CoState: Interpretable Donor Representation Learning from Single-Cell Transcriptomes

Anonymous authors
Paper under double-blind review

Abstract

Single-cell RNA sequencing offers new opportunities for individual-level disease classification. Disease-associated differences between donors can reflect changes in cell-type composition, within-type expression states, or both. However, existing approaches that average cell features to represent donors do not distinguish between these two sources of variation. We introduce **CoState** (**Co**mposition and **State**), which jointly learns cell and donor representations by combining observed cell-type proportions with learned within-type state distributions into a joint probability distribution. Using Hellinger geometry, we map this distribution to a donor vector, enabling interpretable comparisons that quantify each cell type's contribution to differences between donors. On a benchmark comprising 1,007 donors from 23 Inflammation Atlas studies spanning 19 diseases and healthy controls, **CoState** outperforms the evaluated baselines in cross-study disease retrieval, achieving an approximately 16% relative improvement in retrieval macro-F1 over the strongest baseline in independent-study evaluation. Compared with baseline methods in the training-cell scaling experiment, **CoState** benefits more from increasing the number of training cells, achieving greater performance gains under a fixed training schedule. Code is available at `https://anonymous.4open.science/r/costate-0D40`.

1 Introduction

Single-cell RNA sequencing offers new opportunities to identify disease at cellular resolution, with potential applications in precision diagnostics. Previous studies have used single-cell expression profiles to detect rare malignant cells and identify immune cell states that distinguish patients from healthy controls, demonstrating that cellular populations contain information relevant to disease identification (Ledergor et al., 2018; Reyes et al., 2020). Peripheral blood, which is accessible and amenable to repeated sampling, provides an important setting for exploring these applications. Recent peripheral blood single-cell atlases bring together cohorts from multiple studies spanning infectious diseases, immune-mediated inflammatory diseases, and cancer (Jiménez-Gracia et al., 2026). These studies differ in sampling procedures, experimental protocols, and patient populations, so donor differences can reflect both disease biology and study of origin. When diseases are unevenly represented across studies, models may exploit study-specific signals that do not transfer to new cohorts. A key challenge is therefore to learn donor representations that preserve disease-related differences across heterogeneous studies while limiting the influence of study-specific variation.

Donors can differ in both cell-type abundance and the distribution of states within each type. For example, an increased proportion of monocytes and an enhanced interferon response within monocytes without a change in their proportion represent distinct sources of immune variation (Wen et al., 2020; Kang et al., 2018). Cell-type proportions, expression means, and pooled cell embeddings provide simple and effective summaries for donor comparison (Buettner et al., 2021; Murphy & Skene, 2022). Methods such as PaSCient further learn patient representations through multicellular aggregation for disease classification, patient stratification (Liu et al., 2026). However, our evaluation shows that several existing approaches retain strong study-specific structure and struggle to retrieve same-disease donors across studies, including independent cohorts. Averaging introduces a further limitation: distinct cell-state mixtures can produce similar mean expression profiles, and

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.

signals from small disease-associated cell populations can be diluted. For example, a small population with high expression of an activation marker and a larger population with moderate expression can yield the same average. These limitations call for donor representations that retain disease-relevant cellular heterogeneity and support comparison across studies.

To address these challenges, we introduce **CoState** (**Co**mposition and **State**), which learns disease-informative donor representations with interpretable comparison distances. **Cellular heterogeneity informs donor comparison, while donor-level learning improves cellular representations.** For each donor, **CoState** estimates a cell-state distribution for each cell type from gene expression, then combines these distributions with the observed cell-type proportions. We use Hellinger geometry (Jebara et al., 2004) to convert these joint distributions into donor vectors that retain cell-state heterogeneity beyond mean expression. **Donor distances can be traced to changes in cell-type abundance and within-type state distributions.** To support disease comparison across studies, we use disease-supervised contrastive learning to bring donors with the same disease closer together in this space. Count reconstruction and cell-type supervision guide the learning of cellular features. We also encourage consistent state distributions between subsamples from the same donor to reduce sensitivity to under limited cell sampling. Ablations further show that donor-level training improves cell-type clustering and annotation.

To test whether the learned representations support disease comparison across study boundaries, we evaluate **CoState** on three splits of a multi-study inflammation atlas (Jiménez-Gracia et al., 2026): internal test and validation (unseen donors from studies represented in training), and independent external (donors from studies absent from training). Because shared study-specific signals can make donors appear similar, we test whether a query donor and its nearest match from another study share the same disease or healthy status. **CoState outperforms the evaluated baselines in disease retrieval macro-F1 on all three splits, including independent studies absent from training.** We also illustrate how observed cell-type proportions and expression profiles associated with learned states can be examined to aid interpretation of donor representations. **With donor numbers and training schedules fixed, larger training-cell pools improve both donor disease prediction and cell-type prediction.**

Our main contributions are as follows:

1. **Structured donor representations with decomposable geometry. CoState** uses square-root coordinates for joint type–state distributions, whose squared Hellinger distances decompose into composition and abundance-weighted, type-specific state contributions.
2. **Hierarchical learning objectives. CoState** combines count reconstruction and cell-type supervision at the cell level with disease discrimination and subsample consistency at the donor level. Ablations show that donor-level objectives also improve cell-type clustering and annotation.
3. **Disease discrimination and data scaling.** Across a 23-study benchmark from the Inflammation Atlas, **CoState** outperforms the evaluated baselines in cross-study disease retrieval, especially on previously unseen studies, and achieves the largest gains in disease prediction as the number of training cells increases.

2 Related Work

Cell-level representation learning and integration. scVI and scANVI learn generative cell representations (Lopez et al., 2018; Xu et al., 2021); scPoli and MrVI extend generative modeling to sample-level analysis (De Donno et al., 2023; Boyeau et al., 2025). FlatVI, IRVAE, scPhere, and InfoGlobe explicitly structure representation geometry (Palma et al., 2025; Lee et al., 2022; Ding & Regev, 2021; Wang et al.), while PCA and Harmony provide linear embeddings and batch correction (Korsunsky et al., 2019). These objectives differ from explicitly separating composition and within-type state contributions to donor distances.

Single-cell foundation models. Geneformer, scGPT, and scFoundation learn transferable cell embeddings (Theodoris et al., 2023; Cui et al., 2024; Hao et al., 2024). UCE and Nicheformer extend pretraining across tissues and species and to spatial transcriptomics, respectively (Rosen et al., 2026; Tejada-Lapuerta et al., 2025). However, transferability alone does not ensure that pooling preserves

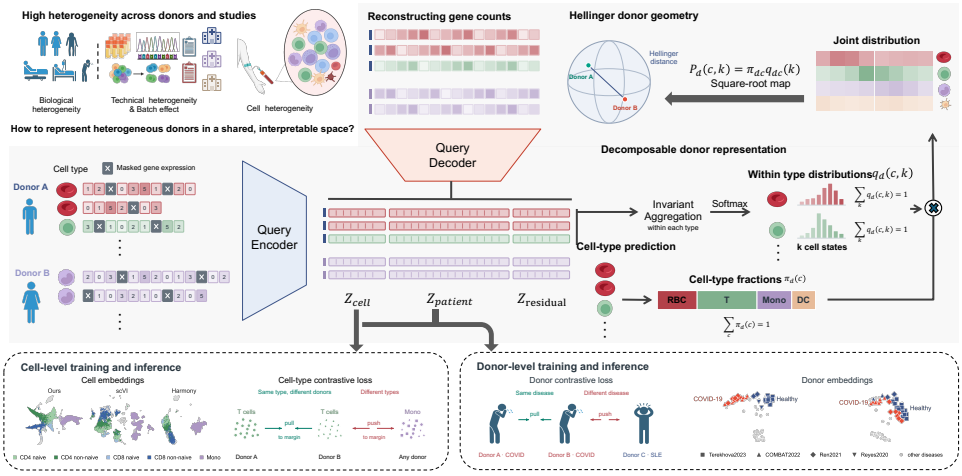

Figure 1: Overview of **CoState**. Joint cell-level and donor-level learning connects cellular expression to donor representations with decomposable distances.

within-type state distributions. **CoState** uses donor-level objectives to improve cellular representations, as supported by clustering and annotation ablations (Section 4.5).

From cell sets to donor representations. DeepSets formalizes invariant aggregation (Zaheer et al., 2017); PaSCient learns attention-based patient representations (Liu et al., 2026), and multicellular factor analysis extracts interpretable factors from type-specific profiles (Ramirez Flores et al., 2023). PhEMD and PILOT compare cellular distributions through optimal transport (Chen et al., 2020; Joodaki et al., 2023); PILOT-GM and PILOT-GM-VAE use sample-specific Gaussian mixtures (Joodaki et al., 2025). By contrast, **CoState** uses square-root coordinates of joint type–state distributions to decompose donor distances exactly into composition and abundance-weighted, type-specific state contributions.

3 Method

Figure 1 outlines **CoState**; Algorithms 1 and 2 in Appendix A detail training and new-donor encoding. We construct joint type–state distributions (Sections 3.1–3.2), map them to donor vectors with Hellinger geometry (Section 3.3), and jointly train cellular and donor representations (Section 3.4).

3.1 Problem setup and donor representation

We encode variable-sized cell collections as fixed-dimensional donor vectors for disease comparison across studies. Donor d provides N_d observed cells, $\mathcal{X}_d = \{(x_{di}, c_{di})\}_{i=1}^{N_d}$, with gene counts $x_{di} \in \mathbb{N}_0^G$, annotated types $c_{di} \in \{1, \dots, C\}$, and an optional training disease label y_d. For type c, S_{dc} is its cell-index set within donor d, n_{dc} its cell count, and π_{dc} its fraction of the donor's cells:

$$S_{dc} = \{i : c_{di} = c\}, \qquad n_{dc} = |S_{dc}|, \qquad \pi_{dc} = \frac{n_{dc}}{N_d}. \tag{1}$$

To retain within-type heterogeneity, we estimate $q_{dc} \in \Delta^{K-1} \subset \mathbb{R}^K$, a probability vector with nonnegative entries summing to one. Its K learned state codes are shared across donors of the same type; $q_{dc}(k)$ averages the probability of code k over donor d's type-c cells (Section 3.2). Combining abundance and conditional states gives the joint distribution

$$P_d(c,k) = \pi_{dc} q_{dc}(k), \qquad \sum_{c=1}^{C} \sum_{k=1}^{K} P_d(c,k) = 1. \tag{2}$$

The fixed-size table preserves abundance and state heterogeneity beyond mean expression.

3.2 NETWORK FOR CONDITIONAL CELL-STATE DISTRIBUTIONS

To compare cellular heterogeneity across donors, we represent cells of the same type in the same learned state space. A shared network maps expression and cell-type identity to probabilities over the same learned state codes. Averaging these probabilities within each donor and type yields q_{dc}.

Cell encoder. The encoder maps gene identities and expression values to tokens, aggregates them with eight learnable cross-attention queries (Appendix C.1), and combines the result with learned expression-profile projections to produce cell, patient, and residual features:

$$F_\theta(x_{di}) = [b_{di}^{\text{cell}};\, b_{di}^{\text{pat}};\, b_{di}^{\text{res}}]. \tag{3}$$

The cell and patient blocks support type supervision and donor learning. An auxiliary residual block adds reconstruction capacity but is excluded from the donor head; all three blocks enter reconstruction (Appendix C.1).

Type-centered feature fusion. The donor head combines RMS-normalized patient features with detached cell features centered on a shared cell-type reference s_c:

$$\widetilde{h}_{di} = \phi_\eta(\widetilde{b}_{di}^{\text{pat}}) + W_{\text{cell}}\big(\widetilde{b}_{di}^{\text{cell}} - s_{c_{di}}\big) \in \mathbb{R}^{D_h}. \tag{4}$$

Here ϕ_η and the bias-free W_{cell} are shared across donors and types.

Shared state network and probability pooling. Let $t_c \in \mathbb{R}^{D_t}$ be a learned type embedding. A shared network g_η transforms the fused feature and its type embedding into K state logits:

$$\alpha_{di} = \text{softmax}\Big(g_\eta([\widetilde{h}_{di}; t_{c_{di}}])\Big), \qquad q_{dc} = \frac{1}{n_{dc}} \sum_{i \in S_{dc}} \alpha_{di} \quad (n_{dc} > 0). \tag{5}$$

Within-type averaging is permutation-invariant (Zaheer et al., 2017); Appendix C.4 details the shared state coordinates.

3.3 HELLINGER DONOR GEOMETRY

To explain why two donors are close or far apart, we need to quantify the contributions of cell-type abundance and within-type states. We form the donor vector $u_d = \text{vec}\sqrt{P_d} \in \mathbb{R}^{CK}$, whose type-$c$ block is $u_{d,c} = \sqrt{\pi_{dc}}\sqrt{q_{dc}}$. Each block encodes abundance in its magnitude and state distribution in its direction.

Square roots are elementwise; blocks follow a fixed type order, with zero blocks for absent types. These unit vectors preserve the joint distribution (Appendix B.1). We compare donors using the squared Hellinger distance between their joint distributions:

$$D(a,b) := H^2(P_a, P_b) = \frac{1}{2}\|u_a - u_b\|_2^2 = 1 - u_a^\top u_b. \tag{6}$$

Here $H^2(p,q) = 1 - \sum_j \sqrt{p_j q_j}$ for probability vectors p and q.

Proposition 1 (Composition-state decomposition)**.** *For any two donors a and b,*

$$D(a,b) = \underbrace{H^2(\pi_a, \pi_b)}_{\textit{composition}} + \underbrace{\sum_{c=1}^{C} \sqrt{\pi_{ac}\pi_{bc}}\, H^2(q_{ac}, q_{bc})}_{\textit{type-specific state}}. \tag{7}$$

Every contribution is nonnegative. A type contributes a conditional-state discrepancy only if it has positive observed mass in both donors.

This decomposition quantifies the absolute and relative contributions of cell-type abundance and within-type state differences to each donor-pair distance, with further resolution by cell type. Appendix B derives the decomposition by expanding the donor inner product $u_a^\top u_b$, and Appendix B.2 gives per-type distance bounds and the corresponding composition-only distance. Figure 9 (Appendix G.4) reports both terms for donor pairs from different studies.

3.4 HIERARCHICAL LEARNING OF DONOR GEOMETRY

Disease labels describe whole donors, whereas the cellular variation underlying donor differences is observed in individual cells. This motivates a hierarchical objective that connects cell-level reconstruction and type supervision with donor-level disease learning, while consistency between same-donor subsamples stabilizes state estimation under finite cell sampling. The ablation in Section 4.5 (Figure 5) shows that jointly removing the donor losses weakens disease prediction, cell-type clustering, and annotation, supporting the benefit of donor-level training for cellular representations.

Disease supervision. We sample M donors per batch, with multiple donors per disease, and encode each donor's full sampled cell bag as u_d^F. For labeled anchor a, let $\mathcal{C}(a)$ contain the other labeled batch donors and $\mathcal{P}(a)$ its same-disease subset. Self-comparisons are excluded, and only anchors with $|\mathcal{P}(a)| > 0$ contribute. On full-bag vectors, supervised contrastive learning (Khosla et al., 2020) uses

$$\ell_a = -\frac{1}{|\mathcal{P}(a)|} \sum_{p \in \mathcal{P}(a)} \log \frac{\exp[-D(a,p)/\tau]}{\sum_{b \in \mathcal{C}(a)} \exp[-D(a,b)/\tau]}, \qquad \tau > 0, \tag{8}$$

Averaging within and then across diseases gives $\mathcal{L}_{\text{dis}}$. Unlabeled donors are excluded. Since $-D(a,b) = (u_a^F)^\top u_b^F - 1$, these logits are equivalent to dot-product logits.

Donor consistency. For each donor, we split sampled cells by cell type into views A and B, sharing the training-pool composition π_d. With $P_d^r(c,k) = \pi_{dc} q_{dc}^r(k)$ for $r \in \{A, B\}$, we align the views through

$$\begin{aligned} \mathcal{L}_{\text{same}} &= \frac{1}{M} \sum_{d=1}^{M} H^2(P_d^A, P_d^B) = \frac{1}{2M} \sum_{d=1}^{M} \|u_d^A - u_d^B\|_2^2 \\ &= \frac{1}{M} \sum_{d=1}^{M} \sum_{c=1}^{C} \pi_{dc} H^2(q_{dc}^A, q_{dc}^B). \end{aligned} \tag{9}$$

Sharing π_d isolates variability in state estimation under cell resampling; Appendix D details the sampling protocol.

Cell decoder and count reconstruction. For reconstruction training, the encoder receives either a randomly gene-masked or full input, denoted by $\widetilde{x}_{di}$. The cell decoder reads the concatenated cell, patient, and residual features. A learned projection maps these features to a vector whose inner products with gene-specific query vectors produce expression logits. A softmax converts these logits into gene proportions ρ_ψ, scaled by the cell's full count total ℓ_{di}:

$$\mu_{di} = \ell_{di} \rho_\psi(F_\theta(\widetilde{x}_{di})). \tag{10}$$

These mean counts parameterize a negative-binomial likelihood with learned gene-specific dispersions. The reconstruction loss $\mathcal{L}_{\text{NB}}$ (Lopez et al., 2018) averages the negative log-likelihood over masked genes for masked inputs and over all genes for full inputs. Appendix C.2 details the decoder, masking strategy, and loss.

Cell-type supervision. We train the cell-feature block with a margin-based contrastive loss $\mathcal{L}_{\text{cell}}$. Positive pairs contain cells of the same type from different donors, whereas negative pairs contain cells of different types. Using cosine distance, we penalize positive pairs whose distance exceeds an intra-type margin and negative pairs whose distance falls below an inter-type margin. The loss averages these hinge penalties separately over positive and negative pairs (Appendix C.2).

Joint objective and gradient flow. The total objective combines negative-binomial count reconstruction $\mathcal{L}_{\text{NB}}$ with cell-type supervision and the donor objectives:

$$\mathcal{L} = \mathcal{L}_{\text{NB}} + \lambda_{\text{cell}} \mathcal{L}_{\text{cell}} + \lambda_{\text{dis}} \mathcal{L}_{\text{dis}} + \lambda_{\text{same}} \mathcal{L}_{\text{same}}. \tag{11}$$

Both donor losses update conditional states, feature fusion, and the patient path; counted π is fixed. Cell losses are specified in Appendix C.2. Neither disease supervision nor consistency guarantees removal of study effects.

4 EXPERIMENTS

Our experiments address five questions. **(Q1)** First, do **CoState** donor vectors support accurate disease prediction (Table 1)? **(Q2)** Second, can they retrieve matching diseases across studies, including studies absent from training (Figures 2a–c and 3)? **(Q3)** Third, do the cell representations preserve cell-type identity across studies (Table 2; Figure 2d)? **(Q4)** Fourth, how does increasing the number of training cells affect performance (Figure 4)? **(Q5)** Finally, what do donor supervision and the composition and state readouts contribute (Figure 5)?

Experimental setup. We benchmark 1,007 donors from 23 studies in the Inflammation Atlas (Jiménez-Gracia et al., 2026), covering 19 diseases and healthy controls. Training uses 735 donors from 16 studies. Test and validation contain 82 and 144 unseen donors from training studies; external evaluation uses 46 donors from seven studies absent from training. **CoState** uses 2,000 highly variable genes and $K = 16$ unless noted. Donor readouts use 250-cell views. Appendix E details preprocessing, splits, supervision, and evaluation.

Baselines. We compare atlas-trained cell models, donor-supervised PaSCient, frozen foundation models, integration methods, and simple PCA, composition, and pseudobulk summaries. Appendix E.4 describes each implementation, supervision, and donor readout.

4.1 DISEASE PREDICTION FROM DONOR REPRESENTATIONS

We fit readouts on reference donors and predict disease in disjoint query donors within each cohort, averaging five repeats. **CoState** achieves the highest disease macro-F1 across nearest-neighbour, prototype, and linear readouts in all three cohorts (Table 1). This consistency suggests that disease-related structure is reflected in both local donor neighbourhoods and broader class organization. Shared study-specific signals can also improve within-cohort prediction. We therefore exclude same-study references in the next analysis to test disease matching across studies.

Table 1: Donor disease macro-F1 (250 cells/view). kNN: 1-NN; Proto.: nearest class prototype; Linear: logistic probe. Bold/underline: best/second best.[1]

	Test donors			Validation cohort			External studies		
Method	kNN ↑	Proto. ↑	Linear ↑	kNN ↑	Proto. ↑	Linear ↑	kNN ↑	Proto. ↑	Linear ↑
Trained on the atlas									
scVI	0.677	0.653	0.616	0.551	0.497	0.454	0.661	0.644	0.586
scANVI	0.402	0.444	0.421	0.385	0.369	0.307	0.604	0.505	0.442
scANVI (disease)	0.519	0.539	0.525	0.598	0.521	0.480	0.740	0.629	0.514
FlatVI	0.505	0.515	0.427	0.481	0.443	0.341	0.724	0.677	0.504
IRVAE	0.478	0.452	0.428	0.413	0.380	0.324	0.603	0.529	0.458
scPhere	0.629	0.598	0.587	0.510	0.460	0.404	0.741	0.656	0.607
PILOT-GM	0.564	0.562	0.523	0.553	0.527	0.495	0.720	0.733	0.582
PaSCient	<u>0.825</u>	<u>0.855</u>	0.662	<u>0.706</u>	<u>0.737</u>	<u>0.563</u>	0.762	0.739	0.574
Linear floor and batch correction									
PCA	0.668	0.669	0.666	0.633	0.533	0.534	0.793	0.771	0.709
Harmony†	0.542	0.537	0.420	0.366	0.308	0.248	0.331	0.304	0.222
InfoGlobe†	0.532	0.452	0.543	0.327	0.309	0.197	0.589	0.491	0.553
Transductive									
MrVI†	0.316	0.257	0.181	0.240	0.184	0.139	0.488	0.399	0.301
Foundation models (frozen)									
scGPT	0.683	0.661	0.581	0.703	0.628	0.447	0.701	0.755	0.757
Geneformer	0.823	0.769	<u>0.687</u>	0.656	0.570	0.491	0.759	0.729	0.687
Nicheformer	0.678	0.663	0.665	0.638	0.534	0.457	0.713	0.686	0.551
UCE	0.596	0.610	0.585	0.613	0.557	0.501	<u>0.834</u>	0.802	<u>0.806</u>
Donor-level baselines									
Composition only	0.222	0.263	0.196	0.258	0.282	0.228	0.538	0.619	0.563
Pseudobulk + PCA	0.715	0.734	0.658	0.637	0.517	0.443	<u>0.834</u>	<u>0.874</u>	0.678
CoState (Ours)	**0.891**	**0.900**	**0.849**	**0.860**	**0.888**	**0.867**	**0.840**	**0.887**	**0.823**

[1]Five disjoint donor/view repeats; three training seeds where available. **CoState** uses Hellinger distance, other donor vectors cosine distance. scANVI (disease) and PaSCient use disease labels. †Transductive: evaluation cells or batch labels are used.

4.2 DISEASE RETRIEVAL ACROSS STUDIES

We evaluate cross-study retrieval by assigning each query the disease or healthy label of its nearest reference donor from another study. Scored categories occur in at least two studies, with at least two donors per study: healthy/COVID-19 on test and validation, and healthy/SLE externally. Evaluation uses three cell resamples (Appendix E.3).

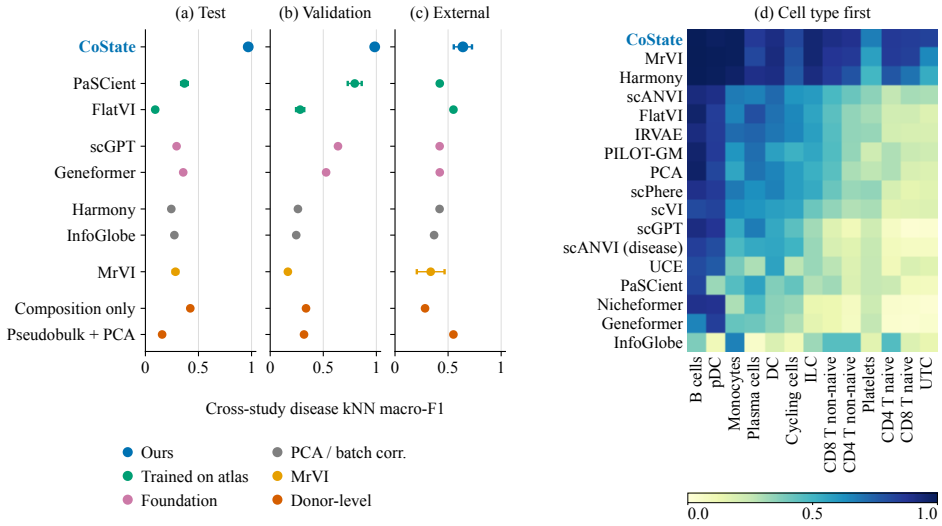

Figure 2: Cross-study evaluation. (a)–(c): disease retrieval macro-F1 with same-study references excluded; bars: seed s.d. Top two methods per family by mean F1 across splits. (d): fraction of cells whose nearest same-type cell in another study is closer than their nearest other-type cell in their own study. Full results: Figures 6 and 7.

CoState achieves the highest cross-study retrieval F1 in all three cohorts (Figure 2a–c). With same-study matches excluded, this result indicates that disease-related donor similarity extends across study boundaries. The advantage persists on external studies absent from training, supporting disease matching in new data sources.

For cross-study disease comparison, disease status (colour) is the biological label of interest, while study identity (marker shape) is a nuisance label (Figure 3). scVI organizes COVID-19 donors largely by shape, separating COMBAT2022 (Ahern et al., 2022) and Ren2021 (Ren et al., 2021) shared colour. In **CoState**, these donors group together across shapes and remain clearly separated by colour from healthy donors. This disease-centred organization illustrates how **CoState** brings donors from different studies together through disease-related similarities, providing a visual explanation for its cross-study retrieval advantage.

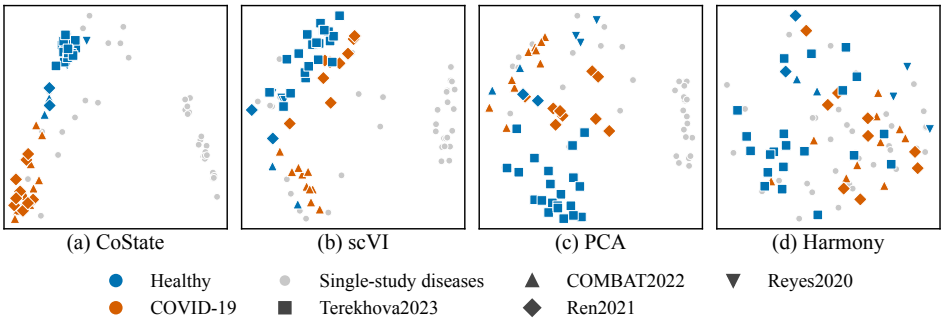

Figure 3: Donor distances visualized by classical MDS (82 test donors, 13 studies; **CoState**: $K = 8$, first seed). Colour: disease; shape: study; grey: diseases outside the cross-study comparison.

4.3 CELL-TYPE PRESERVATION ACROSS STUDIES

Cell-type evaluation uses manual annotations for 82 held-out donors from the atlas main cohort; all 13 studies appear in training. **CoState** leads all seven cell-level metrics (Table 2). Its cross-study label-transfer macro-F1 of 0.901 supports accurate cell annotation using references from other studies, with alignment spanning multiple cell types (Figure 2d).

Table 2: Cell-type readout on test donors. L/k: Leiden/k-means; kNN: 15-NN macro-F1; Transfer: cross-study F1. Type first: Figure 2d. Bold/underline: best/second best. Details: Appendix E.3.

	Cell-type structure					Same type across studies	
Method	NMI_L ↑	ARI_L ↑	NMI_k ↑	ARI_k ↑	kNN F1 ↑	Type first ↑	Transfer F1 ↑
Trained on the atlas							
scVI	0.577	0.317	0.507	0.285	0.817	0.451	0.819
scANVI*	0.623	0.368	0.567	0.315	<u>0.832</u>	0.553	0.823
scANVI (disease)	0.568	0.267	0.565	0.313	0.788	0.375	0.769
FlatVI	0.609	0.449	0.578	0.337	0.793	0.529	0.793
IRVAE	0.610	0.432	0.576	0.311	0.794	0.529	0.781
scPhere	0.592	0.351	0.542	0.300	0.782	0.495	0.772
PILOT-GM	0.657	0.475	0.569	0.354	0.766	0.504	0.789
PaSCient	0.570	0.378	0.427	0.218	0.683	0.340	0.650
Linear floor and batch correction							
PCA	0.657	0.464	0.548	0.314	0.764	0.498	0.789
Harmony†	0.689	0.558	0.628	0.417	0.754	0.812	0.792
InfoGlobe†	0.105	0.111	0.080	0.073	0.110	0.280	0.077
Transductive							
MrVI†	<u>0.762</u>	<u>0.699</u>	<u>0.690</u>	<u>0.513</u>	0.816	<u>0.852</u>	<u>0.849</u>
Foundation models (frozen)							
scGPT	0.550	0.287	0.511	0.268	0.731	0.378	0.706
Geneformer	0.476	0.210	0.353	0.187	0.763	0.282	0.753
Nicheformer	0.504	0.220	0.495	0.260	0.760	0.299	0.751
UCE	0.545	0.234	0.508	0.304	0.816	0.351	0.823
CoState (Ours)*	**0.796**	**0.762**	**0.703**	**0.520**	**0.846**	**0.883**	**0.901**

4.4 EFFECT OF TRAINING-CELL SCALE

With donors and training schedules fixed, we enlarge each donor's training pool by uniform sampling without replacement; per-step cell budgets remain unchanged (Appendix E.4). **CoState**'s disease prediction improves up to the largest pool, whereas most baselines plateau or fluctuate (Figure 4a–c; Table 7). These gains, accompanied by better cell-type prediction (Figure 4d), suggest that **CoState** better uses additional cellular variation within the same donors.

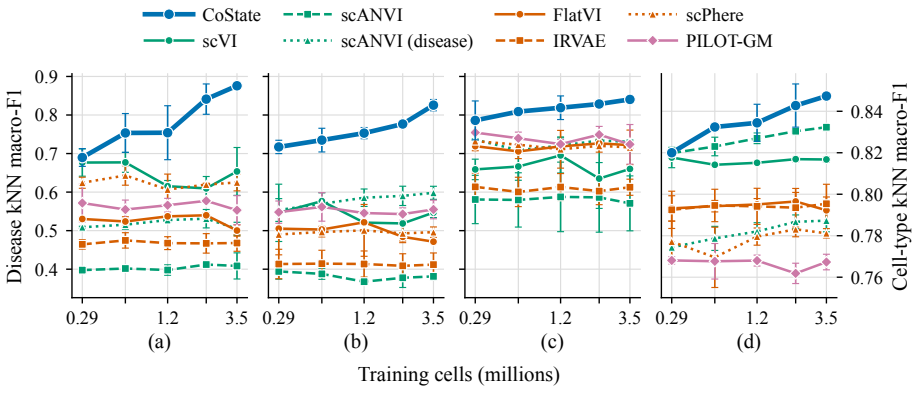

Figure 4: Training-cell scaling. (a)–(c): disease prediction on test/validation/external donors; (d): cell-type prediction. Means ± s.d. over two seeds.

4.5 Ablation study

Donor losses in hierarchical learning. Removing the disease-contrastive and same-donor losses reduces donor disease prediction as well as cell-type clustering and annotation (Figure 5). Donor supervision therefore also benefits the shared cell representation, supporting hierarchical learning at both cell and donor levels.

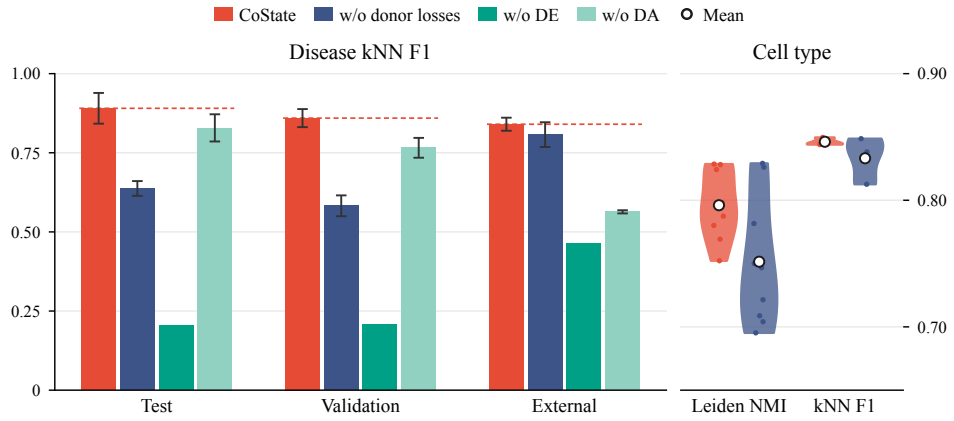

Figure 5: **CoState** ablations. Left: donor disease prediction; right: cell-type clustering and annotation. DA: cell-type abundance; DE: within-type expression state. Error bars: s.d. over three seeds; dashed lines: **CoState**; white circles: means.

Block roles. The cell block emphasizes cell-type structure, the patient block captures donor-level disease information, and within-cell-type shuffling increases reconstruction NLL most for the residual block (14.6%, versus 1.3% and 0.6% for the cell and patient blocks; Appendix Figure 10).

Complementary roles of abundance and state. DA (Differential Abundance) and DE (Differential Expression) in Figure 5 denote cell-type abundance and within-type expression-state information, respectively. With the encoder fixed, removing state information sharply reduces disease prediction across cohorts. Distance decomposition likewise attributes most of the additional distance between diseases to within-type states (Appendix Figure 9). Removing abundance also lowers performance, most strongly on unseen studies. Together, these results show that state heterogeneity supports disease separation, while abundance provides complementary information for transfer across studies.

Exploratory biological characterization. Observed plasma-cell and cycling-cell fractions are higher, and plasmacytoid dendritic-cell fractions lower, in the analyzed COVID-19 donors than in healthy donors. Within-donor expression comparisons suggest interferon-associated and activation-associated annotations for selected monocyte and non-naive CD4 T-cell states, respectively (Appendix G.4).

5 Discussion and conclusion

CoState represents donors through joint distributions of cell types and expression states, retaining heterogeneity beyond mean expression. Hierarchical learning brings same-disease donors together while preserving cell-type structure. Hellinger geometry quantifies abundance and state contributions to each donor-pair distance, with resolution by cell type.

The retrieval advantage on studies absent from training supports disease matching across data sources. Ablations show that donor supervision also improves cell annotation, linking disease-level learning to cellular representations. With donor numbers and training budgets fixed, gains from larger cell pools further support sampling cellular variation more deeply within existing cohorts.

Across the evaluated cohorts, **CoState** combines the strongest cross-study disease retrieval with cell-type preservation and interpretable donor comparisons based on cellular abundance and expression state. Future work will test whether learned states track disease progression and treatment response.

AI USE STATEMENT

We used generative AI tools to help find and identify literature, develop and carry out research ideas, and refine the language when preparing this work. We did not use AI tools to create synthetic datasets, to prove mathematical statements, or to produce the authors' manuscript. All AI-assisted content was carefully reviewed and verified by the authors, who take full responsibility for the final content, including all text, claims, and experimental results presented in this paper.

ETHICS STATEMENT

This study analyzes publicly available single-cell RNA-sequencing datasets. No new participants were recruited and no new biological samples were collected for this work. **CoState** is developed and evaluated as a research method, rather than a clinically validated diagnostic tool. Differences in cohort representation and study-specific data collection may limit its generalizability, and further validation is needed before any clinical application.

REPRODUCIBILITY STATEMENT

To support reproducibility during peer review, the source code for **CoState** is available in an anonymized repository at `https://anonymous.4open.science/r/costate-0D40`.

REFERENCES

David J Ahern, Zhichao Ai, Mark Ainsworth, Chris Allan, Alice Allcock, Brian Angus, M Azim Ansari, Carolina V Arancibia-Cárcamo, Dominik Aschenbrenner, Moustafa Attar, et al. A blood atlas of covid-19 defines hallmarks of disease severity and specificity. *Cell*, 185(5):916–938, 2022.

John Aitchison. The statistical analysis of compositional data. *Journal of the Royal Statistical Society: Series B (Methodological)*, 44(2):139–160, 1982.

Prabhu S Arunachalam, Florian Wimmers, Chris Ka Pun Mok, Ranawaka APM Perera, Madeleine Scott, Thomas Hagan, Natalia Sigal, Yupeng Feng, Laurel Bristow, Owen Tak-Yin Tsang, et al. Systems biological assessment of immunity to mild versus severe covid-19 infection in humans. *Science*, 369(6508):1210–1220, 2020.

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(11):2264–2274, 2025.

Maren Buettner, Johannes Ostner, Christian L Mueller, Fabian J Theis, and Benjamin Schubert. sccoda is a bayesian model for compositional single-cell data analysis. *Nature communications*, 12(1):6876, 2021.

William S Chen, Nevena Zivanovic, David Van Dijk, Guy Wolf, Bernd Bodenmiller, and Smita Krishnaswamy. Uncovering axes of variation among single-cell cancer specimens. *Nature methods*, 17(3):302–310, 2020.

Anthony R Cillo, Cornelius HL Kürten, Tracy Tabib, Zengbiao Qi, Sayali Onkar, Ting Wang, Angen Liu, Umamaheswar Duvvuri, Seungwon Kim, Ryan J Soose, et al. Immune landscape of viral-and carcinogen-driven head and neck cancer. *Immunity*, 52(1):183–199, 2020.

Haotian Cui, Chloe Wang, Hassaan Maan, Kuan Pang, Fengning Luo, Nan Duan, and Bo Wang. scgpt: toward building a foundation model for single-cell multi-omics using generative ai. *Nature methods*, 21(8):1470–1480, 2024.

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(11):1683–1692, 2023.

Jiarui Ding and Aviv Regev. Deep generative model embedding of single-cell rna-seq profiles on hyperspheres and hyperbolic spaces. *Nature communications*, 12(1):2554, 2021.

Minsheng Hao, Jing Gong, Xin Zeng, Chiming Liu, Yucheng Guo, Xingyi Cheng, Taifeng Wang, Jianzhu Ma, Xuegong Zhang, and Le Song. Large-scale foundation model on single-cell transcriptomics. *Nature methods*, 21(8):1481–1491, 2024.

Tony Jebara, Risi Kondor, and Andrew Howard. Probability product kernels. *Journal of Machine Learning Research*, 5(Jul):819–844, 2004.

Yale Jiang, Brian R Rosborough, Jie Chen, Sudipta Das, Georgios D Kitsios, Bryan J McVerry, Rama K Mallampalli, Janet S Lee, Anuradha Ray, Wei Chen, et al. Single cell rna sequencing identifies an early monocyte gene signature in acute respiratory distress syndrome. *JCI insight*, 5 (13):e135678, 2020.

Laura Jiménez-Gracia, Davide Maspero, Sergio Aguilar-Fernández, Francesco Craighero, Maria Boulougouri, Max Ruiz, Domenica Marchese, Ginevra Caratù, Jose Liñares-Blanco, Miren Berasategi, et al. Interpretable inflammation landscape of circulating immune cells. *Nature medicine*, 32(2):633–644, 2026.

Ian T Jolliffe and Jorge Cadima. Principal component analysis: a review and recent developments. *Philosophical transactions. Series A, Mathematical, physical, and engineering sciences*, 374(2065):20150202, 2016.

Mehdi Joodaki, Mina Shaigan, Victor Parra, Roman D Bülow, Christoph Kuppe, David L Hölscher, Mingbo Cheng, James S Nagai, Michaël Goedertier, Nassim Bouteldja, et al. Detection of patient-level distances from single cell genomics and pathomics data with optimal transport (pilot). *Molecular systems biology*, 20(2):57, 2023.

Mehdi Joodaki, Mina Shaigan, Samaneh Samiei, James Nagai, Tiago Maié, Christoph Kuppe, and Ivan G Costa. Pilot-gm-vae: patient-level analysis of single-cell disease atlas with optimal transport of gaussian mixture variational autoencoders. *Briefings in Bioinformatics*, 26(5):bbaf547, 2025.

Hyun Min Kang, Meena Subramaniam, Sasha Targ, Michelle Nguyen, Lenka Maliskova, Elizabeth McCarthy, Eunice Wan, Simon Wong, Lauren Byrnes, Cristina M Lanata, et al. Multiplexed droplet single-cell rna-sequencing using natural genetic variation. *Nature biotechnology*, 36(1): 89–94, 2018.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673, 2020.

Ilya Korsunsky, Nghia Millard, Jean Fan, Kamil Slowikowski, Fan Zhang, Kevin Wei, Yuriy Baglaenko, Michael Brenner, Po-ru Loh, and Soumya Raychaudhuri. Fast, sensitive and accurate integration of single-cell data with harmony. *Nature methods*, 16(12):1289–1296, 2019.

Guy Ledergor, Assaf Weiner, Mor Zada, Shuang-Yin Wang, Yael C Cohen, Moshe E Gatt, Nimrod Snir, Hila Magen, Maya Koren-Michowitz, Katrin Herzog-Tzarfati, et al. Single cell dissection of plasma cell heterogeneity in symptomatic and asymptomatic myeloma. *Nature medicine*, 24 (12):1867–1876, 2018.

Yonghyeon Lee, Sangwoong Yoon, Minjun Son, and Frank C Park. Regularized autoencoders for isometric representation learning. In *International Conference on Learning Representations*, 2022.

Tianyu Liu, Edward De Brouwer, Archit Verma, Alsu Missarova, Tony Kuo, Nathaniel Diamant, Hanchen Wang, Minsheng Hao, Tommaso Biancalani, Hector Corrada Bravo, et al. Learning multi-cellular representations of single-cell transcriptomics data enables characterization of patient-level disease states. *Cell Systems*, 17(5), 2026.

Yang Liu, Shuai He, Xi-Liang Wang, Wan Peng, Qiu-Yan Chen, Dong-Mei Chi, Jie-Rong Chen, Bo-Wei Han, Guo-Wang Lin, Yi-Qi Li, et al. Tumour heterogeneity and intercellular networks of nasopharyngeal carcinoma at single cell resolution. *Nature communications*, 12(1):741, 2021.

Romain Lopez, Jeffrey Regier, Michael B Cole, Michael I Jordan, and Nir Yosef. Deep generative modeling for single-cell transcriptomics. *Nature methods*, 15(12):1053–1058, 2018.

Jerome C Martin, Christie Chang, Gilles Boschetti, Ryan Ungaro, Mamta Giri, John A Grout, Kyle Gettler, Ling-shiang Chuang, Shikha Nayar, Alexander J Greenstein, et al. Single-cell analysis of crohn's disease lesions identifies a pathogenic cellular module associated with resistance to anti-tnf therapy. *Cell*, 178(6):1493–1508, 2019.

Divij Mathew, Josephine R Giles, Amy E Baxter, Derek A Oldridge, Allison R Greenplate, Jennifer E Wu, Cécile Alanio, Leticia Kuri-Cervantes, M Betina Pampena, Kurt D'Andrea, et al. Deep immune profiling of covid-19 patients reveals distinct immunotypes with therapeutic implications. *Science*, 369(6508):eabc8511, 2020.

Pragnesh Mistry, Shuichiro Nakabo, Liam O'Neil, Rishi R Goel, Kan Jiang, Carmelo Carmona-Rivera, Sarthak Gupta, Diana W Chan, Philip M Carlucci, Xinghao Wang, et al. Transcriptomic, epigenetic, and functional analyses implicate neutrophil diversity in the pathogenesis of systemic lupus erythematosus. *Proceedings of the National Academy of Sciences*, 116(50):25222–25228, 2019.

Alan E Murphy and Nathan G Skene. A balanced measure shows superior performance of pseudobulk methods in single-cell rna-sequencing analysis. *Nature communications*, 13(1):7851, 2022.

Alessandro Palma, Sergei Rybakov, Leon Hetzel, Stephan Günnemann, and Fabian J Theis. Enforcing latent euclidean geometry in single-cell vaes for manifold interpolation. *arXiv preprint arXiv:2507.11789*, 2025.

Mukta G Palshikar, Rohith Palli, Alicia Tyrell, Sanjay Maggirwar, Giovanni Schifitto, Meera V Singh, and Juilee Thakar. Executable models of immune signaling pathways in hiv-associated atherosclerosis. *NPJ systems biology and applications*, 8(1):35, 2022.

Richard K Perez, M Grace Gordon, Meena Subramaniam, Min Cheol Kim, George C Hartoularos, Sasha Targ, Yang Sun, Anton Ogorodnikov, Raymund Bueno, Andrew Lu, et al. Single-cell rna-seq reveals cell type–specific molecular and genetic associations to lupus. *Science*, 376(6589): eabf1970, 2022.

P Ramachandran, R Dobie, JR Wilson-Kanamori, EF Dora, BEP Henderson, NT Luu, JR Portman, KP Matchett, M Brice, JA Marwick, et al. Resolving the fibrotic niche of human liver cirrhosis at single-cell level. *Nature*, 575(7783):512–518, 2019.

Ricardo Omar Ramirez Flores, Jan David Lanzer, Daniel Dimitrov, Britta Velten, and Julio Saez-Rodriguez. Multicellular factor analysis of single-cell data for a tissue-centric understanding of disease. *Elife*, 12:e93161, 2023.

Xianwen Ren, Wen Wen, Xiaoying Fan, Wenhong Hou, Bin Su, Pengfei Cai, Jiesheng Li, Yang Liu, Fei Tang, Fan Zhang, et al. Covid-19 immune features revealed by a large-scale single-cell transcriptome atlas. *Cell*, 184(7):1895–1913, 2021.

Miguel Reyes, Michael R Filbin, Roby P Bhattacharyya, Kianna Billman, Thomas Eisenhaure, Deborah T Hung, Bruce D Levy, Rebecca M Baron, Paul C Blainey, Marcia B Goldberg, et al. An immune-cell signature of bacterial sepsis. *Nature medicine*, 26(3):333–340, 2020.

Yanay Rosen, Yusuf Roohani, Ayush Agrawal, Leon Samotorčan, Tabula Sapiens Consortium, Stephen R Quake, and Jure Leskovec. Universal cell embedding provides a foundation model for cell biology. *Nature*, pp. 1–9, 2026.

Adam K Savage, Miriam V Gutschow, Tony Chiang, Kathy Henderson, Richard Green, Monica Chaudhari, Elliott Swanson, Alexander T Heubeck, Nina Kondza, Kelli C Burley, et al. Multimodal analysis for human ex vivo studies shows extensive molecular changes from delays in blood processing. *Iscience*, 24(5), 2021.

David Schafflick, Chenling A Xu, Maike Hartlehnert, Michael Cole, Andreas Schulte-Mecklenbeck, Tobias Lautwein, Jolien Wolbert, Michael Heming, Sven G Meuth, Tanja Kuhlmann, et al. Integrated single cell analysis of blood and cerebrospinal fluid leukocytes in multiple sclerosis. *Nature communications*, 11(1):247, 2020.

Jonas Schulte-Schrepping, Nico Reusch, Daniela Paclik, Kevin Baßler, Stephan Schlickeiser, Bowen Zhang, Benjamin Krämer, Tobias Krammer, Sophia Brumhard, Lorenzo Bonaguro, et al. Severe covid-19 is marked by a dysregulated myeloid cell compartment. *cell*, 182(6):1419–1440, 2020.

Alejandro Tejada-Lapuerta, Anna C Schaar, Robert Gutgesell, Giovanni Palla, Lennard Halle, Mariia Minaeva, Larsen Vornholz, Leander Dony, Francesca Drummer, Till Richter, et al. Nicheformer: a foundation model for single-cell and spatial omics. *Nature methods*, 22(12): 2525–2538, 2025.

Marina Terekhova, Amanda Swain, Pavla Bohacova, Ekaterina Aladyeva, Laura Arthur, Anwesha Laha, Denis A Mogilenko, Samantha Burdess, Vladimir Sukhov, Denis Kleverov, et al. Single-cell atlas of healthy human blood unveils age-related loss of nkg2c+ gzmb- cd8+ memory t cells and accumulation of type 2 memory t cells. *Immunity*, 56(12):2836–2854, 2023.

Christina V Theodoris, Ling Xiao, Anant Chopra, Mark D Chaffin, Zeina R Al Sayed, Matthew C Hill, Helene Mantineo, Elizabeth M Brydon, Zexian Zeng, X Shirley Liu, et al. Transfer learning enables predictions in network biology. *Nature*, 618(7965):616–624, 2023.

Cheng Wang, Jinpu Cai, Chongxiao Mao, Yuxuan Wang, Xinzhu Jiang, Yunhao Qiao, Luqi Yang, Luting Zhou, Qiuyu Lian, and Hongyi Xin. Infoglobe: Local-and-global information-preserving statistical manifold learning for single-cell transcriptomics. In *Forty-third International Conference on Machine Learning*.

Shaobo Wang, Qiong Zhang, Hui Hui, Kriti Agrawal, Maile Ann Young Karris, and Tariq M Rana. An atlas of immune cell exhaustion in hiv-infected individuals revealed by single-cell transcriptomics. *Emerging Microbes & Infections*, 9(1):2333–2347, 2020.

Wen Wen, Wenru Su, Hao Tang, Wenqing Le, Xiaopeng Zhang, Yingfeng Zheng, Xiuxing Liu, Lihui Xie, Jianmin Li, Jinguo Ye, et al. Immune cell profiling of covid-19 patients in the recovery stage by single-cell sequencing. *Cell discovery*, 6(1):31, 2020.

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):MSB20209620, 2021.

Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Russ R Salakhutdinov, and Alexander Smola. Deep sets. *Advances in neural information processing systems*, 30, 2017.

Chao Zhang, Jiesheng Li, Yongqian Cheng, Fanping Meng, Jin-Wen Song, Xing Fan, Hongtao Fan, Jing Li, Yu-Long Fu, Ming-Ju Zhou, et al. Single-cell rna sequencing reveals intrahepatic and peripheral immune characteristics related to disease phases in hbv-infected patients. *Gut*, 72(1): 153–167, 2023.

A TRAINING AND DONOR ENCODING ALGORITHMS

Algorithm 1 CoState training

Require: Training pools $\{\mathcal{X}_d, y_d\}$; K states, M donors per batch, T steps.
Ensure: Trained parameters Θ and reference buffers.

Initialize Θ and reference buffers.
$\pi_{dc} \leftarrow n_{dc}^{\text{train}} / N_d^{\text{train}}$ ▷ eligible training pool
for $t = 1, \ldots, T$ **do**
  $\mathcal{B} \leftarrow \text{DiseaseBalancedSample}(M)$ ▷ equal donors per disease
  $F_d \leftarrow \text{TypeCoveringSample}(\mathcal{X}_d), d \in \mathcal{B}$ ▷ type coverage
  $b_{di} \leftarrow F_\theta(x_{di}), d \in \mathcal{B},\ i \in F_d$ ▷ full gene inputs
  $\sigma_{\text{cell}}, \sigma_{\text{pat}}, \{\mu_c\} \leftarrow \text{UpdateReferences}(\{b_{di}\})$ ▷ Appendix C.3
  $\widetilde{h}_{di} \leftarrow \phi_\eta(b_{di}^{\text{pat}}/\sigma_{\text{pat}}) + W_{\text{cell}}\big(\text{sg}(b_{di}^{\text{cell}})/\sigma_{\text{cell}} - \mu_{c_{di}} - \Delta_{c_{di}}\big)$
  $\alpha_{di} \leftarrow \text{softmax}(g_\eta([\widetilde{h}_{di}; t_{c_{di}}]))$
  $(A_d, B_d) \leftarrow \text{TypeStratifiedSplit}(F_d), d \in \mathcal{B}$ ▷ split; reuse singletons
  for $d \in \mathcal{B}, r \in \{F, A, B\}$ **do**
    $S_{dc}^r \leftarrow \{i \in r_d : c_{di} = c\}$; $q_{dc}^r \leftarrow |S_{dc}^r|^{-1} \sum_{i \in S_{dc}^r} \alpha_{di} \quad (|S_{dc}^r| > 0)$
    $P_d^r(c,k) \leftarrow \pi_{dc} q_{dc}^r(k)$; $u_d^r \leftarrow \text{vec}\sqrt{P_d^r}$ ▷ zero absent-type rows
  end for
  $(\widetilde{x}_{di}, T_{di}) \leftarrow \text{MaskOrFull}(x_{di})$ ▷ masking protocol
  $z_{di} \leftarrow F_\theta(\widetilde{x}_{di})$
  $\mathcal{L}_{\text{NB}}, \mathcal{L}_{\text{cell}}, \mathcal{L}_{\text{dis}} \leftarrow$ Equations (25), (26), (30)
  $\mathcal{L}_{\text{same}} \leftarrow \frac{1}{2M} \sum_d \|u_d^A - u_d^B\|_2^2$
  $\mathcal{L} \leftarrow \mathcal{L}_{\text{NB}} + \lambda_{\text{cell}}\mathcal{L}_{\text{cell}} + \lambda_{\text{dis}}\mathcal{L}_{\text{dis}} + \lambda_{\text{same}}\mathcal{L}_{\text{same}}$
  Update Θ using $\nabla_\Theta \mathcal{L}$.
end for

Algorithm 2 CoState new-donor encoding

Require: Nonempty, vocabulary-filtered $\mathcal{X}_d$; trained Θ and reference buffers.
Ensure: Donor vector $u_d \in \mathbb{R}^{CK}$.

Freeze Θ and buffers; disable dropout.
$\{\alpha_{di}\} \leftarrow \text{StateHead}(F_\theta(\mathcal{X}_d))$ ▷ Equations (4)–(5)
$P_d(c,k) \leftarrow N_d^{-1} \sum_{i:c_{di}=c} \alpha_{di}(k)$ ▷ Equation (20)
return $u_d = \text{vec}\sqrt{P_d}$

B PROPERTIES OF THE HELLINGER REPRESENTATION

B.1 NORMALIZATION, RECOVERY, AND DISTANCE DECOMPOSITION

For each nonempty type, averaging simplex-valued assignments gives $\sum_k q_{dc}(k) = 1$. Hence $\sum_{c,k} P_d(c,k) = \sum_c \pi_{dc} = 1$ and $\|u_d\|_2^2 = 1$. For a type block,

$$\|u_{d,c}\|_2^2 = \sum_k \pi_{dc} q_{dc}(k) = \pi_{dc}, \qquad q_{dc}(k) = \frac{u_{d,ck}^2}{\pi_{dc}} \quad \text{if } \pi_{dc} > 0. \tag{12}$$

If $\pi_{dc} = 0$, the block is zero and its conditional state cannot be recovered, nor does it affect any distance.

Proof of Proposition 1. Expanding the donor inner product and using the definition of squared Hellinger distance,

$$u_a^\top u_b = \sum_c \sqrt{\pi_{ac}\pi_{bc}} \sum_k \sqrt{q_{ac}(k) q_{bc}(k)} \tag{13}$$

$$= \sum_c \sqrt{\pi_{ac}\pi_{bc}} \big[1 - H^2(q_{ac}, q_{bc})\big]. \tag{14}$$

Subtracting this expression from one gives

$$1 - u_a^\top u_b = \left(1 - \sum_c \sqrt{\pi_{ac}\pi_{bc}}\right) + \sum_c \sqrt{\pi_{ac}\pi_{bc}} H^2(q_{ac}, q_{bc}), \tag{15}$$

which is Equation (7). Each term is nonnegative, since squared Hellinger distances and the weights are nonnegative. □

The decomposition concerns the finite probability table defined by the learned code assignments. It does not require density estimation in the original gene-expression space and does not equate the code distribution with a complete model of cellular expression. If every donor has the same conditional code distribution for a given type, that type's state contribution is zero. If this occurs for all types, donor distances reduce to composition distances even though donor vectors need not be globally identical.

B.2 DISTANCE CONTRIBUTIONS AND COMPOSITION-ONLY COMPARISON

Write $D_{\mathrm{comp}}(a,b) = H^2(\pi_a, \pi_b)$ and $S_c(a,b) = \sqrt{\pi_{ac}\pi_{bc}} H^2(q_{ac}, q_{bc})$. Since each conditional squared Hellinger distance lies in $[0,1]$,

$$0 \le \sum_c S_c(a,b) \le 1 - D_{\mathrm{comp}}(a,b), \qquad D_{\mathrm{comp}}(a,b) \le D(a,b) \le 1. \tag{16}$$

Thus state differences add to, but cannot cancel, the observed composition discrepancy. Equality $D = D_{\mathrm{comp}}$ holds precisely when the conditional code distributions agree for every type with positive mass in both donors. If no type is shared, $D = D_{\mathrm{comp}} = 1$ and there is no observed within-type state comparison. A zero state contribution for an absent type therefore does not establish biological state agreement.

The complete contribution of type c to the distance between donor coordinates is

$$\frac{1}{2}\|u_{a,c} - u_{b,c}\|_2^2 = \frac{1}{2}(\sqrt{\pi_{ac}} - \sqrt{\pi_{bc}})^2 + S_c(a,b) \le \frac{\pi_{ac} + \pi_{bc}}{2}. \tag{17}$$

This follows by expanding the square and using the nonnegative affinity $\sum_k \sqrt{q_{ac}(k) q_{bc}(k)}$. A rare type can account for a large *fraction* of the total distance when other types are similar, despite this bound on its absolute contribution. Such a distance share is not evidence that rare types are automatically amplified or biologically more important.

A matched composition-only comparison follows by fixing each conditional state distribution to an arbitrary reference r_c shared across donors:

$$P_d^{\mathrm{comp}}(c,k) = \pi_{dc} r_c(k), \qquad H^2(P_a^{\mathrm{comp}}, P_b^{\mathrm{comp}}) = H^2(\pi_a, \pi_b) = \frac{1}{2}\|\sqrt{\pi_a} - \sqrt{\pi_b}\|_2^2. \tag{18}$$

No reference-state fitting is required to evaluate this distance. Comparing it with D under the same distance-based readout, using identical cells and reference/query donors, isolates the inclusion of the state-distance term in that readout. This is a mathematical control, not an additional experimental result; the centered-log composition baseline in Appendix E.4 uses a different geometry.

B.3 Permutation invariance and exact accumulation

Fix the neural parameters, preprocessing, scale references, and type references, and disable stochastic training layers. Every cell then has a deterministic assignment α_i independent of the other cells processed alongside it. Type counts and assignment sums are unchanged by permutations of the cells, so q, P, and u are permutation-invariant.

For disjoint nonempty subsets S_1, S_2 of one type, with $q(S) = |S|^{-1} \sum_{i \in S} \alpha_i$,

$$q(S_1 \cup S_2) = \frac{|S_1| q(S_1) + |S_2| q(S_2)}{|S_1| + |S_2|}. \tag{19}$$

Thus an inference implementation maintains, for each type c, a count n_c and a sum $A_c = \sum_{i:c_i=c} \alpha_i$. After all chunks have been processed, $q_c = A_c/n_c$ for $n_c > 0$, and $\pi_c = n_c / \sum_{c'} n_{c'}$. Equivalently, the joint table satisfies

$$P(c,k) = \frac{A_{ck}}{N}. \tag{20}$$

The square root is taken only after accumulation. Averaging chunk-level square-root vectors generally does not reproduce the full representation. The accumulator requires $O(CK)$ memory in addition to the cell-encoder chunk; pairwise similarities for M donors require $O(M^2CK)$ arithmetic and $O(M^2)$ storage if the full similarity matrix is materialized. Permutation invariance and mergeability hold for a fixed cell collection; they do not imply invariance to sampling a different collection.

C Network architecture and cell-level objectives

C.1 Gene-token encoder and block-specific readouts

Let s_{ig} be the preprocessed expression input for cell i and gene g, and let $o_{ig} \in \{0,1\}$ indicate whether that gene is observed in the current input view. The encoder F_θ includes input preprocessing; masked and full reconstruction inputs are defined in Appendix C.2. The encoder constructs

$$t_{ig} = \mathrm{LN}\left(f_{\mathrm{tok}}([e_g;\, o_{ig} s_{ig};\, o_{ig}])\right), \tag{21}$$

where e_g is a learned gene embedding and f_{tok} is a two-layer GELU MLP. Gene identities distinguish masked genes from one another, while the indicator distinguishes an observed zero from a masked value. Learned queries v_j pool these tokens over genes:

$$a_{ijg} = \frac{\exp(v_j^\top W_K t_{ig}/\sqrt{D_a})}{\sum_{g'} \exp(v_j^\top W_K t_{ig'}/\sqrt{D_a})}, \qquad r_{ij} = \sum_g a_{ijg} W_V t_{ig}. \tag{22}$$

This softmax operates over genes inside a cell and is distinct from the state-code softmax in Equation (5). It maps G gene tokens to eight query summaries $r_{ij} \in \mathbb{R}^{64}$. The queries are partitioned into four groups of two. Let s_i^{std} denote the expression profile after applying the model's stored per-gene centering and scaling buffers, with masked entries set to zero. These buffers remain fixed at evaluation. For each group j, a separate direct-path MLP d_j maps this profile to 32 dimensions, and a readout MLP h_j combines it with the two query summaries:

$$b_i^{(j)} = h_j\big([r_{i,2j-1}; r_{i,2j}; d_j(s_i^{\mathrm{std}})]\big) \in \mathbb{R}^{64}, \qquad j = 1, \ldots, 4.$$

Thus each readout takes $2 \times 64 + 32 = 160$ inputs and returns 64 features; concatenation yields the 256-dimensional vector in Equation (3). The three functional groups in the main text correspond to $b_i^{\text{cell}} = b_i^{(1)}$, $b_i^{\text{pat}} = [b_i^{(2)}; b_i^{(3)}]$, and $b_i^{\text{res}} = b_i^{(4)}$. The two patient readouts are combined for the donor head, without assigning them distinct biological meanings. The donor readout uses the cell and patient features, while the count decoder also uses the residual features. These roles specify information routing, not a guaranteed separation of biological or technical factors.

The two softmax operations should not be conflated: query pooling aggregates gene information into cell features, whereas the donor state head maps each cell to probabilities over state codes.

C.2 Count reconstruction and cell-type supervision

Gene masking. At each training step, $\mathrm{MaskOrFull}$ independently assigns each cell a masked input with probability 0.75 or a full input with probability 0.25. For a masked input, we uniformly sample 25% of the panel genes without replacement, set their input values and observation indicators to zero, and use these genes as the reconstruction target set T_i. The mask is resampled at each step. For a full input, all genes are visible and $T_i = \{1, \dots, G\}$. Denote the resulting input view by $\widetilde{x}_i$. With o_{ig} indicating visibility, expression inputs are normalized using the visible count total and log-transformed:

$$\ell_i^{\text{vis}} = \max\left(1, \sum_g o_{ig} x_{ig}\right), \qquad s_{ig} = \log\left(1 + 10^4 \frac{o_{ig} x_{ig}}{\ell_i^{\text{vis}}}\right).$$

Gene identities remain available in masked tokens, and the observation indicator distinguishes masked entries from observed zeros. Reconstruction targets retain their raw counts, and the decoder uses the full panel count total $\ell_i = \sum_g x_{ig}$ as its exposure. Cell-type supervision and donor representation learning use a separate encoder pass on the full gene inputs of the same sampled cells.

Cell decoder and reconstruction loss. The cell decoder reads the complete cellular representation $z_i = F_\theta(\widetilde{x}_i)$, including the cell, patient, and residual blocks. It shares the gene embedding table e_g with the encoder. A learned linear projection forms a query vector w_g for each gene, and a two-layer GELU MLP r_ψ projects the cell representation into the same space. Their inner product, together with a learned gene bias β_g, defines the expression logit:

$$w_g = W_q e_g, \qquad [f_\psi(z_i)]_g = \beta_g + w_g^\top r_\psi(z_i). \tag{23}$$

The decoder MLP uses widths $256 \to 128 \to 128$, and W_q maps the 64-dimensional gene embeddings to 128-dimensional queries. A softmax over genes gives the predicted proportions $\rho_\psi(z_i)$, which are scaled by ℓ_i to obtain expected counts. Each gene has a learned inverse dispersion $\vartheta_g > 0$, optimized in log space:

$$\rho_\psi(z_i) = \mathrm{softmax}(f_\psi(z_i)), \qquad x_{ig} \sim \mathrm{NB}(\ell_i \rho_{\psi,g}(z_i), \vartheta_g). \tag{24}$$

The negative binomial is parameterized by mean μ and variance $\mu + \mu^2/\vartheta$. For a cellular batch $\mathcal{I}$ and the target sets T_i defined above,

$$\mathcal{L}_{\text{NB}} = -\frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} \frac{1}{|T_i|} \sum_{g \in T_i} \log p_{\text{NB}}(x_{ig}; \ell_i \rho_{\psi,g}(z_i), \vartheta_g). \tag{25}$$

The loss first averages over target genes within each cell, then over cells, giving masked and full-input reconstruction equal weight per cell.

Cell-type supervision. For cell-type supervision, normalize nonzero cell features as $\bar{b}_i = b_i^{\text{cell}} / \|b_i^{\text{cell}}\|_2$ and let $\delta_{ij} = 1 - \bar{b}_i^\top \bar{b}_j$. Positive pairs $\mathcal{E}_+$ have the same type but come from different donors; negative pairs $\mathcal{E}_-$ have different types. With margins $0 \le m_{\text{in}} < m_{\text{out}} \le 2$,

$$\mathcal{L}_{\text{cell}} = \frac{1}{|\mathcal{E}_+|} \sum_{(i,j) \in \mathcal{E}_+} [\delta_{ij} - m_{\text{in}}]_+ + \frac{1}{|\mathcal{E}_-|} \sum_{(i,j) \in \mathcal{E}_-} [m_{\text{out}} - \delta_{ij}]_+. \tag{26}$$

An empty pair-set term is omitted. The positive hinge stops rewarding further attraction once the margin is satisfied; this does not establish an entropy lower bound or guarantee preservation of all within-type variation.

C.3 Reference statistics and gradient paths

The normalization and type references used in Equation (4) are

$$\widetilde{b}^{\mathrm{pat}}_{di} = b^{\mathrm{pat}}_{di}/\sigma_{\mathrm{pat}}, \qquad \widetilde{b}^{\mathrm{cell}}_{di} = \mathrm{sg}(b^{\mathrm{cell}}_{di})/\sigma_{\mathrm{cell}}, \tag{27}$$

where sg is stop-gradient. The scalar RMS references σ_{pat} and σ_{cell} are estimated from training cells, shared by all donor views in a step, and frozen at inference. For type c, let $s_c = \mu_c + \Delta_c$ be a shared reference: μ_c is a running training centroid in normalized cell-feature coordinates and Δ_c is a learned correction.

$\mathrm{UpdateReferences}$ updates the block RMS scales and cell-type centroids from training features, outside backpropagation. Type references are shared by every donor. The learned correction Δ_c is initialized to zero, while the type embeddings and both MLPs are ordinary trainable parameters. A centroid is updated only when enough valid cells of that type are available in the current batch; sparse batches otherwise retain its existing estimate. Parameters and buffers are fixed at evaluation, and dropout is disabled.

Stop-gradient in Equation (27) applies to the cell-feature input, not to W_{cell} or Δ_c. The patient-feature path and the cell path share upstream encoder parameters, so stopping the direct cell-input gradient is not equivalent to freezing the entire cell encoder. Reconstruction and cell-type supervision continue to train their respective paths. Since π is computed by counting, its parameter derivatives are zero; its values nevertheless influence donor similarities and the weighting of state gradients.

C.4 Shared state coordinates across donors

For a given cell type c, all donors use the same learned type embedding t_c, type reference s_c, and state network g_η. With the trained parameters fixed, the probability assigned to state k is a common function of the fused cell feature h:

$$a_{ck}(h) = \frac{\exp([g_\eta([h; t_c])]_k)}{\sum_{\ell=1}^{K} \exp([g_\eta([h; t_c])]_\ell)}. \tag{28}$$

For a donor d with $n_{dc} > 0$, the coordinate $q_{dc}(k)$ averages this same function over its type-c cells:

$$q_{dc}(k) = \frac{1}{n_{dc}} \sum_{i \in S_{dc}} a_{ck}(\widetilde{h}_{di}).$$

Thus, index k refers to the same learned assignment rule for type c across donors, making $q_{ac}(k)$ and $q_{bc}(k)$ directly comparable. For example, values of 0.3 and 0.1 indicate greater average membership in state k among donor a's type-c cells than among donor b's. The donor losses train these assignment rules jointly across donors, while shared preprocessing and type references keep their inputs in common feature coordinates. This correspondence enables the type-specific Hellinger distance to compare the two donors coordinate by coordinate.

D Sampling, inference, and reporting conventions

Observed composition and shared vocabulary. All donors use the same gene panel and annotated cell-type vocabulary, with a fixed type order in the final vector. Counts refer to cells retained after quality control and vocabulary filtering. The resulting π_d describes the composition of the observed collection; it is not asserted to be an unbiased estimate of in-vivo abundance. Its count-based definition is separate from the learned conditional state summaries q_{dc}. Neither these summaries nor the final donor coordinates are free parameters assigned to donor identities. The benchmark evaluation pipeline uses supplied manual cell-type annotations for `main-test-inner` and supplied scANVI-transferred annotations for `validation` and `ext-independent`. These annotations are inputs to **CoState**, not predictions of its cell encoder. The following training and evaluation conventions specify which cells supply the counts.

Training composition and view construction. Training proportions are computed on each donor's eligible training pool, after removing held-out cells and applying the training-pool definition, but before drawing a per-step bag. They are not estimated anew from the type-stratified bag. $\mathrm{TypeCoveringSample}$ reserves at least one cell for each observed type when its budget covers the vocabulary, and draws the remaining cells from the remaining pool. $\mathrm{TypeStratifiedSplit}$ randomly divides each type's sampled cells between two nonempty views; a singleton is placed in both. The same encoded cellular features are used to construct the full-bag and half-bag readouts. Full-bag vectors enter disease supervision, and half-bag vectors enter the positive-only consistency term.

Writing F for the full sampled bag and A, B for its two stratified views, all three share the training-pool composition:

$$P_d^r(c,k) = \pi_{dc} q_{dc}^r(k), \qquad u_d^r = \mathrm{vec}\sqrt{P_d^r}, \quad r \in \{F, A, B\}. \tag{29}$$

Because both views reuse the same encoded feature for a singleton, that type's consistency contribution is exactly zero; this does not measure its sampling stability. A budget smaller than the number of observed types also cannot satisfy complete type coverage. Such settings must not be interpreted as the stated coverage-preserving protocol. In particular, inserting a uniform q_c into an unsampled type with positive externally supplied π_c creates an imputed state, not evidence of a uniform biological state. This differs from the harmless absent-type convention in the main text, where $\pi_c = 0$.

Class-balanced disease objective. Let $\mathcal{I}_y$ contain labeled anchors of disease y with an eligible same-disease positive, and let $\mathcal{Y}$ contain the diseases with valid anchors. The reduction used in the main objective is

$$\mathcal{L}_{\mathrm{dis}} = \frac{1}{|\mathcal{Y}|} \sum_{y \in \mathcal{Y}} \frac{1}{|\mathcal{I}_y|} \sum_{a \in \mathcal{I}_y} \ell_a. \tag{30}$$

Self comparisons and unlabeled donors are excluded. $\mathrm{DiseaseBalancedSample}$ selects disease classes and draws an equal number of distinct donors per class, with at least two classes and two donors per class. This supplies same-disease positives and different-disease comparisons. The network receives no donor-ID embedding. Disease labels are used only by the training loss, and neither they nor training-gallery comparisons are needed to construct a new donor vector.

Independent new-donor encoding. All preprocessing parameters, state-network weights, normalization scales, and type references are frozen. Counts are accumulated from the evaluation collection, so an absent type contributes no mass. Cells outside the predefined valid type vocabulary are excluded consistently from both counts and state sums; a donor with no remaining valid cells has no defined representation and must be excluded. An annotation model, if used, must assign the same vocabulary without accessing donor disease labels.

E Data and experimental protocol

E.1 Data source, preprocessing, and splits

We use the three post-QC count matrices and sample metadata released with the Inflammation Atlas (Jiménez-Gracia et al., 2026) (Zenodo record 14851902). Our benchmark includes 817 main-cohort donors, 144 validation donors, and 46 independent external donors. We additionally remove cells annotated as `Doublets` or `LowQuality_cells` and align genes by Ensembl ID. The 2,000-gene panel is a voting consensus of disease-specific highly variable gene rankings, excluding immunoglobulin/T-cell-receptor variable segments and haemoglobin genes. This panel was selected on the entire main cohort before donor splitting, including the internal test donors, and is therefore label-informed preprocessing with internal-test exposure. The validation and external cohorts did not enter gene selection.

For **CoState**, visible counts are normalized to 10,000 counts per cell and log-transformed: $s_{ig} = \log(1 + 10^4 x_{ig}/\ell_i)$ for a full input, where ℓ_i is the sum over panel genes. Masked inputs use only the visible-gene sum in this normalization. The negative-binomial decoder receives raw-count targets and the full panel library size as its exposure. The encoder's direct path uses the per-gene normalization buffers stored with the model, which remain fixed at evaluation.

The main cohort is split by donor and sequencing-library components, stratified by study and disease. The 653 training and 82 development donors form the final 735-donor training pool (16 studies, 20 disease/control classes). A deterministic 10% cell holdout within these donors leaves 3,499,783 training cells for the uncapped setting. This cell holdout is distinct from the atlas validation cohort. The remaining 82 main donors form the internal test set. Validation contains 144 unseen donors from six training studies. External evaluation contains 46 donors from seven studies absent from training. No evaluation donor contributes to neural-model training, except for the explicitly transductive baselines below.

Manual level-1 annotations are available for main-cohort cells. Validation and external cells instead carry the atlas's scANVI-transferred annotations, which serve as inputs to the donor readout. We consequently evaluate cell-type accuracy only on the manually annotated internal test set. Disease and study are strongly associated in these cohorts, so an unrestricted disease score alone does not establish study-independent disease information.

E.2 Contributing studies and source publications

Our benchmark uses 23 studies from the Inflammation Atlas. We retain the released identifiers when defining study boundaries and trace each source to the atlas's Supplementary Table 1 and its original publication or data release (Jiménez-Gracia et al., 2026). Table 3 reports the donors retained in our benchmark splits, rather than the full recruitment counts of the source publications. Sixteen studies contribute to model training, seven supply the independent external evaluation. Validation donors come from six of the training studies. Only the atlas-selected blood-cell data are used here, even when the original publication also profiles tissue, cerebrospinal fluid, or other assays. Additional donors used for biological characterization are specified in Appendix G.4.

E.2.1 Atlas-generated studies represented in training

These six studies were generated for the Inflammation Atlas and are described in its primary publication (Jiménez-Gracia et al., 2026). They share the SCGT prefix but remain distinct study identifiers in our analysis.

SCGT00. The DoCTIS cohort contains rheumatoid arthritis (RA), psoriatic arthritis (PsA), Crohn's disease (CD), ulcerative colitis (UC), psoriasis (PS), systemic lupus erythematosus (SLE), and healthy controls. It supplies a multidisease cohort collected with the Vall d'Hebron Research Institute (Jiménez-Gracia et al., 2026). Raw-data access is administered through EGA DAC EGAC50000000566.

SCGT01. This healthy-control cohort was collected with the Institut Hospital del Mar d'Investigacions Mèdiques (Jiménez-Gracia et al., 2026). It contributes healthy reference donors to training; its data are deposited as GSE248689.

SCGT02. This University Medical Center Groningen cohort includes asthma, chronic obstructive pulmonary disease (COPD), and healthy controls (Jiménez-Gracia et al., 2026). All retained donors enter our training pool. The source accession is GSE248695.

SCGT03. This cohort contains circulating immune cells from patients with breast cancer (BRCA), collected with the Vall d'Hebron Institute of Oncology (Jiménez-Gracia et al., 2026). The data are deposited as GSE248685.

SCGT04. This cirrhosis cohort was collected with the Biomedical Research Institute Sant Pau (Jiménez-Gracia et al., 2026). Its blood-cell profiles provide the training-study source for cirrhosis, with data available as GSE248693.

SCGT05. This colorectal cancer (CRC) cohort was collected with KU Leuven (Jiménez-Gracia et al., 2026). The raw-data study is EGAS50000000590, with dataset EGAD50000000841. We use the processed PBMC profiles distributed with the atlas.

Table 3: All 23 source-study identifiers and donor counts in our benchmark. Train includes the 653 training and 82 development donors. Test is the internal donor test set; Val is the atlas validation cohort; Ext is the independent external cohort. HC denotes the atlas healthy label; clinical abbreviations and original-study context are defined in Appendix E.2. Counts refer to our retained donor units, not the full cohorts recruited by the source publications.

Study ID	Retained atlas labels	Train	Test	Val	Ext
Studies represented in training					
SCGT00	RA, PsA, CD, UC, PS, SLE, HC	137	20	27	0
SCGT01	HC	7	0	0	0
SCGT02	Asthma, COPD, HC	32	0	0	0
SCGT03	BRCA	5	1	0	0
SCGT04	Cirrhosis	6	2	0	0
SCGT05	CRC	13	1	0	0
Reyes2020	Sepsis, HC	26	4	0	0
Cillo2020	HNSCC, HC	26	1	5	0
Zhang2023	HBV, HC	17	1	0	0
Schafflick2020	MS, HC	9	1	0	0
Liu2021	NPC	9	1	0	0
Wang2020	HIV, HC	6	1	0	0
Perez2022	SLE, HC	197	0	50	0
COMBAT2022	COVID-19, influenza, sepsis, HC	84	17	20	0
Ren2021	COVID-19, HC	49	11	9	0
Terekhova2023	HC	112	21	33	0
Independent external studies					
Jiang2020	Sepsis	0	0	0	7
Palshikar2022	HIV	0	0	0	8
Ramachandran2019	Cirrhosis	0	0	0	4
Martin2019	CD	0	0	0	8
Mistry2019	SLE	0	0	0	3
Savage2021	SLE, HC	0	0	0	8
10XGenomics	HC	0	0	0	8
Total		735	82	144	46

E.2.2 PUBLISHED STUDIES REPRESENTED IN TRAINING

Reyes2020. Reyes et al. (2020) profiled circulating immune cells in bacterial sepsis and identified a sepsis-associated monocyte state. The atlas retains sepsis and healthy-control samples from this study. Its PBMC expression data are available through the Broad Single Cell Portal as SCP548.

Cillo2020. Cillo et al. (2020) compared immune populations in viral- and carcinogen-associated head and neck squamous cell carcinoma (HNSCC). The atlas uses the peripheral-blood component, including HNSCC and healthy donors, from GSE139324.

Zhang2023 (Zhang2022 in the atlas study table). Zhang et al. (2023) studied liver and peripheral immune cells across phases of hepatitis B virus (HBV) infection. The atlas retains blood samples with HBV and healthy labels, sourced from HRA001730. The study table uses the 2022 online-publication year, whereas the released sample metadata and our split manifest use the 2023 journal-issue year. These identifiers refer to the same study.

Schafflick2020. Schafflick et al. (2020) compared blood and cerebrospinal-fluid leukocytes in multiple sclerosis (MS). We use the atlas's blood-cell subset from GSE138266. The atlas labels its comparator samples as healthy, although the original control cohort comprised individuals with idiopathic intracranial hypertension. We retain the released labels for benchmark consistency.

Liu2021. Liu et al. (2021) characterized tumor heterogeneity and cellular interactions in nasopharyngeal carcinoma (NPC). The atlas-selected peripheral-blood samples contribute NPC donors; the sequencing study is SRP293799.

Wang2020. Wang et al. (2020) examined immune-cell exhaustion in people with HIV using single-cell PBMC profiles. The atlas retains HIV and healthy-control samples from SRP282057, providing the HIV study represented during training.

Perez2022. Perez et al. (2022) used multiplexed single-cell RNA sequencing to study cell-type-specific molecular and genetic associations with SLE. The atlas includes SLE and healthy donors from SRP319208. This study contributes training and validation donors but no donors to our internal test split.

COMBAT2022. The COMBAT study compared blood immune profiles across COVID-19 severity, influenza, sepsis, and healthy controls (Ahern et al., 2022). All four diagnostic categories occur in our retained data. The atlas source is the COMBAT CELLxGENE collection. Together with Ren2021, it supplies COVID-19 donors for cross-study retrieval on the internal test and validation cohorts.

Ren2021. Ren et al. (2021) assembled a large single-cell atlas of immune responses in COVID-19. The retained blood-cell data contain COVID-19 and healthy donors and originate from the Ren CELLxGENE collection. Its distinct study identifier allows COVID-19 retrieval across studies.

Terekhova2023. Terekhova et al. (2023) profiled healthy human blood to characterize age-associated changes in immune populations. This study supplies healthy donors across training, internal test, and validation. Its source data are deposited at Synapse as syn49637038.

E.2.3 INDEPENDENT EXTERNAL STUDIES

The following seven study identifiers are absent from encoder training. Their 46 retained donors form the independent external evaluation cohort.

Jiang2020. Jiang et al. (2020) investigated early monocyte transcriptional signatures in sepsis-associated acute respiratory distress syndrome. The atlas contributes seven donors labeled sepsis, with sequencing data under SRP264928. ARDS status is not a separate prediction label in our benchmark.

Palshikar2022. Palshikar et al. (2022) developed executable immune-signaling models using PBMCs from people living with HIV, with or without atherosclerosis. The eight retained donors carry the HIV label in the atlas. Their data are deposited as GSE198339.

Ramachandran2019. Ramachandran et al. (2019) characterized the cellular organization of the fibrotic niche in human liver cirrhosis. The atlas uses four cirrhosis PBMC donors from this study, whose data are deposited as GSE136103. The liver-tissue profiles are outside our input dataset.

Martin2019. Martin et al. (2019) studied cellular programs in Crohn's disease lesions and their association with resistance to anti-TNF therapy. We use the eight atlas-selected blood donors with CD labels from GSE134809. Lesion profiles and treatment-response outcomes are not inputs to our donor representation.

Mistry2019. Mistry et al. (2019) investigated neutrophil heterogeneity in SLE using transcriptomic, epigenetic, and functional analyses. The atlas-selected single-cell data contribute three external SLE donors, with source accession GSE142016.

Savage2021. Savage et al. (2021) examined how blood-processing delays affect molecular measurements in human ex vivo studies. The atlas retains eight donors with healthy or SLE labels from GSE156989. Its SLE donors and those from Mistry2019 permit external cross-study retrieval of the same disease.

10XGenomics. This identifier pools eight healthy donors from official 10x Genomics PBMC data releases. The atlas's Supplementary Table 1 lists Chromium Connect and manual preparations, Chromium Controller and Chromium X datasets, and a 5′ cell-hashing example (Jiménez-Gracia et al., 2026). These are technical data releases; the source table assigns no separate research article to this group. They remain one study identifier in our evaluation.

E.3 Task and metric definitions

Within-cohort disease readout. For each repeat, each donor's exported pool (at most 1,000 cells) supplies two disjoint random subsamples A_d, B_d of $n = 250$ cells each ($n = 125$ for the smaller-budget analysis). Donors with fewer than $2n$ exported cells are excluded. Each subsample supplies its own composition and state estimates. We randomly divide eligible donors into reference and query halves, fit readouts on reference A views, and score query B views, retaining classes present on both sides. Thus this evaluates information in frozen representations within each cohort, rather than transfer of a classifier trained on the atlas training cohort.

CoState uses $D(a, b) = 1 - u_a^\top u_b = H^2(P_a, P_b)$; other donor vectors use cosine distance. kNN copies the label of the nearest reference donor. Proto. selects the nearest normalized class-mean vector. Linear is class-balanced logistic regression ($C = 0.1$) on unit-normalized vectors, without coordinate-wise standardization. Macro-F1 is $|\mathcal{Y}|^{-1} \sum_{y \in \mathcal{Y}} 2\mathrm{TP}_y/(2\mathrm{TP}_y + \mathrm{FP}_y + \mathrm{FN}_y)$, where $\mathcal{Y}$ is the scored class set. We average five sampling/readout repeats per fitted model. Reported training-seed standard deviations use three seeds for **CoState**, the trained baselines, and MrVI; deterministic or single-run methods have no training-seed error bar.

Cross-study disease retrieval. Each donor is represented by 250 sampled cells, with three resamples. For query d, the reference gallery contains all donors in the evaluation cohort with $s_j \neq s_d$, including other diseases. Retrieval F1 is 1-NN disease macro-F1. A scored disease must have at least two donors in each of at least two studies. This leaves 47 scored test donors from four studies and 75 validation donors from four (healthy/COVID-19), and 19 independent external donors from three studies (healthy/SLE). These are subsets of the 82-, 144-, and 46-donor cohorts, respectively. The encoder remains frozen; gallery exclusion does not retrain it. Figure 2a–c displays the two methods per family with the highest mean retrieval F1 over the three splits; the full ranking is in Figure 6 and Table 5.

Disease over study is the per-query AUROC comparing all same-disease donors from other studies against all other-disease donors from its own study, using negative distance as the score. Both sets must be nonempty; 26, 33, and 8 queries qualify. A value of 0.5 indicates no preference. *Study*

mixing is the within-disease mean of $1 - |s|$, where s is the silhouette coefficient for study labels; higher values indicate more mixing. These diagnostics accompany retrieval, since mixing alone can also result from losing disease information.

Stability and study information. R@1 is the fraction of donor views retrieving their matching view as the nearest neighbour among all opposite-half views, averaged over $A \to B$ and $B \to A$. *Closer* is the fraction for which the nearest other donor with the same disease is closer than the donor's own other half (lower indicates better identity consistency). *Study* is the same 1-NN macro-F1 readout with study labels; a high value indicates accessible study information. High R@1 can reflect donor-specific technical effects as well as stable biology.

Cell-type structure and transfer. We evaluate **CoState**'s 64-dimensional cell block and each baseline's full cell embedding. In Table 2, * denotes cell-type supervision and † denotes transductive methods. NMI (normalized mutual information) measures shared information between clusters and manual types; ARI (adjusted Rand index) measures pairwise clustering agreement corrected for chance. Both are reported for Leiden and k-means. Leiden resolution is optimized against the labels in the scoring procedure, so this is a label-assisted clustering evaluation. The separate cell kNN probe uses $k = 15$ and excludes the query cell itself. Transfer F1 restricts that gallery to other studies. *Type first* is the fraction satisfying

$$\min_{j:c_j=c_i,\, s_j\neq s_i} \|z_i - z_j\|_2^2 < \min_{j:c_j\neq c_i,\, s_j=s_i} \|z_i - z_j\|_2^2.$$

Both candidate sets must be nonempty; ties count as failures. Cross-study cell metrics use 250 cells per donor and three resamples, with equal weight over 13 types having at least 200 cells in the exported pool. Rarer types remain in the gallery. All 13 studies in this test set were represented during encoder training.

E.4 Baseline implementations and supplementary-analysis conventions

Panel-based baselines use the same 2,000-gene panel and training-donor partition. Harmony, InfoGlobe, and MrVI additionally access evaluation cells or their study/sample identifiers as specified below and are marked transductive. In the formulas below, x denotes a cell's count vector, z its baseline embedding, and v_d a donor vector. Unless stated otherwise, we use $v_d = |V_d|^{-1} \sum_{i \in V_d} z_{di}$ for a 250-cell view V_d and compare donor vectors using cosine distance. Foundation models use full-transcriptome inputs from the same evaluation cell pools, with model-specific preprocessing and frozen pretrained weights. Their gene access and external pretraining differ from the panel-based models; pretraining overlap with this atlas has not been excluded. For these models, T_θ denotes the frozen transformer and the CLS output denotes its cell-summary token.

scVI. We train the scvi-tools variational autoencoder (Lopez et al., 2018) with a negative-binomial decoder and a 256-dimensional latent space. Its count model is $p_\theta(x \mid z, \ell) = \prod_g \mathrm{NB}(x_g; \ell \rho_{\theta g}(z), \vartheta_g)$, where ℓ is sequencing depth, $\rho_\theta(z)$ gives gene-expression proportions, and ϑ_g controls dispersion. Thus the latent state predicts mean expression while the likelihood accounts for count noise. Its adapter uses the shared masked/full reconstruction protocol and reference-depth normalization, with no study or cell-type covariates. Posterior means provide the cell embeddings used for donor averaging.

scANVI. We use the scvi-tools semi-supervised variational model (Xu et al., 2021) with annotated cell types as the classification targets. The latent hierarchy $p(z, h, c) = p(z \mid h, c)p(h)p(c)$ allows the embedding z to vary within each supervised type c through an auxiliary latent variable h. It shares scVI's input panel, latent dimension, and reconstruction protocol. Cell-type labels supervise training; donor vectors average the resulting cell posterior means.

scANVI (disease). This variant replaces cell-type targets with the disease label of each cell's donor, while retaining the scANVI architecture (Xu et al., 2021) and training protocol. The classification term is $\mathcal{L}_{\mathrm{cls}} = -\mathbb{E}_{d,i} \log q_\phi(y_d \mid x_{di})$, where y_d is the donor's disease label and q_ϕ the classifier. This encourages individual cell embeddings to carry donor-level disease information. Donor vectors are obtained by mean pooling.

FlatVI. We use the released negative-binomial VAE (Palma et al., 2025) with its Fisher-metric flattening penalty and a 256-dimensional latent space. The penalty is $\mathcal{R}_{\text{flat}} = \mathbb{E}_z\|G(z) - I\|_F^2$, where $G(z)$ is the decoder's pullback of the negative-binomial Fisher metric. It measures how latent displacements change the decoded count distribution; bringing $G(z)$ towards the identity makes these changes more uniform across latent directions and locations. Reconstruction follows the shared masking and reference-depth conventions; the geometric penalty is evaluated on a cell subset to control memory use. Posterior means are exported without study or donor conditioning.

IRVAE. We adapt IRVAE's released relaxed-distortion regularizer (Lee et al., 2022) to the same negative-binomial VAE backbone used for the geometric baselines. The penalty is $\mathcal{R}_{\text{iso}} = \mathbb{E}_z[\text{tr}(G(z)^2)]/(\mathbb{E}_z[\text{tr}\, G(z)])^2$, with $G(z) = J_f(z)^\top J_f(z)$ and $f(z) = \log \mu_\theta(z)$ at fixed reference depth. Here J_f is the decoder Jacobian. The penalty discourages unequal stretching of latent directions while allowing an overall scale factor; expectations are estimated using sampled latent points and random Jacobian probes. The reported representation consists of 256-dimensional posterior means from this count-data adaptation.

scPhere. We implement scPhere's von Mises–Fisher latent distribution (Ding & Regev, 2021) in the negative-binomial VAE backbone, using a uniform hyperspherical prior. Its posterior has the form $q(z \mid x) \propto \exp\{\kappa(x)\mu(x)^\top z\}$ with $\|z\|_2 = \|\mu(x)\|_2 = 1$. The direction $\mu(x)$ represents cell state, while the concentration $\kappa(x)$ controls angular uncertainty. The cell representation is the posterior mean direction on the unit sphere in $\mathbb{R}^{256}$. This adaptation uses the shared count reconstruction protocol and gene-specific dispersion, with no batch covariates or auxiliary depth regularization.

PILOT-GM. We train the released Gaussian-mixture VAE (Joodaki et al., 2025) on cell coordinates from the training-fitted PCA basis, using its unsupervised reconstruction and mixture-prior objective. Its prior is a mixture $p(z) = \sum_{r=1}^{R} \omega_r \mathcal{N}(z; \mu_r, \Sigma_r)$, where ω_r are normalized component weights. Multiple Gaussian components allow distinct groups in the latent space. Deterministic latent means are averaged within donor views. This evaluates PILOT-GM's cell representation; the full native donor-level mixture fitting and pairwise optimal-transport pipeline is outside this readout.

PaSCient. We retrain PaSCient (Liu et al., 2026) on the atlas gene panel and disease labels, retaining its gene-to-cell projection, nonlinear attention pooling, and patient MLP. For cell features h_{di}, the donor representation is $v_d = g_\theta(\sum_i a_{di} h_{di})$, with $a_{di} = \exp s_\theta(h_{di}) / \sum_j \exp s_\theta(h_{dj})$. The attention scorer s_θ learns which cells to emphasize, and g_θ maps their weighted summary to the patient representation. Training uses donor-level disease cross-entropy as its only loss. The donor vector is the patient MLP output after attention pooling, before the disease-classification head.

PCA. We normalize each cell to 10,000 total counts, apply $\log(1 + x)$, and fit a 256-component PCA basis (Jolliffe & Cadima, 2016) on training cells. The projection is $z = W^\top(a - \bar{a})$, where $a = \log(1 + 10^4 x / \sum_g x_g)$, $\bar{a}$ is the training mean, and the columns of W are the leading training-covariance eigenvectors. These directions retain the largest expression variance. Evaluation cells use the fixed training basis. Averaging their coordinates gives a linear baseline for pooling cell-level expression summaries.

Harmony. We apply Harmony (Korsunsky et al., 2019) to the PCA cell coordinates using study identity as the batch variable. Its correction takes the form $z_i^{\text{corr}} = z_i^{\text{PCA}} - \sum_k r_{ik} \beta_{k,s_i}$, where r_{ik} is soft cluster membership and β_{k,s_i} is the estimated offset for study s_i within cluster k. Each cell therefore receives a correction weighted by its local cluster assignments. Correction is performed separately within each evaluation cohort, and corrected coordinates are averaged by donor. This procedure uses evaluation cells and study labels and is therefore transductive.

InfoGlobe. InfoGlobe (Wang et al.) takes cell counts normalized to gene-probability vectors. Its representation is summarized by $p_i \approx A w_i$, $z_i = \sqrt{w_i}$, where $p_i = x_i / \sum_g x_{ig}$, w_i is a probability vector, and the columns of A are gene-probability profiles. Each cell is represented as a mixture of shared profiles, with square-root coordinates for its mixture weights. We fit a 256-component probability basis on training cells, capped at 400 cells per donor, then hold that basis fixed while fitting

cell-specific coefficients within each evaluation cohort. Square-root coefficient vectors provide the cell embeddings. Fitting evaluation coefficients makes this baseline transductive.

MrVI. We fit the scvi-tools implementation of MrVI (Boyeau et al., 2025) on raw panel counts, supplying donor identity as the sample variable. Its defining map is $z = f_\theta(u, s)$: shared cell-state coordinates u and sample identity s jointly determine the sample-aware coordinates z used to decode expression. This represents sample-associated changes through a conditional transformation of a shared cell state. Each evaluation cohort is combined with training cells for a separate fit. Reported donor vectors average the sample-unaware u coordinates within donors. This baseline is transductive.

scGPT. We use scGPT's released human checkpoint (Cui et al., 2024) and official cell-embedding routine with gene-symbol inputs and checkpoint-specific preprocessing. Gene tokens combine identity and expression as $t_g = e_{\mathrm{gene}}(g) + e_{\mathrm{expr}}(\widetilde{x}_g)$, and the cell vector is $z = [T_\theta(\mathrm{CLS}, \{t_g\}_g)]_{\mathrm{CLS}}$. Here $\widetilde{x}_g$ is the processed expression value; attention combines gene identities and their expression into a contextual cell summary. These cell embeddings are averaged within donor views.

Geneformer. We use Geneformer (Theodoris et al., 2023), specifically its V2 104M checkpoint, with the official transcriptome tokenizer and embedding extractor. Full-transcriptome counts are mapped to human Ensembl identifiers and converted into ranked gene-token sequences. Writing m_g for the pretrained gene-normalization factor, the input order is $r = \mathrm{argsort}^{\downarrow}_{g:x_g>0}(x_g/m_g)$ and $z = [T_\theta(\mathrm{CLS}, r)]_{\mathrm{CLS}}$. This encodes which genes are relatively prominent in a cell through their rank in the sequence. Frozen CLS embeddings are restored to the original cell order and averaged by donor.

Nicheformer. We align genes to Nicheformer's released checkpoint vocabulary (Tejada-Lapuerta et al., 2025) and apply its technology-specific mean normalization. Species, assay, and modality context tokens accompany the expression tokens. With context tokens κ_i and normalized gene ranks r_i, we write $H_i = T_\theta(\kappa_i, r_i)$ and $z_i = |\mathcal{J}_i|^{-1} \sum_{j \in \mathcal{J}_i} H_{ij}$, where $\mathcal{J}_i$ contains the post-context positions pooled by the official routine. The context influences gene-token representations, whose average provides the cell summary. We average these cell vectors by donor.

UCE. We use UCE's released 33-layer model (Rosen et al., 2026) with human protein-derived gene embeddings and full-transcriptome gene-symbol inputs. Genes are sampled with replacement with probabilities $P_i(g) = \log(1 + x_{ig}) / \sum_h \log(1 + x_{ih})$. Their protein embeddings are arranged by genomic position, with chromosome tokens, and passed through the transformer and its CLS readout. Thus expression controls which gene products contribute to the cell summary, while protein embeddings encode their biological relationships. The official preprocessing and inference pipeline produces frozen cell embeddings, which are aligned by cell identifier and averaged within donor views. Cells excluded by the native preprocessing are recorded in export metadata.

Composition only. We count annotated cell types within each donor view, add a pseudocount of 0.5 to each type, giving $\widetilde{\pi}_{dc} = (n_{dc} + 0.5)/(\sum_{c'} n_{dc'} + 0.5C)$ for C types. The centered log-ratio transform (Aitchison, 1982) is $v_{dc} = \log \widetilde{\pi}_{dc} - C^{-1} \sum_{c'} \log \widetilde{\pi}_{dc'}$. Each coordinate measures one type's abundance relative to the geometric mean abundance across types. This baseline uses cell-type abundances without expression features.

Pseudobulk + PCA. We form a pseudobulk profile (Murphy & Skene, 2022) by summing raw gene counts over each donor view, normalize the aggregate to 10,000 total counts, and apply $\log(1 + x)$. This donor profile is projected using the same 256-component basis fitted to training cells for PCA (Jolliffe & Cadima, 2016). In the notation above, $v_d = W^\top [\log(1 + 10^4 t_d / \sum_g t_{dg}) - \bar{a}]$, with $t_d = \sum_{i \in V_d} x_{di}$. The vector summarizes a donor's aggregate expression profile. Aggregating counts before normalization and logarithms yields a different summary from averaging cell-level PCA coordinates.

Training-cell scaling protocol. We vary the per-donor cell cap while keeping donors, gene panel, preprocessing, and evaluation cohorts unchanged. After reserving a fixed cell holdout, we sample

each training pool uniformly without replacement using a shared selection seed across methods and runs, and keep it fixed throughout training. Within each method, architecture, optimization, training steps, and per-step cell budgets remain unchanged. Checkpoints are selected by holdout masked negative-binomial loss. Frozen representations are evaluated following Appendix E.3; Figure 4 and Table 7 report means and standard deviations across training seeds after averaging within-seed evaluation repeats.

Ablation protocols. The w/o donor losses variant is retrained without the disease-contrastive and same-donor objectives and uses mean cell embeddings as donor vectors. With cell embeddings fixed, w/o DE retains only annotated composition distance, whereas w/o DA sums conditional-state distances over types shared by both donors. Scores aggregate training seeds and applicable resampling or clustering repeats; the composition-only readout has no training-seed variance. White circles in Figure 5 show means.

F COMPACT NOTATION REFERENCE

Table 4: Core quantities. Only neural parameters and reference corrections are optimized; donor probabilities and coordinates are computed from cells.

Symbol	Shape	Meaning
x_{di}, c_{di}, y_d	G; scalar labels	Cell counts, annotated type, and donor disease label
n_{dc}, π_{dc}	Scalars	Observed type count and proportion
b_{di}^{cell}	64	Type-supervised cellular feature
b_{di}^{pat}	128	Concatenated patient-feature blocks
$\sigma_{\mathrm{cell}}, \sigma_{\mathrm{pat}}$	Scalars	Training-estimated RMS references
$s_c = \mu_c + \Delta_c$	64	Training centroid plus learned type correction
$\widetilde{h}_{di}, t_c$	64; 16	Fused feature and learned type embedding
α_{di}	K	Per-cell probabilities over state codes
q_{dc}, P_d	K; $C \times K$	Conditional state probabilities and joint donor table
$u_d, u_{d,c}$	CK; K	Donor square-root vector and one type block
$D(a, b)$	Scalar	Squared Hellinger discrepancy between donors
F, A, B	View indices	Full sampled bag and its two stratified views
$M, \tau, \lambda_\bullet$	Scalars	Donor batch size, disease temperature, and loss weights

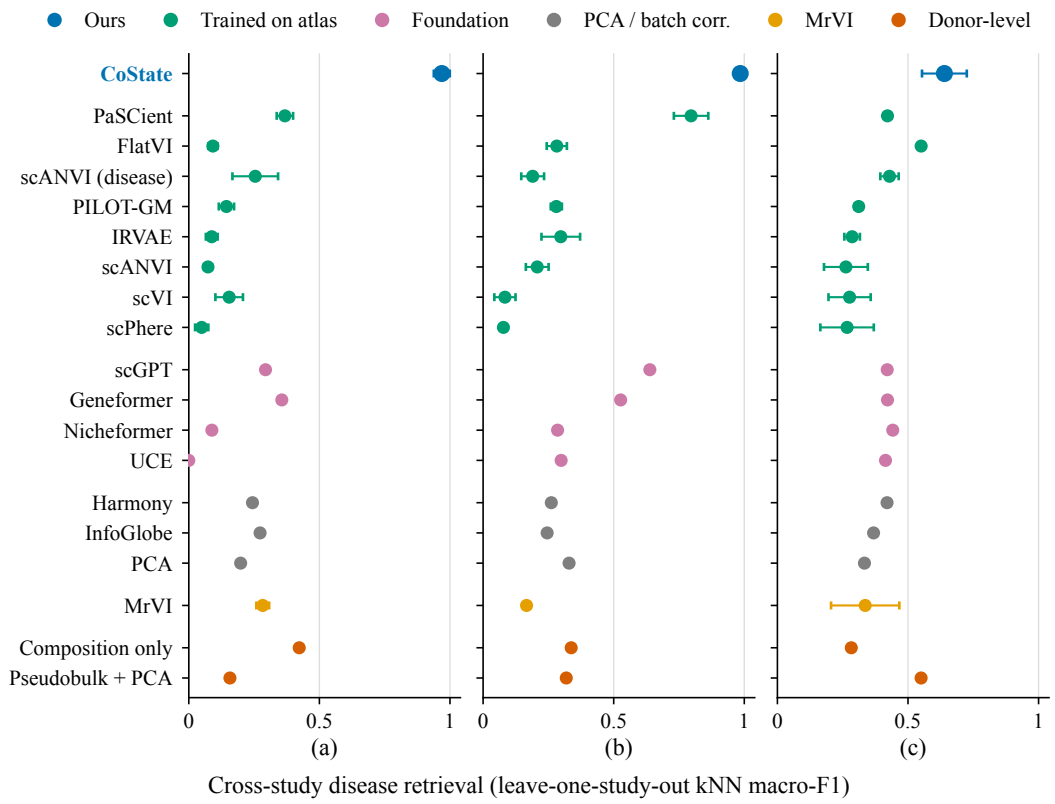

Figure 6: Cross-study disease retrieval for all methods: 1-NN disease macro-F1 with galleries restricted to other studies. Panels: (a) held-out test donors from training studies, (b) unseen validation donors, and (c) donors from studies absent from the training pool. Colours indicate method families; bars show ± 1 s.d. across training seeds where available.

G Supplementary Experimental Results

G.1 Donor Readouts and Cross-Study Comparison

Table 1 compares the three within-cohort disease readouts at 250 cells per view. Table 6 adds the 125-cell setting and training-seed variability. Figure 6 gives the complete cross-study retrieval comparison; Table 5 adds disease-over-study AUROC and within-disease study mixing under the same eligibility rule. Figure 3 visualizes the donor geometry; its **CoState** panel uses $K = 8$ and is illustrative rather than the default $K = 16$ result.

Table 5: Does disease, rather than study, organise the donor space (250 cells per donor, 3 resamples)? Retrieval F1: cross-study disease retrieval, the 1-NN disease macro-F1 when the gallery only holds donors from other studies. Disease over study: for each donor, the probability (AUROC) that a same-disease donor from another study is closer than an other-disease donor from its own study, over all such pairs; 0.5 = no pairwise preference. Study mixing: the batch silhouette of the scIB benchmark with the study as batch, computed within each disease, $1 - |s|$ averaged over donors; 1 corresponds to zero study silhouette, not a guarantee of study independence. Higher mixing need not imply better disease retrieval. **Bold**: best; underlined: second best.

	Retrieval F1 ↑			Disease over study ↑			Study mixing ↑		
Method	Test	Valid.	External	Test	Valid.	External	Test	Valid.	External
Trained on the atlas									
scVI	0.155	0.084	0.276	0.247	0.034	0.423	0.488	0.490	0.810
scANVI	0.073	0.207	0.262	0.219	0.188	0.415	0.634	0.625	0.740
scANVI (disease)	0.254	0.190	0.429	0.255	0.025	0.419	0.615	0.554	0.766
FlatVI	0.092	0.282	0.550	0.353	0.276	0.438	0.653	0.708	0.745
IRVAE	0.088	0.297	0.286	0.266	0.239	0.454	0.673	0.708	0.791
scPhere	0.049	0.078	0.267	0.248	0.066	0.460	0.522	0.578	0.777
PILOT-GM	0.144	0.280	0.311	0.315	0.210	0.391	0.697	0.720	0.725
PaSCient	0.368	0.796	0.421	0.459	0.232	0.439	0.528	0.676	0.774
Linear floor and batch correction									
PCA	0.199	0.329	0.333	0.170	0.035	0.234	0.653	0.681	0.740
Harmony†	0.244	0.261	0.420	0.633	0.514	**0.543**	0.885	0.917	0.849
InfoGlobe†	0.273	0.245	0.368	0.384	0.264	0.334	**0.957**	**0.953**	**0.889**
Transductive									
MrVI†	0.283	0.166	0.336	0.502	0.534	0.442	0.803	0.828	0.715
Foundation models (frozen)									
scGPT	0.294	0.638	0.420	0.144	0.109	0.246	0.536	0.640	0.600
Geneformer	0.356	0.527	0.422	0.072	0.047	0.163	0.548	0.559	0.532
Nicheformer	0.088	0.285	0.442	0.027	0.050	0.158	0.460	0.529	0.515
UCE	0.000	0.299	0.414	0.002	0.000	0.136	0.422	0.472	0.632
Donor-level baselines									
Composition only	0.423	0.337	0.283	0.445	0.430	0.528	0.860	0.837	0.835
Pseudobulk + PCA	0.157	0.318	0.550	0.147	0.139	0.426	0.643	0.686	0.828
CoState (Ours)	**0.969**	**0.984**	**0.639**	**0.829**	**0.683**	0.508	0.817	0.843	0.850

Test (donors): held-out donors of the training studies; Validation (cohort): the atlas' validation set of unseen donors; External (studies): donors from studies absent from the training pool. Donors are evaluated only if their disease has at least 2 donors in each of at least two studies of the split (Test: healthy, COVID-19; Valid.: healthy, COVID-19; External: healthy, SLE). Disease over study additionally needs another disease in the donor's own study: 26, 33, 8 donors on the three splits, against 47, 75, 19 for the other two columns. scANVI (disease) and PaSCient are supervised with disease labels (PaSCient: retrained on this atlas, disease cross-entropy as its only loss). Harmony is run on the PCA coordinates with the study as the batch. †Transductive: fit on the evaluation cells or given their batch labels.

Table 6: Disease kNN macro-F1 for every evaluated method at both half-bag sizes. Mean ± sample s.d. over three training seeds where a method is trained with seeds; other methods are single runs or deterministic. **Bold**: best; underlined: second best.

	Test donors		Validation cohort		External studies	
Method	125 cells	250 cells	125 cells	250 cells	125 cells	250 cells
Trained on the atlas						
scVI	0.632 ± 0.032	0.677 ± 0.060	0.512 ± 0.005	0.551 ± 0.011	0.702 ± 0.006	0.661 ± 0.033
scANVI	0.427 ± 0.047	0.402 ± 0.026	0.379 ± 0.050	0.385 ± 0.010	0.591 ± 0.074	0.604 ± 0.076
scANVI (disease)	0.560 ± 0.002	0.519 ± 0.027	0.553 ± 0.009	0.598 ± 0.012	0.755 ± 0.013	0.740 ± 0.017
FlatVI	0.496 ± 0.007	0.505 ± 0.013	0.406 ± 0.017	0.481 ± 0.017	0.752 ± 0.023	0.724 ± 0.006
IRVAE	0.486 ± 0.025	0.478 ± 0.023	0.422 ± 0.014	0.413 ± 0.021	0.656 ± 0.024	0.603 ± 0.022
scPhere	0.615 ± 0.005	0.629 ± 0.017	0.464 ± 0.026	0.510 ± 0.025	0.754 ± 0.050	0.741 ± 0.054
PILOT-GM	0.551 ± 0.039	0.564 ± 0.032	0.517 ± 0.035	0.553 ± 0.017	0.774 ± 0.012	0.720 ± 0.037
PaSCient	0.779 ± 0.013	0.825 ± 0.052	0.683 ± 0.034	0.706 ± 0.027	0.785 ± 0.049	0.762 ± 0.041
Linear floor and batch correction						
PCA	0.642	0.668	0.601	0.633	0.821	0.793
Harmony†	0.532	0.542	0.351	0.366	0.341	0.331
InfoGlobe†	0.412	0.532	0.238	0.327	0.466	0.589
Transductive						
MrVI (u)†	0.304 ± 0.009	0.316 ± 0.028	0.174 ± 0.007	0.240 ± 0.023	0.496 ± 0.044	0.488 ± 0.026
Foundation models (frozen)						
scGPT	0.681	0.683	0.654	0.703	0.723	0.701
Geneformer	0.778	0.823	0.612	0.656	0.759	0.759
Nicheformer	0.648	0.678	0.548	0.638	0.691	0.713
UCE	0.556	0.596	0.605	0.613	0.828	0.834
Donor-level baselines						
Composition only	0.213	0.222	0.225	0.258	0.543	0.538
Pseudobulk + PCA	0.756	0.715	0.569	0.637	**0.877**	0.834
CoState (Ours)	**0.833 ± 0.021**	**0.891 ± 0.049**	**0.839 ± 0.027**	**0.860 ± 0.029**	0.860 ± 0.020	**0.840 ± 0.021**

Test (donors): held-out donors of the training studies; Validation (cohort): the atlas' validation set of unseen donors; External (studies): donors from studies absent from the training pool. MrVI (u) is the sample-unaware latent space used in the main text. scANVI (disease) and PaSCient are supervised with disease labels (PaSCient: retrained on this atlas, disease cross-entropy as its only loss). Harmony is run on the PCA coordinates with the study as the batch. †Transductive: fit on the evaluation cells or given their batch labels.

G.2 PER-TYPE CROSS-STUDY TRANSFER

Figure 7 resolves cross-study alignment by cell type. These results use manually annotated test donors; the supplied labels on validation and external cohorts are transferred annotations rather than manual ground truth.

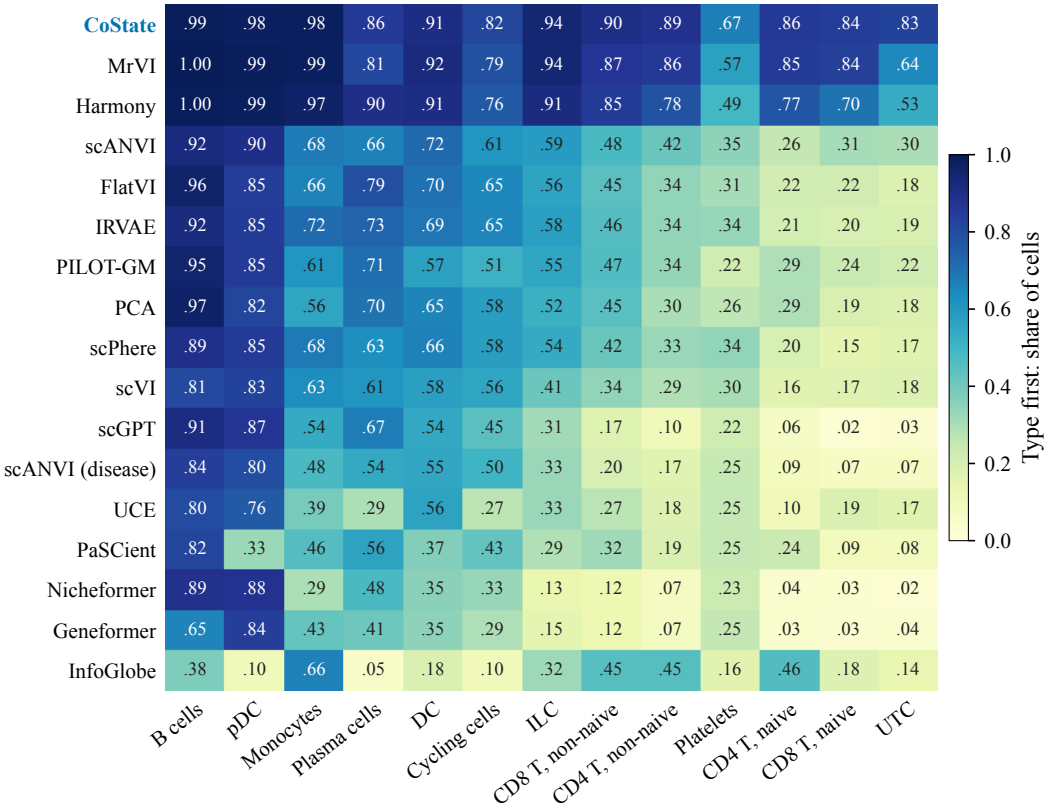

Figure 7: Full numerical version of Figure 2d. Each entry is the fraction of cells whose nearest same-type cell from another study is closer than their nearest other-type cell from their own study, using squared Euclidean distance. Higher values indicate stronger alignment by type under this criterion. Evaluation uses 250 cells per donor, three resamples, and 13 scored types, with means over training seeds where available. Row means give the Type first column in Table 2.

G.3 EFFECT OF THE NUMBER OF TRAINING CELLS

Figure 4 and Table 7 vary only the per-donor training-cell cap, retaining all 735 donors and the same training schedule. **CoState** uses $K = 8$, with two seeds per setting. Figure 8 checks the trend with prototype and linear readouts. The endpoint improvement is more consistent than intermediate changes, for which the two seeds can disagree. Increasing cell counts therefore supports better readout in this setting, without establishing a scaling law or a donor-count effect.

Table 7: Disease kNN macro-F1 against the size of the training pool (250 cells per half-bag). All 735 training donors are kept; only the per-donor cell cap changes (400, 800, 1600, 3200, none). Mean ± s.d. over two training seeds; Δ: largest minus smallest pool, computed before rounding. PaSCient is listed here but not drawn in Figure 4. **Bold**: best; underlined: second best. **CoState** is run with $K=8$ state codes in this experiment.

Training cells	0.29M	0.59M	1.15M	2.14M	3.50M	Δ
Test (held-out donors)						
CoState (Ours)	0.690 ± 0.023	0.753 ± 0.050	0.754 ± 0.070	0.841 ± 0.039	**0.876 ± 0.004**	**+0.186**
scVI	0.677 ± 0.035	0.677 ± 0.026	0.615 ± 0.015	0.610 ± 0.031	0.654 ± 0.062	−0.023
scANVI	0.398 ± 0.001	0.402 ± 0.008	0.398 ± 0.014	0.412 ± 0.003	0.408 ± 0.034	+0.011
scANVI (disease)	0.509 ± 0.003	0.515 ± 0.012	0.528 ± 0.007	0.531 ± 0.024	0.505 ± 0.017	−0.004
FlatVI	0.530 ± 0.005	0.524 ± 0.012	0.537 ± 0.004	0.540 ± 0.007	0.500 ± 0.015	−0.030
IRVAE	0.464 ± 0.012	0.475 ± 0.020	0.468 ± 0.017	0.467 ± 0.025	0.468 ± 0.021	+0.003
scPhere	0.624 ± 0.014	0.643 ± 0.025	0.607 ± 0.040	0.620 ± 0.004	0.625 ± 0.022	+0.001
PILOT-GM	0.572 ± 0.049	0.555 ± 0.025	0.566 ± 0.018	0.577 ± 0.001	0.553 ± 0.038	−0.019
PaSCient	**0.727 ± 0.018**	**0.880 ± 0.013**	**0.842 ± 0.029**	**0.878 ± 0.034**	0.851 ± 0.036	+0.124
Validation cohort						
CoState (Ours)	**0.717 ± 0.017**	**0.735 ± 0.031**	**0.753 ± 0.015**	**0.777 ± 0.002**	**0.826 ± 0.015**	**+0.108**
scVI	0.546 ± 0.074	0.577 ± 0.021	0.521 ± 0.031	0.519 ± 0.026	0.547 ± 0.013	+0.001
scANVI	0.394 ± 0.020	0.388 ± 0.014	0.368 ± 0.007	0.378 ± 0.025	0.381 ± 0.009	−0.013
scANVI (disease)	0.554 ± 0.030	0.570 ± 0.009	0.585 ± 0.023	0.590 ± 0.025	0.598 ± 0.017	+0.045
FlatVI	0.505 ± 0.001	0.503 ± 0.008	0.522 ± 0.029	0.484 ± 0.011	0.471 ± 0.005	−0.034
IRVAE	0.413 ± 0.038	0.414 ± 0.035	0.413 ± 0.034	0.409 ± 0.031	0.412 ± 0.030	−0.002
scPhere	0.490 ± 0.053	0.496 ± 0.013	0.501 ± 0.067	0.493 ± 0.024	0.496 ± 0.013	+0.006
PILOT-GM	0.548 ± 0.029	0.562 ± 0.037	0.546 ± 0.000	0.543 ± 0.016	0.555 ± 0.024	+0.007
PaSCient	0.692 ± 0.021	0.722 ± 0.032	0.707 ± 0.039	0.725 ± 0.014	0.720 ± 0.015	+0.028
External, independent studies						
CoState (Ours)	**0.786 ± 0.050**	**0.809 ± 0.003**	**0.819 ± 0.031**	**0.829 ± 0.002**	**0.840 ± 0.008**	**+0.054**
scVI	0.659 ± 0.027	0.666 ± 0.036	0.695 ± 0.046	0.635 ± 0.077	0.660 ± 0.047	+0.001
scANVI	0.581 ± 0.063	0.580 ± 0.070	0.588 ± 0.090	0.586 ± 0.090	0.571 ± 0.071	−0.010
scANVI (disease)	0.736 ± 0.002	0.709 ± 0.000	0.721 ± 0.023	0.734 ± 0.003	0.730 ± 0.005	−0.006
FlatVI	0.719 ± 0.000	0.706 ± 0.019	0.718 ± 0.012	0.726 ± 0.001	0.722 ± 0.008	+0.003
IRVAE	0.614 ± 0.036	0.600 ± 0.038	0.613 ± 0.035	0.603 ± 0.035	0.613 ± 0.021	−0.001
scPhere	0.732 ± 0.025	0.723 ± 0.000	0.711 ± 0.049	0.720 ± 0.016	0.716 ± 0.045	−0.016
PILOT-GM	0.754 ± 0.001	0.740 ± 0.016	0.724 ± 0.012	0.749 ± 0.021	0.724 ± 0.051	−0.031
PaSCient	0.724 ± 0.009	0.734 ± 0.001	0.739 ± 0.011	0.781 ± 0.021	0.738 ± 0.010	+0.014

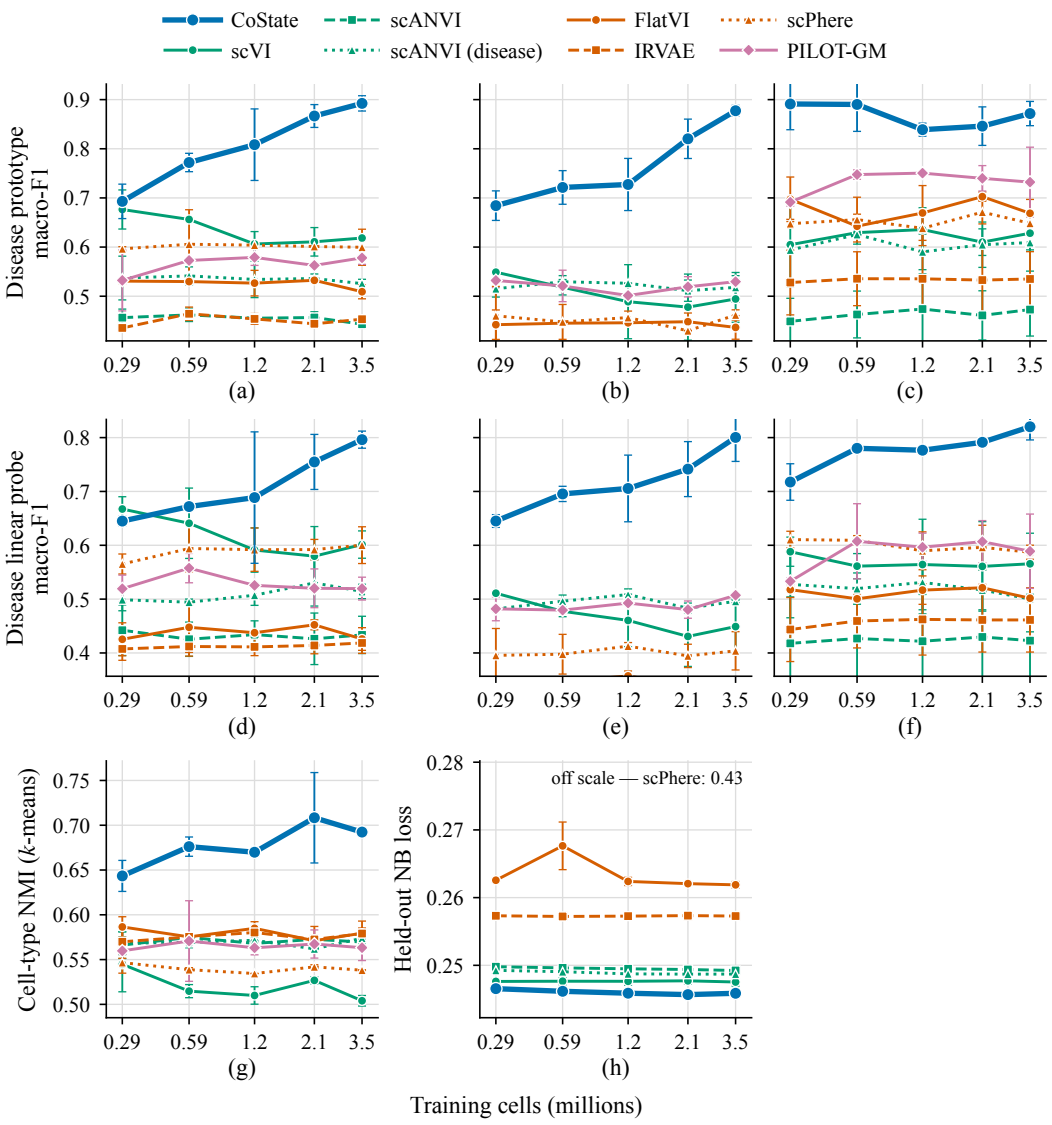

Figure 8: Further data-size readouts (250 cells per half-bag). (a)–(c) nearest class prototype and (d)–(f) linear probe, each on the test donors, the validation cohort and the external studies; (g) k-means NMI of the level-1 cell types on the test donors; (h) held-out negative-binomial loss at the end of training (lower is better). Means over 2 training seeds, bars ± 1 s.d. **CoState** is run with $K=8$ state codes in this experiment.

G.4 Composition-state decomposition

Figure 9 applies Proposition 1 to donor pairs from different studies. Different-disease pairs are farther apart than same-disease pairs in every cohort (mean D of 0.64 vs. 0.30 on the test donors, 0.67 vs. 0.25 on validation, and 0.40 vs. 0.29 on the external studies), and the state term forms 74–78% of the different-disease distance. For these eligible donor pairs, the state term accounts for approximately 85% of the mean distance increase externally and 91% on validation. On test, the state term increases by 0.350 while the composition term decreases by 0.006, giving a net increase of 0.344. These describe contributions to the learned distance, rather than causal disease effects.

Exploratory characterization of abundance and state. We examine observed cell-type proportions and expression profiles associated with selected learned state coordinates. Composition analy-

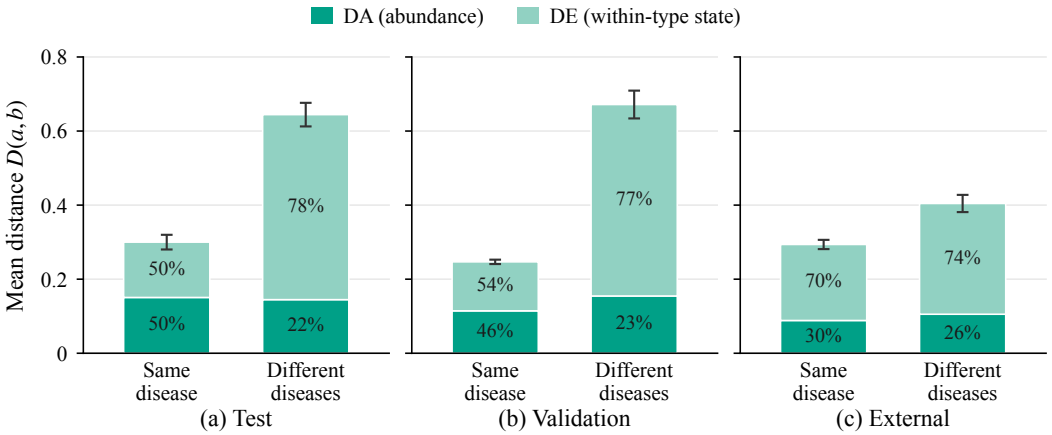

Figure 9: Composition-state decomposition of cross-study donor distances (Proposition 1). Bars show the mean **CoState** distance $D(a, b)$ over unordered pairs of donors from different studies, with the same or with different diseases, in (a) the test donors, (b) the validation cohort and (c) the external studies. Donors and disease classes follow cross-study retrieval (healthy/COVID-19 in (a, b), healthy/SLE in (c); 47, 75 and 19 donors; 253/494, 729/1224 and 49/63 same-/different-disease pairs). Each bar splits into DA, $H^2(\pi_a, \pi_b)$, and DE, $\sum_c \sqrt{\pi_{ac}\pi_{bc}}\, H^2(q_{ac}, q_{bc})$; percentages are each term's share of the mean. π and q are recomputed from 250 sampled cells per donor with the model frozen. Values average three resamples within each of three training seeds; error bars show the s.d. of the total over seeds (DA depends only on the sampled cells and has no seed variance).

sis compares 145 COVID-19 and 19 healthy donors from COMBAT2022, Ren2021, and SCGT06, including training donors, using observed cell counts. State characterization uses expression comparisons within held-out donors across disease categories, with the model frozen.

Observed cell-type proportions. Mean plasma-cell and cycling-cell fractions in the COVID-19 group are 6.5 and 4.4 times those in the healthy group, while the plasmacytoid dendritic-cell (pDC) fraction is 0.47-fold that of healthy donors. All three changes pass study-stratified tests with adjusted $p < 1.1 \times 10^{-4}$. The ratios describe pooled group means, and the fractions refer to retained annotated cells. Their directions are compatible with reported plasmablast and proliferating lymphocyte responses (Mathew et al., 2020) and reduced pDC frequencies (Arunachalam et al., 2020) in COVID-19, although cell definitions and sampling differ across analyses. Agreement with the source COMBAT study (Ahern et al., 2022) reflects reanalysis of overlapping data.

Interferon-associated expression in a learned monocyte state. Within-donor comparisons in 142 held-out donors across disease and healthy groups associate monocyte state 3 with higher *ISG15*, *IFI44L*, *IFIT1*, and *MX1* expression (all adjusted $p < 0.03$). The 20 top-ranked significant upregulated genes are enriched for Reactome Interferon Signaling (adjusted $p = 3.9 \times 10^{-6}$). This expression pattern suggests an interferon-associated annotation, compatible with programs reported in COVID-19 (Schulte-Schrepping et al., 2020).

Activation-associated expression in a learned CD4 T-cell state. In 48 held-out donors with sufficient cells for paired comparisons, non-naive CD4 T-cell state 2 shows higher *CD70* and *CD38* expression together with *CHEK1* and *STMN1* (all adjusted $p < 0.05$). These observations suggest an activation-associated annotation, with some cell-cycle-related expression. CD4 T-cell activation has also been reported in COVID-19 (Mathew et al., 2020). The comparison is concentrated in

COMBAT2022 and Ren2021 (47 of 48 donors). The marker pattern alone is insufficient to establish a proliferating state.

Illustrating cell-type contributions to an individual distance. A same-study donor pair illustrates how abundance and state differences enter an individual comparison. For healthy donor `Ren2021_PHC014_T0` and COVID-19 donor `Ren2021_PM054_T0`, the total distance is 0.300, comprising 0.038 from abundance and 0.262 from within-type states. The non-naive CD4 T-cell fraction changes from 19.5% to 17.6%, yet its state term contributes 0.059 to the distance and its abundance term contributes less than 0.001. The observed monocyte fractions are 14.6% and 31.5%, respectively, contributing through both abundance (0.016) and state (0.022). The decomposition quantifies contributions from observed proportions and learned state distributions under the chosen cell-type annotation for this donor pair.

Analysis protocol and scope. State characterization draws on 312 held-out donors: 82 test, 144 validation, 46 independent external, and 40 additional atlas donors (SCGT00val: 32; SCGT06: 8) (Jiménez-Gracia et al., 2026). The additional 40 enter neither **CoState** training nor benchmark scoring. The 142 monocyte and 48 CD4 T-cell donors are subsets meeting the state-comparison cell-count criteria. Composition tests permute donor labels within studies. State characterization compares cells assigned to a state with other cells of the same annotated type within donors, requiring at least 20 cells in each group. Gene tests use donor-level paired Wilcoxon tests with Benjamini–Hochberg correction within each state comparison. Reactome enrichment uses expressed panel genes as background and corrects over tested pathways. State assignment and characterization use the same expression modality; we therefore treat these results as exploratory expression-based annotations. Disease-specific state enrichment, functional activity, and contributions to retrieval performance require separate validation.

G.5 Roles of the representation blocks

Figure 10 evaluates the cell block (64 dimensions), patient block (128 dimensions), and residual block (64 dimensions) of **CoState** with $K = 16$ over three training seeds. Let b_i^ℓ denote cell i's vector in block $\ell \in \{\mathrm{cell}, \mathrm{pat}, \mathrm{res}\}$. Each diagnostic is computed separately for each block and training seed; bars and error bars report the mean and standard deviation across seeds.

(a) Cell-type share of variance. This analysis uses 81,170 manually annotated cells from 82 test donors, covering 13 studies and 15 level-1 cell types. Let μ^ℓ and μ_c^ℓ be the overall and type-c mean vectors, and n_c the number of cells of type c. The reported fraction is

$$V_\ell = \frac{\sum_c n_c \|\mu_c^\ell - \mu^\ell\|_2^2}{\sum_i \|b_i^\ell - \mu^\ell\|_2^2}. \tag{31}$$

The numerator measures variation explained by type means and the denominator measures total variation, both summed over block dimensions.

(b) Disease prediction from one block. This panel uses test donors only. In repeat r, each donor d is represented by the mean of one block over its 250-cell view $V_{d,r}$:

$$v_{d,r}^\ell = \frac{1}{250} \sum_{i \in V_{d,r}} b_i^\ell. \tag{32}$$

These vectors bypass the donor head. We apply the nearest-neighbour disease readout of Table 1 with disjoint reference and query donors. Disease macro-F1 is averaged over five cell resamples within each training seed before summarizing across seeds.

(c) Reconstruction sensitivity to block shuffling. We randomly shuffle one block among cells of the same type, keeping the other blocks and original count targets fixed. The decoder then evaluates negative-binomial reconstruction NLL on held-out cells. If $\mathcal{L}_0$ and $\mathcal{L}_\ell^{\mathrm{shuffle}}$ are the unperturbed and perturbed NLL, the relative increase is

$$\Delta_\ell = 100\% \times \frac{\mathcal{L}_\ell^{\mathrm{shuffle}} - \mathcal{L}_0}{\mathcal{L}_0}. \tag{33}$$

The reported unperturbed NLL is 0.246 nats per gene per cell. This measures reconstruction sensitivity to each block under within-cell-type shuffling.

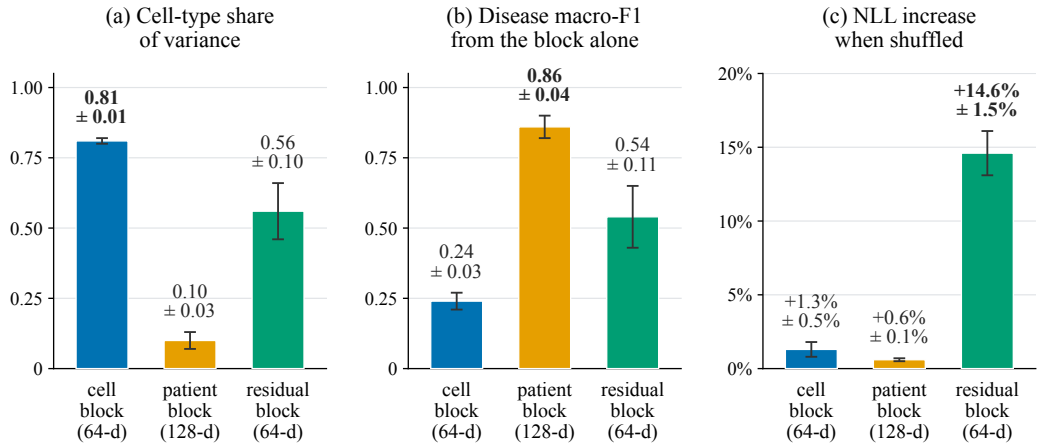

Figure 10: Block-wise diagnostics for **CoState** ($K = 16$) on test donors; bars and error bars show means and standard deviations over three training seeds (calculation details in Appendix G.5). (a) Fraction of total block variance, summed over dimensions, explained by cell-type means; (b) disease macro-F1 from each block alone after averaging 250-cell donor views, using the nearest-neighbour readout from Table 1 without the donor head and averaging five cell resamples; (c) relative increase in held-out negative-binomial reconstruction NLL when one block is shuffled within cell types and the other blocks remain fixed. Each block has the largest mean in a different panel; the residual block also retains cell-type information.