# HIERARCHICAL WEAK SUPERVISION FOR MULTI-LEVEL REPRESENTATION LEARNING WITH SCPACE

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Scientific and clinical datasets often consist of groups of exchangeable instances, such as thousands of cells sampled from a patient or donor. Yet information is unevenly distributed across levels: integer-valued molecular counts are measured for individual cells, whereas phenotypes such as age, diagnosis, or disease severity are typically measured only for the donor. Cell-level phenotypes and cell-type labels may be sparse or missing. Although modern generative models represent single-cell count data, they do not typically capture how donor-level biological variation manifests in cellular states, cell-type composition, and gene expression. We introduce scPACE, a hierarchical generative model that represents each donor by an underlying biological state and each cell by its position along a related continuum, such as developmental pseudotime, cellular maturity, or disease-associated progression. This structure captures how donor-level variation is expressed heterogeneously across cells. Under full observation and the stated conditional-independence assumptions for phenotype proxies, the latent state hierarchy is identifiable up to a sign convention. We apply scPACE to large-scale real-world single-cell studies of mouse cortical development, Alzheimer's, and Huntington's disease. scPACE provides a principled framework for modeling phenotype-linked variation across donors and cells, enabling the study of development and disease progression across biological scales.

## 1 INTRODUCTION

Standard deep representation learning typically assumes independent and identically distributed observations. In many scientific and clinical domains, however, data naturally arrive as groups or exchangeable sets. For example, document corpora are composed of sentences, speech sessions of acoustic frames, and samples from brain donors of thousands of single cells (Zaheer et al., 2017; Edwards & Storkey, 2017; Bouchacourt et al., 2018). Such grouped data exhibit statistical variation across multiple resolutions. A local representation must capture instance-level heterogeneity, a group representation must summarize an unordered collection of instances of varying size, and both levels must remain statistically coupled. Furthermore, real-world groups often carry supervision targets at two levels: global targets describe the entire group (e.g. patient age, disease stage, or diagnosis), and local targets describe the individual instances (e.g. cell maturity, somatic mutation burden, or activation state). In single-cell studies, integer-valued molecular counts are measured for individual cells, whereas phenotypes such as age, disease stage, or diagnosis are typically measured only for the donor. Cell-level phenotypes and cell-type labels may therefore be sparse or missing.

Learning from such complex structured data presents multiple machine learning challenges. First, assigning a group phenotype to every instance creates pseudoreplication and obscures within-group heterogeneity, whereas ignoring instance-level targets discards informative local variation. Second, instance-level phenotypes are often sparse, noisy, or unavailable, requiring models to exploit them when present while remaining able to infer instance-level states when they are absent at test time. Third, instance classes, such as cell types, may be only partially observed, requiring a model that works in semi-supervised settings. Finally, biologically interpretable latent states require identifiable structure: without suitable inductive biases or supervision, the coordinates of nonlinear latent-variable models are not uniquely determined (Locatello et al., 2019; Khemakhem et al., 2020; Brehmer et al., 2022). Hierarchical latent-variable models provide a natural probabilistic framework for grouped data. Structured variational autoencoders combine neural networks with graphi-



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.

cal models (Johnson et al., 2016), while Neural Statistician, ML-VAE, and FHVAE separate shared group-level variation from instance-specific variation (Edwards & Storkey, 2017; Bouchacourt et al., 2018; Hsu et al., 2017). Deep Sets, Set Transformers, and attention-based multiple-instance learning provide permutation-invariant representations of variable-sized groups (Zaheer et al., 2017; Lee et al., 2019; Ilse et al., 2018). However, these approaches generally do not jointly connect donor phenotypes, cell-level states, cell-type composition, and molecular counts in a generative model whose donor- and cell-level states admit an explicit identifiability result.

Here, we introduce scPACE (Personalized Autoencoder for Cell Evolution), a hierarchical generative model for grouped count single-cell data. scPACE represents each donor by an underlying biological state and each cell by its position along a related biological continuum, such as developmental pseudotime, cellular maturity, or disease-associated progression. The model connects donor phenotypes, cellular states, cell-type composition, and gene expression within a single probabilistic framework. Donor- and cell-level phenotypes are modeled as noisy observations of their corresponding states, while amortized inference estimates these states from counts when phenotypes are unavailable. The framework also accommodates incomplete cell-type labels. Because phenotype targets never enter the inference network, both states remain predictable when those targets are unobserved. Under explicit assumptions, the latent state hierarchy is identifiable up to a sign convention (Appendix A).

We evaluate scPACE under held-out-donor splits in four settings across three large-scale real-world single-cell studies comprising more than 1.3 million cells. A mouse visual-cortex cohort links chronological age to cellular maturation (Gao et al., 2025). An Alzheimer's disease cohort uses a donor-level pathology progression score without cell-level supervision (Gabitto et al., 2024). Lastly, a Huntington's disease cohort pairs donor-level disease burden with cell-specific somatic $HTT$ CAG repeat length (Handsaker et al., 2025; Warner et al., 2022). Together, these studies show that scPACE recovers phenotype-linked variation across donors and cells, generalizes to unseen donors, and outperforms existing methods that can only address subsets of these tasks.

## 2 Related Work

**Hierarchical generative models represent variation across multiple scales.** Variational autoencoders (Kingma & Welling, 2014) provide scalable amortized inference, while structured VAEs (Johnson et al., 2016) combine neural networks with probabilistic graphical models. For exchangeable groups, Neural Statistician (Edwards & Storkey, 2017) and ML-VAE (Bouchacourt et al., 2018) separate shared group-level factors from instance-specific variation. Deep hierarchical VAEs, including Ladder VAE (Sønderby et al., 2016), NVAE (Vahdat & Kautz, 2020), and VDVAE (Child, 2021), further demonstrate the ability of latent hierarchies to capture multi-scale structure. scPACE builds on this tradition by linking an interpretable donor state to related cell-level states within a generative model of molecular counts.

**Permutation-invariant aggregation enables fixed-dimensional representations of unordered, variable-size groups.** Deep Sets (Zaheer et al., 2017) introduced a general architecture based on instance-wise transformations and symmetric pooling, while Set Transformer (Lee et al., 2019) uses attention to capture interactions among instances. Attention-based deep MIL (Ilse et al., 2018) learns adaptive weights for constructing group representations. scPACE extends this approach by pooling cells separately within each class and combining the resulting summaries with empirical class proportions to infer the donor state.

**Phenotype supervision can anchor latent states that would otherwise be non-identifiable.** Unsupervised disentanglement is generally impossible without additional assumptions or information (Locatello et al., 2019). Approaches such as iVAE (Khemakhem et al., 2020) and weakly supervised causal representation learning (Brehmer et al., 2022) use auxiliary variables or paired observations to constrain latent representations. scPACE uses donor-level and, when available, cell-level phenotype channels to anchor its scalar hierarchy.

**Single-cell generative models and trajectory methods address complementary parts of multi-scale biological variation.** Count-based VAEs such as scVI (Lopez et al., 2018) model gene counts and technical effects, while scANVI (Xu et al., 2021) adds semi-supervised cell-type inference. Multi-group methods, including multiGroupVI (Weinberger et al., 2022), scPoli (De Donno et al., 2023), and MrVI (Boyeau et al., 2025), learn sample-specific representations, whereas scVIP (Lai

et al., 2026) connects single-cell profiles to donor phenotypes through attention-based MIL. Separately, Slingshot (Street et al., 2018), PhenoPath (Campbell & Yau, 2018), psupertime (Macnair et al., 2022), and Sceptic (Li et al., 2025) estimate cell-level trajectories, while CNA (Reshef et al., 2022) and MELD (Burkhardt et al., 2021) associate cellular neighborhoods with sample conditions. To our knowledge, these approaches do not jointly model donor phenotypes, cell-level continua, cellular composition, and molecular counts within a single hierarchical generative framework.

## 3 METHOD

scPACE (Personalized Autoencoder for Cell Evolution) is a hierarchical generative model for grouped count data, fitted by amortized variational inference (Figure 1). Each donor is represented by a scalar biological state $t_d$, and each cell by a related scalar state $s_n$. The donor state generates the donor phenotype, cell-type composition, and the distribution of cell states. The cell state generates an optional cell-level phenotype and, together with cell type and additional cellular variation, determines the latent representation from which molecular counts are generated. Phenotypes are excluded from the inference networks, allowing the latent states to be estimated when phenotype measurements are unavailable at deployment.

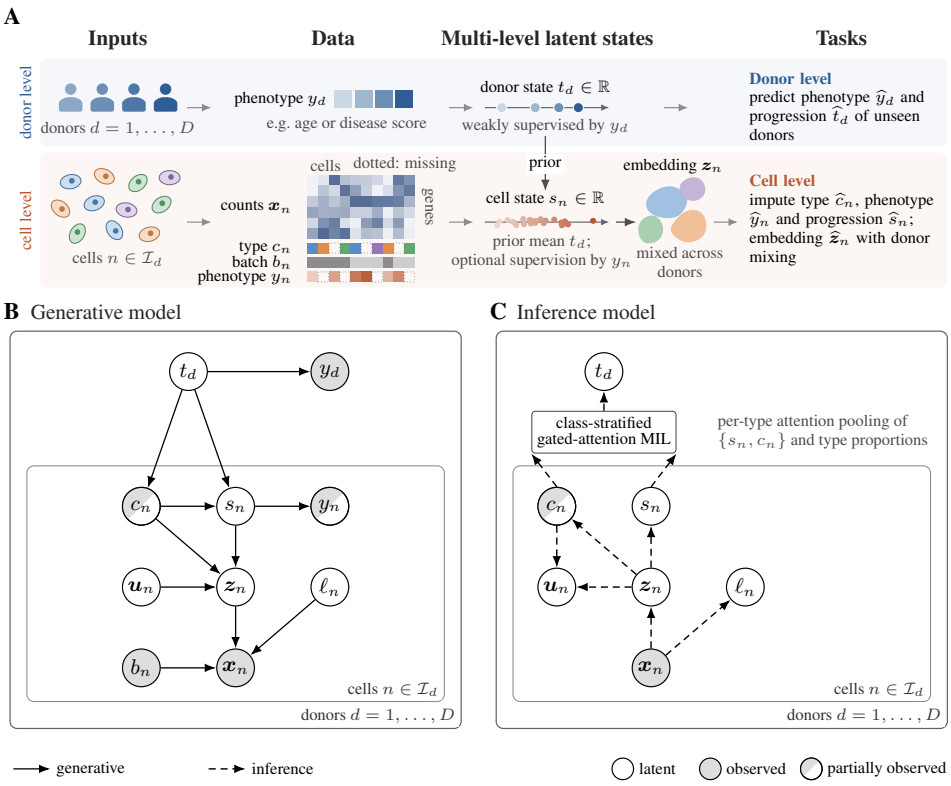



Figure 1: **Overview of scPACE.** (A) scPACE models grouped single-cell counts with donor phenotypes and partially observed cell annotations, inferring donor states $t_d$, cell states $s_n$, and embeddings $\boldsymbol{z}_n$ for phenotype prediction and cell-type imputation. (B) The hierarchical generative model links donor and cell states to cell types, phenotypes, and molecular counts. Solid arrows denote generative dependencies. Donor and optional cell phenotypes are generated through affine Gaussian observation models. (C) Amortized inference estimates the latent states, with class-stratified MIL pooling cells within observed or predicted types to infer $t_d$. Dashed arrows denote inference pathways. Shaded and half-shaded nodes indicate fully and partially observed variables.

### 3.1 PROBLEM FORMULATION

We study hierarchical representation learning from grouped single-cell count data with incomplete cell-level annotations, donor-level phenotype supervision, and optional cell-level phenotype supervision. Donor $d \in \{1, \dots, D\}$ contains an unordered set $\mathcal{I}_d$ of $N_d$ cells. Each cell has a count vector $\boldsymbol{x}_n \in \mathbb{Z}_{\geq 0}^G$, an observed batch label $b_n \in \{1, \dots, B\}$, and a potentially unobserved cell type $c_n \in \{1, \dots, K\}$. Each donor has a phenotype $y_d$, and a subset $\mathcal{Y}_d \subseteq \mathcal{I}_d$ may additionally have cell-level phenotypes $y_n$. We denote cells with observed and unobserved cell types by $\mathcal{H}_d$ and $\mathcal{U}_d$, respectively. Missingness patterns and batch labels are treated as fixed, observed design information. Our objective is to learn donor states, cell states, cell representations, and cell-type probabilities that generalize to unseen donors without using their phenotype measurements as inference inputs.

### 3.2 HIERARCHICAL DEEP GENERATIVE MODELING FOR GROUPED SINGLE-CELL DATA

Conditional on the observed batch labels, the generative process is

$$
\begin{aligned}
& t_d \sim \mathcal{N}(0,1), \quad y_d \mid t_d \sim \mathcal{N}\left(\delta_D + a_D t_d, \tau_D^2\right), \\
& c_n \mid t_d \sim Cat\left(\boldsymbol{\pi}(t_d)\right), \qquad \boldsymbol{\pi}(t_d) = \mathrm{softmax}(\boldsymbol{\alpha} + t_d \boldsymbol{\beta}), \\
& s_n \mid t_d, c_n = k \sim \mathcal{N}\left(t_d + m_k, \sigma_{s,k}^2\right), \qquad y_n \mid s_n \sim \mathcal{N}\left(\delta_C + a_C s_n, \tau_C^2\right), \\
& \boldsymbol{u}_n \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \qquad \ell_n \sim \mathcal{N}(\mu_\ell, v_\ell), \\
& \boldsymbol{z}_n \mid \boldsymbol{u}_n, c_n, s_n \sim \mathcal{N}\left(\boldsymbol{\mu}_\theta^p([\boldsymbol{u}_n; \mathrm{onehot}(c_n); s_n]), \mathrm{diag}\{(\boldsymbol{\sigma}_\theta^p)^2\}\right), \\
& \boldsymbol{\rho}_n = \mathrm{softmax}(g_\theta([\boldsymbol{z}_n; \mathrm{onehot}(b_n)])), \\
& x_{ng} \mid \boldsymbol{z}_n, \ell_n, b_n \sim NB(e^{\ell_n} \rho_{ng}, \theta_g).
\end{aligned} \tag{1}
$$

Here, $\boldsymbol{\alpha}, \boldsymbol{\beta} \in \mathbb{R}^K$ parameterize baseline cell-type composition and its dependence on donor state. The bridge $p(s_n \mid t_d, c_n)$ places the cell states of type $k$ around the donor state shifted by a type-specific intercept $m_k \in \mathbb{R}$, with type-specific variance $\sigma_{s,k}^2$. The intercepts absorb the different baseline positions of the cell types along the continuum (e.g. progenitors and neurons), so a change in the donor state moves the cell states of every type together. The auxiliary variable $\boldsymbol{u}_n \in \mathbb{R}^{d_u}$ captures cellular variation not represented by the scalar state or cell type. The observational model includes library size, batch effects, and gene-specific count overdispersion. We treat $\tau_D^2$ and $\tau_C^2$ as fixed observation variances as they determine the strength of the corresponding likelihood contributions.

Let $\boldsymbol{b}_d = \{b_n : n \in \mathcal{I}_d\}$. The joint density of the latent variables and observed data for donor $d$ factorizes as

$$
\begin{aligned}
& p_\theta(t_d, y_d, \{c_n, s_n, \boldsymbol{u}_n, \boldsymbol{z}_n, \ell_n, \boldsymbol{x}_n\}_{n \in \mathcal{I}_d}, \{y_n\}_{n \in \mathcal{Y}_d} \mid \boldsymbol{b}_d) = \\
& \quad p(t_d)\, p_\theta(y_d \mid t_d) \prod_{n \in \mathcal{I}_d} \Big[ p_\theta(c_n \mid t_d) p_\theta(s_n \mid t_d, c_n) p(\boldsymbol{u}_n) p_\theta(\boldsymbol{z}_n \mid \boldsymbol{u}_n, c_n, s_n) p(\ell_n) \\
& \quad p_\theta(\boldsymbol{x}_n \mid \boldsymbol{z}_n, \ell_n, b_n) \Big] \prod_{n \in \mathcal{Y}_d} p_\theta(y_n \mid s_n).
\end{aligned} \tag{2}
$$

Unobserved cell types and phenotypes are marginalized from this joint model. Under the assumptions stated in Appendix A, including fully observed cell types and independent phenotype channels, the latent hierarchy is identifiable up to a common sign convention.

### 3.3 AMORTIZED INFERENCE AND CLASS-STRATIFIED ATTENTION MIL

The variational proposal for the model in Eq. 2 is

$$
\begin{aligned}
q_\phi = q_\phi(t_d \mid \boldsymbol{h}_d) \prod_{n \in \mathcal{I}_d} & q_\phi(\boldsymbol{z}_n \mid \boldsymbol{x}_n)\, q_\phi(\ell_n \mid \boldsymbol{x}_n)\, q_\phi(s_n \mid \boldsymbol{z}_n) \\
& \times \prod_{n \in \mathcal{H}_d} q_\phi(\boldsymbol{u}_n \mid \boldsymbol{z}_n, c_n) \prod_{n \in \mathcal{U}_d} q_\phi(c_n \mid \boldsymbol{z}_n)\, q_\phi(\boldsymbol{u}_n \mid \boldsymbol{z}_n, c_n).
\end{aligned} \tag{3}
$$

All continuous factors are diagonal Gaussians parameterized by neural networks, and $q_\phi(c_n \mid \boldsymbol{z}_n)$ is categorical. Neither donor nor cell phenotypes are provided to the inference networks, allowing the states to be inferred when phenotype measurements are unavailable. To infer the donor state, scPACE summarizes cell states separately within each cell type. Each state is mapped to a type-specific feature vector $\boldsymbol{v}_{nk} = \mathrm{ReLU}(W_k s_n + \boldsymbol{b}_k) \in \mathbb{R}^{32}$. Let $\mathcal{P}_{dk}$ denote the cells assigned to type $k$ in donor $d$. Observed labels are used when available, and otherwise the assignment is obtained from $q_\phi(c_n \mid \boldsymbol{z}_n)$. Gated attention gives

$$e_{nk} = \boldsymbol{w}_k^\top \left[\tanh(V_k \boldsymbol{v}_{nk} + \boldsymbol{b}_k^V) \odot \sigma(U_k \boldsymbol{v}_{nk} + \boldsymbol{b}_k^U)\right] + b_k^a,$$
$$a_{nk} = \frac{\exp(e_{nk})}{\sum_{j \in \mathcal{P}_{dk}} \exp(e_{jk})}, \qquad \boldsymbol{h}_{dk} = \sum_{n \in \mathcal{P}_{dk}} a_{nk} \boldsymbol{v}_{nk}. \tag{4}$$

The type-specific summaries are concatenated with the empirical cell-type proportions $\boldsymbol{p}_d \in \Delta^{K-1}$: $\boldsymbol{h}_d = [\boldsymbol{h}_{d1}; \ldots; \boldsymbol{h}_{dK}; \boldsymbol{p}_d]$. Two neural networks map $\boldsymbol{h}_d$ to the mean and variance of $q_\phi(t_d \mid \boldsymbol{h}_d)$. The resulting donor representation is invariant to cell order and has fixed dimension for donors with different numbers of cells.

The scPACE's generative model defines the full-data evidence lower bound (ELBO) below and reported training uses the within-donor cellwise averaging described in Appendix D. Let $\boldsymbol{o}_d$ collect the observed variables for donor $d$, including the counts, donor phenotype, observed cell types, and observed cell phenotypes, and let $\boldsymbol{\xi}_d$ collect all latent variables, including unobserved cell types. Batch labels are treated as fixed covariates, missing phenotypes contribute no likelihood factor, and unobserved cell types are marginalized through the variational distribution. For cells with observed types, we additionally supervise $q_\phi(c_n \mid \boldsymbol{z}_n)$. The complete expansion of the ELBO and the supervised classification term are provided in Appendix B. The ELBO is

$$\mathcal{L}_{\mathrm{ELBO}}(\theta, \phi) = \sum_{d=1}^{D} \mathbb{E}_{q_\phi(\boldsymbol{\xi}_d)} \left[\log p_\theta(\boldsymbol{o}_d, \boldsymbol{\xi}_d \mid \boldsymbol{b}_d) - \log q_\phi(\boldsymbol{\xi}_d)\right]. \tag{5}$$

# 4 Experiments

We evaluate scPACE in four held-out-donor settings across three single-cell cohorts. The experiments ask whether scPACE can (i) recover donor- and cell-level phenotype variation in unseen donors, (ii) use cell-level supervision when available, and (iii) reveal biologically meaningful cellular variation when only donor phenotypes are observed. Each setting partitions donors into training, validation, and query groups stratified by donor phenotype. Query donors are excluded from model fitting, gene filtering, decoder-batch definition, target standardization, and calibration. Detailed cohort construction, target definitions, and model configurations are provided in Appendices D to F.

The cell-level targets differ across cohorts. The mouse maturity score and the SEAAD disease-associated microglia (DAM) score are expression-derived summaries of marker programs (Appendix E). They are reproducible proxies for maturation and for the disease-associated microglial state, and they are not independent measurements of the underlying biology. We use them as evaluation targets, and in Setting 2 as an auxiliary supervision signal. The Huntington's disease target, somatic *HTT* CAG repeat length, is measured independently of the transcriptome, by single-cell PacBio long-read sequencing of the same nuclei. It is a genomic quantity, not a summary of the counts, so an association between the inferred cell state and CAG length provides evidence about biology (Handsaker et al., 2025).

## 4.1 Evaluation protocol

**Baselines.** We compare scPACE with the representation-learning methods scVI (Lopez et al., 2018), scANVI (Xu et al., 2021), and scVIP (Lai et al., 2026), and with six methods that produce scalar cell-level scores, termed pseudotime: psupertime (Macnair et al., 2022), Sceptic (Li et al., 2025), PhenoPath (Campbell & Yau, 2018), Slingshot (Street et al., 2018), CNA (Reshef et al., 2022), and MELD (Burkhardt et al., 2021). All methods use the same donor partitions and are fitted without query donors. Query cells are scored using each method's native projection procedure when available. Otherwise, scores are interpolated from nearby reference cells in a fixed PCA space. Full baseline specifications are given in Appendix F.

**Evaluation.** Donor- and cell-level states are evaluated using absolute Pearson correlation $|r|$. We additionally report donor RMSE and cell nRMSE after an affine calibration estimated without query phenotypes. Cell-type imputation is measured by accuracy on annotated query cells, and donor integration is measured by within-type donor mixing. Metric definitions and calibration are provided in Appendix F.

The methods differ in the supervision they receive. Five of the six scalar methods, psupertime, Sceptic, PhenoPath, CNA, and MELD, receive donor-level labels and no cell-level target. The sixth, an unsupervised method named Slingshot, uses the donor labels to select the root cluster of its lineage tree. scVI receives neither phenotypes nor cell types, scANVI receives cell types, and scVIP receives cell types and the donor phenotype. scPACE receives the donor phenotype and the available cell types in every setting. In Settings 2 and 4 it also receives a cell-level target. In Settings 1 and 3 no method receives a cell-level target, so the cell-level comparison rests on donor-level supervision alone. Among the methods in Table 1, scPACE alone is trained on the evaluated cell-level target in Settings 2 and 4, and its cell-level margin there reflects that supervision (Appendix F).

Across the four settings, scPACE attains a strong query cell-level association and low error in each cohort while maintaining cell-type accuracy of at least 93%. It also produces higher donor-mixing ratios than scVI, scANVI, and scVIP (Table 1).

### 4.2 Mouse cortical development: effect of cell-level supervision

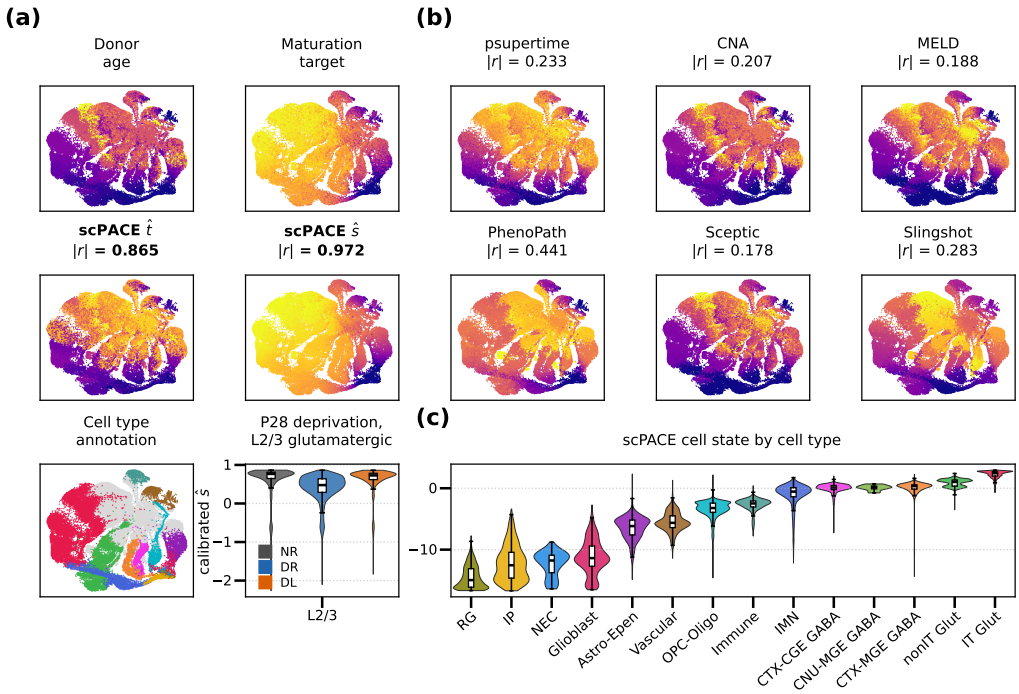



Figure 2: **Cell-state estimation in mouse visual cortex development.** **(a)** Reference variables and Setting 2 scPACE estimates in 60,000 subsampled query cells. Reported correlations use all 176,614 query cells. The P28 panel compares L2/3 glutamatergic cells under normal rearing (NR), dark rearing (DR), and dark rearing followed by light exposure (DL). **(b)** Scalar baseline scores. **(c)** scPACE cell-state distributions by annotated cell type.

**Dataset and task.** We use 865,012 cells from 88 mouse sample units, combining single-cell profiles across embryonic and adult development (Gao et al., 2025) with single-nucleus profiles spanning postnatal development and visual deprivation (Cheng et al., 2022). Chronological age is the donor phenotype. Cell-level evaluation uses an expression-derived mature-versus-progenitor score (Ap-

Table 1: **Reference and query performance across three cohorts and four settings.** Entries report reference/query values; reference includes fitting and validation data. Boldface marks the strongest query value among the listed methods within each cohort block. Supervised maturity-score prediction from scVI representations is evaluated in Appendix G. Dashes indicate unavailable or inapplicable outputs.

| Method | Donor $\|r\|$ | Donor RMSE | Cell $\|r\|$ | Cell nRMSE | Type acc. ↑ | Mixing ↑ |
|---|---|---|---|---|---|---|
| *(a) Mouse visual cortex development, Settings 1 and 2* | | | | | | |
| *68 reference and 20 query donors; 688,398 reference and 176,614 query cells* | | | | | | |
| scPACE (S1, $\lambda_{\text{cell}}$=0) | **0.97/0.91** | **0.24/0.40** | 0.13/0.08 | 0.99/1.00 | 0.93/0.93 | **0.81/0.54** |
| scPACE (S2, $\lambda_{\text{cell}}$=500) | 0.96/0.86 | 0.30/0.50 | **0.98/0.97** | **0.22/0.23** | 0.93/0.95 | **0.81/0.54** |
| scVIP | 0.69/0.73 | 0.76/0.65 | – | – | – | 0.78/0.50 |
| psupertime | – | – | 0.28/0.23 | 0.96/0.97 | – | – |
| Sceptic | – | – | 0.22/0.18 | 0.98/0.99 | – | – |
| Slingshot | – | – | 0.28/0.28 | 0.96/0.96 | – | – |
| PhenoPath | – | – | 0.49/0.44 | 0.87/0.90 | – | – |
| CNA | – | – | 0.21/0.21 | 0.98/0.98 | – | – |
| MELD | – | – | 0.22/0.19 | 0.97/0.98 | – | – |
| scVI | – | – | – | – | – | 0.79/0.50 |
| scANVI | – | – | – | – | **0.99/0.99** | 0.78/0.50 |
| *(b) Alzheimer's disease, Setting 3* | | | | | | |
| *67 reference and 17 query donors; 150,798 reference and 37,509 query cells;* | | | | | | |
| *cell metrics on 26,790 reference and 6,417 query microglia* | | | | | | |
| scPACE ($\lambda_{\text{cell}}$=0) | 0.23/0.13 | 1.01/1.13 | **0.53/0.55** | **0.85/0.84** | **1.00/1.00** | **0.95/0.85** |
| scVIP | **0.36/0.58** | **0.96/0.98** | – | – | – | 0.87/0.68 |
| psupertime | – | – | 0.06/0.09 | 1.00/1.00 | – | – |
| Sceptic | – | – | 0.11/0.19 | 0.99/0.99 | – | – |
| Slingshot | – | – | 0.09/0.05 | 1.00/1.00 | – | – |
| PhenoPath | – | – | 0.04/0.05 | 1.00/1.00 | – | – |
| CNA | – | – | 0.17/0.25 | 0.99/0.97 | – | – |
| MELD | – | – | 0.07/0.17 | 1.00/0.99 | – | – |
| scVI | – | – | – | – | – | 0.90/0.70 |
| scANVI | – | – | – | – | **1.00/1.00** | 0.91/0.71 |
| *(c) Huntington's disease, Setting 4* | | | | | | |
| *5 reference and 3 query donors; 183,987 reference and 118,204 query cells;* | | | | | | |
| *cell metrics on 17,983 reference and 11,047 CAG-measured query nuclei* | | | | | | |
| scPACE ($\lambda_{\text{cell}}$=500) | –[†] | **0.91/1.03** | **0.31/0.25** | **0.95/0.97** | **0.99/0.99** | **0.68/0.66** |
| scVIP | –[†] | 0.28/1.29 | – | – | – | 0.50/0.33 |
| psupertime | – | – | 0.05/0.07 | 1.00/1.01 | – | – |
| Sceptic | – | – | 0.08/0.03 | 1.00/1.01 | – | – |
| Slingshot | – | – | 0.21/0.19 | 0.98/0.98 | – | – |
| PhenoPath | – | – | 0.14/0.15 | 0.99/0.99 | – | – |
| CNA | – | – | – | – | – | – |
| MELD | – | – | 0.07/0.01 | 1.00/1.00 | – | – |
| scVI | – | – | – | – | – | 0.57/0.33 |
| scANVI | – | – | – | – | **0.99/0.99** | 0.59/0.33 |

[†] Donor $|r|$ is omitted in this cohort. A Pearson correlation over five reference or three query donors is unstable, so donor RMSE carries the donor-level comparison here.

pendix E). Setting 1 uses only donor-level phenotypes, whereas Setting 2 adds the maturity score as cell-level supervision. The settings share the same data, split, architecture, and seed.

**Results.** Without cell-level supervision, the donor state reaches $|r| = 0.9050$ with chronological age across 20 query donors and RMSE $0.4027$, compared with $|r| = 0.7340$ for scVIP. Adding maturity supervision increases the query cell-state association to $|r| = 0.9724$ with nRMSE $0.23$. PhenoPath is the strongest scalar baseline ($|r| = 0.4407$, nRMSE $= 0.8980$), while the other scalar methods reach correlations between $0.1784$ and $0.2826$. Setting 2 also achieves cell-type accuracy of $0.9464$. The query donor-mixing ratios are $0.5422$ and $0.5384$ for Settings 1 and 2, compared with $0.5022$ for scVI, $0.4974$ for scANVI, and $0.4988$ for scVIP (Figure 2).

With the same cell target, an MLP on frozen scVI embeddings reaches $|r| = 0.959$ and nRMSE $0.285$, compared with $0.972$ and $0.233$ for scPACE. Type-demeaned correlation is also higher for scPACE ($0.846$ versus $0.783$; Appendix G).

**Biological analysis.** Within P28 L2/3 glutamatergic cells, the median inferred state decreases from $0.767$ under normal rearing to $0.477$ under dark rearing and returns toward the normal-rearing level after light exposure ($0.728$) (Figure 2). Because all cells in this comparison have the same chronological age, the shift captures experience-dependent maturation beyond donor age. Across cell types, neuroepithelial cells and radial glia occupy the lowest-state region, whereas differentiated neuronal populations occupy higher states, consistent with known cortical developmental organization (La Manno et al., 2021; Gao et al., 2025). Together, the panels show that scPACE captures cell-type-associated maturation, heterogeneity within cell populations, and the maturation-contrast shifts accompanying visual deprivation and recovery.

To assess the imputed labels without reference to the annotation, we grouped query cells by imputed label and computed the mean expression of canonical markers within each class for every query donor (Appendix F.7). As a result, the marker gene reaches its maximum in the imputed class: *Pax6* and *Top2a* in RG; *Eomes* and *Neurog2* in IP; *Nhlh2* and *Neurod1* in IMN; *Slc30a3* in IT Glut; *Syt6* and *Foxp2* in non-IT Glut; *Lhx6* and *Sst* in CTX-MGE GABA; *Adarb2* and *Vip* in CTX-CGE GABA; *Tnc* in Glioblast; *Aqp4* in Astro-Epen; *Pdgfra* and *Olig1* in OPC-Oligo; *C1qa* and *Csf1r* in Immune cells; and *Flt1* in Vascular cells.

**Ablation study.** We ablate scPACE components on the mouse cohort in Setting 1, which uses donor supervision only (Appendix G). This setting tests how donor information supports cell-state inference without direct cell-level supervision. We retain the same data and donor split and change only the specified model or training components. Removing both generative donor-to-cell connections reduces the median absolute within-type cell correlation from the full-model value of $0.45$ to $0.18$, while donor correlation remains $0.91$. These results support a structural contribution of the generative hierarchy to cell-state organization. Using cell-type proportions alone reduces donor correlation from $0.91$ to $0.60$, showing the contribution of within-type state summaries to donor prediction. Finally, adding cell-level supervision with a small cell-channel weight substantially improves pooled cell-state correlation while retaining donor-level performance (Appendix Figure 8).

### 4.3 Alzheimer's disease: cellular variation from donor-level supervision

**Dataset and task.** Our Alzheimer's disease experiment uses 188,307 glial nuclei from 84 SEAAD brain donors, sampled from the middle temporal gyrus across 177 technical batches (Gabitto et al., 2024). The donor phenotype is the continuous pseudoprogression score, a score created by Gabitto et al. (2024), which orders donors by disease severity burden. No cell-level phenotype is used during fitting. We evaluate cell states in 6,417 query Microglia-PVM nuclei against an expression-derived DAM-versus-homeostatic score. This score captures the transition from a homeostatic to a DAM state and is withheld from all methods during training (Appendix E).

**Results.** scPACE attains the strongest query cell-level association ($|r| = 0.548$), the lowest cell nRMSE ($0.84$), perfect broad glial-type accuracy, and the highest within-type donor mixing ($0.85$) (Table 1; Figure 3). scVIP performs better at the donor level in this cohort, with query donor $|r| = 0.58$ and RMSE $0.98$, compared with $0.13$ and $1.13$ for scPACE.

**Biological analysis.** The inferred cell state varies across Micro-PVM supertypes despite the absence of cell-level supervision (Figure 3). Cells in the upper third of both the scPACE state and the DAM score account for $43.3\%$ of Micro-PVM_3-SEAAD nuclei, compared with $23.8\%$ of Micro-PVM_1, $15.1\%$ of Micro-PVM_2_3-SEAAD, and $8.4\%$ of Micro-PVM_2. Micro-PVM_3-SEAAD is the disease-associated type (DAM) identified by the original atlas, whereas Micro-PVM_2 anchors the homeostatic end of the inferred continuum (Gabitto et al., 2024).

### 4.4 Huntington's disease: multilevel phenotype supervision

**Dataset and task.** The Huntington's disease experiment uses 302,191 caudate nuclei from eight donors (Handsaker et al., 2025). Donor supervision uses the CAP score (Warner et al., 2022), and cell supervision uses somatic *HTT* CAG repeat length measured by single-cell PacBio sequencing. Measurements are available for 29,030 nuclei, including 11,047 nuclei from three query donors. The

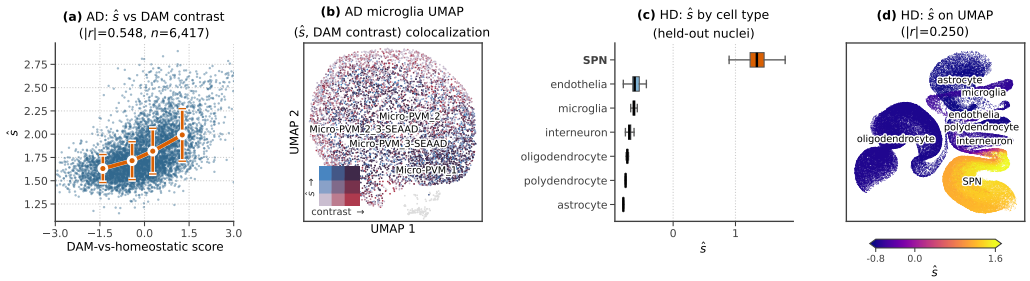



Figure 3: **Disease-associated cell states in Alzheimer's and Huntington's disease.** **(a)** Inferred cell state versus the DAM score in 6,417 query SEAAD nuclei, without cell-level supervision. **(b)** Query Microglia-PVM nuclei colored by joint state and DAM-score terciles. **(c)** Cell-state distributions across HD classes in 118,204 query nuclei. **(d)** HD query nuclei colored by inferred cell state.

remaining donors are divided into four training donors and one validation donor. Appendix Figure 5 shows the measured CAG repeat lengths in query nuclei and their distributions across cell types.

**Results.** scPACE achieves lower query CAP error than scVIP (RMSE 1.03 versus 1.29). Across CAG-measured query nuclei, it obtains cell-level association $|r| = 0.2501$ and nRMSE 0.9688. It also improves within-type donor mixing over scVI and scANVI while retaining cell-type accuracy of 0.9860, compared with 0.9903 for scANVI (Table 1).

**Biological analysis.** Spiny projection neurons have the highest median inferred cell state among the seven annotated classes, while all other classes have negative medians (Figure 3). This ordering is consistent with the selective vulnerability of SPNs and with the reported association between somatic *HTT* CAG expansion, loss of neuronal identity, and degeneration (Handsaker et al., 2025).

## 5 Discussion and Conclusion

We introduced scPACE, a hierarchical generative model that links donor phenotypes, cell-level phenotype, cell-type composition, and molecular counts. Its central methodological contribution is a coupled donor-cell state hierarchy that jointly models these quantities within a single probabilistic framework and accommodates incomplete supervision. Class-stratified attention combines within-type cell-state summaries with cell-type proportions, allowing donor predictions to reflect both state and composition differences. Identifiability provides supporting theory for interpreting the scalar donor and cell states. Under the assumptions in Appendix A, these states are identifiable up to a common sign convention.

Across mouse cortical development, Alzheimer's disease, and Huntington's disease, scPACE captured biologically coherent variation at donor and cell levels. The inferred cell states reflected developmental and experience-dependent maturation in mouse cortex, disease-associated microglial variation in AD, and the elevated state of vulnerable spiny projection neurons in HD. The quantitative comparisons supported these findings. In the mouse cohort, scPACE outperformed the compared methods in predicting donor age and cell maturity. In AD, its cell states tracked the DAM score more closely than the scalar baselines, and in HD it reduced donor-level prediction error. Across all three cohorts, it also achieved the highest donor-mixing ratios among the compared methods. These results support the value of jointly modeling phenotype-linked variation at donor and cell levels.

**Limitations.** scPACE currently relies on parametric phenotype observation models, which may limit its flexibility across biological settings. Future work could consider more expressive observation models, such as monotone nonlinear mappings, to capture richer phenotype–state relationships while potentially preserving identifiability. In addition, the mouse and AD cell-level scores are expression-derived, and the AD and HD findings require biological validation. Future studies should validate scPACE in larger cohorts using independent multimodal measurements.

## Reproducibility Statement

Data preparation, feature selection, donor splits, baseline adapters, hyperparameters, and evaluation scripts are recorded in versioned configurations, and all experiments use a single Apptainer image. Splits are reproduced from a recorded seed (1024 for all four settings). All reported runs use training seed 0 and the settings in Section 4. Checkpoints, configurations, and evaluation manifests are retained. Code and environment specifications will be released upon acceptance.

## Ethics Statement

This study reanalyzes published, publicly available, de-identified postmortem human brain datasets (SEAAD middle temporal gyrus (Gabitto et al., 2024) and Huntington's disease caudate (Handsaker et al., 2025)) and published mouse visual-cortex single-cell data (Gao et al., 2025). No new human or animal subjects were involved.

## Use of Generative AI

In this work, we used generative AI tools for code development, for the scripting of experiments and evaluations, and for drafting descriptive text in the appendices. Additionally, we used generative AI as adversarial reviewers of the identifiability analysis in Appendix A, helping us challenge and refine the claims and their scope. We have not used generative AI tools for research ideation, for the design of the model. Generative AI was not used to generate novel scientific claims. Moreover, we used generative AI tools for copy-editing, for formatting tables, and for producing figure code. We have reviewed all AI-assisted work. The literature survey and every citation were assembled and verified by the authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

## References

Sara Aibar, Carmen Bravo González-Blas, Thomas Moerman, Vân Anh Huynh-Thu, Hana Imrichova, Gert Hulselmans, Florian Rambow, Jean-Christophe Marine, Pierre Geurts, Jan Aerts, et al. SCENIC: Single-cell regulatory network inference and clustering. *Nature Methods*, 14(11):1083–1086, 2017. doi: 10.1038/nmeth.4463. URL https://doi.org/10.1038/nmeth.4463.

Diane Bouchacourt, Ryota Tomioka, and Sebastian Nowozin. Multi-level variational autoencoder: Learning disentangled representations from grouped observations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, pp. 2095–2102, 2018. doi: 10.1609/aaai.v32i1.11867. URL https://ojs.aaai.org/index.php/AAAI/article/view/11867.

Pierre Boyeau, Justin Hong, Adam Gayoso, Martin Kim, José L. McFaline-Figueroa, Michael I. Jordan, Elham Azizi, Can Ergen, and Nir Yosef. Deep generative modeling of sample-level heterogeneity in single-cell genomics. *Nature Methods*, 22:2264–2274, 2025. doi: 10.1038/s41592-025-02808-x. URL https://doi.org/10.1038/s41592-025-02808-x.

Johann Brehmer, Pim de Haan, Phillip Lippe, and Taco S. Cohen. Weakly supervised causal representation learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pp. 38319–38331, 2022. doi: 10.52202/068431-2776. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/fa567e2b2c870f8f09a87b6e73370869-Abstract-Conference.html.

Daniel B. Burkhardt, Jay S. Stanley, III, Alexander Tong, Ana Luisa Perdigoto, Scott A. Gigante, Kevan C. Herold, Guy Wolf, Antonio J. Giraldez, David van Dijk, and Smita Krishnaswamy. Quantifying the effect of experimental perturbations at single-cell resolution. *Nature Biotechnology*, 39:619–629, 2021. doi: 10.1038/s41587-020-00803-5. URL https://doi.org/10.1038/s41587-020-00803-5.

Oleg Butovsky, Mark P Jedrychowski, Craig S Moore, Ron Cialic, Amanda J Lanser, Galina Gabriely, Thomas Koeglsperger, Ben Dake, Pauline M Wu, Camille E Doykan, Zain Fanek, LiPing Liu, Zhuoxun Chen, Jeffrey D Rothstein, Richard M Ransohoff, Steven P Gygi, Jack P Antel,

and Howard L Weiner. Identification of a unique tgf-$\beta$-dependent molecular and functional signature in microglia. *Nature Neuroscience*, 17(1):131–143, 2014.

Kieran R. Campbell and Christopher Yau. Uncovering pseudotemporal trajectories with covariates from single cell and bulk expression data. *Nature Communications*, 9:2442, 2018. doi: 10.1038/s41467-018-04696-6. URL https://doi.org/10.1038/s41467-018-04696-6.

Raymond J. Carroll, David Ruppert, Leonard A. Stefanski, and Ciprian M. Crainiceanu. *Measurement Error in Nonlinear Models: A Modern Perspective*. Chapman & Hall/CRC, Boca Raton, second edition, 2006.

Sarah Cheng, Salwan Butrus, Liming Tan, Runzhe Xu, Srikant Sagireddy, Joshua T. Trachtenberg, Karthik Shekhar, and S. Lawrence Zipursky. Vision-dependent specification of cell types and function in the developing cortex. *Cell*, 185(2):311–327.e24, 2022. doi: 10.1016/j.cell.2021.12.022. URL https://doi.org/10.1016/j.cell.2021.12.022.

Rewon Child. Very deep VAEs generalize autoregressive models and can outperform them on images. In *Proceedings of the 9th International Conference on Learning Representations (ICLR)*, 2021. URL https://openreview.net/forum?id=RLRXCV6DbEJ.

Carlo De Donno, Soroor Hediyeh-Zadeh, Amir Ali Moinfar, Marco Wagenstetter, Luke Zappia, Mohammad Lotfollahi, and Fabian J. Theis. Population-level integration of single-cell datasets enables multi-scale analysis across samples. *Nature Methods*, 20:1683–1692, 2023. doi: 10.1038/s41592-023-02035-2. URL https://doi.org/10.1038/s41592-023-02035-2.

Harrison Edwards and Amos Storkey. Towards a neural statistician. In *Proceedings of the 5th International Conference on Learning Representations (ICLR)*, 2017. URL https://openreview.net/forum?id=HJDBUF5le.

Mariano I. Gabitto, Kyle J. Travaglini, Victoria M. Rachleff, Eitan S. Kaplan, Brian Long, Jeanelle Ariza, Yi Ding, Joseph T. Mahoney, Nick Dee, Jeff Goldy, et al. Integrated multimodal cell atlas of alzheimer's disease. *Nature Neuroscience*, 27:2366–2383, 2024. doi: 10.1038/s41593-024-01774-5. URL https://doi.org/10.1038/s41593-024-01774-5.

Yuan Gao, Cindy TJ Van Velthoven, Changkyu Lee, Emma D Thomas, Rémi Mathieu, Angela P Ayala, Stuard Barta, Darren Bertagnolli, Jazmin Campos, Trangthanh Cardenas, et al. Continuous cell-type diversification in mouse visual cortex development. *Nature*, 647(8088):127–142, 2025.

Robert E. Handsaker, Seva Kashin, Nora M. Reed, Steven Tan, Won-Seok Lee, Tara M. McDonald, Kiely Morris, Nolan Kamitaki, Christopher D. Mullally, Neda R. Morakabati, et al. Long somatic DNA-repeat expansion drives neurodegeneration in huntington's disease. *Cell*, 188(3):623–639.e19, 2025. doi: 10.1016/j.cell.2024.11.038. URL https://doi.org/10.1016/j.cell.2024.11.038.

Wei-Ning Hsu, Yu Zhang, and James Glass. Unsupervised learning of disentangled and interpretable representations from sequential data. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, pp. 1878–1889, 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/hash/0a0a0c8aaa00ade50f74a3f0ca981ed7-Abstract.html.

Maximilian Ilse, Jakub M. Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2127–2136. PMLR, 2018. URL https://proceedings.mlr.press/v80/ilse18a.html.

Matthew J. Johnson, David K. Duvenaud, Alex Wiltschko, Ryan P. Adams, and Sandeep R. Datta. Composing graphical models with neural networks for structured representations and fast inference. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 29, pp. 2946–2954, 2016. URL https://proceedings.neurips.cc/paper/2016/hash/7d6044e95a16761171b130dcb476a43e-Abstract.html.

Hadas Keren-Shaul, Amit Spinrad, Assaf Weiner, Orit Matcovitch-Natan, Raz Dvir-Szternfeld, Tyler K Ulland, Eyal David, Kuti Baruch, David Lara-Astaiso, Beata Toth, Shalev Itzkovitz, Marco Colonna, Michal Schwartz, and Ido Amit. A unique microglia type associated with restricting development of alzheimer's disease. *Cell*, 169(7):1276–1290.e17, 2017.

Ilyes Khemakhem, Diederik P. Kingma, Ricardo Pio Monti, and Aapo Hyvärinen. Variational autoencoders and nonlinear ICA: A unifying framework. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 108 of *Proceedings of Machine Learning Research*, pp. 2207–2217. PMLR, 2020. URL https://proceedings.mlr.press/v108/khemakhem20a.html.

Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *Proceedings of the 2nd International Conference on Learning Representations (ICLR)*, 2014. URL https://arxiv.org/abs/1312.6114.

Gioele La Manno, Kimberly Siletti, Alessandro Furlan, Daniel Gyllborg, Elin Vinsland, Alejandro Mossi Albiach, Christoffer Mattsson Langseth, Irina Khven, Alex R Lederer, Lisa M Dratva, Anna Johnsson, Mats Nilsson, Peter Lonnerberg, and Sten Linnarsson. Molecular architecture of the developing mouse brain. *Nature*, 596(7870):92–96, 2021.

Hsin-Yu Lai, Yehchan Yoo, Andreas Tjärnberg, Ruoxin Li, Ziyuan He, Kyle J. Travaglini, Qi Qiao, Anamika Agrawal, Omar Kana, et al. scVIP: Personalized modeling of single-cell transcriptomes for developmental and disease phenotypes. *bioRxiv*, 2026. doi: 10.64898/2026.04.20.717759. URL https://doi.org/10.64898/2026.04.20.717759. Preprint, version 2.

Juho Lee, Yoonho Lee, Jungtaek Kim, Adam Kosiorek, Seungjin Choi, and Yee Whye Teh. Set transformer: A framework for attention-based permutation-invariant neural networks. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3744–3753. PMLR, 2019. URL https://proceedings.mlr.press/v97/lee19d.html.

Gang Li, Hyeon-Jin Kim, Sriram Pendyala, Ran Zhang, Jean-Philippe Vert, Christine M. Disteche, Xinxian Deng, Douglas M. Fowler, and William Stafford Noble. Sceptic: Pseudotime analysis for time-series single-cell sequencing and imaging data. *Genome Biology*, 26 (1):209, 2025. doi: 10.1186/s13059-025-03679-3. URL https://doi.org/10.1186/s13059-025-03679-3.

Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Rätsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, volume 97 of *Proceedings of Machine Learning Research*, pp. 4114–4124. PMLR, 2019. URL https://proceedings.mlr.press/v97/locatello19a.html.

Romain Lopez, Jeffrey Regier, Michael B. Cole, Michael I. Jordan, and Nir Yosef. Deep generative modeling for single-cell transcriptomics. *Nature Methods*, 15:1053–1058, 2018. doi: 10.1038/s41592-018-0229-2. URL https://doi.org/10.1038/s41592-018-0229-2.

Will Macnair, Revant Gupta, and Manfred Claassen. psupertime: Supervised pseudotime analysis for time-series single-cell RNA-seq data. *Bioinformatics*, 38(Supplement 1):i290–i298, 2022. doi: 10.1093/bioinformatics/btac227. URL https://doi.org/10.1093/bioinformatics/btac227.

Yakir A. Reshef, Laurie Rumker, Joyce B. Kang, Aparna Nathan, Ilya Korsunsky, Samira Asgari, Megan B. Murray, D. Branch Moody, and Soumya Raychaudhuri. Co-varying neighborhood analysis identifies cell populations associated with phenotypes of interest from single-cell transcriptomics. *Nature Biotechnology*, 40(3):355–363, 2022. doi: 10.1038/s41587-021-01066-4. URL https://doi.org/10.1038/s41587-021-01066-4.

Casper Kaae Sønderby, Tapani Raiko, Lars Maaløe, Søren Kaae Sønderby, and Ole Winther. Ladder variational autoencoders. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 29, pp. 3738–3746, 2016.

Kelly Street, Davide Risso, Russell B. Fletcher, Diya Das, John Ngai, Nir Yosef, Elizabeth Purdom, and Sandrine Dudoit. Slingshot: Cell lineage and pseudotime inference for single-cell transcriptomics. *BMC Genomics*, 19(1):477, 2018. doi: 10.1186/s12864-018-4772-0. URL https://doi.org/10.1186/s12864-018-4772-0.

Arash Vahdat and Jan Kautz. NVAE: A deep hierarchical variational autoencoder. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 19667–19679, 2020.

John H Warner, Jeffrey D Long, James A Mills, Douglas R Langbehn, Jennifer Ware, Amrita Mohan, and Cristina Sampaio. Standardizing the cap score in huntington's disease by predicting age-at-onset. *Journal of Huntington's disease*, 11(2):153–171, 2022.

Ethan Weinberger, Romain Lopez, Jan-Christian Huetter, and Aviv Regev. Disentangling shared and group-specific variations in single-cell transcriptomics data with multiGroupVI. In *Proceedings of the 17th Machine Learning in Computational Biology Meeting*, volume 200 of *Proceedings of Machine Learning Research*, pp. 16–32. PMLR, 2022. URL https://proceedings.mlr.press/v200/weinberger22a.html.

Quanhan Xi and Benjamin Bloem-Reddy. Indeterminacy in generative models: Characterization and strong identifiability. In *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 206 of *Proceedings of Machine Learning Research*, pp. 6912–6939, 2023.

Chenling Xu, Romain Lopez, Edouard Mehlman, Jeffrey Regier, Michael I. Jordan, and Nir Yosef. Probabilistic harmonization and annotation of single-cell transcriptomics data with deep generative models. *Molecular Systems Biology*, 17(1):e9620, 2021. doi: 10.15252/msb.20209620. URL https://doi.org/10.15252/msb.20209620.

Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabás Póczos, Ruslan Salakhutdinov, and Alexander J. Smola. Deep sets. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, pp. 3391–3401, 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/hash/f22e4747da1aa27e363d86d40ff442fe-Abstract.html.

# A IDENTIFIABILITY OF THE LATENT STATE HIERARCHY

The coordinates of the multivariate cell representation $\boldsymbol{z}_n$ are not uniquely defined. Invertible reparameterizations can be absorbed by the conditional prior and count decoder. The scalar donor and cell states are more constrained because they enter the model through a standard-Normal donor prior, a unit-slope donor-to-cell bridge, cell-type composition, and two phenotype observation channels. This appendix characterizes the remaining indeterminacy of the scalar hierarchy.

Generic nonlinear latent-variable models require additional structure for identification (Locatello et al., 2019). iVAE uses an observed auxiliary variable to modulate a latent prior (Khemakhem et al., 2020). scPACE uses its scalar donor-to-cell hierarchy and phenotype observation models. We characterize the resulting equivalence class in the framework of Xi & Bloem-Reddy (2023).

The model in the main text includes type-specific bridge intercepts $m_k$. These intercepts require one additional gauge. Without it, the transformation

$$S \mapsto S + h, \qquad m_k \mapsto m_k + h, \qquad \delta_C \mapsto \delta_C - a_C h,$$

together with the corresponding translation of the state-indexed count kernel, leaves the observable law unchanged. We therefore choose a reference type $K$ and impose $m_K = 0$. An equivalent centered constraint could be used instead.

## A.1 GENERATIVE FAMILY AND FIXED BATCH DESIGN

For donor $d$ and cell $n$, let $T_d$ and $S_{dn}$ denote the scalar donor and cell states. Let $C_{dn}$ denote the observed cell type, $X_{dn}$ the count vector, and $Y_d^D$ and $Y_{dn}^C$ the continuous donor and cell phenotypes. The theorem treats the cell types and both phenotype channels as observed. The estimator in Appendix B handles missing values, which are not part of the full identifiability claim. We treat the technical batch vector $\boldsymbol{b}_d = (b_{d1}, \ldots, b_{dN_d})$ as fixed design information and assign it no probability law. Conditional on this design,

$$T_d \sim \mathcal{N}(0,1), \qquad C_{dn} \mid T_d = t \sim \mathrm{Cat}\{\boldsymbol{\pi}(t)\}, \tag{6}$$

$$S_{dn} \mid T_d = t, C_{dn} = k \sim \mathcal{N}(t + m_k, \sigma_k^2), \qquad Y_d^D \mid T_d = t \sim \mathcal{N}(\delta_D + a_D t, \tau_D^2), \tag{7}$$

$$Y_{dn}^C \mid S_{dn} = s \sim \mathcal{N}(\delta_C + a_C s, \tau_C^2), \qquad \pi_k(t) = \frac{\exp(\alpha_k + \beta_k t)}{\sum_j \exp(\alpha_j + \beta_j t)}. \tag{8}$$

The remaining count hierarchy is

$$\begin{aligned} \boldsymbol{U}_{dn} &\sim \mathcal{N}(\boldsymbol{0}, \mathbf{I}), & \boldsymbol{Z}_{dn} \mid \boldsymbol{U}_{dn}, C_{dn}, S_{dn} &\sim \mathcal{N}\big(\boldsymbol{\mu}_\theta^p, \mathrm{diag}\{(\boldsymbol{\sigma}_\theta^p)^2\}\big), \\ \ell_{dn} &\sim \mathcal{N}(\mu_\ell, v_\ell), & X_{dng} \mid \boldsymbol{Z}_{dn}, \ell_{dn}; b_{dn} &\sim \mathrm{NB}\big(e^{\ell_{dn}} \rho_{dng}, \theta_g\big), \end{aligned} \tag{9}$$

where $(\boldsymbol{\mu}_\theta^p, \boldsymbol{\sigma}_\theta^p)$ are evaluated at $(\boldsymbol{U}_{dn}, C_{dn}, S_{dn})$ and $\boldsymbol{\rho}_{dn}$ at $(\boldsymbol{Z}_{dn}, b_{dn})$.

Conditional on $\boldsymbol{b}_d$, the joint density is

$$\begin{aligned} &p_\theta\Big(t_d, y_d^D, \{c_{dn}, s_{dn}, y_{dn}^C, \boldsymbol{u}_{dn}, \boldsymbol{z}_{dn}, \ell_{dn}, \boldsymbol{x}_{dn}\}_{n=1}^{N_d} \mid \boldsymbol{b}_d\Big) \\ &\quad = p(t_d) p_\theta(y_d^D \mid t_d) \prod_{n=1}^{N_d} \Big[ p_\theta(c_{dn} \mid t_d) p_\theta(s_{dn} \mid t_d, c_{dn}) p_\theta(y_{dn}^C \mid s_{dn}) p(\boldsymbol{u}_{dn}) \\ &\qquad \times p_\theta(\boldsymbol{z}_{dn} \mid \boldsymbol{u}_{dn}, c_{dn}, s_{dn}) p(\ell_{dn}) p_\theta(\boldsymbol{x}_{dn} \mid \boldsymbol{z}_{dn}, \ell_{dn}; b_{dn}) \Big]. \end{aligned} \tag{10}$$

Let

$$\mathcal{O}_d^{\boldsymbol{b}} = \{C_{dn}, Y_{dn}^C, X_{dn}\}_{n=1}^{N_d}$$

denote the observed cellular record under the fixed design. It excludes $Y_d^D$, which serves as the donor proxy in the first deconvolution step.

After integrating out $\boldsymbol{U}$, $\boldsymbol{Z}$, and $\ell$, we define the induced state-indexed count kernel

$$K_{k,b}(A \mid s) = \Pr_\theta(X_{dn} \in A \mid S_{dn} = s, C_{dn} = k, b_{dn} = b). \tag{11}$$

The kernel may depend on the known batch label $b$. We assume that the composition, bridge, and phenotype channels are batch-invariant.

### A.2 Observational equivalence, gauges, and assumptions

For each fixed admissible batch design $\boldsymbol{b}$, let $P_\theta^{\boldsymbol{b}}$ denote the law induced by Eq. (10). Two parameterizations are *observationally equivalent* when they induce the same law of $(Y_d^D, \mathcal{O}_d^{\boldsymbol{b}})$ for every admissible design.

We use the following gauges:

1. $T \sim \mathcal{N}(0,1)$ fixes donor location and scale;
2. $\mathbb{E}(S \mid T, C = k) = T + m_k$ with unit coefficient fixes the scale of every cell state relative to $T$;
3. $m_K = 0$ fixes the remaining common translation of the cell-state axis;
4. $\alpha_K = \beta_K = 0$ fixes the reference category of the multinomial logit;
5. $0 < \tau_D^2, \tau_C^2 < \infty$ are fixed and known, while $0 < \sigma_k^2 < \infty$ are learned;
6. optionally, $a_D > 0$ fixes the orientation.

Because the cell types are observed in the theorem regime, there is no cell-type label switching.

**Assumption 1** (Identifiability conditions)**.** The following conditions hold.

1. $a_D a_C \neq 0$.
2. For every fixed design $\boldsymbol{b}$, the factorization in Eq. (10) is correctly specified. In particular,
$$Y^D \perp\!\!\!\perp \mathcal{O}^{\boldsymbol{b}} \mid T, \qquad Y_n^C \perp\!\!\!\perp (X_n, T, C_n) \mid S_n, \qquad X_n \perp\!\!\!\perp (T, Y_n^C) \mid (S_n, C_n, b_n).$$
3. Every batch value for which identification is claimed appears as a cell label in at least one admissible design.
4. The observable spaces are standard Borel and their sigma-fields are countably generated, so regular conditional distributions exist.
5. The compared parameterizations belong to this family and induce the same population law of $(Y_d^D, \mathcal{O}_d^{\boldsymbol{b}})$ for every admissible design.

### A.3 The common sign reversal is an exact symmetry

For a parameter $\theta$, define its reversed parameter $\theta^-$ by

$$a_D^- = -a_D, \qquad a_C^- = -a_C, \qquad \boldsymbol{\beta}^- = -\boldsymbol{\beta}, \qquad m_k^- = -m_k, \tag{12}$$

$$K_{k,b}^-(A \mid s) = K_{k,b}(A \mid -s), \tag{13}$$

with $\delta_D, \delta_C, \boldsymbol{\alpha}, \{\sigma_k^2\}$ unchanged. At the neural parameter level, we obtain Eq. (13) by replacing $\boldsymbol{\mu}_\theta^p(\boldsymbol{u}, k, s)$ and $\boldsymbol{\sigma}_\theta^p(\boldsymbol{u}, k, s)$ with their values at $-s$. The count decoder remains unchanged.

**Proposition 1** (Exact reflection symmetry)**.** *For every fixed design $\boldsymbol{b}$,*

$$\begin{aligned}\mathcal{L}_{\theta^-}^{\boldsymbol{b}}\Big(T, \{S_n, C_n, Y_n^C, X_n\}_{n=1}^{N_d}, Y^D\Big) \\ = \mathcal{L}_\theta^{\boldsymbol{b}}\Big(-T, \{-S_n, C_n, Y_n^C, X_n\}_{n=1}^{N_d}, Y^D\Big).\end{aligned} \tag{14}$$

*Consequently, $\theta$ and $\theta^-$ are observationally equivalent.*

*Proof.* The standard-Normal prior is invariant under $T \mapsto -T$. The donor channel is unchanged because $(-a_D)t = a_D(-t)$. The composition probabilities are unchanged because $\boldsymbol{\alpha} + t(-\boldsymbol{\beta}) = \boldsymbol{\alpha} + (-t)\boldsymbol{\beta}$. The bridge satisfies

$$\mathcal{N}(s; t - m_k, \sigma_k^2) = \mathcal{N}(-s; -t + m_k, \sigma_k^2),$$

which is the original bridge evaluated at the reflected states after $m_k^- = -m_k$. The cell channel and count kernel are invariant by Eqs. (12)–(13). Multiplying the factors and changing variables from $(t, \boldsymbol{s})$ to $(-t, -\boldsymbol{s})$ proves the claim. □

### A.4 A GAUSSIAN DECONVOLUTION LEMMA

The proof uses each phenotype channel as a noisy linear measurement of its state. The following lemma gives the classical known-error Gaussian deconvolution argument (Carroll et al., 2006). For a finite measure $\nu$, we write $\widehat{\nu}(\omega) = \int e^{i\omega t}\nu(dt)$.

**Lemma 1** (Gaussian deconvolution)**.** *Let $S$ be real-valued, let $R \geq 0$ be bounded, and let $Y = aS + c + \tau\varepsilon$, where $a \neq 0$, $\tau \geq 0$, and $\varepsilon \sim \mathcal{N}(0,1)$ is independent of $(S,R)$. Define $\nu_R(B) = \mathbb{E}[R\,\mathbb{I}\{S \in B\}]$. Then*

$$\mathbb{E}\Big[Re^{i\omega(Y-c)}\Big]\, e^{\tau^2\omega^2/2} = \widehat{\nu}_R(a\omega), \qquad \omega \in \mathbb{R}. \tag{15}$$

*Thus, the law of $(R,Y)$ together with $(a,c,\tau)$ determines $\nu_R$.*

*Proof.* Independence gives $\mathbb{E}[Re^{i\omega(aS+\tau\varepsilon)}] = \widehat{\nu}_R(a\omega)e^{-\tau^2\omega^2/2}$. Because $a \neq 0$, varying $\omega$ supplies the Fourier transform at every frequency. The uniqueness of Fourier transforms of finite measures completes the argument. □

### A.5 IDENTIFIABILITY THEOREM

**Theorem 1** (Proxy-anchored identifiability)**.** *Under the gauges above and Assumption 1, let $\theta$ and $\widetilde{\theta}$ be observationally equivalent. Then there is a single $\varepsilon \in \{-1,+1\}$ such that, for every cell type $k$,*

$$\widetilde{\delta}_D = \delta_D, \qquad \widetilde{\delta}_C = \delta_C, \qquad \widetilde{a}_D = \varepsilon a_D, \qquad \widetilde{a}_C = \varepsilon a_C, \tag{16}$$

$$\widetilde{m}_k = \varepsilon m_k, \qquad \widetilde{\sigma}_k^2 = \sigma_k^2, \quad \widetilde{\pi}_k(t) = \pi_k(\varepsilon t), \tag{17}$$

$$\widetilde{K}_{k,b}(\cdot \mid s) = K_{k,b}(\cdot \mid \varepsilon s) \quad \textit{for Lebesgue-a.e. } s \tag{18}$$

*for every observed batch $b$ covered by the assumptions. Under the reference-type gauge, $\widetilde{\boldsymbol{\alpha}} = \boldsymbol{\alpha}$ and $\widetilde{\boldsymbol{\beta}} = \varepsilon\boldsymbol{\beta}$. Moreover, for every fixed design,*

$$\begin{aligned}\mathcal{L}^{\boldsymbol{b}}_{\widetilde{\theta}}\Big(T, \{S_n, C_n, Y_n^C, X_n\}_{n=1}^{N_d}, Y^D\Big) \\ = \mathcal{L}^{\boldsymbol{b}}_{\theta}\Big(\varepsilon T, \{\varepsilon S_n, C_n, Y_n^C, X_n\}_{n=1}^{N_d}, Y^D\Big).\end{aligned} \tag{19}$$

*Thus, the scalar hierarchy is identifiable up to one common reflection. If $a_D > 0$ is imposed, then $\varepsilon = 1$.*

*Proof.* Let $\phi_\sigma$ denote the $\mathcal{N}(0,\sigma^2)$ density and $\phi = \phi_1$.

**Step 0: donor moments.** Equation (7) gives

$$\mathbb{E}(Y^D) = \delta_D, \qquad \mathrm{Var}(Y^D) = a_D^2 + \tau_D^2.$$

Because $\tau_D^2$ is common and known, observational equivalence gives $\widetilde{\delta}_D = \delta_D$ and $\widetilde{a}_D^2 = a_D^2$. We define $\varepsilon = \widetilde{a}_D/a_D \in \{-1,+1\}$.

**Step 1: donor deconvolution.** For any $E \in \sigma(\mathcal{O}_d^{\boldsymbol{b}})$ and $\omega \in \mathbb{R}$,

$$\mathbb{E}^{\boldsymbol{b}}_{\theta}\Big[\mathbb{I}_E e^{i\omega(Y^D-\delta_D)}\Big] = e^{-\tau_D^2\omega^2/2}\int e^{i\omega a_D t} P^{\boldsymbol{b}}_{\theta}(E \mid T=t)\phi(t)\,dt. \tag{20}$$

We equate this identity under the two models, use $\widetilde{a}_D = \varepsilon a_D$, and change variables by reflection. The Fourier transforms of the corresponding finite measures then agree. A countable generating $\pi$-system, uniqueness of Fourier transforms, positivity of $\phi$, and the $\pi$–$\lambda$ theorem yield

$$P^{\boldsymbol{b}}_{\widetilde{\theta}}(E \mid T=t) = P^{\boldsymbol{b}}_{\theta}(E \mid T=\varepsilon t) \quad \text{for Lebesgue-a.e. } t. \tag{21}$$

Equivalently, $\mathcal{L}^{\boldsymbol{b}}_{\widetilde{\theta}}(T, \mathcal{O}^{\boldsymbol{b}}) = \mathcal{L}^{\boldsymbol{b}}_{\theta}(\varepsilon T, \mathcal{O}^{\boldsymbol{b}})$.

**Step 2: composition and bridge.** Applying Eq. (21) to $C_n$ gives $\widetilde{\pi}_k(t) = \pi_k(\varepsilon t)$ almost everywhere and therefore, by continuity, the equality holds everywhere. Reference-type log odds identify $\widetilde{\boldsymbol{\alpha}} = \boldsymbol{\alpha}$ and $\widetilde{\boldsymbol{\beta}} = \varepsilon\boldsymbol{\beta}$.

The conditional moments of the cell phenotype are

$$\begin{aligned}\mathbb{E}(Y_n^C \mid T = t, C_n = k) &= \delta_C + a_C(t + m_k),\\ \mathrm{Var}(Y_n^C \mid T = t, C_n = k) &= a_C^2\sigma_k^2 + \tau_C^2.\end{aligned} \tag{22}$$

Equation (21) equates the affine conditional mean under $\widetilde{\theta}$ at $t$ with the one under $\theta$ at $\varepsilon t$. Matching slopes gives $\widetilde{a}_C = \varepsilon a_C$. For the reference type $K$, the gauge $m_K = \widetilde{m}_K = 0$ then gives $\widetilde{\delta}_C = \delta_C$. Matching the remaining intercepts yields $\widetilde{m}_k = \varepsilon m_k$ for every $k$. Finally, matching conditional variances and using the fixed $\tau_C^2$ gives $\widetilde{\sigma}_k^2 = \sigma_k^2$.

**Step 3: state-indexed count law.** Fix a coordinate with $b_n = b$, a type $k$, and a regular conditional value $t$ outside the preceding null sets. For a measurable count event $A$,

$$\begin{aligned}&\mathbb{E}_\theta^{\boldsymbol{b}}\Big[\mathbb{I}\{X_n \in A\}e^{i\omega(Y_n^C - \delta_C)} \mid T = t, C_n = k\Big]\\ &\qquad = e^{-\tau_C^2\omega^2/2}\int e^{i\omega a_C s}K_{k,b}(A \mid s)\phi_{\sigma_k}(s - t - m_k)\,ds.\end{aligned} \tag{23}$$

The left side is a conditional functional of the observed cellular record. Equation (21) therefore equates its value under $\widetilde{\theta}$ at $t$ with its value under $\theta$ at $\varepsilon t$. We substitute the identities from Step 2 and reflect the integration variable. Fourier uniqueness gives

$$\widetilde{K}_{k,b}(A \mid s)\phi_{\sigma_k}(s - t - \varepsilon m_k) = K_{k,b}(A \mid \varepsilon s)\phi_{\sigma_k}(s - t - \varepsilon m_k)$$

for almost every $s$. Positivity of the Gaussian density removes the weight. Repeating the argument on a countable generating $\pi$-system and applying the $\pi$–$\lambda$ theorem proves Eq. (18) as equality of probability measures.

**Step 4: common sign.** The bridge has fixed unit slope. Reflecting $S$ without $T$ would require a bridge with slope $-1$, which is outside the family. The same holds for reflecting $T$ alone. Therefore, the only remaining scalar transformation is the common reflection in Eq. (19). The identified factor relations then give the complete latent-observable law. □

### A.6 Discussion on Identifiability Results

We now discuss the scope of the result and its relation to the implementations and examples in this paper.

**Population result.** Theorem 1 is a population identifiability statement for the generative family. It does not establish consistency for a particular amortized variational estimator or for an auxiliary or tempered objective.

**Independent phenotype observations** The full theorem assumes an independently measured continuous cell phenotype. A score computed deterministically from the same counts, such as an expression-derived maturity score, is not an independent proxy. It does not satisfy the conditional independence required by Eq. (23).

**Latent Cell Representation.** We make no identifiability claim for the coordinates of $\boldsymbol{U}$ or $\boldsymbol{Z}$, library-size or dispersion parameters, neural-network weights, or attention weights. Constraining the decoder or the law governing the relationship between the latent variable $\boldsymbol{Z}$ and the observations $\boldsymbol{X}$ could imply identifiability results similar to iVAE. However, the latent variable $\boldsymbol{U}$ should be removed in that case.

## B Complete generative model and variational inference

We expand the factorization and learning objective summarized in Section 3. For donor $d$, let $\mathcal{I}_d$ index its cells; $\mathcal{H}_d$ and $\mathcal{U}_d$ index cells with observed and unobserved types, respectively; and $\mathcal{Y}_d \subseteq \mathcal{I}_d$ index cells with an observed cell phenotype. We treat batch labels and missingness patterns as fixed design information.

### B.1 FULL JOINT DISTRIBUTION

Conditional on $\boldsymbol{b}_d = \{b_n : n \in \mathcal{I}_d\}$, the joint distribution is

$$p_\theta(t_d, y_d, \{c_n, s_n, \boldsymbol{u}_n, \boldsymbol{z}_n, \ell_n, \boldsymbol{x}_n\}_{n\in\mathcal{I}_d}, \{y_n\}_{n\in\mathcal{Y}_d} \mid \boldsymbol{b}_d)$$
$$= p(t_d)p_\theta(y_d \mid t_d) \prod_{n\in\mathcal{I}_d} p_\theta(c_n \mid t_d)p_\theta(s_n \mid t_d, c_n)p(\boldsymbol{u}_n)p_\theta(\boldsymbol{z}_n \mid \boldsymbol{u}_n, c_n, s_n)p(\ell_n)p_\theta(\boldsymbol{x}_n \mid \boldsymbol{z}_n, \ell_n, b_n)$$
$$\prod_{n\in\mathcal{Y}_d} p_\theta(y_n \mid s_n). \tag{24}$$

An unobserved phenotype contributes no likelihood factor. For a binary donor phenotype, we replace the Gaussian donor observation model with

$$y_d \mid t_d \sim \text{Bernoulli}(\text{expit}(\delta_D + a_D t_d)).$$

### B.2 VARIATIONAL FAMILY

The variational family is

$$q_\phi(\boldsymbol{\xi}_d \mid \boldsymbol{o}_d) = q_\phi(t_d \mid \boldsymbol{h}_d) \prod_{n\in\mathcal{I}_d} q_\phi(\boldsymbol{z}_n \mid \boldsymbol{x}_n)q_\phi(\ell_n \mid \boldsymbol{x}_n)q_\phi(s_n \mid \boldsymbol{z}_n)$$
$$\times \prod_{n\in\mathcal{H}_d} q_\phi(\boldsymbol{u}_n \mid \boldsymbol{z}_n, c_n^{\text{obs}}) \prod_{n\in\mathcal{U}_d} q_\phi(c_n \mid \boldsymbol{z}_n)q_\phi(\boldsymbol{u}_n \mid \boldsymbol{z}_n, c_n), \tag{25}$$

where $\boldsymbol{\xi}_d$ contains all continuous latent variables and the unobserved cell types. All continuous factors are diagonal Gaussians, and $q_\phi(c_n \mid \boldsymbol{z}_n)$ is categorical. The inference networks receive no donor or cell phenotypes.

For an unobserved type and any type-dependent quantity $F(c_n)$, we evaluate the expectation exactly as

$$\mathbb{E}_{q_\phi(c_n|\boldsymbol{z}_n)}[F(c_n)] = \sum_{k=1}^{K} q_\phi(c_n = k \mid \boldsymbol{z}_n)F(k). \tag{26}$$

This enumeration affects only the low-dimensional latent terms. We evaluate the count decoder once per cell because it does not condition directly on $c_n$.

### B.3 OBSERVED-DATA ELBO

Let $\boldsymbol{o}_d$ collect the observed counts, donor phenotype, observed cell types, and observed cell phenotypes of donor $d$. The donor-level observed-data evidence lower bound is

$$\mathcal{L}_d(\theta, \phi) = \mathbb{E}_{q_\phi(\boldsymbol{\xi}_d|\boldsymbol{o}_d)} \left[\log p_\theta(\boldsymbol{o}_d, \boldsymbol{\xi}_d \mid \boldsymbol{b}_d) - \log q_\phi(\boldsymbol{\xi}_d \mid \boldsymbol{o}_d)\right]. \tag{27}$$

The complete generative objective is

$$\mathcal{L}_{\text{ELBO}}(\theta, \phi) = \sum_{d=1}^{D} \mathcal{L}_d(\theta, \phi). \tag{28}$$

For a cell with observed type, the ELBO contains $\mathbb{E}_q \log p_\theta(c_n^{\text{obs}} \mid t_d)$. For a cell with unobserved type, its categorical contribution is

$$\mathbb{E}_{q_\phi(t_d,\boldsymbol{z}_n)} \left[\sum_{k=1}^{K} q_\phi(c_n = k \mid \boldsymbol{z}_n) \log \frac{p_\theta(c_n = k \mid t_d)}{q_\phi(c_n = k \mid \boldsymbol{z}_n)}\right]$$
$$= -\mathbb{E}_{q_\phi(t_d,\boldsymbol{z}_n)} \text{KL}[q_\phi(c_n \mid \boldsymbol{z}_n) \,\|\, p_\theta(c_n \mid t_d)]. \tag{29}$$

We average the remaining type-dependent latent terms using Eq. (26).

For a continuous phenotype channel $Y \mid V = v \sim \mathcal{N}(\delta + av, \tau^2)$ and $q(V) = \mathcal{N}(\mu_V, \sigma_V^2)$, the expected negative log likelihood is available in closed form:

$$-\mathbb{E}_q \log p(Y = y \mid V) = \frac{(y - \delta - a\mu_V)^2 + a^2\sigma_V^2}{2\tau^2} + \frac{1}{2}\log(2\pi\tau^2). \tag{30}$$

### B.4 Supervised inference and cell phenotype likelihood

For cells with observed types, we also supervise the amortized classifier through

$$\mathcal{A}_{\text{class}} = \sum_{d=1}^{D} \sum_{n \in \mathcal{H}_d} \mathbb{E}_{q_\phi(\boldsymbol{z}_n \mid \boldsymbol{x}_n)} \log q_\phi(c_n^{\text{obs}} \mid \boldsymbol{z}_n). \tag{31}$$

The classifier-assisted estimator maximizes

$$\mathcal{L}_{\text{train}} = \mathcal{L}_{\text{ELBO}} + \lambda_{\text{class}} \mathcal{A}_{\text{class}}. \tag{32}$$

The additional term trains the inference classifier on observed labels. The corresponding generative factor for cell type remains $p_\theta(c_n \mid t_d)$.

Observed cell targets, including the expression-derived maturity score in Setting 2 and measured CAG repeat length in Setting 4, enter through the Gaussian likelihood $p_\theta(y_n \mid s_n)$ in Eq. (24). Its quadratic contribution can be written using

$$\mathcal{A}_{\text{score}} = -\sum_{d} \sum_{n \in \mathcal{S}_d} \mathbb{E}_{q_\phi(s_n \mid \boldsymbol{z}_n)} \left(y_n - g_\omega(s_n)\right)^2, \tag{33}$$

where $\mathcal{S}_d$ indexes cells with the score. Independent measurements, such as somatic *HTT* CAG repeat length, enter through the likelihood factor $p_\theta(y_n \mid s_n)$ in Eq. (24). In Setting 2, we also use the expression-derived maturity score as $y_n$ in the same Gaussian likelihood.

### B.5 Subsampling and unbiased rescaling

For a uniformly sampled subset $\mathcal{B}_d \subset \mathcal{I}_d$ of $M_d$ cells and fixed cellwise contributions $f_n$, the full sum has the unbiased estimator

$$\sum_{n \in \mathcal{I}_d} f_n \approx \frac{N_d}{M_d} \sum_{n \in \mathcal{B}_d} f_n. \tag{34}$$

Donor-level terms enter once per sampled donor. The within-donor averaging reported in Appendix D additionally divides cellwise sums by $N_d$, giving a sample mean over the sampled cells.

## C Simulation study and weight sensitivity

We used simulated data to select the donor- and cell-phenotype weights. The simulator draws a donor state $t_d^*$ for each donor $d$ and a cell state $s_n^*$ for each cell $n$. It then generates counts from these states, the cell type, the library size, and nuisance variation. Both latent states are therefore known, so we score each weighting by how well the fitted model recovers them. The symbols below describe the simulator and are local to this appendix.

We simulate six cohorts by crossing the number of donors $D \in \{5, 30, 100\}$ with the number of cells per donor $N_d \in \{100, 1000\}$ at data seed 0. These two quantities govern how much information each donor carries and how many donors contribute that information. Every cohort has $G = 20{,}000$ genes, $K = 6$ cell types, and twelve nuisance factors. The six cohorts contain $148{,}500$ cells in total. The cohorts are generated separately, so a small cohort is not a subsample of a larger one.

For donor $d$ and cell $n$ nested within that donor, the generator draws

$$\begin{aligned}
&t_d^* \sim \mathcal{N}(0,1), && y_d = 0.5 + 2t_d^* + \epsilon_d, && \epsilon_d \sim \mathcal{N}(0, 0.3^2),\\
&\boldsymbol{\pi}_d = \text{softmax}(\boldsymbol{\alpha} + t_d^* \boldsymbol{\beta}), && c_n \sim \text{Cat}(\boldsymbol{\pi}_d), && s_n^* = 0.4\, t_d^* + \delta_{c_n} + \eta_n,\\
&\eta_n \sim \mathcal{N}(0, \sigma_{c_n}^2), && y_n^{\text{cell}} = 3 + s_n^* + \xi_n, && \xi_n \sim \mathcal{N}(0, 0.5^2),
\end{aligned} \tag{35}$$

where $y_d$ is the donor phenotype and $\epsilon_d$ its observation noise; $\boldsymbol{\pi}_d \in \Delta^{K-1}$ is the cell-type composition of donor $d$, with $\alpha_k \sim \mathcal{N}(0, 0.5^2)$ and $\beta_k \sim \mathcal{N}(0, 0.8^2)$; $\delta_{c_n}$ is the state offset of the cell's type, a seed-fixed permutation of $\{-2.5, -1.5, -0.5, 0.5, 1.5, 2.5\}$; $\sigma_{c_n}$ is the within-type state spread, linearly spaced from $0.20$ to $1.10$ across the six types; and $y_n^{\text{cell}}$ is the cell phenotype with noise $\xi_n$. We provide $y_d$ and $y_n^{\text{cell}}$ during fitting. We reserve the two structural states $t_d^*$ and $s_n^*$ for evaluation. The type offset $\delta_{c_n}$ lies in the model class through the bridge intercepts $m_k$. The coefficient $0.4$ on

$t_d^*$ is absorbed by the scale of the fitted donor state, because the donor-head slope is free and the recovery score is invariant to scale. The standard-Normal prior on $t_d$ is the only term that penalizes the smaller scale.

The cell state affects expression through a progression program with a smooth part and a rough part. The smooth part maps the scalar state to two fixed features, $h_1(s) = s$ and $h_2(s) = \frac{1}{2}(s^2 - 1)$, collected in $\boldsymbol{h}(s) \in \mathbb{R}^2$. Here $h_1$ carries a linear progression effect and $h_2$ a quadratic one. Subtracting 1 centers the quadratic feature when $s$ has unit second moment. An indicator $\iota_g \sim \mathrm{Bernoulli}(0.30)$ marks whether gene $g$ belongs to the progression pool. Thus, an expected $30\%$ of genes respond to the state. A pooled gene carries a shared coefficient pair $\boldsymbol{\kappa}_g^{\mathrm{sh}} \sim \mathcal{N}(\mathbf{0}, 0.9^2\mathbf{I}_2)$, and a gene outside the pool has $\boldsymbol{\kappa}_g^{\mathrm{sh}} = \mathbf{0}$. Each cell type also modulates part of the pool. A second indicator $\iota_{kg} \sim \mathrm{Bernoulli}(0.50)$ selects which pooled genes type $k$ modulates. A selected gene carries a private pair $\boldsymbol{\kappa}_{kg}^{\mathrm{pr}} \sim \mathcal{N}(\mathbf{0}, 0.9^2\mathbf{I}_2)$, and every other gene has $\boldsymbol{\kappa}_{kg}^{\mathrm{pr}} = \mathbf{0}$. The effective coefficient pair of gene $g$ in type $k$ is $\boldsymbol{\kappa}_{kg} = 0.5\,\boldsymbol{\kappa}_g^{\mathrm{sh}} + 0.5\,\boldsymbol{\kappa}_{kg}^{\mathrm{pr}}$.

The rough part adds cell-specific variation that still depends on the state. A rough gene is drawn from within the progression pool with probability $\frac{1}{3}$, which leaves an expected $10\%$ of all genes rough. A rough gene carries a coefficient $r_g \sim \mathcal{N}(0, 0.9^2)$, and every other gene has $r_g = 0$. Each cell carries a jitter $\zeta_n \sim \mathcal{N}(0, 1)$. The two parts combine into the progression effect of gene $g$ in cell $n$,

$$M_{ng} = \sqrt{0.75}\,\boldsymbol{h}(s_n^*)^\top \boldsymbol{\kappa}_{c_n g} + \sqrt{0.25}\,s_n^* r_g(1 + \zeta_n). \tag{36}$$

The weight $0.25$ is a nominal squared amplitude. It does not make the rough term explain exactly a quarter of the realized expression variance. Because the rough term varies from cell to cell at fixed state, transcriptomically neighboring cells are less likely to share a state. This makes recovery harder without changing the underlying states or the supervision.

Let $d(n)$ denote the donor containing cell $n$. The unnormalized log-expression score of gene $g$ in cell $n$ is

$$q_{ng} = \beta_{0g} + a_{gc_n} + M_{ng} + b_{gd(n)} + \boldsymbol{f}_n^\top \boldsymbol{\Lambda}_g, \tag{37}$$

where $\beta_{0g} \sim \mathcal{N}(0, 1.2^2)$ is the baseline level of gene $g$; $a_{gk}$ is its type effect; $b_{gd}$ is its donor intercept; and $\boldsymbol{f}_n^\top \boldsymbol{\Lambda}_g$ is its nuisance term. Cell type 0 is the reference, so $a_{g0} = 0$. For every other type, $a_{gk}$ adds a small effect drawn from $\mathcal{N}(0, 0.25^2)$ to a marker boost drawn from $\mathcal{N}(2, 0.5^2)$. The marker boost is present with probability $0.15$, so an expected $15\%$ of nonreference gene-type pairs carry a marker. The donor intercepts share a per-gene spread, with $\tau_g = |\widetilde{\tau}_g|$ for $\widetilde{\tau}_g \sim \mathcal{N}(0, 0.45^2)$ and $b_{gd} \mid \tau_g \sim \mathcal{N}(0, \tau_g^2)$. The nuisance term uses twelve factors. The loading vector is $\boldsymbol{\Lambda}_g \sim \mathcal{N}(\mathbf{0}, 0.5^2\mathbf{I})$, and the factor score is $\boldsymbol{f}_n = \boldsymbol{\nu}_{c_n} + \boldsymbol{\chi}_n$. Here $\boldsymbol{\nu}_k \sim \mathcal{N}(\mathbf{0}, 1.1^2\mathbf{I})$ gives cells of one type a common location in factor space, and $\boldsymbol{\chi}_n \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ adds cell-specific variation. The product $\boldsymbol{f}_n^\top \boldsymbol{\Lambda}_g$ therefore creates expression that is correlated across genes and unrelated to the state.

We then generate counts from these scores. A softmax over genes maps $\boldsymbol{q}_n$ to relative expression $\rho_{ng}$. The expected count is $\mu_{ng} = L_n \rho_{ng}$, where $\log L_n \sim \mathcal{N}(\log 3000, 0.35^2)$ is the log expected library size of cell $n$. Gene-specific inverse dispersion is $\theta_g = 4\exp\{0.25\, z_g^{(0)} + \epsilon_g^{(\theta)}\}$, where $z_g^{(0)}$ is the baseline level $\beta_{0g}$ standardized across genes and $\epsilon_g^{(\theta)} \sim \mathcal{N}(0, 0.4^2)$. The dispersion therefore has base scale 4 and varies with the baseline level of the gene, with additional gene-specific log-normal variation. The generator then draws

$$x_{ng} \mid \mu_{ng}, \theta_g \sim \mathrm{NB}(\mu_{ng}, \theta_g), \qquad \mathbb{E}(x_{ng}) = \mu_{ng}, \qquad \mathrm{Var}(x_{ng}) = \mu_{ng} + \frac{\mu_{ng}^2}{\theta_g}, \tag{38}$$

as a Gamma-Poisson mixture with shape $\theta_g$ and scale $\mu_{ng}/\theta_g$. The simulation contains no batch effect.

We partition each cohort by donor. The number of query donors is $Q = \min\{D - 1, \max(1, \mathrm{round}(0.2D))\}$. We sample these donors without replacement from a stream derived from data seed 0. Among the remaining donors, the number of training donors is $F = \min\{D - Q - 1, \lceil 0.9(D - Q)\rceil\}$, and the rest become validation donors. This gives training, validation, and query counts of $3/1/1$, $22/2/6$, and $72/8/20$ for $D = 5$, 30, and 100. We exclude query donors from fitting and score orientation, as in the real-data settings.

We fix $\lambda_{\text{comp}} = \lambda_{\text{class}} = 1$ and search the weights of the two phenotype channels. We search $\lambda_{\text{donor}}$ and $\lambda_{\text{cell}}$ over $\{0.1, 1, 10, 100, 500, 1000\}$. This gives $6^2 = 36$ weightings per cohort and 216 fitted models in total. We hold the composition weight at one because that term is the untempered likelihood of the observed cell-type labels. We hold the classification weight at one because that term lies outside the bound. Every run uses the architecture and optimizer of Appendix D and the schedules of Appendix D.

We score a weighting by the mean of two absolute Pearson correlations on query donors. The first is $|r(\widehat{s}, s^*)|$ between the inferred and simulated cell states, and the second is $|r(\widehat{t}, t^*)|$ between the inferred and simulated donor states. We calibrate each cell state on training cells before computing the correlation. We call the mean of the two correlations the recovery score. The two five-donor cohorts hold out a single donor, so their donor correlation is undefined. For these two cohorts, the recovery score is the cell term alone.

Appendix Figure 4 summarizes the result. Panels (a) to (c) use the selected weighting, $\lambda_{\text{comp}} = \lambda_{\text{class}} = 1$, $\lambda_{\text{donor}} = 10$ and $\lambda_{\text{cell}} = 500$. The inferred cell state tracks the simulated cell state with $|r|$ between $0.969$ and $0.992$ across the six cohorts. The inferred donor state tracks the simulated donor state with $|r|$ between $0.783$ and $0.996$ in the four cohorts where that correlation is defined. The lowest correlation comes from the cohort with 30 donors of 100 cells, whose six query donors give the least stable correlation. As expected, recovery improves with cohort size at both levels.

Panels (d) to (f) report the sensitivity analysis. The weight $\lambda_{\text{cell}}$ dominates the recovery score. Averaged over cohorts, its marginal rises from $0.58$ at $\lambda_{\text{cell}} = 0.1$ to $0.94$ at $1000$. It plateaus from $\lambda_{\text{cell}} = 100$ onward. We select $\lambda_{\text{cell}} = 500$ from this plateau. The weight $\lambda_{\text{donor}}$ has little effect on the score, because every simulated cohort includes cell phenotypes and the cell channel already anchors the hierarchy. Its pooled marginal spans $0.77$ to $0.86$, and the per-cohort marginals peak at $\lambda_{\text{donor}} = 10$ in three of the six cohorts. Settings 1 and 3 train without the cell channel, so the donor channel is their only phenotype channel. We set $\lambda_{\text{donor}} = 10$. The selected weighting ranks between second and ninth of the 36 weightings in every cohort. Its recovery score is within $0.003$ of the per-cohort best in five of the six cohorts and within $0.08$ in the sixth. Panel (f) shows the score of every weighting in each cohort, with the selected weighting and the per-cohort best marked.

# D Implementation, optimization, and prediction

## D.1 Architecture

The expression encoder parameterizes $q_\phi(\boldsymbol{z}_n \mid \boldsymbol{x}_n)$ and $q_\phi(\ell_n \mid \boldsymbol{x}_n)$ from $\log(1 + \boldsymbol{x}_n)$. Separate networks parameterize $q_\phi(s_n \mid \boldsymbol{z}_n)$, $q_\phi(c_n \mid \boldsymbol{z}_n)$, and $q_\phi(\boldsymbol{u}_n \mid \boldsymbol{z}_n, c_n)$. The conditional prior of $\boldsymbol{z}_n$ takes $[\boldsymbol{u}_n; \text{onehot}(c_n); s_n]$ as input, and the count decoder takes $[\boldsymbol{z}_n; \text{onehot}(b_n)]$. Class-stratified gated attention maps the inferred cell states and type assignments to the donor posterior $q_\phi(t_d \mid \boldsymbol{h}_d)$. Appendix Table 2 gives the specific architecture and optimization configuration.

We evaluate the negative-binomial likelihood on raw counts. The decoder produces logits for all retained genes and applies a single softmax across genes. We then evaluate the log likelihood in chunks to reduce peak memory without changing the objective.

## D.2 Optimization

Training uses the terms in Eq. (32), with the within-donor averaging specified below. Each training step samples donors and, when needed, cells within donors. Cellwise terms are rescaled according to Eq. (34). We use the reparameterization estimator for continuous latent variables and evaluate sums over unobserved cell types exactly.

The donor-state divergence coefficient follows

$$\rho(e) = \min\{1, e/W\}, \tag{39}$$

where $e$ is the epoch and $W$ the warm-up length, with $W = 25$. The coefficient equals one after the warm-up. The final objective is the ELBO in Eq. (28) with an additional cell type supervision term. In the reported runs, we fix the channel variances at $\tau_D^2 = 1/20$ and $\tau_C^2 = 1/1000$, so the squared channel errors carry the weights $\lambda_{\text{donor}} = 10$ and $\lambda_{\text{cell}} = 500$. We average every cellwise term of

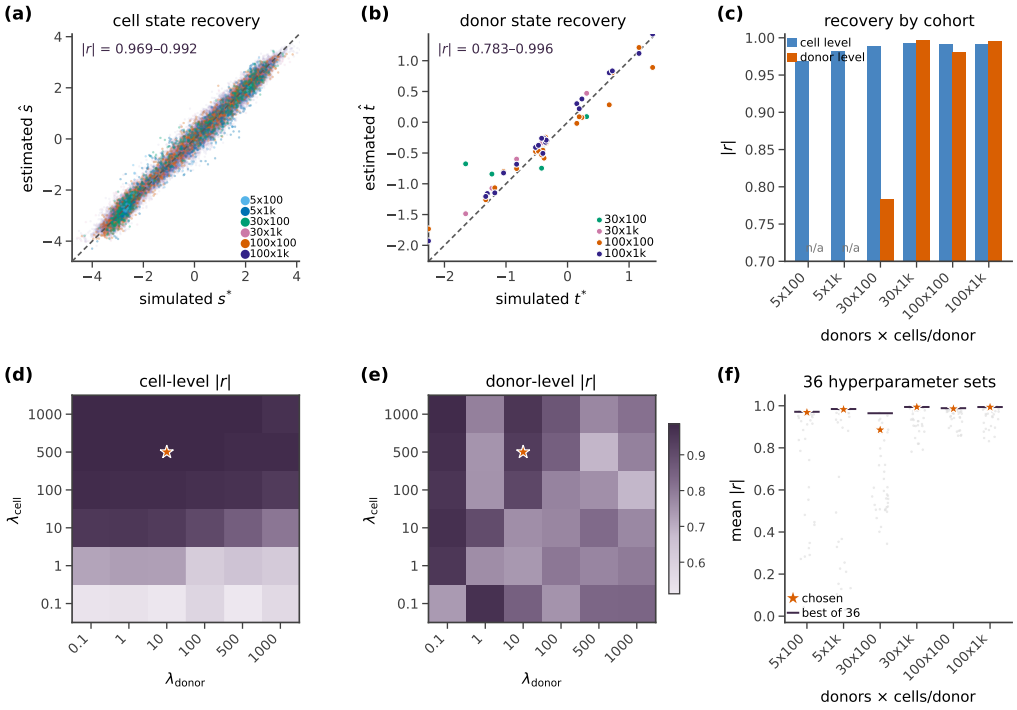



Figure 4: **Simulation recovery and weight sensitivity.** **(a)** Held-out cells of all six simulated cohorts, showing the train-calibrated cell state against the simulated cell state. **(b)** Held-out donors of the four cohorts that hold out more than one donor. **(c)** Cell-level and donor-level $|r|$ by cohort, labeled donors $\times$ cells per donor. **(d)** Cell-level recovery $|r(\widehat{s}, s^*)|$ over $\lambda_{\text{donor}}$ and $\lambda_{\text{cell}}$, averaged over cohorts, with the selected weighting starred. **(e)** Donor-level recovery $|r(\widehat{t}, t^*)|$ over the same grid, averaged over the four cohorts where it is defined. **(f)** All 36 weightings scored per cohort, with the selected weighting and the per-cohort best marked.

Table 2: scPACE architecture and optimization configuration

| Parameter | Configuration |
|---|---|
| Cell representation / auxiliary dimensions | $d_z = d_u = 10$ |
| Hidden networks | 2 layers, width 128 |
| Hidden-layer operations | LayerNorm, ReLU, dropout 0.1 |
| MIL feature / attention width | 32 / 32 |
| Gene chunk size | 2,048 |
| Optimizer / learning rate | Adam / $10^{-3}$ |
| Gradient clipping | global norm 5 |
| Numerical precision | BF16 |
| Maximum epochs / early-stopping patience | 100 / 45 |
| Training seed | 0 |
| Classifier coefficient $\lambda_{\text{class}}$ | 1 |
| Auxiliary-score coefficient, if used | $\lambda_{\text{cell}} = 500$ in Setting 2; not used in Settings 1, 3, and 4 |
| Donor channel variance $\tau_D^2$ | $1/20$, i.e. weight $\lambda_{\text{donor}} = 10$ on the squared error |
| Cell channel variance $\tau_C^2$ (Setting 4) | $1/1000$, i.e. weight $\lambda_{\text{cell}} = 500$ |
| $\text{KL}_t$ warm-up $W$ / channel warm-start | 25 epochs / 1000 decaying geometrically to the target over 25 epochs |
| Donor split seed | 1024 |
| Label-transfer threshold $T$ | 0.6 |

a donor over its $N_d$ cells and count the donor-level terms once. Both phenotype channels start at an effective weight of 1000 that decays geometrically to the target weight over the first 25 epochs. We base early stopping on the unscheduled validation objective.

### D.3 Class-stratified donor aggregation

Let

$$r_{nk} = \begin{cases} \mathbb{I}(c_n^{\text{obs}} = k), & n \in \mathcal{H}_d, \\ 0, & n \in \mathcal{U}_d \text{ and } |\mathcal{H}_d| > 0, \\ \mathbb{I}\{\arg\max_{k'} q_\phi(c_n = k' \mid \boldsymbol{z}_n) = k\}, & n \in \mathcal{U}_d \text{ and } |\mathcal{H}_d| = 0, \end{cases} \tag{40}$$

be the type assignment indicator. For a bag with observed types, we pool only its typed cells. For an entirely untyped bag, we pool every cell under its most probable type. At query time, we include an unannotated cell under its transferred type when its confidence is at least $T$ and exclude it from pooling otherwise. A type-specific transformation and gated-attention score are

$$\begin{aligned} \boldsymbol{v}_{nk} &= \mathrm{ReLU}(W_k s_n + \boldsymbol{b}_k), \\ e_{nk} &= \boldsymbol{w}_k^\top \left[\tanh(V_k \boldsymbol{v}_{nk} + \boldsymbol{b}_k^V) \odot \sigma(U_k \boldsymbol{v}_{nk} + \boldsymbol{b}_k^U)\right] + b_k^a. \end{aligned} \tag{41}$$

Attention pooling within each type gives

$$\begin{aligned} a_{nk} &= \frac{r_{nk} \exp(e_{nk})}{\sum_{j \in \mathcal{I}_d} r_{jk} \exp(e_{jk})}, & \boldsymbol{h}_{dk} &= \sum_{n \in \mathcal{I}_d} a_{nk} \boldsymbol{v}_{nk}, \\ p_{dk} &= \frac{\sum_{n \in \mathcal{I}_d} r_{nk}}{\sum_{n \in \mathcal{I}_d} \sum_{k'} r_{nk'}}, & \boldsymbol{h}_d &= [\boldsymbol{h}_{d1}; \ldots; \boldsymbol{h}_{dK}; \boldsymbol{p}_d]. \end{aligned} \tag{42}$$

If a type has no pooled cell in a donor, we set its summary and proportion to zero. The aggregation is invariant to cell order and has fixed dimension across donors.

### D.4 Reported states and phenotype predictions

We use posterior means as point estimates:

$$\widehat{\boldsymbol{z}}_n = \mathbb{E}_q(\boldsymbol{z}_n \mid \boldsymbol{x}_n), \qquad \widehat{s}_n = \mathbb{E}_q(s_n \mid \widehat{\boldsymbol{z}}_n), \qquad \widehat{t}_d = \mathbb{E}_q(t_d \mid \widehat{\boldsymbol{h}}_d). \tag{43}$$

For continuous phenotypes, the corresponding posterior-mean predictions are

$$\widehat{y}_d = \delta_D + a_D \widehat{t}_d, \qquad \widehat{y}_n = \delta_C + a_C \widehat{s}_n. \tag{44}$$

For a binary donor phenotype, the predicted probability is $\mathrm{expit}(\delta_D + a_D \widehat{t}_d)$.

For an unannotated cell, the transferred type is

$$\widehat{c}_n = \arg\max_k q_\phi(c_n = k \mid \widehat{\boldsymbol{z}}_n) \tag{45}$$

when the maximum probability is at least $T$; otherwise, the model abstains. We retain the full probability vector for evaluation.

## E Cohorts, preprocessing, and phenotype construction

We define all partitions at the donor or mouse-sample level and assign every cell from the same biological unit to the same split (Appendix Table 3). We exclude query donors from model fitting, feature selection, target standardization, and hyperparameter selection (Appendix Table 4).

### E.1 Cohorts and supervision regimes

**Mouse visual cortex.** The mouse cohort combines 568,654 single-cell profiles from 53 sample units in the developmental atlas of Gao et al. (2025) with 296,358 single-nucleus profiles from 35 GSE190940 sample units (Cheng et al., 2022). The Gao atlas spans embryonic through adult development. GSE190940 covers postnatal development under normal rearing, dark rearing, and

Table 3: Cohorts and phenotype information. Expression-derived scores are not independently measured phenotypes.

| Cohort | Cells | Donors | Genes | Types | Donor phenotype | Cell-level information |
|---|---|---|---|---|---|---|
| Mouse visual cortex | 865,012 | 88 | 30,566 | 15 | chronological age | maturity score derived from expression for evaluation (Setting 1) or cell phenotype supervision (Setting 2) |
| SEAAD MTG | 188,307 | 84 | 36,601 | 3 | continuous pseudoprogression score | DAM-versus-homeostatic expression score used for evaluation |
| Huntington's disease | 302,191 | 8 | 29,907 | 7 | CAP score | independently measured somatic *HTT* CAG repeat length |

Table 4: Donor partitions and query populations.

| Cohort | Split seed | Training | Validation | Query | Query cells | Scored cells |
|---|---|---|---|---|---|---|
| Mouse visual cortex | 1024 | 62 | 6 | 20 | 176,614 | 176,614 |
| SEAAD MTG glia | 1024 | 60 | 7 | 17 | 37,509 | 6,417 |
| HD caudate | 1024 | 4 | 1 | 3 | 118,204 | 11,047 |

dark rearing followed by light re-exposure. Dataset identity defines the two decoder batches. The Gao atlas provides the 15 cell-type annotations. We treat GSE190940 cells as unannotated for label transfer. Chronological age relative to birth is the donor phenotype.

We use the expression-derived maturity score defined below for cell-level evaluation. In Setting 2, we also use it to supervise training cells through the Gaussian cell-phenotype likelihood, with weight $\lambda_{\text{cell}} = 500$ in Eq. (33) (Appendix B).

**SEAAD middle temporal gyrus.** The Alzheimer's disease cohort contains 188,307 glial nuclei from 84 SEAAD middle temporal gyrus donors across 177 technical batches (Gabitto et al., 2024). We retain astrocytes, oligodendrocytes, and Microglia-PVM for the annotation task. We use the continuous pseudoprogression score for donor supervision. No cell-level phenotype is used for fitting. We withhold the DAM-versus-homeostatic score defined below from every method and use it only to evaluate cell states in query Microglia-PVM nuclei.

**Huntington's disease caudate.** The Huntington's disease cohort combines caudate single-nucleus expression profiles with single-cell PacBio measurements of somatic *HTT* CAG repeat length (Handsaker et al., 2025). It contains 302,191 nuclei from eight donors. The seven broad classes are spiny projection neurons, interneurons, astrocytes, oligodendrocytes, polydendrocytes, microglia, and endothelial cells. We use the CAP score for donor supervision (Warner et al., 2022). CAG repeat lengths are available for 29,030 nuclei, including 11,047 query nuclei. Cells without measurements remain in the count model but contribute no cell-phenotype likelihood.

### E.2 Count preprocessing and technical covariates

We retain a gene when it has nonzero counts in at least

$$\max\{10, \lfloor 0.001 N_{\text{train}} \rfloor\} \tag{46}$$

training cells. We learn this filter from training donors and apply it unchanged to validation and query cells (Appendix Table 5). We supply the retained raw counts to the negative-binomial likelihood and use $\log(1+x)$ only as an encoder input. We also define decoder batch categories from training donors. We estimate the library-size prior moments from training cells using $\log \max(1, \sum_g x_{ng})$.

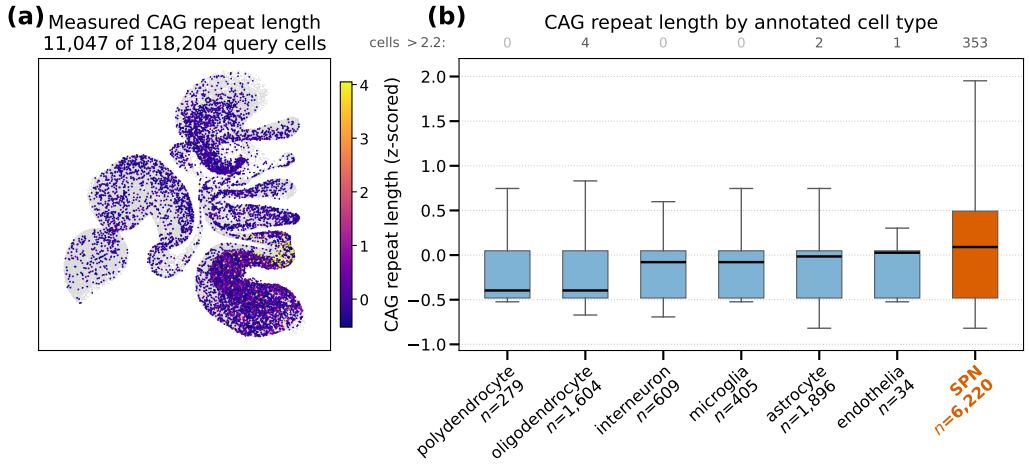



Figure 5: **Measured CAG repeat length in Huntington's disease query nuclei.** **(a)** Query-cell embedding colored by measured somatic *HTT* CAG repeat length, shown as z-scores. Measurements are available for 11,047 of 118,204 query nuclei; nuclei without measurements are shown in gray. **(b)** CAG repeat length distributions across the seven annotated cell types. Labels below the boxes give the number of measured nuclei in each type. Numbers above the panel give the number of nuclei with a CAG repeat length z-score greater than 2.2. Spiny projection neurons (SPNs) are highlighted in orange.

Table 5: Prepared and fitted feature dimensions.

| Cohort | Prepared genes | Fitted genes | Prepared batches | Fitted batches |
|---|---|---|---|---|
| Mouse visual cortex | 30,566 | 21,847 | 2 | 2 |
| SEAAD MTG glia | 36,601 | 26,462 | 177 | 125 |
| HD caudate | 29,907 | 29,245 | 38 | 15 |

### E.3 Phenotype scaling

We standardize donor phenotypes using the mean and standard deviation of the training donors. We standardize independent cell phenotypes using measured cells from training donors. We then apply these transformations unchanged to validation and query data. Missing values are never imputed for likelihood evaluation.

### E.4 Rank-based expression scores

For the mouse maturity and SEAAD DAM analyses, we use an AUCell-like rank score (Aibar et al., 2017). Within cell $n$, let $r_{ng} \in \{0, \dots, G-1\}$ be the rank of gene $g$ by decreasing expression, and let $L = \max\{1, \mathrm{round}(0.05G)\}$. For marker set $\mathcal{S}$, the score is

$$A_n(\mathcal{S}) = \frac{1}{|\mathcal{S}|} \sum_{g \in \mathcal{S}} \max\left(1 - \frac{r_{ng}}{L}, 0\right). \tag{47}$$

We compute these scores before scPACE-specific gene filtering.

**Mouse maturity score.** The mature-marker set is

$$\mathcal{M} = \{\textit{Rbfox3}, \textit{Satb2}, \textit{Cux2}, \textit{Mef2c}, \textit{Tbr1}, \textit{Snap25}, \textit{Syt1}, \textit{Grin1}, \textit{Camk2a}, \textit{Mapt}\},$$

and the progenitor-marker set is

$$\mathcal{P} = \{\textit{Sox2}, \textit{Pax6}, \textit{Hes1}, \textit{Hes5}, \textit{Nes}, \textit{Vim}, \textit{Mki67}, \textit{Top2a}, \textit{Ccnd2}, \textit{Eomes}, \textit{Neurog2}, \textit{Btg2}\}.$$

Let $z_{\mathcal{R}}(\cdot)$ denote standardization using reference population $\mathcal{R}$. The maturity contrast is

$$\Delta_n^{\mathrm{mat}} = \frac{z_{\mathcal{R}}(A_n(\mathcal{M})) - z_{\mathcal{R}}(A_n(\mathcal{P}))}{2}, \qquad m_n = z_{\mathcal{R}}(\Delta_n^{\mathrm{mat}}). \tag{48}$$

Higher values indicate stronger mature relative to progenitor expression.

**SEAAD DAM-versus-homeostatic score.** The DAM set is

$\mathcal{D} = \{$*APOE*, *TYROBP*, *TREM2*, *LPL*, *CST7*, *CTSB*, *CTSD*, *CD9*, *ITGAX*, *GPNMB*, *SPP1*, *LGALS3*, *APOC1*, *AXL*$\}$

(Keren-Shaul et al., 2017), and the homeostatic set is

$$\mathcal{H} = \{\textit{P2RY12}, \textit{P2RY13}, \textit{TMEM119}, \textit{CX3CR1}, \textit{SALL1}, \textit{GPR34}, \textit{HEXB}\}$$

(Butovsky et al., 2014). The evaluation score is

$$\Delta_n^{\mathrm{AD}} = \frac{z_{\mathcal{R}_{\mathrm{AD}}}(A_n(\mathcal{D})) - z_{\mathcal{R}_{\mathrm{AD}}}(A_n(\mathcal{H}))}{2}, \qquad S_n^{\mathrm{AD}} = z_{\mathcal{R}_{\mathrm{AD}}}(\Delta_n^{\mathrm{AD}}). \tag{49}$$

Higher values indicate stronger DAM relative to homeostatic expression. We exclude monocytes, lymphocytes, and nuclei with extreme PVM/monocyte marker scores using thresholds estimated from non-query donors.

**Measured HD target.** The HD cell phenotype is the PacBio-measured somatic *HTT* CAG repeat length. Because it is measured independently of the RNA counts, we model it through $p_\theta(y_n \mid s_n)$. We omit missing measurements without imputing them.

# F Baselines and evaluation protocol

## F.1 Held-out-donor benchmark

We evaluate all methods on the same query donors. We fit the representation-learning baselines on the scPACE training donors and use validation donors only for early stopping. When the software for a scalar trajectory or neighborhood method does not provide a separate validation mechanism, we fit it on the non-query reference population. We learn every preprocessing transformation without query cells or query phenotypes (Appendix Table 6).

Table 6: Outputs included in the benchmark. A dash denotes an output that the method does not provide or that is not evaluated.

| Method | Cell scalar | Donor prediction | Type accuracy | Donor mixing |
|---|---|---|---|---|
| Sceptic | Yes | – | – | – |
| psupertime | Yes | – | – | – |
| PhenoPath | Yes | – | – | – |
| Slingshot | Yes | – | – | – |
| CNA | Yes | – | – | – |
| MELD | Yes | – | – | – |
| scVI | – | – | – | Yes |
| scANVI | – | – | Yes | Yes |
| scVIP | – | Yes | – | Yes |

## F.2 Representation-learning baselines

**scVI.** We fit scVI to raw counts with the same training-cell indices, gene set, and technical batch labels as scPACE (Lopez et al., 2018). We supply neither phenotypes nor cell types. We report its posterior-mean cell representation for within-type donor-mixing evaluation.

**scANVI.** scANVI extends scVI with semi-supervised cell-type inference (Xu et al., 2021). We treat training cells with supplied labels as labeled and all query cells as unlabeled. We report the posterior-mean representation and the complete cell-type probability vector.

**scVIP.** scVIP links single-cell profiles and type composition to donor phenotypes using a donor latent and type-stratified attention (Lai et al., 2026). We fit it with the same donor partition and supplied cell-type labels as scPACE. We report its donor phenotype prediction and cell representation. Because scVIP does not produce a nested scalar cell state, we do not assign it a cell-level phenotype metric.

### F.3 Scalar cell-state baselines

For the scalar methods, we first apply $\log(1+x)$ and select highly variable genes on the reference cells. We then standardize these genes and compute a reference PCA basis. We apply the same basis to query cells.

***psupertime*** learns a sparse proportional-odds model from ordered labels (Macnair et al., 2022). ***Sceptic*** learns a supervised ordering from sequential labels (Li et al., 2025). We convert continuous reference-donor phenotypes to ordered stages using quantiles computed from reference donors only. psupertime scores query cells directly with its learned linear coefficients. Sceptic applies its trained classifier to the fixed query features. ***PhenoPath*** fits a covariate-dependent latent trajectory to reference cells (Campbell & Yau, 2018). ***Slingshot*** fits lineage curves in a reference embedding and projects query cells onto those fixed curves (Street et al., 2018). For cells associated with multiple Slingshot lineages, we average lineage pseudotimes using the native lineage weights. Slingshot also requires a root cluster for its lineage tree. We cluster the reference cells by $k$-means. We then take the root to be the cluster with the lowest mean donor phenotype computed from training donors. ***CNA*** estimates phenotype-associated cellular neighborhoods from reference donors (Reshef et al., 2022). ***MELD*** estimates condition-associated density on a cell graph (Burkhardt et al., 2021). We extend their native reference-cell scores to query cells using the common projection below. We omit CNA when its minimum reference-donor requirement is not met.

### F.4 Projection of reference-only scores

For methods without a native out-of-sample operator, let $\boldsymbol{r}_n$ be the fixed PCA coordinates of query cell $n$. Let $\mathcal{N}_k(n)$ contain its $k = 30$ nearest scored reference cells. If reference cell $j$ has score $v_j$, we use inverse-distance interpolation,

$$\widehat{s}_n^{\mathrm{proj}} = \frac{\sum_{j\in\mathcal{N}_k(n)} w_{nj} v_j}{\sum_{j\in\mathcal{N}_k(n)} w_{nj}}, \qquad w_{nj} = \frac{1}{\max\{\|\boldsymbol{r}_n - \boldsymbol{r}_j\|_2, \epsilon\}}. \tag{50}$$

We obtain the PCA basis, neighbors, and reference scores without query phenotypes.

### F.5 Association and prediction metrics

For eligible query units $i \in \mathcal{Q}$, we compute Pearson correlation as

$$r = \frac{\sum_{i\in\mathcal{Q}} (\widehat{v}_i - \bar{v})(y_i - \bar{y})}{\sqrt{\sum_{i\in\mathcal{Q}} (\widehat{v}_i - \bar{v})^2}\sqrt{\sum_{i\in\mathcal{Q}} (y_i - \bar{y})^2}}. \tag{51}$$

We report $|r|$ because the scalar-state orientation is arbitrary unless an orientation constraint is imposed.

Before computing the scPACE donor and cell errors and the cell errors of every scalar baseline, we calibrate each inferred state to its target. We fit this affine calibration on the reference population, separately from the learned channel heads:

$$(\widehat{a}, \widehat{b}) = \arg\min_{a,b} \sum_{i\in\mathcal{R}} \left[y_i - (a + b\widehat{v}_i)\right]^2, \qquad \widetilde{y}_j = \widehat{a} + \widehat{b}\widehat{v}_j, \quad j \in \mathcal{Q}. \tag{52}$$

We fit the calibration once on training and validation units and apply it unchanged to query units.

We compute donor root mean squared error as

$$\mathrm{RMSE}_{\mathrm{donor}} = \sqrt{\frac{1}{|\mathcal{Q}_D|} \sum_{d\in\mathcal{Q}_D} (y_d - \widetilde{y}_d)^2}. \tag{53}$$

We normalize cell error by the query-target standard deviation,

$$\mathrm{nRMSE}_{\mathrm{cell}} = \frac{\sqrt{|\mathcal{Q}_C|^{-1} \sum_{n \in \mathcal{Q}_C} (y_n - \widetilde{y}_n)^2}}{\sigma_{\mathcal{Q}_C}(y)}. \tag{54}$$

We include only pairs with finite predictions and targets.

### F.6 Cell-type accuracy

For annotated query cells, we compute cell-type accuracy as

$$\mathrm{Acc} = \frac{1}{|\mathcal{Q}_{\mathrm{ann}}|} \sum_{n \in \mathcal{Q}_{\mathrm{ann}}} \mathbb{I}\{\widehat{c}_n = c_n^{\mathrm{obs}}\}. \tag{55}$$

We use the highest-probability class to compute accuracy regardless of the reporting threshold. We may additionally report selective accuracy and coverage at threshold $T$.

### F.7 Marker-gene validation of imputed cell types

Table 1 reports cell-type accuracy against the observed annotation. That measure covers only the annotated Gao et al. (2025) mouse cohort, so we also ask whether the imputed labels are biologically valid on their own. We group the 176,614 query cells of Setting 2 by the imputed cell type label, the argmax of $q_\phi(c \mid \widehat{z}_n)$. Expression is the observed count matrix, normalized to 10,000 counts per cell and log-transformed. Appendix Figure 6 shows the marker genes under different cell types. As a result, predicted type classes on query cells carry a marker whose expression is highest in the class it names.

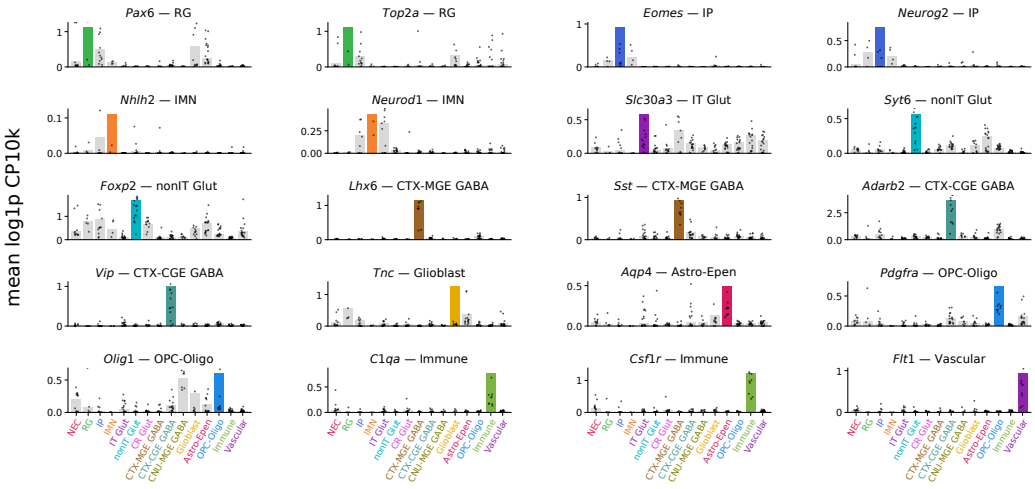



Figure 6: **Gene markers by imputed cell type on Setting 2 query cells.** One panel per marker gene, ordered by target class along the developmental lineage. Bars give the mean log1p CP10k of that gene in each imputed class, with the target class in its class color. Black points are per-donor means over donors contributing at least 20 cells to the class. Cells are grouped by the imputed cell type label.

### F.8 Within-type donor mixing

We sample at most 15,000 query cells using a fixed seed. Within each supplied cell-type group, let $\mathcal{N}_n$ be the 15 nearest other cells in the representation. We compute the observed cross-donor fraction as

$$m_n = \frac{1}{|\mathcal{N}_n|} \sum_{j \in \mathcal{N}_n} \mathbb{I}\{d(j) \neq d(n)\}. \tag{56}$$

For type $k$, we define the random-mixing reference as

$$e_k = 1 - \sum_d p_{dk}^2, \tag{57}$$

where $p_{dk}$ is donor $d$'s proportion within the sampled cells of type $k$. We report the normalized mixing ratio

$$M = \frac{\sum_{n\in\mathcal{E}} m_n}{\sum_{n\in\mathcal{E}} e_{k(n)}}, \tag{58}$$

where $\mathcal{E}$ contains cells from groups represented by at least two donors. Values near one indicate mixing close to the random-donor reference.

## G Ablation study on the mouse cohort

We evaluate the contributions of individual scPACE components on the mouse cohort. We focus on Setting 1, which uses donor-level supervision without cell-level phenotypes. In this setting, donor-level information reaches the cells through the bridge $p(s_n \mid t_d, c_n)$, the composition model $p(c_n \mid t_d)$, and the donor aggregator. We then examine Setting 2 and vary the weight of the cell channel.

**Protocol.** All variants use the data, donor partition, gene set, and training configuration of the run in Table 1. We change only the named component and perform no additional tuning. The remaining architecture, loss weights, schedules, optimizer, numerical precision, and early-stopping rule are unchanged. Each reported variant is trained for at most 100 epochs with seed 0. We estimate sampling uncertainty using percentile bootstrap intervals. For donor metrics, we resample the 20 query donors 2000 times. For cell metrics, we resample donors 500 times and retain all cells from each sampled donor, preserving dependence among cells from the same donor. Appendix Figure 7 shows the 95% intervals for the ablated variants and the full-model point estimate.

**Measuring cell-state association in Setting 1.** The pooled correlation in Table 1 measures association between the inferred cell state and the maturity score across all 176,614 query cells. It reflects both within-type ordering and the relative positions of cell types along the continuum. These components rely on different information. The bridge $p(s_n \mid t_d, c_n)$ and donor phenotype provide training signals for within-type cell-state association. The type intercepts $m_k$ determine the relative positions of types, and the cell channel identifies these intercepts (Appendix A). Without this channel in Setting 1, the counts do not identify the intercepts. The conditional prior of $\boldsymbol{z}_n$ receives $\mathrm{onehot}(c_n)$, so $\boldsymbol{\mu}_\theta^p$ can absorb a type-specific shift in $s_n$ without changing the likelihood. Setting 1 therefore motivates evaluation within types. Setting 2 uses the expression-derived score through the cell-phenotype channel and improves pooled correlation from $0.08$ to $0.97$ as well as within-type association (Appendix Table 8). Cell type explains $0.863$ of the maturity-score variance among annotated query cells. The pooled correlation in Setting 1 therefore depends strongly on the relative positions of types.

To evaluate components that support within-type ordering, we use two within-type association measures. Let $\mathcal{Q}_{\mathrm{ann}}$ denote the annotated query cells and $\mathcal{Q}_k = \{n \in \mathcal{Q}_{\mathrm{ann}} : c_n = k\}$ the cells of type $k$. The type-specific means are

$$\bar{s}_k = \frac{1}{|\mathcal{Q}_k|}\sum_{n\in\mathcal{Q}_k} \widehat{s}_n, \qquad \bar{y}_k = \frac{1}{|\mathcal{Q}_k|}\sum_{n\in\mathcal{Q}_k} y_n, \tag{59}$$

where $\widehat{s}_n$ is the inferred cell state and $y_n$ is the maturity score. The demeaned correlation subtracts the corresponding type means and computes a single correlation across all annotated query cells,

$$r_{\mathrm{dm}} = \left| r\big(\{\widehat{s}_n - \bar{s}_{c_n}\}_{n\in\mathcal{Q}_{\mathrm{ann}}},\, \{y_n - \bar{y}_{c_n}\}_{n\in\mathcal{Q}_{\mathrm{ann}}}\big)\right|, \tag{60}$$

where $r(\cdot,\cdot)$ is the Pearson correlation defined in Eq. (51). Each cell contributes once, so abundant types contribute more observations. The median correlation computes the absolute correlation within each type and takes the median,

$$r_{\mathrm{med}} = \underset{k:\,|\mathcal{Q}_k|\geq 200}{\mathrm{median}} \left| r\big(\{\widehat{s}_n\}_{n\in\mathcal{Q}_k}, \{y_n\}_{n\in\mathcal{Q}_k}\big)\right|, \tag{61}$$

including types with at least 200 query cells. Each eligible type contributes one value. Appendix Table 8 reports both measures and the pooled correlation from Table 1 for all methods evaluated on the mouse cohort.

For the scPACE Setting 1 run in Table 1, the demeaned correlation is $0.22$ and the median correlation is $0.45$. All twelve annotated query types show positive correlations with the maturity score, ranging from $0.60$ for intermediate progenitors to $0.02$ for IT glutamatergic neurons. This common direction is consistent with the bridge. IT glutamatergic neurons account for 46,513 of the 104,327 annotated query cells, but the model recovers little maturity signal within this type. The difference between the summaries reflects how they combine these heterogeneous within-type associations. The demeaned correlation depends on both type abundance and within-type variation in the inferred state and maturity score. The median reflects association within a typical cell type. By this measure, scPACE in Setting 1 ranks highest among methods receiving only donor-level supervision ($0.45$, compared with $0.40$ for PhenoPath), although it ranks lowest by the demeaned measure. We use the median as the primary ablation metric because the Setting 1 ablations assess within-type ordering and each eligible type contributes equally. We also report the demeaned correlation to show how the choice of summary affects the results.

**Hierarchy variants.** *No bridge and* $\boldsymbol{\beta} = \mathbf{0}$ replaces $p(s_n \mid t_d, c_n) = \mathcal{N}(t_d + m_k, \sigma^2_{s,k})$ with $\mathcal{N}(m_k, \sigma^2_{s,k})$ and fixes $\boldsymbol{\beta} = \mathbf{0}$ in $p(c_n \mid t_d) = \mathrm{Cat}(\mathrm{softmax}(\boldsymbol{\alpha} + t_d\boldsymbol{\beta}))$. Both generative donor-to-cell edges are removed while retaining the aggregator, and the donor phenotype provides the supervision for learning the donor state. *No type intercepts* fixes $m_k = 0$, giving every type the same prior mean $t_d$. These intercepts allow cell types, such as progenitors and neurons, to occupy different baseline positions while shifting together with donor state. *No auxiliary variable* sets $d_u = 0$ and removes $\boldsymbol{u}_n$ from $p(\boldsymbol{z}_n \mid \boldsymbol{u}_n, c_n, s_n)$. Cellular variation beyond the scalar state and type label then has no separate latent variable. *Standard-normal prior on* $\boldsymbol{z}_n$ replaces this conditional prior with $\mathcal{N}(\mathbf{0}, \mathbf{I})$, leaving only the inference pathway to connect the hierarchy to the counts.

**Donor aggregator variants.** The aggregator in Section 3 uses gated attention to pool per-cell scalar states within each annotated type. It concatenates these summaries with donor cell-type proportions and maps the result to $t_d$. *Mean pooling within type* retains type stratification and uses uniform pooling weights. *Proportions only* removes the per-type summaries, making the donor state a function of composition alone. *Head input detached* blocks gradients from the donor objective at the aggregator input while preserving its input values and architecture. This tests whether donor-loss gradients reaching the encoder are needed, separately from the information available to the aggregator.

**Objective and schedule variants.** Setting $\lambda_{\mathrm{donor}} = 0$ removes the donor observation term, leaving the counts and hierarchy to shape the donor state. Setting $\lambda_{\mathrm{class}} = 0$ removes the supervised cell-type term, leaving the categorical ELBO contribution to train $q_\phi(c_n \mid \boldsymbol{z}_n)$. *No warm-up schedules* removes both schedules described in Appendix D. The donor-state KL term is annealed over the first 25 epochs. Over the same period, the effective weight of the donor observation term decreases from 1000 to its target value. The donor slope is initialized at zero, so the donor observation term initially provides no gradient to the inferred donor state. This gradient pathway becomes active as the trainable slope moves away from zero. The schedules control the relative contributions of donor regularization and supervision during early training. Removing them tests sensitivity to these initial training weights.

**The hierarchy supports cell-state organization.** Appendix Table 7 reports the ablation results. The full model achieves a median correlation of $0.45$. Removing both donor-to-cell edges reduces the median to $0.18$ while retaining donor correlation of $0.91$ (Appendix Figure 7). Thus, the joint ablation separates retained donor-level prediction from reduced cell-state association. Detaching the aggregator input gives a median correlation of $0.46$ and a donor correlation of $0.91$. The within-type signal therefore persists without donor-loss gradients reaching the encoder. Removing the auxiliary variable gives a median of $0.34$, while the standard-normal prior gives $0.43$. The latter reduces cell-type accuracy from $0.93$ to $0.87$, indicating that the conditional prior on $\boldsymbol{z}_n$ supports cell-type inference. Removing the type intercepts gives a median correlation of $0.47$ and reduces donor correlation to $0.84$, the lowest among the hierarchy variants. Without $m_k$, the prior constrains all cell types to share the same baseline position.

**The aggregator acts on the donor channel.** Aggregator variants yield median correlations between $0.31$ and $0.48$, close to the full-model result, with larger effects at the donor level. Retaining only the proportions produces the largest donor-level reduction in the study. Donor correlation falls to $0.60$, with a bootstrap interval of $[0.19, 0.80]$, and the standard deviation of $\widehat{t}$ across query donors is $0.45$. Composition alone therefore does not fully determine donor state. We retain stratification because it represents within-type state summaries and donor composition as separate blocks in $\boldsymbol{h}_d$. Mean pooling within type gives a median of $0.31$, the lowest among aggregator variants.

**Objective weights and warm-up schedules.** Removing donor supervision by setting $\lambda_{\text{donor}} = 0$ gives a donor correlation of $0.79$, a standard deviation of $\widehat{t}$ of $0.35$, and median cell correlation of $0.42$. The retained cell association indicates that the hierarchy ablations assess structural organization rather than uniquely identifying transfer of donor-phenotype information. Setting $\lambda_{\text{class}} = 0$ reduces cell-type accuracy from $0.93$ to $0.82$, with little change in donor correlation and median within-type cell correlation. Removing both warm-up schedules does not cause donor-channel collapse on this cohort: donor correlation is $0.97$, and the standard deviation of $\widehat{t}$ is $1.20$. Thus, donor-level performance is maintained without warm-up in this run.

**Setting 2 and the weight of the cell channel.** Appendix Table 9 reports the results for Setting 2, which includes cell-level phenotypes. Removing both hierarchy edges leaves cell-level association largely unchanged: pooled and median correlations are $0.97$ and $0.82$, compared with $0.97$ and $0.81$ for the full model. Here, the cell channel directly anchors $s_n$, providing an additional source of cell-level supervision. The effect of the joint ablation appears at the donor level, where correlation decreases from $0.84$ to $0.73$. Thus, removing both edges primarily reduces cell-level association under donor-only supervision and donor-level association when cell supervision is also available.

The largest change in cell-level association occurs when $\lambda_{\text{cell}}$ increases from $0$ to $10$. Median correlation rises from $0.45$ to $0.51$, and pooled correlation rises from $0.08$ to $0.91$. Donor correlation is $0.97$ versus $0.91$ without cell supervision. A small cell-channel weight therefore anchors the relative positions of types without reducing donor-level performance. Increasing the weight to $100$ and $500$ further raises the median cell correlation to $0.70$ and $0.81$, while donor correlation decreases to $0.80$ and $0.84$, respectively. These results indicate competing effects of donor and cell supervision on $s_n$ as the cell channel becomes stronger, consistent with the difference between Settings 1 and 2 in Table 1. Appendix Figure 8 shows the four weights.

**Supervised prediction from scVI representations.** We extend the mouse ablation study by comparing scPACE with predictors trained on fixed scVI representations. The mouse cohort provides donor-only and donor-plus-cell supervision settings on the same data, allowing us to examine cell-level prediction alongside the component ablations. Here, we focus on Setting 2 and retain the same donor split and maturity-score target.

We fit ridge regression and a multilayer perceptron (MLP) to the existing 10 dimensional scVI representations. These representations remain fixed during predictor training. The MLP has one hidden layer with 64 ReLU units. Both predictors use the same cell-level maturity scores as scPACE. scPACE also uses donor phenotypes and cell-type labels, and learns its cell representation during training.

We train the predictors on 62 donors (625,704 cells) and validate them on six donors (62,694 cells). We select the ridge penalty from $\{0.1, 1, 10, 100\}$ using the lowest validation mean squared error, which selects $0.1$. We train the MLP for 30 epochs using Adam, a learning rate of $10^{-3}$, a batch size of 4,096, and training seed 0. We retain the model from epoch 29, which has the lowest validation error. Query phenotypes are used only for evaluation. Prediction errors are calculated after calibration using reference data, as described in Eq. (52).

scPACE achieves higher correlations and lower prediction error than both predictors (Table 10). Its advantage is present both across all query cells and within cell types.

We estimate uncertainty in differences between methods by sampling the 20 query donors with replacement 1,000 times. Each sample includes all cells from the sampled donors, and both methods are evaluated on the same sample. The trained models and reference calibration remain fixed throughout. We report 95% intervals using the 2.5th and 97.5th percentiles of the resulting differ-

Table 7: **Ablation on the mouse query cohort, Setting 1.** Each row reports one training run (seed 0). Median $|r|$, the primary cell metric, is the median absolute within-type correlation (Eq. (61)). Demeaned $|r|$ is computed across cells after subtracting type-specific means (Eq. (60)). Acc. denotes cell-type accuracy on annotated query cells. $\mathrm{sd}(\widehat{t})$ is the standard deviation of inferred donor states across query donors before calibration. Appendix Figure 7 shows the bootstrap intervals for the ablated variants. A dash indicates that the full-model state-spread statistic is not available for the retained run.

| | | Cell $\lvert r\rvert$ | | | |
|---|---|---|---|---|---|
| Variant | Donor $\lvert r\rvert$ | Median | Demeaned | Acc. | $\mathrm{sd}(\widehat{t})$ |
| Full model | 0.91 | 0.45 | 0.22 | 0.93 | – |
| *Hierarchy* | | | | | |
| No bridge and $\boldsymbol{\beta} = \mathbf{0}$ | 0.91 | 0.18 | 0.14 | 0.95 | 1.48 |
| No type intercepts | 0.84 | 0.47 | 0.35 | 0.94 | 1.16 |
| No auxiliary variable | 0.95 | 0.34 | 0.18 | 0.93 | 1.76 |
| Standard-normal $\boldsymbol{z}_n$ prior | 0.92 | 0.43 | 0.33 | 0.87 | 1.04 |
| *Donor aggregator* | | | | | |
| Mean pooling within type | 0.96 | 0.31 | 0.04 | 0.94 | 1.55 |
| Proportions only | 0.60 | 0.48 | 0.49 | 0.96 | 0.45 |
| Head input detached | 0.91 | 0.46 | 0.39 | 0.95 | 1.41 |
| *Objective and schedules* | | | | | |
| $\lambda_{\mathrm{donor}} = 0$ | 0.79 | 0.42 | 0.34 | 0.93 | 0.35 |
| $\lambda_{\mathrm{class}} = 0$ | 0.91 | 0.44 | 0.37 | 0.82 | 1.36 |
| No warm-up schedules | 0.97 | 0.42 | 0.36 | 0.95 | 1.20 |

Table 8: **Three summaries of cell-state association on query cells.** Pooled uses all 176,614 query cells, as in Table 1; the two within-type measures use 104,327 annotated query cells. It reflects both within-type ordering and the relative positions of types. Demeaned pools the values after subtracting type-specific means (Eq. (60)). Median summarizes the absolute within-type correlations, with one value per type containing at least 200 query cells (Eq. (61)).

| | Query-cell $\lvert r\rvert$ | | |
|---|---|---|---|
| Method | Pooled | Demeaned | Median |
| scPACE, Setting 2 | 0.97 | 0.85 | 0.82 |
| scPACE, Setting 1 | 0.08 | 0.22 | 0.45 |
| PhenoPath | 0.44 | 0.42 | 0.40 |
| psupertime | 0.23 | 0.35 | 0.37 |
| Slingshot | 0.28 | 0.30 | 0.40 |
| MELD | 0.19 | 0.33 | 0.38 |
| Sceptic | 0.18 | 0.31 | 0.36 |
| CNA | 0.21 | 0.30 | 0.31 |

ences. These intervals quantify uncertainty from sampling query donors for the fitted models and the chosen donor split.

All differences below are scPACE minus the comparator, with 95% intervals in parentheses. In Setting 2, scPACE exceeds the MLP in pooled cell correlation by $0.014$ ($[0.008, 0.025]$), correlation after subtracting cell-type means by $0.063$ ($[0.049, 0.104]$), and median within-type correlation by $0.084$ ($[0.061, 0.152]$). Applying the same procedure to donor prediction in Setting 1, scPACE exceeds scVIP in donor correlation by $0.171$ ($[0.052, 0.321]$).

Table 9: **Ablation on the mouse cohort, Setting 2.** The lower block varies the cell-channel weight. $\lambda_{\text{cell}} = 0$ uses the Setting 1 reference run in Table 1; $\lambda_{\text{cell}} = 500$ uses the Setting 2 ablation run.

| | | Cell $\lvert r \rvert$ | | |
|---|---|---|---|---|
| Variant | Donor $\lvert r \rvert$ | Pooled | Median | Demeaned |
| Full model | 0.84 | 0.97 | 0.81 | 0.84 |
| No bridge and $\boldsymbol{\beta} = \mathbf{0}$ | 0.73 | 0.97 | 0.82 | 0.85 |
| *Cell-channel weight* | | | | |
| $\lambda_{\text{cell}} = 0$ (Setting 1) | 0.91 | 0.08 | 0.45 | 0.22 |
| $\lambda_{\text{cell}} = 10$ | 0.97 | 0.91 | 0.51 | 0.50 |
| $\lambda_{\text{cell}} = 100$ | 0.80 | 0.95 | 0.70 | 0.77 |
| $\lambda_{\text{cell}} = 500$ (Setting 2) | 0.84 | 0.97 | 0.81 | 0.84 |

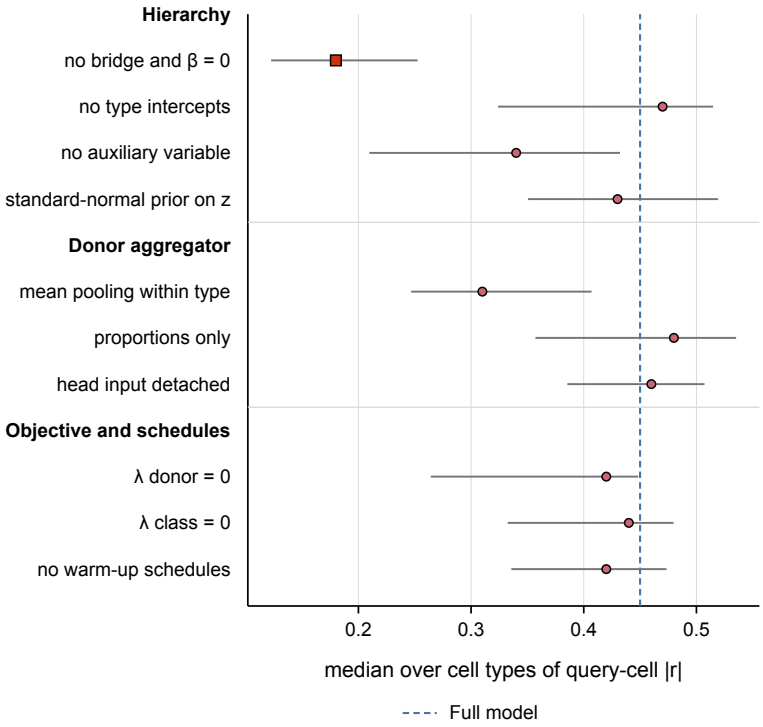



Figure 7: **Median cell-state correlation for Setting 1 ablations on held-out donors.** Points show the single-seed results for the variants retained in Appendix Table 7, and bars show 95% donor-cluster bootstrap intervals. The dashed line marks the full-model median correlation of 0.45.

Table 10: **Maturity-score prediction on the mouse query cohort, Setting 2.** Pooled correlation and nRMSE use all 176,614 query cells. Demeaned $\lvert r \rvert$ is the absolute correlation after subtracting cell-type means from both predictions and targets. Median $\lvert r \rvert$ is the median absolute correlation calculated separately within each eligible cell type. These two within-type measures use 104,327 annotated cells across 12 types, as defined in Eqs. (60)–(61). Boldface marks the best value among the three models.

| Method | Pooled $\lvert r \rvert$ | nRMSE | Demeaned $\lvert r \rvert$ | Median $\lvert r \rvert$ |
|---|---|---|---|---|
| scPACE | **0.972** | **0.233** | **0.846** | **0.818** |
| MLP on scVI | 0.959 | 0.285 | 0.783 | 0.734 |
| Ridge on scVI | 0.875 | 0.484 | 0.530 | 0.556 |

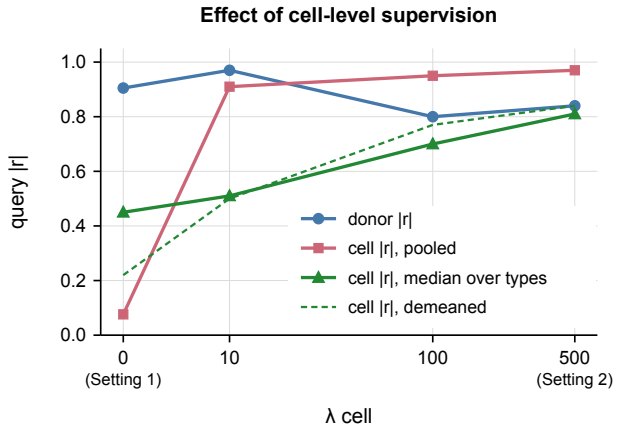



Figure 8: **Effect of the cell-channel weight $\lambda_{\text{cell}}$ on the mouse cohort.** The leftmost and rightmost points correspond to Settings 1 and 2, respectively.