

A vision foundation model for single-cell biology via spatial gene cartography

Ridvan Yesiloglu¹, Sakib Mostafa², James Zou^{1,3,4}, Ash Alizadeh⁵, Jiajun Wu⁴, Lei Xing^{1,2,6}, Ehsan Adeli^{3,4,7,*}, and Md Tauhidul Islam^{2,*}

¹Department of Electrical Engineering, Stanford University, Stanford, California, USA

²Department of Radiation Oncology, Stanford University, Stanford, California, USA

³Department of Biomedical Data Science, Stanford University School of Medicine, Stanford, California, USA

⁴Department of Computer Science, Stanford University, Stanford, California, USA

⁵Department of Medicine, Division of Oncology, Stanford University, Stanford, California, USA

⁶Institute of Computational and Mathematical Engineering, Stanford University, Stanford, California, USA

⁷Department of Psychiatry and Behavioral Sciences, Stanford University, Stanford, California, USA

*correspondence: tauhid@stanford.edu and eadeli@stanford.edu

ABSTRACT

Single-cell transcriptomics has made it possible to measure gene expression in tens of millions of individual cells, revealing cellular diversity that bulk profiling cannot resolve. Foundation models aim to learn general representations from these large datasets that can be reused across many biological tasks. However, most current single-cell foundation models are adapted from language models and represent each cell as a set or sequence of gene tokens. This design has two limitations. It treats genes as largely unordered inputs, even though genes act together in coordinated programs, and it often requires expression values to be discretized or ranked, losing quantitative information about expression magnitude. Here we present scVision, a vision foundation model for single-cell biology. Instead of converting genes into tokens, scVision represents each cell as a continuous gene-expression image. It assigns genes to fixed spatial positions using optimal transport, so that genes with related expression patterns are placed near one another and coordinated gene programs form local image regions. The resulting image preserves both the quantitative expression level of each gene and the biological relationships among genes. We pretrain a vision transformer with masked image modelling on 72 million human cells, creating one of the largest pretrained models for single-cell analysis. In zero-shot evaluations across six independent, held-out studies, frozen scVision representations outperform existing foundation models and classical baselines in cell-type annotation and gene-program discovery, without task-specific retraining. On multi-study integration, scVision matches the strongest token-based foundation model on the combined benchmark score and conserves more biological structure than any method tested. The spatial organization of scVision also improves interpretability: image regions correspond to groups of co-expressed genes, and attention maps can be read as gene-program activity. This structure also enables spatial masking experiments, in which a neighborhood of related genes is perturbed as a single unit, an operation with no direct counterpart in token-based foundation models. By preserving continuous gene-expression values and giving genes biologically meaningful positions, scVision reframes single-cell representation learning as a vision problem. This approach retains more of the original transcriptomic signal while opening a direct path for applying modern computer vision methods to single-cell biology.

Introduction

Single-cell RNA sequencing (scRNA-seq) has changed the scale at which biology can be studied. Instead of measuring average gene expression across mixed tissues, it now allows individual cells to be profiled across organs, developmental stages, disease states, treatments and patient populations^{1–6}. This shift makes it possible to examine biology at the level where important events first arise. Rare immune populations expand within a tissue. Progenitor cells move through intermediate states toward differentiation. Drug-resistant malignant states emerge before they are detectable clinically⁷. Large cellular atlases built using scRNA-seq serve as reference maps of human biology^{8,9}. They can define the molecular identity of cell types across tissues, reveal rare or previously unrecognized cell populations, reconstruct developmental trajectories, and map how cells change during ageing, infection, inflammation, cancer and treatment response^{10–12}. In disease, such atlases can uncover pathogenic immune and stromal states, identify cell-type-specific gene programs, distinguish primary disease mechanisms from secondary tissue responses, and show how abnormal cells interact with their microenvironment^{13–16}. In translational settings, they provide a framework for comparing patient samples with healthy references, discovering biomarkers, prioritizing therapeutic targets, and

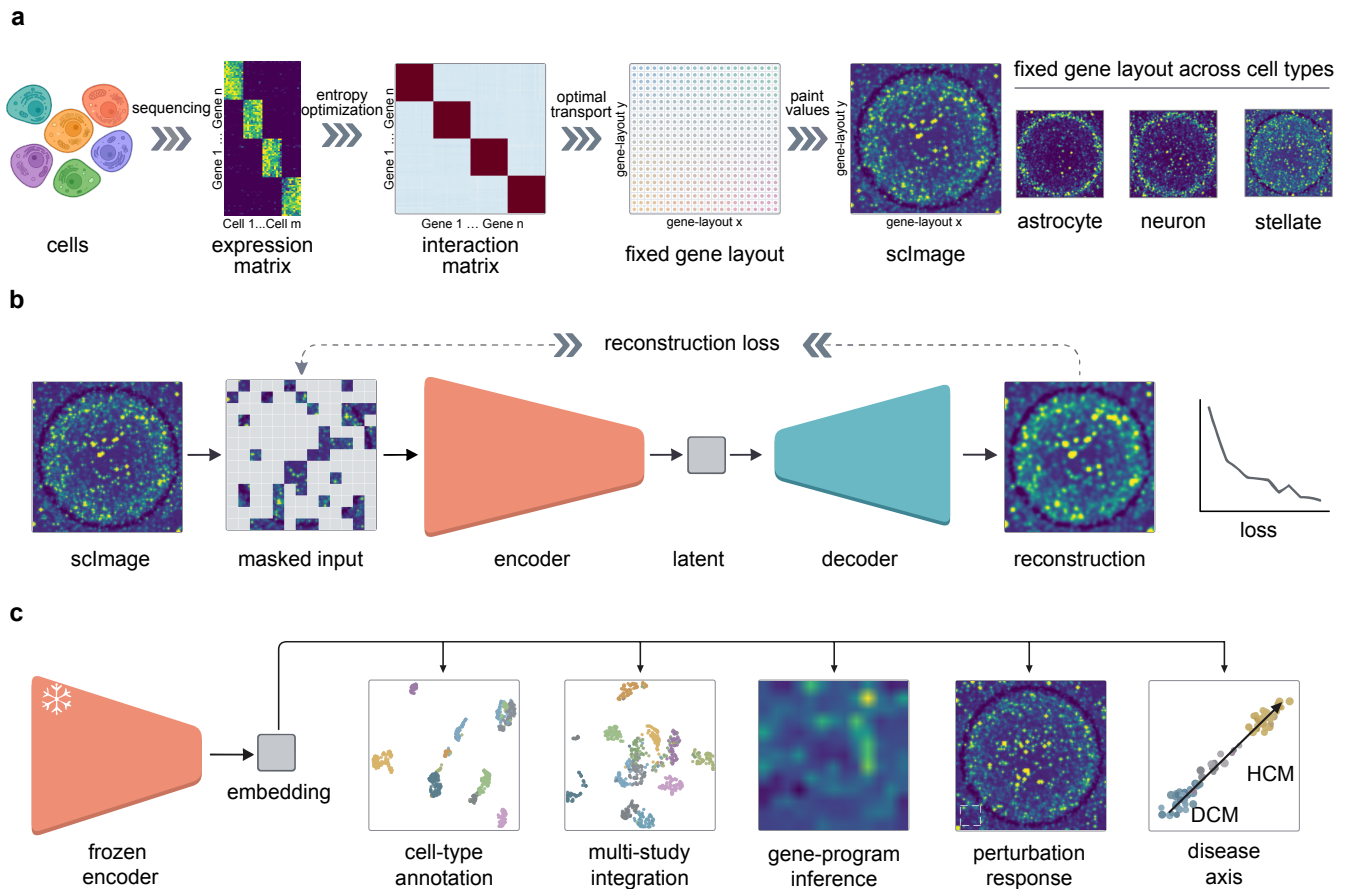


Figure 1. scVision renders each cell as an image and learns one frozen representation by masked image modelling. a, Construction of the scImage. The 10,816 most informative genes define a gene-gene co-expression matrix, and Gromov–Wasserstein optimal transport places the genes on a fixed 104×104 lattice so that co-expressed genes become spatial neighbours. The layout is computed once, pan-tissue, and shared across all cells; painting a single cell’s expression onto it renders the transcriptome as one image, in which co-regulation appears as local texture and cellular identity as global pattern. **b,** Pretraining. A ViT-base encoder (approximately 86 million parameters) and a lightweight decoder are trained by masked image modelling, reconstructing the 75% of scImage patches hidden at the input, on 72 million of a roughly 94 million human-cell dataset. **c,** From the single frozen encoder, a cell embedding supports, with no fine-tuning, cell-type annotation, multi-study integration, gene-program inference read from attention, perturbation response probed by masking a gene neighbourhood (64 co-regulated genes at once, one 8×8 patch), and a disease axis that transfers across assay and subtype, from dilated to hypertrophic cardiomyopathy.

understanding why only some cells or patients respond to a given intervention¹⁷. As these atlases grow, they offer the possibility of building computational models that learn general principles of cellular organization from millions of cells and then apply that knowledge to new biological problems. Such foundation models could help annotate newly generated datasets, harmonize studies collected across different laboratories and technologies, detect rare or transitional cell states, identify disease-associated programs in small patient cohorts, and predict how cells respond to genetic perturbations, drugs or environmental signals.

Most current single-cell foundation models borrow their design from natural-language processing. In these models, a cell is treated like a sentence and genes are treated like words. Models such as scBERT, Geneformer, scGPT, scFoundation and Universal Cell Embeddings are trained by masking or predicting gene tokens across large collections of cells^{18–22}. These approaches have shown that self-supervised learning can be scaled to single-cell data and have established foundation modelling as an important direction for the field. However, independent evaluations have also raised important questions. In common benchmarks such as cell-type classification, several token-based models do not consistently outperform simple regularized linear models; in zero-shot settings, their embeddings can underperform standard dimensionality-reduction methods for clustering or integration; and across multi-task evaluations, they do not always exceed task-specific tools^{23–25}. A key limitation of these token based models lies in how the transcriptome is represented. These models make two simplifying assumptions. First, they

usually do not give genes a biologically meaningful arrangement. A transcriptome is represented as a set or sequence of gene tokens, but genes that participate in the same pathway, regulon or co-expression program are not necessarily placed near one another. Second, many token-based approaches require expression values to be ranked, discretized or otherwise converted into tokens, which can reduce the quantitative information contained in continuous gene-expression measurements. Yet cellular identity is shaped by both of these features: genes act together in coordinated programs, and the strength of those programs depends on expression magnitude. A representation that removes gene-gene structure and compresses continuous expression values may therefore fail to learn the underlying biological relationships and lead to suboptimal results.

Here we introduce scVision, a vision foundation model for single-cell biology that preserves both gene relationships and expression magnitude. Instead of representing a cell as a sequence of tokens, scVision represents each cell as a continuous-valued image. We first assign informative genes to fixed positions on a two-dimensional lattice using optimal transport²⁶, so that genes with similar expression patterns across cells are placed near one another. Each cell's expression profile is then painted onto this shared layout, producing an scImage in which each pixel corresponds to one gene and pixel intensity corresponds to its expression level. In this representation, co-regulated genes form local image regions, gene-program activity appears as spatial texture, and global cellular identity appears as an image-level pattern. This formulation turns single-cell representation learning into a computer-vision problem²⁷. It enables the use of local feature extraction, spatial attention and masked image modelling while retaining the continuous nature of gene expression. We pretrain scVision, a vision transformer with approximately 86 million parameters, as a masked autoencoder on 72 million human cells from a dataset of approximately 94 million cells assembled from public atlases^{6,8,9}. During pretraining, the model learns to reconstruct masked regions of the scImage, encouraging it to capture both local gene-program structure and global cellular state.

We evaluate scVision in a zero-shot setting, where cells from held-out studies are encoded using the frozen pretrained backbone, without fine-tuning, and downstream tasks are performed using the learned representation. Under this cross-study evaluation, scVision provides the strongest cell-type annotation of the methods tested and identifies gene-expression programs without pathway supervision. It also matches the strongest existing methods on multi-study integration and enables spatial masking of gene neighborhoods, an operation with no direct counterpart in token-based models²⁸⁻³³. Because the same spatial gene layout is used across all cells, the representation also supports direct biological interpretation. Contiguous image regions correspond to groups of co-expressed genes, attention maps highlight active gene programs, and spatial masking allows coordinated gene neighborhoods to be perturbed together in silico.

Representation may be a central bottleneck in single-cell foundation modelling. By giving genes biologically meaningful positions while preserving continuous expression values, scVision offers an alternative to gene tokenization and connects single-cell biology to modern computer vision²⁷. This spatial view of the transcriptome enables both transferable representations and interpretable analyses of gene programs, opening a path toward vision foundation models for cellular biology^{19-21,29}.

Results

Label-efficient zero-shot annotation of held-out kidney cortex

We first tested scVision on an external kidney cortex dataset, where the model never sees kidney cells during training. The task is to assign cortical cell types using a frozen embedding and a reference bank containing 87 cell types from other studies. In two-dimensional projections, all four foundation models separated the kidney cortex into coherent lineage structure, but the scVision embedding was the most identifiable: annotating each panel with its matched 20-nearest-neighbour probe balanced accuracy, the common geometry metric we report for every held-out atlas (Fig. 8), ranked scVision first, ahead of scFoundation, scGPT and Geneformer (Fig. 2a). Using a nearest-neighbour probe on the frozen scVision embedding, scVision achieved the highest performance on the external cortex, with 0.47 balanced accuracy and 0.56 present-class macro-F1 (Fig. 2b). These metrics emphasize performance across all cell types, including rare cortical populations. scVision outperformed the next-best model in balanced accuracy, scFoundation²¹ (0.40), and the next-best method in macro-F1, logistic regression (0.50). scGPT²⁰ reached 0.38 balanced accuracy, whereas Geneformer¹⁹ reached 0.15. We report balanced accuracy and present-class macro-F1 because these metrics better reflect annotation performance on rare but biologically important cell types than threshold-free area-under-the-curve scores, which can be dominated by abundant classes.

We then inquired how many labelled examples are needed to annotate a new cell type. For each method, we sampled $k \in \{1, 5, 10, 50\}$ labelled cells per type from the reference bank and evaluated performance on the full external cortex, using the same support cells for all models (Fig. 2c). With only one labelled cell per type, scVision reached 0.27 balanced accuracy. With ten labelled cells per type, it reached 0.34, which is 72% of its fully supervised performance of 0.47. Performance then changed little with additional labels. Under the same nearest-neighbour readout, scVision performed best at every label budget. Remarkably, one labelled cell per type with scVision outperformed fifty labelled cells per type with scGPT, scFoundation and Geneformer (0.27 versus 0.24, 0.23 and 0.04, respectively). This shows that scVision reduces the required labelling effort by more than fifty-fold in this setting.

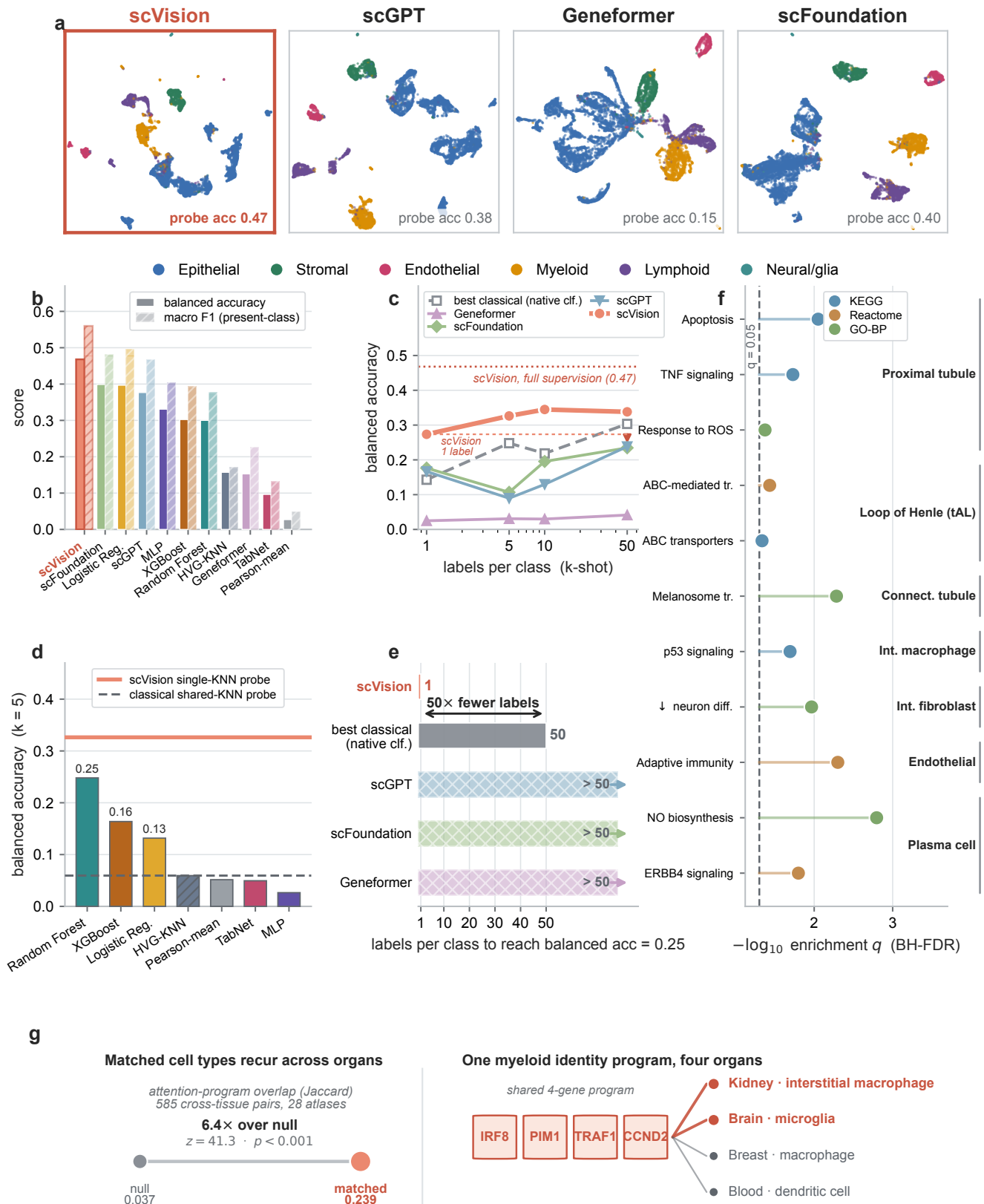


Figure 2. A spatial representation annotates the held-out kidney cortex most accurately, learns to do so from a single labelled cell, and recovers nephron biology without supervision.

Figure 2. Held-out kidney cortex under the cross-study holdout (paradigm A); every method reads a frozen representation and is scored on the full external test pool against an 87-type reference bank. **a**, Per-method embedding atlas: UMAP of the frozen representation for scVision (coral border) and three single-cell foundation models, computed with one matched routine and coloured by lineage (twenty-six native cortical cell types grouped into six lineages). Each panel is annotated with the embedding's matched 20-nearest-neighbour probe balanced accuracy, the common cross-atlas identifiability metric (Fig. 8): scVision 0.47, scFoundation 0.40, scGPT 0.38, Geneformer 0.15. scVision yields the most label-transferable cortical geometry, consistent with the matched probe (**b**). **b**, Annotation accuracy under a matched full-bank 20-nearest-neighbour probe: balanced accuracy (solid) and present-class macro-F1 (hatched) for scVision, three single-cell foundation models and seven classical and tabular baselines. scVision (coral) leads both metrics; we report these present-class metrics, which reward correct calls on the minority cortical types, rather than threshold-free areas under the curve, which are dominated by the abundant classes. **c**, Label efficiency: balanced accuracy versus labelled cells per type (k -shot, log axis). The four foundation models share the same 20-nearest-neighbour probe, isolating representation quality; scVision leads at every budget and, with one labelled cell per type, already exceeds every other foundation model given fifty (horizontal guide). Dotted line, scVision's fully supervised balanced accuracy; grey dashed curve, the best of seven classical baselines each fitting its own native classifier. **d**, Representation versus classifier at $k = 5$: each classical baseline fits its own best classifier on the identical support (bars); the dashed line is the shared nearest-neighbour probe (identical across classical methods, which share the raw highly-variable-gene space), the coral line scVision's single probe. The hatched bar (k -nearest neighbours) coincides with the shared probe by construction. **e**, The same efficiency read as an annotation budget: labelled cells per type needed to reach a balanced accuracy of 0.25. scVision needs one, the best steelmanned classical baseline needs fifty ($50\times$ fewer), and the other foundation models do not reach the target within fifty (hatched, off-axis). **f**, scVision's attention recovers nephron-segment biology: per-cell-type top-attention gene programs, mapped to genes through the scImage, tested for pathway enrichment (hypergeometric, Benjamini–Hochberg FDR) against KEGG, Reactome and GO biological process, with no pathway supervision. Points mark significant cell-type→pathway hits by $-\log_{10} q$; dashed line, $q = 0.05$; cell types grouped at right. **g**, These attention programs are not kidney-private. Across 585 cross-tissue cell-type pairs spanning twenty-eight atlases, attention-program overlap (top-attended-gene Jaccard) for matched cell types exceeds a label-permuted null 6.4-fold (mean 0.239 versus 0.037; $z = 41.3$, $p < 0.001$; left). At right, a single four-gene myeloid program (IRF8, PIM1, TRAF1, CCND2) supplies the top-attended patch for the kidney interstitial macrophage, brain microglia, breast macrophage and autoimmune-blood dendritic cell: one model, one program, four organs.

To test whether this advantage came only from the nearest-neighbour classifier, we also gave each classical baseline its own preferred classifier. These included logistic regression, a multilayer perceptron, gradient-boosted trees, a random forest, k -nearest neighbours, a correlation classifier and TabNet, all trained on the same support cells (Fig. 2d). The k -nearest-neighbour baseline reproduced the shared-probe result, confirming that the comparison isolated the effect of the classifier. Allowing each method to use its own classifier improved the strongest classical baselines, but none matched scVision. At five labelled cells per type, the best classical method reached 0.25 balanced accuracy, compared with 0.33 for scVision. scVision remained best at all label budgets, with the gap narrowing only at fifty labelled cells per type (0.34 versus 0.30). In practical terms, scVision needed only one labelled cell per type to reach 0.25 balanced accuracy, whereas the best classical method needed fifty; the other foundation models did not reach this level within fifty labels (Fig. 2e). The label efficiency therefore reflects the learned spatial representation rather than a particular classifier.

Finally, we examined whether the same frozen representation also captured interpretable kidney biology. We read scVision's last-block attention as gene-expression programs by mapping the most-attended image patches back to genes through the scImage. This analysis used no pathway supervision. The resulting programs matched known nephron biology (Fig. 2f). Proximal-tubule epithelial cells, which are highly susceptible to injury, were enriched for Apoptosis and TNF signalling (KEGG $q = 0.009$ and 0.019) and for the reactive-oxygen-species response (GO biological process $q = 0.042$), consistent with canonical tubular-injury programs. The thin ascending limb of the loop of Henle, which is involved in solute handling, was enriched for ABC transporters (Reactome $q = 0.037$). Interstitial macrophages were enriched for p53 signalling ($q = 0.020$).

These programs were not specific to kidney tissue alone. The same four-gene myeloid program, IRF8, PIM1, TRAF1 and CCND2, formed the top-attended patch in both kidney interstitial macrophages and brain microglia. The model recovered a shared myeloid program across two organs without pathway labels. Across 585 matched cell-type pairs from 28 atlases, attention-derived programs from the same cell type overlapped 6.4-fold more than expected under a label-permuted null model (mean Jaccard 0.239 versus 0.037; $z = 41.3$; Fig. 2g). These results suggest that scVision learns reusable cell-type identity programs that can be recognized across tissues.

A pericyte stress program in held-out ovary

We next tested scVision on an external ovary dataset. This was a smaller transfer test than the kidney cortex analysis. After restricting the dataset to cell types present in the 87-type reference bank, four ovarian populations remained: endothelial cells, leukocytes, pericytes and ovarian stromal cells. The model had not seen ovarian cells during training, so it had to assign these cell types using only the frozen embedding and the reference bank. Chance performance for this four-class task is 0.25.

Using a nearest-neighbour probe on the frozen scVision embedding, scVision achieved the best performance, with 0.83 balanced accuracy and 0.89 present-class macro-F1 (Fig. 3b). scFoundation²¹ and scGPT²⁰ were the closest foundation-model baselines, with balanced accuracies of 0.79 and 0.78, respectively. Across the five matched folds, scVision's improvement over scFoundation was significant (paired t -test, $p < 0.001$), as was its improvement over scGPT ($p < 0.01$). Logistic regression was the strongest classical baseline, reaching 0.70 balanced accuracy, while Geneformer¹⁹ reached 0.59. We focus on balanced accuracy and present-class macro-F1 because they measure performance across all four cell types. In contrast, ROC AUC was above 0.98 for every method and therefore did not distinguish the models in this coarse four-class setting.

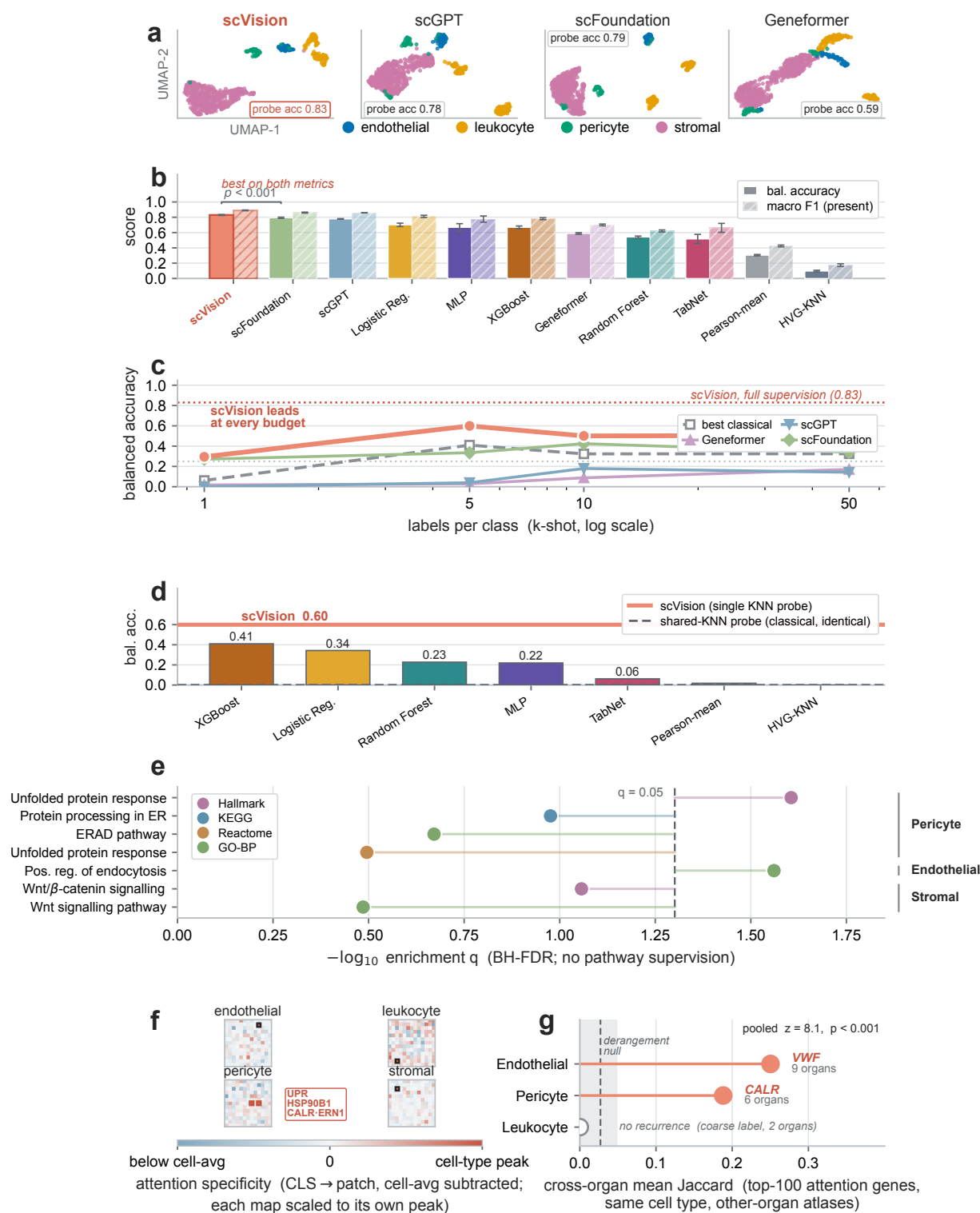


Figure 3. A spatial representation annotates the held-out ovary most accurately, leads at every labelling budget, and recovers a coherent pericyte stress program without supervision. Held-out ovary, cross-study holdout (paradigm A); methods read a frozen representation scored against the 87-type bank (present pool: endothelial, leukocyte, pericyte, stromal; four-class chance 0.25). Panels a–g are described in the continued caption overleaf.

Figure 3. a. Embedding geometry: UMAP of each frozen representation coloured by reference type, annotated with the matched 20-nearest-neighbour probe balanced accuracy, the common cross-atlas identifiability metric (Fig. 8). scVision is the most identifiable embedding (0.83), ahead of scFoundation (0.79), scGPT (0.78) and Geneformer (0.59); this ordering reproduces the accuracy ordering in (b). **b.** Matched full-bank 20-NN probe: balanced accuracy (solid) and present-class macro-F1 (hatched) for scVision, three foundation models and seven classical baselines (bars, five-fold mean; whiskers ± 1 s.e.m.). scVision (coral) leads both; its +3.6-point balanced-accuracy margin over scFoundation is significant (paired t , $df = 4$, $p < 0.001$; versus scGPT $p < 0.01$). ROC-AUC exceeds 0.98 throughout, so present-class metrics are reported. **c.** Label efficiency (k -shot, log axis), the four foundation models sharing the 20-NN probe; scVision leads at every budget and saturates by five shots. Dotted, scVision fully supervised (0.83); grey dashed, best classical; faint guide, chance. **d.** Representation versus classifier at $k = 5$: each classical baseline fits its own best classifier (bars); dashed, the shared NN probe; coral, scVision's. Native classifiers lift the strongest baselines (gradient-boosted trees 0.41), yet none reaches scVision (0.60). **e.** Attention recovers ovarian biology: per-cell-type top-attention programs tested for pathway enrichment (hypergeometric, Benjamini-Hochberg FDR). A pericyte unfolded-protein-response program is significant (Hallmark $q = 0.025$), an endothelial endocytosis program likewise (GO biological process $q = 0.027$). Points, $-\log_{10} q$; dashed line $q = 0.05$. **f.** The programs are spatially localised: per-cell-type attention maps (the 13×13 grid minus the cell-type average) give each population a distinct fingerprint, the pericyte signature concentrating on one region carrying HSP90B1, CALR and ERN1 from (e). **g.** The programs recur across organs: mean Jaccard overlap of each ovarian type's 100 top-attention genes with the same type across twenty-eight atlases. Endothelial and pericyte programs recur far above a thousand-permutation null (grey band; pooled $z = 8.1$, $p < 0.001$): VWF across kidney, brain, blood and fallopian-tube endothelium, CALR across prostate, breast and intestinal pericytes; the coarse leukocyte label does not.

We then tested how well each representation performed with very few labelled examples. For each label budget, $k \in \{1, 5, 10, 50\}$ labelled cells per type were sampled from the reference bank and shared across methods (Fig. 3c). With only one labelled cell per type, scVision reached 0.30 balanced accuracy. With five labelled cells per type, it reached 0.60, which is 72% of its fully supervised value of 0.83. Performance then changed little with more labels. Under the same nearest-neighbour readout, scVision led at every label budget. The gap was smaller than in the kidney cortex task, but scVision still performed best. One labelled cell per type with scVision outperformed fifty labelled cells per type with scGPT and Geneformer (0.30 versus 0.14 and 0.17). We do not make a fifty-fold label-efficiency claim here, because scFoundation with fifty labels reached 0.34, slightly above scVision with one label.

To test whether this advantage was caused by the nearest-neighbour classifier, we also allowed each classical baseline to use its own classifier. These included logistic regression, a multilayer perceptron, gradient-boosted trees, a random forest, k -nearest neighbours, a correlation classifier and TabNet, all trained on the same support cells (Fig. 3d). The k -nearest-neighbour baseline reproduced the shared-probe result, confirming that the comparison was controlled. Allowing each method to use its own classifier improved some baselines. For example, gradient-boosted trees reached 0.41 balanced accuracy at five labelled cells per type. However, this remained below scVision's 0.60. scVision retained the lead at every label budget. At fifty labelled cells per type, scVision reached 0.51, compared with 0.32 for the best classical method. Again, the label efficiency reflects the learned spatial representation rather than a particular classifier.

We also asked whether the same frozen representation captured interpretable ovarian biology. We read scVision's last-block attention as gene-expression programs by mapping the most-attended image patches back to genes through the scImage. This analysis used no pathway supervision. The recovered programs were cell-type specific and biologically coherent (Fig. 3e). In pericytes, the strongest program was enriched for the unfolded-protein response (MSigDB Hallmark $q = 0.025$), an endoplasmic-reticulum stress pathway marked by CALR, ERN1 and HSP90B1. Related endoplasmic-reticulum stress annotations appeared below significance in KEGG, GO and Reactome, suggesting a consistent signal across databases. The endothelial program was enriched for positive regulation of endocytosis (GO biological process $q = 0.027$), consistent with endothelial vesicular transport. The ovarian stromal program showed enrichment for Wnt/ β -catenin signalling (Hallmark $q = 0.088$), a pathway linked to ovarian stromal and follicular biology. The number of recovered programs was smaller than in the kidney analysis, as expected for a four-cell-type dataset, but the programs were tissue appropriate and were obtained without pathway labels.

Two additional analyses showed that these programs were not artefacts of the readout and were not specific only to ovary. First, when attention was displayed as cell-type-specific maps over the image grid, each ovarian population showed a distinct spatial pattern (Fig. 3f). The pericyte signal was concentrated in a contiguous image region containing the unfolded-protein-response genes HSP90B1, CALR and ERN1, showing that the program was spatially localized rather than diffuse. Second, we compared each ovarian attention program with programs from the same cell type across a 28-atlas attention bank, using the overlap of the 100 most-attended genes. The endothelial and pericyte programs recurred far above a permutation null (pooled $z = 8.1$, $p < 0.001$; Fig. 3g). The shared programs were biologically meaningful: VWF anchored the endothelial program across kidney, brain, blood and fallopian-tube endothelium, while CALR recurred in pericytes from prostate, breast and intestine. These results suggest that scVision's unsupervised attention captures conserved cell-type programs rather than tissue-specific noise.

Finally, we examined whether the annotation differences were already visible in the geometry of the frozen embeddings. When the four ovarian populations were projected to two dimensions, scVision separated endothelial cells, leukocytes, pericytes and stromal cells into compact and distinct clusters (Fig. 3a). Annotating each embedding with its matched 20-nearest-neighbour probe balanced accuracy, the common identifiability metric we report across all held-out atlases (Fig. 8), placed scVision first

(0.83), followed by scFoundation (0.79), scGPT (0.78) and Geneformer (0.59), reproducing the annotation ranking in (b). Label-free clustering scores agreed: among the foundation embeddings, scVision also attained the highest 15-nearest-neighbour purity (0.97) and Moran's I (0.93). Even before annotation, the frozen scVision embedding organizes held-out ovarian cells into the most readable cell-type geometry.

A microglial program recurs across organs in held-out dysplastic cortex

We next tested scVision on an external focal cortical dysplasia cortex dataset. Focal cortical dysplasia is a developmental abnormality of the cerebral cortex and a common cause of drug-resistant focal epilepsy. This dataset provides a stringent transfer test because the model had not seen dysplastic-cortex cells during training. The task was to assign cortical cell types using a frozen embedding and a reference bank built from other studies.

Using the same nearest-neighbour probe for all foundation models, scVision achieved the best performance, with 0.65 balanced accuracy and 0.71 present-class macro-F1 (Fig. 4b). It narrowly outperformed scGPT²⁰ on both metrics: scGPT reached 0.64 balanced accuracy and 0.706 macro-F1, against 0.65 and 0.714 for scVision. scVision more clearly outperformed Geneformer¹⁹ (0.57 and 0.64) and scFoundation²¹ (0.46 and 0.53). Because all models used the same readout, these differences reflect the quality of the learned representations.

This performance was also visible in the embedding geometry (Fig. 4a). Under identical UMAP settings, scVision separated the 15 cortical cell types into the clearest clusters, and each panel is annotated with its matched 20-nearest-neighbour probe balanced accuracy, the common identifiability metric we report across held-out atlases (Fig. 8), on which scVision ranks first. Label-free clustering agreed: scVision also formed the tightest neighbourhoods, with a 15-nearest-neighbour purity of 0.87 against 0.83 for scGPT, 0.77 for Geneformer and 0.73 for scFoundation. This ranking matched the annotation performance, showing that scVision produced the most readable cell-type structure.

We also compared scVision with classical and tabular baselines using threshold-free metrics. In this focal cortical dysplasia cortex, scVision achieved the highest macro-averaged average precision (0.78) and AUROC (0.96), ahead of gradient-boosted trees (0.75 and 0.94), a multilayer perceptron (0.74 and 0.94), and other classical baselines (Fig. 4c). Its balanced accuracy of 0.65 was also higher than the best classical method, which reached 0.51. Across all measured metrics, scVision was the most accurate annotator.

We then tested label efficiency. Each embedding was kept frozen, and each method was evaluated using $k = 1, 5, 10, 50$ labelled cells per class. scVision was the most label-efficient foundation model at one, five and fifty labels per class, reaching 0.39, 0.51 and 0.58 balanced accuracy, respectively (Fig. 4d). Geneformer and scFoundation performed lower across the label sweep. The difference from classical baselines was larger: the seven tabular probes reached only 0.25 balanced accuracy even with fifty labelled cells per class, whereas scVision exceeded this level with a single labelled cell per class. The same representation that produced the best full-reference annotation also required the fewest labels.

We next examined whether scVision's attention captured interpretable disease biology. We mapped the most-attended image patches from the last transformer block back to genes through the scImage. This produced cell-type-specific gene programs without using pathway labels (Fig. 4e). Microglia and endothelial cells were enriched for p53 signalling (KEGG $q = 0.020$), consistent with immune and vascular stress in the epileptogenic lesion. Oligodendrocyte precursor cells were enriched for growth-factor and differentiation pathways, including RET signalling (Reactome $q = 0.002$), DARPP-32 events ($q = 0.003$) and SCF-KIT signalling ($q = 0.036$), consistent with oligodendroglial abnormalities in dysplastic cortex. Somatostatin interneurons were enriched for glycosylphosphatidylinositol-anchor biosynthesis (KEGG $q = 0.024$), relevant to neuronal cell-surface biology. Re-evaluated across the same eleven cell-resampling seeds used in (f), this p53 vascular-immune enrichment was the most reproducible of these programs, significant (BH-FDR $q < 0.05$) in all eleven endothelial and nine of eleven microglial seeds, with oligodendrocyte-precursor RET signalling holding in nine of eleven, whereas the remaining precursor and interneuron annotations were more seed-dependent.

These attention programs were mostly specific to individual cell types. Across the 15 cortical cell types, the mean overlap between top-attention programs was low, with a mean Jaccard index of 0.044 (Fig. 4f). This value was stable across eleven cell-resampling seeds (range 0.028–0.055), indicating that the model did not simply learn one broad lesion-wide stress signal. The main exception was the microglia-endothelium pair, whose top programs were nearly identical (Jaccard 1.0 in nine of eleven seeds; at least 64 of 100 genes shared in all eleven, mean Jaccard 0.90), reflecting a shared p53-linked vascular-immune module. This was the only such overlap between distinct lineages; neuronal subtypes showed moderate overlap, as expected for related excitatory and inhibitory cell populations, whereas most other cell-type pairs remained distinct.

Finally, we asked whether the microglial program was specific to focal cortical dysplasia or reflected a broader myeloid identity. The top-attended program in dysplastic-cortex microglia contained four genes: IRF8, PIM1, TRAF1 and CCND2. The same four-gene program was also the top-attended program in healthy-brain microglia, kidney interstitial macrophages, breast macrophages and autoimmune-blood dendritic cells (Fig. 4g). scVision recovered the same myeloid program across five organs without pathway supervision.

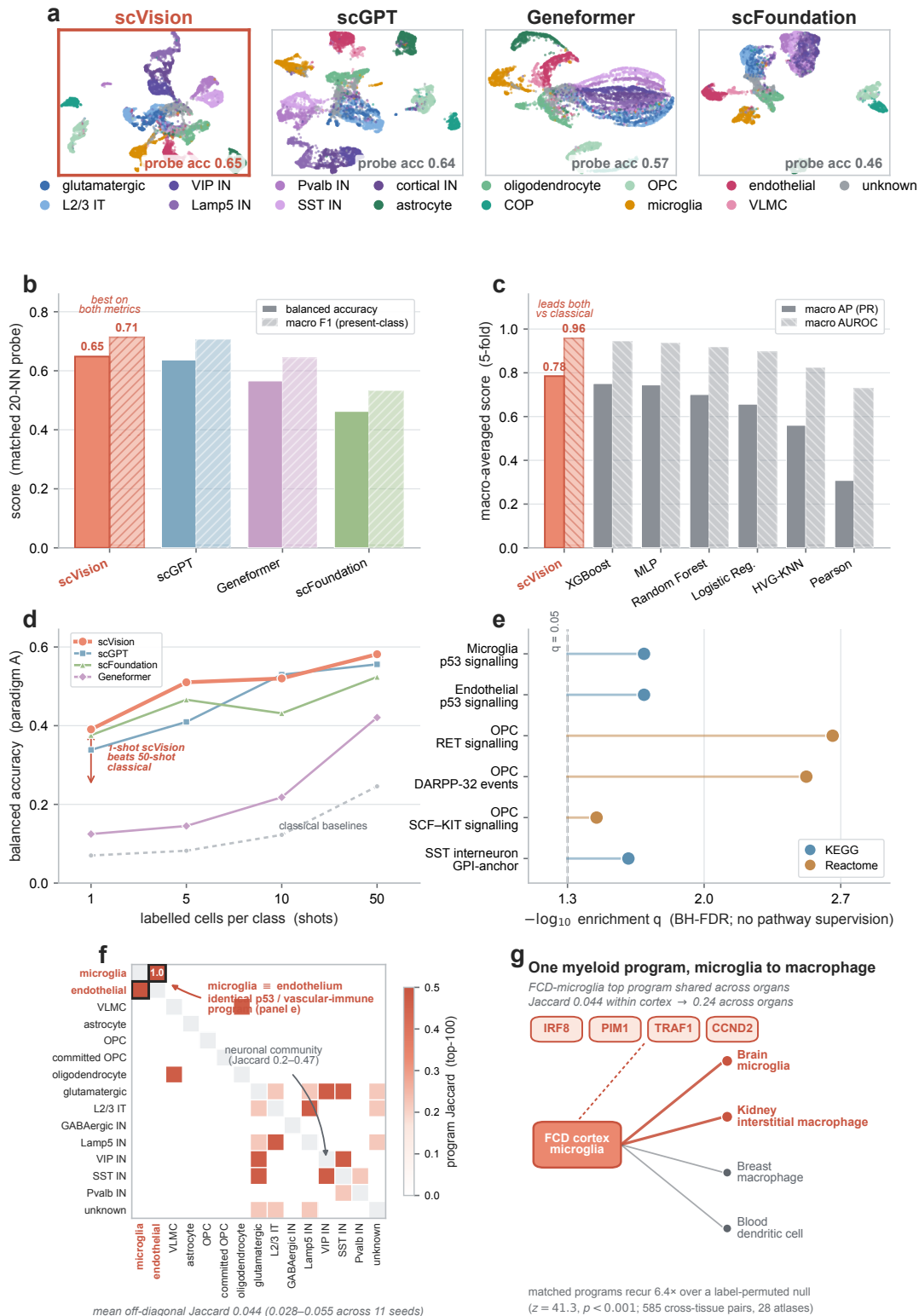


Figure 4. scVision annotates the held-out focal cortical dysplasia cortex most accurately and reads a microglial identity program it shares with macrophages across organs. Held-out focal cortical dysplasia cortex, cross-study holdout (paradigm A); every method reads a frozen representation scored on the full external test pool. (Panel-by-panel description a–g continued on the next page.)

Figure 4 (continued). **a**, Per-method frozen-embedding UMAP of the same held-out external test pool, coloured by the fifteen native cortical cell types (shared legend) and annotated with each model's matched 20-nearest-neighbour probe balanced accuracy, the common cross-atlas identifiability metric (Fig. 8); higher means a more readily annotated embedding. All four panels use identical UMAP settings (30 neighbours, minimum distance 0.3, cosine metric, and a PCA-50 pre-step where the embedding exceeds 100 dimensions), so only the representation being visualised differs; scVision is the most identifiable embedding (0.65, against 0.64/0.57/0.46 for scGPT/Geneformer/scFoundation). **b**, The matched full-bank 20-nearest-neighbour probe reading the same four frozen representations: balanced accuracy (solid) and present-class macro-F1 (hatched), all sharing the identical probe so only the embedding differs. scVision (coral) leads both, and the panel-a badge is exactly this matched-probe balanced accuracy. **c**, Calibrated detection against six classical and tabular baselines: macro-averaged average precision (solid) and AUROC (hatched), five-fold. **d**, Label efficiency ($k = 1, 5, 10, 50$), each method read through the same 20-nearest-neighbour probe. scVision (coral) leads the foundation models at one, five and fifty shots (0.39, 0.51, 0.58) and is edged only by scGPT at ten (0.53 versus 0.52). Every foundation model sits far above the seven classical probes (grey), whose fifty-shot accuracy (0.25) scVision exceeds with one labelled cell per class. **e**, Attention recovers cortical biology: per-cell-type top-attention programs tested for pathway enrichment (hypergeometric, Benjamini–Hochberg FDR against KEGG and Reactome). Points, significant cell-type→pathway hits by $-\log_{10} q$; dashed line, $q = 0.05$. **f**, Programs are cell-type-specific: pairwise Jaccard overlap of the top-100 attention programs across the fifteen cortical cell types (mean off-diagonal 0.044; 0.028–0.055 across eleven cell-resampling seeds). The dominant cross-lineage overlap is the microglia–endothelium pair (Jaccard 1.0; boxed), the same p53-linked vascular–immune module scoring in (e) and the most reproducible off-diagonal feature, an identical top-100 program in nine of eleven seeds, sharing at least 64 of 100 genes in all eleven (mean Jaccard 0.90). Excitatory and inhibitory neurons form a looser community (Jaccard 0.2–0.47). **g**, The dysplastic-cortex microglia's top four-gene program (IRF8, PIM1, TRAF1, CCND2) is also top-attended in healthy-brain microglia, kidney interstitial macrophage, breast macrophage and autoimmune-blood dendritic cell, one myeloid identity program across five organs (matched-type Jaccard 0.24 versus 0.044 within-cortex). Across 585 cross-tissue cell-type pairs spanning twenty-eight atlases, matched programs recur 6.4-fold over a label-permuted null ($z = 41.3$, $p < 0.001$).

This cross-tissue signal was stronger than the overlap between different cell types within the same lesion. Different cortical cell types shared only 0.044 of their attention programs on average, whereas the same cell type across organs shared 0.24, about fivefold higher. Across 585 matched cell-type pairs from 28 atlases, attention programs from the same cell type overlapped 6.4-fold more than expected under a label-permuted null model ($z = 41.3$, $p < 0.001$). These results suggest that scVision learns reusable cell-type identity programs that can be recognized across tissues, including activated microglia in diseased human cortex.

Inflamed-stroma programs in held-out Crohn's-disease ileum

We next tested scVision on an external Crohn's-disease ileum dataset. This was a demanding transfer test because the model had not seen ileal cells during training. The task was to assign each held-out cell to a cell type using a frozen embedding and a fixed reference bank of 87 cell types from other studies. The original ileum dataset contains about 30 expert-annotated populations, including epithelial, stromal, endothelial and immune cells. Of these, 10 cell types were also present in the reference bank and could therefore be evaluated in the matched annotation task (Fig. 5a).

We compared all foundation models using the same 20-nearest-neighbour probe on their frozen embeddings. This controlled comparison isolates the quality of the representation rather than the classifier used for readout. Under this matched readout, scVision achieved the best performance, with 0.51 balanced accuracy and 0.52 present-class macro-F1 (Fig. 5b). It also matched or exceeded the strongest non-foundation baselines, including gradient-boosted trees (0.48 balanced accuracy and 0.51 macro-F1) and a multilayer perceptron (0.46 and 0.49). Other foundation models performed lower under the same readout: scFoundation²¹ reached 0.44 balanced accuracy and 0.47 macro-F1, scGPT²⁰ reached 0.39 and 0.42, and Geneformer¹⁹ reached 0.33 and 0.36. We focus on balanced accuracy and present-class macro-F1 because these metrics evaluate performance across all scored cell types, including rarer mucosal populations that are important for Crohn's disease biology.

We then tested how many labelled examples were needed for annotation. For each method, we sampled $k \in \{1, 5, 10, 50\}$ labelled cells per type and evaluated performance on the full external dataset (Fig. 5c). scVision was the most label-efficient method. With only one labelled cell per type, it reached 0.27 balanced accuracy, which was higher than any classical baseline reached even with 50 labelled cells per type (0.09). scVision also led all other foundation models at every label budget. Because the same readout was used across embeddings, this improvement reflects the learned representation rather than a more powerful classifier.

The five cross-validation folds were fixed and shared across methods, allowing a fold-by-fold comparison (Fig. 5d). At the matched probe, scVision outperformed the best competing foundation model, scFoundation, on all five folds. The mean

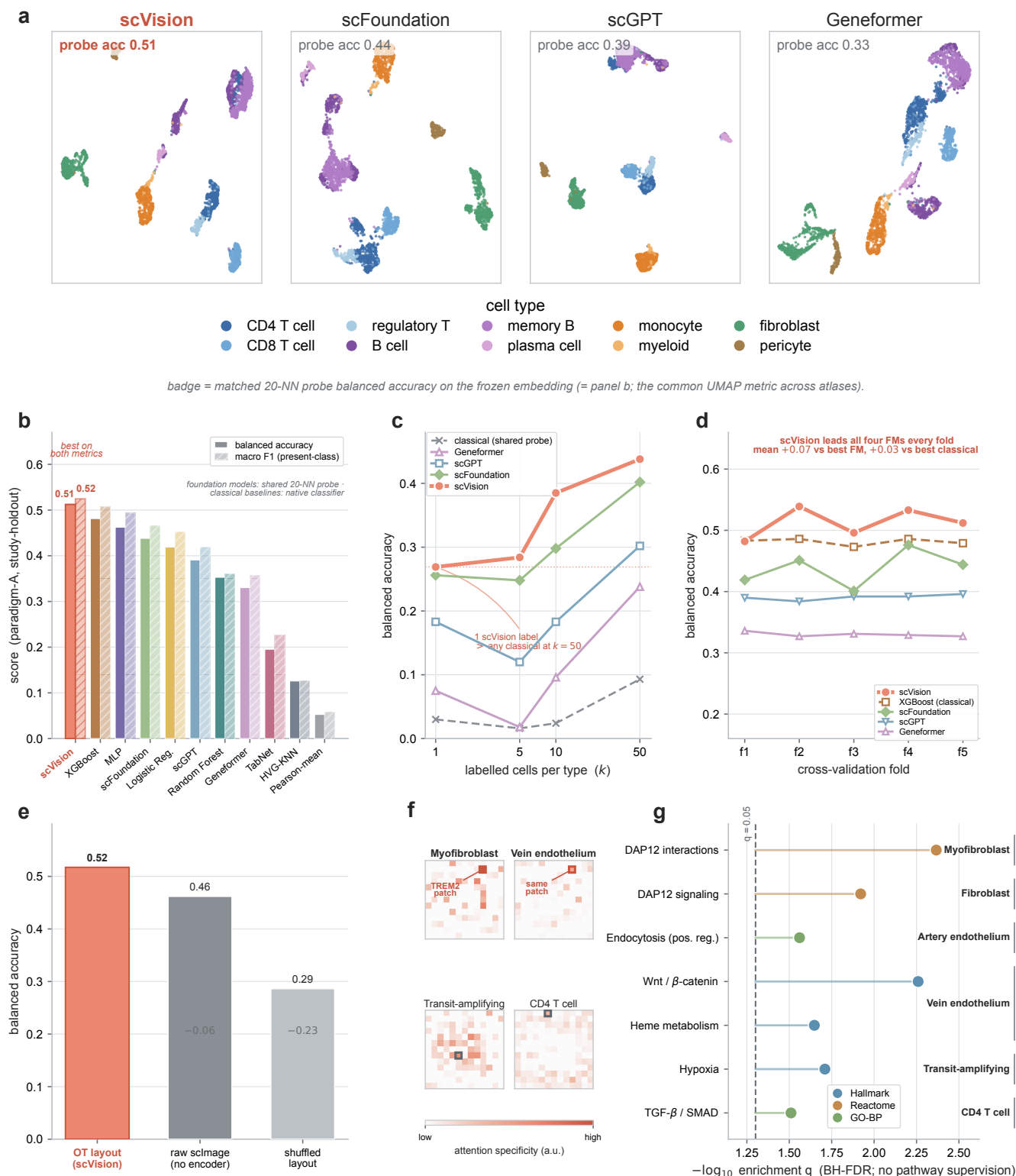


Figure 5. scVision is the most accurate and most label-efficient annotator of the held-out Crohn's-disease ileum, and its frozen attention reads the inflamed-stroma biology. Held-out Crohn's-disease ileum, cross-study holdout (paradigm A); every method reads a frozen representation scored on the full external test pool. (Panel-by-panel description a–g continued on the next page.)

Figure 5 (continued). **a**, Embedding geometry: UMAP of each model’s frozen embedding of the external test cells, coloured by expert type; each badge is the matched 20-nearest-neighbour probe balanced accuracy, the common cross-atlas identifiability metric (Fig. 8): scVision 0.51, scFoundation 0.44, scGPT 0.39, Geneformer 0.33. scVision is the most identifiable embedding; within-test cluster tightness (kNN purity) separates the foundation embeddings far less than cross-study transfer does, so the decisive margin is cross-study transfer (b), not within-test geometry. **b**, Matched read-out: foundation models share the same 20-NN probe, classical and tabular baselines use their own native classifier. Balanced accuracy (solid) and present-class macro-F1 (hatched) for scVision, three foundation models and seven classical baselines; scVision (coral) leads both and alone among single-cell embeddings reaches the strong non-foundation baselines. **c**, Label efficiency at $k \in \{1, 5, 10, 50\}$, foundation models read through the same 20-NN probe, classical through the shared raw-gene probe. One scVision label exceeds every classical baseline given fifty. **d**, Per-fold robustness across five method-matched folds. scVision leads every foundation model on all five and leads or ties the best classical (gradient-boosted trees) on every fold (fold 1 a 0.001 dead heat); mean $+0.07$ over the best foundation model (scFoundation, paired $p=5 \times 10^{-4}$), $+0.03$ over the best classical. **e**, The advantage is in the spatial representation: balanced accuracy for the optimal-transport scImage layout (native head 0.52; 20-NN probe 0.51), the same encoder on a permuted layout (“shuffled”), and a perceptron on the flattened raw scImage (“no encoder”). Destroying the layout costs 0.23, removing the encoder 0.06. **f**, Frozen attention concentrates on compartment-specific patches: specificity-corrected last-block attention for four cell types on the 13×13 grid; myofibroblast and vein endothelium peak on the same TREM2/DAP12 inflamed-stroma patch (outlined). **g**, Attention recovers inflamed-gut biology: per-cell-type top-attention programs tested for pathway enrichment (hypergeometric, Benjamini–Hochberg FDR; Hallmark, Reactome, GO biological process). Points, significant hits by $-\log_{10} q$; dashed line, $q = 0.05$.

improvement was 0.07 balanced accuracy (paired $p = 5 \times 10^{-4}$; 95% confidence interval $[+0.05, +0.10]$). Compared with the strongest non-foundation baseline, gradient-boosted trees, the margin was smaller but still positive, with a mean improvement of 0.03 balanced accuracy across four of five folds. On this difficult atlas, scVision clearly outperformed the other foundation models and performed at least as well as the strongest classical baselines.

We next tested whether the spatial scImage layout was important. When the gene-to-pixel assignment was randomly permuted, balanced accuracy dropped from 0.52 to 0.29 (Fig. 5e). This shows that the learned spatial arrangement of genes is essential. Removing the vision encoder and classifying the flattened raw scImage with a multilayer perceptron caused a smaller drop, from 0.52 to 0.46. Both the vision backbone and the spatial gene layout together contribute to performance, but the gene arrangement carries the larger share of the signal.

The embedding geometry supported these results (Fig. 5a). In two-dimensional projections, scVision organized the external mucosa into compact and well-separated cell-type clusters, and its matched 20-nearest-neighbour probe balanced accuracy on the frozen embedding was the highest of the foundation models (0.51; the common cross-atlas identifiability metric, Fig. 8). A complementary label-free view was more nuanced: within-test k -nearest-neighbour purity was 0.92 for scVision, compared with about 0.70 for the raw-gene space used by classical learners, but scFoundation (0.91), Geneformer (0.88) and scGPT (0.87) nearly matched it. These local purity scores did not fully predict annotation accuracy. For example, scFoundation nearly matched scVision in within-dataset purity but performed worse when cells had to be matched to the external reference bank. This suggests that scVision’s main advantage is not only better clustering within the held-out tissue, but stronger transfer across studies.

Finally, we examined whether scVision’s attention captured disease biology. We mapped last-block attention back onto the scImage and found that attention concentrated on a small number of cell-type-specific patches rather than spreading diffusely (Fig. 5f). Most cell types had distinct peak patches, but myofibroblasts and venous endothelial cells shared the same peak patch. Mapping these patches back to genes, without pathway supervision, recovered programs consistent with Crohn’s disease pathophysiology (Fig. 5g).

The shared myofibroblast–endothelium patch was anchored by TREM2. Both myofibroblasts and fibroblasts were enriched for TREM2–DAP12 signalling (Reactome $q = 0.004$ and 0.012), suggesting a shared inflamed-stroma program in the gut wall. Other cell types showed distinct programs. Venous endothelial cells were enriched for Wnt/ β -catenin signalling and heme metabolism (Hallmark $q = 0.006$ and 0.023), consistent with vascular remodelling. Arterial endothelial cells were enriched for positive regulation of endocytosis (GO biological process $q = 0.028$), consistent with activated vessel-wall transport. Transit-amplifying crypt epithelial cells were enriched for hypoxia (Hallmark $q = 0.020$), consistent with the metabolic stress of inflamed and regenerating gut tissue. CD4 T cells were enriched for TGF- β /SMAD regulation (GO biological process $q = 0.031$), a pathway linked to the Treg–Th17 immune axis in Crohn’s disease. These programs were recovered without pathway labels, showing that the frozen scVision representation captures both transferable cell identity and disease-relevant biology.

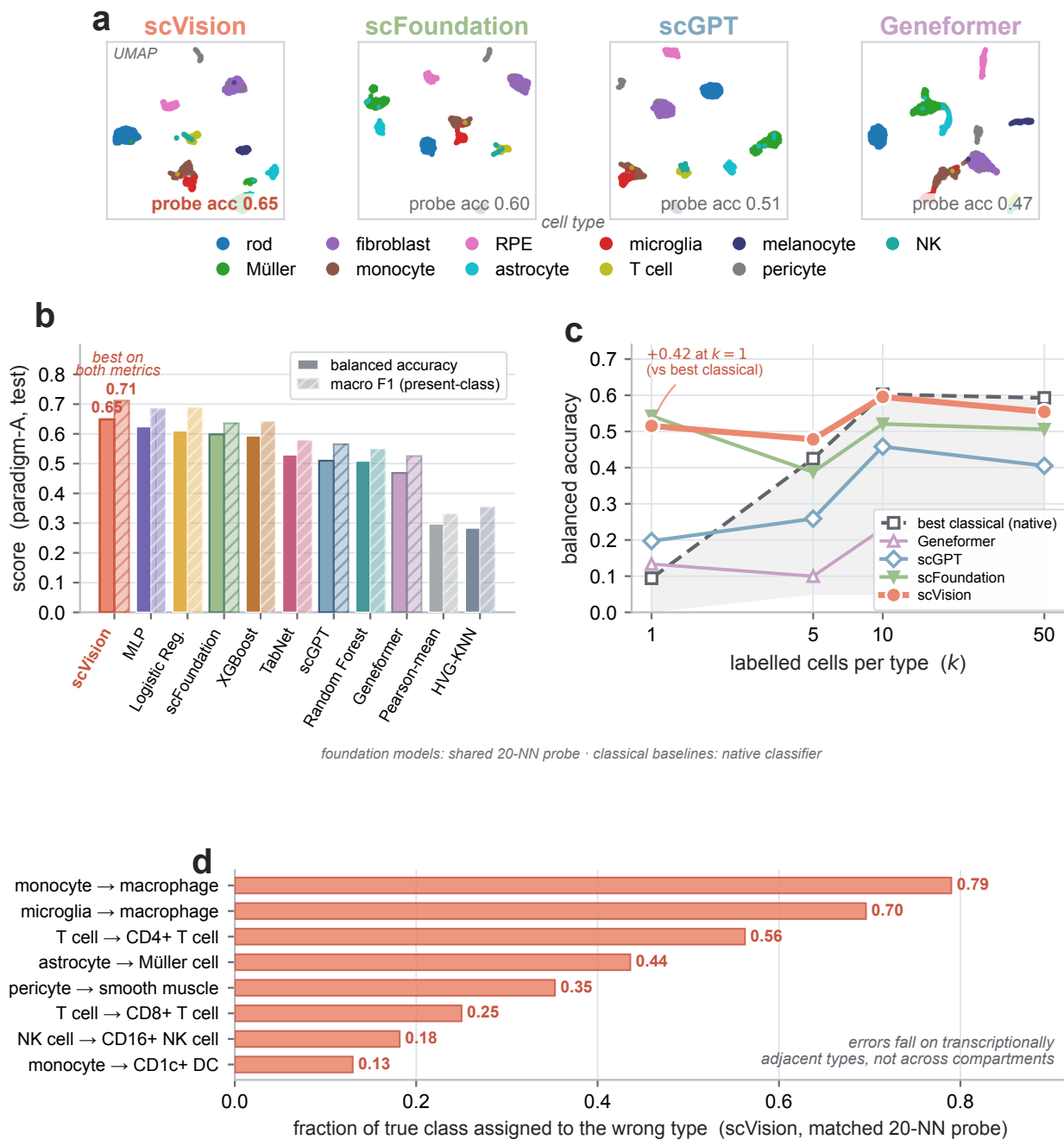


Figure 6. scVision annotates the held-out human retina most accurately and is the strongest foundation-model embedding from a handful of labels per type, while classical learners close the gap as labels accumulate. Held-out human retina, cross-atlas transfer (paradigm A); every method reads a frozen representation scored on the full external test pool. Panels a–d are described in the continued caption overleaf.

Figure 6. a, Frozen-embedding structure: UMAP (balanced per-class subsample) coloured by the eleven coarse retinal cell types, annotated with each embedding's matched 20-nearest-neighbour probe balanced accuracy, the common cross-atlas identifiability metric (Fig. 8). scVision, scFoundation and scGPT separate the neuronal, glial, immune and vascular compartments into distinct islands, and scVision is the most identifiable embedding (0.65, ahead of scFoundation 0.60, scGPT 0.51 and Geneformer 0.47); Geneformer recovers the same structure more diffusely. None saw a retinal cell in training. **b**, Matched read-out: foundation models share the same 20-NN probe, classical and tabular baselines native. Balanced accuracy (solid) and present-class macro-F1 (hatched); scVision (coral) leads both, by a slim margin over the strongest tabular learners. **c**, Label efficiency ($k \in \{1, 5, 10, 50\}$), foundation models through the same 20-NN probe, classical at their native classifier (best-of-seven envelope; band, full spread). At one label per type the foundation embeddings beat the best classical by 0.43, the regime where the frozen embeddings help most. **d**, Where the annotator's residual errors fall: the eight most frequent confusions of the matched-read-out annotator. Errors fall on transcriptionally adjacent types (monocytes and microglia as macrophages, T-cell subsets split among themselves, astrocytes as Müller glia, pericytes as smooth muscle), not on cells from a different compartment.

Label-efficient annotation of held-out human retina

We next tested scVision on an external human retina dataset. This is a challenging transfer setting because the retina contains many closely related neuronal and support cell types, including rods, cones, bipolar cells, amacrine cells, horizontal cells, ganglion cells, Müller glia, astrocytes, microglia and vascular cells. Many retinal subclasses differ by only a small number of transcripts. The model had not seen retinal cells during training, so the task was to assign more than 20 retinal cell types using a frozen embedding and a reference bank of 87 cell types from other atlases.

We compared all foundation models using the same 20-nearest-neighbour probe on their frozen embeddings. This allows a controlled comparison of the representations rather than the classifiers used on top of them. Under this matched readout, scVision achieved the best performance, with 0.65 balanced accuracy and 0.71 present-class macro-F1 (Fig. 6b). The margin was modest because several methods performed well on this atlas. A multilayer perceptron reached 0.62 balanced accuracy and 0.68 macro-F1, logistic regression reached 0.61 and 0.69, scFoundation²¹ reached 0.60 and 0.64, and gradient-boosted trees reached 0.59 and 0.64. scGPT²⁰ and Geneformer¹⁹ performed lower under the same matched readout, reaching 0.51 and 0.47 balanced accuracy, respectively. We focus on balanced accuracy and present-class macro-F1 because these metrics evaluate performance across all retinal cell types, including rare populations, rather than being dominated by abundant rods and cones.

This ranking was also visible in the embedding geometry. When each frozen representation was projected to two dimensions, scVision, scFoundation and scGPT separated the main retinal compartments into clear groups, including rod-dominated neurons, Müller glia, immune cells and vascular support cells (Fig. 6a). Geneformer recovered the same broad structure but with more diffuse separation. The UMAP layouts themselves are qualitative and not aligned across methods, but each panel is annotated with the embedding's matched 20-nearest-neighbour probe balanced accuracy, the common cross-atlas identifiability metric (Fig. 8), on which scVision ranked first (0.65), ahead of scFoundation (0.60), scGPT (0.51) and Geneformer (0.47). Cell types that form clearer groups in the embedding are more easily annotated by the 20-nearest-neighbour probe.

We then tested performance when labels were scarce. For each method, we sampled $k \in \{1, 5, 10, 50\}$ labelled cells per type and evaluated annotation on the full external dataset (Fig. 6c). Foundation-model embeddings performed especially well in the one-shot setting: with only one labelled cell per type, scVision reached 0.52 balanced accuracy, far above the best classical learner (0.09), making it the strongest representation in the low-label regime. On the near-saturated, threshold-free ROC-AUC, scVision reached 0.998.

Finally, we examined the errors made by scVision (Fig. 6d). The most frequent mistakes occurred between biologically related cell types. For example, monocytes and microglia were sometimes called macrophages, CD4 and CD8 T-cell subsets were confused with each other, astrocytes were sometimes called Müller glia, and pericytes were sometimes called smooth muscle cells. These errors occurred within related lineages, such as myeloid cells, T cells, glial cells and perivascular support cells, rather than across unrelated compartments. When scVision made mistakes, they were usually between transcriptionally similar neighbouring cell states.

Zero-shot annotation across a held-out 53-type multi-organ reference

We next tested scVision on a held-out multi-organ single-nucleus reference. This was the most demanding transfer test because the dataset contains many fine-grained cell types from several organs, including cardiomyocytes and vascular endothelial cells from heart, luminal, basal and epithelial cells from prostate, and smooth muscle cells from gut. The model had not seen this test study during training. The task was to assign each cell type using a frozen embedding and a reference bank built from other atlases.

We compared all foundation models using the same 20-nearest-neighbour probe on their frozen embeddings. This controlled

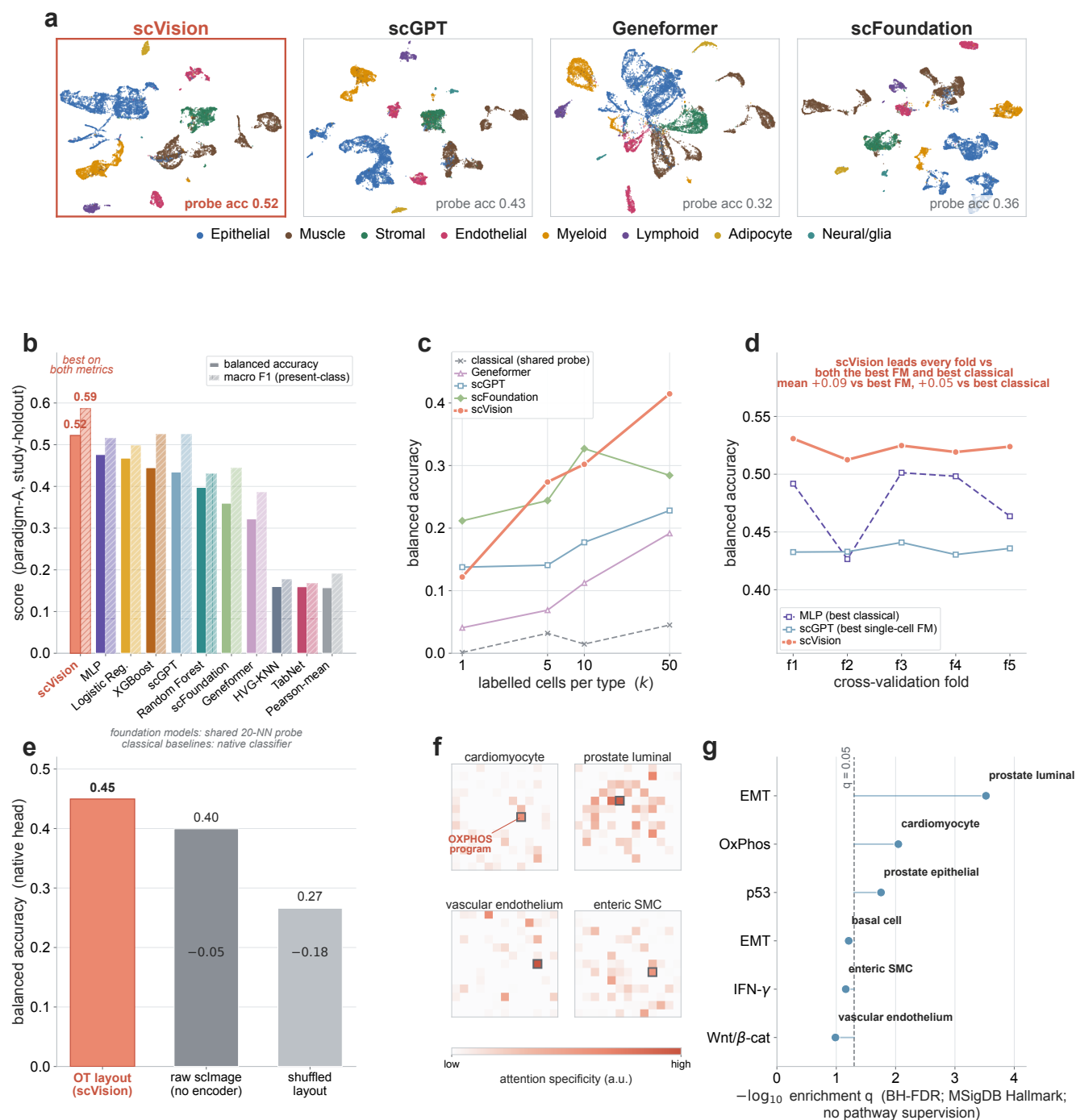


Figure 7. scVision most accurately annotates a held-out, multi-organ single-nucleus reference, is the only single-cell embedding to clear the classical baselines, and its frozen attention assigns each fine-grained cell type its own program. Held-out multi-organ single-nucleus reference under the cross-study holdout (paradigm A); every method reads a frozen representation and is scored on the full external test pool against the reference bank. (Panel-by-panel description a–g continued on the next page.)

Figure 7 (continued). **a**, Per-method frozen-embedding UMAP of the same held-out external test pool, coloured by the eight lineages the fifty-three native cell types collapse into (shared legend) and annotated with each model’s matched 20-nearest-neighbour probe balanced accuracy, the common cross-atlas identifiability metric (Fig. 8), computed at the full native resolution; higher means a more readily annotated embedding. All four panels use one matched UMAP routine (30 neighbours, minimum distance 0.3, cosine metric, and a PCA-50 pre-step where the embedding exceeds 100 dimensions), so only the representation differs. scVision is the most identifiable embedding on this fifty-three-type reference (0.52), ahead of scGPT (0.43), scFoundation (0.36) and Geneformer (0.32). Raw cluster tightness tells a different story: by label-free 15-nearest-neighbour purity scFoundation packs marginally the tightest islands (0.80), ahead of scVision (0.78), yet scFoundation annotates the same reference far worse than scVision (panel b), so raw embedding geometry is not a proxy for cross-study transfer accuracy. **b**, Annotation accuracy at a matched read-out: every foundation model is read with the same 20-nearest-neighbour probe on its frozen embedding (so the bars isolate representation quality rather than the choice of classifier), while the classical and tabular baselines use their own native classifier (their best case). Balanced accuracy (solid) and present-class macro-F1 (hatched) for scVision, three single-cell foundation models and seven classical and tabular baselines; scVision (coral) leads both metrics and is the only single-cell embedding to clear the strong non-foundation baselines, which here outrank all three single-cell foundation models. **c**, Label efficiency: balanced accuracy at $k \in \{1, 5, 10, 50\}$ labelled cells per type for the four single-cell foundation models, each read through the matched probe, against the classical baselines (which share the raw-gene probe and collapse to a single floor band that never clears 0.05). All four foundation embeddings sit above that floor from a single label, and scVision scales best across the budget range, reaching the highest balanced accuracy of the foundation models by $k=50$ (0.41). **d**, Per-fold robustness across the five deterministic, method-matched folds: scVision and the best foundation model (scGPT) at the matched probe, the best classical learner (a multilayer perceptron) native. scVision leads both on all five folds, a mean $+0.09$ over the best foundation model (paired $p=9 \times 10^{-6}$; 95% confidence interval $[+0.08, +0.10]$) and $+0.05$ over the best classical (paired $p=0.02$; $[+0.01, +0.08]$). **e**, The advantage is in the spatial representation. Balanced accuracy for the optimal-transport scImage layout (scVision’s native head, 0.45; the matched 20-NN probe of panel b reads it higher, at 0.52), a multilayer-perceptron probe on the flattened raw scImage (“no encoder”, 0.40), and the same encoder reading a randomly permuted layout (“shuffled”, 0.27). Destroying the layout costs 0.18 balanced accuracy; removing the encoder costs 0.05; the learned gene layout carries most of the signal. **f**, scVision’s frozen attention concentrates on compartment-specific scImage patches: specificity-corrected last-block attention for four cell types drawn from different compartments of the reference (a cardiomyocyte, a prostate luminal cell, a vascular endothelial cell and an enteric smooth-muscle cell) on the 13×13 patch grid. The cardiomyocyte peaks on its oxidative-phosphorylation patch (outlined); each type peaks on a patch of its own. **g**, That attention recovers cell-type-appropriate biology: per-cell-type top-attention gene programs, mapped to genes through the scImage, tested for pathway enrichment (hypergeometric, Benjamini–Hochberg FDR) against the Hallmark collection with no pathway supervision. Points mark cell-type→pathway hits by $-\log_{10} q$, each labelled with its cell type; dashed line, $q = 0.05$. Three programs clear significance (prostate luminal cells with epithelial–mesenchymal transition, cardiomyocytes with oxidative phosphorylation, prostate epithelium with the p53 pathway), and three more corroborate just below threshold.

comparison isolates the quality of the representation rather than the classifier placed on top of it. Under this matched readout, scVision achieved the best performance, with 0.52 balanced accuracy and 0.59 present-class macro-F1 (Fig. 7b). Other foundation models performed lower under the same readout: scGPT²⁰ reached 0.43 balanced accuracy and 0.53 macro-F1, scFoundation²¹ reached 0.36 and 0.44, and Geneformer¹⁹ reached 0.32 and 0.39. The strongest non-foundation baselines were a multilayer perceptron, with 0.48 balanced accuracy and 0.51 macro-F1, and gradient-boosted trees, with 0.44 and 0.52. On this multi-organ reference, scVision was the only single-cell foundation-model embedding that exceeded the strongest classical baselines. We focus on balanced accuracy and present-class macro-F1 because these metrics evaluate performance across rare and abundant cell types more evenly.

We then tested label efficiency by varying the number of labelled examples per type. For each method, we sampled $k \in \{1, 5, 10, 50\}$ support cells per type and evaluated performance on the full external dataset, while keeping the readout fixed (Fig. 7c). The foundation-model embeddings clearly separated from the classical baselines. The classical methods, which read the same raw-gene representation, remained near a shared floor and did not exceed 0.05 balanced accuracy even with 50 labelled cells per type. In contrast, a single labelled cell per type already lifted the foundation-model embeddings above that floor, and scVision scaled best across the budget range.

Among the foundation models, scVision’s lead widened as labels accrued: by 50 labels per type it was clearly strongest, reaching 0.41 balanced accuracy, compared with 0.28 for scFoundation, 0.23 for scGPT and 0.19 for Geneformer. These results show that label efficiency comes from the pretrained representation, and that scVision benefits most as additional labels become available.

Because the five cross-validation folds were fixed and shared across methods, we also compared performance fold by fold (Fig. 7d). At the matched probe, scVision outperformed the best competing foundation model, scGPT, on all five folds. The mean improvement was 0.09 balanced accuracy (paired $p = 9 \times 10^{-6}$; 95% confidence interval [+0.08, +0.10]). scVision also outperformed the strongest non-foundation baseline, the multilayer perceptron, on all five folds, with a smaller mean improvement of 0.05 balanced accuracy (paired $p = 0.02$; 95% confidence interval [+0.01, +0.08]). The advantage over classical learning was modest but consistent across splits.

We next tested whether the spatial scImage layout was important (Fig. 7e). With its native head, the optimal-transport scImage reached 0.45 balanced accuracy, while the matched 20-nearest-neighbour readout in panel b reached 0.52. Replacing the encoder with a multilayer perceptron on the flattened scImage reduced performance by 0.05. Randomly permuting the gene-to-pixel layout before encoding reduced performance by 0.18. The learned spatial arrangement of genes therefore carries much of the signal, more than the encoder alone.

Finally, we examined scVision's attention programs across organs. We mapped last-block attention back to the 13×13 scImage grid for cell types from different compartments, including cardiomyocytes, prostate luminal cells, vascular endothelial cells and enteric smooth muscle cells (Fig. 7f). Each cell type showed a distinct peak attention patch. The genes in these patches matched relevant cell-type biology. Without using pathway labels, pathway enrichment against the Hallmark collection identified significant programs for prostate luminal cells, enriched for epithelial–mesenchymal transition ($q = 3 \times 10^{-4}$), cardiomyocytes, enriched for oxidative phosphorylation ($q = 0.009$), and prostate epithelial cells, enriched for the p53 pathway ($q = 0.02$) (Fig. 7g). Additional programs, including basal-cell epithelial–mesenchymal transition, enteric smooth-muscle interferon- γ response and vascular endothelial Wnt signalling, were just below the significance threshold.

Together, these results show that scVision transfers across organs, improves with limited labels, and assigns distinct attention programs to fine-grained cell types. The pathway signal is modest and unsupervised, but it shows that the model often attends to genes with biology appropriate for the cell type being annotated.

Embedding geometry predicts annotation accuracy across atlases

The link between annotation accuracy and embedding structure held across every held-out atlas. For each frozen embedding we measured a single common identifiability score: the balanced accuracy of a matched 20-nearest-neighbour probe read directly from the embedding at native cell-type resolution, the same metric annotated on every atlas's UMAP panels. Across all six held-out references, spanning 4 to 53 native cell types, scVision produced the most identifiable embedding on every atlas (Fig. 8).

scVision was the most identifiable embedding on all six atlases: the four-type ovary (0.83), the ten-type Crohn's ileum (0.51), the eleven-type retina (0.65), the fifteen-type focal cortical dysplasia cortex (0.65), the twenty-six-type kidney cortex (0.47) and the fifty-three-type multi-organ reference (0.52). No other embedding led any atlas. scFoundation was the most frequent runner-up, ranking second on the ovary (0.79), kidney cortex (0.40), retina (0.60) and Crohn's ileum (0.44), but it fell to third on the multi-organ reference and fourth on the dysplastic cortex. scGPT was second on the dysplastic cortex (0.64) and the multi-organ reference (0.43) but third on the other four atlases. Geneformer ranked third or fourth throughout. scVision was therefore the only frozen single-cell embedding that was most identifiable on every held-out atlas, from a simple four-type ovary to a complex fifty-three-type multi-organ reference.

These embeddings can be inspected directly: we projected the two largest references, the kidney cortex with 26 native cell types and the multi-organ atlas with 53 native cell types, into two dimensions using the same UMAP settings as for the dysplastic cortex (Fig. 2a and Fig. 7a). Each panel carries its matched-probe badge, and scVision is the most identifiable embedding in both, separating the major epithelial, immune, stromal and endothelial compartments into coherent groups while the other embeddings are progressively more diffuse.

Because this identifiability score reads each embedding the same way the annotation probe does, it tracks annotation accuracy directly: scVision is both the most identifiable embedding and the most accurate annotator on every atlas. Label-free clustering scores tell a subtler story: by raw 15-nearest-neighbour purity, scGPT forms marginally the tightest neighbourhoods on the kidney cortex and scFoundation on the multi-organ reference, confirming that placing identical labels closest together is not sufficient for accurate cross-study annotation. scVision instead organizes cell states so that a simple matched probe reads them most reliably across all six held-out atlases.

This layout-free purity is also the comparison most easily dismissed as a projection artifact, so we tested its robustness directly. On four held-out atlases on which scVision does lead label-free purity — the focal cortical dysplasia cortex and Crohn's ileum, plus two further held-out atlases, autoimmune blood (PBMC) and type-I-interferon COVID-19 — we chose one UMAP layout per atlas and applied it identically to all four foundation models, with no per-model tuning (Fig. 9). scVision retained the tightest native cell-type neighbourhoods on all four, both on the UMAP-invariant high-dimensional features and on the two-dimensional embedding shown. The high-dimensional margins over the closest competitor were modest, from +0.014 to +0.038 (scFoundation on three atlases, scGPT on the dysplastic cortex), consistent with purity separating the foundation

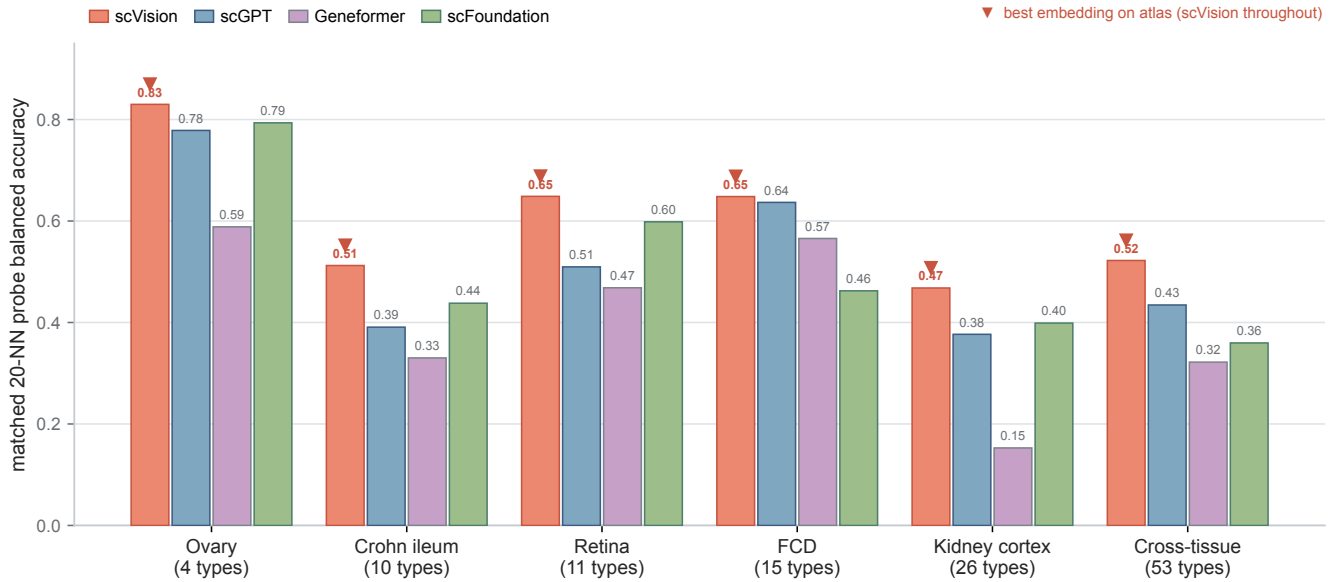


Figure 8. The annotation ordering is written into the embedding geometry across every atlas: scVision yields the most identifiable frozen embedding on all six held-out references. For each held-out atlas (native cell-type count in parentheses, spanning 4 to 53 types), every embedding is scored with one common metric, the balanced accuracy of a matched 20-nearest-neighbour probe read directly from the frozen embedding at native cell-type resolution, computed identically for scVision and the three single-cell foundation models. A triangle marks the best embedding on each atlas. scVision is the most identifiable embedding on every atlas (ovary 0.83, Crohn’s ileum 0.51, retina 0.65, focal cortical dysplasia cortex 0.65, kidney cortex 0.47, multi-organ reference 0.52); scFoundation is second on four atlases but falls to third or fourth on the other two, scGPT is second on two and third on the rest, and Geneformer is third or fourth throughout.

embeddings less sharply than the matched probe does; but they were stable across the shared layout rather than the product of a favourable projection.

scVision's cell-type geometry edge holds under a shared layout applied to every model

One UMAP layout per atlas, the same for every model. raw = high-D kNN-purity (UMAP-invariant, the decider); 2-D = purity of the embedding shown. $k=15$ cosine, 500 cells/class.

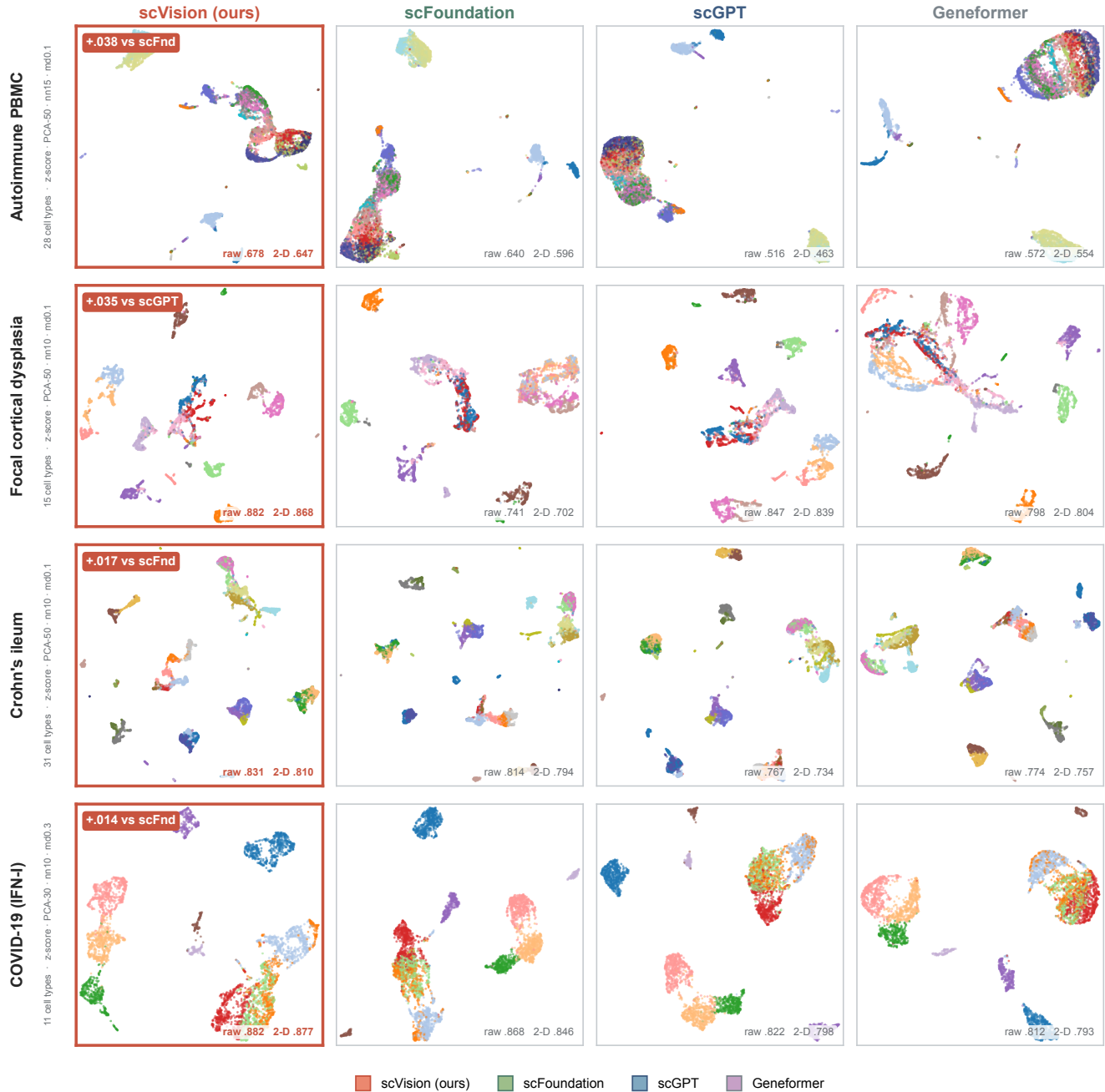


Figure 9. On four held-out atlases, a single UMAP layout shared by every model shows scVision's native cell-type geometry edge is not a projection artifact. Rows, four held-out atlases, the focal cortical dysplasia cortex and Crohn's ileum (also Figs. 4a and 5a) and two further held-out atlases, autoimmune blood (PBMC) and type-I-interferon COVID-19; columns, the four single-cell foundation models. For each atlas one UMAP layout (normalisation, PCA dimension, neighbours and minimum distance; annotated at left) was chosen and applied identically to all four models, so only the representation differs across a row; rows are ordered by scVision's high-dimensional margin. Points are individual cells coloured by native cell type (11–31 types per atlas). Each panel carries two kNN label-purity values ($k = 15$ cosine, balanced 500 cells per class, seed 0): *raw*, on the frozen high-dimensional features (UMAP-invariant, the genuine label-geometry decider), and *2-D*, on the embedding shown. scVision (coral frame) gives the tightest native neighbourhoods on every atlas in both values; the coral flag is its raw-space margin over the closest competitor (scFoundation on three atlases, scGPT on the dysplastic cortex), from +0.014 to +0.038. scGPT and Geneformer are looser on every atlas. These modest margins match the main-text finding that within-test purity separates the foundation embeddings less sharply than the matched cross-study probe (Fig. 8); the point here is that the ordering purity does show is stable to the layout, not a product of per-model tuning.

The frozen embedding integrates cells across studies without batch labels

We next asked whether the same frozen embedding could integrate cells across studies. Annotation tests whether a representation separates cell types. Integration tests a complementary property: whether cells from different studies can be placed into a shared space while preserving biological structure and reducing technical batch effects. This is a standard benchmark for single-cell foundation models³⁰. Because scVision is pretrained only by masked image modelling and is never trained with batch labels, this integration task tests whether its frozen embedding generalizes beyond its training objective.

For each of eight held-out atlases, including blood, hypothalamus, fovea, heart, kidney, lung, ovary and jejunum, we embedded a balanced pool of about 20,000 cells from multiple studies. We compared six methods: scVision, scGPT²⁰, Geneformer¹⁹, scVI²⁹, Harmony²⁸ and principal-component analysis. We evaluated each method using the scIB v2 benchmark, which combines five biological-conservation metrics and four batch-correction metrics into a weighted total score (Fig. 10). This comparison is especially challenging because scVI is trained separately on each atlas with dataset identity provided as a batch covariate, whereas the foundation-model embeddings are used frozen and do not receive batch labels.

scVision preserved biological structure best overall (Fig. 10a). It had the highest mean biological-conservation score, 0.61, with a 95% bootstrap confidence interval of 0.58 to 0.64, and ranked first in five of eight atlases. scVision preserved biology significantly better than Geneformer (paired difference +0.042, $p = 0.008$), scVI (paired difference +0.020, 95% confidence interval [+0.005, +0.036], $p = 0.04$), raw PCA (difference +0.16, $p = 0.004$) and Harmony (difference +0.20, $p = 0.004$). scVision also slightly exceeded scGPT, although this difference was not statistically significant (paired difference +0.013, 95% confidence interval [-0.001, +0.028], $p = 0.07$). The comparison with scVI is particularly important because scVI is designed and trained for integration, while scVision is used as a frozen embedding without batch labels.

On the combined scIB total score, the three learned embeddings were statistically tied. scVI scored 0.520, scGPT scored 0.520 and scVision scored 0.519, with all paired differences from scVision within ± 0.001 and not significant (paired Wilcoxon $p \geq 0.32$). All three outperformed Geneformer, which scored 0.495 ($p = 0.008$), and the classical baselines, raw PCA and Harmony, which scored 0.41 and 0.38, respectively ($p = 0.004$).

The main trade-off was batch mixing. scVI, which is explicitly trained to remove batch effects, achieved the strongest batch-mixing score, 0.413. scVision had the lowest batch-mixing score among the foundation models (Fig. 10b), and this was also supported by a post-hoc kBET acceptance test³⁰ in which scVision scored 0.355, compared with 0.401 for scGPT and 0.417 for Geneformer. scVI and scGPT thus gained more batch mixing, whereas scVision preserved more biological structure. These effects nearly balanced in the total scIB score.

Together, these results show that scVision provides a frozen representation that preserves cell identity strongly across studies and reaches the same overall integration frontier as scGPT and scVI. Although it is not explicitly trained for batch correction and does not mix batches as strongly as scVI, it matches the best learned integration methods in total score while outperforming classical integration baselines.

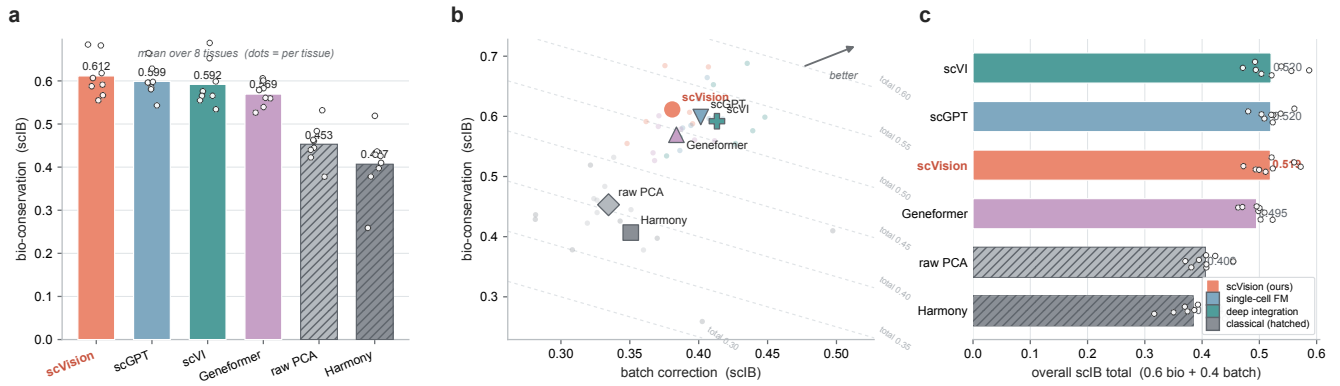


Figure 10. A frozen scVision embedding integrates multi-study atlases: it best conserves biology and matches the strongest token model on the combined score, without being trained for the task. Eight held-out atlases, each a balanced $\sim 20,000$ -cell pool spanning multiple studies (batches); six representations scored with the scIB v2 panel (bio-conservation = mean of NMI, ARI, cell-type and isolated-label silhouette and cLISI; batch correction = mean of batch silhouette, iLISI, graph connectivity and PCR; total = $0.6\text{bio} + 0.4\text{batch}$). **a**, Bio-conservation per method (bar = mean over tissues; white dots = the eight individual atlases); scVision is highest and rank-1 in five of eight tissues. **b**, The bio-batch trade-off: batch correction (x) versus bio-conservation (y), faint dots per tissue and large markers per-method means; grey iso-total contours ($0.6\text{bio} + 0.4\text{batch} = c$) show that scGPT's greater batch mixing offsets scVision's greater biological conservation, leaving their totals tied. **c**, Overall scIB total (horizontal bars, white dots per tissue): scVI 0.520, scGPT 0.520 and scVision 0.519 are statistically tied (paired Wilcoxon $p \geq 0.32$), all ahead of Geneformer and the classical baselines. Coral = scVision; the token foundation models and scVI in solid colour; classical baselines (Harmony, PCA) hatched. Across-tissue confidence intervals are from 10,000 bootstrap resamples.

The frozen embedding is far cheaper to deploy than token foundation models, and its zero-shot readout is stable

A representation meant to be used zero-shot is only useful if it can be computed quickly and read out without per-dataset tuning. Because scVision encodes a cell as a single fixed-size image rather than a variable-length sequence of gene tokens, its end-to-end cost is low. Measured on one NVIDIA RTX 2080 Ti from raw highly-variable-gene counts to the cell embedding (including each model's required preprocessing), scVision rendered an scImage and encoded it with the frozen vision transformer at 528 cells per second, against 14.2 for Geneformer, 12.9 for scFoundation and 1.8 for scGPT, a 37 to nearly $300\times$ throughput advantage over the token models, whose per-cell gene ranking and tokenization dominate their runtime. It reached this throughput at a peak memory of 2.55 GB, below Geneformer (8.6 GB) and scFoundation (4.2 GB), with a parameter count (85 M) in the middle of the field. The spatial formulation therefore turns a frozen general-purpose embedding into one that can be applied to atlas-scale data on a single commodity GPU.

The zero-shot readout is likewise insensitive to its few inference-time choices. Pooling the patch tokens by their mean yielded a more discriminative embedding than the prepended class token across the held-out atlases (present-class macro-F1 0.498 versus 0.471, higher in seven of eight atlases), which is why the mean-pooled embedding is used throughout. Annotation accuracy varied smoothly with the nearest-neighbour probe size and was highest at small neighbourhoods, decreasing monotonically as k grew (for example on ovary from macro-F1 0.74 at $k=5$ to 0.40 at $k=100$); rather than tune k to maximize any method's score we fixed a single conservative neighbourhood ($k=20$) a priori for every method and atlas. Finally, re-rendering each scImage from raw counts reproduced the stored embedding almost exactly (median cosine similarity 1.00), so the representation is deterministic and reproducible from the released layout and weights.

The frozen embedding is robust to missing genes and degrades gracefully with sequencing depth

Real measurements are corrupted in ways a curated atlas is not: genes drop out and libraries are sequenced shallowly. We stressed the zero-shot annotator accordingly, corrupting only the held-out query cells of a cystic-fibrosis atlas (masking a growing fraction of each cell's genes, or downsampling its counts), while the labelled reference bank stayed clean, and re-ran the same 20-nearest-neighbour probe used throughout for every method. scVision re-rendered its scImage from the corrupted counts; the token foundation models scGPT, scFoundation and Geneformer re-tokenized the corrupted counts; and a raw highly-variable-gene baseline (HVG-KNN) read the corrupted genes directly (Fig. 11).

Against random gene dropout the methods split sharply. The three modern foundation models began statistically tied on the clean atlas (present-class macro-F1 ≈ 0.27), well above Geneformer (0.18) and the raw-gene baseline (0.11). As genes were

Table 1. Efficiency and scale. Trainable parameters, embedding dimension, end-to-end encoding throughput and peak GPU memory, all measured on a single NVIDIA RTX 2080 Ti (11 GB) for a fixed sample of 512 cells with the frozen encoder. Throughput is wall-clock from raw highly-variable-gene counts to the cell embedding and includes each model’s required preprocessing (scImage rendering for scVision; gene tokenization for the token models). Parameter counts are read from each released checkpoint’s state dict.

Model	Params (M)	Embed. dim	Throughput (cells/s)	Peak GPU mem (GB)
scVision	85.2	768	528	2.55
scGPT	51.3	512	1.8	1.36
Geneformer	104.4	768	14.2	8.60
scFoundation	119.4	3072	12.9	4.20

masked scVision separated as the clear leader: its score *rose* to 0.34 by 30% masking and held 0.33 even with 70% of genes removed, the only method still above its clean score at this masking, whereas scGPT and scFoundation stayed near their clean level (0.24–0.26), Geneformer fell to 44% of its clean value (0.08 from 0.18) and the raw-gene baseline collapsed to 16% (0.02 from 0.11, Fig. 11a). Fixing each gene to an scImage coordinate makes masking punch holes in the image while leaving its global layout (and the encoder’s read of it) largely intact, a redundancy a flat expression vector lacks. Count downsampling tells a different and more pointed story we report plainly (Fig. 11b): here the rank-based token models were the more robust. scGPT led at every reduced depth, reaching 0.33 at one-fifth of counts and still 0.28 at one-tenth, essentially unaffected by depth; scVision held through half-depth (0.31) but degraded under heavier thinning, retaining 76% of its clean score at one-tenth depth (0.20), below scGPT (0.28) and scFoundation (0.23) though still above Geneformer (0.16) and the raw-gene baseline. This is what the representations predict: the scImage encodes expression as image intensity, so uniform downsampling dims the whole image and pushes it off the frozen encoder’s input distribution, whereas the token models’ rank-based tokenization is by construction near-invariant to total depth. The frozen spatial embedding is thus markedly robust to *which* genes are observed (where it leads every baseline) and degrades gracefully with *how deeply* they are sequenced, which localizes its one sensitivity to sequencing depth and points to count augmentation during rendering as a natural route to close the remaining gap.

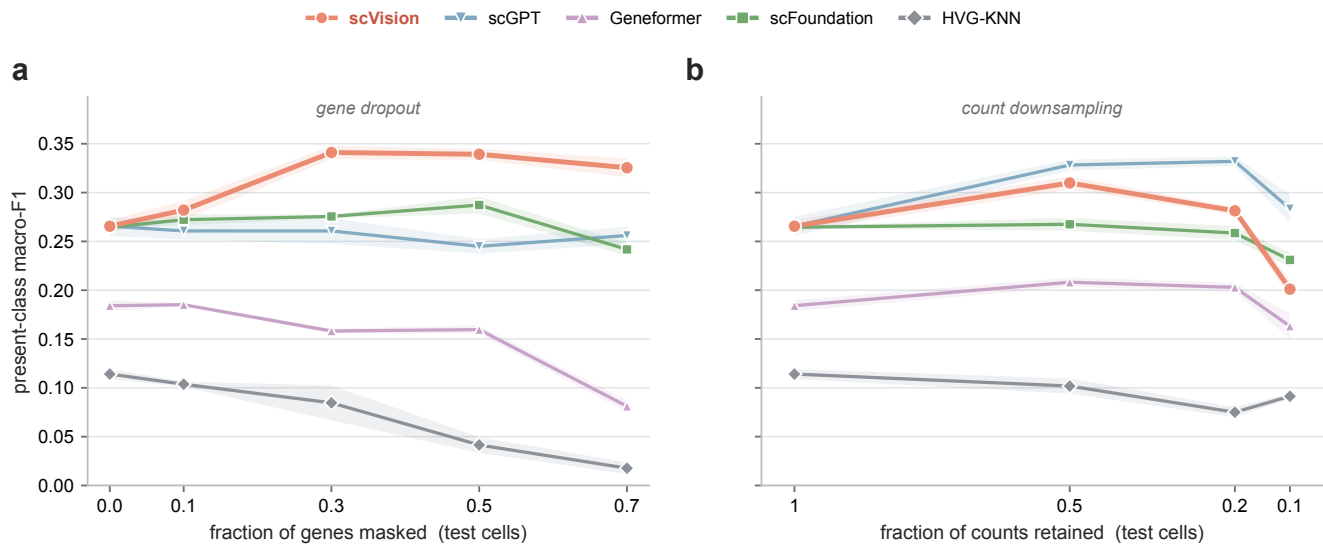


Figure 11. The frozen scVision embedding is robust to missing genes and degrades gracefully with sequencing depth. Held-out query cells of a cystic-fibrosis atlas are corrupted while the labelled reference bank stays clean; present-class macro-F1 of the shared 20-nearest-neighbour probe is plotted for all five methods (mean over three folds; shaded band = ± 1 s.d.). **a**, Random gene dropout (a fraction of each query cell’s genes is masked): scVision (coral), scGPT and scFoundation start tied (≈ 0.27), but scVision separates as the clear leader as masking grows, rising to 0.34 and holding 0.33 at 70% masked, the only method still above its clean-atlas score at this masking, while Geneformer drops to 44% of its clean value and the raw highly-variable-gene baseline (HVG-KNN) falls to 16%. **b**, Count downsampling (counts thinned to the indicated fraction; axis reversed, depth decreasing rightward): the rank-based token models are the more robust, with scGPT leading at every reduced depth; scVision holds to half-depth but retains 76% at one-tenth depth (0.20, below scGPT 0.28 and scFoundation 0.23), where the scImage’s intensity-encoded image moves off the frozen encoder’s input distribution.

Discussion

scVision frames single-cell representation learning as a vision problem. Instead of treating a cell as an unordered set of gene tokens, scVision places each gene at a fixed position using optimal transport. Genes that are often co-expressed are placed close together, so each transcriptome can be represented as an image. A vision transformer is then pretrained by masked image modelling on 72 million human cells to produce a frozen cell representation.

In a strict study-level holdout setting, where entire studies are withheld from training, this frozen representation performed strongly across all six external atlases tested: kidney cortex, ovary, focal cortical dysplasia cortex, Crohn’s-disease ileum, human retina and a multi-organ single-nucleus reference. scVision achieved the highest balanced accuracy and present-class macro-F1 on each atlas. It was also highly label-efficient. In many cases, one labelled cell per type with scVision performed better than other foundation models given many more labelled examples. When another foundation model was stronger at the smallest label budget, scVision often improved more as additional labels were added.

Several controls show that this advantage comes from the spatial gene layout. When the gene-to-pixel arrangement was randomly permuted before encoding, performance dropped substantially. This loss was larger than the drop caused by removing the vision encoder and applying a probe directly to the raw scImage. The learned spatial arrangement of genes is therefore not a cosmetic transformation; it carries much of the useful signal. The model benefits from the fact that genes with related expression patterns are placed near one another.

The results also show that scVision’s advantage is mainly in cross-study transfer. Within a single held-out tissue, some token-based foundation models can organize cells into locally clean neighbourhoods. However, those neighbourhoods do not always transfer well when cells must be assigned to an external reference bank built from other studies. In other words, an embedding can look well organized within one dataset but still fail across the study gap. scVision appears to preserve cell-type structure more reliably across studies that the model never saw during training.

Representing a cell as an image also makes the model easier to interpret. Because co-expressed genes occupy nearby positions, attention maps can be mapped back to genes and read as gene-expression programs. Without any pathway supervision, scVision recovered programs that matched known biology, including proximal-tubule injury programs in kidney, an unfolded-protein-response program in ovarian pericytes, and a TREM2–DAP12 inflamed-stroma program in Crohn’s disease ileum.

Table 2. List of notation. Symbols used throughout the Methods, grouped by the scImage construction (left) and the vision transformer and its zero-shot evaluation (right).

Symbol	Description	Symbol	Description
G	Number of selected genes	P	Patch side length
M	Number of lattice pixels	N	Number of patch tokens
$\mathbf{X}$	scImage of a cell	d	Token embedding dimension
$\mathbf{e}, \tilde{\mathbf{e}}$	Raw / z-scored expression	H	Number of attention heads
$\mathbf{D}^g$	Gene–gene distance matrix	d_h	Per-head width (d/H)
$\mathbf{D}^s$	Pixel–pixel distance matrix	$\mathbf{E}$	Patch-embedding projection
r_{ik}	Pearson correlation of genes i, k	$\mathbf{p}_j$	Positional embedding
$\mathbf{c}_j$	Coordinate of pixel j	$\mathbf{U}$	Token sequence
μ, ν	Gene / pixel marginals	$\mathbf{Q}, \mathbf{K}, \mathbf{V}$	Query, key, value
Π	Coupling polytope	f_θ, g_ϕ	Encoder / decoder
$\mathbf{T}^*$	Optimal coupling (gene layout)	$\mathcal{M}$	Masked-patch index set
$L(a, b)$	Kullback–Leibler loss	$\mathcal{L}_{\text{MAE}}$	Reconstruction loss
ε	Entropic-regularization weight	$\mathbf{h}$	Cell embedding
$\mathbf{P}$	Soft projection matrix	$\alpha_j^{(h)}$	Class-token $\rightarrow$ patch attention
$i, k / j, l$	Gene / pixel indices	$a_j^{(c)}$	Cell-type attention profile

These programs were not limited to one tissue. Across 585 matched cell-type pairs from 28 atlases, attention-derived programs from the same cell type overlapped 6.4-fold more than expected by chance. A four-gene myeloid program, IRF8, PIM1, TRAF1 and CCND2, appeared in microglia, kidney macrophages, breast macrophages and blood dendritic cells. The attention maps capture reusable cell-type programs across organs.

These findings address an important issue in single-cell foundation models. Previous evaluations have found that token-based models can struggle to outperform regularized logistic regression for cell-type annotation without fine-tuning^{23–25}. Our results suggest that the limitation may not be only model scale or training-set size. Instead, the way the transcriptome is represented may be critical. By giving genes biologically meaningful positions, scVision improves transfer under a stricter zero-shot evaluation. This image-based formulation also enables operations that are difficult in token models. For example, one can mask a spatial neighbourhood of genes, perturbing a co-regulated module at once, and then measure how the cell embedding changes. This provides a natural way to perform in silico perturbation at the level of gene programs rather than individual genes.

We emphasize balanced accuracy and present-class macro-F1 throughout because these metrics measure the quality of the actual cell-type calls and give weight to rare but biologically important populations, rather than being dominated by abundant classes. Under this readout scVision is the most accurate annotator on every held-out atlas, though on the Crohn’s-disease ileum and the multi-organ reference its margin over the strongest classical baselines is consistent but modest.

This study also has several limitations. All evaluations used frozen, zero-shot embeddings, which is the right setting for measuring representation quality but does not test the best possible performance after fine-tuning. The training data are human, so performance in other species remains to be tested. The current scImage uses one pan-tissue gene layout shared across all cells, which may not capture every tissue-specific relationship. Finally, attention-derived pathway programs provide useful biological clues, but they should not be treated as calibrated regulatory inference.

By turning each cell into an image, scVision brings computer-vision methods into single-cell biology. This opens several future directions. Fine-tuning the frozen backbone may improve performance when labeled data are available. Masking spatial gene neighbourhoods could support systematic in silico perturbation studies. The image formulation may also extend naturally to spatial transcriptomics and paired modalities such as surface protein or chromatin accessibility. Overall, these results suggest that progress in single-cell foundation models may come not only from larger datasets or larger models, but also from representations that better encode the relational structure of the transcriptome. Giving genes biologically meaningful positions is one such representation, and it makes the cell accessible to the tools of vision.

Methods

Training data and quality control

We assembled the pretraining dataset from the CZ CELLxGENE Census³⁴, drawing primary human cells through the Census interface and keeping only records marked as primary data, so that a cell deposited in several studies is counted once. These

primary cells span the major published human atlases aggregated by the Census, including pan-tissue references^{6,9} and organ- and disease-focused atlases of lung⁸, heart¹⁰, breast^{35,36}, hypothalamus¹², gut¹¹, blood^{13,14,17}, kidney³⁷, cerebellum³⁸ and fallopian tube³⁹. Cells were filtered by transcriptome-wide quality control (at least 200 detected genes, at least 500 total counts and no more than 20% mitochondrial content) and targeted-panel assays (BD Rhapsody Targeted, 10x Targeted Gene Expression) were excluded. The resulting dataset comprises approximately 94 million cells. Data were handled throughout with the scanpy and anndata single-cell ecosystem^{40,41}.

To measure generalization to new studies rather than to new cells of seen studies, we used a study-level holdout, the cross-study holdout denoted paradigm A in the figures: roughly 10% of cells, drawn from approximately 25 entirely held-out studies, form the external test pool, while the remaining studies are partitioned into training and validation by donor, so that no donor crosses the split. All splits use a fixed seed (42). The training split comprises 72 million cells.

scImage construction by optimal transport

Table 2 collects the notation used below. Each cell is rendered as a single-channel image $\mathbf{X} \in \mathbb{R}^{1 \times 104 \times 104}$, one gene per pixel (Fig. 1), so that the lattice carries $M = 104 \times 104 = 10,816$ pixels and we select exactly $G = M$ genes to fill it. The genes are the G most highly variable genes under a Seurat-style criterion, ranking each gene by its variance after a loess fit of variance against mean expression, computed in a single chunked pass over the dataset.

We place genes on the lattice by Gromov–Wasserstein optimal transport^{42,43}, which aligns the relational geometry of genes to that of pixels without requiring the two spaces to be compared point by point. Co-expression defines a gene–gene distance matrix $\mathbf{D}^g \in \mathbb{R}^{G \times G}$ with entries the correlation distance $D_{ik}^g = 1 - r_{ik}$, where r_{ik} is the Pearson correlation of genes i and k across cells; the lattice defines a pixel–pixel distance matrix $\mathbf{D}^s \in \mathbb{R}^{M \times M}$ with $D_{jl}^s = \|\mathbf{c}_j - \mathbf{c}_l\|_2$ the Euclidean distance between the two-dimensional coordinates $\mathbf{c}_j, \mathbf{c}_l$ of pixels j and l . With uniform marginals $\mu = \mathbf{1}_G/G$ over genes and $\nu = \mathbf{1}_M/M$ over pixels, and the coupling polytope

$$\Pi(\mu, \nu) \stackrel{\text{def}}{=} \{ \mathbf{T} \in \mathbb{R}_{\geq 0}^{G \times M} : \mathbf{T} \mathbf{1}_M = \mu, \mathbf{T}^\top \mathbf{1}_G = \nu \}, \quad (1)$$

the layout is the coupling that best matches the two geometries,

$$\mathbf{T}^* = \arg \min_{\mathbf{T} \in \Pi(\mu, \nu)} \sum_{i,j,k,l} L(D_{ik}^g, D_{jl}^s) T_{ij} T_{kl}, \quad (2)$$

penalizing pairs of genes whose co-expression distance is poorly preserved by the spatial distance of the pixels they are sent to. The discrepancy is the Kullback–Leibler loss $L(a, b) = \text{KL}(a \| b) \stackrel{\text{def}}{=} a \log(a/b) - a + b$, solved with no entropic regularization ($\epsilon = 0$)⁴⁴ for 200 iterations on a sample of 100,000 cells stratified by tissue (seed 42). Optimal transport has become a workhorse for single-cell analysis, from trajectory inference²⁶ to multi-omic alignment⁴⁵ and image-based embeddings of single-cell data⁴⁶; here it instead fixes a single shared layout, computed once and applied pan-tissue to every cell.

The coupling is retained as a dense, soft projection matrix $\mathbf{P} = M \mathbf{T}^* \in \mathbb{R}^{G \times M}$, the plan scaled by the number of pixels, with no Hungarian assignment or hard argmax, so that each gene spreads a smoothly distributed mass over the lattice rather than collapsing onto a single pixel. A cell with raw expression $\mathbf{e} \in \mathbb{R}^G$ is first z -scored gene-wise with the layout's stored mean μ_i^g and standard deviation σ_i^g , $\tilde{e}_i = (e_i - \mu_i^g)/\sigma_i^g$, then projected and reshaped,

$$\mathbf{y} = \mathbf{P}^\top \tilde{\mathbf{e}} \in \mathbb{R}^M, \quad \mathbf{X} = \text{reshape}_{104 \times 104}(\mathbf{y}), \quad (3)$$

with the reshape taken in column-major (Fortran) order and the image stored at half precision as a $(1, 104, 104)$ tensor. Because co-expressed genes occupy neighbouring pixels, co-regulation appears as local texture and cellular identity as global pattern. The scImages so produced are reproducible within a cell type and distinct across types (Supplementary Fig. S1).

scVision architecture and masked-image pretraining

scVision is a vision transformer trained as a masked autoencoder^{47,48}, one of several self-supervised vision paradigms that also include contrastive and self-distillation objectives^{49–51} and alternative hierarchical backbones⁵². The scImage $\mathbf{X} \in \mathbb{R}^{1 \times 104 \times 104}$ is partitioned into non-overlapping $P \times P = 8 \times 8$ patches, which tile it as a 13×13 grid of $N = 169$ tokens, each patch flattening $P^2 = 64$ pixels, equivalently 64 genes. Writing $\mathbf{x}_p^{(j)} \in \mathbb{R}^{P^2}$ for the flattened pixels of patch j , each patch is linearly embedded and endowed with a position,

$$\mathbf{u}_j = \mathbf{E} \mathbf{x}_p^{(j)} + \mathbf{p}_j \in \mathbb{R}^d, \quad j = 1, \dots, N, \quad (4)$$

with a shared projection $\mathbf{E} \in \mathbb{R}^{d \times P^2}$ to dimension $d = 768$ and a learnable positional embedding $\mathbf{p}_j$ encoding the patch's location on the 13×13 grid; prepending a learnable class token $\mathbf{u}_0$ forms the input sequence $\mathbf{U}^{(0)} = [\mathbf{u}_0, \mathbf{u}_1, \dots, \mathbf{u}_N]^\top \in \mathbb{R}^{(N+1) \times d}$. The

encoder f_θ is a ViT-base network ($d = 768$, with $L = 12$ transformer blocks, $H = 12$ attention heads and an MLP ratio of 4) for approximately 86 million parameters⁵³. Each block updates the sequence by pre-normalized multi-head self-attention and a token-wise MLP. For head h , with per-head width $d_h = d/H = 64$, the queries, keys and values are the linear projections $\mathbf{Q}_h = \mathbf{U}\mathbf{W}_h^Q$, $\mathbf{K}_h = \mathbf{U}\mathbf{W}_h^K$ and $\mathbf{V}_h = \mathbf{U}\mathbf{W}_h^V$, and tokens are mixed by scaled dot-product attention,

$$\text{Attn}_h(\mathbf{U}) = \text{softmax}\left(\frac{\mathbf{Q}_h\mathbf{K}_h^\top}{\sqrt{d_h}}\right)\mathbf{V}_h. \quad (5)$$

The H heads are concatenated and projected, $\text{MHSA}(\mathbf{U}) = [\text{Attn}_1(\mathbf{U}), \dots, \text{Attn}_H(\mathbf{U})]\mathbf{W}^O$, and, with layer normalization LN and residual connections, a block maps

$$\begin{aligned} \mathbf{U}' &= \mathbf{U} + \text{MHSA}(\text{LN}(\mathbf{U})), \\ \mathbf{U}'' &= \mathbf{U}' + \text{MLP}(\text{LN}(\mathbf{U}')), \end{aligned} \quad (6)$$

so that the encoder f_θ is the composition of $L = 12$ such blocks.

At each step a random subset $\mathcal{M} \subset \{1, \dots, N\}$ of patches, of size $\lfloor 0.75N \rfloor$, is masked; only the visible patches are encoded, and a lightweight decoder g_ϕ (512-dimensional, with 8 blocks and 16 heads, approximately 14 million parameters) reconstructs the masked patches from the encoded visible tokens together with learned mask tokens. Writing $\mathbf{z}_j$ for the per-patch-normalized pixel target of patch j and $\hat{\mathbf{z}}_j$ for its reconstruction, the objective is the mean squared error over masked patches only,

$$\mathcal{L}_{\text{MAE}}(\theta, \phi) = \frac{1}{|\mathcal{M}|} \sum_{j \in \mathcal{M}} \|\hat{\mathbf{z}}_j - \mathbf{z}_j\|_2^2. \quad (7)$$

The model was optimized with AdamW ($\beta_1 = 0.9$, $\beta_2 = 0.95$), a cosine learning-rate schedule with linear warmup, gradient clipping at a norm of 5 and half-precision mixed-precision training, on the 72-million-cell training split. The released backbone is the ViT-base, patch-8 model used for all evaluations.

Zero-shot evaluation and baselines

All evaluations use the frozen pretrained encoder f_{θ^*} with no fine-tuning. Writing $\mathbf{t}_1, \dots, \mathbf{t}_N$ for the encoded patch tokens of a cell, its embedding is their mean, L2-normalized,

$$\mathbf{h} = \frac{\bar{\mathbf{t}}}{\|\bar{\mathbf{t}}\|_2}, \quad \bar{\mathbf{t}} = \frac{1}{N} \sum_{j=1}^N \mathbf{t}_j \in \mathbb{R}^{768}. \quad (8)$$

Cell-type annotation is a 20-nearest-neighbour probe in cosine space against a labelled reference bank, following the reference-based annotation paradigm of established label-transfer tools^{54–56}: with unit-norm embeddings the cosine similarity is the inner product $s(\mathbf{h}, \mathbf{h}') = \mathbf{h}^\top \mathbf{h}'$, and the predicted label aggregates similarity-weighted votes over the set $\mathcal{N}_{20}(\mathbf{h})$ of 20 nearest reference cells,

$$\hat{y} = \arg \max_c \sum_{r \in \mathcal{N}_{20}(\mathbf{h})} s(\mathbf{h}, \mathbf{h}_r) \mathbb{I}[y_r = c]. \quad (9)$$

We report balanced accuracy, the mean per-class recall $\text{BA} = |\mathcal{C}|^{-1} \sum_{c \in \mathcal{C}} \text{TP}_c / (\text{TP}_c + \text{FN}_c)$, and present-class macro-F1, the macro average of the F_1 score over only the cell types $\mathcal{C}_p$ actually present in the held-out atlas,

$$\text{F1}_p = \frac{1}{|\mathcal{C}_p|} \sum_{c \in \mathcal{C}_p} \frac{2\text{Prec}_c \text{Rec}_c}{\text{Prec}_c + \text{Rec}_c}, \quad (10)$$

and, where stated, threshold-free macro-average precision and area under the ROC curve.

Label efficiency was measured by k -shot evaluation at $k \in \{1, 5, 10, 50\}$ labelled cells per type under five-fold cross-validation; the foundation models share the identical nearest-neighbour probe, so that differences isolate representation quality, while each classical baseline is additionally allowed to fit its own native classifier on the same support. Multi-study integration was assessed on eight held-out atlases: for each we embedded a balanced $\sim 20,000$ -cell pool spanning multiple studies and scored batch correction and biological conservation with the scIB v2 panel³⁰, the standard framework for ranking the many available single-cell integration methods^{57–64}, comparing the frozen scVision embedding against the token foundation models scGPT and Geneformer, the classical baselines Harmony²⁸ and principal-component analysis, and the deep-learning integration

model scVI^{29,65}, all read from the matched input. Unlike the foundation-model embeddings, which are read frozen, scVI was trained separately on each atlas's raw counts in the same 10,816-gene space with the dataset identity supplied as its only batch covariate (a negative-binomial variational autoencoder, 30 latent dimensions, two hidden layers, up to 400 epochs with early stopping); its latent posterior mean served as the integrated embedding. Bio-conservation is the mean of NMI, ARI, cell-type and isolated-label silhouette and cLISI; batch correction the mean of batch silhouette, iLISI, graph connectivity and PCR; the total weights them 0.6/0.4. Across-tissue confidence intervals are from 10,000 bootstrap resamples and paired method comparisons use the Wilcoxon signed-rank test over the eight tissues; a supplementary post-hoc kBET acceptance test (pure-Python k -nearest-neighbour χ^2 batch test, $k_0 = 25$) reports per-method batch mixing.

We compared against three single-cell foundation models (Geneformer¹⁹, scGPT²⁰ and scFoundation²¹), each encoded under the matched protocol, and against classical and tabular learners on the raw highly-variable-gene space: regularized logistic regression, a multilayer perceptron, k -nearest neighbours, a correlation classifier, a random forest⁶⁶, gradient-boosted trees⁶⁷ and TabNet⁶⁸, implemented with scikit-learn⁶⁹.

Gene-program inference and cross-tissue recurrence

To read gene-expression programs from the model, we used the attention that the class token pays to each patch in the final transformer block. Writing $\mathbf{q}_0$ for the class-token query and $\mathbf{k}_j$ for the key of patch j in head h of that block, this attention is the softmax

$$\alpha_j^{(h)} = \frac{\exp(\mathbf{q}_0^\top \mathbf{k}_j / \sqrt{d_h})}{\sum_{j'=1}^N \exp(\mathbf{q}_0^\top \mathbf{k}_{j'} / \sqrt{d_h})}, \quad (11)$$

and the patch profile $a_j^{(c)}$ of cell type c averages $\alpha_j^{(h)}$ over the H heads and over the cells of type c . Programs were made cell-type-specific by subtracting the per-patch mean across cell types,

$$\tilde{a}_j^{(c)} = a_j^{(c)} - \frac{1}{|\mathcal{C}|} \sum_{c' \in \mathcal{C}} a_j^{(c')}, \quad (12)$$

which removes the shared attention-sink patches. The eight patches of largest $\tilde{a}_j^{(c)}$ (each an 8×8 , 64-gene spatial neighbourhood of the scImage) define the attended program, from which the top 100 genes, in descending attention order, were mapped back through the layout and tested for pathway enrichment by the hypergeometric test against MSigDB Hallmark^{70,71} and, per atlas, KEGG, Reactome and GO biological process, with Benjamini–Hochberg control of the false-discovery rate. To confirm these programs were not artefacts of the per-patch mean-subtraction, we recomputed them at four cell-resampling seeds under two alternative specificity definitions, the uncentred raw profile $a_j^{(c)}$ and a z-score dividing $\tilde{a}_j^{(c)}$ by its per-patch standard deviation across cell types. Cross-cell-type program overlap fell monotonically with centring stringency (mean off-diagonal Jaccard 0.27, 0.047 and 0.003 for the raw, mean-subtracted and z-scored variants), and the microglia–endothelium program identity (Jaccard 1.0) and its p53 enrichment (KEGG $q = 0.020$, identical leading genes) were recovered under both the raw and mean-subtracted variants but removed by the z-score, which by construction discards patches shared across cell types, indicating that this shared vascular–immune module reflects genuine co-attention rather than the chosen specificity metric. These programs were likewise stable to the number of cells sampled per type: halving or doubling the published budget of 200 cells per type (to 100 or 400) left the mean off-diagonal Jaccard essentially unchanged (0.046, 0.044 and 0.054), and both the endothelial p53 anchor (KEGG $q = 0.020$, the same five leading genes) and oligodendrocyte-precursor RET signalling (Reactome $q \approx 0.002$, the same five genes) were recovered at every budget; only the microglial arm of the vascular–immune module weakened at the smallest budget, its overlap with endothelium falling to Jaccard 0.47 and its p53 enrichment to non-significance at 100 cells, independently mirroring the cell-resampling hierarchy above, in which the endothelial p53 program is the most reproducible and microglia the most sampling-sensitive member of the module. The published top-100 program was itself unchanged for attended-patch budgets of 6, 8 and 12 patches, as expected since the 100 highest-attention genes are drawn from the two most-attended patches, which are selected independently of the total budget.

To test whether the same programs recur across tissues, we matched cell types appearing in more than one of 28 atlases (585 cross-tissue pairs over 77 cell types) and measured the Jaccard overlap $J(\mathcal{G}_1, \mathcal{G}_2) = |\mathcal{G}_1 \cap \mathcal{G}_2| / |\mathcal{G}_1 \cup \mathcal{G}_2|$ of their top-100 gene programs $\mathcal{G}_1, \mathcal{G}_2$ against a null in which cell-type labels are randomly deranged (1000 permutations, seed 20240601).

In silico perturbation and the disease axis

The spatial representation supports an in silico perturbation native to its geometry, complementing supervised perturbation-response models that learn cellular responses from labelled perturbation data^{72–74}. For a target gene we masked the single 8×8 patch (64 genes) containing its co-expression neighbourhood in the scImages of control cells, re-encoded them and measured the mean shift in embedding, $\Delta_{\text{pred}} = \mathbf{h}_{\text{mask}} - \mathbf{h}_{\text{ctrl}}$. We compared its cosine similarity $\cos(\Delta_{\text{pred}}, \Delta_{\text{obs}})$ to the measured

perturbation direction Δ_{obs} from large-scale CRISPR perturbation screens^{31–33} against a null in which a random patch is masked instead, by a paired Wilcoxon test.

To probe disease, we fit a logistic-regression direction $\mathbf{w} \in \mathbb{R}^{768}$ in the frozen embedding separating disease from control cells, scoring a cell by $\mathbf{w}^T \mathbf{h}$, and evaluated whether this axis transfers across assay and disease subtype, from dilated to hypertrophic cardiomyopathy^{15,16,75}, scoring the transferred axis by area under the ROC curve.

Statistical analysis

Group comparisons used the tests named with each analysis: a paired Wilcoxon test for the perturbation null, hypergeometric tests with Benjamini–Hochberg correction for pathway enrichment, and permutation tests for cross-tissue recurrence. Cross-study annotation was evaluated over five folds, and every quantitative value reported in the Results derives from a stored evaluation record.

Data availability

All datasets analysed in this study are publicly available. The pretraining dataset and every held-out evaluation atlas were obtained from the CZ CELLxGENE Census of human single-cell transcriptomes³⁴, itself assembled from public tissue atlases^{6,8,9}; the held-out atlases, their tissues and their cell-type resolutions are listed in Supplementary Table S1. The large-scale CRISPR perturbation screens used for the in silico perturbation analysis are the public Perturb-seq datasets^{31–33} cited in the Methods. All processed expression matrices, the shared scImage layout and the intermediate evaluation records required to reproduce the analyses will be released with the code (see Code availability); detailed preprocessing and rendering procedures are described in the Methods. No new human or animal data were generated for this study.

Code availability

The code for scVision, including data preprocessing, scImage rendering, masked-image pretraining and the zero-shot evaluation pipelines, is reserved for now and will be released publicly, with a persistent identifier, upon publication. Further information and updates are available at the project page, <https://islamlab.org/scvision>.

References

1. Tanay, A. & Regev, A. Scaling single-cell genomics from phenomenology to mechanism. *Nature* **541**, 331–338, DOI: [10.1038/nature21350](https://doi.org/10.1038/nature21350) (2017).
2. Macosko, E. Z., Basu, A., Satija, R. *et al.* Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* **161**, 1202–1214, DOI: [10.1016/j.cell.2015.05.002](https://doi.org/10.1016/j.cell.2015.05.002) (2015).
3. Klein, A. M., Mazutis, L., Akartuna, I. *et al.* Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell* **161**, 1187–1201, DOI: [10.1016/j.cell.2015.04.044](https://doi.org/10.1016/j.cell.2015.04.044) (2015).
4. Cao, J., Spielmann, M., Qiu, X. *et al.* The single-cell transcriptional landscape of mammalian organogenesis. *Nature* **566**, 496–502, DOI: [10.1038/s41586-019-0969-x](https://doi.org/10.1038/s41586-019-0969-x) (2019).
5. The Tabula Muris Consortium. Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. *Nature* **562**, 367–372, DOI: [10.1038/s41586-018-0590-4](https://doi.org/10.1038/s41586-018-0590-4) (2018).
6. The Tabula Sapiens Consortium. The Tabula Sapiens: A multiple-organ, single-cell transcriptomic atlas of humans. *Science* **376**, eabl4896, DOI: [10.1126/science.abl4896](https://doi.org/10.1126/science.abl4896) (2022).
7. Tirosh, I., Izar, B., Prakadan, S. M. *et al.* Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science* **352**, 189–196, DOI: [10.1126/science.aad0501](https://doi.org/10.1126/science.aad0501) (2016).
8. Sikkema, L., Ramírez-Suástegui, C., Strobl, D. C. *et al.* An integrated cell atlas of the lung in health and disease. *Nat. Medicine* **29**, 1563–1577, DOI: [10.1038/s41591-023-02327-2](https://doi.org/10.1038/s41591-023-02327-2) (2023).
9. Eraslan, G., Drokhyansky, E., Anand, S. *et al.* Single-nucleus cross-tissue molecular reference maps toward understanding disease gene function. *Science* **376**, eabl4290, DOI: [10.1126/science.abl4290](https://doi.org/10.1126/science.abl4290) (2022).
10. Litviňuková, M., Talavera-López, C., Maatz, H. *et al.* Cells of the adult human heart. *Nature* **588**, 466–472, DOI: [10.1038/s41586-020-2797-4](https://doi.org/10.1038/s41586-020-2797-4) (2020).
11. Elmentaite, R., Kumasaka, N., Roberts, K. *et al.* Cells of the human intestinal tract mapped across space and time. *Nature* **597**, 250–255, DOI: [10.1038/s41586-021-03852-1](https://doi.org/10.1038/s41586-021-03852-1) (2021).

12. Tadross, J. A., Steuernagel, L., Dowsett, G. K. C. *et al.* A comprehensive spatio-cellular map of the human hypothalamus. *Nature* **639**, 708–716, DOI: [10.1038/s41586-024-08504-8](https://doi.org/10.1038/s41586-024-08504-8) (2025).
13. Stephenson, E., Reynolds, G., Botting, R. A. *et al.* Single-cell multi-omics analysis of the immune response in COVID-19. *Nat. Medicine* **27**, 904–916, DOI: [10.1038/s41591-021-01329-2](https://doi.org/10.1038/s41591-021-01329-2) (2021).
14. COvid-19 Multi-omics Blood ATlas (COMBAT) Consortium. A blood atlas of COVID-19 defines hallmarks of disease severity and specificity. *Cell* **185**, 916–938, DOI: [10.1016/j.cell.2022.01.012](https://doi.org/10.1016/j.cell.2022.01.012) (2022).
15. Chaffin, M., Papangelis, I., Simonson, B. *et al.* Single-nucleus profiling of human dilated and hypertrophic cardiomyopathy. *Nature* **608**, 174–180, DOI: [10.1038/s41586-022-04817-8](https://doi.org/10.1038/s41586-022-04817-8) (2022).
16. Reichart, D., Lindberg, E. L., Maatz, H. *et al.* Pathogenic variants damage cell composition and single cell transcription in cardiomyopathies. *Science* **377**, eabo1984, DOI: [10.1126/science.abo1984](https://doi.org/10.1126/science.abo1984) (2022).
17. Yazar, S., Alquicira-Hernandez, J., Wing, K. *et al.* Single-cell eQTL mapping identifies cell type-specific genetic control of autoimmune disease. *Science* **376**, eabf3041, DOI: [10.1126/science.abf3041](https://doi.org/10.1126/science.abf3041) (2022).
18. Yang, F., Wang, W., Wang, F. *et al.* scBERT as a large-scale pretrained deep language model for cell type annotation of single-cell RNA-seq data. *Nat. Mach. Intell.* **4**, 852–866, DOI: [10.1038/s42256-022-00534-z](https://doi.org/10.1038/s42256-022-00534-z) (2022).
19. Theodoris, C. V., Xiao, L., Chopra, A. *et al.* Transfer learning enables predictions in network biology. *Nature* **618**, 616–624, DOI: [10.1038/s41586-023-06139-9](https://doi.org/10.1038/s41586-023-06139-9) (2023).
20. Cui, H., Wang, C., Maan, H. *et al.* scGPT: toward building a foundation model for single-cell multi-omics using generative AI. *Nat. Methods* **21**, 1470–1480, DOI: [10.1038/s41592-024-02201-0](https://doi.org/10.1038/s41592-024-02201-0) (2024).
21. Hao, M., Gong, J., Zeng, X. *et al.* Large-scale foundation model on single-cell transcriptomics. *Nat. Methods* **21**, 1481–1491, DOI: [10.1038/s41592-024-02305-7](https://doi.org/10.1038/s41592-024-02305-7) (2024).
22. Rosen, Y. *et al.* Universal cell embedding provides a foundation model for cell biology. *Nature* DOI: [10.1038/s41586-026-10689-z](https://doi.org/10.1038/s41586-026-10689-z) (2026).
23. Boiarsky, R., Singh, N., Buendia, A., Getz, G. & Sontag, D. A deep dive into single-cell RNA sequencing foundation models. *bioRxiv* DOI: [10.1101/2023.10.19.563100](https://doi.org/10.1101/2023.10.19.563100) (2023). Preprint.
24. Kedzierska, K. Z., Crawford, L., Amini, A. P. & Lu, A. X. Zero-shot evaluation reveals limitations of single-cell foundation models. *Genome Biol.* **26**, 101, DOI: [10.1186/s13059-025-03574-x](https://doi.org/10.1186/s13059-025-03574-x) (2025).
25. Liu, T., Li, K., Wang, Y., Li, H. & Zhao, H. Evaluating the utilities of foundation models in single-cell data analysis. *Adv. Sci.* **13**, e14490, DOI: [10.1002/advs.202514490](https://doi.org/10.1002/advs.202514490) (2026).
26. Schiebinger, G., Shu, J., Tabaka, M. *et al.* Optimal-transport analysis of single-cell gene expression identifies developmental trajectories in reprogramming. *Cell* **176**, 928–943, DOI: [10.1016/j.cell.2019.01.006](https://doi.org/10.1016/j.cell.2019.01.006) (2019).
27. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444, DOI: [10.1038/nature14539](https://doi.org/10.1038/nature14539) (2015).
28. Korsunsky, I., Millard, N., Fan, J. *et al.* Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat. Methods* **16**, 1289–1296, DOI: [10.1038/s41592-019-0619-0](https://doi.org/10.1038/s41592-019-0619-0) (2019).
29. Lopez, R., Regier, J., Cole, M. B., Jordan, M. I. & Yosef, N. Deep generative modeling for single-cell transcriptomics. *Nat. Methods* **15**, 1053–1058, DOI: [10.1038/s41592-018-0229-2](https://doi.org/10.1038/s41592-018-0229-2) (2018).
30. Luecken, M. D., Büttner, M., Chaichoompu, K. *et al.* Benchmarking atlas-level data integration in single-cell genomics. *Nat. Methods* **19**, 41–50, DOI: [10.1038/s41592-021-01336-8](https://doi.org/10.1038/s41592-021-01336-8) (2022).
31. Replogle, J. M., Saunders, R. A., Pogson, A. N. *et al.* Mapping information-rich genotype-phenotype landscapes with genome-scale Perturb-seq. *Cell* **185**, 2559–2575, DOI: [10.1016/j.cell.2022.05.013](https://doi.org/10.1016/j.cell.2022.05.013) (2022).
32. Norman, T. M., Horlbeck, M. A., Replogle, J. M. *et al.* Exploring genetic interaction manifolds constructed from rich single-cell phenotypes. *Science* **365**, 786–793, DOI: [10.1126/science.aax4438](https://doi.org/10.1126/science.aax4438) (2019).
33. Adamson, B., Norman, T. M., Jost, M. *et al.* A multiplexed single-cell CRISPR screening platform enables systematic dissection of the unfolded protein response. *Cell* **167**, 1867–1882, DOI: [10.1016/j.cell.2016.11.048](https://doi.org/10.1016/j.cell.2016.11.048) (2016).
34. Abdulla, S., Aeversmann, B., Assis, P. *et al.* CZ CELLxGENE Discover: a single-cell data platform for scalable exploration, analysis and modeling of aggregated data. *Nucleic Acids Res.* **53**, D886–D900, DOI: [10.1093/nar/gkaf1142](https://doi.org/10.1093/nar/gkaf1142) (2025).
35. Kumar, T., Nee, K., Wei, R. *et al.* A spatially resolved single-cell genomic atlas of the adult human breast. *Nature* **620**, 181–191, DOI: [10.1038/s41586-023-06252-9](https://doi.org/10.1038/s41586-023-06252-9) (2023).

36. Reed, A. D., Pensa, S., Steif, A. *et al.* A single-cell atlas enables mapping of homeostatic cellular shifts in the adult human breast. *Nat. Genet.* **56**, 652–662, DOI: [10.1038/s41588-024-01688-9](https://doi.org/10.1038/s41588-024-01688-9) (2024).
37. Krishna, C., DiNatale, R. G., Kuo, F. *et al.* Single-cell sequencing links multiregional immune landscapes and tissue-resident T cells in ccRCC to tumor topology and therapy efficacy. *Cancer Cell* **39**, 662–677, DOI: [10.1016/j.ccell.2021.03.007](https://doi.org/10.1016/j.ccell.2021.03.007) (2021).
38. Aldinger, K. A., Thomson, Z., Phelps, I. G. *et al.* Spatial and cell type transcriptional landscape of human cerebellar development. *Nat. Neurosci.* **24**, 1163–1175, DOI: [10.1038/s41593-021-00872-y](https://doi.org/10.1038/s41593-021-00872-y) (2021).
39. Ulrich, N. D., Shen, Y.-c., Ma, Q. *et al.* Cellular heterogeneity of human fallopian tubes in normal and hydrosalpinx disease states identified using scRNA-seq. *Dev. Cell* **57**, 914–929, DOI: [10.1016/j.devcel.2022.02.017](https://doi.org/10.1016/j.devcel.2022.02.017) (2022).
40. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.* **19**, 15, DOI: [10.1186/s13059-017-1382-0](https://doi.org/10.1186/s13059-017-1382-0) (2018).
41. Virshup, I., Rybakov, S., Theis, F. J. *et al.* anndata: Access and store annotated data matrices. *J. Open Source Softw.* **9**, 4371, DOI: [10.21105/joss.04371](https://doi.org/10.21105/joss.04371) (2024).
42. Mémoli, F. Gromov-wasserstein distances and the metric approach to object matching. *Foundations Comput. Math.* **11**, 417–487, DOI: [10.1007/s10208-011-9093-5](https://doi.org/10.1007/s10208-011-9093-5) (2011).
43. Peyré, G. & Cuturi, M. Computational optimal transport: With applications to data science. *Foundations Trends Mach. Learn.* **11**, 355–607, DOI: [10.1561/22000000073](https://doi.org/10.1561/22000000073) (2019).
44. Cuturi, M. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems (NeurIPS)* (2013).
45. Klein, D., Palla, G., Lange, M. *et al.* Mapping cells through time and space with moscot. *Nature* **638**, 1065–1075, DOI: [10.1038/s41586-024-08453-2](https://doi.org/10.1038/s41586-024-08453-2) (2025).
46. Islam, M. T. & Xing, L. Cartography of genomic interactions enables deep analysis of single-cell expression data. *Nat. Commun.* **14**, 679, DOI: [10.1038/s41467-023-36383-6](https://doi.org/10.1038/s41467-023-36383-6) (2023).
47. He, K. *et al.* Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 16000–16009 (2022).
48. Dosovitskiy, A. *et al.* An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)* (2021).
49. Chen, T., Kornblith, S., Norouzi, M. & Hinton, G. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning (ICML)* (2020).
50. Caron, M., Touvron, H., Misra, I. *et al.* Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 9650–9660 (2021).
51. Oquab, M., Darcet, T., Moutakanni, T. *et al.* DINOv2: Learning robust visual features without supervision. *Transactions on Mach. Learn. Res.* (2024).
52. Liu, Z., Lin, Y., Cao, Y. *et al.* Swin Transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)* (2021).
53. Vaswani, A. *et al.* Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)* (2017).
54. Aran, D., Looney, A. P., Liu, L. *et al.* Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. *Nat. Immunol.* **20**, 163–172, DOI: [10.1038/s41590-018-0276-y](https://doi.org/10.1038/s41590-018-0276-y) (2019).
55. Domínguez Conde, C., Xu, C., Jarvis, L. B. *et al.* Cross-tissue immune cell analysis reveals tissue-specific features in humans. *Science* **376**, eabl5197, DOI: [10.1126/science.abl5197](https://doi.org/10.1126/science.abl5197) (2022).
56. Kiselev, V. Y., Yiu, A. & Hemberg, M. scmap: projection of single-cell RNA-seq data across data sets. *Nat. Methods* **15**, 359–362, DOI: [10.1038/nmeth.4644](https://doi.org/10.1038/nmeth.4644) (2018).
57. Stuart, T., Butler, A., Hoffman, P. *et al.* Comprehensive integration of single-cell data. *Cell* **177**, 1888–1902, DOI: [10.1016/j.cell.2019.05.031](https://doi.org/10.1016/j.cell.2019.05.031) (2019).
58. Butler, A., Hoffman, P., Smibert, P. *et al.* Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.* **36**, 411–420, DOI: [10.1038/nbt.4096](https://doi.org/10.1038/nbt.4096) (2018).
59. Haghverdi, L., Lun, A. T. L., Morgan, M. D. *et al.* Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nat. Biotechnol.* **36**, 421–427, DOI: [10.1038/nbt.4091](https://doi.org/10.1038/nbt.4091) (2018).

60. Hie, B., Bryson, B. & Berger, B. Efficient integration of heterogeneous single-cell transcriptomes using Scanorama. *Nat. Biotechnol.* **37**, 685–691, DOI: [10.1038/s41587-019-0113-3](https://doi.org/10.1038/s41587-019-0113-3) (2019).
61. Polański, K., Young, M. D., Miao, Z. *et al.* BBKNN: fast batch alignment of single cell transcriptomes. *Bioinformatics* **36**, 964–965, DOI: [10.1093/bioinformatics/btz625](https://doi.org/10.1093/bioinformatics/btz625) (2020).
62. Xu, C., Lopez, R., Mehlman, E. *et al.* Probabilistic harmonization and annotation of single-cell transcriptomics data with deep generative models. *Mol. Syst. Biol.* **17**, e9620, DOI: [10.15252/msb.20209620](https://doi.org/10.15252/msb.20209620) (2021).
63. Lotfollahi, M., Naghipourfar, M., Luecken, M. D. *et al.* Mapping single-cell data to reference atlases by transfer learning. *Nat. Biotechnol.* **40**, 121–130, DOI: [10.1038/s41587-021-01001-7](https://doi.org/10.1038/s41587-021-01001-7) (2022).
64. Tran, H. T. N., Ang, K. S., Chevrier, M. *et al.* A benchmark of batch-effect correction methods for single-cell RNA sequencing data. *Genome Biol.* **21**, 12, DOI: [10.1186/s13059-019-1850-9](https://doi.org/10.1186/s13059-019-1850-9) (2020).
65. Gayoso, A., Lopez, R., Xing, G. *et al.* A Python library for probabilistic analysis of single-cell omics data. *Nat. Biotechnol.* **40**, 163–166, DOI: [10.1038/s41587-021-01206-w](https://doi.org/10.1038/s41587-021-01206-w) (2022).
66. Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32, DOI: [10.1023/A:1010933404324](https://doi.org/10.1023/A:1010933404324) (2001).
67. Chen, T. & Guestrin, C. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 785–794, DOI: [10.1145/2939672.2939785](https://doi.org/10.1145/2939672.2939785) (2016).
68. Arik, S. O. & Pfister, T. TabNet: Attentive interpretable tabular learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, 6679–6687 (2021).
69. Pedregosa, F., Varoquaux, G., Gramfort, A. *et al.* Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
70. Liberzon, A. *et al.* The molecular signatures database hallmark gene set collection. *Cell Syst.* **1**, 417–425, DOI: [10.1016/j.cels.2015.12.004](https://doi.org/10.1016/j.cels.2015.12.004) (2015).
71. Subramanian, A., Tamayo, P., Mootha, V. K. *et al.* Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci.* **102**, 15545–15550, DOI: [10.1073/pnas.0506580102](https://doi.org/10.1073/pnas.0506580102) (2005).
72. Lotfollahi, M., Wolf, F. A. & Theis, F. J. scGen predicts single-cell perturbation responses. *Nat. Methods* **16**, 715–721, DOI: [10.1038/s41592-019-0494-8](https://doi.org/10.1038/s41592-019-0494-8) (2019).
73. Roohani, Y., Huang, K. & Leskovec, J. Predicting transcriptional outcomes of novel multigene perturbations with GEARS. *Nat. Biotechnol.* **42**, 927–935, DOI: [10.1038/s41587-023-01905-6](https://doi.org/10.1038/s41587-023-01905-6) (2024).
74. Lotfollahi, M., Klimovskaia, A., De Donno, C. *et al.* Predicting cellular responses to complex perturbations in high-throughput screens. *Mol. Syst. Biol.* **19**, e11517, DOI: [10.15252/msb.202211517](https://doi.org/10.15252/msb.202211517) (2023).
75. Koenig, A. L., Shchukina, I., Amrute, J. *et al.* Single-cell transcriptomics reveals cell-type-specific diversification in human heart failure. *Nat. Cardiovasc. Res.* **1**, 263–280, DOI: [10.1038/s44161-022-00028-6](https://doi.org/10.1038/s44161-022-00028-6) (2022).

Supplementary Information

This supplement reports the full per-atlas benchmark panels behind the headline numbers, the significance testing of the cross-study comparison, and the integration, gene-program and confusion analyses. Throughout, scVision is evaluated zero-shot under the study-level holdout (paradigm A); Δ denotes scVision’s frozen-probe score minus the best competing single-cell, classical or tabular baseline, and we report ties and losses alongside wins. Data files label the model TabVisionFM.

scImage construction and reproducibility

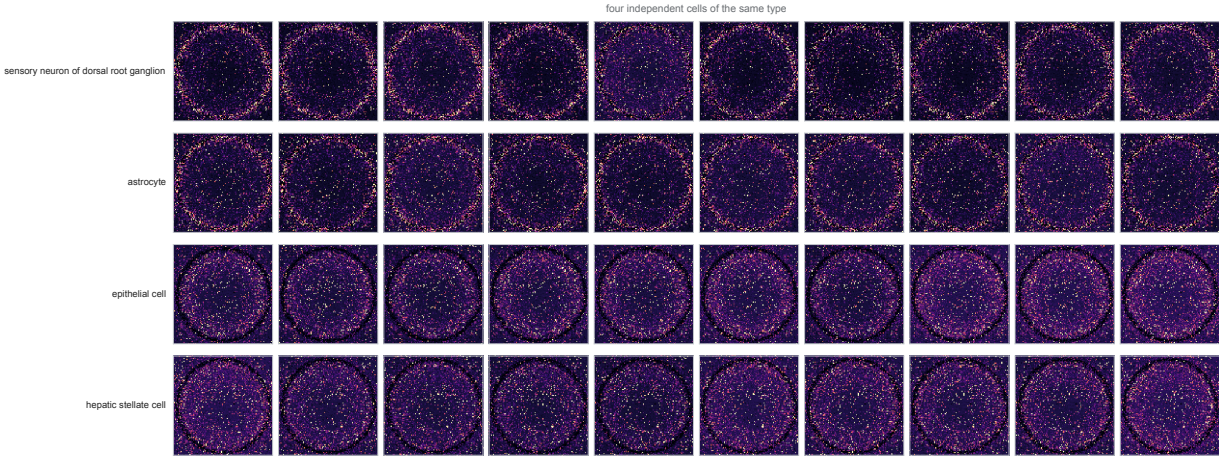


Figure S1. scImages are reproducible within a cell type and distinct across types. Each 104×104 image places the 10,816 most informative genes at optimal-transport coordinates, so co-expressed genes become spatial neighbours; intensity is the cell’s z -scored expression projected onto that fixed layout. Rows are cell types; columns are four independent cells of that type drawn from different tissues. The within-row similarity (same spatial texture) against the across-row contrast is what the vision encoder reads as identity. Intensity is clipped to the 2nd–98th percentile per panel for display.

Zero-shot cell-type annotation: full panels

Table S4. Head-to-head significance on cell-type identity (paradigm A, balanced accuracy). For each baseline, the per-fold balanced-accuracy difference (scVision’s best probe minus the baseline) was pooled across the four deeply-evaluated atlases (ovary, kidney cortex, focal cortical dysplasia, cystic fibrosis) and tested with a paired Wilcoxon signed-rank test; p values are Holm-adjusted across the eleven baselines. Win/tie/loss counts are per atlas. scVision wins every comparison in aggregate, but the counts are honest about the one loss — scFoundation takes cystic fibrosis. Per-atlas overall winners: Ovary: scVision (20-NN), Kidney cortex: scVision, Focal cort. dysplasia: scVision (20-NN), Cystic fibrosis: scFoundation.

Baseline	mean Δ	W/T/L	Wilcoxon p	Holm p	sig.
scGPT	+0.084	3/1/0	3.3×10^{-4}	6.6×10^{-4}	yes
scFoundation	+0.045	3/0/1	0.0013	0.0013	yes
Geneformer	+0.089	4/0/0	5.3×10^{-5}	1.6×10^{-4}	yes
MLP	+0.128	4/0/0	7.6×10^{-6}	8.4×10^{-5}	yes
LogReg	+0.114	4/0/0	7.6×10^{-6}	8.4×10^{-5}	yes
XGBoost	+0.137	4/0/0	1.5×10^{-5}	1.4×10^{-4}	yes
Random Forest	+0.191	4/0/0	1.5×10^{-5}	1.4×10^{-4}	yes
TabNet	+0.317	4/0/0	1.5×10^{-5}	1.4×10^{-4}	yes
HVG-kNN	+0.384	4/0/0	1.5×10^{-5}	1.4×10^{-4}	yes
Pearson	+0.382	4/0/0	1.5×10^{-5}	1.4×10^{-4}	yes

Table S1. Balanced accuracy under cross-study holdout (paradigm A), all probes and atlases. Five-fold mean balanced accuracy for scVision’s frozen 20-NN, logistic-regression and MLP probes against three single-cell foundation models and the classical/tabular baselines. Bold marks the row-best method; Δ is scVision minus the best of scGPT, scFoundation, logistic regression, MLP, XGBoost and random forest (it excludes TabNet, HVG-kNN and Pearson). scVision leads on most atlases, but the panel is honest about exceptions — scVision trails on the obstructive-nephropathy 20-NN block.

Atlas	scVision	scGPT	scFound.	Genef.	LogReg	MLP	XGB	RF	TabNet	HVG-kNN	Pears.	Δ
<i>20-NN probe</i>												
Retina	0.649	0.510	0.598	0.468	0.610	0.624	0.593	0.509	0.529	0.283	0.297	+0.025
Kidney cortex	0.468	0.377	0.399	0.153	0.397	0.331	0.302	0.300	0.096	0.157	0.027	+0.069
Ovary	0.830	0.778	0.793	0.589	0.701	0.668	0.668	0.542	0.516	0.097	0.305	+0.036
Obstructive neph.	0.482	0.334	0.358	0.143	0.503	0.475	0.440	0.455	0.120	0.221	0.087	-0.021
Focal cort. dys.	0.648	0.637	0.462	0.566	0.493	0.489	0.505	0.390	0.277	0.385	0.406	+0.011
scVision best ($\Delta > 0$): 4/5 ; mean $\Delta = +0.024$												
<i>Logistic-regression probe</i>												
Retina	0.680	0.538	0.664	0.680	0.610	0.624	0.593	0.509	0.529	0.283	0.297	+0.016
Kidney cortex	0.444	0.368	0.384	0.448	0.397	0.331	0.302	0.300	0.096	0.157	0.027	+0.047
Ovary	0.718	0.852	0.760	0.623	0.701	0.668	0.668	0.542	0.516	0.097	0.305	-0.134
Obstructive neph.	0.510	0.362	0.457	0.460	0.503	0.475	0.440	0.455	0.120	0.221	0.087	+0.007
Focal cort. dys.	0.584	0.502	0.525	0.528	0.493	0.489	0.505	0.390	0.277	0.385	0.406	+0.059
scVision best ($\Delta > 0$): 4/5 ; mean $\Delta = -0.001$												
<i>MLP probe</i>												
Retina	0.695	0.598	0.645	0.654	0.610	0.624	0.593	0.509	0.529	0.283	0.297	+0.050
Kidney cortex	0.470	0.416	0.368	0.447	0.397	0.331	0.302	0.300	0.096	0.157	0.027	+0.055
Ovary	0.822	0.761	0.725	0.643	0.701	0.668	0.668	0.542	0.516	0.097	0.305	+0.061
Obstructive neph.	0.542	0.453	0.476	0.493	0.503	0.475	0.440	0.455	0.120	0.221	0.087	+0.039
Focal cort. dys.	0.597	0.579	0.484	0.597	0.493	0.489	0.505	0.390	0.277	0.385	0.406	+0.018
scVision best ($\Delta > 0$): 5/5 ; mean $\Delta = +0.045$												

Table S2. Present-class macro-F1 under cross-study holdout (paradigm A). As Table S1 but for the present-class macro-F1 (averaged only over classes present in a fold, the harder and more honest score). Dashes mark baseline/atlas combinations that were not run at full supervision. Macro-specificity (one-vs-rest) was near ceiling for every method and atlas (range 0.978–1.000); it is omitted as a separate table.

Atlas	scVision	scGPT	scFound.	Genef.	LogReg	MLP	XGB	RF	TabNet	HVG-kNN	Pears.	Δ
<i>20-NN probe</i>												
Retina	0.712	0.565	0.636	0.525	0.688	0.683	0.641	0.549	0.579	0.353	0.331	+0.024
Kidney cortex	0.560	0.468	0.481	0.226	–	–	–	–	0.132	–	–	+0.079
Ovary	0.890	0.859	0.861	0.700	–	–	–	–	0.661	–	–	+0.029
Obstructive neph.	0.538	0.399	0.432	0.209	0.575	0.542	0.526	0.488	0.146	0.295	0.106	-0.037
Focal cort. dys.	0.714	0.706	0.532	0.643	–	–	–	–	0.328	–	–	+0.007
scVision best ($\Delta > 0$): 4/5 ; mean $\Delta = +0.020$												
<i>MLP probe</i>												
Retina	0.750	0.675	0.710	0.703	0.688	0.683	0.641	0.549	0.579	0.353	0.331	+0.040
Kidney cortex	0.553	0.510	0.447	0.509	–	–	–	–	0.132	–	–	+0.043
Ovary	0.893	0.846	0.818	0.773	–	–	–	–	0.661	–	–	+0.047
Obstructive neph.	0.601	0.534	0.539	0.551	0.575	0.542	0.526	0.488	0.146	0.295	0.106	+0.026
Focal cort. dys.	0.665	0.652	0.544	0.658	–	–	–	–	0.328	–	–	+0.013
scVision best ($\Delta > 0$): 5/5 ; mean $\Delta = +0.034$												

Table S3. Macro-recall (sensitivity) under cross-study holdout (paradigm A). Per-class recall averaged over present classes. Recall is the metric most sensitive to rare cell types. Bold marks the row-best; Δ as in Table S1.

Atlas	scVision	scGPT	scFound.	Genef.	LogReg	MLP	XGB	RF	TabNet	HVG-kNN	Pears.	Δ
<i>20-NN probe</i>												
Retina	0.238	0.147	0.246	0.138	0.212	0.205	0.194	0.182	0.169	0.060	0.077	-0.008
Kidney cortex	0.091	0.069	0.071	0.030	0.060	0.061	0.051	0.078	0.018	0.024	0.005	+0.012
Ovary	0.338	0.237	0.317	0.155	0.151	0.157	0.240	0.254	0.115	0.008	0.036	+0.021
Obstructive neph.	0.068	0.044	0.050	0.015	0.047	0.048	0.064	0.076	0.021	0.025	0.009	-0.008
Focal cort. dys.	0.177	0.172	0.135	0.163	0.126	0.130	0.130	0.148	0.101	0.083	0.097	+0.005
scVision best ($\Delta > 0$): 3/5 ; mean $\Delta = +0.004$												
<i>MLP probe</i>												
Retina	0.293	0.218	0.225	0.263	0.212	0.205	0.194	0.182	0.169	0.060	0.077	+0.068
Kidney cortex	0.092	0.075	0.056	0.093	0.060	0.061	0.051	0.078	0.018	0.024	0.005	+0.014
Ovary	0.286	0.262	0.249	0.155	0.151	0.157	0.240	0.254	0.115	0.008	0.036	+0.024
Obstructive neph.	0.074	0.053	0.046	0.046	0.047	0.048	0.064	0.076	0.021	0.025	0.009	-0.002
Focal cort. dys.	0.181	0.177	0.125	0.168	0.126	0.130	0.130	0.148	0.101	0.083	0.097	+0.004
scVision best ($\Delta > 0$): 4/5 ; mean $\Delta = +0.022$												

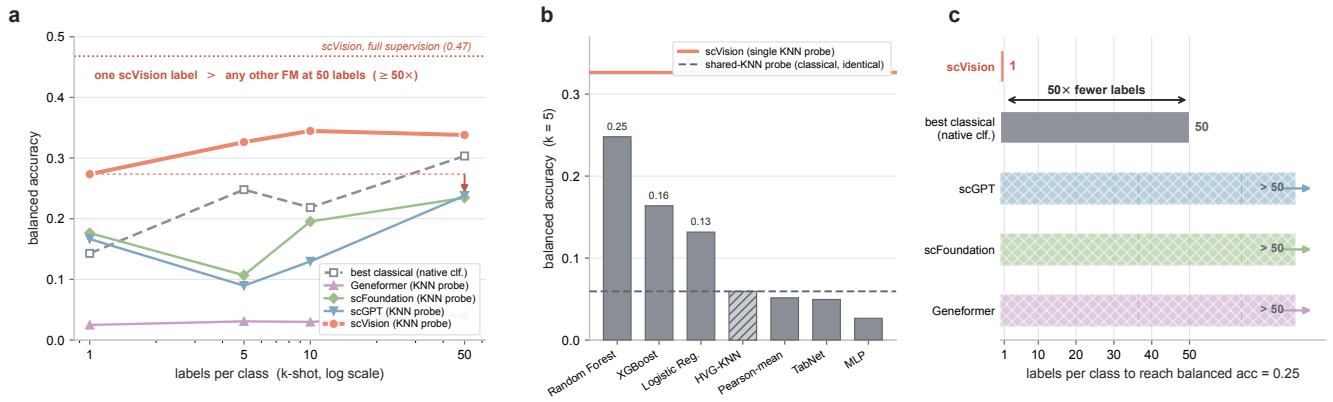


Figure S2. Label efficiency on kidney cortex (paradigm A). Balanced accuracy versus labelled cells per class ($k \in \{1, 5, 10, 50\}$). All foundation models share one 20-NN probe, isolating feature-space quality; classical baselines use their own native classifier (the steelman envelope). (a) scVision leads at every budget, and a single scVision label exceeds every other foundation model given fifty. (b) Even when each classical baseline fits its own best classifier on a 5-shot support, all remain below scVision's single probe. (c) Annotation budget to reach balanced accuracy 0.25.

Multi-study integration

Table S5. Embedding-space integration quality (ovary cross-donor pool). From a frozen embedding of 536 ovary cells spanning four coarse identities, kNN label purity ($k = 15$) and Moran's I spatial autocorrelation of the cell-type label on the embedding graph, each with a shuffled-label null reported as a Z score. scVision attains the highest purity and the strongest label autocorrelation, with scFoundation a close second. Classes: endothelial cell ($n=57$), leukocyte ($n=133$), pericyte ($n=46$), stromal cell of ovary ($n=7052$).

Method	kNN purity	purity Z	Moran's I	Moran Z
scVision	0.973	73.603	0.930	119.8
scGPT	0.957	68.324	0.880	111.9
scFoundation	0.972	68.509	0.913	116.6
Geneformer	0.911	73.625	0.752	93.8

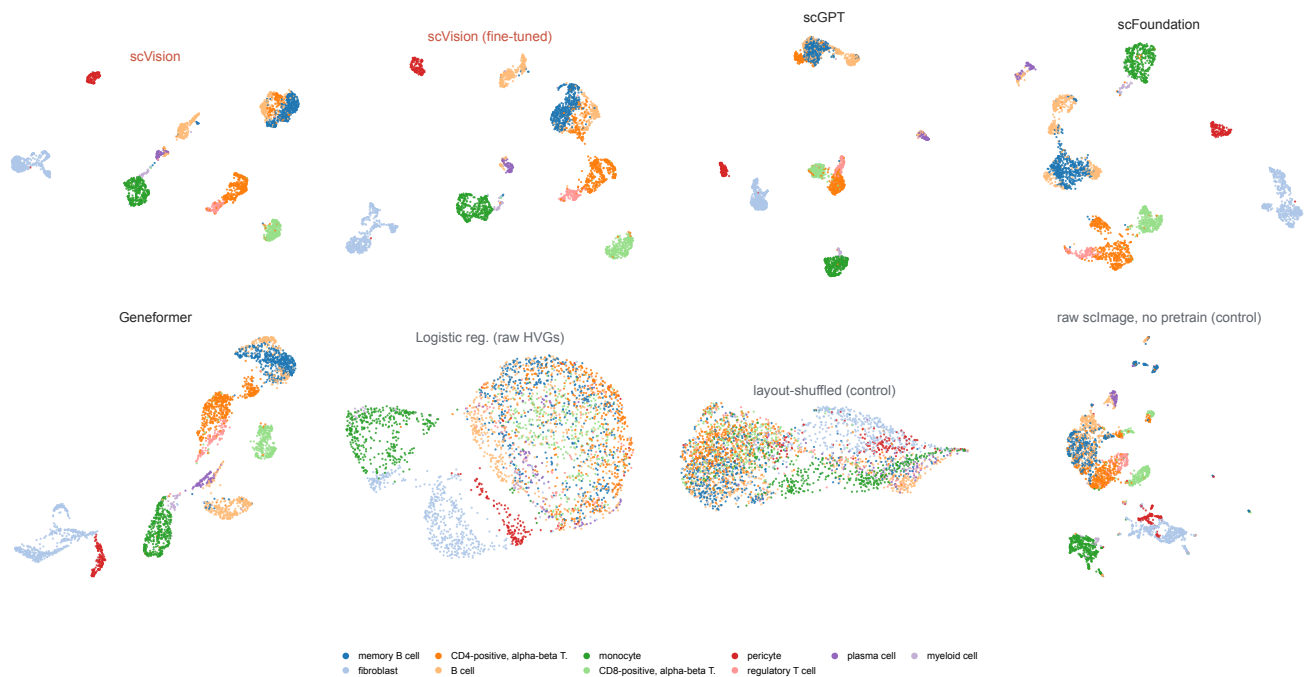


Figure S3. Frozen-embedding integration of the Crohn ileum atlas, by method. UMAP of each method's representation of the same 3,121 cells, coloured by cell type (top-18 by frequency; rarer types grey). scVision and the token foundation models resolve the major lineages; the two controls collapse that structure — shuffling the scImage layout before encoding (*layout-shuffled*) and probing the un-pretrained raw scImage (*raw scImage*) both lose the cell-type geometry, showing the separation comes from the learned spatial arrangement, not the gene set alone. Coordinates from `crohn_umap_per_method.npz`.

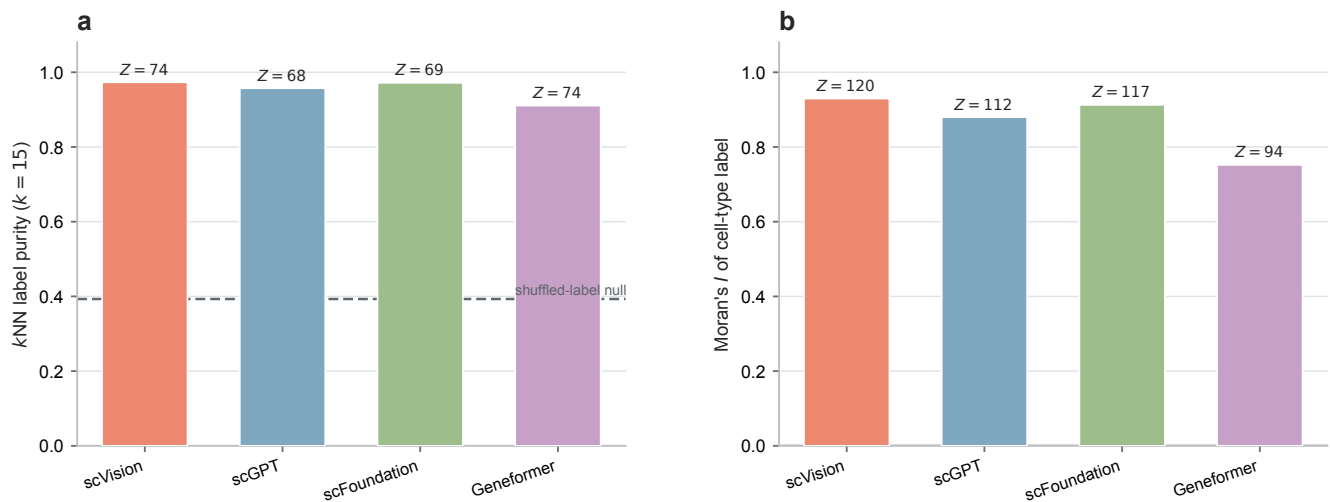


Figure S4. Quantitative integration quality on the ovary cross-donor pool. (a) k NN label purity ($k = 15$) and (b) Moran's I of the cell-type label on the embedding graph, for four foundation models, each annotated with its Z score against a shuffled-label null. scVision attains the highest purity and label autocorrelation, with scFoundation close behind. Values from `ovary_umap4_quality.json`.

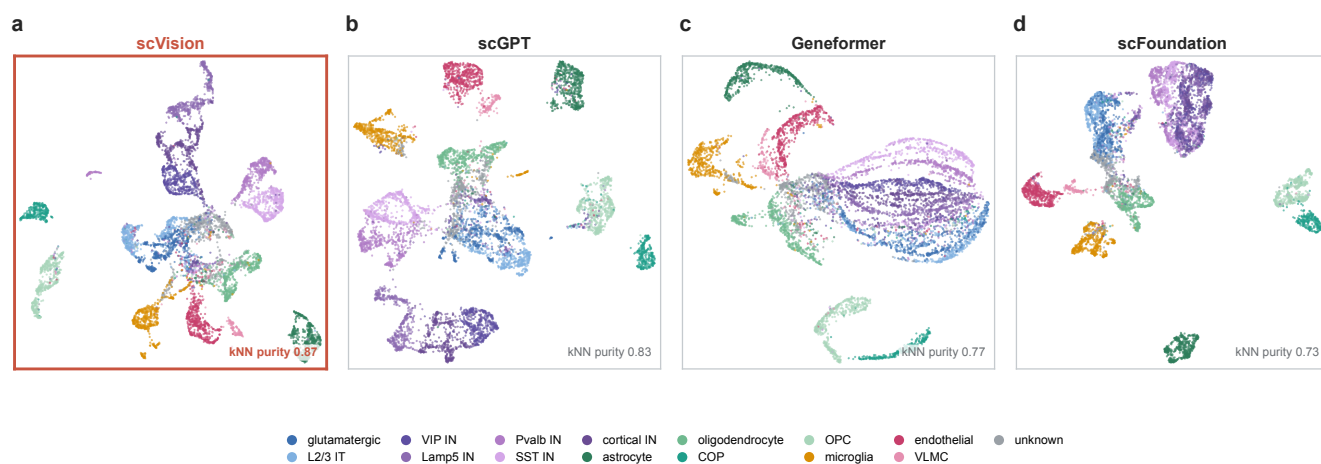


Figure S5. Frozen-embedding UMAP of the focal cortical dysplasia atlas. scVision's representation of held-out epilepsy-surgery tissue, coloured by annotated cell type, recovers the neuronal, glial and vascular compartments without any task-specific training.

Gene programs and cross-tissue recurrence

Table S6. Recurrence of attention-derived gene programs across tissues. The top-100 genes of each cell type’s last-block attention map were compared between every pair of atlases sharing that cell type. Observed mean Jaccard overlap is contrasted with a per-pair label-derangement null over 1000 permutations (seed 20240601). Cross-tissue pairs (different organs) overlap far more than chance ($Z = 41.3$). The most reproducible programs are immune and stromal identities shared organ-wide; single-pair entries ($n_{\text{pairs}} = 1$) reflect one atlas pair and are listed for completeness only.

Set	<i>n</i> pairs	cell types	obs. <i>J</i>	null <i>J</i>	null sd	<i>Z</i>	<i>p</i>
<i>Recurrence statistics (top-100 attention genes; 1000 permutations)</i>							
All cell-type pairs	731	106	0.232	0.036	0.004	48.7	1.0×10^{-3}
Cross-tissue pairs	585	77	0.239	0.037	0.005	41.3	1.0×10^{-3}
<i>Most reproducible cross-tissue cell types (mean pairwise Jaccard of the top-100 gene set)</i>							
IgG plasma cell						1.000	<i>n</i> =1
vein endothelial cell						0.610	<i>n</i> =2
plasma cell						0.592	<i>n</i> =35
natural killer cell						0.587	<i>n</i> =50
erythrocyte						0.517	<i>n</i> =2
CD16-negative, CD56-bright natural killer cell, human						0.480	<i>n</i> =3
vascular associated smooth muscle cell						0.480	<i>n</i> =3
IgA plasma cell						0.471	<i>n</i> =1
early lymphoid progenitor						0.471	<i>n</i> =1
effector memory CD4-positive, alpha-beta T cell						0.471	<i>n</i> =1
plasmacytoid dendritic cell, human						0.471	<i>n</i> =1
myeloid cell						0.418	<i>n</i> =3

Table S7. Gene programs are enriched across independent pathway databases. Top: hypergeometric enrichment (Benjamini–Hochberg $q < 0.05$) of the specificity-subtracted cross-tissue programs, evaluated against four independent databases; columns give the number of cell-type tests, significant hits, and distinct cell types with at least one hit. Bottom: the same test applied per win-atlas. Enrichment reproduces beyond the Hallmark collection used in the main text, indicating the attention programs are not an artifact of a single annotation source.

Database	tests	sig. hits	cell types
<i>Cross-tissue recurrent programs (subtract specificity), $q < 0.05$</i>			
MSigDB Hallmark	300	13	7
Reactome	300	10	5
KEGG	300	15	6
GO biological process	300	3	2
<i>Per-win-atlas extension: significant hits ($q < 0.05$) by database</i>			
Atlas	Reactome	KEGG	GO-BP
Ovary	0	0	1
Kidney cortex	4	4	9
Cystic fibrosis	6	2	7
Focal cort. dys.	3	3	0

Ablation: pretraining objective

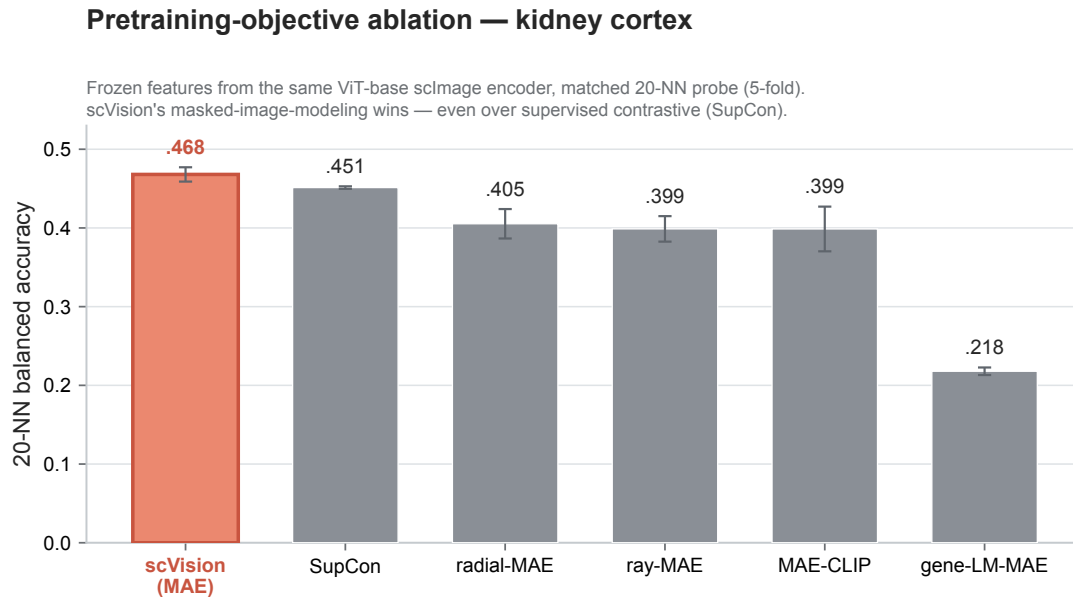


Figure S8. Pretraining-objective ablation: masked-image-modeling gives the best frozen scImage features. The frozen embedding of the *same* ViT-base scImage encoder, pretrained under six self-supervised / pretraining objectives, is probed on the kidney-cortex atlas with the identical matched 20-NN protocol (five-fold, paradigm A; every encoder loaded with zero missing keys). scVision's masked-image-modeling attains the highest balanced accuracy (0.468), ahead of supervised contrastive (SupCon, 0.451, which additionally uses cell-type labels), the MAE-masking and MAE+CLIP variants (0.399–0.405), and MAE with a gene-language model (0.218, where the gene-text conditioning is actively harmful as a frozen encoder). Bars are five-fold means $\pm$ s.d.; values from `results/ablations/ssl_kidney`.

Per-dataset cross-model benchmarks across diverse held-out tissues

Beyond the main-text atlases, we benchmark scVision against the three single-cell foundation models (scGPT, scFoundation and Geneformer) on 15 additional held-out cell atlases spanning diverse organs — brain, retina, breast, kidney, liver, prostate, lymph node, adipose, heart, eye, uterus and blood. Each atlas is a study-level holdout that scVision never saw during pre-training; to broaden tissue coverage beyond the curated collections we draw several directly from otherwise-unused CELLxGENE Census test studies. Every model is evaluated apples-to-apples on the same seeded held-out cells (identical cells, cell-type labels and class vocabulary across models), with a single shared frozen-embedding 20-nearest-neighbour probe (the decisive metric of the main text; paradigm A, five-fold cross-validation), a shared k -shot protocol, and identical k NN-purity and Moran's I geometry metrics. Within each figure the only variable is the learned representation.

On the decisive matched 20-NN balanced accuracy, **scVision's frozen embedding ranks first on all 15 atlases**, with scFoundation the most frequent runner-up; each per-atlas margin over the best competing model is reported verbatim from the on-disk five-fold metrics in the captions below. For every atlas we show (a) the frozen-embedding UMAP of each model, coloured by native cell type (coral frame marks scVision), with the matched balanced accuracy and 2-D k NN purity inset; (b) balanced accuracy and present-class macro-F1; (c) label efficiency at $k \in \{1, 5, 10, 50\}$ labels per class; and (d) 2-D k NN purity and Moran's I .

Adipose tissue

a 2,573 held-out cells · 7 native cell types · 4 single-cell foundation models, matched cells + matched 20-NN probe (paradigm A)

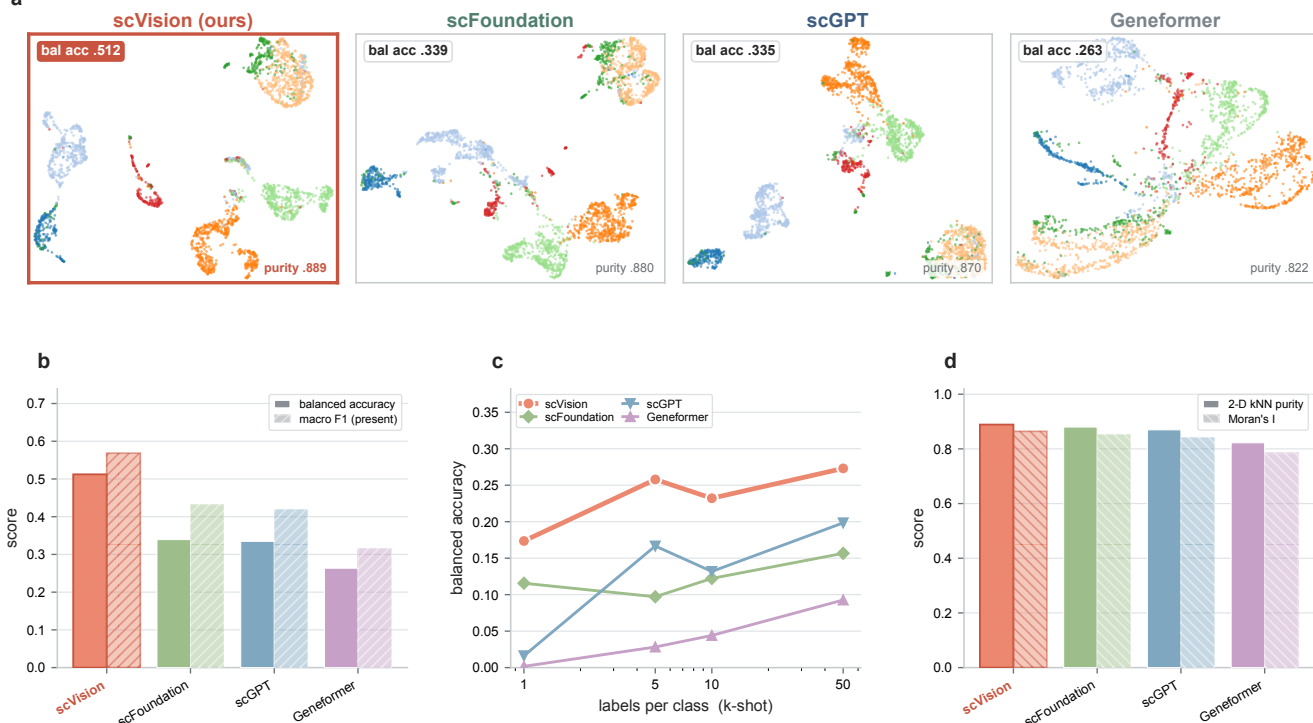


Figure S9. Adipose tissue: matched four-model benchmark. 2,573 held-out cells across 7 native cell types, evaluated apples-to-apples across scVision, scFoundation, scGPT and Geneformer (identical cells and labels, one shared 20-NN probe, paradigm A). (a) Frozen-embedding UMAP per model, cells coloured by native cell type; badges give the matched balanced accuracy (top-left) and 2-D k NN purity (bottom-right), the coral frame marking scVision. (b) Balanced accuracy and present-class macro-F1. (c) Label efficiency ($k \in \{1, 5, 10, 50\}$ labels per class). (d) 2-D k NN purity and Moran's I . scVision attains the highest matched balanced accuracy (0.512 vs. 0.339 for scFoundation). scVision also leads 2-D k NN purity (0.889).

Table S8. Per-dataset matched 20-NN benchmark summary. Held-out cell atlases (study-level holdout, paradigm A), ordered by scVision's margin over the best competing single-cell foundation model. Columns: held-out cells scored, native cell types, scVision's frozen-embedding 20-NN balanced accuracy (five-fold mean on identical matched cells), and Δ , scVision minus the best of scGPT, scFoundation and Geneformer. scVision is first on every atlas.

Atlas	Cells	Cell types	scVision (20-NN)	Δ vs best FM
Adipose tissue	2,573	7	0.512	+0.173
Eye	1,266	5	0.563	+0.145
Obstructive nephropathy	2,000	4	0.482	+0.124
Uterus	1,197	4	0.337	+0.088
Cross-tissue (snRNA)	3,881	11	0.522	+0.088
Lymph node	1,720	5	0.713	+0.083
Kidney cortex	1,916	7	0.468	+0.069
Retina	2,775	11	0.649	+0.050
Brain atlas	6,500	13	0.596	+0.027
Liver	1,698	8	0.451	+0.021
Clear-cell RCC, immune	3,500	7	0.324	+0.014
Heart	2,029	5	0.490	+0.013
Human breast cell atlas	2,167	10	0.243	+0.012
Benign prostatic hyperplasia	2,337	7	0.539	+0.011
Blood	5,476	12	0.560	+0.003

Benign prostatic hyperplasia

a 2,337 held-out cells · 7 native cell types · 4 single-cell foundation models, matched cells + matched 20-NN probe (paradigm A)

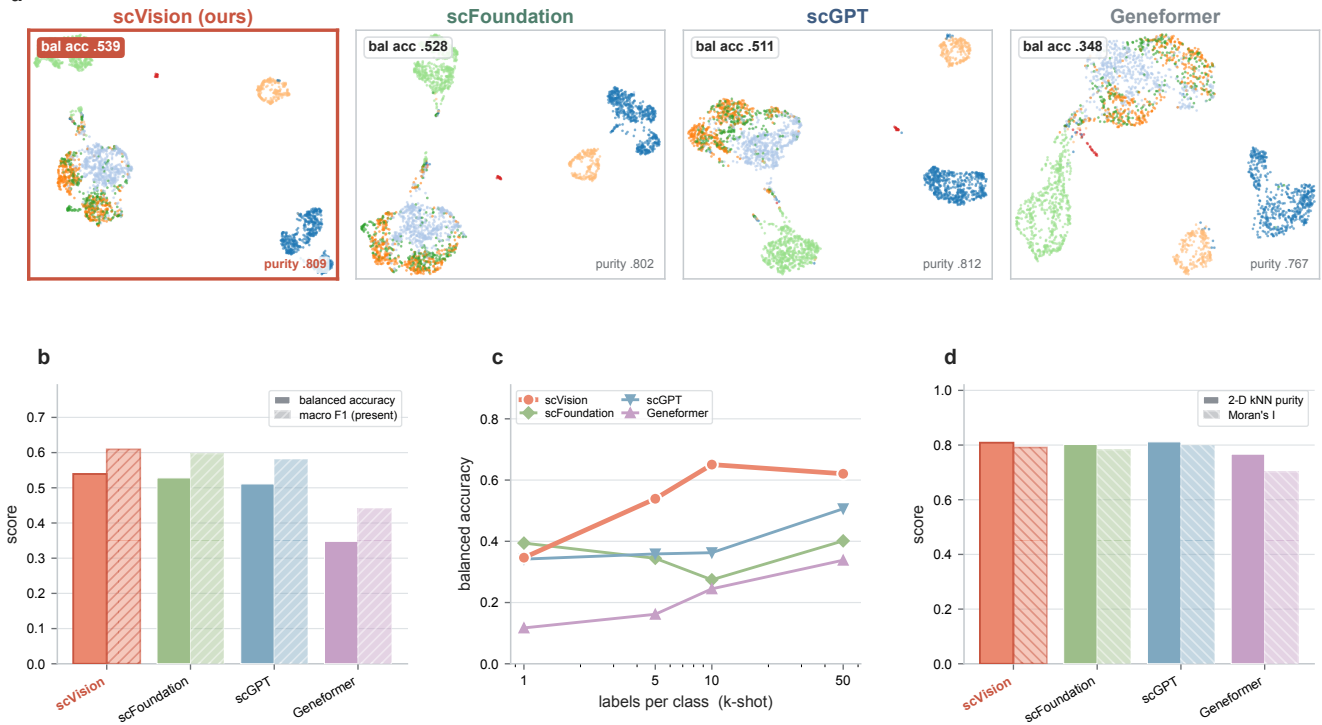


Figure S10. Benign prostatic hyperplasia: matched four-model benchmark. 2,337 held-out cells across 7 native cell types, evaluated apples-to-apples across scVision, scFoundation, scGPT and Geneformer (identical cells and labels, one shared 20-NN probe, paradigm A). (a) Frozen-embedding UMAP per model, cells coloured by native cell type; badges give the matched balanced accuracy (top-left) and 2-D k NN purity (bottom-right), the coral frame marking scVision. (b) Balanced accuracy and present-class macro-F1. (c) Label efficiency ($k \in \{1, 5, 10, 50\}$ labels per class). (d) 2-D k NN purity and Moran's I . scVision attains the highest matched balanced accuracy (0.539 vs. 0.528 for scFoundation). On 2-D geometry scGPT edges scVision on k NN purity (0.812 vs. 0.809).

Blood

a 5,476 held-out cells · 12 native cell types · 4 single-cell foundation models, matched cells + matched 20-NN probe (paradigm A)

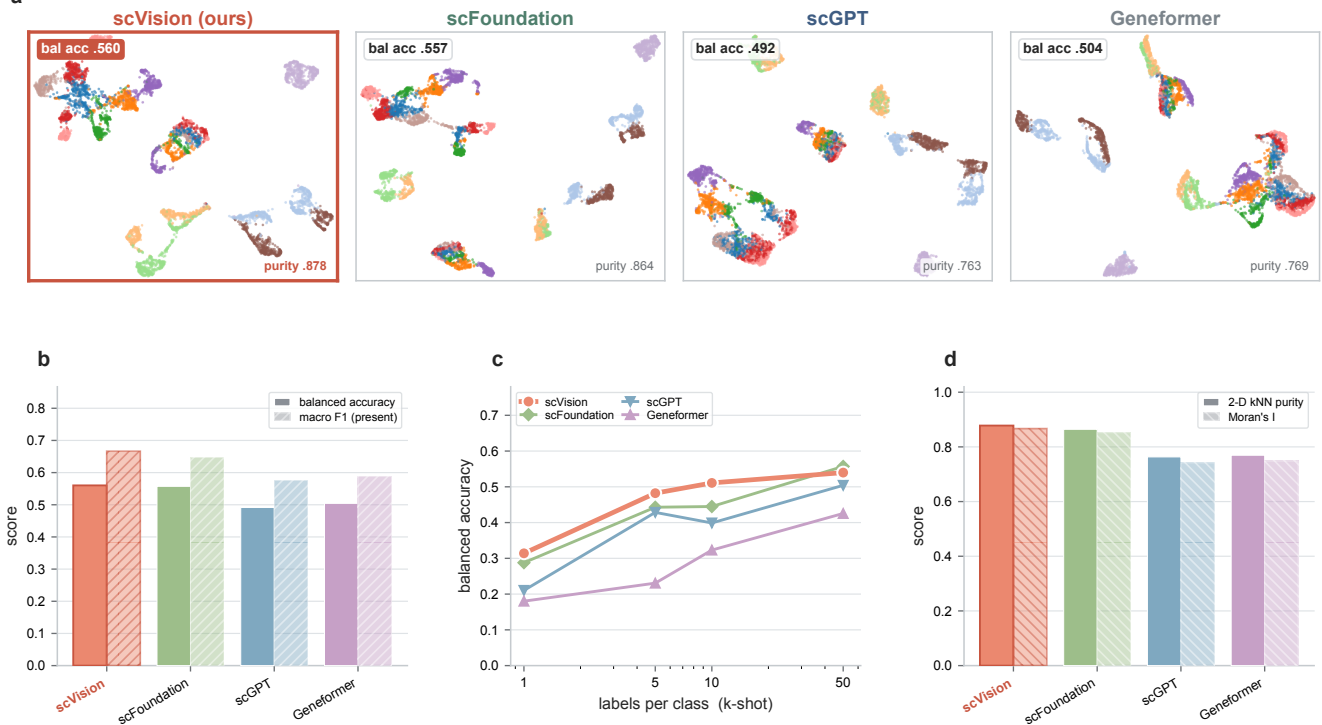


Figure S11. Blood: matched four-model benchmark. 5,476 held-out cells across 12 native cell types, evaluated apples-to-apples across scVision, scFoundation, scGPT and Geneformer (identical cells and labels, one shared 20-NN probe, paradigm A). (a) Frozen-embedding UMAP per model, cells coloured by native cell type; badges give the matched balanced accuracy (top-left) and 2-D k NN purity (bottom-right), the coral frame marking scVision. (b) Balanced accuracy and present-class macro-F1. (c) Label efficiency ($k \in \{1, 5, 10, 50\}$ labels per class). (d) 2-D k NN purity and Moran's I . scVision attains the highest matched balanced accuracy (0.560 vs. 0.557 for scFoundation). scVision also leads 2-D k NN purity (0.878).

Brain atlas

a 6,500 held-out cells · 13 native cell types · 4 single-cell foundation models, matched cells + matched 20-NN probe (paradigm A)

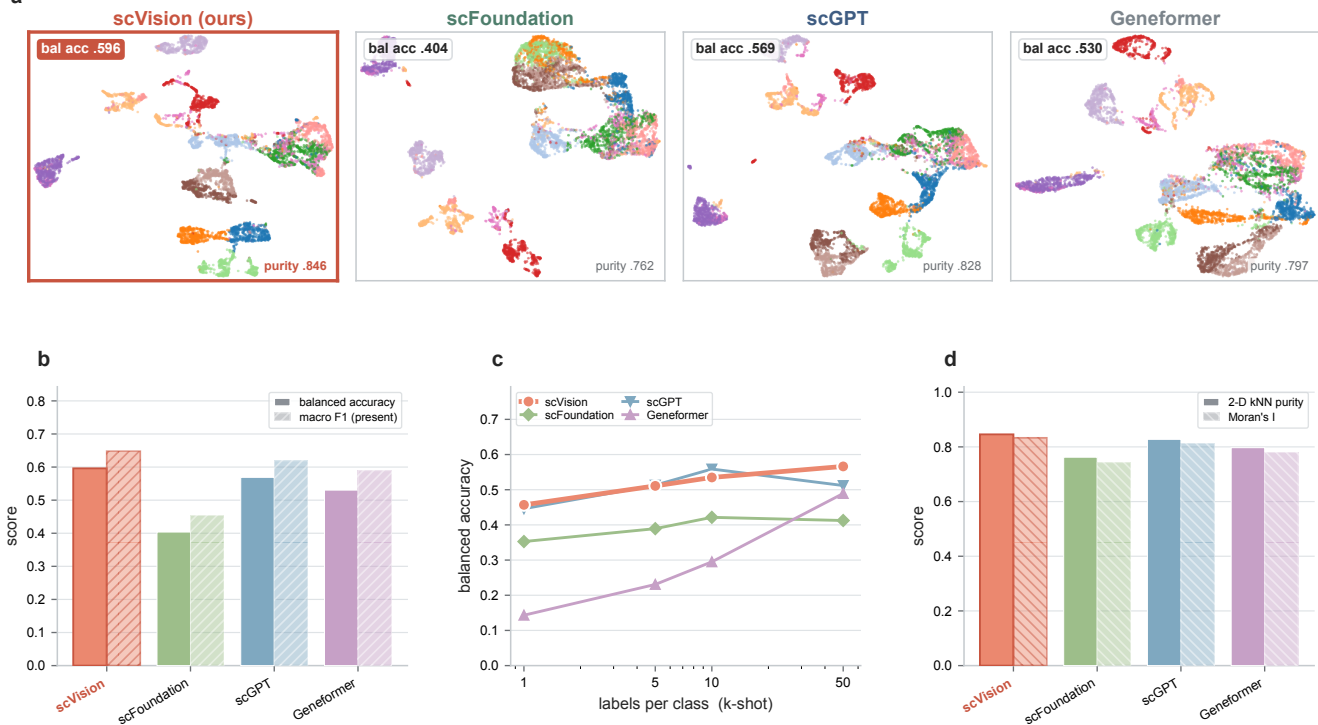


Figure S12. Brain atlas: matched four-model benchmark. 6,500 held-out cells across 13 native cell types, evaluated apples-to-apples across scVision, scFoundation, scGPT and Geneformer (identical cells and labels, one shared 20-NN probe, paradigm A). (a) Frozen-embedding UMAP per model, cells coloured by native cell type; badges give the matched balanced accuracy (top-left) and 2-D k NN purity (bottom-right), the coral frame marking scVision. (b) Balanced accuracy and present-class macro-F1. (c) Label efficiency ($k \in \{1, 5, 10, 50\}$ labels per class). (d) 2-D k NN purity and Moran's I . scVision attains the highest matched balanced accuracy (0.596 vs. 0.569 for scGPT). scVision also leads 2-D k NN purity (0.846).

Clear-cell RCC, immune

a 3,500 held-out cells · 7 native cell types · 4 single-cell foundation models, matched cells + matched 20-NN probe (paradigm A)

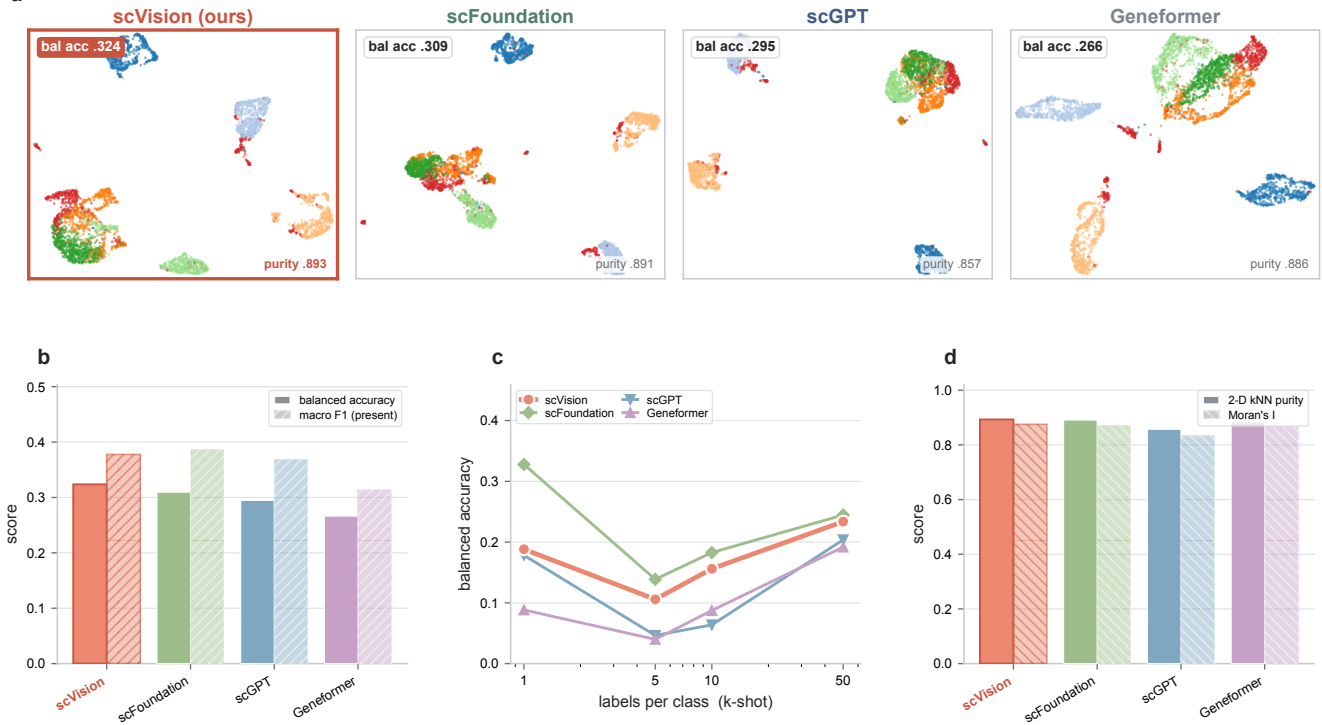


Figure S13. Clear-cell RCC, immune: matched four-model benchmark. 3,500 held-out cells across 7 native cell types, evaluated apples-to-apples across scVision, scFoundation, scGPT and Geneformer (identical cells and labels, one shared 20-NN probe, paradigm A). (a) Frozen-embedding UMAP per model, cells coloured by native cell type; badges give the matched balanced accuracy (top-left) and 2-D k NN purity (bottom-right), the coral frame marking scVision. (b) Balanced accuracy and present-class macro-F1. (c) Label efficiency ($k \in \{1, 5, 10, 50\}$ labels per class). (d) 2-D k NN purity and Moran's I . scVision attains the highest matched balanced accuracy (0.324 vs. 0.309 for scFoundation). scVision also leads 2-D k NN purity (0.893).

Cross-tissue (snRNA)

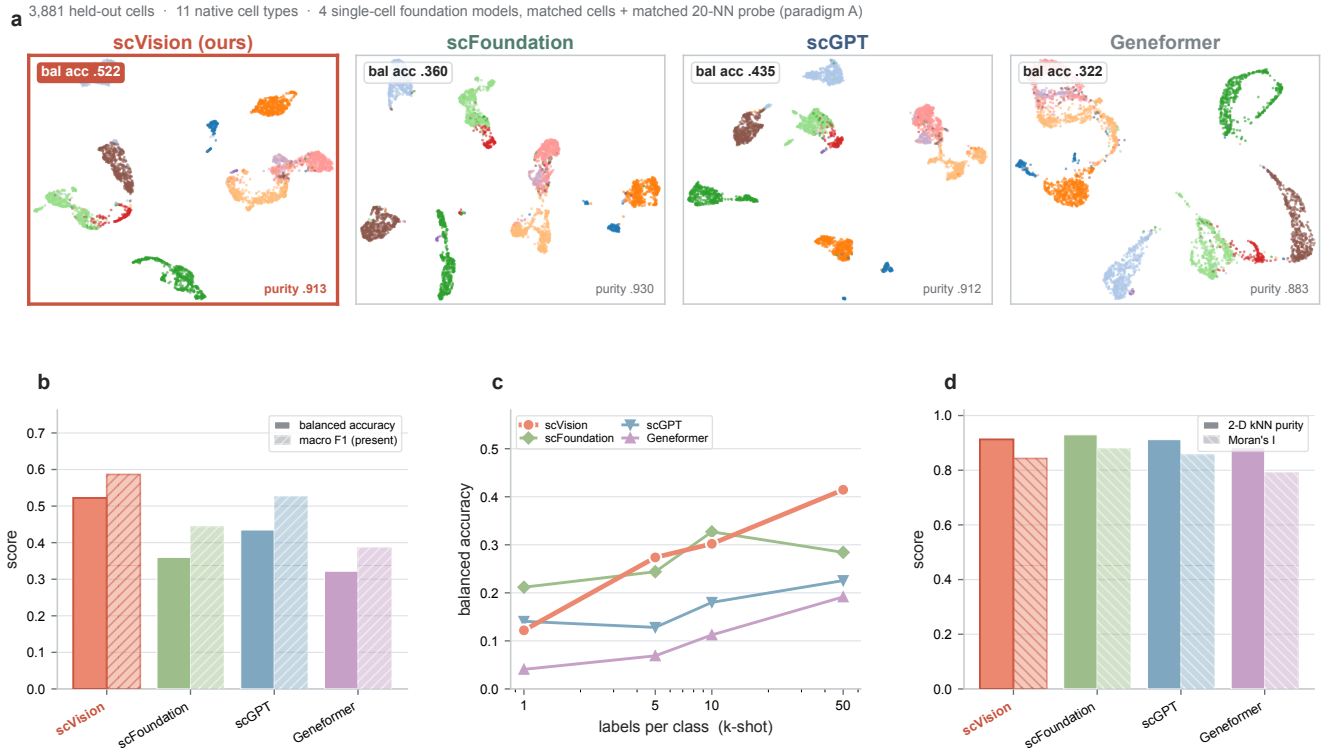


Figure S14. Cross-tissue (snRNA): matched four-model benchmark. 3,881 held-out cells across 11 native cell types, evaluated apples-to-apples across scVision, scFoundation, scGPT and Geneformer (identical cells and labels, one shared 20-NN probe, paradigm A). (a) Frozen-embedding UMAP per model, cells coloured by native cell type; badges give the matched balanced accuracy (top-left) and 2-D k NN purity (bottom-right), the coral frame marking scVision. (b) Balanced accuracy and present-class macro-F1. (c) Label efficiency ($k \in \{1, 5, 10, 50\}$ labels per class). (d) 2-D k NN purity and Moran's I . scVision attains the highest matched balanced accuracy (0.522 vs. 0.435 for scGPT). On 2-D geometry scFoundation edges scVision on k NN purity (0.930 vs. 0.913).

Eye

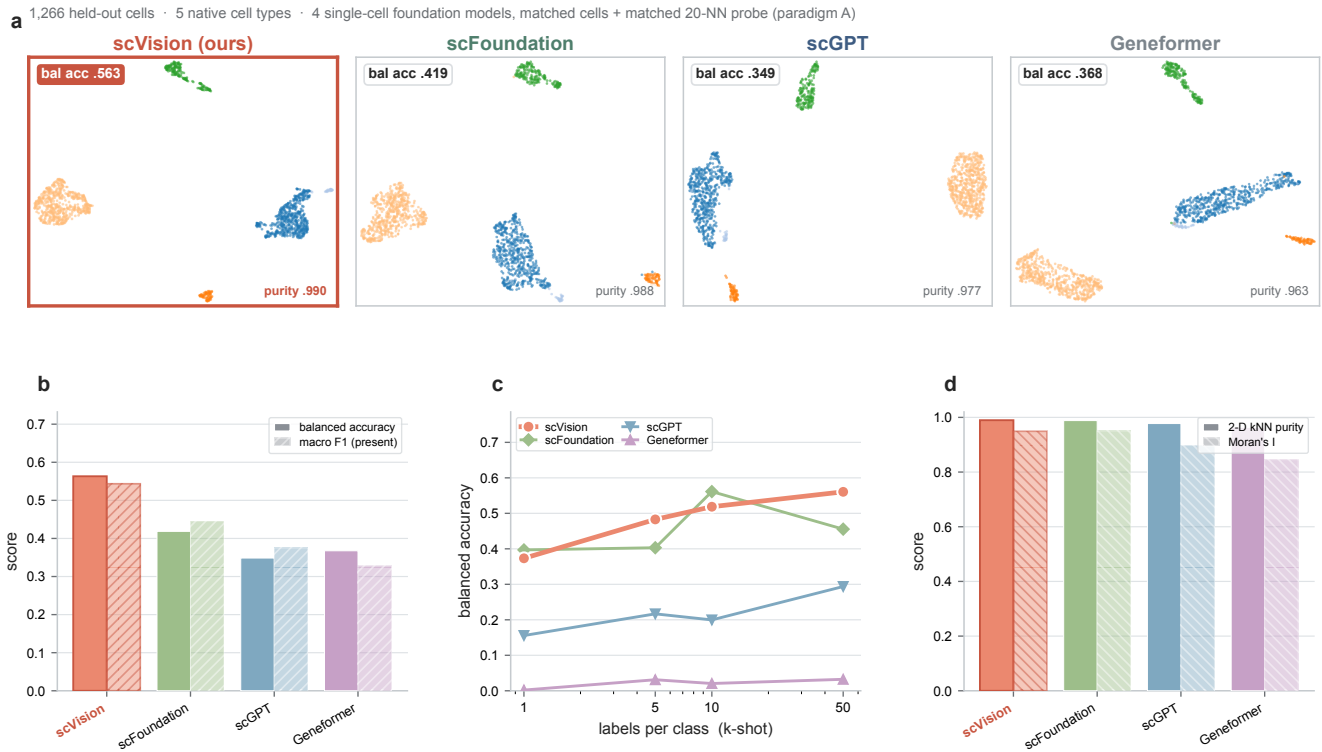


Figure S15. Eye: matched four-model benchmark. 1,266 held-out cells across 5 native cell types, evaluated apples-to-apples across scVision, scFoundation, scGPT and Geneformer (identical cells and labels, one shared 20-NN probe, paradigm A). (a) Frozen-embedding UMAP per model, cells coloured by native cell type; badges give the matched balanced accuracy (top-left) and 2-D k NN purity (bottom-right), the coral frame marking scVision. (b) Balanced accuracy and present-class macro-F1. (c) Label efficiency ($k \in \{1, 5, 10, 50\}$ labels per class). (d) 2-D k NN purity and Moran's I . scVision attains the highest matched balanced accuracy (0.563 vs. 0.419 for scFoundation). scVision also leads 2-D k NN purity (0.990).

Heart

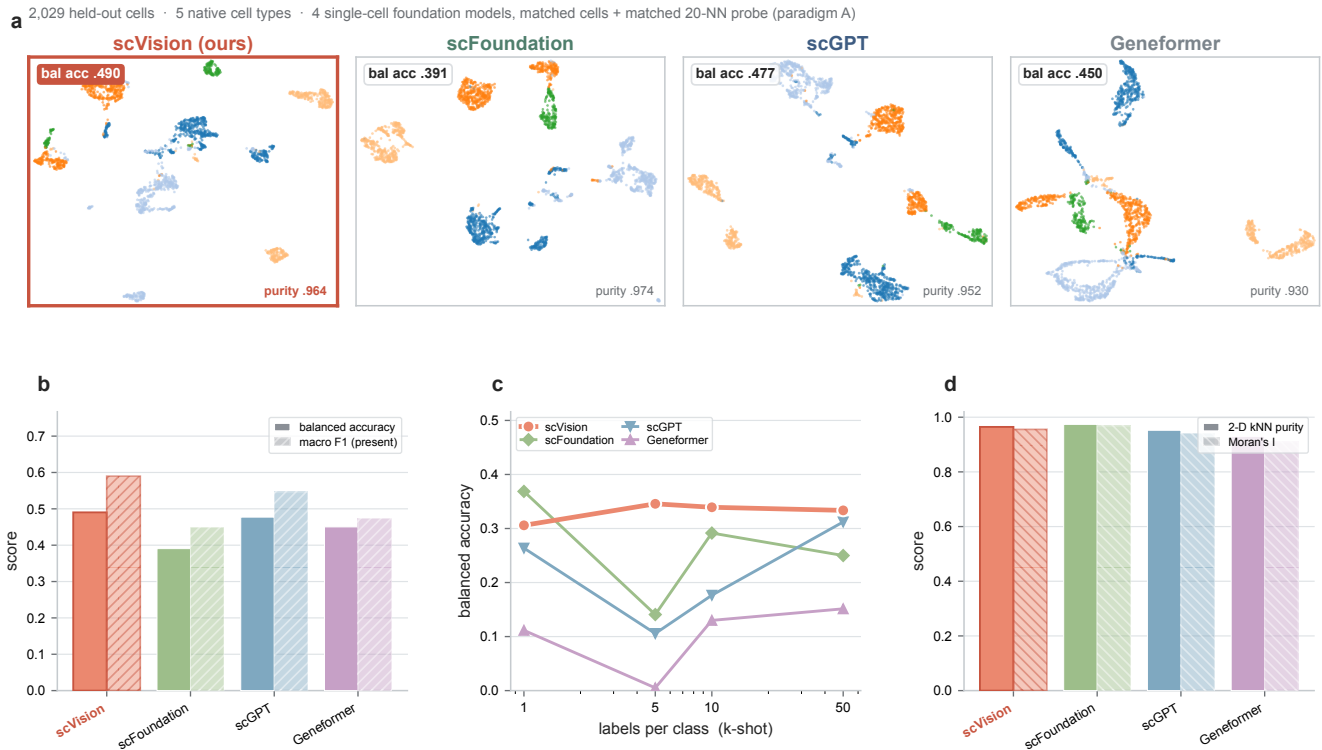


Figure S16. Heart: matched four-model benchmark. 2,029 held-out cells across 5 native cell types, evaluated apples-to-apples across scVision, scFoundation, scGPT and Geneformer (identical cells and labels, one shared 20-NN probe, paradigm A). (a) Frozen-embedding UMAP per model, cells coloured by native cell type; badges give the matched balanced accuracy (top-left) and 2-D k NN purity (bottom-right), the coral frame marking scVision. (b) Balanced accuracy and present-class macro-F1. (c) Label efficiency ($k \in \{1, 5, 10, 50\}$ labels per class). (d) 2-D k NN purity and Moran's I . scVision attains the highest matched balanced accuracy (0.490 vs. 0.477 for scGPT). On 2-D geometry scFoundation edges scVision on k NN purity (0.974 vs. 0.964).

Human breast cell atlas

a 2,167 held-out cells · 10 native cell types · 4 single-cell foundation models, matched cells + matched 20-NN probe (paradigm A)

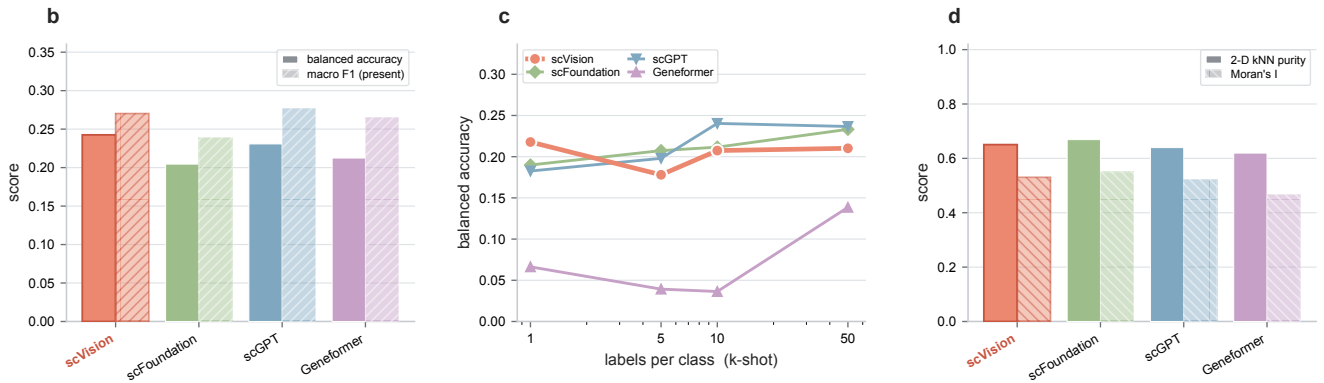
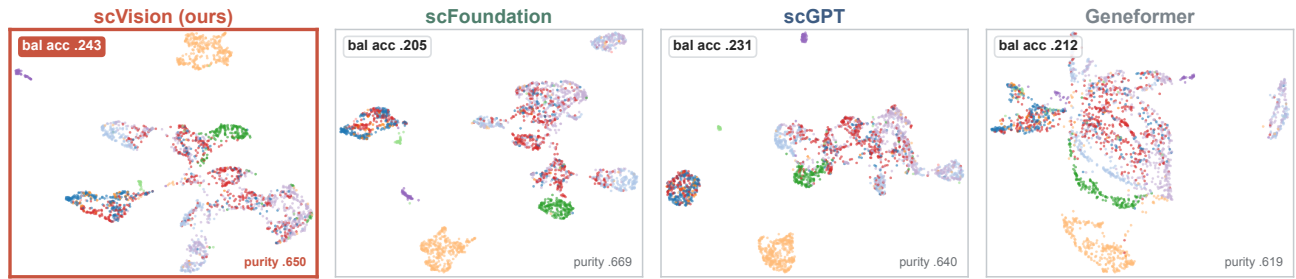


Figure S17. Human breast cell atlas: matched four-model benchmark. 2,167 held-out cells across 10 native cell types, evaluated apples-to-apples across scVision, scFoundation, scGPT and Geneformer (identical cells and labels, one shared 20-NN probe, paradigm A). (a) Frozen-embedding UMAP per model, cells coloured by native cell type; badges give the matched balanced accuracy (top-left) and 2-D k NN purity (bottom-right), the coral frame marking scVision. (b) Balanced accuracy and present-class macro-F1. (c) Label efficiency ($k \in \{1, 5, 10, 50\}$ labels per class). (d) 2-D k NN purity and Moran's I . scVision attains the highest matched balanced accuracy (0.243 vs. 0.231 for scGPT). On 2-D geometry scFoundation edges scVision on k NN purity (0.669 vs. 0.650).

Kidney cortex

a 1,916 held-out cells · 7 native cell types · 4 single-cell foundation models, matched cells + matched 20-NN probe (paradigm A)

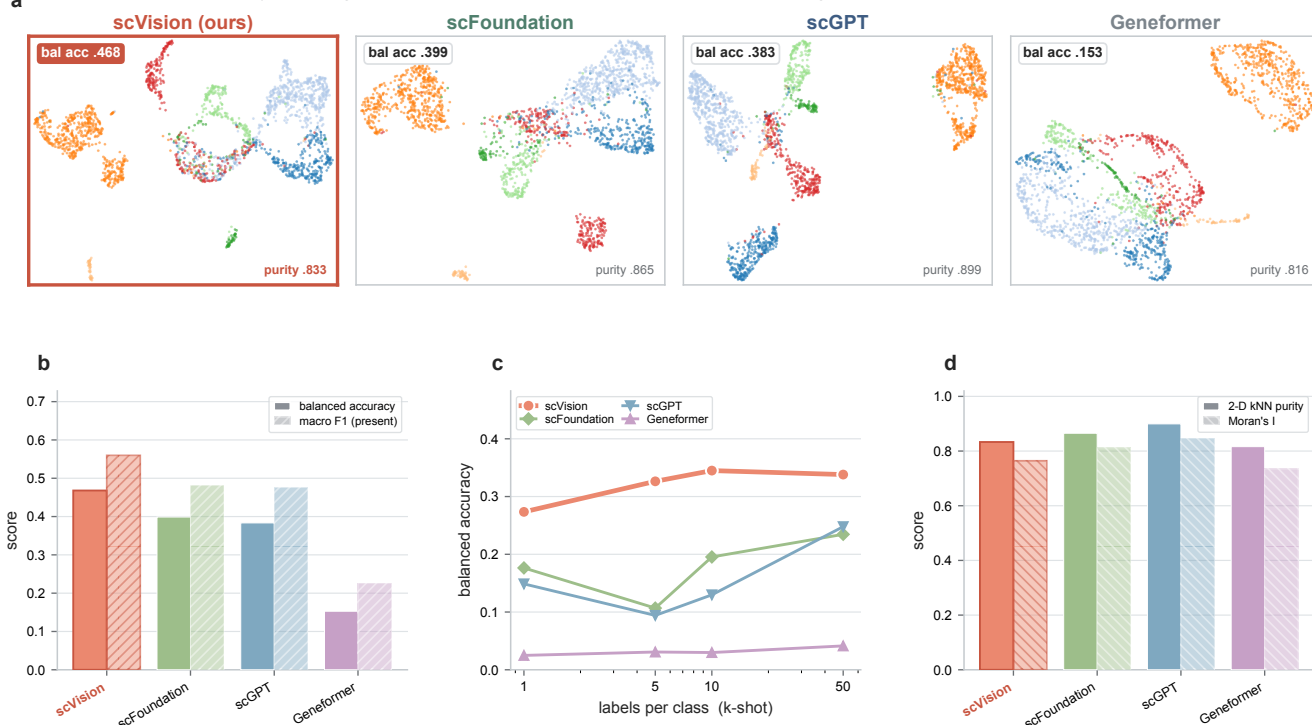


Figure S18. Kidney cortex: matched four-model benchmark. 1,916 held-out cells across 7 native cell types, evaluated apples-to-apples across scVision, scFoundation, scGPT and Geneformer (identical cells and labels, one shared 20-NN probe, paradigm A). (a) Frozen-embedding UMAP per model, cells coloured by native cell type; badges give the matched balanced accuracy (top-left) and 2-D k NN purity (bottom-right), the coral frame marking scVision. (b) Balanced accuracy and present-class macro-F1. (c) Label efficiency ($k \in \{1, 5, 10, 50\}$ labels per class). (d) 2-D k NN purity and Moran's I . scVision attains the highest matched balanced accuracy (0.468 vs. 0.399 for scFoundation). On 2-D geometry scGPT edges scVision on k NN purity (0.899 vs. 0.833).

Liver

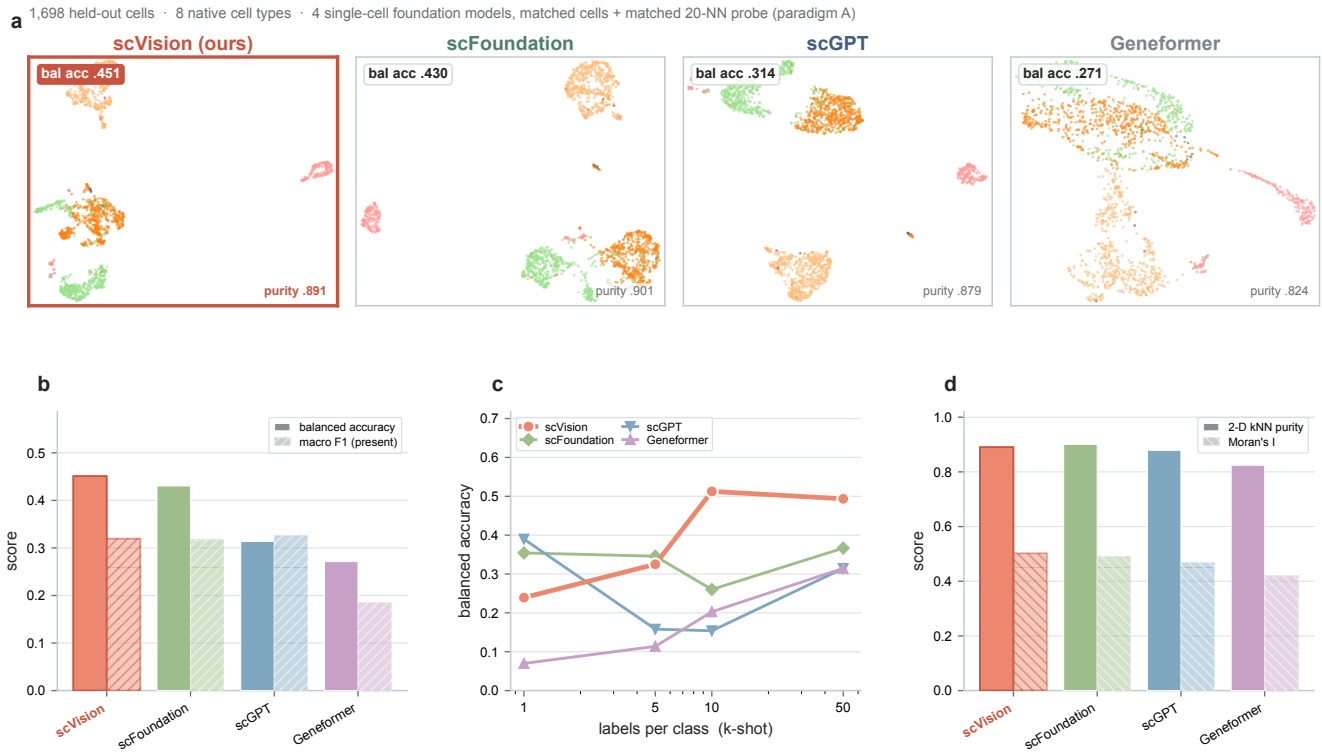


Figure S19. Liver: matched four-model benchmark. 1,698 held-out cells across 8 native cell types, evaluated apples-to-apples across scVision, scFoundation, scGPT and Geneformer (identical cells and labels, one shared 20-NN probe, paradigm A). (a) Frozen-embedding UMAP per model, cells coloured by native cell type; badges give the matched balanced accuracy (top-left) and 2-D k NN purity (bottom-right), the coral frame marking scVision. (b) Balanced accuracy and present-class macro-F1. (c) Label efficiency ($k \in \{1, 5, 10, 50\}$ labels per class). (d) 2-D k NN purity and Moran's I . scVision attains the highest matched balanced accuracy (0.451 vs. 0.430 for scFoundation). On 2-D geometry scFoundation edges scVision on k NN purity (0.901 vs. 0.891).

Lymph node

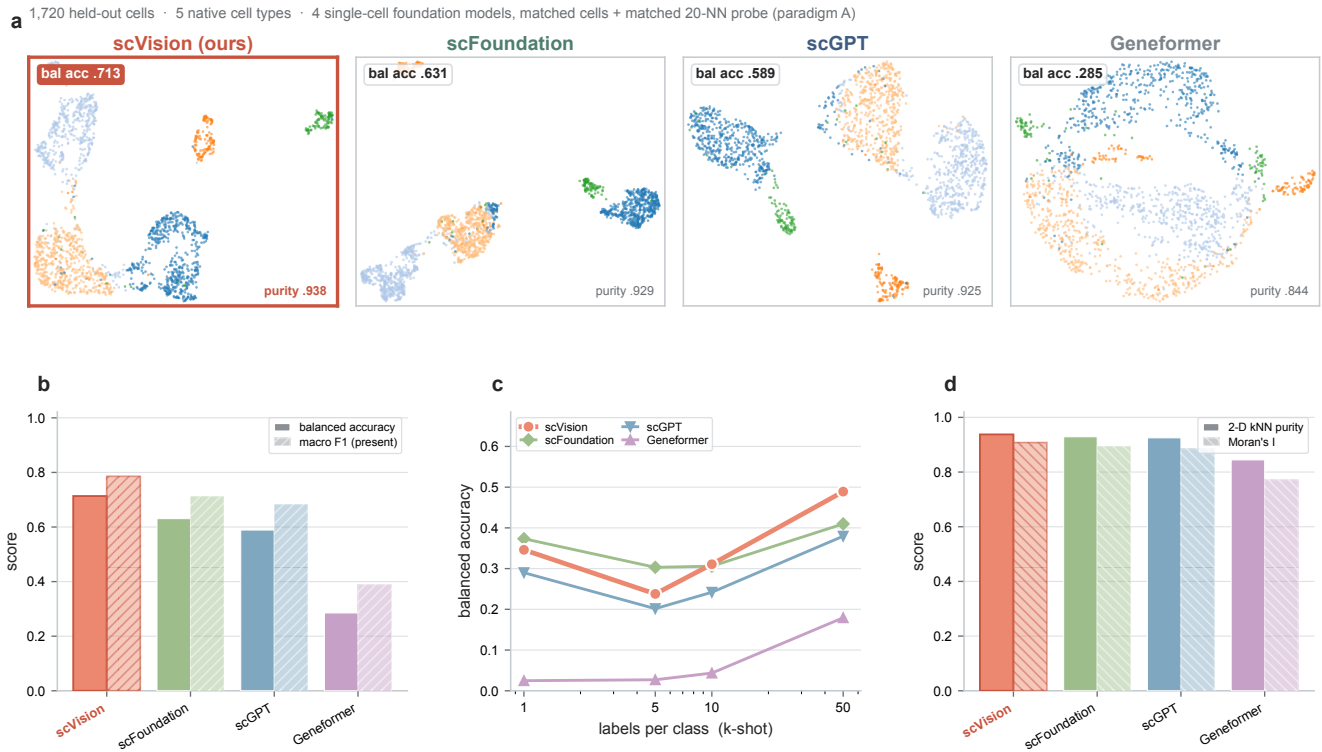


Figure S20. Lymph node: matched four-model benchmark. 1,720 held-out cells across 5 native cell types, evaluated apples-to-apples across scVision, scFoundation, scGPT and Geneformer (identical cells and labels, one shared 20-NN probe, paradigm A). (a) Frozen-embedding UMAP per model, cells coloured by native cell type; badges give the matched balanced accuracy (top-left) and 2-D k NN purity (bottom-right), the coral frame marking scVision. (b) Balanced accuracy and present-class macro-F1. (c) Label efficiency ($k \in \{1, 5, 10, 50\}$ labels per class). (d) 2-D k NN purity and Moran's I . scVision attains the highest matched balanced accuracy (0.713 vs. 0.631 for scFoundation). scVision also leads 2-D k NN purity (0.938).

Obstructive nephropathy

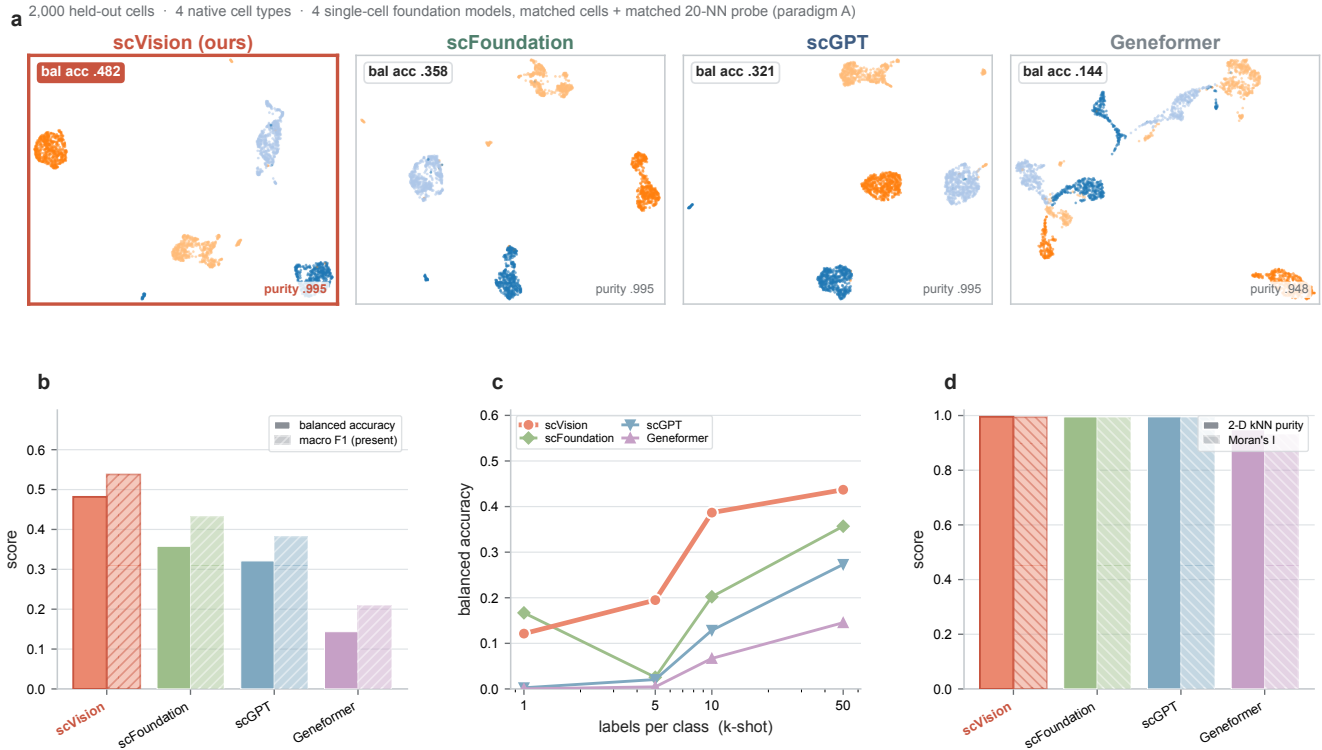


Figure S21. Obstructive nephropathy: matched four-model benchmark. 2,000 held-out cells across 4 native cell types, evaluated apples-to-apples across scVision, scFoundation, scGPT and Geneformer (identical cells and labels, one shared 20-NN probe, paradigm A). (a) Frozen-embedding UMAP per model, cells coloured by native cell type; badges give the matched balanced accuracy (top-left) and 2-D k NN purity (bottom-right), the coral frame marking scVision. (b) Balanced accuracy and present-class macro-F1. (c) Label efficiency ($k \in \{1, 5, 10, 50\}$ labels per class). (d) 2-D k NN purity and Moran's I . scVision attains the highest matched balanced accuracy (0.482 vs. 0.358 for scFoundation). scVision also leads 2-D k NN purity (0.995).

Retina

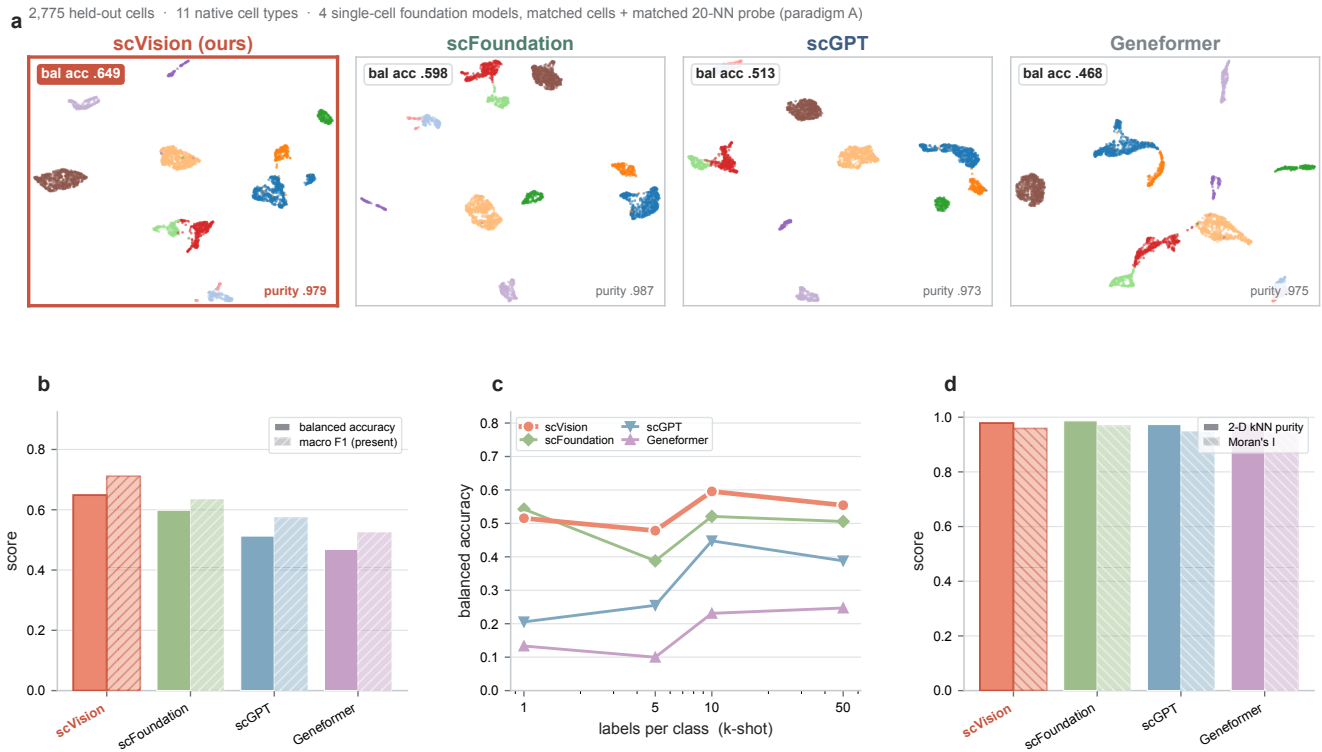


Figure S22. Retina: matched four-model benchmark. 2,775 held-out cells across 11 native cell types, evaluated apples-to-apples across scVision, scFoundation, scGPT and Geneformer (identical cells and labels, one shared 20-NN probe, paradigm A). (a) Frozen-embedding UMAP per model, cells coloured by native cell type; badges give the matched balanced accuracy (top-left) and 2-D k NN purity (bottom-right), the coral frame marking scVision. (b) Balanced accuracy and present-class macro-F1. (c) Label efficiency ($k \in \{1, 5, 10, 50\}$ labels per class). (d) 2-D k NN purity and Moran's I . scVision attains the highest matched balanced accuracy (0.649 vs. 0.598 for scFoundation). On 2-D geometry scFoundation edges scVision on k NN purity (0.987 vs. 0.979).

Uterus

a 1,197 held-out cells · 4 native cell types · 4 single-cell foundation models, matched cells + matched 20-NN probe (paradigm A)

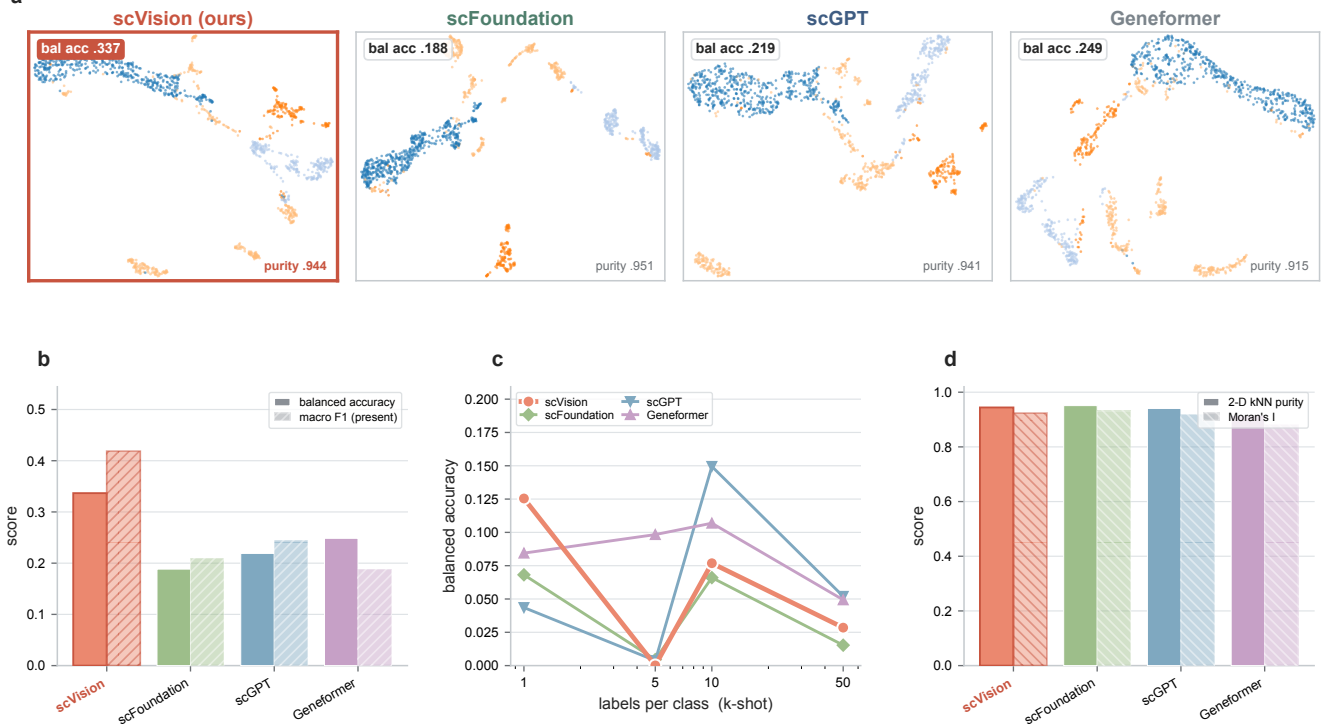


Figure S23. Uterus: matched four-model benchmark. 1,197 held-out cells across 4 native cell types, evaluated apples-to-apples across scVision, scFoundation, scGPT and Geneformer (identical cells and labels, one shared 20-NN probe, paradigm A). (a) Frozen-embedding UMAP per model, cells coloured by native cell type; badges give the matched balanced accuracy (top-left) and 2-D k NN purity (bottom-right), the coral frame marking scVision. (b) Balanced accuracy and present-class macro-F1. (c) Label efficiency ($k \in \{1, 5, 10, 50\}$ labels per class). (d) 2-D k NN purity and Moran's I . scVision attains the highest matched balanced accuracy (0.337 vs. 0.249 for Geneformer). On 2-D geometry scFoundation edges scVision on k NN purity (0.951 vs. 0.944).