HyCell: a Compact Embedding Model for Single-Cell Transcriptomes

Hyra Team

Abstract

Pretrained single-cell models are used as frozen encoders that map the transcriptome of a cell to an embedding for annotation, integration and comparison across studies, and recent models have grown to hundreds of millions of parameters. We present HyCell, a single-cell embedding model with 16 million parameters trained on 79.5 million human cells from CELLxGENE Census. HyCell encodes the expressed genes of a cell as an unordered set with a six-layer transformer and reconstructs its counts through a negative-binomial decoder that also receives learned dataset and donor codes, while a donor discriminator on the embedding removes donor signal that remains; batch and donor information is used only in training. We compare frozen, zero-shot embeddings of HyCell with those of ten published single-cell foundation models, among them Stack, SCimilarity, scFoundation, UCE, scGPT, Geneformer and STATE, on seven atlases excluded from pretraining, with every model rerun on the same cells and scored by the same code. HyCell has the highest scIB biological-conservation score of the eleven models (0.700, against 0.699 for SCimilarity, which is trained with cell-type labels), and its overall score (0.735), which combines biological conservation, batch correction and annotation transfer, is level with that of Stack in its recommended batch-grouped mode (0.736), which uses batch labels at inference and has 14 times more parameters, and above those of the other nine models. In downstream analyses, HyCell embeddings separate naive from effector CD4 T cells across donors and sequencing chemistries (ARI 0.64, against at most 0.03 for every other embedding) and order B cells into plasma cells more faithfully than any other embedding, and the decoder reads out the corresponding gene programs consistently across donors.

1 Introduction

Single-cell foundation models are pretrained on tens to hundreds of millions of transcriptomes and then used as frozen encoders, whose embeddings of new cells serve to annotate cell types, integrate datasets and compare cell states across studies (Cui et al., 2024; Theodoris et al., 2023; Rosen et al., 2026; Hao et al., 2024). Recent models have grown to hundreds of millions or billions of parameters (Adduri et al., 2026; Gandhi et al., 2025; Pearce et al., 2026), yet independent evaluations find that their zero-shot embeddings often separate cell types no better than principal component analysis (PCA) of the counts (Kedzierska et al., 2025). A central difficulty is technical variation. Cells from different donors, laboratories and sequencing chemistries differ systematically. A reconstruction loss is lowered by encoding these differences, and the embedding then mixes cell identity with batch.

HyCell addresses this in the training objective rather than in model size. Its encoder maps the expressed genes of a cell to a 512-dimensional embedding z, and a count decoder reconstructs the full expression profile from z. The decoder also receives a learned code for the dataset and the donor of each cell, so dataset- and donor-specific variation can be explained without passing through z. A donor discriminator trained on z adds pressure to remove what remains. Its transformer has 4.8 million parameters, and the whole model, including a gene embedding table shared by encoder and decoder, has 16 million, fewer than any model we compare with and 42 times fewer than UCE.

We evaluate HyCell against ten published single-cell foundation models on seven atlases that were excluded from pretraining, rerunning every model with its released weights on the same cells and scoring all embeddings with the same code. Our main results are the following.

- **Highest biological conservation.** HyCell has the highest scIB biological-conservation score of the eleven models, 0.700 averaged over the seven atlases, level with SCimilarity (0.699), which is trained with cell-type labels, and at least 0.038 above every other self-supervised model (Figure 1).

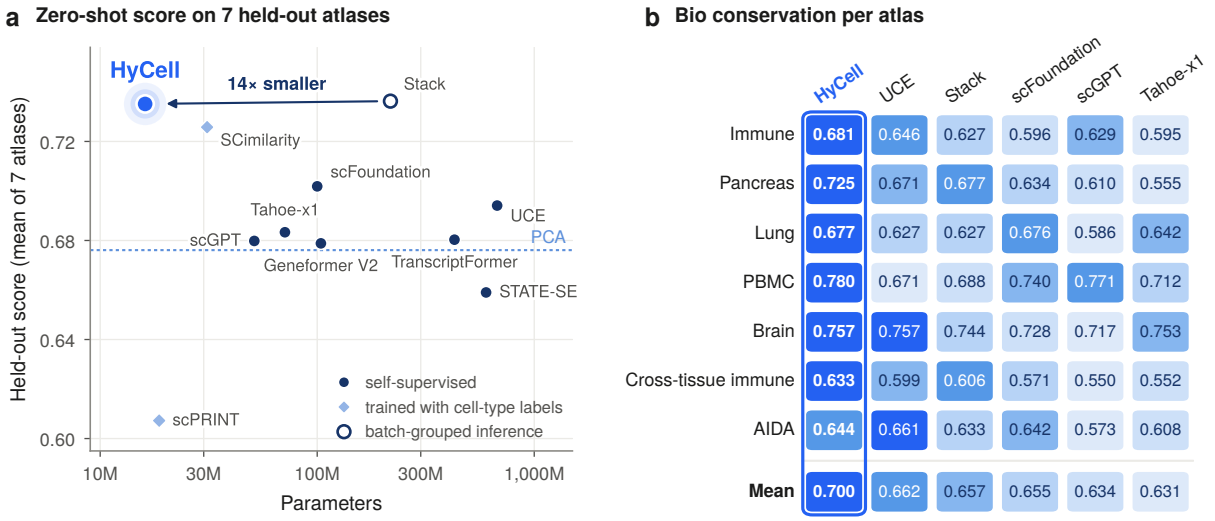


Figure 1: HyCell and ten published single-cell foundation models on seven held-out atlases. (a) Score S (Section 3.1) against the number of parameters, averaged over the seven atlases. Circles are self-supervised models, diamonds models trained with cell-type labels; the open circle is Stack in its recommended mode, which groups cells by batch at inference. The dotted line is PCA fitted to each atlas. (b) scIB biological conservation per atlas for the six self-supervised models with the highest mean; darker cells rank higher within a row.

- **Level with Stack at a fraction of the size.** On the combined score of biological conservation, batch correction and annotation transfer, HyCell (0.735) is level with Stack in its recommended batch-grouped mode (0.736), which uses batch labels at inference and has 14 times more parameters, and ahead of the nine other models (Table 3).
- **Downstream biology.** A two-group split of HyCell embeddings separates naive from effector CD4 T cells (ARI 0.64), where the other eleven embeddings reach at most 0.03 and mostly split by sequencing chemistry, and diffusion pseudotime orders B cells into plasma cells more faithfully than with any other embedding (Spearman $\rho = 0.94$). The count decoder reads out the gene programs of these states from the embeddings, consistently across donors (Section 4).

2 Model

HyCell is an encoder–decoder model trained to reconstruct the expression profile of a cell (Figure 2). The encoder turns the expressed genes of a cell into an embedding $z \in \mathbb{R}^{512}$, which is the representation used in all evaluations. The decoder and the donor discriminator are used only during training.

2.1 Cells as gene sets

The gene vocabulary has 36,495 genes, the 20,080 protein-coding genes together with every gene detected in at least 0.1% of corpus cells, and covers 99.9% of the nonzero entries of the corpus. A cell is represented as an unordered set of tokens with no positional encoding. Each expressed gene contributes a token equal to the gene's embedding plus a two-layer network applied to its $\log(1 + x)$-transformed expression per 10,000 counts. Up to 818 expressed genes enter the set; for cells that express more, genes are sampled without replacement with probability increasing in their expression. To each cell we add measured zeros, genes that the dataset measured but the cell did not express, in number equal to a quarter of the expressed genes. Zeros are drawn only from genes measured in at least 90% of training datasets, so that the gene panel of a dataset does not identify it. With a [CLS] token, a set has at most 1,024 tokens.

During training, two augmentations make the input noisier than the target. With probability 0.3 a random 15–100% subset of the genes is kept, and with probability 0.5 the counts are binomially thinned to 10–100% of their depth. The decoder always reconstructs the full, unaugmented profile. Embeddings for evaluation are computed without augmentation and with a fixed per-cell sampling seed.

2.2 Encoder

The encoder is a pre-normalisation transformer with RMSNorm, six layers of width 256, four attention heads and SwiGLU feed-forward blocks of width 704. Attention is bidirectional over sets of variable size packed without padding. The embedding z is a linear projection of the final [CLS] state to 512 dimensions.

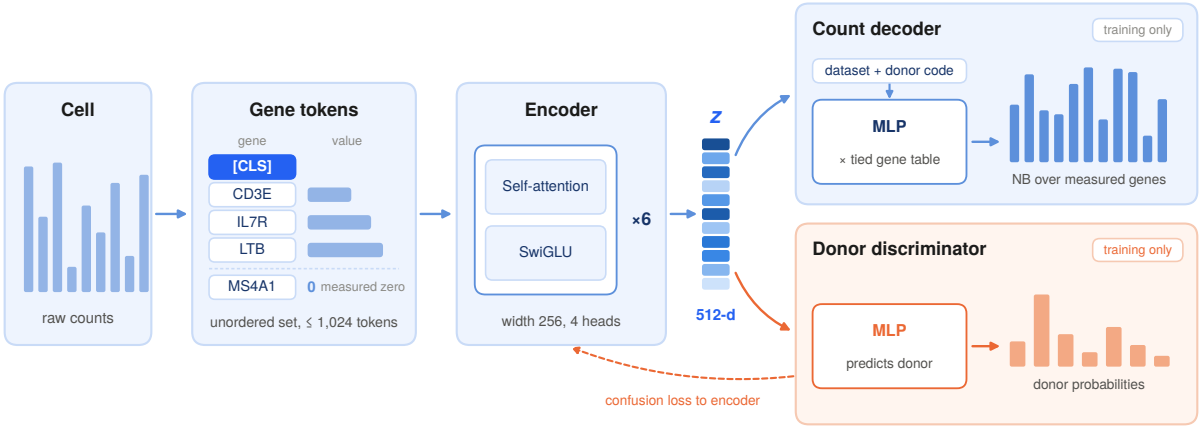


Figure 2: HyCell architecture. A cell is a set of gene tokens, each the sum of a learned gene embedding and an embedding of the gene's normalised expression, together with genes measured as zero and a [CLS] token. A transformer encoder of six attention and SwiGLU blocks maps the set to the embedding z. During training, a count decoder reconstructs the expression of every measured gene under a negative-binomial likelihood from z and a learned code for the cell's dataset and donor, and a donor discriminator predicts the donor from z, while a confusion loss (dashed) trains the encoder to make this prediction uninformative.

The transformer has 4.8 million parameters; with the gene embedding table, which the decoder shares, and the decoder itself, HyCell has 16.1 million (Table 1).

2.3 Count decoder with dataset and donor codes

The decoder maps z through a two-layer network ($512 \to 1{,}024 \to 256$, SiLU) to a vector h, and scores each gene g by $\ell_g = h \cdot t_g + b_g$, where t_g is the gene's input embedding and b_g a gene bias. The mean count of gene g is the library size times the softmax of ℓ over the genes measured in the cell's dataset, and counts follow a negative-binomial distribution with a learned dispersion per gene (Lopez et al., 2018). Genes the dataset did not measure do not enter the likelihood, so missing genes are never treated as zeros.

The decoder is conditioned on a 64-dimensional code $e = E_{\text{dataset}} + E_{\text{donor}}$, the sum of a learned embedding of the cell's dataset (555 entries) and of its donor within that dataset (12,533 entries). The code enters both decoder layers. Differences in depth, chemistry, ambient RNA and donor background that are shared by the cells of a dataset or a donor can be absorbed by e, and reconstructing them does not require z to encode them. A second loss term, with weight 0.1, reconstructs each cell from a stop-gradient copy of the decoder input with the code replaced by a learned unknown-donor or unknown-dataset code, so the decoder also works for cells from datasets outside the corpus.

2.4 Donor adversary

A discriminator, a two-layer network ($512 \to 256 \to 256$, LeakyReLU with slope 0.2, dropout 0.3), predicts the donor of each cell among the donors of its dataset from z. It is trained on detached embeddings with a cross-entropy loss. The encoder is trained, through the discriminator with its weights held fixed, to make the predicted donor distribution match the donor composition of the cell's dataset (Ganin et al., 2016). The weight of this term starts after 2×10^8 tokens and rises smoothly to 0.01. Only datasets with at least two donors contribute.

The training loss is the negative-binomial reconstruction loss, plus 0.1 times the reconstruction loss under the unknown code, plus the weighted donor-confusion loss.

2.5 Pretraining data and schedule

The corpus is the human primary-cell data with raw counts in the CZ CELLxGENE Census release of 8 November 2025 (CZI Cell Science Program et al., 2025), 95.0 million cells from 555 datasets. We remove every dataset and collection that contributes cells to an evaluation atlas (Section 3.1) or to a separate development set of eight tissues, which we used for all design choices, leaving 79.5 million training cells from 532 datasets and 20,000 validation cells. Datasets are sampled with probability proportional to the square root of their size, and cells uniformly within a dataset.

Table 1: HyCell at a glance.

Parameters	16.1M (transformer 4.8M, gene table 9.4M, decoder, codes and projection 1.9M)
Encoder	6 layers, width 256, 4 heads, SwiGLU 704, no positional encoding
Input	up to 818 expressed genes + measured zeros + [CLS]; 36,495-gene vocabulary
Embedding	512-d linear projection of [CLS]
Decoder	$512 \to 1{,}024 \to 256$, tied gene table, negative-binomial likelihood
Conditioning	64-d dataset + donor code (555 datasets, 12,533 donors), decoder only
Adversary	donor discriminator on z, confusion weight 0.01
Data	79.5M cells, 532 datasets, CELLxGENE Census 2025-11-08
Training	6.4×10^{10} tokens, 9.9 h on 2×H20

HyCell was trained for 246,197 steps of 262,144 tokens, 6.4×10^{10} tokens or 78.2 million cell presentations, about one pass over the training cells. We use AdamW ($\beta = (0.9, 0.95)$, weight decay 0.1, gradient clipping at 1.0) with a learning rate of 10^{-3}, a warm-up of 8×10^{7} tokens and a linear decay over the last 20% of training, in bfloat16 mixed precision. Training took 9.9 hours on two NVIDIA H20 GPUs.

3 Benchmark

3.1 Protocol

Atlases. We evaluate on seven atlases with raw integer counts and curated cell-type labels (Table 2). Three are the integer-count parts of the scIB integration benchmarks for human immune cells, pancreas and lung (Luecken et al., 2022). The others are peripheral blood mononuclear cells from two 10x Genomics runs (10x Genomics, 2017), the healthy donors of the MultiVINE atlas of the human brain vasculature, released in 2025 (Reid et al., 2025), a stratified 40,322-cell sample of the cross-tissue immune atlas of Domínguez Conde et al. (2022), and a 41,140-cell sample of the Asian Immune Diversity Atlas (Kock et al., 2025). Every dataset and collection that contributes cells to these atlases is excluded from the pretraining corpus.

Metrics. All embeddings are scored by one implementation, built on scib-metrics (Gayoso et al., 2026), on the absolute scale without rescaling across methods. Biological conservation (Bio) is the mean of seven scIB metrics: isolated-label score, NMI and ARI of k-means and of Leiden clusters against the cell types, cell-type silhouette and cLISI. Batch correction (Batch) is the mean of five: batch silhouette, iLISI, kBET, graph connectivity and principal-component regression. Annotation transfer holds out each batch in turn, trains a logistic-regression classifier (LR) and a 15-nearest-neighbour classifier (kNN) on the remaining batches, and reports macro-F1 averaged over held-out batches. The score $S = 0.3\,\text{Bio} + 0.2\,\text{Batch} + 0.25\,\text{LR} + 0.25\,\text{kNN}$ weights the scIB total ($0.6\,\text{Bio} + 0.4\,\text{Batch}$) and annotation transfer equally. Atlases are weighted equally in all averages.

Models. We compare with Stack (Dong et al., 2026), SCimilarity (Heimberg et al., 2025), scFoundation (Hao et al., 2024), UCE (Rosen et al., 2026), Tahoe-x1 (Gandhi et al., 2025), TranscriptFormer (Pearce et al., 2026), scGPT (Cui et al., 2024), Geneformer V2 (Chen et al., 2026), STATE (Adduri et al., 2026) and scPRINT (Kalfon et al., 2025), each run by us with its released weights and default inference settings on the same cells. Stack encodes sets of 256 cells, and its released tutorial orders cells by donor before encoding; we encode the cells of each batch together, so Stack uses batch labels at inference. SCimilarity and scPRINT are trained with cell-type labels. As a reference without pretraining we fit PCA to each atlas (2,000 highly variable genes of the log-normalised counts, 50 components). HyCell embeds each cell from its counts alone, without batch or donor information.

Table 2: Evaluation atlases. Batches are the unit held out in annotation transfer and the label used by the batch metrics.

Atlas	Tissue and technology	Cells	Batches	Cell types
Immune	bone marrow and blood, 5 studies	32,484	9	16
Pancreas	inDrop and Smart-seq2	10,963	5	14
Lung	Drop-seq	9,701	4	11
PBMC	blood, 10x Chromium	12,039	2	9
Brain	prefrontal cortex, vascular-enriched	21,155	10 donors	14
Cross-tissue immune	17 tissue sites, three 10x chemistries	40,322	12 donors	45
AIDA	blood, 5 sites	41,140	5 sites	32

Table 3: Zero-shot embeddings on seven held-out atlases, averaged over atlases. Params, parameters in millions; Cells, pretraining cells as reported by the authors (n/s, not stated). The best value of each column among the pretrained models (all rows except PCA) is in bold, and the value of HyCell is underlined where it ranks second. [b]Batch-grouped inference. [l]Trained with cell-type labels.

Model	Params	Cells	Score S	Bio	Batch	LR	kNN
HyCell	16	79.5M	<u>0.735</u>	**0.700**	0.527	<u>0.847</u>	<u>0.833</u>
Stack[b]	218	149M	**0.736**	0.657	**0.596**	0.841	**0.839**
SCimilarity[l]	31	7.9M	0.726	0.699	0.542	0.839	0.791
scFoundation	100	>50M	0.702	0.655	0.436	**0.849**	0.823
UCE	675	>36M	0.694	0.662	0.432	0.823	0.814
Tahoe-x1 (70M)	71	n/s	0.683	0.631	0.427	0.833	0.801
TranscriptFormer	429	57M	0.680	0.593	0.475	0.836	0.793
scGPT	51	33M	0.680	0.634	0.460	0.825	0.765
Geneformer V2	104	104M	0.679	0.606	0.465	0.834	0.782
STATE-SE	600	167M	0.659	0.596	0.461	0.842	0.711
scPRINT[l]	19	54M	0.607	0.519	0.434	0.781	0.678
PCA	–	–	0.676	0.654	0.443	0.827	0.738

Table 4: Score S on each held-out atlas. Best pretrained model per atlas in bold; HyCell underlined where it ranks second.

Model	Immune	Pancreas	Lung	PBMC	Brain	Cross-tissue	AIDA
HyCell	**0.735**	<u>0.717</u>	<u>0.686</u>	**0.871**	0.775	**0.697**	0.665
Stack[b]	0.714	**0.743**	**0.690**	0.838	**0.806**	0.689	0.673
SCimilarity[l]	0.729	0.677	0.684	0.848	0.781	0.680	**0.682**
scFoundation	0.669	0.659	0.673	0.840	0.768	0.645	0.660
UCE	0.715	0.668	0.639	0.793	0.758	0.630	0.655
Tahoe-x1 (70M)	0.674	0.594	0.657	0.820	0.770	0.630	0.638
TranscriptFormer	0.681	0.603	0.641	0.819	0.728	0.652	0.639
scGPT	0.689	0.610	0.647	0.834	0.756	0.624	0.599
Geneformer V2	0.683	0.579	0.653	0.816	0.756	0.634	0.630
STATE-SE	0.667	0.623	0.642	0.735	0.746	0.611	0.589
scPRINT[l]	0.642	0.540	0.609	0.772	0.661	0.493	0.532
PCA	0.676	0.710	0.659	0.785	0.779	0.535	0.588

3.2 Results

Table 3 summarises the seven atlases and Table 4 gives the score of every model on each.

Cell identity. HyCell has the highest biological-conservation score of all models, 0.700. SCimilarity, trained with cell-type labels, follows at 0.699, and the best self-supervised model, UCE, at 0.662. Among self-supervised models, HyCell ranks first on five of the seven atlases and shares first place on the brain atlas (Figure 1b).

Overall score. HyCell reaches $S = 0.735$, level with Stack (0.736) and ahead of SCimilarity (0.726), scFoundation (0.702) and UCE (0.694). Stack scores higher on batch correction (0.596 against 0.527), with sets of cells grouped by batch at inference; HyCell scores higher on biological conservation (0.700 against 0.657) with 14 times fewer parameters. Among models that embed each cell on its own, HyCell has the best batch-correction score after SCimilarity. In annotation transfer HyCell is second on both classifiers, by 0.002 behind scFoundation for LR and by 0.006 behind batch-grouped Stack for kNN, and ahead of all other models on both. These comparisons concern cell-level tasks. Stack also draws on the other cells of a sample at inference, which helps tasks defined per sample, such as predicting a donor's disease state or the perturbation applied to a well; such tasks are not part of this benchmark.

Size. The models within 0.03 of HyCell in S have 2 to 14 times more parameters, and the three largest models, UCE, STATE and TranscriptFormer with 429 to 675 million parameters, score between 0.659 and 0.694 (Figure 1a).

Embedding structure. Figure 3 shows the cross-tissue immune atlas, which spans 17 tissue sites, 12 donors and three sequencing chemistries. In PCA of the counts, T and NK cells form one mass partitioned

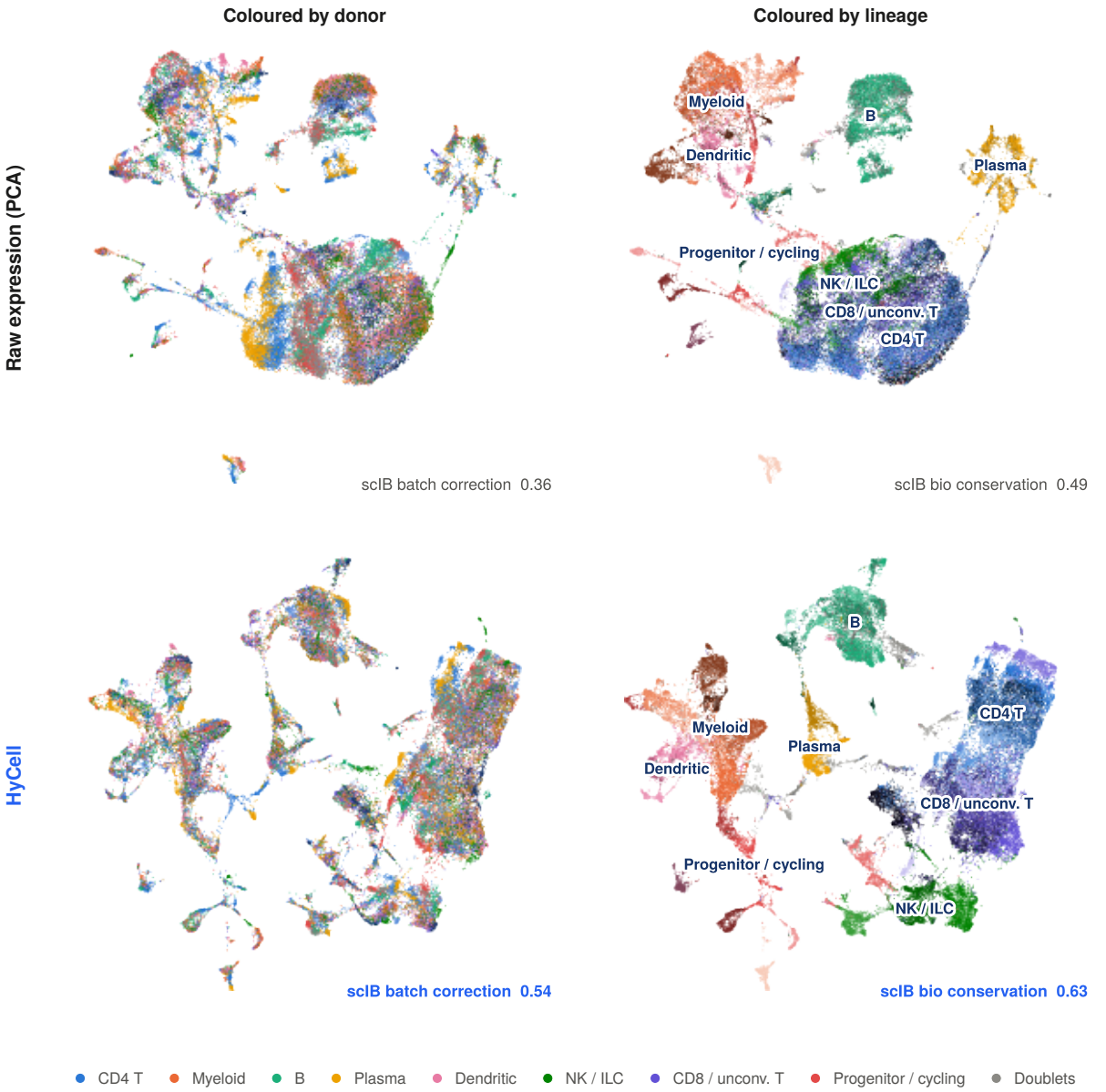


Figure 3: The cross-tissue immune atlas (40,322 cells, 12 donors, 17 tissue sites) embedded by PCA of the counts (top) and by HyCell (bottom), coloured by donor (left) and by lineage (right). UMAP (McInnes et al., 2018) of each representation with the same settings. Each panel gives the scIB score of the property its colouring shows.

by donor. In the HyCell embedding, cells of different donors overlap within each population, and CD4 T cells, CD8 and unconventional T cells, NK cells, B cells, plasma cells, myeloid cells and progenitors form separate regions. Batch correction rises from 0.36 to 0.54 and biological conservation from 0.49 to 0.63.

4 Downstream biological analysis

We apply the frozen embeddings to the cross-tissue immune atlas, where cells of the same state come from 12 donors, 17 tissue sites and three 10x chemistries. We separate naive from effector CD4 T cells, order B cells along their differentiation into plasma cells (Figure 4), and read out the gene programs of these states with the decoder (Figure 5).

4.1 Naive and effector CD4 T cells

We split the 2,778 CD4 T cells of the atlas, 1,343 naive or central-memory and 1,435 effector or effector-memory cells, into two groups by k-means on each embedding. With HyCell the two groups are 85% naive and 96% effector cells (ARI 0.64), and the split is nearly independent of sequencing chemistry (NMI with chemistry 0.012). With every other embedding, Stack and PCA included, the two groups do not

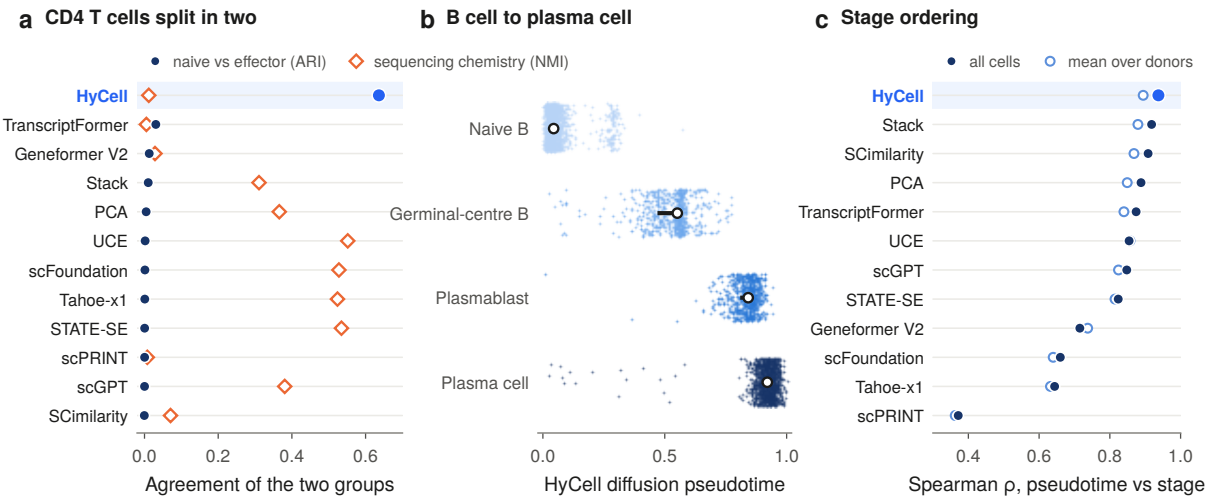


Figure 4: Differentiation states in the cross-tissue immune atlas. (a) CD4 T cells split into two groups by k-means on each embedding; agreement of the groups with the naive–effector annotation (ARI, filled) and with the sequencing chemistry (NMI, open). (b) HyCell diffusion pseudotime of naive B, germinal-centre B, plasmablast and plasma cells, rooted at naive-B marker-high cells; bars show the interquartile range and circles the median. (c) Spearman correlation between pseudotime and stage for every embedding, over all cells (filled) and averaged over the 12 donors (open).

match the naive–effector distinction (ARI at most 0.031); for seven of them they follow chemistry instead, with NMI between 0.31 and 0.55 (Figure 4a). Within single donors, where chemistry no longer varies, HyCell gives the highest agreement as well (mean ARI 0.66 over 12 donors; UCE 0.60, Tahoe-x1 0.58). The same holds for $\gamma\delta$ T cells against tissue-resident CD8 T cells (ARI 0.37 for HyCell, 0.30 for PCA, at most 0.06 for the published models).

The main axis of variation among HyCell embeddings of CD4 T cells is therefore their differentiation state, whereas in most other embeddings it is the sequencing chemistry.

4.2 From B cells to plasma cells

Naive B cells differentiate through germinal-centre B cells into antibody-secreting plasmablasts and plasma cells. For the 3,751 cells of these four states (1,440, 572, 618 and 1,121 cells), we computed diffusion pseudotime (Haghverdi et al., 2016) with Scanpy (Wolf et al., 2018) on each embedding, rooted at the medoid of the 50 cells with the highest naive-B marker score (*TCL1A*, *IGHD*, *FCER2*). In HyCell the median pseudotime rises from 0.04 for naive B cells through 0.55 and 0.84 to 0.92 for plasma cells (Figure 4b), and the Spearman correlation between pseudotime and stage, 0.937, is the highest of all embeddings. Averaged over donors it is 0.894, again the highest, ahead of Stack (0.918 and 0.879), SCimilarity (0.908 and 0.868) and PCA (0.888 and 0.849) (Figure 4c). Eight of the ten published models order the stages less faithfully than PCA of the counts.

4.3 Gene programs from the decoder

Because HyCell is trained to reconstruct counts, its decoder turns an embedding back into expected expression. Decoding the embedding of each cell with the unknown-dataset code and averaging within a state gives a gene-expression profile per state computed from the embeddings, and the ratio between two states gives a decoded fold change for every gene.

From naive B cells to plasma cells, the decoded fold changes recover the antibody-secretion program (Figure 5a). The genes decoded as most increased in plasma cells are *JCHAIN*, *SSR4*, *MZB1*, *XBP1*, *PRDX4* and *HSP90B1*, together with the endoplasmic-reticulum machinery *DERL3*, *FKBP11*, *SEC11C* and *HSPA5*; the most decreased are the HLA class II genes *HLA-DRA*, *HLA-DRB1*, *HLA-DPA1* and *HLA-DPB1*, which plasma cells lose with antigen presentation, and the lymphocyte genes *CD37*, *CD52*, *CXCR4* and *LTB*. Over the 12,879 genes detected in at least 1% of either state, decoded and counted fold changes have a Spearman correlation of 0.81 (Figure 5b).

The decoded programs are stable across donors. We compared the program of each donor with that of all other donors. For the plasma-cell axis the mean Spearman correlation is 0.965 for the decoder and 0.772 for pseudobulk counts; for classical against nonclassical monocytes it is 0.940 against 0.476, and for naive against effector CD4 T cells 0.957 against 0.376 (Figure 5c). The decoder thus provides a gene-level

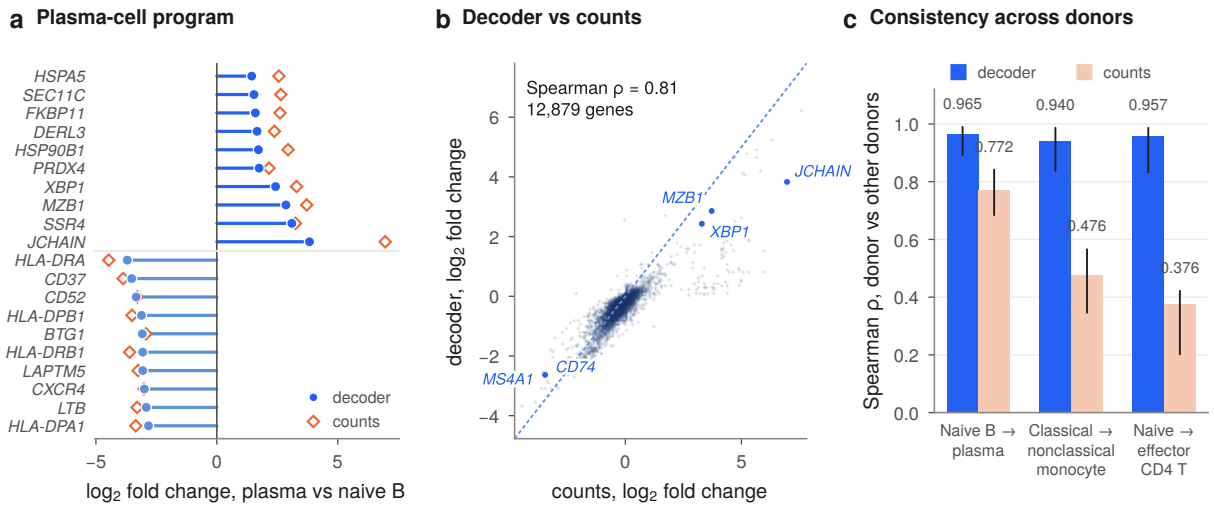


Figure 5: Gene programs read from the HyCell decoder. (a) The ten protein-coding genes with the largest decoded increase and decrease from naive B cells (1,440) to plasma cells (1,121), with the pseudobulk fold change from counts (open diamonds). (b) Decoded against counted $\log_2$ fold change for all genes detected in at least 1% of cells of either state. (c) Spearman correlation between each donor's program and the program of the other donors, for three pairs of states, from the decoder and from pseudobulk counts; bars show the mean and lines the range over donors.

readout of cell states that varies little between individuals, from the same embedding used for annotation and integration.

5 Fine-tuning for cell-type annotation

With a labelled reference, HyCell can be fine-tuned into a cell-type annotator. The encoder and a linear classification head on the embedding are trained together with cross-entropy, starting from the pretrained weights. Each step draws 64 labelled cells and tokenizes them with the pretraining augmentations (Section 2.1); the network is trained with a learning rate of 10^{-5} and the head with 10^{-3}, for at most 2,000 steps, with early stopping on 10% of the reference cells.

We evaluated this procedure by leave-one-batch-out annotation on four of the benchmark atlases. For every batch, the annotator is fine-tuned on the other batches and predicts the held-out batch. The baselines are the logistic-regression and kNN classifiers of the benchmark, trained on frozen HyCell embeddings of the same reference cells. On all four atlases the fine-tuned annotator has the highest mean macro-F1 (Table 5), and it is ahead of logistic regression on 19 of the 20 held-out batches and of kNN on 17. The largest gain, from 0.838 to 0.907, is on the pancreas atlas, whose batches come from different sequencing technologies. Fine-tuning takes one to three minutes on one NVIDIA H20 for 10,000 to 31,000 reference cells.

Table 5: Annotation of a held-out batch after fine-tuning. Macro-F1 over the cell types of the held-out batch, averaged over batches. Frozen, the benchmark classifiers trained on frozen HyCell embeddings of the same reference cells. Best value per atlas in bold.

Atlas	Batches	Fine-tuned	Frozen + LR	Frozen + kNN
Pancreas	5	**0.907**	0.832	0.838
Lung	4	**0.794**	0.768	0.758
PBMC	2	**0.966**	0.961	0.936
Immune	9	**0.877**	0.857	0.873

6 Availability

HyCell is distributed with the `hycell` Python package, which embeds cells from the raw counts of AnnData objects or `.h5ad` files, decodes embeddings into gene expression, and fine-tunes cell-type annotators, from Python and from the command line. On a GPU with flash-attention the package reproduces the embeddings scored in this report bit for bit. HyCell embeds every cell on its own and is trained to discount donor and dataset effects, which suits comparisons of cells across donors and studies; properties shared by all cells of a donor, such as sex or disease state, are better read from embeddings aggregated per donor or sample.

References

10x Genomics. PBMCs from a healthy donor, datasets pbmc4k and pbmc8k (Single Cell 3′ v2, Cell Ranger 2.1.0). https://cf.10xgenomics.com/samples/cell-exp/2.1.0/, 2017.

Abhinav K. Adduri et al. Predicting cellular responses to perturbation across diverse contexts with State. *Cell*, 189:5914–5931.e20, 2026. doi: 10.1016/j.cell.2026.07.052.

Han Chen et al. Scaling and quantization of large-scale foundation model enables resource-efficient predictions in network biology. *Nature Computational Science*, 6:450–463, 2026. doi: 10.1038/s43588-026-00972-4.

Haotian Cui et al. scGPT: toward building a foundation model for single-cell multi-omics using generative AI. *Nature Methods*, 21:1470–1480, 2024. doi: 10.1038/s41592-024-02201-0.

CZI Cell Science Program et al. CZ CELLxGENE Discover: a single-cell data platform for scalable exploration, analysis and modeling of aggregated data. *Nucleic Acids Research*, 53(D1):D886–D900, 2025. doi: 10.1093/nar/gkae1142.

C. Domínguez Conde et al. Cross-tissue immune cell analysis reveals tissue-specific features in humans. *Science*, 376:eabl5197, 2022. doi: 10.1126/science.abl5197.

Mingze Dong et al. Stack: in-context learning of single-cell biology. bioRxiv 2026.01.09.698608, 2026.

Shreshth Gandhi et al. Tahoe-x1: scaling perturbation-trained single-cell foundation models to 3 billion parameters. bioRxiv 2025.10.23.683759, 2025.

Yaroslav Ganin et al. Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 17(59):1–35, 2016. URL http://jmlr.org/papers/v17/15-239.html.

Adam Gayoso et al. scib-metrics: accelerated, Python-only, single-cell integration benchmarking metrics, version 0.5.9. Zenodo, 2026. URL https://github.com/YosefLab/scib-metrics.

Laleh Haghverdi et al. Diffusion pseudotime robustly reconstructs lineage branching. *Nature Methods*, 13:845–848, 2016. doi: 10.1038/nmeth.3971.

Minsheng Hao et al. Large-scale foundation model on single-cell transcriptomics. *Nature Methods*, 21:1481–1491, 2024. doi: 10.1038/s41592-024-02305-7.

Graham Heimberg et al. A cell atlas foundation model for scalable search of similar human cells. *Nature*, 638:1085–1094, 2025. doi: 10.1038/s41586-024-08411-y.

Jérémie Kalfon et al. scPRINT: pre-training on 50 million cells allows robust gene network predictions. *Nature Communications*, 16:3607, 2025. doi: 10.1038/s41467-025-58699-1.

Kasia Z. Kedzierska et al. Zero-shot evaluation reveals limitations of single-cell foundation models. *Genome Biology*, 26:101, 2025. doi: 10.1186/s13059-025-03574-x.

Kian Hong Kock et al. Asian diversity in human immune cells. *Cell*, 188:2288–2306.e24, 2025. doi: 10.1016/j.cell.2025.02.017.

Romain Lopez et al. Deep generative modeling for single-cell transcriptomics. *Nature Methods*, 15:1053–1058, 2018. doi: 10.1038/s41592-018-0229-2.

Malte D. Luecken et al. Benchmarking atlas-level data integration in single-cell genomics. *Nature Methods*, 19:41–50, 2022. doi: 10.1038/s41592-021-01336-8.

Leland McInnes, John Healy, and James Melville. UMAP: uniform manifold approximation and projection for dimension reduction. arXiv:1802.03426, 2018. URL https://arxiv.org/abs/1802.03426.

James D. Pearce et al. TranscriptFormer: a generative cell atlas across 1.5 billion years of evolution. *Science*, 393:eaec8514, 2026. doi: 10.1126/science.aec8514.

Madigan M. Reid et al. Human brain vascular multi-omics elucidates disease-risk associations. *Neuron*, 113:3143–3161.e5, 2025. doi: 10.1016/j.neuron.2025.07.001.

Yanay Rosen et al. Universal cell embedding provides a foundation model for cell biology. *Nature*, 656:183–191, 2026. doi: 10.1038/s41586-026-10689-z.

Christina V. Theodoris et al. Transfer learning enables predictions in network biology. *Nature*, 618:616–624, 2023. doi: 10.1038/s41586-023-06139-9.

F. Alexander Wolf, Philipp Angerer, and Fabian J. Theis. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biology*, 19:15, 2018. doi: 10.1186/s13059-017-1382-0.