{s}_n) + 0.2\gamma \sin(\mathbf{r}^\star_b \odot \mathbf{s}_n). \tag{28}$$

Here $\mathbf{d}^\star_b$ is centered, the sparse mixing matrix $\mathbf{B}^\star_b$ has spectral norm one, and $\mathbf{a}^\star_b, \mathbf{r}^\star_b$ are positive scales. Technology-shifted log-library sizes give $\boldsymbol{\mu}^\star_n = \ell^\star_n \operatorname{softmax}(\boldsymbol{\eta}^\star_n)$ and NB inverse dispersions in $\{2, 10, 50\}$. Four severities $\gamma \in \{0, 0.5, 1, 2\}$ preserve captured-population comparability; severity zero retains depth and dispersion differences. Branching provides a separate non-clustered population. Frozen generator arrays and draws accompany the artifact. Truth is evaluator-only, and the simulator differs from the fitted decoder. Each technology contributes 2,048 cells with eight true factors and 1,000 genes. The $K = 64, H = 128$ internal design crosses generators 1101/2213/3307/4421/5501 with optimizers 7/19/37/61/101 (375 fits). Optimizer averages define five generator units. Paired 95% t intervals for combined-minus-none factor correlation are [0.0772, 0.1033], [0.0412, 0.0804], [0.0299, 0.0703], and [0.0216, 0.0645] across severities 0/0.5/1/2. Corresponding kNN-15 recall changes are $-0.00689/-0.00387/-0.00386/-0.00673$ and NB NLL changes $+0.00234/+0.00250/+0.00189/+0.00164$. Branching gives a correlation change of -0.0167 with interval $[-0.0486, 0.0151]$. At severity 2 the transfer gain is 0.103 [0.021, 0.185]; lower-severity intervals include zero. These costs prevent a claim of uniformly improved biological preservation.

The external comparison uses three generators and one optimizer seed (45 external fits). Table 8 retains the matched internal slice. DRVI has the highest mean factor correlation at severity zero and in branching, and the highest neighborhood recall throughout. scETM has the highest factor correlation at severity two; combined alignment has the highest at severities 0.5 and 1. External full-decoder semantic scores are unavailable, so these comparisons concern recovery and utility rather than identification of a unique biological program basis.

Table 7: E1 internal means: five optimizer runs are averaged within each of five generators. Measurement severity and branching are separate settings. Lower factor MSE and technology BA are preferable; higher factor correlation and transfer BA are preferable. Paired generator intervals and geometry/reconstruction costs are stated in the text.

Setting	Arm	Factor corr.	Factor MSE	Transfer BA	Tech. BA
measurement 0	No alignment	0.594	0.641	0.964	0.834
measurement 0	Marginal	0.640	0.583	0.994	0.715
measurement 0	Marginal + dependence	0.685	0.496	0.984	0.597
measurement 0.5	No alignment	0.625	0.604	0.963	0.820
measurement 0.5	Marginal	0.661	0.550	0.996	0.755
measurement 0.5	Marginal + dependence	0.686	0.494	0.988	0.620
measurement 1	No alignment	0.638	0.591	0.985	0.827
measurement 1	Marginal	0.656	0.565	0.997	0.738
measurement 1	Marginal + dependence	0.688	0.490	0.990	0.650
measurement 2	No alignment	0.671	0.549	0.891	0.825
measurement 2	Marginal	0.658	0.563	0.996	0.725
measurement 2	Marginal + dependence	0.714	0.462	0.994	0.704
branching 1	No alignment	0.520	0.622	0.642	0.816
branching 1	Marginal	0.557	0.576	0.654	0.741
branching 1	Marginal + dependence	0.503	0.504	0.659	0.663

F.2 E2: REAL INTEGRATION AND PROBE SENSITIVITY

Tables 9 and 10 consolidate the complete nine-metric profiles. GSE133549 uses the development split; Ding and pancreas use frozen settings. Primary-model ARI differences from no alignment are 0.0553 ± 0.0559 on Ding and 0.1122 ± 0.0556 on pancreas, while F1 changes are -0.0027 ± 0.0400 and 0.0006 ± 0.0283. These paired seed differences illustrate the clustering/prediction trade-off without donor-level inference. On Ding, the proposed model's F1 difference from CONCORD is -0.0572 ± 0.0297, reversing the development ordering. On pancreas, biological ASW decreases relative to marginals alone in every paired seed. Strong prediction controls remain competitive.

Probe and output controls. The linear logistic and secondary RBF-SVC probes use scikit-learn 1.7.2, training-fitted StandardScaler, no class weighting, and validation-selected $C \in \{0.01, 0.1, 1, 10\}$, with ties choosing the first ascending value. RBF bandwidth uses the `scale` rule on transformed labelled training cells. No representation is refitted. Table 11 retains the mean-order reversal against CONCORD under RBF-SVC; changing both classifier loss and kernel makes this a whole-protocol sensitivity. Table 12 compares scCRAFT's native sampled 256 coordinates with its training-PCA64 export under identical probe access. Neither sensitivity selects the primary model or output from test performance.

Ding query coverage. A batch-only diagnostic includes all 15,793 inferred query cells, including 273 unlabelled Smart-seq2 cells. Across five seeds, iLISI/PCR are approximately 0/0 without alignment, $0.212 \pm 0.034/0.924 \pm 0.017$ for marginals, and $0.198 \pm 0.020/0.907 \pm 0.030$ for combined alignment; scCRAFT gives $0.277 \pm 0.010/0.870 \pm 0.013$ and CONCORD $0.195 \pm 0.007/0.830 \pm 0.029$. This technical coverage diagnostic does not establish biological preservation on cells without labels.

F.3 ALIGNMENT-OBJECT CONTROLS (E3)

All six arms share the frozen backbone, total-variance scale rule, GSE133549 development split, $K = 64$, $H = 512$, $M = 128$, a 300-epoch cap, and patience 30. Seeds 7/19/37/73/101 are paired across objectives; no training technology needs replacement sampling. Only the alignment target changes, without objective-specific retuning. Table 13 includes the mixing endpoints omitted from the compact main-text table. The earlier single-seed $H = 128$ comparison remains separate in the experiment records.

Table 8: E1 external context with matched internal arms: means ± sample SD across generators 1101/2213/3307 and optimizer seed 7. This slice is distinct from the internal five-by-five aggregate. Severity zero retains depth/dispersion differences; branching is a separate population. Native configurations do not form an equal-budget search.

Severity	Method	Factor corr.	kNN-15 recall	Transfer BA
0	No alignment	0.570 ± 0.036	0.078 ± 0.001	0.994 ± 0.002
0	Marginals	0.612 ± 0.013	0.083 ± 0.002	0.995 ± 0.001
0	Combined	0.687 ± 0.003	0.076 ± 0.004	0.983 ± 0.001
0	scVI	0.368 ± 0.031	0.060 ± 0.004	0.990 ± 0.003
0	scETM	0.717 ± 0.008	0.082 ± 0.002	0.996 ± 0.005
0	DRVI	0.858 ± 0.026	0.086 ± 0.007	0.999 ± 0.001
0.5	No alignment	0.603 ± 0.017	0.082 ± 0.004	0.996 ± 0.001
0.5	Marginals	0.660 ± 0.032	0.088 ± 0.009	0.995 ± 0.002
0.5	Combined	0.695 ± 0.003	0.079 ± 0.003	0.988 ± 0.002
0.5	scVI	0.302 ± 0.019	0.056 ± 0.009	0.990 ± 0.006
0.5	scETM	0.620 ± 0.023	0.078 ± 0.003	0.977 ± 0.020
0.5	DRVI	0.627 ± 0.197	0.105 ± 0.011	1.000 ± 0.000
1	No alignment	0.604 ± 0.016	0.086 ± 0.002	0.992 ± 0.008
1	Marginals	0.626 ± 0.024	0.089 ± 0.001	0.994 ± 0.006
1	Combined	0.687 ± 0.032	0.085 ± 0.005	0.988 ± 0.003
1	scVI	0.328 ± 0.034	0.064 ± 0.006	0.996 ± 0.004
1	scETM	0.680 ± 0.016	0.080 ± 0.005	0.989 ± 0.011
1	DRVI	0.501 ± 0.091	0.127 ± 0.000	1.000 ± 0.000
2	No alignment	0.656 ± 0.033	0.100 ± 0.006	0.833 ± 0.152
2	Marginals	0.626 ± 0.028	0.095 ± 0.007	0.993 ± 0.006
2	Combined	0.705 ± 0.004	0.086 ± 0.009	0.990 ± 0.007
2	scVI	0.367 ± 0.012	0.069 ± 0.003	0.996 ± 0.003
2	scETM	0.751 ± 0.017	0.089 ± 0.003	0.978 ± 0.019
2	DRVI	0.489 ± 0.049	0.137 ± 0.009	1.000 ± 0.000
Branching	No alignment	0.510 ± 0.007	0.117 ± 0.008	0.646 ± 0.024
Branching	Marginals	0.515 ± 0.023	0.118 ± 0.005	0.665 ± 0.010
Branching	Combined	0.493 ± 0.005	0.113 ± 0.009	0.660 ± 0.028
Branching	scVI	0.396 ± 0.045	0.069 ± 0.008	0.619 ± 0.017
Branching	scETM	0.412 ± 0.040	0.083 ± 0.006	0.517 ± 0.015
Branching	DRVI	0.560 ± 0.094	0.218 ± 0.016	0.622 ± 0.027

Relative to marginal alignment alone, the combined objective's paired ARI and F1 differences are $+0.0431 \pm 0.0295$ and $+0.0145 \pm 0.0130$ (mean ± sample SD), positive in four of five seeds for each metric. Batch ASW and iLISI increase and exact-rank MMD decreases in all five. Biological ASW decreases in all five, with a mean paired difference of -0.0158 ± 0.0085. Relative to combined alignment, full-latent MMD has higher ARI in two seeds and higher batch ASW in four; joint Wasserstein has higher biological ASW and lower test NB in all five.

F.4 Capacity / collapse controls (E3)

Table 14 varies fitted capacity with eight or 20 true factors under the primary total-variance floor. Almost every coordinate remains active, while full-state prediction is non-monotone in K. One generator and optimizer seed provide a descriptive control, not model-order selection. The activity threshold does not rule out coordinate shrinkage compensated by increased decoder sensitivity; the following frozen-model diagnostic addresses that distinction.

Coordinate scale and decoder sensitivity. We additionally diagnose 24 frozen selected checkpoints: the three primary arms on GSE133549 with five optimizer seeds, and the same arms on three measurement-setting generators (1101/2213/3307, severity 1, eight true factors, $K = 64$, optimizer seed 7). No model is refitted. The training implementations did not save a warm-up-end checkpoint, so this analysis compares fitted endpoints rather than reconstructing a training trajectory.

Table 9: Consolidated GSE133549 profiles on 1,057 test cells and 28,742 genes. Five-run entries are mean ± sample SD across caller seeds; one-fit controls have no seed uncertainty. pmVAE adds a pathway prior, and LIGER has separate joint-query access. Neither is pooled into the no-prior inductive ranking. GC denotes graph connectivity.

Method	n	ARI	NMI	Bio ASW	F1
No alignment	5	0.275 ± 0.008	0.507 ± 0.006	0.550 ± 0.008	0.895 ± 0.007
Marginals only	5	0.258 ± 0.021	0.542 ± 0.011	0.608 ± 0.016	0.886 ± 0.009
Combined	5	0.302 ± 0.026	0.565 ± 0.009	0.599 ± 0.020	0.901 ± 0.007
CONCORD	5	0.292 ± 0.034	0.554 ± 0.016	0.569 ± 0.015	0.886 ± 0.012
scCRAFT	5	0.566 ± 0.020	0.670 ± 0.008	0.603 ± 0.003	0.823 ± 0.027
DRVI	5	0.358 ± 0.014	0.605 ± 0.008	0.611 ± 0.004	0.899 ± 0.001
scETM	5	0.276 ± 0.005	0.510 ± 0.014	0.557 ± 0.005	0.893 ± 0.007
LDVAE	5	0.262 ± 0.028	0.508 ± 0.009	0.549 ± 0.005	0.910 ± 0.008
scVI	5	0.301 ± 0.013	0.540 ± 0.010	0.565 ± 0.001	0.915 ± 0.003
cNMF (20 restarts)	1	0.237	0.430	0.489	0.884
PCA	1	0.272	0.508	0.541	0.919
pmVAE (15 modules)	5	0.258 ± 0.005	0.488 ± 0.002	0.527 ± 0.002	0.863 ± 0.011
LIGER (native projection)	1	0.201	0.429	0.529	0.900

Method	Batch ASW	iLISI	cLISI	GC	PCR
No alignment	0.379 ± 0.043	0.008 ± 0.004	0.928 ± 0.011	0.558 ± 0.050	0.005 ± 0.012
Marginals only	0.722 ± 0.025	0.234 ± 0.058	0.958 ± 0.006	0.879 ± 0.037	0.933 ± 0.039
Combined	0.776 ± 0.008	0.374 ± 0.043	0.962 ± 0.005	0.940 ± 0.024	0.928 ± 0.020
CONCORD	0.774 ± 0.024	0.382 ± 0.014	0.954 ± 0.004	0.932 ± 0.012	0.722 ± 0.041
scCRAFT	0.919 ± 0.002	0.546 ± 0.023	0.959 ± 0.005	0.951 ± 0.005	0.937 ± 0.005
DRVI	0.794 ± 0.004	0.407 ± 0.017	0.965 ± 0.004	0.951 ± 0.006	0.862 ± 0.011
scETM	0.496 ± 0.020	0.010 ± 0.010	0.929 ± 0.023	0.711 ± 0.059	0.029 ± 0.033
LDVAE	0.605 ± 0.011	0.037 ± 0.008	0.949 ± 0.007	0.922 ± 0.022	0.349 ± 0.058
scVI	0.691 ± 0.013	0.102 ± 0.027	0.949 ± 0.008	0.965 ± 0.001	0.593 ± 0.022
cNMF (20 restarts)	0.361	0.000	0.888	0.442	0.000
PCA	0.524	0.017	0.913	0.702	0.000
pmVAE (15 modules)	0.593 ± 0.017	0.018 ± 0.008	0.906 ± 0.007	0.717 ± 0.032	0.037 ± 0.048
LIGER (native projection)	0.736	0.163	0.916	0.930	0.692

Table 10: Ding and pancreas profiles: mean ± sample SD across five caller seeds, except fixed PCA. Ding uses 10,448 labelled query cells; pancreas uses 1,650 test cells. Biological ASW is scaled; GC denotes graph connectivity. Native budgets, partial RNG resets, and query masks follow Appendices D and E.

Method	n	ARI	NMI	Bio. ASW	Batch ASW	GC	iLISI	cLISI	PCR	F1
Ding PBMC										
No alignment	5	0.291 ± 0.013	0.531 ± 0.004	0.526 ± 0.004	0.487 ± 0.039	0.599 ± 0.049	0.000 ± 0.000	0.985 ± 0.003	0.000 ± 0.000	0.763 ± 0.016
Marginals only	5	0.245 ± 0.065	0.535 ± 0.042	0.597 ± 0.043	0.691 ± 0.030	0.887 ± 0.052	0.296 ± 0.050	0.989 ± 0.004	0.939 ± 0.015	0.758 ± 0.025
Combined	5	0.346 ± 0.055	0.588 ± 0.032	0.636 ± 0.030	0.755 ± 0.028	0.950 ± 0.027	0.272 ± 0.034	0.992 ± 0.003	0.919 ± 0.026	0.761 ± 0.027
PCA64	1	0.341	0.586	0.549	0.709	0.867	0.000	0.991	0.000	0.827
scCRAFT	5	0.652 ± 0.024	0.712 ± 0.009	0.583 ± 0.004	0.927 ± 0.004	0.945 ± 0.004	0.375 ± 0.017	0.983 ± 0.003	0.893 ± 0.013	0.669 ± 0.011
CONCORD	5	0.381 ± 0.009	0.621 ± 0.006	0.594 ± 0.010	0.828 ± 0.013	0.976 ± 0.002	0.258 ± 0.007	0.994 ± 0.001	0.833 ± 0.031	0.818 ± 0.004
DRVI	5	0.478 ± 0.043	0.671 ± 0.009	0.603 ± 0.007	0.859 ± 0.006	0.990 ± 0.002	0.204 ± 0.012	0.993 ± 0.001	0.657 ± 0.017	0.820 ± 0.012
scETM	5	0.474 ± 0.047	0.634 ± 0.012	0.566 ± 0.005	0.842 ± 0.005	0.982 ± 0.006	0.157 ± 0.009	0.989 ± 0.001	0.618 ± 0.033	0.792 ± 0.008
scVI	5	0.370 ± 0.022	0.614 ± 0.008	0.578 ± 0.006	0.791 ± 0.002	0.992 ± 0.001	0.068 ± 0.028	0.991 ± 0.001	0.564 ± 0.026	0.807 ± 0.005
Pancreas raw-count-compatible subset										
No alignment	5	0.384 ± 0.037	0.703 ± 0.016	0.629 ± 0.024	0.775 ± 0.039	0.968 ± 0.019	0.142 ± 0.103	0.995 ± 0.003	0.469 ± 0.165	0.926 ± 0.029
Marginals only	5	0.428 ± 0.015	0.730 ± 0.008	0.638 ± 0.015	0.826 ± 0.009	0.974 ± 0.002	0.404 ± 0.018	0.997 ± 0.002	0.956 ± 0.005	0.932 ± 0.013
Combined	5	0.496 ± 0.038	0.743 ± 0.014	0.604 ± 0.016	0.857 ± 0.007	0.973 ± 0.004	0.537 ± 0.012	0.995 ± 0.001	0.952 ± 0.002	0.927 ± 0.038
PCA64	1	0.372	0.694	0.523	0.812	0.980	0.062	0.995	0.000	0.976
scCRAFT	5	0.841 ± 0.030	0.853 ± 0.017	0.641 ± 0.002	0.927 ± 0.003	0.969 ± 0.004	0.581 ± 0.006	0.994 ± 0.003	0.908 ± 0.003	0.783 ± 0.033
CONCORD	5	0.372 ± 0.024	0.671 ± 0.011	0.525 ± 0.024	0.831 ± 0.015	0.974 ± 0.004	0.424 ± 0.023	0.991 ± 0.002	0.754 ± 0.036	0.856 ± 0.067
DRVI	5	0.591 ± 0.052	0.772 ± 0.017	0.616 ± 0.011	0.882 ± 0.006	0.984 ± 0.012	0.460 ± 0.011	0.996 ± 0.003	0.897 ± 0.011	0.921 ± 0.042
scETM	5	0.501 ± 0.039	0.753 ± 0.020	0.692 ± 0.010	0.838 ± 0.005	0.969 ± 0.006	0.265 ± 0.021	0.999 ± 0.000	0.501 ± 0.034	0.960 ± 0.005
scVI	5	0.524 ± 0.030	0.732 ± 0.005	0.547 ± 0.005	0.880 ± 0.005	0.976 ± 0.002	0.337 ± 0.022	0.994 ± 0.002	0.685 ± 0.020	0.970 ± 0.008

Let v_k be the equal-technology pooled training variance. Using up to 32 training reference cells per technology, sampled with seed 261337 and shared cell identities across arms, we compute

$$s_{bk}^2 = \sum_{n \in \mathcal{I}_{\text{ref}}} w_n^{\text{ref}} \frac{1}{G} \sum_g r_{nbkg}^2, \qquad s_k = \sqrt{\frac{1}{B} \sum_b s_{bk}^2}, \qquad a_k = \sqrt{v_k}\, s_k, \tag{29}$$

where r_{nbkg} is the full log-mean response from Equation (31). Each decoder receives the same reference coordinates within a fit; weights assign equal mass to each reference technology. Squaring precedes reference-cell averaging, avoiding cancellation of signed derivatives. a_k measures local response per training standard deviation, not biological information. With d_k denoting the coordinate contribution before averaging over K in Equation (3), we report both $K^{-1} \sum_k d_k$ and the mean of d_k / v_k, using 128 midpoint inverse-CDF quantiles. Held-out discrepancies use training variances

Table 11: Probe sensitivity on the same frozen development representations. The linear logistic and RBF-SVC probes share train-fitted standardization, the same four-value C grid, and validation-only probe selection. Five-run values are mean $\pm$ sample standard deviation; PCA is a single fixed fit. The difference column pairs probes on each identical representation. Both the classifier loss and kernel change between protocols.

Representation	n	Linear F1	RBF F1	RBF $-$ linear
No alignment	5	0.895 ± 0.007	0.888 ± 0.007	-0.008 ± 0.009
Marginals only	5	0.886 ± 0.009	0.889 ± 0.011	0.003 ± 0.012
Marginals + copula	5	0.901 ± 0.007	0.890 ± 0.011	-0.011 ± 0.015
PCA, unintegrated	1	0.919	0.913	-0.006
scCRAFT	5	0.823 ± 0.027	0.809 ± 0.034	-0.014 ± 0.010
CONCORD	5	0.886 ± 0.012	0.895 ± 0.009	0.009 ± 0.012
DRVI	5	0.899 ± 0.001	0.907 ± 0.006	0.008 ± 0.006
scVI	5	0.915 ± 0.003	0.921 ± 0.005	0.005 ± 0.004

Table 12: Exploratory scCRAFT output-dimensionality sensitivity without representation refitting. Values are macro-F1 means $\pm$ sample standard deviations across all five caller seeds. The final column uses within-run native-minus-PCA64 differences. Projection has different effects across datasets and probe families; it does not alone explain the lower linear-probe scores. Upstream partial RNG resets and the task-specific labelled masks are unchanged.

Task	Probe	Train-PCA64	Native sampled 256D	Native minus 64D
GSE133549	Linear	0.8229 ± 0.0272	0.8029 ± 0.0226	-0.0200 ± 0.0111
GSE133549	RBF-SVC	0.8089 ± 0.0340	0.8361 ± 0.0301	0.0272 ± 0.0171
Ding PBMC	Linear	0.6691 ± 0.0115	0.6814 ± 0.0153	0.0124 ± 0.0091
Pancreas subset	Linear	0.7830 ± 0.0328	0.8047 ± 0.0308	0.0217 ± 0.0369

shared across technologies; no technology-specific standardization is applied. Variances at most 10^{-12} are excluded from division and reported separately; none occur here, and all 64 coordinates in every fit exceed the existing 10^{-3} activity threshold.

On GSE133549, smaller coordinate variance and greater decoder sensitivity accompany both alignment objectives (Table 15). However, combined-to-none held-out discrepancy ratios are 0.023–0.035 in raw units and 0.114–0.211 after normalization across all five paired seeds. The improvement therefore cannot be explained entirely by pure coordinate rescaling. Cross-coordinate variance–sensitivity Spearman correlations are 0.271 ± 0.163, -0.300 ± 0.209, and -0.019 ± 0.202 for none, marginals, and combined, respectively; corresponding log-variance/log-sensitivity slopes are 0.093 ± 0.049, -0.371 ± 0.257, and -0.018 ± 0.126. The inverse association is clearest for marginal-only alignment and is not consistent across combined-model fits. These endpoint associations do not identify a causal compensation trajectory or quantify the fraction of improvement attributable to scale.

Synthetic combined models instead have larger variance than no-alignment controls. Their raw held-out discrepancy is larger in two of three generators, whereas normalized discrepancy is smaller in two of three; the remaining generator has a combined-to-none normalized ratio of 1.372. Thus scale changes differ across settings and neither raw nor normalized improvement is universal. The absence of inactive coordinates does not exclude scale compensation, and the normalized diagnostic establishes neither biological preservation nor coordinate identifiability.

F.5 Architecture sensitivity (E3)

In the single-generator/seed sensitivity, a roughly matched batch-conditioned encoder raises state R^2 from 0.892 to 0.906 but lowers factor correlation from 0.708 to 0.582. A linear decoder raises factor correlation to 0.806 while increasing technology-prediction accuracy from 0.571 to 0.982. These trade-offs do not select a preferred architecture class; E3's alignment comparison fixes one backbone.

Table 13: Alignment-object comparison on GSE133549, with $H = 512$. All arms share the frozen backbone, split, 300-epoch cap, patience 30, and quantile count 128. Entries are mean ± sample SD across five paired optimizer seeds (7/19/37/73/101). Test NB is the equal-technology mean negative-binomial NLL; rank MMD is a held-out dependence diagnostic.

Objective	ARI ↑	Bio ASW ↑	Batch ASW ↑	iLISI ↑	Rank MMD ↓	F1 ↑	Test NB ↓
No alignment	0.275 ± 0.008	0.550 ± 0.008	0.379 ± 0.043	0.008 ± 0.004	0.0161 ± 0.0020	0.895 ± 0.007	0.897 ± 0.008
Means/variances	0.259 ± 0.018	0.595 ± 0.026	0.695 ± 0.017	0.154 ± 0.052	0.0164 ± 0.0011	0.878 ± 0.012	0.890 ± 0.007
Full-latent MMD	0.288 ± 0.017	0.613 ± 0.011	0.772 ± 0.013	0.439 ± 0.026	0.0079 ± 0.0007	0.889 ± 0.010	0.890 ± 0.005
Joint Wasserstein	0.277 ± 0.018	0.624 ± 0.009	0.704 ± 0.012	0.134 ± 0.051	0.0117 ± 0.0006	0.879 ± 0.009	0.882 ± 0.003
Program marginals	0.258 ± 0.021	0.608 ± 0.016	0.722 ± 0.025	0.234 ± 0.058	0.0150 ± 0.0022	0.886 ± 0.009	0.893 ± 0.005
Marginals + dependence	0.301 ± 0.023	0.592 ± 0.010	0.770 ± 0.013	0.425 ± 0.038	0.0060 ± 0.0003	0.900 ± 0.006	0.888 ± 0.004

Table 14: Combined alignment with the primary total-variance floor at one generator (1101) and optimizer seed (7). Coverage is matched truths divided by true factor count; inactive coordinates have training variance below 10^{-3}. Factor correlation assigns zero to unmatched truths; state R^2 uses all fitted coordinates.

K^*	K	Factor corr.	State R^2	Coverage	Inactive
8	16	0.684	0.902	1.00	0/16
8	32	0.701	0.905	1.00	0/32
8	64	0.708	0.892	1.00	0/64
8	128	0.714	0.891	1.00	1/128
20	16	0.585	0.920	0.80	0/16
20	32	0.638	0.922	1.00	0/32
20	64	0.632	0.918	1.00	0/64
20	128	0.644	0.904	1.00	0/128

F.6 E4: FIXED-ENCODER SAMPLING

Fixed-encoder CDF sampling and rare-state coverage. Synthetic CDF comparisons use a frozen encoder and a separate 65,536-cell reference per technology. The finite-reference allowance is 0.00839, added to sample allowances 0.1899/0.09496/0.04748/0.02374 at 128/512/2048/8192 sampled cells. All 20 draws fall within the combined allowance at each size and rare-state fraction. Yet at a 0.1% rare-state fraction the state is absent in 16/20 draws at 128 cells per technology and 6/20 at 512. The CDF observation is conditional on this generated source and fixed encoder; it does not establish training or prediction reliability for a rare state. A separate GSE133549 fixed-training-pool bootstrap gives empirical coverage 0.95, 0.94, 0.95, and 0.96 at the same four index-draw sizes. Its 100 draws per size sample with replacement from fixed technology pools, so they do not represent new donors or 8192 distinct biological cells.

F.7 E5: RESPONSE STABILITY AND INDEPENDENT PROTEIN EVIDENCE

GSE133549. Five fits use seeds 7/19/37/73/101, 28,742 genes, and five actual assays. Full-decoder log-mean derivatives use the same 160 training-cell identities (32 per assay), with model-specific inferred states. A Hungarian match based on absolute Pearson correlation of assay-pooled profiles aligns each later fit to seed 7 with one global sign per factor, reused across assays. Table 16 scores the same training reference: its rows are dependent, and near-perfect within-fit agreement is constrained by the shared dictionary.

For disjoint test-state scoring, the training match remains fixed. Seven states satisfy at least ten test cells overall and five in two assays; at most eight cells per assay/state are sampled, and only supported assay/state comparisons are scored. Dendritic cells lack two-assay support. Across 8,448 dependent rows, median correlation/Jaccard are 0.574/0.020; within-fit cross-assay medians are

Table 15: Frozen-model scale diagnostics. Values are means and sample standard deviations over five optimizer seeds on GSE133549 or three synthetic generators, separately. Variance, sensitivity and one-standard-deviation response are coordinate medians within each fit; discrepancies average coordinates.

Data	Alignment	Raw test d	Test d/v	Median v	Median s	Median a
GSE133549	None	0.4759 ± 0.0854	0.3431 ± 0.0787	1.2476 ± 0.3978	0.1179 ± 0.0216	0.1332 ± 0.0445
	Marginals	0.0163 ± 0.0023	0.0601 ± 0.0163	0.2880 ± 0.0963	0.3018 ± 0.0668	0.1658 ± 0.0594
	Combined	0.0135 ± 0.0014	0.0511 ± 0.0091	0.2701 ± 0.0634	0.2913 ± 0.0366	0.1487 ± 0.0300
Synthetic	None	0.000630 ± 0.000691	0.02973 ± 0.02892	0.01680 ± 0.00305	0.1989 ± 0.0108	0.02512 ± 0.00073
	Marginals	0.000077 ± 0.000031	0.00584 ± 0.00233	0.01319 ± 0.00034	0.2149 ± 0.0087	0.02436 ± 0.00107
	Combined	0.001397 ± 0.001748	0.01341 ± 0.01610	0.09823 ± 0.00721	0.2717 ± 0.0043	0.08517 ± 0.00211

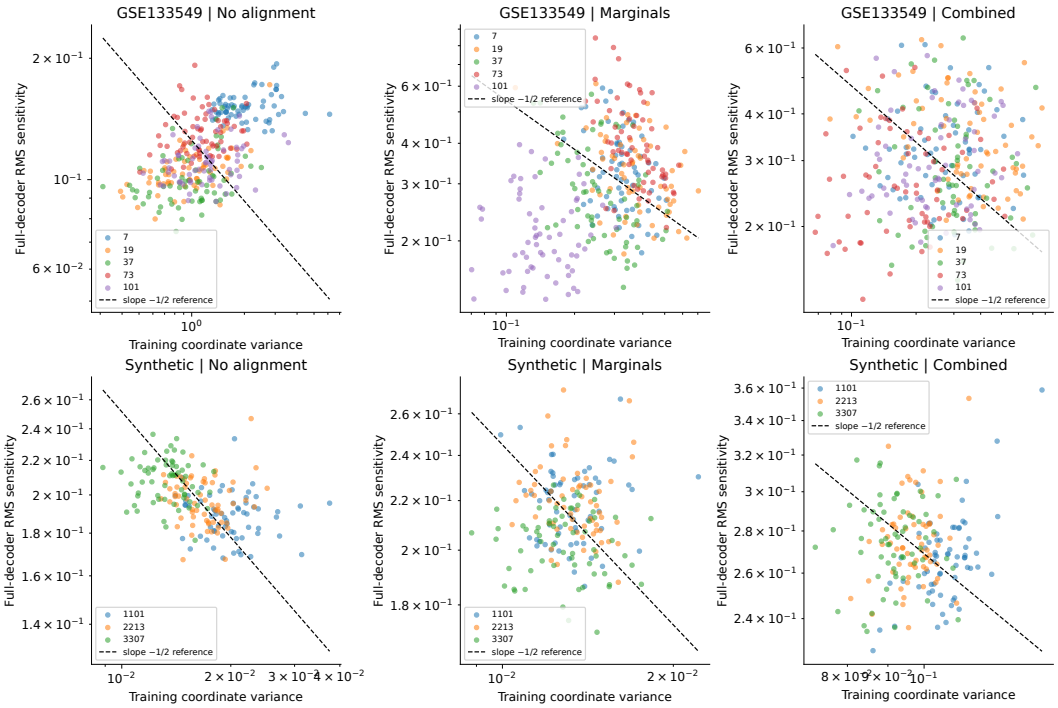

Figure 4: Coordinate variance versus full-decoder RMS sensitivity for the frozen fits in Table 15. Points are dependent coordinates within fits; colors distinguish optimizer seeds (GSE133549) or generators (synthetic). Dashed lines show the slope $-1/2$ expected under pure rescaling with decoder compensation, with reference rather than fitted intercepts.

1.000/1.000 over 19,840 dependent rows. Table 17 retains the state breakdown. This cohort supplies no orthogonal factor-level biological test.

BMMC stability. No-alignment and combined arms each use five fits on the same donor split. One site-pooled training-response match and sign per seed pair are reused across all four sites. Table 18 distinguishes response means over four seed pairs from activation medians over dependent factor/site rows. Better response stability coexists with poorer activation stability; sites are not independent assay protocols.

Protein prediction and factor eligibility. Training donors 1–6 fit RNA-only representations and ridge probes; donor 7 selects penalties, and donors 8–9 test 134 ADTs. Table 2 reports median-across-protein performance. Combined alignment improves R^2 over no alignment and PCA, while Spearman is close to PCA; within-type correlation remains small (median 0.071). Both test donors are at site 4, which is also represented in training.

Table 16: Full-decoder response stability on the five-assay GSE133549 task. Correlations retain signs after one training-derived global factor assignment and sign per fit; top-50 genes use absolute response magnitude. Matching and scoring share 160 training reference cells. Row counts are dependent factor-by-assay comparisons.

Comparison	Rows	Median correlation	Median top-50 Jaccard
Across optimizer seeds, within assay	1,280	0.575	0.020
Across assays, within fitted model	3,200	1.000	1.000

Table 17: E5 GSE133549 held-out state-stratified full-decoder response stability across five actual assays. The same training-derived factor matching is reused in every state. Values are medians over dependent seed/factor/assay rows with at least five held-out source cells for that assay and state.

Annotated test state	Test cells	Signed correlation	Top-50 Jaccard
B cells	80	0.577	0.010
CD14+ monocytes	128	0.575	0.020
CD4 T cells	395	0.575	0.010
CD8 T cells	107	0.574	0.020
FCGR3A+ monocytes	35	0.575	0.020
HEK cells	212	0.574	0.020
NK cells	87	0.574	0.020

Before protein evaluation, factor eligibility required median signed response correlation at least 0.5 and top-50 Jaccard at least 0.2, both across seeds within site and across sites within fit, plus variance at least 0.001 in three of five fits. One no-alignment factor and no combined factors pass. A validation-selected ADT must additionally show absolute within-type Spearman at least 0.1 in validation and 0.05 in each test donor, with consistent sign, in at least three fits. Requiring the same ADT identity across those matched fits leaves no candidates in either arm. These descriptive, non-multiplicity-adjusted filters distinguish whole-embedding prediction from stable factor evidence. Table 2 in the main text presents the protein-prediction comparison. Figure 5 brings together the full response/activity audit and prediction results.

F.8 E6: CAPTURED-POPULATION VIOLATIONS

Figure 6 summarizes the three assumption-violation settings discussed in RQ4; the complete numerical results and capture-strength effects follow below. Two equally frequent source states undergo capture with fixed observation kernels:

$$c_1(\mathrm{A}) = c_2(\mathrm{B}) = 0.6 + 0.3\gamma_c, \qquad c_1(\mathrm{B}) = c_2(\mathrm{A}) = 0.6 - 0.3\gamma_c, \quad \gamma_c \in \{0, 0.5, 1\}. \tag{30}$$

At strength one the expected proportions are $0.75{:}0.25$ and $0.25{:}0.75$. Every method receives the same realized captured cells within each replicate; capture probabilities and truth are evaluator-only. Separate missing-state and equal-marginal/different-dependence scenarios test other violations. In the latter, dependence differences are biological effects to preserve.

The five internal arms retain one backbone and scale rule: none, marginals, combined, matched balanced entropy, and matched unbalanced entropy. The last two share numerical settings and disable dependence, isolating mass relaxation. Generators 2213/3307/4421 use optimizer seed 7, eight true factors, $K = 64, H = 128$, cap 300 and patience 30. The core and capture curve each contain 45 fits; their strength-one block overlaps and is not independent replication.

Paired effects and preservation. Combined-minus-none factor-correlation differences across strengths 0/0.5/1 are $+0.204$ [0.136, 0.272], $+0.042$ [-0.060, 0.144], and -0.156 [-0.186, -0.126], with descriptive paired 95% t intervals over three generators. The effect changes relative to strength zero are -0.162 [-0.315, -0.009] and -0.360 [-0.417, -0.303]. Transfer intervals span zero at every strength; the intermediate no-alignment mean is particularly variable (0.864 ± 0.169). These points do not identify a universal threshold.

Table 18: BMMC response/activity diagnostics on a fixed donor split. Response statistics are mean $\pm$ sample SD over four seed-7 comparison pairs; activation medians use 1,024 dependent factor-by-site rows per arm. Redundancy is the median number of highly correlated pairs among 2,016 pairs per fit. Collection sites are not assay technologies.

Endpoint	No alignment	Marginals + dependence
Across-seed signed response correlation	0.282 ± 0.006	0.505 ± 0.019
Across-seed top-50 response Jaccard	0.050 ± 0.003	0.076 ± 0.007
Matched activation correlation (median)	0.210	0.108
Active coordinates in each fit	64/64	64/64
Activation pairs with $\lvert r\rvert \geq 0.8$ (median count)	4	0
Response pairs with $\lvert r\rvert \geq 0.8$ (median count)	0	1
Factors passing response/activity filters, before ADT	1/64	0/64

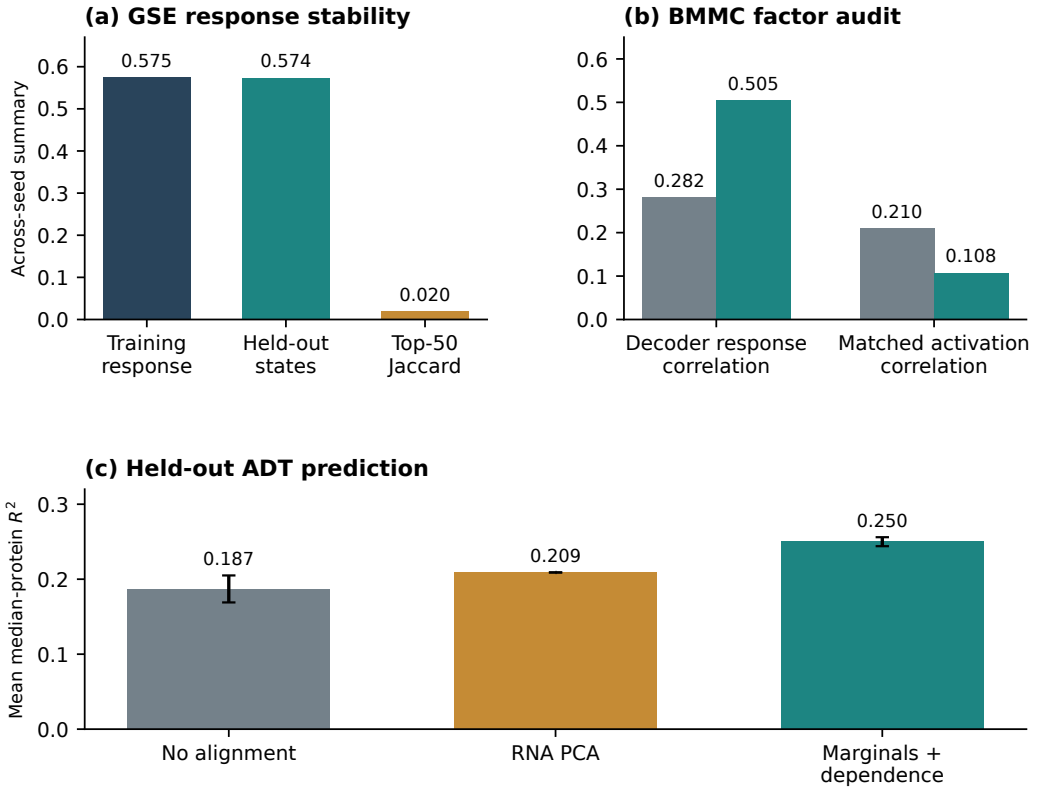

Figure 5: Program-coordinate stability and protein prediction. (a) GSE133549: median signed decoder-response correlation and top-50 gene Jaccard across fits; held-out states reuse the training-derived match. (b) BMMC: response correlation averaged over four seed pairs, activation correlation as a median over dependent factor/site rows, and counts passing response/activity filters. (c) Prediction of 134 ADTs on donors 8–9: mean median-protein R^2 and sample SD across five optimizer seeds; PCA is one fixed control.

Unbalanced-minus-balanced transfer changes under capture and missing states are 0.029 [-0.078, 0.136] and 0.063 [-0.170, 0.296]; both arms remain below no alignment. All capture-strength intervals for their factor-correlation and transfer contrasts span zero. Under genuine dependence change, effect-norm ratios are 1.1735 without alignment, 0.8981 for marginals, 0.0955 for combined, and 0.8574 for unbalanced. Combined and unbalanced direction cosines are 0.2456 and -0.1008, so magnitude retention alone is insufficient. Missing-state false assignments are zero except marginals alone, averaging one third of a query cell; this does not imply recovery of an absent state.

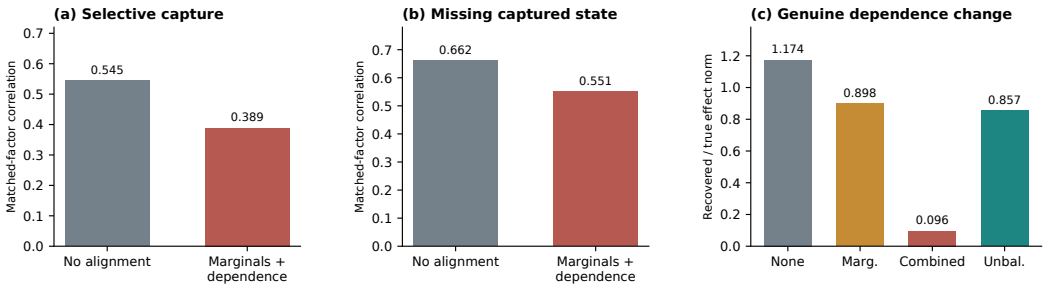

Figure 6: Violations of captured-population comparability. Bars show means over three independent generators with optimizer seed 7. (a–b) Matched-factor correlation under selective capture and a missing captured state. (c) Recovered-to-true correlation-effect norm under genuine dependence change; direction is assessed separately in the text. "Marg. denotes marginal alignment, "Combined adds dependence alignment, and "Unbal. uses unbalanced transport with dependence alignment disabled.

Table 19: E6 internal matrix: means across generators 2213/3307/4421 at optimizer seed 7. Strength-one capture is reported once. Balanced/unbalanced entropic arms share 32 supports, smoothing 0.05, 100 updates, and disabled dependence. Dashes denote endpoints absent from these source summary tables; they are not zeros. Technology BA and composition TV measure different objectives.

Scenario	Arm	Factor corr.	Transfer BA	Comp. TV	Tech. BA	NB NLL
γ_c=0	No alignment	0.518	0.993	0.008	–	2.374
γ_c=0	Marginal	0.619	0.992	0.005	–	2.375
γ_c=0	Combined	0.722	0.990	0.004	–	2.377
γ_c=0	Balanced entropy	0.469	0.997	0.002	–	2.387
γ_c=0	Unbalanced entropy	0.428	0.994	0.003	–	2.386
γ_c=0.5	No alignment	0.544	0.864	0.009	–	2.383
γ_c=0.5	Marginal	0.523	0.977	0.011	–	2.384
γ_c=0.5	Combined	0.586	0.969	0.007	–	2.385
γ_c=0.5	Balanced entropy	0.366	0.900	0.015	–	2.386
γ_c=0.5	Unbalanced entropy	0.326	0.948	0.005	–	2.388
γ_c=1	No alignment	0.545	0.990	0.005	0.880	2.372
γ_c=1	Marginal	0.383	0.852	0.032	0.778	2.373
γ_c=1	Combined	0.389	0.888	0.011	0.837	2.376
γ_c=1	Balanced entropy	0.402	0.844	0.024	0.769	2.375
γ_c=1	Unbalanced entropy	0.414	0.873	0.017	0.814	2.375
missing state	No alignment	0.662	0.827	0.004	0.939	–
missing state	Marginal	0.559	0.821	0.003	0.852	–
missing state	Combined	0.551	0.822	0.004	0.791	–
missing state	balanced entropy	0.459	0.713	0.010	0.866	–
missing state	unbalanced entropy	0.447	0.776	0.008	0.853	–
dependence	No alignment	0.592	0.994	0.005	0.811	–
dependence	Marginal	0.603	0.994	0.005	0.736	–
dependence	Combined	0.682	0.986	0.006	0.532	–
dependence	balanced entropy	0.507	0.989	0.006	0.621	–
dependence	unbalanced entropy	0.530	0.994	0.004	0.646	–

Numerical scope. Unbalanced mean transported mass in capture/missing/dependence fits is 0.997713/0.997603/0.996930. Balanced mass is approximately one, with mean maximum marginal residual 0.000113/0.000104/0.001572. Fixed solver iterations are not assumed converged, and mass is not an estimate of capture probability. The comparison supplies no reliable biological rescue.

External configurations and numerical failures. scVI succeeds on all six native fits. scCRAFT succeeds on three of six; the others produce nonfinite training values. A predetermined rescue changes only its generator Adam learning rate from 10^{-3} to 3×10^{-4} and succeeds in all three

Table 20: E6 external capture results. Native rows report means ± sample SD over successful generators out of three attempts; absent SD means one success. Rescue rows are individual fits with a reduced scCRAFT generator learning rate, kept separate from native averages.

Method/setting	γ_c	Successes or generator	Factor corr.	kNN-15	Transfer BA
scCRAFT	0	1/3	0.167	0.045	0.969
scCRAFT	1	2/3	0.149 ± 0.006	0.047 ± 0.002	0.982 ± 0.004
scVI	0	3/3	0.254 ± 0.032	0.040 ± 0.003	0.991 ± 0.003
scVI	1	3/3	0.350 ± 0.057	0.041 ± 0.002	0.883 ± 0.032
scCRAFT rescue	0	2213	0.172	0.051	0.981
scCRAFT rescue	0	3307	0.155	0.044	0.956
scCRAFT rescue	1	4421	0.137	0.051	0.985

cases. Numerical failure triggers rescue, not query performance. Native successful subsets differ between strengths and cannot estimate a paired capture effect. Table 20 keeps all rescues separate. One optimizer seed leaves optimization variability unmeasured; full-decoder semantic scores are unavailable for these adapters.

G Evaluation Definitions

Standard metrics and probes. The common evaluator uses scIB 1.1.7 with Leiden resolution 1.0, arithmetic-normalized NMI, and biological ASW on $[0, 1]$. Batch ASW averages $1-|s_n^{\mathrm{batch}|c}|$ within eligible classes and then across classes. ARI, graph connectivity, iLISI, cLISI, and PCR follow the named implementation; PCR uses training-fitted PCA as reference. Logistic/RBF probes use scikit-learn 1.7.2 and the access/selection rules in Appendix F. Protein ridge regression fits training-donor feature scaling and an unpenalized intercept, selects penalties on validation donors, and reports R^2 and Spearman on held-out donors. Negative R^2 is retained; constant targets are undefined. Labels and ADTs enter evaluation only.

Alignment diagnostics. D_{prog} uses Equation (3) on frozen held-out coordinates. D_{dep} uses Equation (4) with exact midranks (Equation (21)), separately from the smooth training surrogate. Raw joint MMD/transport are distinct diagnostics with fixed sample sizes and kernels. Small discrepancy alone neither identifies semantics nor implies biological preservation, particularly under selection shift.

Full-decoder gene responses. For a training-derived common reference-state set $\mathcal{I}_{\mathrm{ref}}$ with normalized batch-balanced weights w_n^{ref}, define

$$r_{nbkg} = \left.\frac{\partial \log \mu_g^{(b)}(\mathbf{z}, \ell_n)}{\partial z_k}\right|_{\mathbf{z}=\widehat{\mathbf{z}}_n}, \qquad \overline{r}_{bkg}^{\mathrm{sgn}} = \sum_{n\in\mathcal{I}_{\mathrm{ref}}} w_n^{\mathrm{ref}} r_{nbkg}, \quad \overline{r}_{bkg}^{\mathrm{mag}} = \sum_{n\in\mathcal{I}_{\mathrm{ref}}} w_n^{\mathrm{ref}} |r_{nbkg}|. \tag{31}$$

Library size and other coordinates remain fixed; the derivative includes the technology-specific observation decoder and softmax. Every observation decoder receives the same reference states. These responses describe model sensitivity, not causal regulation.

One global Hungarian assignment and at most one sign per factor are fitted between seeds using pooled training responses and reused across technologies and test states. Top-50 genes use absolute response; Jaccard is intersection size divided by union size. Signed response correlation, activation correlation, and redundant high-correlation factor pairs remain distinct. Zero-response or zero-variance correlations are undefined. Appendix F gives the response/activity and independent ADT eligibility thresholds, including same-ADT consistency across fits. Shared architecture or training-prior enrichment cannot replace independent factor evidence.

Capacity, known-state recovery, and selection. Inactive factors have batch-balanced training variance $v_k < 10^{-3}\tau_V$; V and its within/between-batch components are distinct from useful biological dimension. Known-state evaluation fits one global assignment and coordinatewise affine

calibration using training truth, then freezes both across test technologies. Matched-factor recovery is separate from true-state neighborhood recall and whole-state prediction.

E6 targets the actual post-capture population. Composition TV uses one training-derived state assignment; false assignments to absent states and rare-state recall measure different errors. For genuine dependence changes, the effect-norm ratio measures retained correlation-effect magnitude and the effect-direction cosine measures directional agreement. Transported mass and marginal residuals diagnose the solver rather than capture probabilities. Biological preservation is not inferred from mixing alone. All uncertainty retains the generator, optimizer, donor, or dependent-row units stated with each result.