LAMINAR: QUERYING BIOLOGY AT ANY SCALE WITH TISSUE EMBEDDING FIELDS

Anonymous authors
Paper under double-blind review

ABSTRACT

Biology is organized at nested spatial scales, from molecules to cells to multicellular neighborhoods, whose functions emerge from multi-scale organizations rather than any single cell. Imaging-based spatial transcriptomics (iST) captures all of these scales in a single experiment, measuring up to billions of individual transcripts, each with its gene identity and sub-micrometer position. To handle this volume, existing foundation models pool transcripts into predefined units such as cells, patches, or bins before learning, which fixes the representation to a single spatial scale and discards the molecular geometry within each unit. We introduce *tissue embedding fields*, a representation that assigns contextual embeddings on a shared unit hypersphere to every measured transcript. A region of any size or shape, from a single cell to a whole section, can then be represented by summing and normalizing the embeddings of the transcripts it contains, without re-encoding the tissue. The spatial unit of analysis is therefore chosen at query time rather than fixed during training. To compute such fields at billion-transcript scale, we introduce Laminar, a local-to-global encoder whose cost grows linearly with transcript count. A graph neural network exchanges short-range molecular information over local transcript graphs while a quadtree spatial Transformer propagates information to progressively larger tissue tiles. Each transcript then retrieves this multi-scale context through cross-attention, yielding its final embedding. We train Laminar using a joint-embedding predictive architecture, where a context encoder processes spatially- and gene-masked tissue, while a predictor matches multi-scale regional representations of the complete tissue. A single Laminar encoder pretrained on 632 public human iST sections, evaluated frozen on eight transfer tasks spanning cell, niche, and tissue scales, achieves the best scores on tissue-scale tasks while remaining competitive with fixed-scale foundation models at cell scale, with an overall best mean rank. Laminar shows that biological representations need not be tied to predefined spatial units; a unified representation enables reasoning across the hierarchy of tissue organization.

1 INTRODUCTION

A general-purpose representation of tissue must capture molecular organization across spatial scales (Adler et al., 2023; Palla et al., 2022). RNA molecules occupy subcellular locations at submicrometer scale (Xia et al., 2019; Mah et al., 2024). Cells of $\sim$10 μm form multicellular neighborhoods of $\sim$100 μm (Schürch et al., 2020; Bhate et al., 2022). Critically, these scales are not independent (Adler et al., 2023). Capturing this organization therefore requires resolving molecular context at its native resolution, retaining the dependencies between scales, and supporting queries over arbitrary regions at any of them (Fig. 1a) (Bommasani et al., 2021; Bunne et al., 2024).

Imaging-based spatial transcriptomics (iST) allows profiling tissues across all of these spatial scales in a single experiment. By measuring billions of individual RNA molecules, hereafter denoted transcripts, through gene identity and their spatial coordinates, iST resolves tissue architecture at sub-μm resolution (Chen et al., 2015; Moffitt et al., 2022; Zhuang, 2021). Existing iST corpora now exceed 100 milion cells (Zhang et al., 2023; Plummer et al., 2025; Tejada-Lapuerta et al., 2025; Birk et al., 2026), with a single spatial dataset containing up to a billion of transcript measurements. This motivates training high-capacity foundation models, requiring architectures that can operate on massive molecular datasets (Jiang et al., 2026).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

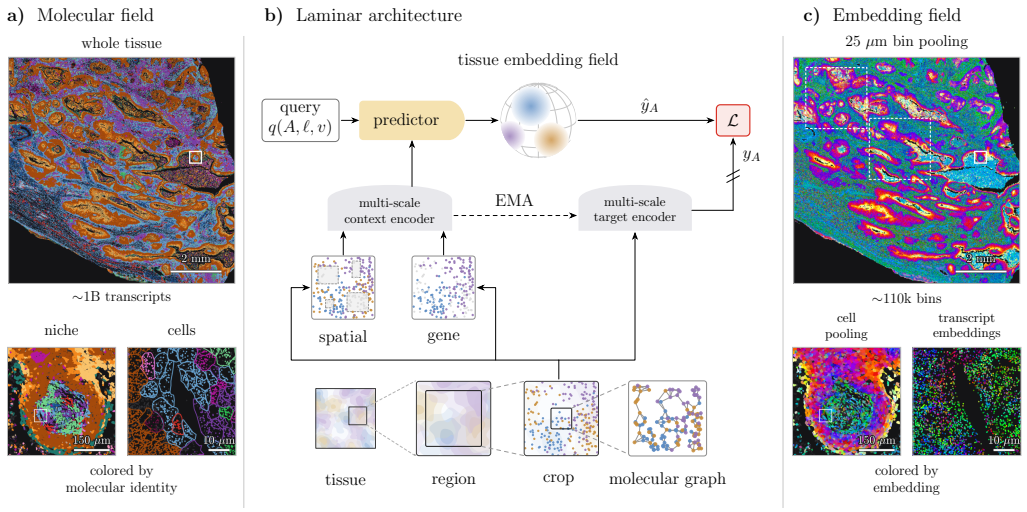

Figure 1: **Tissue embedding fields and Laminar architecture.** **(a)** iST represents tissue as a gene-labeled molecular point field spanning transcripts, cells, niches, and whole tissue. **(b)** Laminar is trained with a multi-scale JEPA objective. The online encoder receives spatially or gene-corrupted views, while an EMA target encoder computes the complete tissue field. A predictor recovers regional target embeddings across physical scales. **(c)** The resulting tissue embedding field assigns a contextual unit embedding to each transcript coordinate. Cells, bins, niches, and other spatial regions are represented by pooling the same field without re-encoding the tissue.

> **Biological terminology.** A **transcript** is an RNA molecule with a gene identity and spatial coordinate, measured as point at sub-μm resolution. The **molecular profile** of a region is the number of detected transcripts for each gene it encompasses. A **cell** is the basic functional unit of tissue and can span around 10 μm in diameter. A **niche** is a local group of interacting cells and can extends over tens to hundreds of micrometers. A **tissue region** contains many cells and niches and extends from hundreds of micrometers to millimeters. A **tissue section** is the imaged slice from which these measurements are obtained.

A model operating on iST data must propagate information across spatial scales. Short-range molecular organization carries subcellular and cellular structure (Xia et al., 2019; Mah et al., 2024), while multicellular and tissue-scale patterns extend over distances orders of magnitude larger (Schürch et al., 2020). Sections can contain 10^7–10^9 transcripts (Moffitt et al., 2022; Ren et al., 2025), resulting in 10^{14}–10^{18} pairwise direct interactions on a single section, far beyond what is computationally feasible. A scalable architecture must therefore couple local molecular interactions with long-range spatial context without attending over all pairs of transcripts.

Beyond scalability, there is a representation choice that precedes architecture design: **what should be the intrinsic spatial unit of tissue representation?** Existing iST-based models typically pool transcripts into predefined cells, spots, neighborhoods, or larger spatial units before encoding (Tejada-Lapuerta et al., 2025; Blampey et al., 2025; Birk et al., 2026; Zhao et al., 2025). Such naive pooling discards the molecular geometry within those units and, for cell-based models, also inherits transcript-to-cell assignments from segmentation (Ishaque et al., 2026). Yet tissue has no single intrinsic unit of analysis. Depending on the downstream biological question, the same section may be studied at the level of cells, multicellular niches, tissue regions, or the whole section (Palla et al., 2022; Adler et al., 2023). Fixing the spatial unit before the question is known rules out finer queries and allows regions that cut across units to be represented only approximately. A tissue foundation model should therefore learn context across scales while deferring the choice of scale to query time (Bommasani et al., 2021; Bunne et al., 2024).

Our contributions. To address these challenges, we introduce the concept of *tissue embedding fields*. The core idea is to assign a contextual embedding on a shared unit hypersphere to every measured transcript in the tissue section. Any spatial region—a cell, a niche, a tissue compartment, or an entire tissue section—can then represented by summing the embeddings of the transcripts it contains and normalizing the result. We ensure these sums compose additively across disjoint regions, so

the spatial unit of analysis is deferred to query time rather than committed during training. Once a tissue section is encoded, a single cached field supports downstream analyses at all scales without re-encoding the tissue (Fig. 1c).

Computing such fields for iST data, where tissue sections can contain billions of transcripts, requires an encoder whose cost does not grow quadratically with transcript count. We introduce LAMINAR, a local-to-global encoder that separates short-range molecular interactions from long-range tissue context. A graph neural network exchanges information among nearby transcripts. A quadtree spatial Transformer then compresses progressively larger physical regions into fixed-size summaries while exchanging information between neighboring regions at each level of the hierarchy. Each transcript retrieves this multi-scale context through cross-attention to produce its final embedding. For a fixed architecture, the total number of interactions grows linearly with the number of transcripts.

We train Laminar with a joint-embedding predictive architecture (JEPA) (LeCun et al., 2022; Assran et al., 2023) matched to the compositional structure of the field. A target encoder, maintained as an exponential moving average of the online weights, computes the embedding field of the complete tissue; its regional embeddings, formed by the same sum-and-normalize readout used downstream, serve as prediction targets at multiple physical scales. The online encoder receives a spatially or gene-masked view and must recover these targets from the remaining context (Fig. 1b).

We pretrain Laminar on a public corpus of 632 human iST sections, containing 96 million cells and 22.2 billion transcripts. We evaluate Laminar's frozen field with a common linear probe on eight transfer tasks spanning cell, niche, and tissue scales, against existing fixed-scale foundation models (Jiang et al., 2026). A single encoder, whose field is pooled at different spatial supports depending on the task, is strongest when the label reflects tissue-scale organization, achieving the best scores on pathology region and methylation subclass, and remains competitive when the label depends on the molecular profile of individual cells. The results show that tissue representations need not be tied to a predetermined spatial unit; a unified embedding field enables reasoning across the full hierarchy of tissue organization.

2 METHOD

An iST tissue section measures individual RNA transcripts, each recorded as a gene identity and a spatial coordinate. We represent a section as a set of $N_{\mathcal{X}}$ such measurements,

$$\mathcal{X} = \{(p_i, g_i)\}_{i=1}^{N_{\mathcal{X}}}, \qquad p_i \in \Omega_{\mathcal{X}} \subset \mathbb{R}^2, \quad g_i \in \{1, \dots, G\}, \tag{1}$$

where p_i is the spatial coordinate of transcript i in micrometers and g_i is its gene identity within a vocabulary of G genes. Typical sections span 10^2–10^4 μm at submicrometer transcript resolution, profile 10^2–10^4 genes, and can contain up to 10^9 detected transcripts (Moffitt et al., 2022).

Our goal is to compute a contextual representation at every observed transcript such that spatial regions can later be represented by pooling the embeddings they contain, without re-encoding the tissue. Section 2.1 formalizes this representation, Section 2.2 introduces the Laminar encoder, and Section 2.3 describes its self-supervised training objective.

2.1 TISSUE EMBEDDING FIELDS

Let d be the embedding dimension, and $\mathbb{S}^{d-1} = \{z \in \mathbb{R}^d : \|z\|_2 = 1\}$ be the unit sphere. We aim to construct a permutation-equivariant encoder E_θ that, per each transcript, assigns an embedding on the unit sphere:

$$E_\theta(\mathcal{X}) = \big(z_i(\mathcal{X}; \theta)\big)_{i=1}^{N_{\mathcal{X}}} \in (\mathbb{S}^{d-1})^{N_{\mathcal{X}}}. \tag{2}$$

Each $z_i(\mathcal{X}; \theta)$ may depend on the surrounding tissue at coordinate p_i. We abbreviate $z_i = z_i(\mathcal{X}; \theta)$.

For any measurable region $A \subseteq \Omega_{\mathcal{X}}$, define its *regional state* $S_{\mathcal{X}}(A) = \sum_{i:p_i \in A} z_i$ as the summed embedding state, and its number of transcripts $n_{\mathcal{X}}(A) = |\{p_i : p_i \in A\}|$. For $S_{\mathcal{X}}(A) \neq 0$, the corresponding *regional embedding* is the (normalized) direction

$$\Phi_{\mathcal{X}}(A) = \frac{S_{\mathcal{X}}(A)}{\|S_{\mathcal{X}}(A)\|_2} \in \mathbb{S}^{d-1}. \tag{3}$$

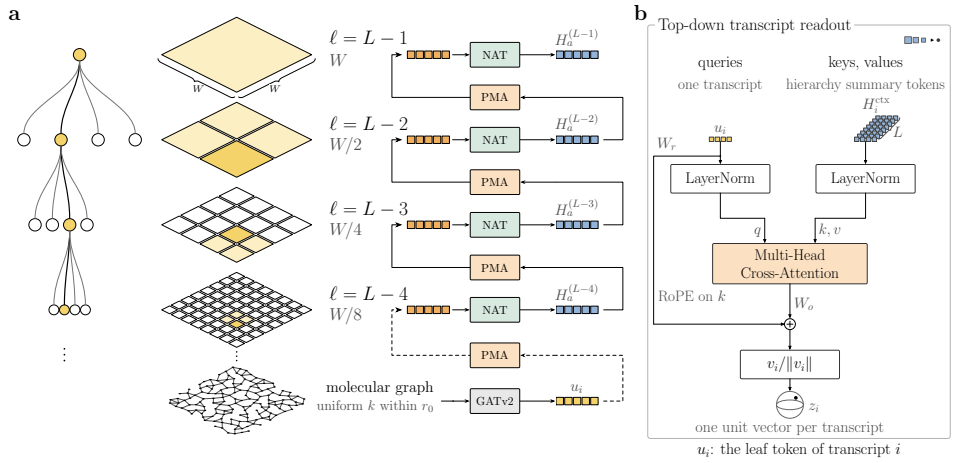

Figure 2: **Laminar encoder. (a)** A radius-capped k-NN molecular graph captures short-range transcript context with GATv2. A bottom-up quadtree then compresses information across progressively larger spatial scales with Pooling by Multihead Attention (PMA), while Neighborhood Attention exchanges information laterally between neighboring tiles at each level. **(b)** A top-down transcript readout cross-attends from each transcript to the selected hierarchy levels and combines this multi-scale context with its local molecular state to produce the final unit embedding z_i.

For $n_{\mathcal{X}}(A) > 0$, the *directional coherence* $R_{\mathcal{X}}(A) = \|S_{\mathcal{X}}(A)\|_2 / n_{\mathcal{X}}(A) \in [0, 1]$ measures alignment among the transcript embeddings in A. We define the *tissue embedding field* as $\mathcal{F}_\theta[\mathcal{X}](A) = (\Phi_{\mathcal{X}}(A), R_{\mathcal{X}}(A), n_{\mathcal{X}}(A))$. The regional state composes additively: for disjoint regions A and B, $S_{\mathcal{X}}(A \cup B) = S_{\mathcal{X}}(A) + S_{\mathcal{X}}(B)$ and $n_{\mathcal{X}}(A \cup B) = n_{\mathcal{X}}(A) + n_{\mathcal{X}}(B)$. Thus $\mathcal{F}_\theta[\mathcal{X}]$ is a single query-independent representation of the tissue. Cells, niches, or larger regions are obtained by summing the same transcript field and normalizing only at readout. No further encoder computation is required. Appendix C.2 gives the complete readout definition and edge cases.

2.2 Laminar Encoder

We introduce Laminar to realize the encoder defined in Equation 2. Laminar exchanges information over distances ranging from neighboring molecules to hundreds of micrometers without applying dense attention to all transcripts. We separate this computation into three stages. A radius-capped k-NN molecular graph captures short-range context. A bottom-up quadtree then compresses information across spatial scales while allowing lateral exchange within each scale. Finally, a top-down readout returns this multi-scale context to each transcript.

Short-range molecular context. Let G be the size of the model's global gene vocabulary and $M_{\text{gene}} \in \mathbb{R}^{G \times d_g}$ its learned embedding table, where d_g is the gene-embedding dimension. Transcript i is initialized as $x_i = M_{\text{gene}}[g_i] \in \mathbb{R}^{d_g}$. We construct a radius-capped k-nearest-neighbor graph in physical space. For transcript i, let

$$\mathcal{B}_i(r_0) = \{j \neq i : \|p_j - p_i\|_2 \leq r_0\}, \qquad \mathcal{N}_i^{\text{mol}} = \text{kNN}_{k_{\text{nn}}}(i \mid \mathcal{B}_i(r_0)) \,. \tag{4}$$

Here $r_0 > 0$ is the maximum physical edge length and $\text{kNN}_{k_{\text{nn}}}(i \mid \mathcal{B}_i)$ returns the $\min(k_{\text{nn}}, |\mathcal{B}_i|)$ transcripts in $\mathcal{B}_i$ closest to i in Euclidean distance. Each transcript therefore exchanges information with at most k_{nn} nearby molecules and no graph edge exceeds r_0.

Applying L_{loc} residual GATv2 layers (Brody et al., 2021) gives a contextual state $h_i \in \mathbb{R}^{d_h}$ (Fig. 2a). We retain a direct path from the original gene embedding and form $u_i = \left[W_h^{\text{loc}} h_i \,\|\, W_x^{\text{loc}} x_i\right] \in \mathbb{R}^{d_{\text{tile}}}$, where $\|$ denotes concatenation and W_h^{loc} and W_x^{loc} are learned projections. Appendix C.3 gives the exact graph construction and local attention.

Bottom-up multi-scale hierarchy. The molecular graph captures short-range context. We process long-range information limited to square crops $C \subseteq \mathbb{R}^2$ of fixed width W_{crop} to set the physical context range and control per-pass computational cost. To propagate information over larger physical distances, we partition each crop into a quadtree with tile widths $s_\ell = 2^\ell s_0$ for $\ell \in \{0, \ldots, L-1\}$,

where s_ℓ is the tile width at hierarchy level ℓ, s_0 is the finest width, and L is the number of levels (Fig. 2a). Each tile a is represented by K tokens $H_a^{(\ell)} \in \mathbb{R}^{K \times d_{\text{tile}}}$. Information propagates upward by repeated cross-attention pooling. At the finest level, K learned queries summarize the transcript features inside a tile. At every higher level, the same operation summarizes the tokens of its children

$$X_a^{(\ell)} = \begin{cases} (u_i)_{p_i \in a}, & \ell = 0, \\ \mathrm{cat}_{b \in \mathrm{ch}(a)}\, H_b^{(\ell-1)}, & \ell > 0, \end{cases} \qquad \widetilde{H}_a^{(\ell)} = \mathrm{PMA}_K\Big(X_a^{(\ell)}\Big). \tag{5}$$

Here $\mathrm{ch}(a)$ denotes the child tiles, cat concatenates their token sequences, and PMA_K denotes cross-attention from K learned pooling queries to the input sequence (Lee et al., 2019). This recursively compresses increasingly large physical regions into a fixed number of tokens.

The quadtree imposes rigid partitions that are favored computationally but lack biological meaning. We therefore follow each pooling stage with Neighborhood Attention (Hassani et al., 2023) over the 3×3 tile neighborhood at the same level. This lets information propagate laterally across tile boundaries before the next compression step and reduces sensitivity to the particular quadtree partition. The resulting state $H_a^{(\ell)}$ is retained for readout and, unless $\ell = L - 1$, passed upward to the parent. The hierarchy therefore alternates cross-scale compression with same-scale spatial exchange. Count and level embeddings, attention blocks, and spatial positional encoding are given in Appendix C.4–C.5.

Top-down transcript readout. Finally, to produce the contextual transcript embedding, we read information back from the spatial hierarchy to each transcript via cross-attention. Let $\mathcal{J}_{\text{out}} \subseteq \{0, \ldots, L-1\}$ be the hierarchy levels exposed to the transcript readout. For transcript i, let H_i^{ctx} collect the hierarchy tokens from the four-tile neighborhood around p_i at every $\ell \in \mathcal{J}_{\text{out}}$, and let w_i denote the associated spatial weights. We return the selected context to the transcript by cross-attention

$$c_i = \mathrm{Attn}\big(W_q u_i, H_i^{\text{ctx}}; w_i\big), \qquad v_i = W_r u_i + W_o c_i, \qquad z_i = \frac{v_i}{\|v_i\|_2}. \tag{6}$$

Here W_q projects the transcript query, W_r maps its local state directly to the field space, and W_o projects the attended hierarchy context. The direct term preserves short-range molecular information, while c_i returns context from multiple physical scales. Appendix C.6 gives the exact readout, attention logits, and positional terms. In Appendix C.7, we demonstrate that the number of neural interactions grows linearly with the number of transcripts in a crop.

2.3 Training Tissue Embedding Fields with JEPA

The hierarchy exposes representations at several physical scales. We train those representations to describe the same compositional field that is queried after pretraining. In particular, every hierarchy level predicts complete regional state at its own scale, while the final transcript field is required to produce the same regional directions under corruption. The training objective is

$$\mathcal{L} = \underbrace{\mathcal{L}_{\text{pred}}}_{\text{multi-scale JEPA prediction}} + \lambda_{\text{field}} \underbrace{\mathcal{L}_{\text{field}}}_{\text{field composition}} + \lambda_{\text{iso}} \underbrace{\mathcal{L}_{\text{iso}}}_{\text{anti-collapse}}. \tag{7}$$

Here $\lambda_{\text{field}} \geq 0$ and $\lambda_{\text{iso}} \geq 0$ weight the field and isotropy terms. This objective makes the hierarchy multi-scale by construction, rather than relying on architectural depth alone.

Online and target fields. For each crop, an online encoder E_θ receives a corrupted view $\mathcal{X}^v$, while a target encoder E_ξ receives the complete crop. We use $v \in \mathcal{V} = \{\text{spatial}, \text{gene}\}$. Spatial corruption removes transcripts inside sampled rectangles. Gene corruption removes complete gene identities. Each corrupted view rebuilds its molecular graph, tiles, and tile counts from the transcripts that remain. The target encoder is updated only by exponential moving average $\xi \leftarrow m_t \xi + (1 - m_t)\theta$, where θ and ξ denote the online and target parameters and $m_t \in [0, 1)$ is the EMA momentum at optimization step t.

For a query region A, let $\mathcal{I}_A = \{i : p_i \in A\}$ be its transcripts in the complete crop. If $z_i^\xi \in \mathbb{S}^{d-1}$ is the target embedding of transcript i, the regional target is $y_A = \mathrm{sg}(S_A^\xi / \|S_A^\xi\|_2) \in \mathbb{S}^{d-1}$ where $S_A^\xi = \sum_{i \in \mathcal{I}_A} z_i^\xi$ and sg denotes stop-gradient. The target is therefore exactly the regional readout of the complete target field.

We assign each target a nonnegative stop-gradient weight ω_A derived from its directional coherence. For any query set $\mathcal{Q}$, write $\langle f_A \rangle_{\mathcal{Q},\omega} = (\sum_{A\in\mathcal{Q}} \omega_A f_A)/\max\{\sum_{A\in\mathcal{Q}} \omega_A, \epsilon\}$, where $\epsilon > 0$ is a numerical stabilizer. The exact coherence correction, weight schedule, and query sampling distributions are given in Appendix C.8.

Multi-scale prediction. For each view v and hierarchy level ℓ, we sample regional queries $\mathcal{Q}^{\mathrm{pred}}_{v,\ell}$ at the physical scale represented by that level. For each region $A \in \mathcal{Q}^{\mathrm{pred}}_{v,\ell}$, an encoder P_ϕ takes a query token $q(A,\ell,v)$ and the online hierarchy tokens $H^{(\ell)}_{\theta,v}$. The query encodes the region's center, side lengths, level, and masking type, and cross-attends to the hierarchy tokens as keys and values. We normalize $P_\phi(q(A,\ell,v), H^{(\ell)}_{\theta,v})$, apply a stochastic angular perturbation, and decode the perturbed vector with a two-layer perceptron h_ψ. Normalizing the decoder output gives the regional prediction $\widehat{y}_{A,v,\ell} \in \mathbb{S}^{d-1}$. We compare these predictions with the complete-crop targets y_A using $\mathcal{L}_{\mathrm{pred}} = \sum_{\ell=0}^{L-1} \sum_{v\in\mathcal{V}} \left\langle 1 - \widehat{y}^{\top}_{A,v,\ell} y_A \right\rangle_{\mathcal{Q}^{\mathrm{pred}}_{v,\ell},\omega}$. Thus, each hierarchy level is trained to predict regional embeddings at its own physical scale. Predictor queries, the stochastic prediction channel, and the training-only decoder are specified in Appendix C.8.

Field composition. The predictor supervises hierarchy tokens before the final transcript readout. We therefore impose the same regional target directly on the deployed field. Let $\mathcal{I}_{A,v} \subseteq \mathcal{I}_A$ index the transcripts in region A that survive view v, with online embeddings $z^{\theta}_{i,v}$. Their regional sum is $S^{\theta}_v(A) = \sum_{i\in\mathcal{I}_{A,v}} z^{\theta}_{i,v}$, and their regional direction is $\Phi^{\theta}_v(A) = S^{\theta}_v(A)/\|S^{\theta}_v(A)\|_2$. For the valid composition-query set $\mathcal{Q}^{\mathrm{cons}}_v$, we define $\mathcal{L}_{\mathrm{field}} = \sum_{v\in\mathcal{V}} \left\langle 1 - \Phi^{\theta}_v(A)^{\top} y_A \right\rangle_{\mathcal{Q}^{\mathrm{cons}}_v,\omega}$. This ties the internal multi-scale hierarchy to the representation actually used downstream. Regional sums of corrupted transcript embeddings must recover the same direction as the corresponding region in the complete target field.

Anti-collapse. Finally, regional embeddings should use the representation space rather than converge to a common direction. For view v, let $\{\phi_b\}_{b=1}^{B_v} \subset \mathbb{S}^{d-1}$ be the online regional directions of the B_v valid composition queries in the global training step. Define their mean $\bar{\phi}_v = B_v^{-1} \sum_{b=1}^{B_v} \phi_b$ and second moment $M_v = B_v^{-1} \sum_{b=1}^{B_v} \phi_b \phi_b^{\top}$. We penalize departures from zero mean and an isotropic second moment

$$\mathcal{L}_{\mathrm{iso}} = \sum_{v\in\mathcal{V}} \left(\|\bar{\phi}_v\|_2^2 + \lambda_{\mathrm{mom}} \left\| M_v - \frac{1}{d} I_d \right\|_F^2 \right). \tag{8}$$

Here $\lambda_{\mathrm{mom}} \geq 0$ weights the second-moment term, $I_d \in \mathbb{R}^{d\times d}$ is the identity matrix, and $\|\cdot\|_F$ is the Frobenius norm.

Together, $\mathcal{L}_{\mathrm{pred}}$ trains Laminar to capture efficient scale embeddings at each hierarchy level, $\mathcal{L}_{\mathrm{field}}$ guarantees that the tissue embedding field described in Equation 3 is realized, and $\mathcal{L}_{\mathrm{iso}}$ prevents a degenerate collapsed solution. Downstream readout requires only the online encoder E_θ. Corruption distributions, region sampling, target weighting, predictor architecture, stochastic prediction, validity rules, distributed moments, and all objective hyperparameters are given in Appendix C.8.

3 Experiments

We evaluate a single frozen Laminar encoder on eight transfer tasks spanning cells, multicellular neighborhoods, tissue regions, and tissue-microarray cores (cylindrical biopsies of $\sim$1 mm punched from larger tumors and arrayed on common slides for parallel profiling). The pretraining corpus and every evaluation cohort are fully disjoint—no section, study, or donor appears in both, and all evaluation labels are withheld during pretraining. The experiments therefore test whether the tissue embedding field, learned from transcript coordinates and gene identities alone, transfers to unseen tissues, organs, and biological endpoints across spatial scales.

Because existing foundation models produce one embedding per cell, we adopt a cell-level evaluation path for a like-for-like comparison. Each cell is represented as the regional embedding of its transcripts (Eq. 3), and for larger units these cell embeddings are averaged with the same rule applied to every

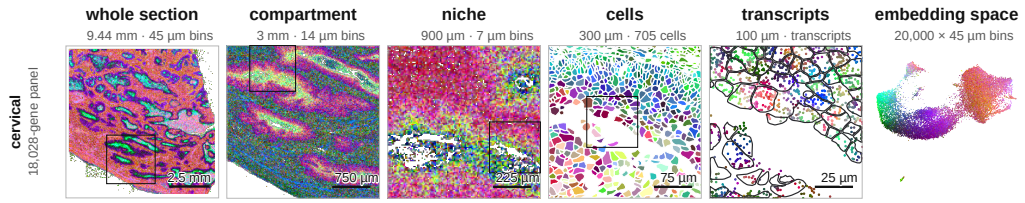

Figure 3: **Querying the tissue embedding field across spatial supports.** A whole-transcriptome FFPE cervical carcinoma section (18,028-gene panel) is encoded once. The cached transcript embeddings are then pooled from 45 μm bins down to individual transcripts without re-encoding the tissue. RGB channels show the three leading PCA components with the highest spatial autocorrelation at each support.

comparator. Cell segmentations enter only after encoding, through this shared aggregation, and never as encoder input. Readout details are in Appendix G.

Pretraining. The evaluated encoder (95.1M parameters, 1,024-dimensional field) is pretrained on 632 human iST sections from 76 public studies spanning 13 organs and five platforms, totaling 22.2 billion transcripts and 96.8 million cells (Appendix D). Cell counts describe the corpus but segmentations are not encoder inputs—the encoder receives only each transcript's gene identity and spatial coordinate. H&E images and all pathology, clinical, or molecular labels are withheld. Training follows a four-stage curriculum from a single whole-transcriptome section to the full corpus (Appendices E–F).

Tasks. The benchmark spans four held-out cohorts—94 sections and 11 tissue-microarray slides—totaling approximately 1.7 million annotated units across eight tasks at three physical scales. None of these cohorts appear in the pretraining manifest. At cell scale, *cell type* classifies six broad identities for 1.1M cells across 59 sections from 13 organs (Jaume et al., 2024), and *gene imputation* recovers the expression of 149–158 genes withheld from the encoder input for 146,056 cells in five lung sections. At multicellular scale, *pathology region* classifies expert-annotated tissue compartments (Vannan et al., 2025) for 231,546 cells across 20 lung-fibrosis sections, *breast niche* classifies annotated spatial niches for 11,579 cells from four breast-cancer sections, and *region composition* recovers the expression profile (339 genes) of 935 held-out 200 μm tiles from the surrounding tissue. At tissue-core scale, meningioma tissue-microarray cores are classified by expert-assigned *WHO grade* (300 cores, 3 classes), by *methylation subclass* (298 cores, 5 classes), and by *malignancy class* (312 cores, 3 classes). These three endpoints are clinically defined tumor classifications assigned by neuropathologists, and the methylation labels are confirmed by an independent DNA methylation array. The evaluation therefore covers a wide range of tissues, cohort sizes, label sources, and physical scales, all under a single frozen encoder and a uniform protocol. Full cohort descriptions and fold construction are given in Appendix G.

Comparison methods. The primary comparison is with frozen pretrained models—Nicheformer (Tejada-Lapuerta et al., 2025), scGPT-spatial (Wang et al., 2025), TERRA (Birk et al., 2026), UCE (Rosen et al., 2026), Novae (Blampey et al., 2025), and SToFM (Zhao et al., 2025)—each used with its released checkpoint and encoder frozen. All produce one embedding per cell, and for larger targets their per-cell embeddings are averaged with the same rule used for Laminar. We separately report unsupervised in-domain methods—scVI (Lopez et al., 2018), BANKSY (Singhal et al., 2024), and CellCharter (Varrone et al., 2023)—which learn representations from the evaluation cohort without task labels and therefore do not perform frozen transfer. Normalized raw counts and their 64-component PCA serve as non-learned baselines. Comparator checkpoints and preprocessing are in Appendix G.

Probe protocol. Every method is assessed with the same single-hidden-layer probe,

$$\mathrm{Linear}(d, 256) \rightarrow \mathrm{GELU} \rightarrow \mathrm{Linear}(256, d_{\mathrm{out}}),$$

trained for 2,000 AdamW steps from three independent initializations. Standardization is fitted on the training fold only. Representations wider than 4,096 dimensions are passed through a fixed random projection before probing, which applies only to BANKSY and the raw-counts baseline. Cross-validation folds hold out entire tissue sections, or entire slides for the meningioma microarray (Appendix G).

Table 1: Transfer across biological scales. Classification tasks report chance-corrected balanced accuracy, gene and region imputation report profile cosine, and entries are the median across evaluation folds ± the sample SD of fold scores; **1st**, **2nd**, and **3rd** indicate the three highest scores, and lower mean rank is better.

		HEST-1k, 13 organs 59 sections	lung fibrosis 20 sections			breast cancer 4 sections	meningioma TMA 11 slides			
method	**setting**	**cell type** cell	**gene imputation** cell	**pathology region** niche	**region composition** 200 μm tile	**breast niche** niche	**WHO grade** core	**methylation subclass** core	**malignancy class** core	**mean rank** ↓
Nicheformer	Frozen pretrained	0.688±.08	0.664±.08	0.373±.19	0.811±.05	0.660±.21	0.133±.20	0.507±.21	0.478±.14	7.1
scGPT-spatial	Frozen pretrained	**0.733**±.09	**0.721**±.07	0.369±.20	0.824±.08	0.696±.19	0.314±.22	0.431±.23	0.400±.15	**5.5**
TERRA-96	Frozen pretrained	0.679±.11	0.632±.08	0.343±.21	0.810±.06	**0.729**±.29	0.248±.21	0.438±.23	0.483±.18	7.1
TERRA-112	Frozen pretrained	0.677±.10	0.606±.08	0.384±.21	**0.834**±.05	0.703±.23	0.128±.27	0.460±.23	0.379±.23	8.1
UCE-4L	Frozen pretrained	0.694±.09	**0.686**±.07	0.350±.19	0.802±.05	0.609±.19	0.308±.22	0.391±.20	0.457±.20	7.8
Novae	Frozen pretrained	0.373±.08	0.611±.09	0.305±.22	**0.862**±.04	**0.794**±.13	0.154±.24	0.460±.15	0.339±.14	8.7
SToFM	Frozen pretrained	0.640±.10	0.658±.09	0.316±.22	0.801±.04	0.542±.19	0.194±.14	0.471±.22	**0.511**±.25	8.8
SToFM-CE	Frozen pretrained	**0.697**±.09	**0.669**±.09	0.392±.19	0.821±.06	0.566±.18	0.193±.16	0.477±.22	**0.507**±.26	**6.0**
scVI	Active	0.674±.09	0.521±.11	**0.432**±.20	0.779±.05	0.694±.13	**0.327**±.19	0.518±.20	0.461±.21	6.9
BANKSY	Active	0.688±.12	–	0.299±.23	0.733±.05	0.686±.12	0.239±.17	**0.529**±.21	**0.559**±.19	7.5
CellCharter	Active	0.152±.07	0.509±.11	0.105±.26	0.820±.05	0.029±.05	0.165±.16	0.300±.19	0.162±.17	12.2
Raw counts	Baseline	**0.698**±.10	0.208±.04	0.371±.19	0.780±.07	0.663±.20	**0.450**±.19	**0.540**±.23	0.467±.16	6.4
Counts PCA-64	Baseline	**0.698**±.10	0.208±.04	**0.398**±.19	0.780±.07	0.665±.19	**0.448**±.16	0.480±.20	0.389±.16	6.9
Laminar (ours)	Frozen pretrained	0.596±.12	0.660±.09	**0.445**±.18	**0.869**±.03	**0.705**±.21	0.278±.23	**0.580**±.26	0.253±.24	**5.2**

Evaluation and reporting. Accuracy for classification tasks is quantified as chance-corrected balanced accuracy, which is 0 for random guessing and 1 for perfect classification. The two expression-recovery tasks are evaluated by cosine similarity between predicted and measured profiles (gene imputation per cell, region composition per tile). Within each fold, scores of the three probe initializations are first averaged. Table entries are then the median of fold scores, and values after ± give the sample standard deviation of fold scores, describing between-section variability rather than a confidence interval. Full metric definitions appear in Appendix G.

Results. Table 1 presents the results by evaluation cohort. Laminar achieves the lowest mean rank (5.2) among all methods across the eight tasks, placing first on pathology region (0.445), region composition (0.869), and methylation subclass (0.580), and third on breast niche (0.705).

The results reveal a pattern tied to the spatial scale of the target. At cell scale, Laminar reaches 0.596 on cell type. Raw counts (0.698) and scGPT-spatial (0.733) score substantially higher, showing that the molecular profile of an isolated cell already carries a strong signal for cell identity and that cell-level pretrained models exploit this effectively. This is consistent with the ablation study (Table 2), where gene identity alone achieves 0.589 on cell type, confirming that molecular identity rather than spatial context drives this task.

The tissue embedding field becomes most competitive when the target depends on spatial organization beyond a single cell. On pathology region, Laminar (0.445) is top-scoring and outperforms scVI (0.432), which is fitted directly on the evaluation cohort. On breast niche, Laminar ranks third (0.705), behind Novae (0.794) and TERRA-96 (0.729). At tissue-core scale, Laminar achieves the best score on methylation subclass (0.580). For WHO grade and malignancy class, the non-learned baselines and BANKSY lead, respectively. The advantage of the tissue embedding field is task-dependent and strongest when the downstream label reflects tissue-scale organization rather than the molecular profile of isolated cells.

These results need two qualifications. For region composition, predicting the training-fold mean already reaches 0.876, while copying the surrounding ring reaches 0.963. The task is therefore strongly driven by spatial smoothness, and we treat it mainly as a diagnostic. The mean rank of 5.2 averages over the eight task columns, which cover different endpoints and scales. For this reason, we focus our interpretation on the individual task results.

Ablations. We assess Laminar's main design choices with a reduced model (3.2M parameters) trained from scratch on a single whole-transcriptome breast-cancer section for 1,500 steps (8.4×10^9 transcripts processed). The Laminar row in Table 2 refers to this single-section model, not the full pretrained encoder. Variant definitions and full results are in Appendix H.

The ablations separate two functional roles of the representation. The transcript-resolved readout preserves cellular identity—assigning one embedding to all transcripts in a 16 μm tile drops cell-type accuracy from 0.567 to 0.469, while pathology-region accuracy is nearly unchanged. Gene identity alone shows the complementary pattern, with cell-type accuracy rising to 0.589 but pathology-region accuracy falling to 0.350. Cell identity therefore depends on preserving transcript-level gene information, while tissue-region classification requires the spatial context provided by the encoder hierarchy.

For pathology region, the largest drops come from replacing the quadtree hierarchy with a single level of 128 μm patches (0.332 vs. 0.387) and from removing the field composition objective (0.345 vs. 0.387). Both the multi-scale hierarchy and the composition training signal contribute independently to encoding tissue-scale organization into the field.

These ablations isolate the effect of individual design choices under controlled single-section training. Whether these effects persist, amplify, or interact differently at full model scale and over longer multi-tissue pretraining remains an open empirical question.

Table 2: Ablations. Chance-corrected balanced accuracy (median ± s.d.).
Shading: Δ vs. Laminar neg. pos.

	cell type	pathology region
no composition loss	0.547±.12	0.345±.20
gene identity only	0.589±.11	0.350±.19
molecular graph only	0.557±.12	0.387±.20
tile-level readout	0.469±.11	0.382±.20
flat ViT (128 μm)	0.505±.13	0.332±.20
frozen gene table	0.567±.12	0.379±.19
genes shuffled	0.523±.11	0.400±.20
Laminar (3.2M)	**0.567**±.12	**0.387**±.19

4 Discussion and Conclusion

The core advance of the proposed tissue embedding fields is to move the choice of representational support from pretraining to query time. Laminar estimates a transcript-resolved contextual field, without committing to a cell, niche, or tissue representation *a priori*. The representation of any region, from a single cell to a whole section, is simply the normalized sum of its cached transcript embeddings; the encoder is neither re-run nor retrained. Laminar makes computing such fields tractable at iST scale with billions of transcripts. A local transcript graph and a quadtree hierarchy that compresses progressively larger regions into a fixed number of summary tokens keep the cost linear in the number of transcripts, and a cross-attention readout then returns this multi-scale context to every transcript. Self-supervised training with JEPA is matched to the same interface. One complete target field provides regional targets at many locations and physical scales, while masked views require those query-defined representations to be predicted from context. The field is therefore the reusable model output, while the downstream representations are instantiated by the query. This carries the embedding-field concept of AlphaEarth Foundations (Brown et al., 2025) from Earth observation to tissue biology, where the field is defined at measured molecules and equipped with an explicit compositional readout.

Limitations and future directions. Laminar is a first step toward tissue embedding fields, and several limitations remain. We introduce the representation and a scalable self-supervised framework for learning it, but how encoders of such fields are best trained and adapted is still an open question. The current model operates on two-dimensional transcript point fields, its context is bounded by the encoded crop, and long-range information is compressed through the hierarchy. Longer pretraining over more diverse tissues, biological states, cohorts, and measurement technologies is an immediate direction, following evidence that broad self-supervised pretraining improves transfer across biological settings (Rosen et al., 2026; Wenckstern et al., 2026). More fundamentally, the input side should become as flexible as the query side through a cross-modal tissue representation. Different spatial measurements, such as transcripts, proteins, and morphology, could be encoded into a common field that retains the same query-defined regional readout. A natural next step is to model how tissue responds to perturbation. Because the field represents each cell in its spatial context, it could serve as a substrate for predicting how genetic, chemical, or environmental perturbations affect cells and propagate across niches, a central goal of AI virtual cells (Bunne et al., 2024; Chen et al., 2025). Finally, we evaluate frozen transfer only. Task-specific post-training can unlock substantial additional information from pretrained biological representations (Theodoris et al., 2023).

AI USE STATEMENT

We used generative AI tools as research assistants at several stages of this work. Claude was used mainly for literature retrieval, implementation support, organizing public datasets, and assisting with execution and debugging of the benchmark pipeline. ChatGPT, Claude, and Gemini were used to help express research ideas in precise mathematical form, examine alternative formulations, and check mathematical statements and derivations. ChatGPT was also used extensively to draft, restructure, and edit the manuscript.

The research questions, Laminar architecture and training design, benchmark construction, and scientific interpretation of the results were determined by the authors. AI-assisted code and mathematical formulations were reviewed and tested by the authors, and AI-assisted text was revised before inclusion in the manuscript. All numerical results reported in the paper were obtained from executed experiments rather than generated by an AI system.

REPRODUCIBILITY STATEMENT

Appendix C provides the complete encoder, field readout, training objective, crop sampling procedure, and whole-section inference pipeline. Appendix D details the public pretraining corpus, including the source, accession, and licence of each study, along with cohorts excluded from pretraining for downstream evaluation. Appendices E and F specify model configuration, training schedule, hardware, and checkpoint used in experiments. Appendix G fully describes the evaluation pipeline, including comparator checkpoints, section encoding, task definitions, fold construction, probes, metrics, and software versions. Appendix H gives the ablation protocol.

Our release will link reported results with the corresponding checkpoint hash, resolved configuration, data manifest, and scorer version. When permitted by source data licences, we will also release out-of-fold predictions.

ETHICS STATEMENT

This study uses publicly available spatial transcriptomics datasets together with an anonymized in-house cohort for downstream evaluation. Public datasets were used under the access conditions and licences specified by their source repositories. The in-house data were anonymized before analysis and were used only for benchmarking within this study. We do not release patient-level labels or molecular measurements from that cohort.

Clinical and pathological annotations are used only as downstream evaluation targets. The experiments do not establish clinical validity or suitability for diagnostic use. Differences in acquisition protocols, cohort composition, and available metadata may also limit generalization beyond the evaluated datasets.

REFERENCES

Dimitris Achlioptas. Database-friendly random projections: Johnson-lindenstrauss with binary coins. *Journal of Computer and System Sciences*, 66(4):671–687, 2003. ISSN 0022-0000. doi: 10.1016/s0022-0000(03)00025-4. URL http://dx.doi.org/10.1016/s0022-0000(03)00025-4.

Miri Adler, Arun R. Chavan, and Ruslan Medzhitov. Tissue biology: In search of a new paradigm. *Annual Review of Cell and Developmental Biology*, 39(1):67–89, 2023. ISSN 1530-8995. doi: 10.1146/annurev-cellbio-120420-113830. URL http://dx.doi.org/10.1146/annurev-cellbio-120420-113830.

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15619–15629. IEEE, 2023. doi: 10.1109/cvpr52729.2023.01499. URL http://dx.doi.org/10.1109/cvpr52729.2023.01499.

Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300, 1995.

Salil S. Bhate, Graham L. Barlow, Christian M. Schürch, and Garry P. Nolan. Tissue schematics map the specialization of immune tissue motifs and their appropriation by tumors. *Cell Systems*, 13(2): 109–130.e6, 2022. ISSN 2405-4712. doi: 10.1016/j.cels.2021.09.012. URL http://dx.doi.org/10.1016/j.cels.2021.09.012.

Sebastian Birk, Mohammad Vali Sanian, Amirhossein Vahidi, Samuel Ogden, Daniyal J Jafree, Adib Miraki Feriz, Carlo Leonardi, Arpit Merchant, Zichen He, Lloyd Steele, et al. Multi-scale modeling of human tissues from spatial transcriptomics with terra. *BioRxiv*, pp. 2026–07, 2026.

Quentin Blampey, Hakim Benkirane, Nadège Bercovici, Kevin Mulder, Grégoire Gessain, Florent Ginhoux, Fabrice André, and Paul-Henry Cournède. Novae: a graph-based foundation model for spatial transcriptomics data. *Nature Methods*, 22(12):2539–2550, 2025. ISSN 1548-7105. doi: 10.1038/s41592-025-02899-6. URL http://dx.doi.org/10.1038/s41592-025-02899-6.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. On the opportunities and risks of foundation models, 2021. URL https://arxiv.org/abs/2108.07258.

Shaked Brody, Uri Alon, and Eran Yahav. How attentive are graph attention networks? *arXiv preprint arXiv:2105.14491*, 2021.

Christopher F. Brown, Michal R. Kazmierski, Valerie J. Pasquarella, William J. Rucklidge, Masha Samsikova, Chenhui Zhang, Evan Shelhamer, Estefania Lahera, Olivia Wiles, Simon Ilyushchenko, Noel Gorelick, Lihui Lydia Zhang, Sophia Alj, Emily Schechter, Sean Askay, Oliver Guinan, Rebecca Moore, Alexis Boukouvalas, and Pushmeet Kohli. Alphaearth foundations: An embedding field model for accurate and efficient global mapping from sparse label data, 2025. URL https://arxiv.org/abs/2507.22291.

Charlotte Bunne, Yusuf Roohani, Yanay Rosen, Ankit Gupta, Xikun Zhang, Marcel Roed, Theo Alexandrov, Mohammed AlQuraishi, Patricia Brennan, Daniel B Burkhardt, et al. How to build the virtual cell with artificial intelligence: Priorities and opportunities. *Cell*, 187(25):7045–7063, 2024.

Kok Hao Chen, Alistair N. Boettiger, Jeffrey R. Moffitt, Siyuan Wang, and Xiaowei Zhuang. Spatially resolved, highly multiplexed rna profiling in single cells. *Science*, 348(6233), 2015. ISSN 1095-9203. doi: 10.1126/science.aaa6090. URL http://dx.doi.org/10.1126/science.aaa6090.

Richard J. Chen, Chengkuan Chen, Yicong Li, Tiffany Y. Chen, Andrew D. Trister, Rahul G. Krishnan, and Faisal Mahmood. Scaling vision transformers to gigapixel images via hierarchical self-supervised learning. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16123–16134. IEEE, 2022. doi: 10.1109/cvpr52688.2022.01567. URL http://dx.doi.org/10.1109/cvpr52688.2022.01567.

Richard J. Chen, Tong Ding, Ming Y. Lu, Drew F. K. Williamson, Guillaume Jaume, Andrew H. Song, Bowen Chen, Andrew Zhang, Daniel Shao, Muhammad Shaban, Mane Williams, Lukas Oldenburg, Luca L. Weishaupt, Judy J. Wang, Anurag Vaidya, Long Phi Le, Georg Gerber, Sharifa Sahai, Walt Williams, and Faisal Mahmood. Towards a general-purpose foundation model for computational pathology. *Nature Medicine*, 30(3):850–862, 2024. ISSN 1546-170X. doi: 10.1038/s41591-024-02857-3. URL http://dx.doi.org/10.1038/s41591-024-02857-3.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PmLR, 2020.

Weiqing Chen, Pengzhi Zhang, Tu N Tran, Yiwei Xiao, Shengyu Li, Vrutant V. Shah, Hao Cheng, Kristopher W. Brannan, Keith Youker, Li Lai, Longhou Fang, Yu Yang, Nhat-Tu Le, Jun-ichi Abe, Shu-Hsia Chen, Qin Ma, Ken Chen, Qianqian Song, John P. Cooke, and Guangyu Wang. A visual–omics foundation model to bridge histopathology with spatial transcriptomics. *Nature Methods*, 22(7):1568–1582, 2025. ISSN 1548-7105. doi: 10.1038/s41592-025-02707-1. URL http://dx.doi.org/10.1038/s41592-025-02707-1.

Mostafa Dehghani, Josip Djolonga, Basil Mustafa, Piotr Padlewski, Jonathan Heek, Justin Gilmer, Andreas Peter Steiner, Mathilde Caron, Robert Geirhos, Ibrahim Alabdulmohsin, et al. Scaling vision transformers to 22 billion parameters. In *International conference on machine learning*, pp. 7480–7512. PMLR, 2023.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, Bilal Piot, koray kavukcuoglu, Remi Munos, and Michal Valko. Bootstrap your own latent - a new approach to self-supervised learning. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 21271–21284. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/f3ada80d5c4ee70142b17b8192b2958e-Paper.pdf.

Ali Hassani, Steven Walton, Jiachen Li, Shen Li, and Humphrey Shi. Neighborhood attention transformer. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6185–6194. IEEE, 2023. doi: 10.1109/cvpr52729.2023.00599. URL http://dx.doi.org/10.1109/cvpr52729.2023.00599.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollar, and Ross Girshick. Masked autoencoders are scalable vision learners. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15979–15988. IEEE, 2022. doi: 10.1109/cvpr52688.2022.01553. URL http://dx.doi.org/10.1109/cvpr52688.2022.01553.

Tri Huynh, Simon Kornblith, Matthew R. Walter, Michael Maire, and Maryam Khademi. Boosting contrastive self-supervised learning with false negative cancellation. In *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 986–996. IEEE, 2022. doi: 10.1109/wacv51458.2022.00106. URL http://dx.doi.org/10.1109/WACV51458.2022.00106.

Naveed Ishaque, Peter Kharchenko, Daria Lazic, Jieran Sun, Jean Yee Hwa Yang, Martin Emons, Florian Heyl, Wolfgang Huber, Daniel Jones, Louis B Kuemmerle, et al. The challenge of cell segmentation in spatially resolved transcriptomics. *arXiv preprint arXiv:2606.09675*, 2026.

Amanda Janesick, Robert Shelansky, Andrew D. Gottscho, Florian Wagner, Stephen R. Williams, Morgane Rouault, Ghezal Beliakoff, Carolyn A. Morrison, Michelli F. Oliveira, Jordan T. Sicherman, Andrew Kohlway, Jawad Abousoud, Tingsheng Yu Drennon, Seayar H. Mohabbat, and Sarah E. B. Taylor. High resolution mapping of the tumor microenvironment using integrated single-cell, spatial and in situ analysis. *Nature Communications*, 14(1), 2023. ISSN 2041-1723. doi: 10.1038/s41467-023-43458-x. URL http://dx.doi.org/10.1038/s41467-023-43458-x.

Guillaume Jaume, Paul Doucet, Andrew Song, Ming Lu, Cristina Almagro-Pérez, Sophia Wagner, Anurag Vaidya, Richard Chen, Drew Williamson, Ahrong Kim, and Faisal Mahmood. Hest-1k: A dataset for spatial transcriptomics and histology image analysis. In *Advances in Neural Information Processing Systems 37*, NeurIPS 2024, pp. 53798–53833. Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2024. doi: 10.52202/079017-1704. URL http://dx.doi.org/10.52202/079017-1704.

Shiyu Jiang, Zhaoyu Fang, Yujie Zhang, Xuting Zhang, Jérémie Kalfon, Weixu Wang, Xi Fu, Aakash Patel, Syed Rizvi, Yuancheng Ryan Lu, Siyu He, Yixin Wang, Kejun Ying, Peter Pao-Huang, Yifan Lu, Weize Xu, Mengchen Wang, Ziyang Miao, Jianhui Lin, Jimmy Ding, Zitong Jerry Wang, Wei Ouyang, Tianlong Chen, Guoxian Yu, Min Li, Jiayi Ma, Fei Wang, Yuying Xie, Jiliang Tang, Raul Rabadan, David van Dijk, Pengtao Xie, Peng He, Emily B. Fox, Le Song, Fabian J. Theis, Eric Xing, Christina V. Theodoris, Xiaojie Qiu, and Jiayuan Ding. The landscape of single-cell foundation models: Design principles, applications, and open challenges. *Preprints*, August 2026. doi: 10.20944/preprints202608.1166.v1. URL https://doi.org/10.20944/preprints202608.1166.v1.

Yann LeCun et al. A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27. *Open Review*, 62(1):1–62, 2022.

Juho Lee, Yoonho Lee, Jungtaek Kim, Adam Kosiorek, Seungjin Choi, and Yee Whye Teh. Set transformer: A framework for attention-based permutation-invariant neural networks. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3744–3753. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/lee19d.html.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9992–10002. IEEE, 2021. doi: 10.1109/iccv48922.2021.00986. URL http://dx.doi.org/10.1109/ICCV48922.2021.00986.

Romain Lopez, Jeffrey Regier, Michael B. Cole, Michael I. Jordan, and Nir Yosef. Deep generative modeling for single-cell transcriptomics. *Nature Methods*, 15(12):1053–1058, 2018. ISSN 1548-7105. doi: 10.1038/s41592-018-0229-2. URL http://dx.doi.org/10.1038/s41592-018-0229-2.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Clarence K. Mah, Noorsher Ahmed, Nicole A. Lopez, Dylan C. Lam, Avery Pong, Alexander Monell, Colin Kern, Yuanyuan Han, Gino Prasad, Anthony J. Cesnik, Emma Lundberg, Quan Zhu, Hannah Carter, and Gene W. Yeo. Bento: a toolkit for subcellular analysis of spatial transcriptomics data. *Genome Biology*, 25(1), 2024. ISSN 1474-760X. doi: 10.1186/s13059-024-03217-7. URL http://dx.doi.org/10.1186/s13059-024-03217-7.

Jeffrey R. Moffitt, Emma Lundberg, and Holger Heyn. The emerging landscape of spatial profiling technologies. *Nature Reviews Genetics*, 23(12):741–759, 2022. ISSN 1471-0064. doi: 10.1038/s41576-022-00515-3. URL http://dx.doi.org/10.1038/s41576-022-00515-3.

Giovanni Palla, David S. Fischer, Aviv Regev, and Fabian J. Theis. Spatial components of molecular tissue biology. *Nature Biotechnology*, 40(3):308–318, 2022. ISSN 1546-1696. doi: 10.1038/s41587-021-01182-1. URL http://dx.doi.org/10.1038/s41587-021-01182-1.

Jasmine T. Plummer, Felipe Segato Dezem, David P. Cook, Jiwoon Park, Luke Zhang, Yutian Liu, Maycon Marção, Hannah DuBose, Arjumand Wani, Kellie Wise, Michael Roach, Kate Harvey, Taopeng Wang, Kirk B. Jensen, Natalia Morosini, Roberto De Gregorio, Alicia Alonso, Shauna Lee Houlihan, Robert E. Schwartz, Erika Hissong, Catherine Snopkowski, Jeffrey L. Wrana, Natalie Ryan, Lisa M. Butler, George Church, Alexander Swarbrick, Christopher E. Mason,

and Luciano G. Martelotto. Standardized metrics for assessment and reproducibility of imaging-based spatial transcriptomics datasets. *Nature Biotechnology*, 44(7):1213–1225, 2025. ISSN 1546-1696. doi: 10.1038/s41587-025-02811-9. URL http://dx.doi.org/10.1038/s41587-025-02811-9.

Ladislav Rampášek, Mikhail Galkin, Vijay Prakash Dwivedi, Anh Tuan Luu, Guy Wolf, and Dominique Beaini. Recipe for a general, powerful, scalable graph transformer, 2022. URL https://arxiv.org/abs/2205.12454.

Pengfei Ren, Rui Zhang, Yunfeng Wang, Peng Zhang, Ce Luo, Suyan Wang, Xiaohong Li, Zongxu Zhang, Yanping Zhao, Yufeng He, Haorui Zhang, Yufeng Li, Zhidong Gao, Xiuping Zhang, Yahui Zhao, Zhihua Liu, Yuanguang Meng, Zhe Zhang, and Zexian Zeng. Systematic benchmarking of high-throughput subcellular spatial transcriptomics platforms across human tumors. *Nature Communications*, 16(1), 2025. ISSN 2041-1723. doi: 10.1038/s41467-025-64292-3. URL http://dx.doi.org/10.1038/s41467-025-64292-3.

Yanay Rosen, Yusuf Roohani, Ayush Agrawal, Leon Samotorčan, Stephen R. Quake, and Jure Leskovec. Universal cell embedding provides a foundation model for cell biology. *Nature*, 656 (8126):183–191, 2026. ISSN 1476-4687. doi: 10.1038/s41586-026-10689-z. URL http://dx.doi.org/10.1038/s41586-026-10689-z.

Christian M Schürch, Salil S Bhate, Graham L Barlow, Darci J Phillips, Luca Noti, Inti Zlobec, Pauline Chu, Sarah Black, Janos Demeter, David R McIlwain, et al. Coordinated cellular neighborhoods orchestrate antitumoral immunity at the colorectal cancer invasive front. *Cell*, 182(5):1341–1359, 2020.

Ruth L Seal, Bryony Braschi, Kristian Gray, Tamsin E M Jones, Susan Tweedie, Liora Haim-Vilmovsky, and Elspeth A Bruford. Genenames.org: the hgnc resources in 2023. *Nucleic Acids Research*, 51(D1):D1003–D1009, 2022. ISSN 1362-4962. doi: 10.1093/nar/gkac888. URL http://dx.doi.org/10.1093/nar/gkac888.

Vipul Singhal, Nigel Chou, Joseph Lee, Yifei Yue, Jinyue Liu, Wan Kee Chock, Li Lin, Yun-Ching Chang, Erica Mei Ling Teo, Jonathan Aow, Hwee Kuan Lee, Kok Hao Chen, and Shyam Prabhakar. Banksy unifies cell typing and tissue domain segmentation for scalable spatial omics data analysis. *Nature Genetics*, 56(3):431–441, 2024. ISSN 1546-1718. doi: 10.1038/s41588-024-01664-3. URL http://dx.doi.org/10.1038/s41588-024-01664-3.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024. ISSN 0925-2312. doi: 10.1016/j.neucom.2023.127063. URL http://dx.doi.org/10.1016/j.neucom.2023.127063.

Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/68053af2923e00204c3ca7c6a3150cf7-Paper.pdf.

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

Theis Lab. Cameo-breast, 2026. URL https://huggingface.co/datasets/theislab/CAMEO-Breast.

Christina V. Theodoris, Ling Xiao, Anant Chopra, Mark D. Chaffin, Zeina R. Al Sayed, Matthew C. Hill, Helene Mantineo, Elizabeth M. Brydon, Zexian Zeng, X. Shirley Liu, and Patrick T. Ellinor. Transfer learning enables predictions in network biology. *Nature*, 618(7965):616–624, 2023. ISSN 1476-4687. doi: 10.1038/s41586-023-06139-9. URL http://dx.doi.org/10.1038/s41586-023-06139-9.

Annika Vannan, Ruqian Lyu, Arianna L. Williams, Nicholas M. Negretti, Evan D. Mee, Joseph Hirsh, Samuel Hirsh, Niran Hadad, David S. Nichols, Carla L. Calvi, Chase J. Taylor, Vasiliy. V. Polosukhin, Ana P. M. Serezani, A. Scott McCall, Jason J. Gokey, Heejung Shim, Lorraine B. Ware, Matthew J. Bacchetta, Ciara M. Shaver, Timothy S. Blackwell, Rajat Walia, Jennifer M. S. Sucre, Jonathan A. Kropski, Davis J. McCarthy, and Nicholas E. Banovich. Spatial transcriptomics identifies molecular niche dysregulation associated with distal lung remodeling in pulmonary fibrosis. *Nature Genetics*, 57(3):647–658, 2025. ISSN 1546-1718. doi: 10.1038/s41588-025-02080-x. URL http://dx.doi.org/10.1038/s41588-025-02080-x.

Marco Varrone, Daniele Tavernari, Albert Santamaria-Martínez, Logan A. Walsh, and Giovanni Ciriello. Cellcharter reveals spatial cell niches associated with tissue remodeling and cell plasticity. *Nature Genetics*, 56(1):74–84, 2023. ISSN 1546-1718. doi: 10.1038/s41588-023-01588-4. URL http://dx.doi.org/10.1038/s41588-023-01588-4.

Chloe Wang, Haotian Cui, Andrew Zhang, Ronald Xie, Hani Goodarzi, and Bo Wang. scgpt-spatial: continual pretraining of single-cell foundation model for spatial transcriptomics. *biorxiv*, pp. 2025–02, 2025.

Johann Wenckstern, Eeshaan Jain, Benedikt von Querfurth, Yexiang Cheng, Kiril Vasilev, Matteo Pariset, Phil F. Cheng, Petros Liakopoulos, Olivier Michielin, Andreas Wicki, Gabriele Gut, and Charlotte Bunne. The virtual tissues foundation model resolves spatial proteomics across scales. *Nature*, 657(8132):744–754, 2026. ISSN 1476-4687. doi: 10.1038/s41586-026-10884-y. URL http://dx.doi.org/10.1038/s41586-026-10884-y.

Andrew T.A Wood. Simulation of the von mises fisher distribution. *Communications in Statistics - Simulation and Computation*, 23(1):157–164, 1994. ISSN 1532-4141. doi: 10.1080/03610919408813161. URL http://dx.doi.org/10.1080/03610919408813161.

Xiaoyang Wu, Li Jiang, Peng-Shuai Wang, Zhijian Liu, Xihui Liu, Yu Qiao, Wanli Ouyang, Tong He, and Hengshuang Zhao. Point transformer v3: Simpler, faster, stronger. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4840–4851. IEEE, 2024. doi: 10.1109/cvpr52733.2024.00463. URL http://dx.doi.org/10.1109/cvpr52733.2024.00463.

Chenglong Xia, Jean Fan, George Emanuel, Junjie Hao, and Xiaowei Zhuang. Spatial transcriptome profiling by merfish reveals subcellular rna compartmentalization and cell cycle-dependent gene expression. *Proceedings of the National Academy of Sciences*, 116(39):19490–19499, 2019. ISSN 1091-6490. doi: 10.1073/pnas.1912459116. URL http://dx.doi.org/10.1073/pnas.1912459116.

Hanwen Xu, Naoto Usuyama, Jaspreet Bagga, Sheng Zhang, Rajesh Rao, Tristan Naumann, Cliff Wong, Zelalem Gero, Javier González, Yu Gu, et al. A whole-slide foundation model for digital pathology from real-world data. *Nature*, 630(8015):181–188, 2024.

Meng Zhang, Xingjie Pan, Won Jung, Aaron R. Halpern, Stephen W. Eichhorn, Zhiyun Lei, Limor Cohen, Kimberly A. Smith, Bosiljka Tasic, Zizhen Yao, Hongkui Zeng, and Xiaowei Zhuang. Molecularly defined and spatially resolved cell atlas of the whole mouse brain. *Nature*, 624(7991):343–354, 2023. ISSN 1476-4687. doi: 10.1038/s41586-023-06808-9. URL http://dx.doi.org/10.1038/s41586-023-06808-9.

Suyuan Zhao, Yizhen Luo, Ganbo Yang, Yan Zhong, Hao Zhou, and Zaiqing Nie. Stofm: a multi-scale foundation model for spatial transcriptomics. *arXiv preprint arXiv:2507.11588*, 2025.

Maksim Zhdanov, Max Welling, and Jan-Willem van de Meent. Erwin: A tree-based hierarchical transformer for large-scale physical systems. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=MrphqqwnKv.

Xiaowei Zhuang. Spatially resolved single-cell genomics and transcriptomics by imaging. *Nature methods*, 18(1):18–22, 2021.

Appendix

CONTENTS

Appendix A **Related Work** 16
Appendix B **Tissue Embedding Fields Across Spatial Scales** 17
Appendix C **Extended Methods** 20
Appendix D **Public Pretraining Corpus** 27
Appendix E **Baseline Model Configuration** 30
Appendix F **Pretraining** 30
Appendix G **Transfer Evaluation** 33
Appendix H **Ablations** 39

A RELATED WORK

Large-scale tissue representation learning. Large-scale self-supervised histopathology has shown that tissue can yield transferable representations from local morphology to whole-slide context (Chen et al., 2024; Xu et al., 2024). iST provides a finer molecular view, but raises a different question of what should remain individually represented after encoding. Existing spatial foundation models fix this unit in advance. Nicheformer represents segmented cells (Tejada-Lapuerta et al., 2025), Novae uses cell- or spot-level graphs (Blampey et al., 2025), TERRA encodes cells within predefined neighborhoods (Birk et al., 2026), and SToFM uses multi-scale sub-slices (Zhao et al., 2025). These choices aggregate transcripts and remove within-unit molecular geometry, while cell-based models also inherit segmentation uncertainty (Ishaque et al., 2026). Laminar instead holds transcripts as the canonical support of a transferable representation, leaving cells, niches, and larger regions to be defined when the field is queried.

Multi-scale spatial encoders. Hierarchical representations have demonstrated effectiveness across vision, histopathology, and point-cloud learning, particularly when structural information spans multiple spatial scales. Swin constructs multiscale visual features using shifted-window attention (Liu et al., 2021). HIPT (Chen et al., 2022) and GigaPath (Xu et al., 2024) link local morphology to broader tissue context. Point Transformer V3 (Wu et al., 2024) and Erwin (Zhdanov et al., 2025) extend scalable hierarchical modeling to irregular point sets. GraphGPS integrates local graph propagation with global attention (Rampášek et al., 2022). Laminar explicitly defines these scales in physical space by employing a local transcript graph and a quadtree over progressively larger tissue regions, subsequently returning the aggregated context to each transcript.

Self-supervised representation learning. The scale of transcript-resolved iST is growing much faster than expert annotation of cellular states, niches, and tissue phenotypes, making self-supervised pretraining a natural way to exploit these data. Contrastive objectives require defining which views should be matched and which examples should be separated, a difficult choice when biological states vary continuously and related samples may form false negatives (Chen et al., 2020; Huynh et al., 2022). Masked reconstruction avoids explicit negatives but places the learning target on the withheld observations themselves (He et al., 2022). Joint-embedding predictive learning instead predicts latent targets from visible context (LeCun et al., 2022; Assran et al., 2023). Laminar applies this principle to regional representations, predicting complete-field targets across physical scales from spatially and gene-masked tissue using the same readout exposed downstream.

Embedding fields. AlphaEarth Foundations introduced *embedding fields* as reusable spatial representations that can be precomputed once and queried across many downstream mapping tasks (Brown et al., 2025). Laminar brings this idea to iST, where the field is supported directly on measured transcript coordinates rather than a fixed raster. Regions are represented by summing their transcript embeddings and normalizing once at readout. These states compose exactly before normalization, so the same encoded tissue can support different boundaries and spatial scales without re-encoding.

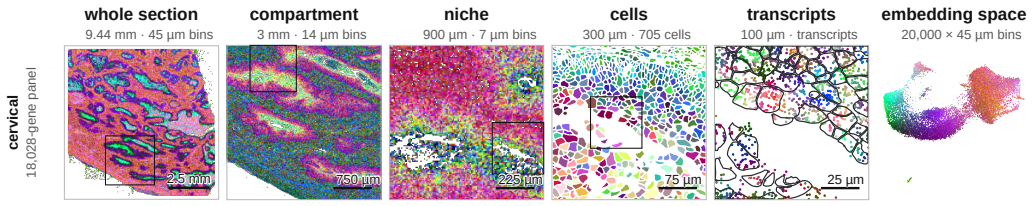

Figure 4: **One field queried across spatial supports.** The same encoded cervical carcinoma section is read from 45 μm regions to individual transcripts.

B TISSUE EMBEDDING FIELDS ACROSS SPATIAL SCALES

We next examine the tissue embedding field directly. These analyses show how one cached field can be read at different spatial supports, whether its geometry is stable across training seeds, and how it relates to expert annotations where available. They are qualitative analyses and are not included in the benchmark.

Laminar encodes every available transcript in a section or core once. At a chosen support, we form each regional embedding by summing the cached transcript vectors in that region and normalizing once. PCA is fit to the regional embeddings at that support; the three leading components with the highest Moran's I are mapped to RGB. Within a support, the same basis and contrast are reused across nested windows. For figures containing multiple tissues, the basis is fit jointly on the stated collection and reused across tissues. PCA is refit when the support changes unless the caption states that two embedding spaces were aligned. Subsampling is used only for dense scatter plots and embedding-space summaries; the field itself is always computed from all available transcripts. Cell boundaries, where shown, are used only after encoding to define a readout or provide spatial reference. RGB values indicate directions in embedding space, not biological classes.

B.1 QUERYING ONE ENCODED FIELD AT DIFFERENT SPATIAL SUPPORTS

We first use a whole-transcriptome FFPE human cervical carcinoma section to show how one encoded field can be read at different spatial supports. Laminar encodes all measured transcripts once. The nested windows in Figure 4 change only the displayed region and the query support.

B.2 STABILITY OF FIELD GEOMETRY ACROSS TRAINING SEEDS

We next encode the same complete cervical carcinoma section with two Laminar models trained with the same architecture and training setup and different random seeds. Their embedding coordinates need not be axis-aligned. We therefore align matched regional units separately at each support with canonical correlation analysis before comparing the spatial patterns.

B.3 COMPARISON WITH EXPERT CELL-TYPE ANNOTATIONS

The same cervical carcinoma section also has expert cell-type annotations generated for this study. We use these annotations only as an external reference for the learned field. Laminar first encodes the complete transcript measurements. Cell-level embeddings are then recomputed from the cached transcript vectors assigned to each cell, while the transcript view uses the stored transcript embeddings directly.

B.4 SHARED FIELD STRUCTURE ACROSS WHOLE TISSUE SECTIONS

We then examine seven anonymized human meningioma sections from an in-house iST cohort. Each row in Figure 7 is a different tissue section. The same Laminar checkpoint is used throughout, and every available transcript in each section is encoded before any regional readout is formed.

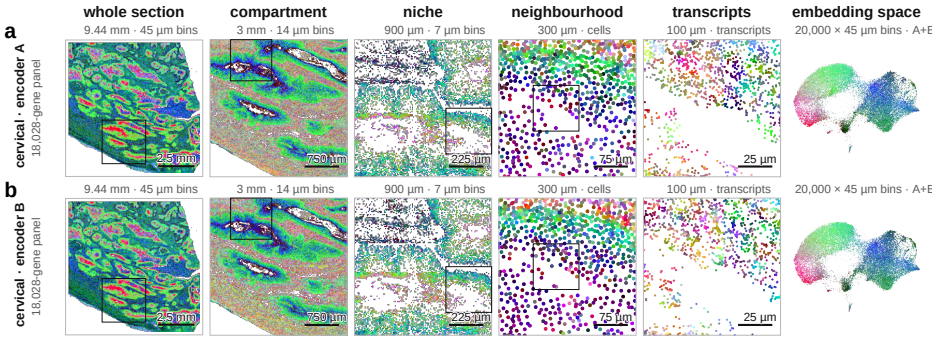

Figure 5: **Field geometry across training seeds.** The same section is encoded by two independently trained Laminar models and aligned at each support with canonical correlation analysis.

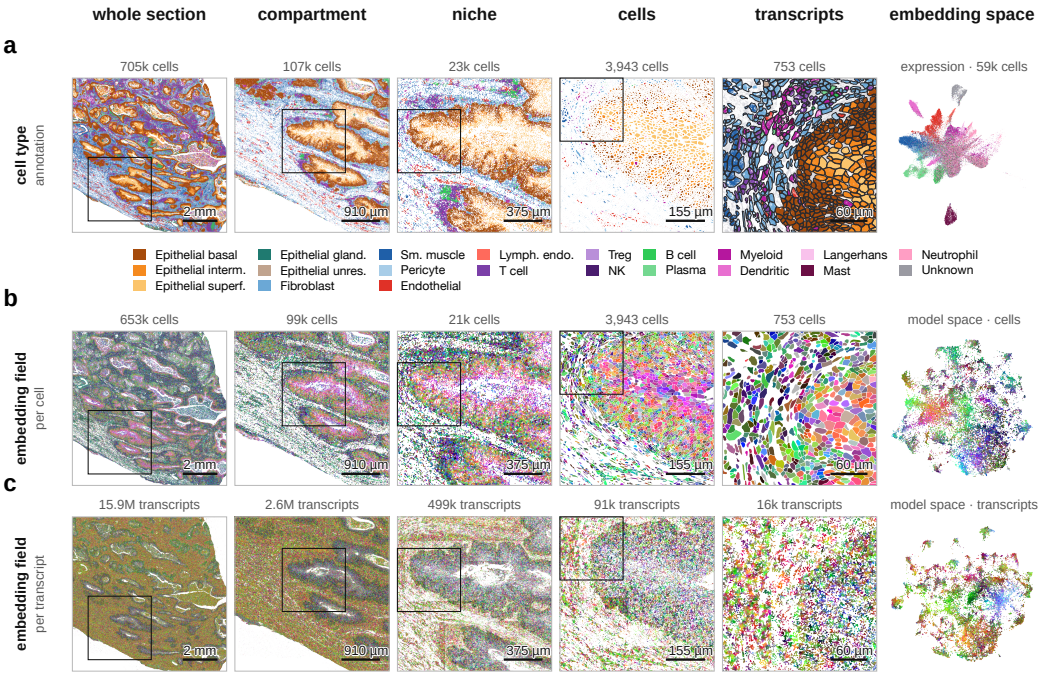

Figure 6: **Field readout and expert cell-type annotations.** Annotations are shown only as an external reference after encoding.

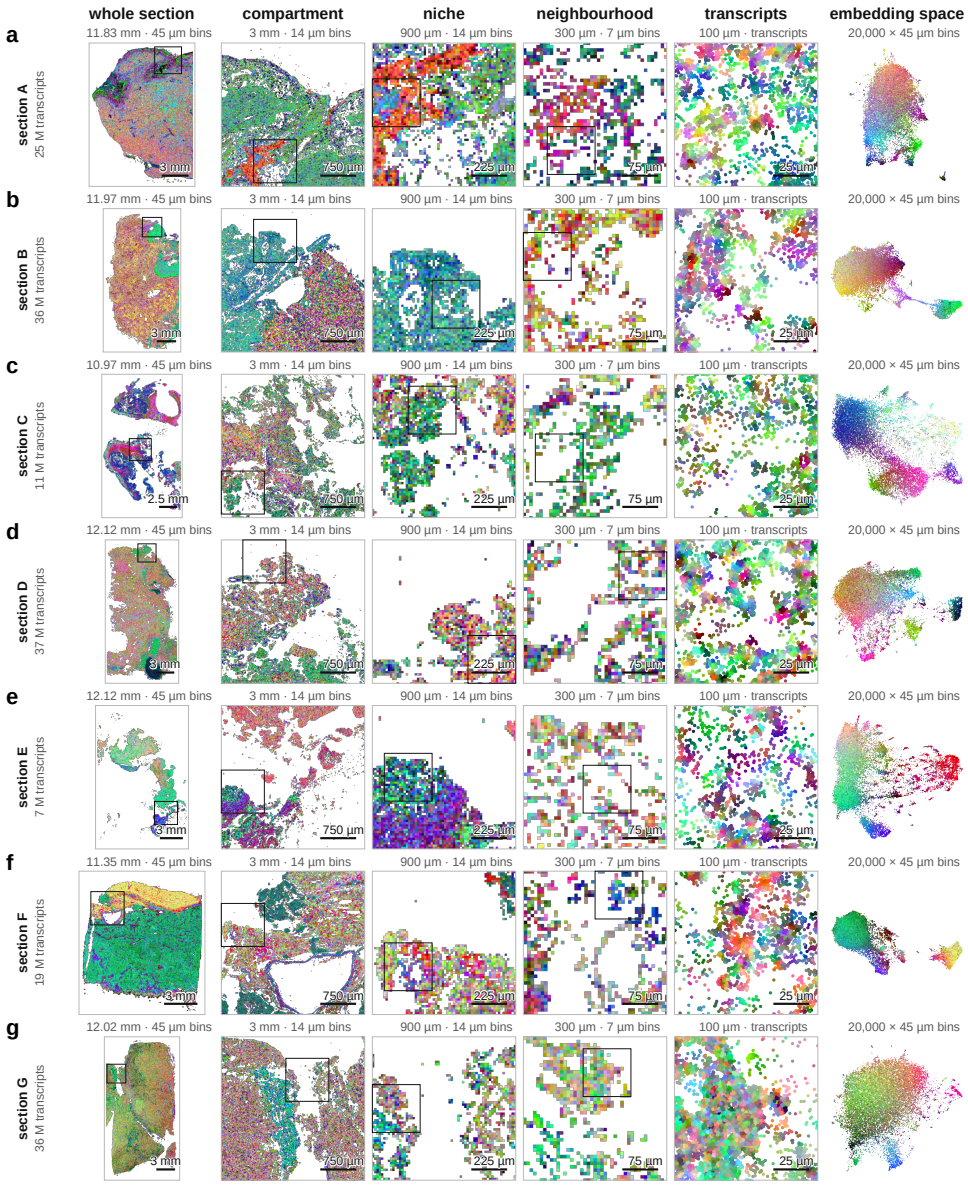

Figure 7: **Shared field structure across tissue sections.** Seven anonymized meningioma sections are encoded with the same Laminar checkpoint and a support-specific visualization basis shared across sections.

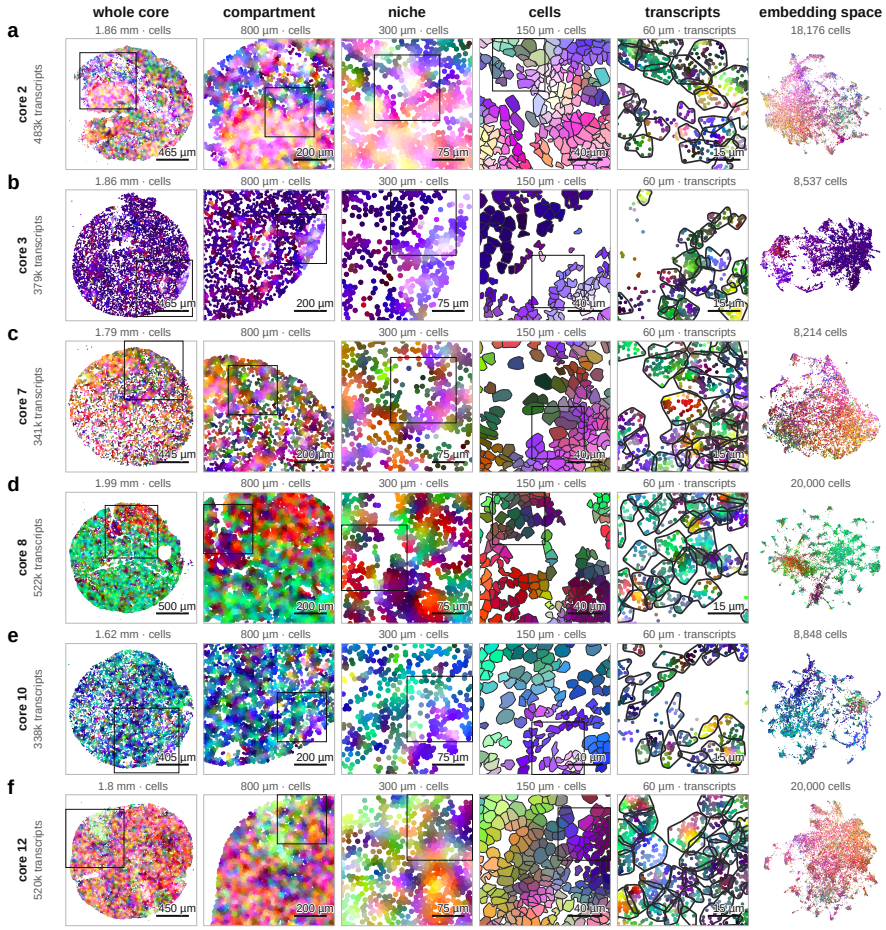

Figure 8: **Shared field structure across tissue-microarray cores.** Six anonymized meningioma cores are encoded with the same checkpoint and shown from core to transcript scale.

B.5 Shared field across tissue-microarray cores

Finally, we examine an in-house meningioma tissue microarray. A tissue microarray contains small cores sampled from larger tumors and arranged on common slides so that multiple specimens can be profiled in parallel. All available transcripts in each core are encoded with the same Laminar checkpoint. Twelve cores are used to fit the shared visualization basis and six are shown in Figure 8. The reduction to six rows and the subsampling in the embedding-space summary are only for visualization.

For the first four spatial views, each cell is drawn at its measured location and colored from a local field readout around that cell. We sum the cached transcript embeddings within a 30 μm FWHM neighborhood and normalize once. The first four columns share one PCA basis fitted on within-core-centered embeddings from all twelve cores. The transcript view uses a separate basis because the readout support changes.

C Extended methods

In this section, we specify the encoder and pretraining procedure used in the experiments in more detail. Table 3 collects the main notation and dimensions. Baseline numerical settings are given in Tables 7 and 8, with stage-dependent sampling values in Appendix F.

C.1 Input and spatial notation

A tissue section contains $N \in \mathbb{N}$ detected transcripts,

$$\mathcal{X} = \{(p_i, g_i)\}_{i=1}^{N}, \qquad p_i = (x_i^{\mathrm{pos}}, y_i^{\mathrm{pos}}) \in \mathbb{R}^2, \qquad g_i \in \{1, \dots, G\}. \tag{9}$$

Here p_i is the physical position of transcript i in micrometres, g_i is its categorical gene index, and G is the vocabulary size. The encoder receives no cell identifier, segmentation, image, or expression count associated with a transcript.

The encoder operates on a square crop $C \subset \mathbb{R}^2$ of physical width W (in μm). Let $\mathcal{I}_C = \{i : p_i \in C\}$ and $N_C = |\mathcal{I}_C|$. A regular quadtree contains $L \in \mathbb{N}$ levels with tile widths

$$s_\ell = 2^\ell s_0, \qquad \ell \in \{0, \dots, L-1\}, \qquad W = 2^{L-1} s_0, \tag{10}$$

where s_0 is the leaf-tile width. Only occupied tiles are materialized. We denote the occupied tiles at level ℓ by $\mathcal{T}_\ell$; $a_i^{(\ell)} \in \mathcal{T}_\ell$ is the tile containing transcript i; $c_a \in \mathbb{R}^2$ is the centre of tile a; $\mathrm{ch}(a)$ is the set of its occupied children; and $n_a \in \mathbb{N}$ is its transcript count. For any physical point $q \in \mathbb{R}^2$, the dimensionless offset relative to a level-ℓ tile is

$$\delta_a(q) = \frac{q - c_a}{s_\ell} \in \mathbb{R}^2, \qquad a \in \mathcal{T}_\ell. \tag{11}$$

Table 3: **Core notation and baseline values.** Shapes are given for one transcript, tile, or region unless otherwise stated. A dash means that the quantity depends on the input section rather than the model configuration.

Symbol	Meaning	Domain / shape	Baseline value
N	transcripts in a tissue section	scalar, $\mathbb{N}$	–
p_i	physical coordinate of transcript i	$\mathbb{R}^2$, μm	–
g_i	gene index of transcript i	$\{1, \dots, G\}$	–
G	rows in the gene embedding table	scalar, $\mathbb{N}$	45 104
W	crop width	scalar, μm	1 024
s_ℓ	tile width at hierarchy level ℓ	scalar, μm	$2^\ell \times 16$
L	number of hierarchy levels	scalar, $\mathbb{N}$	7
M_{gene}	learned gene embedding table	$\mathbb{R}^{G \times d_g}$	$d_g = 128$
x_i	gene embedding of transcript i	$\mathbb{R}^{d_g}$	128
h_i	local graph state	$\mathbb{R}^{d_h}$	1 024
u_i	transcript feature passed to the hierarchy	$\mathbb{R}^{d_{\mathrm{tile}}}$	1 024
K	summary tokens per occupied tile	scalar, $\mathbb{N}$	4
$H_a^{(\ell)}$	state of occupied tile a at level ℓ	$\mathbb{R}^{K \times d_{\mathrm{tile}}}$	$4 \times 1\,024$
$\mathcal{J}_{\mathrm{out}}$	hierarchy levels used by transcript readout	subset of $\{0, \dots, L-1\}$	$\{0, 1, 2\}$
z_i	normalized transcript embedding	$\mathbb{S}^{d-1}$	$d = 1\,024$
$\Phi(A)$	normalized embedding of region A	$\mathbb{S}^{d-1}$	$d = 1\,024$
d_p	predictor token width	scalar, $\mathbb{N}$	384

C.2 Tissue embedding field and regional readout

Fix a section $\mathcal{X}$ and the unit transcript embeddings z_i produced by E_θ. The tissue embedding field is queried through the regional quantities defined below. We suppress the subscript $\mathcal{X}$ for readability. For a measurable region $A \subseteq \Omega_{\mathcal{X}}$, let $\mathcal{I}(A) = \{i : p_i \in A\}$ and $n(A) = |\mathcal{I}(A)|$. The regional state is $(S(A), n(A))$; its embedding sum, regional embedding, and directional coherence for $n(A) > 0$ are

$$S(A) = \sum_{i \in \mathcal{I}(A)} z_i, \quad \Phi(A) = \frac{S(A)}{\|S(A)\|_2} \in \mathbb{S}^{d-1}, S(A) \neq 0, \quad R(A) = \frac{\|S(A)\|_2}{n(A)} \in [0, 1]. \tag{12}$$

If $n(A) = 0$, then $S(A) = 0$ and neither $\Phi(A)$ nor $R(A)$ is defined. Cancellation can also give $S(A) = 0$ when $n(A) > 0$; then $R(A) = 0$, but $\Phi(A)$ remains undefined.

For $n(A) \geq 2$, define the *count-corrected coherence*

$$\bar{c}(A) = \frac{n(A)R(A)^2 - 1}{n(A) - 1} = \frac{1}{n(A)(n(A)-1)} \sum_{\substack{i,j \in \mathcal{I}(A) \\ i \neq j}} z_i^\top z_j. \tag{13}$$

This identity follows by expanding $\|S(A)\|_2^2$ and removing the $n(A)$ diagonal terms $z_i^\top z_i = 1$. Thus $\bar{c}(A)$ gives the mean cosine similarity between distinct transcript pairs without explicitly computing all pairwise similarities. For $n(A) = 1$, $R(A) = 1$, but $\bar{c}(A)$ is undefined.

For disjoint regions $A, B \subseteq \Omega_{\mathcal{X}}$,

$$S(A \cup B) = S(A) + S(B), \qquad n(A \cup B) = n(A) + n(B).$$

All embeddings in these expressions come from the same fixed field. We obtain $\Phi(A \cup B)$ by normalizing $S(A) + S(B)$ when this sum is nonzero. Normalized regional embeddings alone generally do not suffice for this aggregation. Once the transcript embeddings have been computed, regional queries use these embeddings without running the encoder again.

C.3 Molecular graph and local encoder

Molecular graph. For transcript i, define the radius-capped candidate set

$$\mathcal{B}_i(r_0) = \{j \neq i : \|p_j - p_i\|_2 \leq r_0\}, \qquad \mathcal{N}_i = \mathrm{kNN}_{k_{\mathrm{nn}}}(i \mid \mathcal{B}_i(r_0)), \tag{14}$$

where $r_0 > 0$ is the maximum physical edge length in μm and $k_{\mathrm{nn}} \in \mathbb{N}$ is the maximum number of neighbors per transcript. The operator $\mathrm{kNN}_{k_{\mathrm{nn}}}(i \mid \mathcal{B}_i)$ returns the $\min(k_{\mathrm{nn}}, |\mathcal{B}_i|)$ transcripts in $\mathcal{B}_i$ closest to i in Euclidean distance, with a stable transcript-index tie break. A self-loop is added to every transcript. The implementation uses a spatial cell list to retrieve within-radius candidates and then selects the nearest neighbors without materializing a dense pairwise-distance matrix.

Gene embedding and local layers. The trainable gene lookup is

$$M_{\mathrm{gene}} \in \mathbb{R}^{G \times d_g}, \qquad x_i = M_{\mathrm{gene}}[g_i] \in \mathbb{R}^{d_g}, \tag{15}$$

where d_g is the gene-embedding width. A learned affine map $W_{\mathrm{in}} \in \mathbb{R}^{d_h \times d_g}$ produces $h_i^{(0)} = W_{\mathrm{in}} x_i + b_{\mathrm{in}} \in \mathbb{R}^{d_h}$. The local encoder has L_{loc} residual GATv2 layers with H_{loc} heads (Brody et al., 2021). Let $d_{\mathrm{head}}^{\mathrm{loc}} = d_h / H_{\mathrm{loc}}$. At layer t, write $\hat{h}_i = \mathrm{LN}(h_i^{(t-1)}) \in \mathbb{R}^{d_h}$. For head h, $W_{R,h}, W_{L,h} \in \mathbb{R}^{d_{\mathrm{head}}^{\mathrm{loc}} \times d_h}$ and $a_h, b_h \in \mathbb{R}^{d_{\mathrm{head}}^{\mathrm{loc}}}$, and

$$\begin{aligned} e_{ij}^{(h)} &= a_h^\top \mathrm{LeakyReLU}_{0.2}\Big(W_{R,h}\hat{h}_i + W_{L,h}\hat{h}_j\Big) \in \mathbb{R}, \\ \alpha_{ij}^{(h)} &= \frac{\exp e_{ij}^{(h)}}{\sum_{j' \in \mathcal{N}_i \cup \{i\}} \exp e_{ij'}^{(h)}} \in [0,1], \\ m_{i,h} &= \sum_{j \in \mathcal{N}_i \cup \{i\}} \alpha_{ij}^{(h)} W_{L,h}\hat{h}_j + b_h \in \mathbb{R}^{d_{\mathrm{head}}^{\mathrm{loc}}}. \end{aligned} \tag{16}$$

Concatenating the heads gives $m_i = \mathrm{cat}_{h=1}^{H_{\mathrm{loc}}} m_{i,h} \in \mathbb{R}^{d_h}$ and

$$h_i^{(t)} = h_i^{(t-1)} + \mathrm{GELU}(m_i), \qquad t = 1, \ldots, L_{\mathrm{loc}}. \tag{17}$$

Coordinates determine the graph but are not supplied as features to the GATv2 updates. Let $h_i = h_i^{(L_{\mathrm{loc}})} \in \mathbb{R}^{d_h}$ denote the final graph-contextualized state. We combine this state with a direct projection of the original gene embedding to form the spatial Transformer input:

$$u_i = \big[\, W_h^{\mathrm{loc}} h_i \,\big\|\, W_x^{\mathrm{loc}} x_i \,\big] \in \mathbb{R}^{d_{\mathrm{tile}}}, \tag{18}$$

where $W_h^{\mathrm{loc}} \in \mathbb{R}^{d_{u,h} \times d_h}$ and $W_x^{\mathrm{loc}} \in \mathbb{R}^{d_{u,x} \times d_g}$ are learned projections, $\|$ denotes concatenation, and $d_{u,h} + d_{u,x} = d_{\mathrm{tile}}$. Before the spatial hierarchy, both h_i and u_i have physical receptive radius at most $L_{\mathrm{loc}} r_0$.

C.4 ATTENTION BLOCKS AND RELATIVE POSITION

Attention block. For a block of width D, let $x \in \mathbb{R}^{n_q \times D}$ be the query stream and $y \in \mathbb{R}^{n_k \times D}$ the key/value stream. The hierarchy uses $D = d_{\text{tile}}$; the predictor uses its own width $D = d_p$. Every block is pre-norm and residual:

$$x \leftarrow x + \gamma_1 \odot W_O \,\mathrm{Attn}(\mathrm{LN}_q(x), \mathrm{LN}_{kv}(y)), \qquad x \leftarrow x + \gamma_2 \odot W_2 \,\mathrm{GELU}(W_1 \mathrm{LN}(x)), \quad (19)$$

where $\gamma_1, \gamma_2 \in \mathbb{R}^D$ are LayerScale vectors, $W_O \in \mathbb{R}^{D \times D}$, $W_1 \in \mathbb{R}^{4D \times D}$, and $W_2 \in \mathbb{R}^{D \times 4D}$. Self-attention uses $y = x$; cross-attention uses separate query and key/value LayerNorms.

For H attention heads, $d_{\text{head}} = D/H$. Within a head, $q, k, g_q, g_k \in \mathbb{R}^{d_{\text{head}}}$ and the normalized query-key logit is

$$\lambda(q,k) = \frac{(g_q \odot \hat{q})^\top (g_k \odot \hat{k})}{\sqrt{d_{\text{head}}}}, \qquad \hat{q} = \frac{q}{\|q\|_2}, \qquad \hat{k} = \frac{k}{\|k\|_2}. \quad (20)$$

The learnable gains g_q, g_k are initialized at $\sqrt{d_{\text{head}}}$ (Dehghani et al., 2023). LayerScale starts at 0.1 in neighborhood-attention blocks and at 1 in pooling and predictor blocks. The output projection and feed-forward output of the nth block in a stack are scaled at initialization by $1/\sqrt{2n}$ (Bao et al., 2021).

Two-dimensional rotary position. We use a two-dimensional rotary position parameterization (Su et al., 2024). For a dimensionless offset $\delta = (\delta_x, \delta_y) \in \mathbb{R}^2$, each head allocates $q_{\text{rope}} = d_{\text{head}}/4$ coordinate pairs to each spatial axis. Pair $r \in \{0, \ldots, q_{\text{rope}} - 1\}$ uses

$$\varphi_{r,x} = 2\pi f_r \delta_x, \qquad \varphi_{r,y} = 2\pi f_r \delta_y, \qquad f_r = f_{\min} \left(\frac{f_{\max}}{f_{\min}}\right)^{r/(q_{\text{rope}}-1)}, \quad (21)$$

where f_r is measured in cycles per tile width. The same frequency band is reused at all hierarchy levels; because δ is normalized by s_ℓ, a fixed rotation corresponds to a physical displacement that scales with the tile width.

C.5 SPATIAL HIERARCHY

Pooling. Each occupied tile $a \in \mathcal{T}_\ell$ is represented by K tokens $H_a^{(\ell)} \in \mathbb{R}^{K \times d_{\text{tile}}}$. The child sequence is

$$X_a^{(0)} = \mathrm{cat}_{i:p_i \in a}\, u_i \in \mathbb{R}^{n_a \times d_{\text{tile}}}, \qquad X_a^{(\ell)} = \mathrm{cat}_{b \in \mathrm{ch}(a)}\, H_b^{(\ell-1)} \in \mathbb{R}^{K|\mathrm{ch}(a)| \times d_{\text{tile}}} \quad (\ell \geq 1). \quad (22)$$

Let $Q_0 \in \mathbb{R}^{K \times d_{\text{tile}}}$ be K learned pooling seeds, shared across tiles and levels, as in Pooling by Multihead Attention (Lee et al., 2019). With P_ℓ pooling blocks at level ℓ,

$$\widetilde{H}_a^{(\ell)} = \mathcal{P}_{P_\ell} \circ \cdots \circ \mathcal{P}_1\left(Q_0; X_a^{(\ell)}\right) \in \mathbb{R}^{K \times d_{\text{tile}}}. \quad (23)$$

The seed positions are the tile centre. Transcript positions at level 0 use $\delta_a(p_i)$; child-tile tokens use $\delta_a(c_b)$. A scalar count is embedded by a learned two-layer map $\mathrm{Emb}_{\text{count}} : \mathbb{R} \to \mathbb{R}^{d_{\text{tile}}}$, and $e_\ell \in \mathbb{R}^{d_{\text{tile}}}$ is a learned level embedding:

$$H_{a,0}^{(\ell)} = \widetilde{H}_a^{(\ell)} + \mathbf{1}_K \left[\mathrm{Emb}_{\text{count}}(\log(1 + n_a)) + e_\ell\right]^\top \in \mathbb{R}^{K \times d_{\text{tile}}}. \quad (24)$$

Parent counts are exact sums of child counts.

Neighborhood attention. We use shared Neighborhood Attention blocks (Hassani et al., 2023). The same-level spatial neighborhood of tile a is

$$\mathcal{M}_\ell(a) = \{b \in \mathcal{T}_\ell : \|c_b - c_a\|_\infty \leq s_\ell\}. \quad (25)$$

For D_{nat} shared neighborhood-attention blocks,

$$H_{a,t}^{(\ell)} = \mathcal{A}_t\left(H_{a,t-1}^{(\ell)}; \mathrm{cat}_{b \in \mathcal{M}_\ell(a)}\, H_{b,t-1}^{(\ell)}\right), \qquad t = 1, \ldots, D_{\text{nat}}, \quad (26)$$

where the K tokens of a are queries and the tokens in $\mathcal{M}_\ell(a)$ are keys and values. The final tile state is $H_a^{(\ell)} = H_{a,D_{\text{nat}}}^{(\ell)}$ and, for $\ell < L - 1$, becomes an input to the parent tile. Pooling and neighborhood-attention weights are shared across physical scales; e_ℓ identifies the level.

C.6 TOP-DOWN TRANSCRIPT READOUT

Let $\mathcal{J}_{\text{out}} \subseteq \{0, \dots, L-1\}$ be the hierarchy levels read by each transcript. At level $\ell \in \mathcal{J}_{\text{out}}$, write $a = a_i^{(\ell)}$ and $\delta = \delta_a(p_i) = (\delta_x, \delta_y) \in [-\frac{1}{2}, \frac{1}{2})^2$. At each level, the readout uses the four tiles surrounding p_i under bilinear interpolation. These are the containing tile and the three adjacent tiles in the direction of p_i within that tile. Their four bilinear weights are

$$w = \big((1-|\delta_x|)(1-|\delta_y|),\ |\delta_x|(1-|\delta_y|),\ (1-|\delta_x|)|\delta_y|,\ |\delta_x||\delta_y|\big) \in [0,1]^4. \tag{27}$$

Unoccupied neighboring tiles are omitted. Across all read levels, let $\{(h_c, w_c)\}_{c=1}^{K_i}$ denote the retained hierarchy keys and their scalar bilinear priors, with $h_c \in \mathbb{R}^{d_{\text{tile}}}$ and $w_c \in (0,1]$. With $W_q, W_v \in \mathbb{R}^{d_{\text{tile}} \times d_{\text{tile}}}$ and $W_r, W_o \in \mathbb{R}^{d \times d_{\text{tile}}}$,

$$\pi_{ic} = \frac{w_c \exp \lambda_{ic}}{\sum_{c'=1}^{K_i} w_{c'} \exp \lambda_{ic'}}, \qquad v_i = W_r u_i + W_o \sum_{c=1}^{K_i} \pi_{ic} W_v h_c \in \mathbb{R}^d, \qquad z_i = \frac{v_i}{\|v_i\|_2}, \tag{28}$$

where λ_{ic} is the multi-head query-key logit of Eq. (20) between the projected query $W_q u_i$ and the projected, rotary-positioned key h_c. The key offset is measured from p_i to the corresponding tile centre in that level's tile units. Equivalently, w_c contributes an additive $\log w_c$ prior to the attention logits, which makes the readout continuous across tile boundaries. The normalization producing z_i is evaluated in 32-bit precision.

C.7 ROUTING, RECEPTIVE FIELD, AND INTERACTION COUNT

Fix a crop, its molecular graph, and its occupied tiles. For transcripts i and j, let $R = \|p_i - p_j\|_\infty$ (in μm) and $\ell^\star = \min\{\ell : a_i^{(\ell)} = a_j^{(\ell)}\}$ be the level of their lowest common ancestor. Let $T_\ell = |\mathcal{T}_\ell|$, let $P_{\max} = \max_\ell P_\ell$, and define the physical radius

$$\rho_\ell = L_{\text{loc}} r_0 + D_{\text{nat}} \sum_{t=0}^{\ell} s_t = L_{\text{loc}} r_0 + D_{\text{nat}} (2^{\ell+1} - 1) s_0. \tag{29}$$

Proposition 1: Routing, reach, and interaction count

For every crop, radius-capped k-NN graph, and encoder parameter setting:

(i) *Lowest-common-ancestor route.* $\log_2(R/s_0) < \ell^\star \le L-1$. The state $H^{(\ell^\star)}_{a_i^{(\ell^\star)}}$ depends on u_j through $\ell^\star + 1$ pooling stages and summarizes $4^{\ell^\star} > (R/s_0)^2$ leaf tiles. If $\ell^\star \in \mathcal{J}_{\text{out}}$, this state is on a path from u_j to z_i.

(ii) *Reach.* $H_a^{(\ell)}$ depends only on transcripts within ℓ_∞ distance ρ_ℓ of tile a. The transcript embedding z_i depends only on transcripts j satisfying

$$\|p_j - p_i\|_\infty \le \rho_{\text{out}} = \max_{\ell \in \mathcal{J}_{\text{out}}} \left(\rho_\ell + \tfrac{3}{2} s_\ell\right).$$

(iii) *Interaction count.* After graph construction, one forward pass evaluates

$$O\left(N_C(k_{\text{nn}}+1)L_{\text{loc}} + N_C K + K^2(P_{\max} + 9D_{\text{nat}}) \sum_{\ell=0}^{L-1} T_\ell + 4N_C K |\mathcal{J}_{\text{out}}| \right) \tag{30}$$

message-passing and attention interactions. Since $T_\ell \le N_C$ for every ℓ, the count is $O(N_C)$ for fixed architecture parameters, versus $\Theta(N_C^2)$ for dense self-attention over transcripts.

Proof. For (i), two points in the same level-$\ell^\star$ half-open tile satisfy $R < s_{\ell^\star} = 2^{\ell^\star} s_0$, while the coarsest tile is the crop. Transcript j enters the leaf pooling of its level-0 tile and each ancestor adds one pooling stage. If $\ell^\star \in \mathcal{J}_{\text{out}}$, the common ancestor is among the hierarchy states available to the readout around transcript i.

For (ii), each local graph layer expands dependence by at most r_0. Each neighborhood-attention block at level t expands it by at most one tile width s_t in ℓ_∞ distance, while pooling takes unions

within the parent and adds no additional displacement. Induction over levels yields ρ_ℓ. Every tile in the four-tile readout neighborhood at level ℓ lies within $\frac{3}{2}s_\ell$ of p_i.

For (iii), the radius-capped k-NN graph has at most $N_C(k_{\mathrm{nn}}+1)$ directed edges including self-loops. Level-0 pooling attends once from K seeds to each transcript. Above level 0, a tile has at most $4K$ child tokens; neighborhood attention uses at most $9K$ keys per tile; and the transcript readout uses at most $4K$ keys per read level. Summing over occupied tiles gives Eq. (30). □

For the baseline configuration, Table 7 gives $\rho_2 = 493\,\mu m$ and $\rho_{\mathrm{out}} = 589\,\mu m$.

C.8 Pretraining objective

Laminar uses a joint-embedding predictive objective (LeCun et al., 2022; Assran et al., 2023). The online encoder E_θ has parameters θ and processes corrupted crops. The target encoder E_ξ has parameters ξ, is initialized from θ, receives the complete crop, and is updated only by exponential moving average (EMA) (Tarvainen & Valpola, 2017; Grill et al., 2020). A predictor P_ϕ with parameters ϕ and a decoder h_ψ with parameters ψ are training-only modules.

Corrupted views. Each crop produces two views $v \in \{\mathrm{spatial}, \mathrm{gene}\}$. The spatial view removes transcripts inside the union of J_{rect} randomly placed rectangles. A rectangle has geometric-mean side length $\ell_s > 0$ and aspect ratio $a > 0$, giving side lengths $\sqrt{a}\,\ell_s$ and $\ell_s/\sqrt{a}$; ℓ_s and a are sampled log-uniformly over the ranges in Table 8. The draw is accepted only when the retained and removed transcript fractions fall inside the stated bounds. The gene view samples a removal fraction ρ_{gene} and deletes whole genes in a random order until that observation fraction is reached. Each view rebuilds its graph, occupied tiles, and tile counts from the surviving transcripts.

Prediction and composition regions. For every crop, view v, and hierarchy level ℓ, let $\mathcal{Q}^{\mathrm{pred}}_{v,\ell}$ contain Q_{pred} prediction regions and let $\mathcal{Q}^{\mathrm{comp}}_{v,\ell}$ contain Q_{comp} composition regions. A level-ℓ region A is an axis-aligned rectangle with nominal side

$$b = \min(s_\ell e^X, b_{\max}), \qquad X \sim \mathcal{U}(-\log 2, \log 2), \tag{31}$$

and aspect perturbation $Y \sim \mathcal{U}(0, \frac{1}{2}\log 2)$, giving side lengths (be^Y, be^{-Y}) in either axis order with equal probability. Here b and $b_{\max}$ are lengths in μm. Prediction regions of the spatial view are centred within masked rectangles. Composition-region centres are sampled near masks or uniformly in the crop interior, excluding a boundary margin m_{crop}. Table 8 gives Q_{pred}, Q_{comp}, $b_{\max}$, and m_{crop}.

Target regional state and weighting. For a region A, let $\mathcal{I}_A = \{i : p_i \in A\}$ and $n_A = |\mathcal{I}_A|$ in the complete crop. Regions with $n_A = 0$ are omitted. For $n_A > 0$, the target field gives

$$S^\xi_A = \sum_{i \in \mathcal{I}_A} z^\xi_i \in \mathbb{R}^d, \qquad y_A = \mathrm{sg}\left(\frac{S^\xi_A}{\|S^\xi_A\|_2}\right) \in \mathbb{S}^{d-1}, \qquad R_A = \frac{\|S^\xi_A\|_2}{n_A}. \tag{32}$$

Regions for which $\|S^\xi_A\|_2 = 0$ are also omitted because y_A is undefined. For $n_A \geq 2$, let $\bar{c}_A = (n_A R_A^2 - 1)/(n_A - 1)$, and define the scalar target weight

$$\omega_A(t) = \begin{cases} 1, & n_A = 1, \\ (1-\eta_t) + \eta_t[\max(0, \bar{c}_A)]^\beta, & n_A \geq 2, \end{cases} \qquad \eta_t = \min\left(1, \frac{t}{T_\omega}\right). \tag{33}$$

Here $t \in \mathbb{N}$ is the optimizer step, $\beta > 0$ is the coherence exponent, and T_ω is the weight-ramp duration in steps. No gradient passes through ω_A.

Region query and predictor. Let $o_A \in [0,1]^2$ be the region centre in crop-normalized coordinates, $b_A \in \mathbb{R}^2_+$ its two physical side lengths, and d_p the predictor width. The query token $q(A, \ell, v) \in \mathbb{R}^{d_p}$ is

$$q(A,\ell,v) = U_{\mathrm{pos}}\left[\cos(2\pi 2^k o_A), \sin(2\pi 2^k o_A)\right]_{k=0}^{L-1} + U_{\mathrm{size}}\log(b_A/W) + E_{\mathrm{level}}[\ell] + E_{\mathrm{view}}[v], \tag{34}$$

where $U_{\mathrm{pos}} \in \mathbb{R}^{d_p \times 4L}$, $U_{\mathrm{size}} \in \mathbb{R}^{d_p \times 2}$, $E_{\mathrm{level}} \in \mathbb{R}^{L \times d_p}$, and $E_{\mathrm{view}} \in \mathbb{R}^{2 \times d_p}$ are learned. The predictor cross-attends from this query to the online level-ℓ hierarchy tokens and returns $r_A \in \mathbb{R}^{d_p}$. With $W_\mu \in \mathbb{R}^{d \times d_p}$,

$$\mu_A = \frac{W_\mu r_A}{\|W_\mu r_A\|_2} \in \mathbb{S}^{d-1}. \tag{35}$$

Stochastic prediction channel. A fixed concentration $\kappa_0 > 0$ defines

$$\widetilde{z}_A \sim \mathrm{vMF}(\mu_A, \kappa_0), \qquad \widehat{y}_A = \frac{h_\psi(\widetilde{z}_A)}{\|h_\psi(\widetilde{z}_A)\|_2} \in \mathbb{S}^{d-1}. \tag{36}$$

The decoder $h_\psi : \mathbb{R}^d \to \mathbb{R}^d$ is a two-layer MLP whose input is only the sampled direction. Sampling follows the standard von Mises–Fisher construction (Wood, 1994) and has no learned parameters.

Prediction loss. The hierarchy levels are weighted equally; equivalently $\alpha_\ell = 1$ for every $\ell \in \{0, \dots, L-1\}$. With $\epsilon = 10^{-6}$,

$$\mathcal{L}_{\mathrm{pred}} = \sum_{\ell=0}^{L-1} \sum_{v} \frac{\sum_{A \in \mathcal{Q}_{v,\ell}^{\mathrm{pred}}} \omega_A \left(1 - \widehat{y}_A^\top y_A\right)}{\max\left\{\sum_{A \in \mathcal{Q}_{v,\ell}^{\mathrm{pred}}} \omega_A, \epsilon\right\}}. \tag{37}$$

Field composition. The predictor supervises the internal hierarchy before the final transcript readout. We therefore impose the same regional target directly on the deployed field. Let $\mathcal{I}_{A,v} \subseteq \mathcal{I}_A$ be the transcripts of region A that survive corruption v. The online regional direction is

$$\Phi_v^\theta(A) = \frac{\sum_{i \in \mathcal{I}_{A,v}} z_{i,v}^\theta}{\left\|\sum_{i \in \mathcal{I}_{A,v}} z_{i,v}^\theta\right\|_2} \in \mathbb{S}^{d-1}. \tag{38}$$

A composition region is valid when the target direction is defined, the corrupted view contains at least one transcript with a nonzero regional sum, and, for the spatial view, the retained fraction is at least τ. Let $\mathcal{Q}_v^{\mathrm{comp}}$ denote the union of valid composition regions over crops and hierarchy levels. Then

$$\mathcal{L}_{\mathrm{field}} = \sum_{v} \frac{\sum_{A \in \mathcal{Q}_v^{\mathrm{comp}}} \omega_A \left(1 - \Phi_v^\theta(A)^\top y_A\right)}{\max\left\{\sum_{A \in \mathcal{Q}_v^{\mathrm{comp}}} \omega_A, \epsilon\right\}}. \tag{39}$$

Isotropy. For view v, let $\{\phi_b\}_{b=1}^{B_v} \subset \mathbb{S}^{d-1}$ be the unweighted online regional directions of all valid composition regions in the global training step. Define

$$\bar{\phi}_v = \frac{1}{B_v} \sum_{b=1}^{B_v} \phi_b \in \mathbb{R}^d, \qquad M_v = \frac{1}{B_v} \sum_{b=1}^{B_v} \phi_b \phi_b^\top \in \mathbb{R}^{d \times d}, \tag{40}$$

and

$$\mathcal{L}_{\mathrm{iso}} = \sum_{v} \left(\|\bar{\phi}_v\|_2^2 + \lambda_{\mathrm{mom}} \left\| M_v - \frac{1}{d} I_d \right\|_F^2 \right), \tag{41}$$

where $\lambda_{\mathrm{mom}} \geq 0$ is the second-moment weight and I_d is the $d \times d$ identity. Under distributed training, B_v, $\sum_b \phi_b$, and $\sum_b \phi_b \phi_b^\top$ are all-reduced before the loss is evaluated.

Complete objective and update. The training objective is

$$\mathcal{L} = \mathcal{L}_{\mathrm{pred}} + \lambda_{\mathrm{field}} \mathcal{L}_{\mathrm{field}} + \lambda_{\mathrm{iso}} \mathcal{L}_{\mathrm{iso}}, \tag{42}$$

with coefficients given in Table 8. At each optimizer step, the target encoder first processes the complete crops under stop-gradient and forms the regional targets. The online encoder then processes both corrupted views, after which gradients are clipped and θ, ϕ, ψ are updated. Finally, the target encoder is updated as

$$\xi \leftarrow m_t \xi + (1 - m_t)\theta, \tag{43}$$

where $m_t \in [0, 1)$ follows the EMA schedule in Table 8. The predictor and decoder are discarded after pretraining; downstream experiments use the online encoder.

C.9 Sampling training crops

A pass over a section translates the regular grid of width-W windows by one random two-dimensional offset and visits each occupied shifted-grid window once. A window is eligible when its transcript density exceeds the training floor. If an eligible window contains $n > n_{\max}$ transcripts, where

Table 4: **Public pretraining corpus by repository.** Sections are physical tissue sections and are the sampling unit during pretraining. Transcript and cell totals are after corpus preprocessing.

Repository	Studies	Sections	Tx (10^9)	Cells (10^6)	Licence / terms
NCBI GEO	42	357	13.521	41.161	repository terms
BioImage Archive	7	208	4.602	41.607	CC0
Zenodo	13	51	2.181	9.241	CC BY 4.0
figshare	1	1	0.036	0.155	CC BY 4.0
10x Genomics	13	15	1.819	4.612	CC BY 4.0
Total	76	632	22.159	96.777	

$n_{\max} \in \mathbb{N}$ is the segment-specific cap, transcript i carries a fixed pseudorandom $u_i \in [0, 1)$ and is retained when

$$u_i \in [o_r, o_r + f) \bmod 1, \qquad f = \frac{n_{\max}}{n}, \qquad o_r = \mathrm{frac}(r\varphi + o_s), \qquad \varphi = \frac{\sqrt{5}-1}{2}. \tag{44}$$

Here $r \in \mathbb{N}$ is the section-offering index and $o_s \in [0, 1)$ is a fixed section-specific offset. The offset advances across repeated visits so that dense windows do not repeatedly expose the same subset of transcripts. The same retained subset is used to construct the complete crop and both corrupted views.

Sections are visited according to the curriculum of Appendix F.1. The crop density threshold, global batch size, stage-dependent sampling exponent, and cap used by the baseline run are reported there and in Table 8.

C.10 Whole-section inference

A complete section is encoded with overlapping square windows of width W. Let m_{inf} be the physical context margin and $s_{\mathrm{inf}} = W - 2m_{\mathrm{inf}}$ the stride. For section origin $(x_0, y_0) \in \mathbb{R}^2$, transcript i is owned by window

$$k(i) = \left(\left\lfloor \frac{x_i^{\mathrm{pos}} - x_0}{s_{\mathrm{inf}}} \right\rfloor, \left\lfloor \frac{y_i^{\mathrm{pos}} - y_0}{s_{\mathrm{inf}}} \right\rfloor \right) \in \mathbb{Z}^2. \tag{45}$$

Each transcript is stored once, from its owner window, while the overlapping window provides context. Windows above the inference transcript cap are encoded with a deterministic uniform subsample, and the radius-capped k-NN graph is constructed deterministically. Regional states that cross window boundaries are summed before the final normalization in Eq. (12). The baseline margin, stride, and transcript cap are listed in Table 7.

D Public pretraining corpus

Laminar is pretrained on 632 human imaging-based spatial transcriptomics sections from 76 public studies, comprising 22,158,504,713 detected transcripts and 96,776,773 vendor-segmented cells. Cell counts are reported to describe the corpus but cell segmentations are not encoder inputs. The corpus contains molecule-resolved in situ data only. The 10x Genomics contribution consists of 14 Xenium sections and one Atera whole-transcriptome section; no Visium or dissociated Chromium data are included. Table 4 summarizes the sources, and Table 6 lists every study.

D.1 Acquisition and preprocessing

For pretraining we use the molecule-level transcript table from each section and pass only gene identity and physical coordinates to the encoder. Morphology images, cell-by-gene matrices, pathology annotations, and cell segmentations are not used. Duplicate republications are counted once. Gene identifiers are standardized to approved HGNC symbols and control probes are removed; when a transcript-quality field is available, transcripts with quality below 20 are discarded. Coordinates remain in the physical units supplied by the assay. Repository access, archive handling, gene-symbol harmonization, and storage are described in Appendix G.9.

Table 5: **Cohorts excluded from pretraining.** The colored use tag distinguishes transfer evaluation, label-free checkpoint selection, qualitative visualization, and training monitoring. Public accessions are shown where applicable.

Cohort	Accession / source	Description	Use
Lung fibrosis	GSE250346; HEST-1k	20 Xenium sections from 15 donors	transfer evaluation
HEST-1k Xenium	HEST-1k	59 annotated sections from 13 organs	transfer evaluation
Breast cancer block	GSE243280; HEST-1k / CAMEO-Breast	4 sections from one tissue block	transfer evaluation
Meningioma microarray	in-house	14 slides; 11 retained for transfer tasks	transfer evaluation
Anchor regions	in-house	5 held-out microarray regions	checkpoint selection
Atera cervical carcinoma	10x Genomics Atera preview dataset	1 whole-transcriptome FFPE section	visualization
Validation sections	public; excluded from corpus by density criterion	15 sections	monitoring

D.2 Corpus inclusion and held-out cohorts

We include human imaging-based spatial transcriptomics sections with molecule-level transcript tables after removing duplicate records and cohorts reserved for evaluation or model selection. We also exclude 66 public sections whose mean occupied-grid density is below 3,000 transcripts per $192\,\mu m$ grid cell, which is too sparse for the pretraining crop sampler. Evaluation, checkpoint-selection, visualization, and monitoring cohorts are removed by accession or section identifier before the pretraining manifest is built.

D.3 Per-study inventory

Table 6 lists every study and the sections, transcripts, and cells that entered pretraining. Platform labels follow the released panel associated with each section; studies spanning both panel classes are marked "v1 and 5K". Tissue labels follow repository metadata or the study title, and a dash denotes metadata unavailable at the study level.

Table 6: **Every study in the public pretraining corpus.** Studies are grouped by host, and the group heading gives the licence. Platforms are Xenium v1, Xenium Prime 5K and Atera WTA. Transcripts are given in millions and cells in thousands. 10x Genomics sections are listed by their resource identifier, which is also the file prefix of the downloaded bundle, and the three human brain preview datasets share one row.

Accession	Platform	Tissue	Sections	Tx (M)	Cells (k)
NCBI GEO, repository terms					
GSE315435	Prime 5K	kidney, developing	11	3,047.6	3,034.6
GSE290468	v1 and 5K	–	51	2,892.0	6,264.7
GSE260452	v1	brain, ependymoma	56	1,219.1	5,916.0
GSE317381	Prime 5K	brain; ovary	4	1,049.3	532.0
GSE316917	v1	brain, glioblastoma	9	604.4	2,276.1
GSE328422	Prime 5K	breast; lymph node	5	543.0	1,047.6
GSE335898	v1	brain	11	504.4	966.4
GSE326837	Prime 5K	synovium	9	443.4	609.6
GSE296852	v1	urethra	4	331.5	2,874.0
GSE300147	v1	head and neck	18	327.6	1,183.7

Table 6, continued

Accession	Platform	Tissue	Sections	Tx (M)	Cells (k)
GSE313662	Prime 5K	pancreas	6	268.8	970.6
GSE299207	v1 and 5K	bone marrow	22	264.4	2,126.7
GSE335761	v1	lung	14	263.5	2,819.7
GSE319763	Prime 5K	lung	3	232.4	1,021.5
GSE325489	v1	brain, nucleus accumbens	3	218.8	587.8
GSE291246	v1	lymph node	16	195.5	1,768.1
GSE290041	v1	–	9	125.7	722.4
GSE326660	v1	skin, Merkel cell carcinoma	6	119.4	681.5
GSE286964	v1	skin	2	95.4	285.1
GSE319771	v1	heart	9	94.1	723.3
GSE305629	v1	bladder	2	81.3	786.3
GSE307661	v1	bone marrow	3	59.0	290.1
GSE326753	Prime 5K	–	4	51.3	239.9
GSE302642	v1	brain, glioma	6	47.5	134.3
GSE315295	Prime 5K	skin	11	45.9	145.3
GSE320142	v1	ovary; lung	2	45.9	465.6
GSE302264	v1	kidney	4	41.6	374.6
GSE312420	v1 and 5K	intestine	10	37.7	353.3
GSE341250	v1	–	2	34.8	149.8
GSE271936	v1	brain, glioma	2	29.2	150.8
GSE333685	Prime 5K	coronary artery	1	28.0	49.6
GSE285158	Prime 5K	–	3	25.5	206.3
GSE336556	v1	lung	8	24.0	329.2
GSE319755	v1	lung	2	22.2	156.7
GSE310386	v1	heart	5	21.1	106.0
GSE322974	Prime 5K	bone marrow	1	19.0	75.2
GSE324934	v1	head and neck, organotypic culture	6	17.1	132.4
GSE306150	v1	aorta	5	16.0	251.6
GSE313090	v1	lung	1	13.8	255.1
GSE293970	v1	synovium	2	8.3	27.2
GSE302502	v1	brain, glioma	7	5.7	40.2
GSE315246	v1	artery	2	5.3	30.0
EBI BioImage Archive, CC0					
S-BIAD3618	v1	lung	30	1,651.9	30,546.1
S-BIAD3788	v1 and 5K	skin	125	996.1	2,831.0
S-BIAD2208	v1	colorectum	19	908.3	4,387.5
S-BIAD3797	Prime 5K	lung	17	456.0	1,685.6
S-BIAD3103	v1 and 5K	thyroid	9	392.2	1,073.2
S-BIAD1093	v1	brain, medulloblastoma	5	175.5	1,015.0
S-BIAD2214	Prime 5K	skin	3	21.8	68.6
Zenodo, CC BY 4.0					
zenodo.18000256	v1 and 5K	breast; colorectum; lung; lymph node; ovary	10	1,003.3	3,164.8
zenodo.20199229	v1	bladder; breast; colorectum; lung; lymph node; ovary	6	238.8	2,689.6
zenodo.15393965	v1	brain	4	159.3	242.7
zenodo.11534381	v1	brain, caudate and putamen	3	134.9	256.1
zenodo.15838935	v1	small intestine	8	111.1	885.1
zenodo.15395905	v1	brain	2	83.2	225.7
zenodo.11612060	v1	brain, caudate and putamen	2	82.2	200.5
zenodo.11615246	v1	brain, caudate and putamen	1	72.9	133.1
zenodo.17155546	v1	lung, postnatal	5	72.7	817.0
zenodo.15396670	v1	brain	1	70.0	162.9
zenodo.11609973	v1	brain, caudate and putamen	1	67.3	144.5
zenodo.17508360	v1	skin	4	49.8	228.9
zenodo.15186190	v1	brain	4	35.2	90.3

Table 6, continued

Accession	Platform	Tissue	Sections	Tx (M)	Cells (k)
figshare, CC BY 4.0					
figshare.30415409	v1	liver	1	36.2	155.1
10x Genomics datasets, CC BY 4.0					
WTA_Preview_FFPE_Breast_Cancer	Atera WTA	breast	1	624.3	170.1
Xenium_Prime_Human_Ovary_Cancer_FF	Prime 5K	ovary	1	343.0	200.9
Xenium_V1_Human_Ovary_Cancer_FF	v1	ovary	1	167.8	205.0
Xenium_Prime_Human_Lung_Cancer_FFPE	Prime 5K	lung	1	149.4	278.1
Xenium_Prime_Ovarian_Cancer_FFPE_XRrun	Prime 5K	ovary	1	120.5	406.6
Xenium_Prime_Cervical_Cancer_FFPE	Prime 5K	cervix	1	89.6	837.0
Xenium_Prime_Breast_Cancer_FFPE	Prime 5K	breast	1	80.2	683.8
Xenium_Preview_Human_Lung_Cancer_With_Add_on_2_FFPE	v1	lung	1	67.3	530.9
Xenium_V1_Human_Kidney_FFPE_Protein_updated	v1	kidney	1	62.9	465.2
Xenium_V1_FFPE_Human_Brain_{Alzheimers, Glioblastoma, Healthy}_With_Addon	v1	brain	3	33.3	110.2
Xenium_V1_Human_Lung_Cancer_FFPE	v1	lung	1	30.9	278.4
Xenium_V1_hLung_cancer_section	v1	lung	1	25.8	150.2
Xenium_Preview_Human_Non_diseased_Lung_With_Add_on_FFPE	v1	lung	1	24.0	295.8

E Baseline model configuration

Tables 7 and 8 summarize the baseline encoder and pretraining configuration. Stage-dependent sampling values are given in Appendix F.

Weight decay excludes biases, normalization parameters, embeddings, one-dimensional parameters, and PMA seed vectors. Gradients of the local graph layers and the remainder of the network are clipped separately. A step with a non-finite gradient is skipped together with its EMA update.

F Pretraining

F.1 Curriculum and sampling

Pretraining uses four curriculum stages. Stage 1 trains for 800 steps on the public Atera whole-transcriptome FFPE breast-cancer section. Stage 2 adds the remaining 14 10x Genomics sections, all Xenium, for 2,500 steps. Stage 3 trains for 4,000 steps on the other 617 public sections, and Stage 4 samples the complete 632-section corpus. Each sampling track stays on one section for four consecutive steps, and every optimizer step combines two tracks. Studies are traversed round-robin, with section multiplicity proportional to $n_{\text{cells}}^{\alpha}$ using $\alpha = 0, 0.3, 0.3, 0.5$ in Stages 1–4. In Stage 3, 15 section from Stage 2.

Crops are sampled as in Appendix C.9, with a density floor of 10,486 transcripts per $1{,}024\,\mu m$ crop. Windows above the execution cap use the rotating subsampling rule in Eq. (44); the cap limits memory use without changing which windows are selected. The final recorded 40 GB segment used

Table 7: **Baseline encoder architecture.** Thin spaces separate thousands in parameter counts. Dimensions refer to the online encoder; predictor modules are used only during pretraining.

Symbol	Quantity	Baseline value
Input and local graph		
G	gene-embedding rows	45 104
d_g	gene-embedding width	128
k_{nn}	maximum k-NN neighbors per transcript	8
r_0	molecular-graph radius	15 μm
L_{loc}	residual GATv2 layers	3
d_h	local graph width	1 024 (8×128 heads)
$d_{u,h}, d_{u,x}, d_{\mathrm{tile}}$	local-context, direct-gene, concatenated widths	896, 128, 1 024
Spatial hierarchy		
W, s_0, L	crop width, leaf-tile width, hierarchy levels	1 024 μm, 16 μm, 7
K	summary tokens per occupied tile	4
d_{tile}	hierarchy token width	1 024
H	attention heads (head width)	16 (64)
P_ℓ	PMA blocks per level	1 at $\ell = 0$; 2 otherwise
D_{nat}	neighborhood-attention blocks per level	4
r_N	neighborhood radius	1 tile
Transcript readout		
$\mathcal{J}_{\mathrm{out}}$	hierarchy levels read per transcript	$\{0, 1, 2\}$
$K_i^{\max}$	maximum hierarchy keys per transcript	48
d	transcript-field dimension	1 024
Pretraining-only predictor		
d_p	predictor width	384
P_ϕ	predictor blocks	8
H_p	predictor attention heads	6
h_ψ	decoder hidden/output widths	1 024, 1 024, 1 024
Parameter counts		
	encoder total	95 057 920
	pretraining-only modules	20 663 296
	encoder + pretraining modules	115 721 216

Table 8: **Baseline pretraining settings.** Symbols are defined in Appendix C.8. Prediction levels are weighted equally ($\alpha_\ell = 1$), so the factor is omitted from Eq. (37).

Quantity	Baseline value
Corruption and region sampling	
spatial-mask rectangles per crop	8
rectangle geometric-mean side	64–512 μm
rectangle aspect ratio	$1/2$–2
removed transcript fraction, spatial view	0.25–0.75
removed observation fraction, gene view	0.10–0.40
minimum retained fraction	0.15
prediction regions per level and crop, Q_{pred}	8
composition regions per level and crop, Q_{comp}	160
query-side cap, $b_{\max}$	512 μm
crop margin for region centers, m_{crop}	256 μm
spatial-view visibility threshold, τ	0.9
coherence exponent, β	1
coherence-weight ramp, T_ω	10 000 steps
vMF concentration, κ_0	2 560
Objective	
field-composition coefficient, λ_{field}	0.5
isotropy coefficient, λ_{iso}	0.25
second-moment coefficient, λ_{mom}	1
Optimization	
optimizer	AdamW, $\beta = (0.9, 0.98)$, $\epsilon = 10^{-8}$
peak learning rate	4.3×10^{-5}
linear warmup	1 000 steps
weight decay	0.04 on matrix weights
global-norm clipping	1.0 local layers; 5.0 remainder
EMA momentum	0.996 $\rightarrow$ 0.999 over 100 000 steps
global batch	8 crops
numerical precision	bfloat16 autocast; FP32 normalizations
random seed	0

Table 9: **Pretraining throughput by curriculum stage.** Values are medians over measurements logged every 50 optimizer steps. "Model time" is forward/backward/optimizer time; loader wait is the remaining fraction of step wall time spent waiting for the input pipeline.

Stage	Steps	Tx/crop	Step (s)	Model (s)	Loader wait	Peak GB
1	0–800	149 986	34.7	32.1	10.3%	37.8
2	800–3 300	226 255	28.0	27.5	1.3%	34.9
3	3 300–7 300	122 952	20.5	20.4	0.03%	38.9
4	7 300–10 000	112 649	19.3	19.2	0.03%	38.9

$n_{\max} = 123{,}000$ transcripts per crop. By the selected checkpoint, sampled windows had offered 1.382×10^{11} transcripts, of which 1.847×10^{10} were processed after capping; the median optimizer step processed 941,538 transcripts.

F.2 Hardware and throughput

Training used NVIDIA A100-SXM4 (40 GB), L40S (45 GB), and H200 (141 GB) GPUs under distributed data parallelism. The global batch was fixed at eight crops and isotropy statistics were synchronized across GPUs. Up to the baseline checkpoint, accumulated optimizer-step time was 111.6 h; the median step took 27.2 s (10th–90th percentile: 16.8–75.9 s). The molecular graph is constructed on GPU, and gradient checkpointing is used in both the local graph layers and the spatial hierarchy. Table 9 reports stage-level throughput and memory.

F.3 Checkpoint selection

Because the target encoder moves during training, we do not use the training loss for checkpoint selection. Instead, every 500 steps we compute a label-free anchor score on 16 fixed crops from the five anchor regions in Table 5. For each crop, the transcripts are split at random into two halves and each half is pooled into 25 μm bins with at least 8 transcripts. The score is the mean cosine between paired bins of the two halves minus the mean cosine after the bins are re-paired at random, averaged over crops. The score uses no biological label or downstream probe. The baseline checkpoint is the one with the highest anchor score, 0.256, reached at step 10,000 of a run that stopped at 10,300. A teacher-referenced loss on the 15 held-out validation sections of Table 5 was logged every 100 steps for monitoring and took no part in selection.

G Transfer evaluation

All pretrained encoders are frozen during evaluation; only the common probe in Appendix G.4 is optimized. We define the tasks first, followed by representation construction, probe fitting, fold construction, and metrics. Comparator checkpoints and software sources are listed in Appendix G.2.

G.1 Tasks and data

The spatial support in the main results table refers to the biological scale of the prediction target, not necessarily the unit on which the metric is computed. Cell type and gene imputation are cell-level tasks. Pathology region and breast niche are multicellular targets evaluated cell-wise. Region composition is evaluated on 200 μm tiles, and the meningioma endpoints are evaluated on tissue cores.

Cell type. The 59 annotated sections of HEST-1k (Jaume et al., 2024) contribute at most 16,384 cells each. Labels are the six major classes of the released annotation (270,697 malignant, 174,060 lymphoid, 216,657 mesenchymal, 176,242 myeloid, 157,747 epithelial and 103,210 endothelial cells). Quality-control calls and unresolved doublet calls are excluded upstream. We evaluate two fixed 20-fold section partitions, giving 40 held-out fold evaluations.

Gene imputation. Five sections of the lung fibrosis cohort (NCBI857, 859, 861, 867 and 880, one donor each) hold out between 163 and 168 genes each, and every method is re-encoded with the

Table 10: **Downstream tasks and primary metrics.** Folds hold out sections, or slides for the microarray. "Scored unit" is the unit on which the metric is computed, and "Fold evals." is the number of scored held-out evaluations after class eligibility is applied.

Task	Cohort	Scored unit	Units	Fold evals.	Scored output	Primary metric
Cell type	HEST-1k, 59 sections / 13 organs	cell	1 098 613	40	6 classes	chance-corrected BA
Gene imputation	lung fibrosis, 5 sections / 5 donors	cell	146 056	5	149–158 genes/fold	profile cosine
Pathology region	lung fibrosis, 20 sections / 15 donors	cell	231 546	20	22/25 classes	chance-corrected BA
Region composition	lung fibrosis, 20 sections	200 μm tile	935	20	339 genes/fold	profile cosine
Breast niche	one breast-cancer block, 4 sections	cell	11 579	4	2/10 classes	chance-corrected BA
WHO grade	meningioma microarray, 11 slides	core	300	11	3 classes	chance-corrected BA
Methylation subclass	same 11 slides	core	298	10	5 classes	chance-corrected BA
Malignancy class	same 11 slides	core	312	10	3 classes	chance-corrected BA

held-out genes removed from that section's input. Because the held-out sets differ between sections, the probe for held-out section E predicts only genes of E that at least one training section also held out, and a training cell contributes to a gene's loss only if that gene was hidden from the encoder for that cell. This leaves 149–158 scored genes per fold. Cells need at least five visible transcripts. The target is the cell's composition over the scored genes. Counts baselines use the 320-gene common candidate set with each cell's own held-out genes set to zero. BANKSY is excluded because its feature width differs between sections after gene removal.

Pathology region. The 20 lung fibrosis sections carry 25 region classes drawn by pathologists on the registered H&E images and transferred to cells by the source study (Vannan et al., 2025). The task scores 231,546 annotated cells. Twenty-two of the 25 labels satisfy the scorable-class criterion in at least one held-out fold.

Region composition. For each of 935 held-out 200 μm tiles, with at most 100 tiles per section and centers at least 400 μm apart, the input support is the surrounding 3×3 tile ring without the center, excluding cells within 15 μm of the center tile. The target is the measured composition of the center tile over 339 scored genes. Each method constructs its input representation from the surrounding ring according to its readout in Appendix G.3. Two model-free controls are also scored, one predicting the training-fold mean composition for every held-out tile and one copying the measured composition of the surrounding ring.

Breast niche. Cells of the four sections TENX94–TENX97 of one breast cancer block (Janesick et al., 2023) take the CAMEO-Breast (Theis Lab, 2026) niche label of the nearest annotated cell within 2.0 μm. Only two of ten classes (947 and 10,632 cells) occur in all four sections. The task is reported as a diagnostic because all four sections come from one tissue block and therefore do not provide a donor-disjoint transfer test.

Meningioma microarray. Tumor cores come from 11 of 14 in-house tissue-microarray slides. Three slides were excluded because fewer than 60 % of their transcripts matched an annotated cell. A transcript joins the core of its nearest annotated cell within 25 μm. WHO grade contains 112 grade-2, 155 grade-1, and 33 grade-3 cores. Methylation subclass contains ben-1 (72), ben-2 (89), ben-3 (45), Int-A (62), and mal (30) cores. Malignancy class groups 312 cores into benign (206), intermediate

Table 11: **Comparator checkpoints and software.** "Source" identifies the exact public artifact used for the frozen representation or the package used for methods fitted on the evaluation cohort.

Model	Source	Pretraining / fitting	d
Nicheformer	HF `theislab/Nicheformer`; `model.safetensors`	SpatialCorpus-110M	512
scGPT-spatial	figshare 28356068; `best_model.pt`	30M profiles, 821 slides	512
TERRA-96	HF `lotfollahi-lab/TERRA-96M`; `model_checkpoint.pt`	HST-Corpus, 96M cells	384
TERRA-112	HF `lotfollahi-lab/TERRA-112M`	HST-Corpus, 112M cells	384
UCE-4L	GitHub `snap-stanford/UCE`; Figshare file 42706576, `4layer_model.torch`	CELLxGENE, 36M cells	1 280
Novae	HF `MICS-Lab/novae-human-0`	30M cells, 18 tissues	64
SToFM	GitHub `PharMolix/SToFM`; checkpoint link in the official README	SToCorpus-88M	256
SToFM-CE	same release; cell-encoder module only	same	256
scVI	`scvi-tools` 1.3.3	fitted on evaluation cohort	30
BANKSY	`prabhakarlab/Banksy_py`	fitted per section	1 425–54 090
CellCharter	`cellcharter` 0.3.7	fitted per section	40

(76), and malignant (30). The methylation labels come from DNA methylation arrays, an assay independent of the transcript input. WHO grade has 11 scored slide folds. Methylation subclass and malignancy class have 10 because one held-out slide contains fewer than two scorable classes for those endpoints.

G.2 Comparator models

Table 11 lists the checkpoint or software source used for each comparator. Frozen pretrained models use released checkpoints without representation fitting on the evaluation cohort. scVI, BANKSY, and CellCharter are reported separately as evaluation-fit methods. Raw counts and PCA-64 are non-pretrained baselines. For tasks other than gene imputation, frozen comparator embeddings are computed once per section and cached at cell level. Gene imputation reconstructs every representation from inputs with the fold-specific held-out genes removed. If a method does not provide an embedding for a task, that entry is left missing rather than replaced by a different readout.

The comparators are Nicheformer (Tejada-Lapuerta et al., 2025), scGPT-spatial (Wang et al., 2025), TERRA (Birk et al., 2026), UCE (Rosen et al., 2026), Novae (Blampey et al., 2025), SToFM (Zhao et al., 2025), scVI (Lopez et al., 2018), BANKSY (Singhal et al., 2024) and CellCharter (Varrone et al., 2023). We use each model's released inference path and native readout. Nicheformer uses the mean over the 1,497 last-layer token slots; scGPT-spatial uses the released gene-mean readout; UCE uses the released four-layer checkpoint; and Novae uses the cell-neighborhood representation without its prototype head. scVI was fitted without task labels on the evaluation cells of each cohort. On HEST-1k, section was used as the batch key and donor as a categorical covariate, with 7 training epochs. BANKSY concatenates a cell's expression with its neighborhood mean and an azimuthal Gabor filter. On the 18,030-gene meningioma panel this gives 54,090 columns. BANKSY and CellCharter were fitted inside each evaluation section and cannot cross a fold boundary.

G.3 Representation construction

Laminar fields. For all tasks except gene imputation, each section is encoded once with the frozen online encoder using the windowed inference of Appendix C.10. Gene imputation is the exception. Each of its five sections is re-encoded after that section's held-out genes have been removed from the input. The random-initialization control uses the same architecture at seed 0.

Laminar and cell-level comparators. A cell is the set of transcripts carrying one vendor cell identifier. Laminar represents each cell by the regional embedding $\Phi(c)$ of its transcripts, Eq. (12). Cells with fewer than three encoded transcripts are dropped. For units larger than one cell, Laminar and all cell-level comparison methods use the same unweighted mean of the included ℓ_2-normalized cell vectors, with no second normalization. Region composition applies this rule to cells in the surrounding ring defined in Appendix G.1. Segmentation enters Laminar only through this evaluation readout and never through the encoder.

G.4 PROBE

Every method, baseline and control is scored with the same probe,

$$h(x) = W_2\,\mathrm{GELU}\big(W_1(x-\mu_{\mathrm{tr}})/\sigma_{\mathrm{tr}} + b_1\big) + b_2, \qquad W_1 \in \mathbb{R}^{256\times d}, \tag{46}$$

where μ_{tr} and σ_{tr} are the per-dimension mean and standard deviation of the training fold, with σ_{tr} clipped below at 10^{-6} and frozen during fitting. Weights and biases start from the uniform draw on $(-1/\sqrt{\text{fan-in}},\ 1/\sqrt{\text{fan-in}})$ that PyTorch uses for linear layers. The probe is trained for 2,000 steps with AdamW (Loshchilov & Hutter, 2019) at learning rate 10^{-3} and weight decay 10^{-4}, a cosine schedule to zero, and gradient clipping at norm 1. It uses no dropout, no validation split and no early stopping. Each step draws a batch with replacement from the training fold. The batch size $B = \min(1024, \min_f n_f^{\mathrm{tr}})$ is shared by all folds of a task, so every fold receives the same number of gradient samples. For efficiency, folds of a task are fit in one batched module with a leading fold dimension; parameters are not shared across folds. Every fold is fit from three initializations with head seeds 0, 1, and 2.

Classification uses cross-entropy with the class weights

$$w_c = \frac{n^{\mathrm{tr}}}{|\mathcal{C}^{\mathrm{tr}}|\, n_c^{\mathrm{tr}}} \quad \text{for } c \in \mathcal{C}^{\mathrm{tr}}, \qquad w_c = 0 \text{ otherwise}, \tag{47}$$

where n_c^{tr} counts the training units of class c, n^{tr} all training units, and $\mathcal{C}^{\mathrm{tr}}$ the classes present in the training fold. The predicted class of a held-out unit is the argmax of the softmax restricted to the scorable classes of its fold (Appendix G.5), so prediction and truth range over the same class set.

Regression targets are compositions. A unit u with counts c_{ug} over the genes $\mathcal{G}_u$ observed for it has target $t_{ug} = c_{ug}/\sum_{g'\in\mathcal{G}_u} c_{ug'}$. The loss is the squared error averaged over the scored (unit, gene) entries. For gene imputation these are the entries of genes hidden from the encoder for that unit (Appendix G.1).

Inputs wider than 4,096 dimensions, which occur only for BANKSY and the raw-count baseline, first pass through a fixed sparse Achlioptas projection to 4,096 dimensions (Achlioptas, 2003). Projection entries are $\sqrt{3}$, 0, and $-\sqrt{3}$ with probabilities $1/6$, $2/3$, and $1/6$, scaled by $1/\sqrt{4096}$. The random seed is the input width, and the projection is independent of data, labels, and folds.

G.5 SECTION-LEVEL FOLDS AND SCORABLE CLASSES

All splits hold out entire tissue sections, or entire slides for the meningioma microarray. They are therefore section-disjoint rather than generally donor-disjoint. The breast-niche task is the clearest exception to a specimen-level interpretation because its four sections come from one tissue block. For cell type, pathology region, region composition, and breast niche, 33 %, 42 %, 37 %, and 100 % of held-out section evaluations, respectively, have a donor represented in the corresponding training fold. Gene imputation uses five sections from five donors. For meningioma, 263 of 265 donors occur on only one slide, so holding out slides nearly separates patients.

When the requested number of folds is at least the number of sections, leave-one-section-out folds are used. Otherwise, sections are assigned to folds while balancing the class distribution

$$\min \sum_c \mathrm{sd}_f\left(\frac{n_{c,f}}{n_c}\right), \tag{48}$$

where $n_{c,f}$ is the number of units of class c assigned to fold f and n_c is its total count.

For classification, a class is *scorable* in fold f if it contains at least 20 training units and at least 5 held-out units. A fold is scored only when at least two classes are scorable. These thresholds avoid class-balanced scores supported by essentially no examples on either side of the split. For the cell-type task, two deterministic balanced 20-fold partitions are evaluated, giving 40 held-out fold evaluations. Pathology region and region composition use 20 section-held-out evaluations, breast niche uses four, and gene imputation uses five. WHO grade uses 11 slide-held-out folds. Methylation subclass and malignancy class each have 10 scored folds because one slide fold contains fewer than two scorable classes. Table 10 summarizes the resulting counts.

G.6 Metrics and aggregation

Metrics are computed separately in each held-out fold after restricting classification to the fold's scorable class set $\mathcal{C}_f$. Classification tasks are imbalanced and the number of scorable classes can differ between folds, so the primary score is chance-corrected balanced accuracy (the "adjusted" balanced accuracy implemented by scikit-learn). Let $C \in \mathbb{N}^{K\times K}$ be the confusion matrix for $K = |\mathcal{C}_f|$ classes, with rows denoting true classes and columns predictions. For class c, recall and precision are

$$R_c = \frac{C_{cc}}{\sum_j C_{cj}}, \qquad P_c = \frac{C_{cc}}{\sum_i C_{ic}}, \tag{49}$$

with $F1_c = 2P_cR_c/(P_c + R_c)$ when $P_c + R_c > 0$. Balanced accuracy and its chance-corrected form are

$$\mathrm{BA}_f = \frac{1}{K}\sum_{c=1}^{K} R_c, \qquad \mathrm{BA}_f^{\mathrm{adj}} = \frac{\mathrm{BA}_f - 1/K}{1 - 1/K}. \tag{50}$$

Uniform random guessing has expectation 0 on this scale and perfect classification has value 1. We also report unadjusted balanced accuracy, macro F1 $\frac{1}{K}\sum_c F1_c$, ordinary accuracy $\sum_c C_{cc}/\sum_{ij} C_{ij}$, and the multiclass Matthews correlation coefficient

$$\mathrm{MCC} = \frac{cs - \sum_k p_k t_k}{\sqrt{(s^2 - \sum_k p_k^2)\,(s^2 - \sum_k t_k^2)}}, \tag{51}$$

where $c = \sum_k C_{kk}$, $s = \sum_{ij} C_{ij}$, $t_k = \sum_j C_{kj}$, and $p_k = \sum_i C_{ik}$. For probabilistic predictions we also report one-vs-rest average precision (AP). For class c, $\mathrm{AP}_c = \sum_n (R_{c,n} - R_{c,n-1})P_{c,n}$ over successive score thresholds; macro AP is $K^{-1}\sum_c \mathrm{AP}_c$, while micro AP pools all held-out (u, c) pairs before computing the same quantity. Classification metrics use scikit-learn 1.9.1: `balanced_accuracy_score(adjusted=True)`, `f1_score(average="macro")`, `accuracy_score`, `matthews_corrcoef`, and `average_precision_score`.

For composition regression, negative predictions are clipped and each prediction is renormalized,

$$p_{ug} = \frac{\max(\hat{y}_{ug}, 0)}{\sum_{g'} \max(\hat{y}_{ug'}, 0)}. \tag{52}$$

The primary metric is the per-unit profile cosine

$$\cos_u = \frac{\langle p_u, t_u\rangle}{\|p_u\|_2 \|t_u\|_2}, \tag{53}$$

because the targets are nonnegative compositions and the evaluation concerns relative profile shape rather than total mass. Within each fold, the primary regression statistic is the median $\cos_u$ over held-out units. We also report per-unit Spearman correlation, $\mathrm{RMSE}_u = [G^{-1}\sum_g (p_{ug} - t_{ug})^2]^{1/2}$, and

$$R_u^2 = 1 - \frac{\sum_g (p_{ug} - t_{ug})^2}{\sum_g (t_{ug} - \bar{t}_g^{\mathrm{tr}})^2}, \tag{54}$$

where $\bar{t}^{\mathrm{tr}}$ is the training-fold mean composition. The analogous Pearson correlation, Spearman correlation, RMSE, and R^2 are also computed per gene across held-out units; genes with zero held-out variance are omitted from correlation summaries. Correlations use SciPy 1.18.1 and RMSE uses scikit-learn 1.9.1. The out-of-sample R^2 in Eq. (54) is computed directly because its denominator is centered at the training-fold mean rather than the held-out mean.

Table 12: **Measured Laminar encoding cost.** Dashes mark quantities not required for the reported throughput comparison.

Cohort	Encoded data	GPU	Wall time	Tx/s
HEST-1k	23 sections	A100 40 GB	–	23 280
HEST-1k	36 sections	L40S 45 GB	–	43 036
HEST-1k, total	59 sections; 1.75×10^9 transcripts	mixed	12.92 h	37 707
Meningioma microarray	11 slides	–	26–46 min/slide	–
Breast block	4 sections	–	15 min/section	–

Each probe is fitted from three initialization seeds. Within a fold, metric values are averaged across the three seeds. Across folds we report the median, which limits the influence of a single atypical tissue section, together with the sample standard deviation across fold estimates, using denominator $n_f - 1$ for n_f scored folds, as a descriptive measure of between-section heterogeneity. The evaluation output additionally stores the interquartile range, minimum, maximum, every per-fold value, and every secondary metric above. Fold standard deviation is not interpreted as a standard error because training sets overlap across folds. Mean rank is computed across the eight primary task columns, with ties assigned their average rank; lower mean rank is better.

G.7 Controls and baselines

Raw counts are library-size normalized to the median library size and log-transformed. PCA-64 fits 64 components on the training fold and applies the fitted transform to the held-out fold. The random-initialization control is described in Appendix G.3. A label-permuted control fits each method's own features to labels permuted within folds, with one initialization. Units whose features are not finite are dropped and counted. A method losing more than 15 % of a task's units, or 5 % for gene imputation, is not scored on that task.

G.8 Compute for encoding and scoring

Encoding is performed independently by tissue section on a single GPU. For the 59 HEST-1k cell-typing sections, Laminar encodes 1.75×10^9 transcripts in 12.92 h of aggregate section wall time (37,707 transcripts/s); the median peak GPU allocation is 11.1 GB. Probe fitting is inexpensive relative to encoding: across method–task pairs, wall time ranges from 63 s to 7,831 s. Table 12 summarizes the measured cohort-level costs.

G.9 Code, software, and data handling

Training, encoding, and probe fitting use Python 3.10, PyTorch 2.5.1 with CUDA 12.1, and PyTorch Geometric 2.6.1. Classification metrics use scikit-learn 1.9.1 (`https://scikit-learn.org/stable/`); correlation metrics use SciPy 1.18.1 (`https://docs.scipy.org/doc/scipy-1.18.1/`). Variational baselines use `scvi-tools` 1.3.3 and CellCharter uses version 0.3.7.

Public transcript tables are resolved from NCBI GEO, the EBI BioImage Archive, Zenodo, figshare, and the 10x Genomics dataset catalogue. When a repository provides both CSV and Parquet representations, the compact molecule-level table is used. Archive members are retrieved directly when possible and checked against the repository-provided byte count or archive checksum. Gene symbols are harmonized to approved HGNC symbols (Seal et al., 2022) using the HGNC alias and previous-symbol tables; ambiguous aliases map to an unknown token. Three non-biological rows (padding, mask, unknown) complete the $G = 45\,104$ lookup table. Preprocessed sections are stored as memory-mapped coordinate/gene arrays indexed by a coarse spatial grid for crop retrieval. These storage and retrieval choices do not change the model input, which remains only (p_i, g_i).

Public comparator artifacts are taken from the sources in Table 11. Nicheformer is loaded from `theislab/Nicheformer` on Hugging Face; scGPT-spatial uses Figshare record 28356068 (`best_model.pt`); TERRA uses the released Hugging Face bundles; UCE uses the four-layer checkpoint from `snap-stanford/UCE` (Figshare file 42706576); and No-

vae uses `MICS-Lab/novae-human-0`. For SToFM we use the two-module checkpoint (cell encoder and SE(2)-Transformer) linked from the official `PharMolix/SToFM` README; `get_embeddings.py` defines the released loading path for both modules. Each benchmark output stores the task definition, probe configuration, scoring configuration, and per-fold metric values used to produce the tables.

H Ablations

We evaluate the main architectural choices with a reduced configuration, **Laminar (3.2M)**. Its full training model uses a 256-dimensional field and 3,173,504 trainable parameters, compared with a 1,024-dimensional field and 115,721,216 parameters for the baseline training model in Appendix E. The reduced configuration is used only for the ablation study.

H.1 Training setup

All variants are trained from scratch on the same Atera whole-transcriptome FFPE human breast cancer section, containing 624,332,640 detected transcripts in 170,055 vendor-segmented cells. Each run uses 1,500 optimization steps, a global batch of 8 crops, and a per-crop cap of 781,000 transcripts, corresponding to approximately 8.4×10^9 delivered transcripts over training. The tissue section, crop sampler, number of optimization steps, and transcript budget are fixed across variants.

Laminar (3.2M) retains the structure of the baseline model at reduced width. It uses gene-embedding width $d_g = 32$, local width $d_h = 256$ with four 64-dimensional attention heads, tile width $d_{\text{tile}} = 128$, two neighborhood-attention blocks per hierarchy level, and one pooling block per level. The transcript readout uses

$$\mathcal{J}_{\text{out}} = \{0, 1, 2, 3\}.$$

The predictor contains two blocks of width 128, with von Mises–Fisher concentration $\kappa_0 = 640$. Optimization uses a peak learning rate of 1.65×10^{-4} after 150 warmup steps and $Q_{\text{comp}} = 64$ composition queries. Unless stated below, all remaining masking, optimization, EMA, and objective settings are those of Table 8. The ablation variants are the following:

Table 13: **Ablation variants.** Each row gives one change relative to the Laminar architecture. All settings not listed are held fixed.

Variant	Change relative to Laminar (3.2M)
No field composition loss	Set $\lambda_{\text{field}} = 0$; architecture and remaining objective terms unchanged.
Gene identity only	Replace molecular message passing by per-transcript linear layers, bypass the hierarchy in the readout, and set $\lambda_{\text{pred}} = 0$.
Molecular graph only	Retain local GATv2 message passing, bypass the hierarchy in the readout, and set $\lambda_{\text{pred}} = 0$.
Tile-level readout	Replace the transcript-specific readout by the mean state of the containing 16 μm tile; all transcripts in one tile receive the same embedding.
Flat ViT, patch level	Replace the multi-scale hierarchy by one 128 μm patch level with two blocks of global self-attention and one register token.
Frozen random gene table	Keep M_{gene} fixed at its random initialization throughout pretraining.
Genes shuffled within crops	Randomly permute gene identities among transcript coordinates independently in every training crop.

H.2 Evaluation

After training, each encoder is frozen and applied to the held-out sections using the same inference procedure as in the main transfer evaluation. Cell representations are obtained by applying the regional readout to the encoded transcripts assigned to each vendor-segmented cell. The fold construction, scorable-class criterion, probe, and metrics are otherwise identical to those in Appendix G.

For every model and evaluation fold, the downstream probe is fitted independently from three random initializations. We average the three probe scores within each fold before computing any across-fold summary. The main-paper ablation table reports the median of these fold-level scores. The

accompanying "±" is the standard deviation across folds, and therefore reflects variation across held-out tissue sections and fold composition rather than variation across probe initializations.

Cell type contains six scored classes across 59 tissue sections and is evaluated using two fixed 20-fold section partitions, giving 40 fold evaluations. Pathology region contains 22 region classes across 20 sections and is evaluated leave-one-section-out, giving 20 fold evaluations. The same folds and probe procedure are used for every variant. Each task evaluates more than 10^5 held-out cells across its folds.

For the statistical comparisons in the main ablation table, we compare each variant with Laminar (3.2M) on the paired fold-level scores using a two-sided Wilcoxon signed-rank test. Probe initializations are averaged before testing and are not treated as independent observations. We apply Benjamini–Hochberg correction across the 14 ablation–task comparisons (Benjamini & Hochberg, 1995).