# SUDOKU: UNDERSTANDING SPATIAL ORGANIZATION OF TISSUE BY LEARNING TO GENERATE

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

## ABSTRACT

Tissue function emerges from the spatial organization of cells across multiple scales. Local interactions shape the state of each cell, and their collective effects give rise to multicellular structures. Spatial transcriptomics (ST) now profiles millions of cells at known coordinates, making it possible to learn this organization from data. Most existing pre-training methods, however, focus on masked reconstruction at fixed masking rates to learn representations. Others generate cells autoregressively along an order that tissue does not have. Here we present SUDOKU, a spatial foundation model that learns tissue organization through spatial discrete denoising. SUDOKU represents cell states with a hierarchical vocabulary, and learns the distribution of hidden cell states given spatial context across a full masking range. This conditional formulation captures dependencies from local cellular interactions to coherent multicellular configurations without imposing a predefined order, and naturally unifies representation and generation as two inference regimes under full and partial observation. Pre-trained on a rigorously quality-controlled corpus of 49 million cells from 12 organs across human and mouse, SUDOKU outperforms existing pre-trained models and specialized methods on niche retrieval, multi-scale clustering, spatial generation and immune-recruitment prediction, taking a step toward a virtual tissue model built from ST data.

## 1 INTRODUCTION

Tissue function emerges from the spatial organization of cells. At the local level, the state of a cell is shaped by the identities, states, and positions of surrounding cells through contact and secreted signals (Fischer et al., 2023; Cang et al., 2023). Beyond the cellular level, spatially neighboring cells organize into higher-order tissue structures such as layers, boundaries, and functional niches (Schürch et al., 2020; Bhate et al., 2022). Tissue organization therefore reflects a hierarchy of spatial dependencies: local interactions constrain individual cell states, while their coordinated effects give rise to collective multicellular organization. Such structures are not simple aggregates of cells, but emerge from coordinated spatial relationships among them. A model of tissue must therefore capture both local cellular dependencies and the higher-order organization arising from their collective arrangement.

The rapidly growing scale of spatial transcriptomics (ST) makes it increasingly possible to learn tissue organization directly from data. Most spatial pre-training methods, however, have focused on representation learning. They reconstruct masked gene expression or cell states at one or a few predefined masking levels (Wen et al., 2024; Tejada-Lapuerta et al., 2025; Wang et al., 2025), regress latent embeddings (Zhao et al., 2025; Liu et al., 2026; Birk et al., 2026), or use contrastive objectives (Blampey et al., 2025) to learn representations of cells and neighborhoods. While representation learning serves important downstream analyses, such objectives expose the model to only one or a limited set of views of spatial dependency. Yet tissue organization is shaped by complex dependencies spanning individual cells, local neighborhoods, and collective multicellular configurations. Learning these dependencies from only a few predefined settings therefore risks capturing particular spatial relationships rather than the broader organizing principles of tissue.

Generative modeling offers a complementary perspective by asking which cellular configurations are compatible with a given spatial context. GeST (Hao et al., 2025) models this autoregressively by



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.

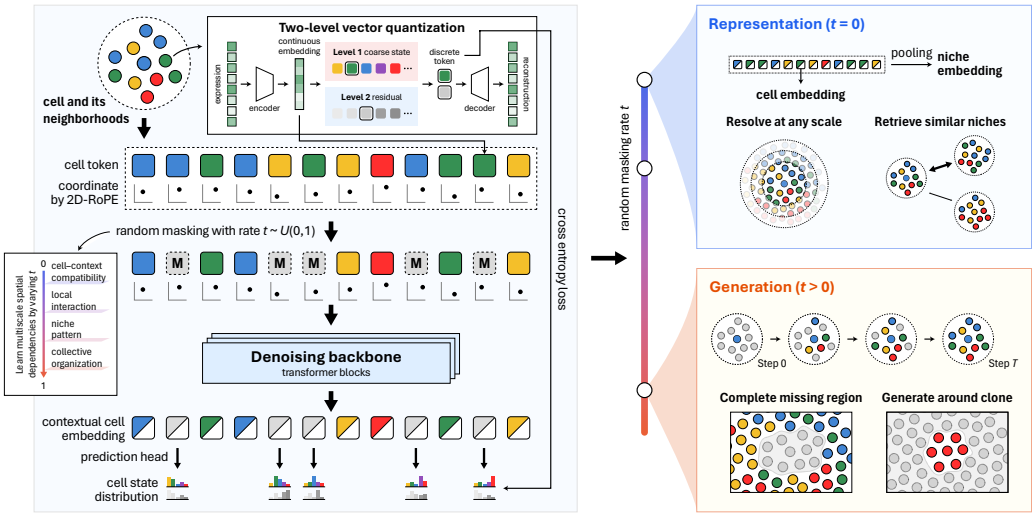



Figure 1: Spatial discrete denoising. A coordinate-aware spatial transformer predicts masked cellular states from visible spatial context across variable masking rates, learning dependencies from local cell–context relationships to multicellular organization. The same model provides contextual representations under full observation and generates missing cellular configurations under partial observation.

serializing cells into a one-dimensional sequence. However, a tissue has no natural traversal order, and an imposed one lets a cell condition only on the cells listed before it.

We argue that tissue organization is best learned by varying the observed spatial context. For a set of spatially neighboring cells with cell states $X$ at coordinates $P$, a general tissue model should learn

$$p(X_M \mid X_{\bar{M}}, P), \tag{1}$$

for arbitrary hidden subsets $M$. With only a few cells hidden, the task primarily probes local cell–context dependencies; as more cells are hidden, it increasingly requires recovering coherent multicellular configurations. Varying the amount of observed context therefore exposes the model to a continuum from local cellular constraints to collective tissue organization. A model trained on this conditional then needs no separate objective for representation downstream tasks, as effective generative pre-training has been shown to yield transferable representations in language (Radford et al., 2018; Sahoo et al., 2024) and vision (Chen et al., 2020; Li et al., 2023a; Gabeur et al., 2026).

Motivated by this view, we propose **SUDOKU**[1], a spatial foundation model trained with **spatial discrete denoising** (Figure 1). A residual-quantized tokenizer maps whole-cell transcriptomes into a hierarchical vocabulary of discrete cell states, and a coordinate-aware spatial transformer is trained to recover masked cells from their visible context. By sampling the masking rate across nearly the full interval from zero to one, SUDOKU learns spatial dependencies under continuously varying amounts of observed context, from local cell–context relationships to coherent multicellular organization, without imposing an artificial order on cells. Our main contributions are:

(1) We formulate spatial tissue pre-training as learning conditional cellular distributions under arbitrary partial observation, providing a unified view of local cellular dependence and collective tissue organization.

(2) We introduce spatial discrete denoising as a new pre-training objective for ST. This approach combines a hierarchical cell-state vocabulary with a bidirectional, order-free spatial transformer trained across the full masking-rate spectrum. The pre-trained backbone serves as an encoder when the neighborhood is fully observed and performs generation when there are missing regions.

(3) We pre-train SUDOKU on a high-quality corpus of 49 million cells from 12 organs across human and mouse, and show broad transfer across both representation and generation, outperforming exist-

[1]Named after the puzzle, in which each blank cell is filled with one of a fixed set of symbols, constrained by the filled cells around it.

ing methods on niche retrieval, multi-scale clustering, spatial generation, and immune-recruitment prediction.

## 2 Background and related work

### 2.1 Key properties of ST data and challenges of data modeling

ST profiles each tissue section as a gene expression matrix, where each row is a cell and each column one of about twenty thousand genes, along with an additional two-dimensional physical coordinate for every cell. Here we show three key properties of such data, which bring challenges and call for specialized design of the training objective and model architecture.

**Expression is high-dimensional and noisy, and cell states are hierarchical.** Gene-expression profiles are high-dimensional, sparse, and noisy: most entries of the expression matrix are zero (Svensson, 2020; Gong et al., 2023), while observed counts also reflect technical artifacts (Hafemeister & Satija, 2019; Hartman & Satija, 2024). A neighborhood of hundreds of cells amounts to millions of measurements, too long to model as gene sequences. Meanwhile, cell states are hierarchically organized, from broad cell types to finer subtypes and functional states (Zeng, 2022; Domcke & Shendure, 2023; Bian et al., 2025). Predicting cell states as a flat set of unrelated categories thus incorrectly penalizes a closely related state as heavily as an unrelated one.

**Spatial dependence is both local and collective.** Cell states are shaped by the identities and positions of surrounding cells through contact and secreted signals (Fischer et al., 2023; Cang et al., 2023). These local relationships collectively organize neighboring cells into recurrent multicellular structures that define higher-order tissue organization (Schürch et al., 2020; Bhate et al., 2022). A model trained to predict limited missing cells may learn how each cell depends on its surroundings, but can hardly understand how cells are jointly arranged into these structures.

**Tissues are not naturally one-dimensional.** A tissue section is a set of cells embedded in physical space. Each cell can be jointly influenced by neighbors from all directions (Fischer et al., 2023; Cang et al., 2023; Li et al., 2023b; Shao et al., 2022), and their relative spatial position does not change with how cells are listed. Therefore, serializing tissue into one-dimensional cell sequences for autoregressive generation imposes an artificial dependency structure, in which each cell is conditioned only on the cells placed before it, although the neighbors after it influence it just as much.

Together, these properties require a tissue model to represent cellular states compactly, capture spatial dependencies from local cell–context relationships to collective organization, and preserve the native two-dimensional structure of tissue.

### 2.2 Existing pre-training paradigms for ST

**Representation learning.** Most existing methods are designed to learn transferable representations of cells or their spatial context. Within this paradigm, several types of pre-training objectives have been explored.

*Masked reconstruction in expression space* predicts the expression of masked genes or cells from the remaining context. CellPLM (Wen et al., 2024), Nicheformer (Tejada-Lapuerta et al., 2025) and scGPT-spatial (Wang et al., 2025) adopt this strategy with one or a few fixed masking rates. While CellPLM and scGPT-spatial are able to impute missing gene expression profiles, their capability is limited to imputation of partly observed cells rather than generating completely missing cell states.

*Masked reconstruction in latent space* moves the reconstruction target to a cell embedding. SToFM (Zhao et al., 2025) regresses the embeddings of masked cells from a single-cell foundation model, and CellWorld (Liu et al., 2026) and TERRA (Birk et al., 2026) predict the latent from a teacher encoder. While this keeps prediction in a compact latent space with less noise, the output is still a point estimate that cannot be sampled.

*Contrastive learning* avoids explicit reconstruction. Novae learns neighborhood representations through graph contrastive learning and prototype assignment (Blampey et al., 2025). These meth-

Table 1: Pre-training objectives of spatial transcriptomics models. Mask rate: masking rate during training; Generative: whether the output is a distribution over cell states that can be sampled; Order-free: whether generation imposes no order on the cells; Multi-scale: whether the pre-trained model gives representations at more than one spatial scale.

| Method | Token | Pre-training task | Mask rate | Generative | Order-free | Multi-scale |
|---|---|---|---|---|---|---|
| CellPLM | cell | masked expression regression | 0.75 cells, 0.25 genes | ✗ | – | ✗ |
| Nicheformer | gene | masked gene prediction | 0.15 | ✗ | – | ✗ |
| scGPT-spatial | gene | masked expression regression | 0.25 / 0.5 / 0.75 | ✗ | – | ✗ |
| SToFM | cell | masked embedding regression | 0.10 | ✗ | – | ✓ |
| CellWorld | cell | masked latent regression | 0.6 | ✗ | – | ✗ |
| TERRA | gene | masked latent regression | 0.6 | ✗ | – | ✗ |
| Novae | – | contrastive learning | – | ✗ | – | ✓ |
| GeST | cell | autoregressive generation | – | ✓ | ✗ | ✗ |
| SUDOKU | cell | discrete denoising | $U(0,1)$ | ✓ | ✓ | ✓ |

ods are common for representation learning, but do not explicitly model the distribution of cellular configurations conditioned on spatial context.

**Generative modeling.** Generative models instead aim to learn a distribution over cellular states or configurations. GeST (Hao et al., 2025) represents cells with discrete meta cells obtained from pre-clustering and generates them autoregressively. To apply autoregression, however, cells must first be serialized into a one-dimensional sequence with an imposed traversal order.

Table 1 summarizes these spatial pre-training paradigms and their differences in prediction target, masking strategy, generative capability, and dependence on cell ordering.

### 2.3 MASKED DISCRETE GENERATIVE MODELING

Before spatial transcriptomics, masked discrete modeling had proved effective as an alternative generative paradigm to autoregressive factorization in other data modalities, including vision and natural language. D3PM introduced diffusion processes in discrete state spaces (Austin et al., 2021), while MaskGIT showed that masked tokens can be generated iteratively by repeatedly predicting and committing high-confidence states (Chang et al., 2022). More recently, MDLM and RADD established that variable-rate masked prediction corresponds to the variational objective of an absorbing-state discrete diffusion process (Sahoo et al., 2024; Ou et al., 2025), and LLaDA demonstrated that this formulation can scale to large language models (Nie et al., 2025).

SUDOKU builds on this line of work but differs in several aspects that follow from the properties of ST data in Section 2.1: what a token is, how tokens are placed in space, and how the visible context enters the model.

## 3 METHOD

A tissue section is a set $S = \{(\mathbf{p}_i, \mathbf{x}_i)\}_{i=1}^{n}$, where $\mathbf{p}_i \in \mathbb{R}^2$ is the physical coordinate of cell $i$ on the section and $\mathbf{x}_i \in \mathbb{R}^G$ is its expression vector over $G$ genes. The neighborhood $N_i$ of cell $i$ is the set of cells within radius $r$ of it. The task is to infer the states of the cells hidden in a neighborhood from the states of the visible ones and the coordinates of all of them.

SUDOKU consists of two components: a hierarchical tokenizer that maps continuous cellular profiles into discrete cell states, and a coordinate-aware spatial transformer trained by spatial discrete denoising. Together, they define this conditional distribution.

### 3.1 HIERARCHICAL CELL-STATE TOKENIZER

Discrete denoising requires a compact state space over which masked cellular states can be predicted probabilistically. We therefore encode each cell into a discrete representation of its transcriptional state, reducing a high-dimensional expression profile to a small number of cell-state codes while preserving relationships among related states. This also makes a neighborhood of hundreds of cells tractable as a collection of cellular states rather than millions of gene-level measurements.

Because cellular states are hierarchically organized, we use residual quantization rather than a flat codebook. The first-level code captures a coarse state and the second-level code refines the residual variation within it. This structure makes nearby cellular states partially share their discrete representation rather than treating all state mismatches equally.

The tokenizer maps the expression $\mathbf{x}_i \in \mathbb{R}^G$ of a cell over $G$ genes to a two-level residual-quantized code. The encoder $f_\varphi$ is an MLP over normalized log expression that outputs a 128-dimensional latent $\mathbf{z}_i = f_\varphi(\mathbf{x}_i)$. Two codebooks $\mathbf{e}^{(1)}, \mathbf{e}^{(2)}$ each hold $K = 256$ vectors of dimension 128, making the vocabulary size $K^2 = 65{,}536$. The first level takes the code nearest to $\mathbf{z}_i$, the second level the code nearest to the residual, and the quantized latent is their sum:

$$c_i^{(1)} = \mathrm{NN}\big(\mathbf{z}_i; \mathbf{e}^{(1)}\big), \qquad c_i^{(2)} = \mathrm{NN}\big(\mathbf{z}_i - \mathbf{e}^{(1)}_{c_i^{(1)}}; \mathbf{e}^{(2)}\big), \qquad \mathbf{z}_{q,i} = \mathbf{e}^{(1)}_{c_i^{(1)}} + \mathbf{e}^{(2)}_{c_i^{(2)}}, \tag{2}$$

where $\mathrm{NN}(\cdot; \mathbf{e})$ returns the index of the nearest code. A decoder $g_\varphi$ maps $\mathbf{z}_{q,i}$ to the mean and dispersion of a negative-binomial distribution per gene (Lopez et al., 2018), $(\boldsymbol{\mu}_i, \mathbf{r}_i) = g_\varphi(\mathbf{z}_{q,i})$, with the mean scaled by the total count $s_i$ of the cell:

$$\mathrm{NB}\big(x_{ig} \mid s_i\mu_{ig}, r_{ig}\big) = \frac{\Gamma(x_{ig} + r_{ig})}{\Gamma(r_{ig})\,\Gamma(x_{ig}+1)} \Big(\frac{r_{ig}}{r_{ig} + s_i\mu_{ig}}\Big)^{r_{ig}} \Big(\frac{s_i\mu_{ig}}{r_{ig} + s_i\mu_{ig}}\Big)^{x_{ig}}. \tag{3}$$

The training loss is the negative log-likelihood summed over genes plus a commitment term with $\beta = 1$,

$$\mathcal{L}_{\mathrm{tok}}(\varphi) = -\sum_{g=1}^{G} \log \mathrm{NB}\big(x_{ig} \mid s_i\mu_{ig}, r_{ig}\big) + \beta \,\big\|\mathbf{z}_i - \mathrm{sg}(\mathbf{z}_{q,i})\big\|_2^2, \tag{4}$$

where sg is the stop-gradient. Codebooks track the exponential moving average of the latents assigned to each code. The joint code is written $c_i = (c_i^{(1)}, c_i^{(2)})$. The distance used at each quantization level, network widths, moving-average decay and batch composition are in Appendix A.

### 3.2 Spatial discrete denoising

We now train the spatial transformer to recover arbitrary subsets of cellular states from the remaining spatial context. By varying the masking rate across neighborhoods, the model is exposed to conditional inference problems ranging from local cell-state recovery to reconstruction of largely missing multicellular configurations.

**Forward process.** The forward process corrupts a neighborhood by masking part of its cells, and the spatial transformer is trained to reverse it. For each neighborhood $N_i$ a masking rate $t \sim U(\epsilon, 1)$ is drawn independently, with $\epsilon = 0.001$. Each cell is masked independently with probability $t$, and both levels of a masked cell's code are replaced by the mask symbol [M]:

$$m_j \sim \mathrm{Bernoulli}(t), \qquad M = \{\, j \in N_i : m_j = 1 \,\}, \qquad V = N_i \setminus M. \tag{5}$$

**Input.** The input to the spatial transformer consists of gene expression information and spatial information. For gene expression, visible cells are fed the linear projection $\mathbf{W}\mathbf{z}_k$ of their continuous tokenizer latent and masked cells a learned mask vector $\mathbf{u}$, and together they form the layer-0 input $\mathbf{h}_j^0$. For spatial information, coordinates are first divided by the measured median nearest-neighbor spacing $d_s$ of section $s$, $\hat{\mathbf{p}}_j = \mathbf{p}_j / d_s$, to align physical scale across platforms. Then $\hat{\mathbf{p}}$ enters every attention layer through an axial rotary positional encoding, in which the head dimension $D$ is split in half, the first half rotated by the $x$ component of $\hat{\mathbf{p}}$ and the second by the $y$ component, with frequency $\omega_m = \theta^{-2m/(D/2)}$ for the $m$-th pair of dimensions and $\theta = 10$. Writing $\mathbf{R}(\hat{\mathbf{p}}) = \mathrm{diag}(\mathbf{R}_x(\hat{p}_x), \mathbf{R}_y(\hat{p}_y))$ for the rotation and $\mathbf{q}_j, \mathbf{k}_k$ for the query of cell $j$ and the key of cell $k$ at a layer, the attention score is

$$a_{jk} = \big(\mathbf{R}(\hat{\mathbf{p}}_j)\mathbf{q}_j\big)^\top \big(\mathbf{R}(\hat{\mathbf{p}}_k)\mathbf{k}_k\big) = \mathbf{q}_j^\top \mathbf{R}(\hat{\mathbf{p}}_k - \hat{\mathbf{p}}_j)\,\mathbf{k}_k, \tag{6}$$

so the score depends only on the relative displacement of the two cells. Theoretically, this way of positional encoding is invariant to translation but not to rotation. But in practice, we show that rotating the coordinates has almost no impact on the model (Appendix B). Continuous input with discrete targets follows BEiT v2 (Peng et al., 2022). This keeps more information of the visible cells and asks the model to learn the conditional cell state distribution at missing positions.

**Backbone and output.** The backbone is a transformer with 10 layers, width 256 and 8 heads, and the final hidden state of cell $j$ is $\mathbf{h}_j$. Two classification heads give the $K$-way distributions of the first- and second-level codes, $p_\theta(c_j^{(\ell)} \mid \mathbf{z}_V, \mathbf{p}_N)$ for $\ell = 1, 2$. Within a single forward pass, the code levels and masked cells are factorized conditionally on the visible context.

**Loss.** The cross-entropy of both code levels on the masked cells, weighted by $1/t$:

$$\mathcal{L}(\theta) = \mathbb{E}_t \, \mathbb{E}_{M \sim \mathrm{Bern}(t)} \, \frac{1}{t} \sum_{j \in M} \sum_{\ell=1}^{2} - \log p_\theta\big(c_j^{(\ell)} \,\big|\, \mathbf{z}_V, \mathbf{p}_N\big). \tag{7}$$

With the $1/t$ weighting, Equation 7 is the variational bound of absorbing-state discrete diffusion on the codes of a neighborhood (Sahoo et al., 2024; Ou et al., 2025). Minimizing it therefore trains a generative model of the whole neighborhood. Note that visible cells enter through their continuous latents $\mathbf{z}_V$ rather than through their codes, different from the canonical MDLM objective (Sahoo et al., 2024). We show in Appendix C that the bound is unchanged, since the latent determines the code, thus the generative modeling property is preserved. The $1/t$ weighting balances contributions from different masking levels, so the model learns conditional distributions across nearly the full range of observed context, from almost complete neighborhoods to largely missing configurations. The objective factorizes over the masked subset and needs no traversal order.

**Pre-training corpus.** We curate a high-quality pre-training corpus of 396 single-cell resolution whole-transcriptome tissue sections, 161 human and 235 mouse, from 12 organs and from both normal and tumor tissue, 49.1 million cells in all, filtered by per-cell and per-section quality control. The corpus composition, quality control and hyperparameters are in Appendix A.

### 3.3 Inference under partial and full observation

**Generation.** Starting from the missing set fully masked, unmask in $T$ parallel steps. At step $s$ let $M_s$ be the set still masked, with $M_1 = M$. One forward pass gives the two-level distribution of every $j \in M_s$. Each level's code is drawn from the tempered distribution by the Gumbel-max trick, and the confidence of a cell is the probability of its drawn codes,

$$\hat{c}_j^{(\ell)} = \arg\max_k \Big[ \tfrac{1}{\tau} \log p_\theta\big(c_j^{(\ell)} = k \,\big|\, \cdot\big) + g_{jk} \Big], \qquad \sigma_j = \sum_{\ell=1}^{2} \log p_\theta\big(\hat{c}_j^{(\ell)} \,\big|\, \cdot\big), \tag{8}$$

with $g_{jk} \sim \mathrm{Gumbel}(0, 1)$ drawn independently. The $k_s$ cells with the highest $\sigma_j$ are committed with their drawn codes. At the next step they enter the conditioning through their codebook reconstruction $\mathbf{z}_q(\hat{c}_j)$, and the rest remain in $M_{s+1}$. The commit count is near-uniform with the remainder placed in the last steps, $k_s = \lfloor |M|/T \rfloor$ plus one when $s > T - (|M| \bmod T)$, so that the hardest cells are decided with the most context. $\tau$ is a temperature. At $\tau = 1$ every committed code is a sample from $p_\theta$, and as $\tau \to 0$ the draw becomes the argmax and decoding is greedy. Within a step the committed cells are drawn independently given the current context, while dependence between generated cells is carried across steps, since every committed cell conditions all later draws, so the $T$ steps carry dependence across generated cells and approximate a sample of the joint distribution over $M$.

**Representation.** When performing representation tasks, the neighborhood is fully visible. The final hidden state $\mathbf{h}_i$ of the center cell $i$ is its representation in context, and the neighborhood representation is the mean of hidden states within it,

$$\bar{\mathbf{h}}_i = \frac{1}{|N_i|} \sum_{k \in N_i} \mathbf{h}_k. \tag{9}$$

The neighborhood radius $r$ controls the spatial scale of the readout and can be varied to obtain representations at multiple levels of tissue organization.

## 4 Experiments

We evaluate SUDOKU on five benchmarks that cover the two regimes of Section 3.3. Under full observation, we evaluate SUDOKU's representation quality with niche retrieval on the human embryo atlas HESTA (Pan et al., 2026) and the human cortex dataset DLPFC (Maynard et al., 2021),

and multi-scale clustering on MERFISH mouse brain data (Cohen et al., 2025). Under partial observation, we test SUDOKU's spatial generation on in-house sections of human and mouse brain and mouse kidney. Furthermore, we test SUDOKU's generative ability in modeling tumor microenvironment and predicting immune recruitment of tumor clones in SPAC-seq, a spatial CRISPR screen of mouse lung metastases (Zhang et al., 2026) (Section 4.3). All benchmark sections are held out of pre-training. More details of datasets and evaluation protocols are in Appendix D and Appendix E.

### 4.1 REPRESENTING TISSUE ORGANIZATION ACROSS SECTIONS AND SCALES

**Niche retrieval.** We first test whether SUDOKU's representations recognize the same niche, a spatially localized group of cells with certain features, across independent tissue sections. Given a niche of interest (NOI) in a query section, the task is to retrieve the matching niches from other sections by the similarity of their representations, an operation behind atlas query and label transfer. HESTA provides organ-level NOIs from human embryo Stereo-seq sections at single-cell resolution, scored by cross-section retrieval AUPRC. DLPFC provides finer NOIs at cortical-layer junctions from 12 human cortex Visium sections at spot resolution, a platform absent from pre-training, scored by PCC against a matching score computed from the layer labels. The task, the metrics and intuitive visualizations are in Appendix D.

On HESTA, SUDOKU achieves the highest AUPRC on seven of the eight organs, with GeST leading only on jaw and teeth. Notably, SUDOKU outperforms the specialized niche-retrieval method QueST (Chen et al., 2024) on all eight organs (Table 2). On DLPFC, SUDOKU achieves the highest PCC on five of the seven niche classes, ties QueST on one, and is second only to QueST on the remaining class. These results show that the representations of SUDOKU support niche retrieval at both the organ and the layer-junction level, on single-cell data and on spot-level data from a platform unseen in pre-training.

Table 2: Niche retrieval on HESTA, AUPRC per organ, and DLPFC, PCC per NOI type (the junction of two or three cortical layers including L3, L4, L5, L6 and White Matter). Rows are grouped into task-specific methods, spatial pre-trained models and ours. **Bold**: best; underline: second best.

| | HESTA, AUPRC per organ | | | | | | | | DLPFC, PCC per neighborhood class | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Method | Brain | Liver | Spinal | Cartilage | Ganglion | Muscle | Heart | Jaw | L3L4 | L4L5 | L5L6 | L6WM | L3L4L5 | L4L5L6 | L5L6WM |
| QueST | 0.490 | 0.797 | 0.430 | 0.352 | 0.147 | 0.277 | 0.710 | 0.126 | **0.566** | 0.673 | 0.762 | 0.746 | **0.628** | 0.720 | 0.696 |
| CellPLM | 0.323 | 0.731 | 0.303 | 0.278 | 0.129 | 0.276 | 0.471 | 0.148 | 0.344 | 0.285 | 0.238 | 0.015 | 0.424 | 0.269 | 0.126 |
| Novae | 0.423 | 0.333 | 0.396 | 0.238 | 0.073 | 0.147 | 0.101 | 0.162 | 0.508 | 0.603 | 0.511 | 0.620 | 0.580 | 0.557 | 0.479 |
| Nicheformer | 0.334 | 0.589 | 0.350 | 0.268 | 0.121 | 0.166 | 0.208 | 0.136 | 0.265 | 0.240 | 0.120 | 0.009 | 0.352 | 0.213 | −0.005 |
| scGPT-spatial | 0.328 | 0.624 | 0.302 | 0.250 | 0.111 | 0.270 | 0.391 | 0.149 | 0.312 | 0.212 | 0.123 | −0.005 | 0.368 | 0.191 | 0.013 |
| SToFM | 0.228 | 0.467 | 0.234 | 0.214 | 0.075 | 0.121 | 0.063 | 0.098 | 0.073 | 0.128 | 0.135 | 0.123 | 0.120 | 0.124 | 0.112 |
| CellWorld | 0.224 | 0.574 | 0.207 | 0.159 | 0.082 | 0.135 | 0.081 | 0.086 | 0.261 | 0.325 | 0.222 | −0.094 | 0.382 | 0.305 | 0.059 |
| TERRA | 0.387 | 0.532 | 0.381 | 0.235 | 0.154 | 0.186 | 0.190 | 0.155 | 0.326 | 0.282 | 0.214 | −0.030 | 0.409 | 0.261 | 0.096 |
| GeST | 0.446 | 0.847 | 0.395 | 0.315 | 0.251 | 0.410 | 0.657 | **0.253** | 0.465 | 0.629 | 0.680 | 0.626 | 0.539 | 0.646 | 0.606 |
| SUDOKU | **0.517** | **0.932** | **0.503** | **0.359** | **0.264** | **0.427** | **0.758** | 0.207 | **0.566** | **0.698** | **0.765** | **0.751** | 0.604 | **0.723** | **0.705** |

**Multi-scale clustering.** We next ask whether the same backbone resolves tissue at multiple levels via representations at different spatial scales. We use six whole-transcriptome MERFISH sections of mouse brain annotated at three levels: cell type, cortical laminar band and Allen brain division (Wang et al., 2020). For SUDOKU, cell-type clustering uses the contextual cell representation within 50 $\mu$m, while laminar and regional clustering use pooled neighborhood representations at 50 and 100 $\mu$m, respectively. For the other methods, we read out one embedding per cell and pool it over the same radii. Clustering performance is evaluated by ARI and NMI.

Table 3: Multi-scale clustering on mouse brain, median ARI and NMI at three levels: Region (Allen brain division), Band (laminar band) and Cell type. Rows: task-specific methods, pretrained models, ours. **Bold**: best; underline: second best.

| | Region | | Band | | Cell type | |
|---|---|---|---|---|---|---|
| Method | ARI | NMI | ARI | NMI | ARI | NMI |
| STAGATE | 0.312 | 0.469 | 0.420 | **0.634** | 0.154 | 0.383 |
| BANKSY | 0.111 | 0.190 | 0.091 | 0.186 | 0.440 | 0.630 |
| CellPLM | 0.113 | 0.170 | 0.115 | 0.140 | 0.438 | 0.671 |
| Novae | 0.264 | 0.440 | 0.334 | 0.535 | 0.121 | 0.345 |
| Nicheformer | 0.246 | 0.294 | 0.144 | 0.180 | 0.345 | 0.567 |
| scGPT-spatial | 0.177 | 0.235 | 0.223 | 0.253 | 0.364 | 0.605 |
| SToFM | 0.176 | 0.237 | 0.229 | 0.357 | 0.283 | 0.494 |
| CellWorld | 0.232 | 0.309 | 0.026 | 0.043 | 0.267 | 0.435 |
| TERRA | 0.207 | 0.304 | 0.212 | 0.355 | 0.154 | 0.390 |
| SUDOKU | **0.342** | **0.491** | **0.493** | 0.629 | **0.520** | **0.776** |

SUDOKU achieves the highest ARI at all three organizational levels and the highest NMI at the cell-type and regional levels, and nearly ties with STAGATE (Dong & Zhang, 2022) at the laminar level (Table 3). Existing methods mostly capture part of the hierarchy: STAGATE and Novae score well on region and band but poorly on cell type, whereas BANKSY (Singhal et al., 2024) and CellPLM show the opposite pattern. These results show that a single pre-trained backbone can support representations from local cell identity to higher-order tissue organization by simply changing the spatial scale of the readout.

### 4.2 SPATIAL GENERATION OF MISSING CELLULAR CONFIGURATIONS

We next evaluate the generative capability of SUDOKU by asking it to recover missing cell states from partially observed neighborhoods. From 31 human and mouse, brain and kidney sections, we construct 972 questions, each masking part of a 500-cell window in one of four missing geometries (scattered points, interior block, shell, multiple blocks) at one of three missing fractions (0.2, 0.4, 0.6). Questions are further split into easy, medium and hard tertiles by the cell-type composition complexity of the window. Generated cells are mapped to cell types through a code-to-type table built outside the window, and the metric is the cell-type accuracy of the generated cells. Besides GeST, the only existing spatial generative model, we compare with baselines that assign each masked cell the majority type of its $k$ nearest visible cells ($k$-NN, $k = 1, 5$). Details on benchmark construction, code-to-type mapping and intuitive visualization of the questions are in Appendix D.

Though accuracy falls with composition difficulty for every method, SUDOKU achieves the highest accuracy in every geometry-by-difficulty setting and every tissue group (Table 4). GeST stays below the $k$-NN baselines in many settings. These results show that SUDOKU recovers missing cellular configurations across organs, species and geometries.

Table 4: Spatial generation, cell-type accuracy of the generated cells on the 972 questions, split by missing geometry and difficulty (left) and by tissue group (right). AD: Alzheimer's disease; Ctrl: control; M1, M2: two mice; Inj.: injured. **Bold**: best; underline: second best.

| | By missing geometry and difficulty | | | | | | | | | | | | By tissue group | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Scattered | | | Interior block | | | Shell | | | Multiple blocks | | | Human brain | | Mouse brain | | Mouse kidney | |
| Method | Easy | Med. | Hard | Easy | Med. | Hard | Easy | Med. | Hard | Easy | Med. | Hard | AD | Ctrl | M1 | M2 | Inj. | Ctrl |
| 1-NN | 0.450 | 0.339 | 0.215 | 0.433 | 0.307 | 0.198 | 0.434 | 0.323 | 0.202 | 0.448 | 0.315 | 0.199 | 0.439 | 0.372 | 0.249 | 0.297 | 0.269 | 0.382 |
| 5-NN | 0.474 | 0.346 | 0.251 | 0.502 | 0.368 | 0.232 | 0.488 | 0.324 | 0.253 | 0.504 | 0.330 | 0.246 | 0.514 | 0.457 | 0.266 | 0.314 | 0.296 | 0.377 |
| GeST | 0.503 | 0.321 | 0.249 | 0.476 | 0.326 | 0.261 | 0.486 | 0.282 | 0.249 | 0.483 | 0.296 | 0.252 | 0.485 | 0.444 | 0.228 | 0.294 | 0.317 | 0.396 |
| SUDOKU | **0.508** | **0.378** | **0.267** | **0.534** | **0.390** | **0.263** | **0.498** | **0.376** | **0.282** | **0.521** | **0.373** | **0.278** | **0.563** | **0.469** | **0.290** | **0.344** | **0.320** | **0.406** |

### 4.3 PREDICTING IMMUNE RECRUITMENT AFTER SPATIAL PERTURBATION

Finally, we ask whether the learned conditional distribution predicts the immune recruitment around perturbed tumor clones from the state of the clone alone. We focus on immune recruitment around perturbed tumor clones: whether a tumor draws immune cells into its surroundings separates hot tumors from cold ones and shapes the response to immunotherapy (Galon & Bruni, 2019; Tumeh et al., 2014). SPAC-seq profiles mouse lung metastases whose tumor clones each carry a CRISPR perturbation. For each of 125 clones, we mask cells around it and ask the model to generate them, without telling the model which perturbation the clone carries. The immune-cell fraction among the generated cells is the predicted recruitment of the clone. We report the Spearman correlation between predicted and measured recruitment across the clones of a section, one value for each of the three sections, and the rank error of each clone, the absolute difference between its predicted and measured rank divided by the number of clones in the section, one value per clone. GeST and a simple expression-interpolation baseline are used for comparison. Details of benchmark construction, baseline settings and metrics are in Appendix D.

SUDOKU reaches a Spearman correlation of 0.77 averaged over the three sections, compared with 0.49 for GeST and 0.11 for expression interpolation (Figure 2a). It also ranks the clones by immune recruitment more accurately, with a median rank error of 0.12, compared with 0.20 for GeST and 0.28 for interpolation (Figure 2b). These results show that SUDOKU recovers immune recruitment variation across perturbed clones from the clone state. As a case study, Figure 2c shows a Cd274

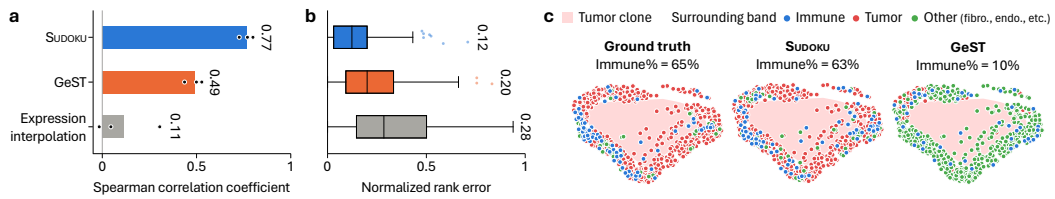



Figure 2: Immune recruitment in SPAC-seq. (a) Spearman correlation between predicted and measured recruitment, mean over the three sections, one dot per section. (b) Rank error of each clone; boxes span the quartiles, whiskers 1.5 times the interquartile range, medians written beside the boxes. (c) A Cd274 (PD-L1) knockout clone (grey) and its 50-$\mu$m band, colored by measured type and by the types generated by SUDOKU and GeST, with the immune fraction of the band above each.

knockout clone. Cd274 encodes PD-L1, through which tumor cells switch off PD-1-bearing T cells, and the PD-1/PD-L1 axis is the target of checkpoint blockade immunotherapy (Tumeh et al., 2014; Galon & Bruni, 2019), so a clone that has lost it is expected to be heavily infiltrated. The measured immune fraction in its band is 0.65. SUDOKU generates 0.63 with a spatial pattern close to the measured one, whereas GeST generates 0.10 and collapses much of the band onto one non-immune type.

### 4.4 ABLATIONS

We ablate SUDOKU along three axes (Table 5), all variants sharing the backbone, corpus and schedule: the two-level vocabulary is replaced by one level with 256 or 65,536 codes; the generative objective is replaced by two predictive controls, regressing the latent of a masked cell instead of predicting a distribution over its codes, or fixing the masking rate at 0.15; and visible cells enter as their codes instead of their continuous latents.

Table 5: Ablations on niche retrieval (HESTA AUPRC, mean over all 84,977 scoreable pairs) and spatial generation (type accuracy, mean over all 972 questions); each variant changes only the item named.

| Variant | Representation | Generation |
|---|---|---|
| w/o residual quantization | | |
| one level, 256 codes | 0.467 | 0.305 |
| one level, 65,536 codes | 0.441 | 0.247 |
| w/o generative objective | | |
| latent regression | 0.429 | 0.368 |
| fixed masking rate 0.15 | 0.489 | 0.379 |
| w/o continuous visible input | | |
| visible cells as codes | 0.401 | 0.323 |
| SUDOKU | **0.503** | **0.405** |

Residual quantization matters most for generation, where one level of either size loses up to 0.16 accuracy. Regressing the latent instead of predicting a distribution hurts both representation and generation, and a fixed masking rate mainly hurts generation. Feeding visible cells as codes costs 0.10 niche retrieval AUPRC and 0.08 cell type generation accuracy. The hierarchical vocabulary, the generative objective over the whole masking-rate axis, and the continuous visible input thus each carry part of the performance of SUDOKU.

## 5 DISCUSSION

We present SUDOKU, a spatial foundation model that learns tissue organization through spatial discrete denoising. Beyond a new pre-training objective, the results support the view of the introduction: tissue organization can be learned as conditional cellular distributions under arbitrary partial observation. Varying how much of a neighborhood is hidden exposes a continuum of spatial dependencies, from local cellular constraints to collective multicellular organization, and representation and generation arise as two regimes of this one conditional model rather than as separate objectives. On representation, generation and perturbation-response benchmarks, SUDOKU outperforms existing pre-trained models, specialized methods and autoregressive generation.

SUDOKU still has several limitations. The tokenizer is an MLP over the whole transcriptome and therefore requires an input of fixed dimension. A design that accepts partial gene panels should be explored. Applying SUDOKU to new sections still requires self-supervised fine-tuning, and a larger pre-training corpus covering more platforms and spatial resolutions may improve its zero-shot generalization. Finally, the model is limited to static transcriptomic profiles. Extending SUDOKU to tissue dynamics and to other modalities such as histology images is an important future direction.

### REPRODUCIBILITY STATEMENT

The model code of SUDOKU, covering the tokenizer, the spatial transformer, the forward process, the loss, the sampler and the representation readout, together with the configuration files holding the hyperparameters of the paper, is provided in the supplementary material. Upon publication, we will release the pre-trained weights, the fine-tuning code, and the benchmarks built from public data.

### AI USE STATEMENT

We used generative AI tools to refactor and simplify the code base and to remove redundant code. The methods themselves were designed and implemented by the authors. We have not used generative AI tools to design the methodology or the experiments, to clean or reformat data, to formulate or prove mathematical claims, or to interpret results. The remaining tasks requiring disclosure are not applicable. We also used LLMs to polish the writing of the paper, including finding grammar mistakes and improving overall readability. All AI-assisted code and suggested edits were reviewed by the authors, who take full responsibility for the final content of this work.

## REFERENCES

Jacob Austin, Daniel D. Johnson, Jonathan Ho, Daniel Tarlow, and Rianne van den Berg. Structured Denoising Diffusion Models in Discrete State-Spaces. In *Advances in Neural Information Processing Systems 34 (NeurIPS 2021)*, 2021.

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. doi: 10.1016/j.cels.2021.09.012.

Haiyang Bian, Yinxin Chen, Lei Wei, and Xuegong Zhang. uHAF: a unified hierarchical annotation framework for cell type standardization and harmonization. *Bioinformatics*, 41(4):btaf149, 2025. doi: 10.1093/bioinformatics/btaf149.

Sebastian Birk, Mohammad Vali Sanian, Amirhossein Vahidi, Samuel Ogden, Daniyal Jafree, Adib Miraki Feriz, Carlo Leonardi, Arpit Merchant, Zichen He, Lloyd Steele, Adam Boxall, Marta Rosa Sallese, Jonas Maaskola, Ciro Ramirez-Suastegui, Sevda Ogut, et al. Multi-scale modeling of human tissues from spatial transcriptomics with TERRA. *bioRxiv*, 2026. doi: 10.64898/2026.07.29.741565. Preprint.

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*, 22:2539–2550, 2025. doi: 10.1038/s41592-025-02899-6.

Zixuan Cang, Yanxiang Zhao, Axel A Almet, Adam Stabell, Raul Ramos, Maksim V Plikus, Scott X Atwood, and Qing Nie. Screening cell–cell communication in spatial transcriptomics via collective optimal transport. *Nature Methods*, 20:218–228, 2023. doi: 10.1038/s41592-022-01728-4.

Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T. Freeman. MaskGIT: Masked Generative Image Transformer. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11305–11315, 2022. doi: 10.1109/CVPR52688.2022.01103.

Mark Chen, Alec Radford, Rewon Child, Jeffrey Wu, Heewoo Jun, David Luan, and Ilya Sutskever. Generative Pretraining From Pixels. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, volume 119 of *Proceedings of Machine Learning Research*, pp. 1691–1703, 2020.

Mo Chen, Minsheng Hao, Xinquan Liu, Lin Deng, Peng Liu, Chen Li, Dongfang Wang, Kui Hua, Liang Guo, Xuegong Zhang, and Lei Wei. Querying structural and functional niches on spatial transcriptomics data. *arXiv preprint arXiv:2410.10652*, 2024. doi: 10.48550/arXiv.2410.10652.

Limor Cohen, Aaron Halpern, Timothy R. Blosser, Phil Che, Xingjie Pan, and Xiaowei Zhuang. Whole-transcriptome-scale and isoform-resolved spatial imaging of single cells in complex tissues. *bioRxiv*, 2025. doi: 10.1101/2025.08.27.672533.

Silvia Domcke and Jay Shendure. A reference cell tree will serve science better than a reference cell atlas. *Cell*, 186(6):1103–1114, 2023. doi: 10.1016/j.cell.2023.02.016.

Kangning Dong and Shihua Zhang. Deciphering spatial domains from spatially resolved transcriptomics with an adaptive graph attention auto-encoder. *Nature Communications*, 13:1739, 2022. doi: 10.1038/s41467-022-29439-6.

David S Fischer, Anna C Schaar, and Fabian J Theis. Modeling intercellular communication in tissues using spatial graphs of cells. *Nature Biotechnology*, 41:332–336, 2023. doi: 10.1038/s41587-022-01467-z.

Valentin Gabeur, Shangbang Long, Songyou Peng, Paul Voigtlaender, Shuyang Sun, Yanan Bao, Karen Truong, Zhicheng Wang, Wenlei Zhou, Jonathan T. Barron, Kyle Genova, Nithish Kannen, Sherry Ben, Yandong Li, Mandy Guo, et al. Image Generators are Generalist Vision Learners. *arXiv preprint arXiv:2604.20329*, 2026. doi: 10.48550/arXiv.2604.20329.

Jérôme Galon and Daniela Bruni. Approaches to treat immune hot, altered and cold tumours with combination immunotherapies. *Nature Reviews Drug Discovery*, 18(3):197–218, 2019. doi: 10.1038/s41573-018-0007-y.

Jing Gong, Minsheng Hao, Xingyi Cheng, Xin Zeng, Chiming Liu, Jianzhu Ma, Xuegong Zhang, Taifeng Wang, and Le Song. xTrimoGene: An Efficient and Scalable Representation Learner for Single-Cell RNA-Seq Data. In *Advances in Neural Information Processing Systems 36 (NeurIPS 2023)*, pp. 69391–69403, 2023. doi: 10.52202/075280-3039.

Christoph Hafemeister and Rahul Satija. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. *Genome Biology*, 20:296, 2019. doi: 10.1186/s13059-019-1874-1.

Minsheng Hao, Nan Yan, Haiyang Bian, Yixin Chen, Jin Gu, Lei Wei, and Xuegong Zhang. GeST: Towards Building A Generative Pretrained Transformer for Learning Cellular Spatial Context. *bioRxiv*, 2025. doi: 10.1101/2025.04.09.648072. Preprint.

Austin Hartman and Rahul Satija. Comparative analysis of multiplexed in situ gene expression profiling technologies. *eLife*, 13:RP96949, 2024. doi: 10.7554/eLife.96949.1. Reviewed preprint.

Tianhong Li, Huiwen Chang, Shlok Kumar Mishra, Han Zhang, Dina Katabi, and Dilip Krishnan. MAGE: MAsked Generative Encoder to Unify Representation Learning and Image Synthesis. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2142–2152, 2023a. doi: 10.1109/CVPR52729.2023.00213.

Zhuoxuan Li, Tianjie Wang, Pentao Liu, and Yuanhua Huang. SpatialDM for rapid identification of spatially co-expressed ligand–receptor and revealing cell–cell communication patterns. *Nature Communications*, 14:3995, 2023b. doi: 10.1038/s41467-023-39608-w.

Haiping Liu, Qian Zhao, Lijing Lin, Jingyuan Sun, and Hongpeng Zhou. CellWorld: From Gene-Level Reconstruction to Latent Cell Prediction in Spatial Transcriptomics Foundation Models. *arXiv preprint arXiv:2608.06659*, 2026. doi: 10.48550/arXiv.2608.06659.

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. doi: 10.1038/s41592-018-0229-2.

Kristen R Maynard, Leonardo Collado-Torres, Lukas M Weber, Cedric Uytingco, Brianna K Barry, Stephen R Williams, Joseph L Catallini, Matthew N Tran, Zachary Besich, Madhavi Tippani, Jennifer Chew, Yifeng Yin, Joel E Kleinman, Thomas M Hyde, Nikhil Rao, Stephanie C Hicks, Keri Martinowich, and Andrew E Jaffe. Transcriptome-scale spatial gene expression in the human dorsolateral prefrontal cortex. *Nature Neuroscience*, 24:425–436, 2021. doi: 10.1038/s41593-020-00787-0.

Shen Nie, Fengqi Zhu, Zebin You, Xiaolu Zhang, Jingyang Ou, Jun Hu, Jun Zhou, Yankai Lin, Ji-Rong Wen, and Chongxuan Li. Large Language Diffusion Models. In *Advances in Neural Information Processing Systems 38 (NeurIPS 2025)*, pp. 56354–56392, 2025. doi: 10.52202/085713-1689.

Jingyang Ou, Shen Nie, Kaiwen Xue, Fengqi Zhu, Jiacheng Sun, Zhenguo Li, and Chongxuan Li. Your Absorbing Discrete Diffusion Secretly Models the Conditional Distributions of Clean Data. In *International Conference on Learning Representations (ICLR)*, 2025. arXiv:2406.03736.

Jiexue Pan, Yuejiao Li, Zhongliang Lin, Qing Lan, et al. Spatiotemporal transcriptome atlas of human embryos after gastrulation. *Nature*, 654(8119):751–761, 2026. doi: 10.1038/s41586-026-10545-0.

Zhiliang Peng, Li Dong, Hangbo Bao, Qixiang Ye, and Furu Wei. BEiT v2: Masked Image Modeling with Vector-Quantized Visual Tokenizers. *arXiv preprint arXiv:2208.06366*, 2022. doi: 10.48550/arXiv.2208.06366.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving Language Understanding by Generative Pre-Training. Technical report, OpenAI, 2018.

Subham Sekhar Sahoo, Marianne Arriola, Yair Schiff, Aaron Gokaslan, Edgar Marroquin, Justin T. Chiu, Alexander M. Rush, and Volodymyr Kuleshov. Simple and Effective Masked Diffusion Language Models. In *Advances in Neural Information Processing Systems 37 (NeurIPS 2024)*, pp. 130136–130184, 2024. doi: 10.52202/079017-4135.

Christian M Schürch, Salil S Bhate, Graham L Barlow, et al. Coordinated cellular neighborhoods orchestrate antitumoral immunity at the colorectal cancer invasive front. *Cell*, 182(5):1341–1359.e19, 2020. doi: 10.1016/j.cell.2020.07.005.

Xin Shao, Chengyu Li, Haihong Yang, Xiaoyan Lu, Jie Liao, Jingyang Qian, Kai Wang, Junyun Cheng, Penghui Yang, Huajun Chen, Xiao Xu, and Xiaohui Fan. Knowledge-graph-based cell–cell communication inference for spatially resolved transcriptomic data with SpaTalk. *Nature Communications*, 13:4429, 2022. doi: 10.1038/s41467-022-32111-8.

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*, 56:431–441, 2024. doi: 10.1038/s41588-024-01664-3.

Valentine Svensson. Droplet scRNA-seq is not zero-inflated. *Nature Biotechnology*, 38:147–150, 2020. doi: 10.1038/s41587-019-0379-5.

Alejandro Tejada-Lapuerta, Anna C Schaar, Robert Gutgesell, Giovanni Palla, Lennard Halle, Mariia Minaeva, Larsen Vornholz, Leander Dony, Francesca Drummer, Till Richter, Mojtaba Bahrami, and Fabian J Theis. Nicheformer: a foundation model for single-cell and spatial omics. *Nature Methods*, 22:2525–2538, 2025. doi: 10.1038/s41592-025-02814-z.

Paul C. Tumeh, Christina L. Harview, Jennifer H. Yearley, I. Peter Shintaku, Emma J. M. Taylor, Lidia Robert, Bartosz Chmielowski, Marko Spasic, Gina Henry, Voicu Ciobanu, Alisha N. West, Manuel Carmona, Christine Kivork, Elizabeth Seja, Grace Cherry, et al. PD-1 blockade induces responses by inhibiting adaptive immune resistance. *Nature*, 515(7528):568–571, 2014. doi: 10.1038/nature13954.

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*, 2025. doi: 10.1101/2025.02.05.636714. Preprint.

Quanxin Wang, Song-Lin Ding, Yang Li, et al. The Allen Mouse Brain Common Coordinate Framework: A 3D Reference Atlas. *Cell*, 181(4):936–953.e20, 2020. doi: 10.1016/j.cell.2020.04.007.

Hongzhi Wen, Wenzhuo Tang, Xinnan Dai, Jiayuan Ding, Wei Jin, Yuying Xie, and Jiliang Tang. CellPLM: Pre-training of Cell Language Model Beyond Single Cells. In *International Conference on Learning Representations (ICLR)*, 2024. bioRxiv preprint 10.1101/2023.10.03.560734.

Zizhen Yao, Cindy T J van Velthoven, Michael Kunst, Meng Zhang, Delissa McMillen, Changkyu Lee, Won Jung, Jeff Goldy, Aliya Abdelhak, Matthew Aitken, Katherine Baker, Pamela Baker, Eliza Barkan, Darren Bertagnolli, Ashwin Bhandiwad, et al. A high-resolution transcriptomic and spatial atlas of cell types in the whole mouse brain. *Nature*, 624:317–332, 2023. doi: 10.1038/s41586-023-06812-z.

Hongkui Zeng. What is a cell type and how to define it? *Cell*, 185(15):2739–2755, 2022. doi: 10.1016/j.cell.2022.06.031.

Haorui Zhang, Zongxu Zhang, Peiyu Wang, Tian Xu, Xiaoyu Chen, Yanping Zhao, Siyu Lin, Wenjie Cai, Pengfei Ren, Ce Luo, Peng Zhang, Yunfeng Wang, Sen Hou, Yahui Zhao, Hu Zeng, et al. Uncovering spatially resolved functional genomics with CRISPR screen sequencing. *Cell*, 189: 4594–4618.e48, 2026. doi: 10.1016/j.cell.2026.04.049.

Suyuan Zhao, Yizhen Luo, Ganbo Yang, Yan Zhong, Hao Zhou, and Zaiqing Nie. SToFM: a Multi-scale Foundation Model for Spatial Transcriptomics. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, 2025. arXiv:2507.11588.

# A PRE-TRAINING DETAILS

**Pre-training corpus.** The corpus is 396 cell-segmented whole-transcriptome Stereo-seq sections, 161 human and 235 mouse, covering 12 organs (brain, skin, intestine, eye, heart, liver, stomach, breast, nasopharynx, lung, bone and ovary), 356 normal and 40 tumor, 49.1 million cells. This corpus is what remains after the following quality control: a cell with fewer than 100 detected genes is dropped, and a section is dropped when this removes more than 20% of its cells. Genes are restricted to the 15,903 protein-coding genes with a one-to-one human–mouse ortholog, so one vocabulary serves both species. A cell is a neighborhood center in pre-training only when at least 10 cells lie within 150 $\mu$m of it.

**Hyperparameters.** Table A.1 lists the tokenizer, the backbone and the pre-training schedule. The tokenizer uses cosine distance at the first level and Euclidean distance on the residual at the second, and each of its batches draws one cell from each of 1,024 sections with replicate, so that no section dominates the codebook statistics of a batch. The tokenizer trains for 1.8 hours and the backbone for 3.4 hours, each on one A100 80G GPU.

Table A.1: Hyperparameters of the tokenizer, the backbone, pre-training and decoding.

| | |
|---|---|
| *Tokenizer* | |
| Encoder / decoder widths | 512, 256, 128 / 256, 512 |
| Latent dimension | 128 |
| Codebooks | 2 levels $\times$ 256 codes |
| Distance per level | cosine, Euclidean |
| Codebook update | exponential moving average, decay 0.99 |
| Commitment weight $\beta$ | 1.0 |
| Normalization | BatchNorm |
| Batch | 1,024 sections $\times$ 1 cell |
| Steps, learning rate | 90,000, $10^{-3}$ |
| Weight decay, gradient clipping | 0.01, 1.0 |
| *Backbone* | |
| Layers, width, heads, MLP width | 10, 256, 8, 1,024 |
| Positional encoding | axial RoPE, $\theta = 10$ |
| Neighborhood radius, minimum cells | 150 $\mu$m, 10 |
| *Pre-training* | |
| Masking rate | $t \sim U(0.001, 1)$ |
| Batch | 768 neighborhoods $\times$ 2 accumulation steps |
| Steps, learning rate | 32,500, $10^{-4}$ constant |
| Optimizer | $\beta = (0.9, 0.95)$, weight decay 0.01, gradient clipping 1.0 |
| *Decoding* | |
| Steps $T$, temperature $\tau$ | 32, 0 |

## B ROTATION OF THE COORDINATES AT TEST TIME

Theoretically, the 2D RoPE of Section 3.2 is not rotation-invariant, since the two halves of each head are rotated by the $x$ and $y$ components of the displacement separately. In practice, sections are placed on the slide at arbitrary orientations, so we test how much rotating the coordinates at test time affects the model performance. We keep the fine-tuned SUDOKU weights and rotate only the center-relative coordinates that enter the positional encoding, at degrees including 45, 90, 158, 180 and 279 (158 and 279 degrees are drawn at random). The neighborhood graphs, the measured spacing and the tokenizer codes remain unchanged. The 0-degree setting thus refers to the original model. Table A.2 reports the results. The representation moves little at every stage of the readout. The cosine similarity between the rotated and the unrotated run is near 1 in all settings. As a comparison, we also randomly draw representations of different items without rotation, whose median cosine similarity is only 0.19, 0.30 and 0.07. In downstream performance metrics, HESTA AUPRC almost has no change as well. For the generation accuracy, it moves by at most 0.008. These results show that in practice, the model is robust to arbitrary rotations of the coordinates.

Table A.2: Impact of test-time coordinate rotation on the model. Top: median cosine similarity between the rotated and the unrotated run of the same item at each stage of the HESTA readout; the last column is the median cosine between two different items at 0 degrees. Bottom: benchmark scores at each angle. –: not exported.

| | Rotation angle (degrees) | | | | | | Different |
|---|---|---|---|---|---|---|---|
| | 0 | 45 | 90 | 158 | 180 | 279 | items |
| *Cosine similarity to the unrotated run, median* | | | | | | | |
| Cell representation $\mathbf{h}$ | 1 | 0.9976 | 0.9953 | – | 0.9941 | – | 0.19 |
| Neighborhood representation $\bar{\mathbf{h}}$ | 1 | 0.9999 | 0.9998 | 0.9997 | 0.9997 | 0.9998 | 0.30 |
| Niche query vector | 1 | 0.99999 | 0.99999 | 0.99998 | 0.99998 | 0.99999 | 0.07 |
| *Benchmark readings* | | | | | | | |
| HESTA AUPRC | 0.503 | 0.503 | 0.503 | 0.502 | 0.502 | 0.503 | |
| Generation accuracy | 0.405 | 0.401 | 0.397 | 0.397 | 0.397 | 0.399 | |

## C TRAINING OBJECTIVE AS A VARIATIONAL BOUND

Section 3.2 feeds visible cells as continuous latents but predicts discrete codes, which differs from masked diffusion language models (Sahoo et al., 2024; Ou et al., 2025), where both the input and the target are codes. This appendix shows that Equation 7 is nevertheless the variational bound of an absorbing-state diffusion model over the neighborhood, so minimizing it trains a generative model. The argument has three steps: the token is defined as the pair of a latent and its code, so that a diffusion over tokens may condition on latents; the reverse process is factorized so that the network only predicts the code; and the extra term this introduces into the bound does not depend on the network.

**State space.** Fix the neighborhood $N$ and its coordinates $\mathbf{p}_N$; only cell states diffuse, and every distribution below is conditioned on $\mathbf{p}_N$. The token of cell $j$ is the pair $\mathbf{x}_j = (\mathbf{z}_j, c_j)$ of its tokenizer latent and its two-level code $c_j = Q(\mathbf{z}_j)$, a deterministic function of the latent under the frozen tokenizer. The data distribution is the distribution $q(\mathbf{x}_N)$ of these pairs over corpus neighborhoods.

**Forward and reverse process.** The absorbing forward process replaces each token independently by the mask with probability $t$; it acts on whole tokens and does not depend on what a token contains. Writing $M_t$ for the masked set and $\mathbf{x}_t$ for the partially masked neighborhood, the reverse process of Sahoo et al. (2024) unmasks each masked cell independently and draws its token from a model $p_\theta(\mathbf{x}_j \mid \mathbf{x}_t)$. We take

$$p_\theta(\mathbf{x}_j \mid \mathbf{x}_t) = p_\theta(c_j \mid \mathbf{x}_t)\; q_{\mathrm{tok}}(\mathbf{z}_j \mid c_j), \tag{10}$$

where $q_{\mathrm{tok}}(\mathbf{z} \mid c)$ is the distribution of tokenizer latents within the cell of code $c$ over the corpus, a fixed distribution that does not depend on $\theta$. Since $c_V = Q(\mathbf{z}_V)$, the partially masked neighborhood $\mathbf{x}_t$ carries exactly the information of $(\mathbf{z}_V, \mathbf{p}_N)$, so $p_\theta(c_j \mid \mathbf{x}_t)$ is the network of Section 3.2, read at the masked cells.

**The bound.** For the absorbing process with the linear schedule, the negative evidence lower bound of Sahoo et al. (2024) and Ou et al. (2025) in continuous time is

$$-\log p_\theta(\mathbf{x}_N) \;\le\; \mathbb{E}_{t\sim U(0,1)}\, \mathbb{E}_{\mathbf{x}_t}\; \frac{1}{t} \sum_{j \in M_t} -\log p_\theta(\mathbf{x}_j \mid \mathbf{x}_t). \tag{11}$$

Substituting Equation 10 splits the summand into $-\log p_\theta(c_j \mid \mathbf{z}_V, \mathbf{p}_N) - \log q_{\mathrm{tok}}(\mathbf{z}_j \mid c_j)$. The second term does not depend on $\theta$, so the right-hand side equals Equation 7 plus a constant, and minimizing Equation 7 minimizes the bound. The $1/t$ weighting is the schedule term of the bound. Equation 7 is therefore the negative evidence lower bound of a generative model over neighborhoods, up to a constant, which is the property used in Section 3.2.

**Relation to the sampler.** The bound above concerns the model with the reverse process of Equation 10, in which the unmasked cells are chosen at random and the latent of an unmasked cell is drawn from $q_{\mathrm{tok}}(\mathbf{z} \mid c)$. The sampler of Section 3.3 uses the same network but approximates this reverse process in three ways: a committed cell enters the conditioning through its codebook reconstruction $\mathbf{z}_q(\hat{c}_j)$ rather than through a draw from $q_{\mathrm{tok}}(\mathbf{z} \mid c)$, cells are committed in order of confidence rather than at random, and the cells committed within one step, as well as the two code levels of a cell, are drawn independently given the current context. Pre-training also draws $t$ from $U(\epsilon, 1)$ with $\epsilon = 0.001$ rather than from $U(0, 1)$.

# D BENCHMARKS

Table A.3 lists the data of the five benchmarks.

Table A.3: The evaluation suite. Protocols are in Appendix D and Appendix E, SPAC-seq in three folds by section, and no benchmark section is in the pre-training corpus.

| Benchmark | Species | Tissue | Platform | Resolution | Sections | Cells | Units |
|---|---|---|---|---|---|---|---|
| HESTA | human | embryo | Stereo-seq | cell | 23 | 9.7M | 5,650 NOIs |
| DLPFC | human | cortex | 10x Visium | spot | 12 | 47,329 spots | 252 NOIs |
| Multi-scale clustering | mouse | brain | MERFISH | cell | 6 | 235,857 | 3 label levels |
| Spatial generation | human, mouse | brain, kidney | Stereo-seq | cell | 31 | 1,938,543 | 972 questions |
| SPAC-seq | mouse | lung metastasis | Stereo-seq | cell | 3 | 597,007 | 125 clones |

**Protocols.** No benchmark section is in the pre-training corpus. SUDOKU is fine-tuned on the benchmark sections with its own objective and no benchmark label; how each baseline is run is stated in Appendix E. Benchmark sections enter SUDOKU through the pre-training vocabulary of 15,903 genes: genes are matched by identifier or through the one-to-one ortholog table, and missing genes are set to zero. A Visium spot of DLPFC is treated as one cell. On HESTA, DLPFC and multi-scale clustering the sections fine-tuned on are the sections scored, all 23 of HESTA, all 12 of DLPFC and the six MERFISH sections. On spatial generation the sections fine-tuned on are 61 sections of the same three cohorts with no questions. On SPAC-seq each fold fine-tunes the model on its two training sections, builds its code-to-type table there, and generates on the held-out section.

**Niche retrieval, HESTA.** HESTA is 23 Stereo-seq sections of human embryos at cell resolution, 9.7 million cells with an organ label each, from the atlas of Pan et al. (2026). We define NOIs in spatially contiguous regions where all cells are purely labeled as the same organ. On each section and for each organ we take the cells of that organ, keep the largest connected component of their 8-nearest-neighbor spatial graph, and split it by k-means on the coordinates into pieces of about 1,000 cells, keeping the pieces of 300 to 3,000 cells. This gives 5,650 niches of interest over 47 organs. For an NOI and a target section, every candidate niche on the target sections is scored by the cosine similarity of its representation to the query NOI representation. The cosine similarity then serves as a binary classification metric to identify regions with the correct organ label. The score of the pair (NOI, target section) is the AUPRC score of this binary classification problem. Figure A.1a walks through one pair for SUDOKU and CellPLM, a heart NOI queried against a section of an earlier stage. SUDOKU's similarity is high on the heart of the target and low elsewhere, and CellPLM's is high over most of the target. Among the top-ranked target cells, as many as there are heart cells, 94% are heart for SUDOKU and 67% for CellPLM, and the AUPRC of the pair is 0.99 and 0.61. Pairs whose target section has no cell of the query organ are dropped, leaving 84,977 scoreable pairs. Table 2 reports the mean AUPRC of the eight main organs, which hold 86% of the scoreable pairs.

**Niche retrieval, DLPFC.** DLPFC is twelve 10x Visium sections of human dorsolateral prefrontal cortex from three donors with a cortical layer label per spot, 47,329 labeled spots in all (Maynard et al., 2021). We use the niche query benchmark released with QueST (Chen et al., 2024), whose NOIs are sets of 100 spots at the junction of two or three layers, seven classes in all (L3L4, L4L5, L5L6, L6WM, L3L4L5, L4L5L6, L5L6WM), with three replicates per class on every section. For an NOI and a target section, every candidate niche on the target section is scored by the cosine similarity of its representation to the query NOI representation. The ground truth is the matching score between the NOI and each candidate niche, computed by a Weisfeiler-Lehman graph kernel on the layer labels and released with the benchmark. The score of the pair (NOI, target section) is the Pearson correlation between similarity and matching score over the candidate niches of the target section. Figure A.1b walks through one pair for SUDOKU and CellPLM. The matching score on the target section is a band along the border of L4 and L5, with a mean of 0.62 in L4, 0.67 in L5, 0.13 in L6 and near zero in the other layers. SUDOKU's similarity forms the same band, with layer means of 0.32 in L4, 0.42 in L5, 0.13 in L6 and below zero elsewhere, PCC 0.87. CellPLM's similarity varies by less than 0.08 across the layers, PCC 0.24. Every NOI is paired with the other eleven sections, giving 2,772 pairs, and Table 2 reports the mean PCC over the pairs of each class.

**Multi-scale clustering.** The data are six coronal sections of mouse brain from three mice measured by whole-transcriptome MERFISH with 22,930 genes and 235,857 cells (Cohen et al., 2025).

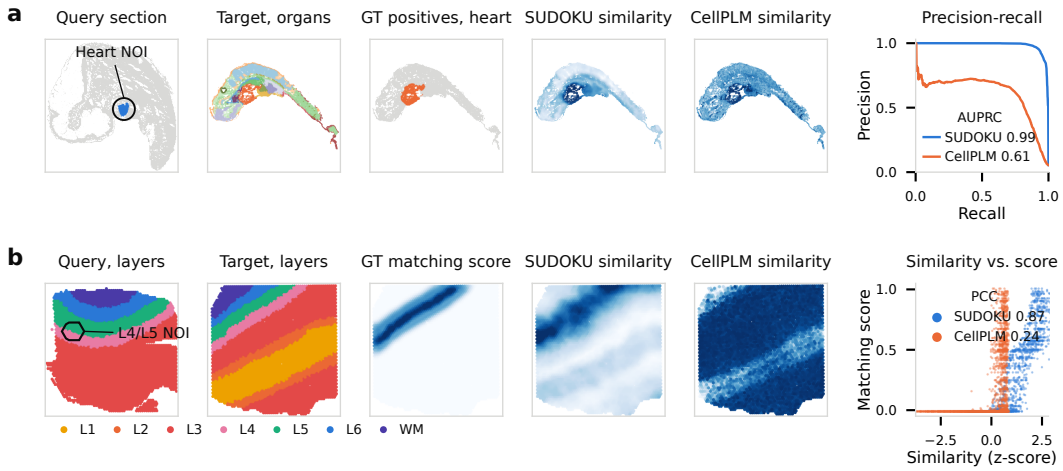



Figure A.1: Niche retrieval, one worked example per scenario, SUDOKU next to CellPLM. (a) HESTA. A heart niche of interest (NOI) on one section is the query, and the target section is shown colored by organ. Every candidate niche of the target is scored by the cosine similarity of its representation to the query NOI representation, shown as a map per model. Candidate niches of the correct organ are the positives, and others are the negatives for the binary classification. (b) DLPFC. An NOI at the L4/L5 junction is the query, and the target section is shown with its cortex layer labels. Every candidate niche of the target is scored by cosine similarity to the query NOI, per model. The ground truth is the matching score between the NOI and each candidate niche, computed by a Weisfeiler-Lehman graph kernel on the layer labels. The score of the pair is the Pearson correlation between similarity and matching score over the candidate niches of the target. Similarity maps run from light to dark over the 1st to 99th percentile of each model's values.

Every cell carries labels at three levels. Region is the Allen brain division of the cell's registered position (isocortex, hippocampal formation, olfactory areas, striatum, pallidum, thalamus, hypothalamus, midbrain, pons, medulla and others), with all fiber tracts merged into one class and all ventricles into another, 12 to 13 classes per section. Band is a laminar band read from the tissue itself. The whole transcriptome of every cell is mapped to the Allen whole-mouse-brain taxonomy (Yao et al., 2023), and the neurons whose subclass belongs to one cortical layer (L2/3, L4/5, L5, L6, L6b) or one hippocampal subfield (CA1, CA2 and IG, CA3, DG) vote for the band of every cell within 75 $\mu$m of them. A cell takes the winning band when at least five such neurons are within reach and the winner holds at least 60% of the votes, so glia and vessels carry the band of the layer they sit in, 6 to 8 classes per section. Cell type is the clustering label of the data provider, 30 to 36 classes per section. Figure A.2 shows the three levels on one section. At every level a class with fewer than 20 cells or 0.1% of the cells of a section is undefined there, and undefined cells take part in clustering but not in scoring. A method returns one vector per cell and scale. For each section and each representation we build a 15-nearest-neighbor graph in the representation space and cluster it with Leiden, searching the resolution by bisection so that the number of clusters among the defined cells equals the number of classes of the level being scored. ARI and NMI are computed on the defined cells and Table 3 reports the median over the six sections.

**Spatial generation.** The benchmark has 31 sections in six groups: human Alzheimer brain (7), human control brain (7), two mouse brains (5 each), injured kidney (4), control kidney (3). Windows of 500 cells are cut per group, 81 in total; four missing geometries times three missing fractions give 972 questions. Scattered points leave visible cells next to every masked cell, the interior block and the shell leave context on one side of the masked region only, and multiple blocks fall in between. Within each tissue group, windows are ranked by a composite heterogeneity score, mainly the entropy of the cell-type composition and the share of the most frequent type, together with the same statistics on the tokenizer codes, and split into easy, medium and hard tertiles. Figure A.3 shows the four geometries on an easy, a medium and a hard window at missing fraction 0.4, and one window per tissue group. The ground truth is the annotated type of the masked cells. For SUDOKU and GeST the code distribution of each generated cell at the step it is committed is folded into a type

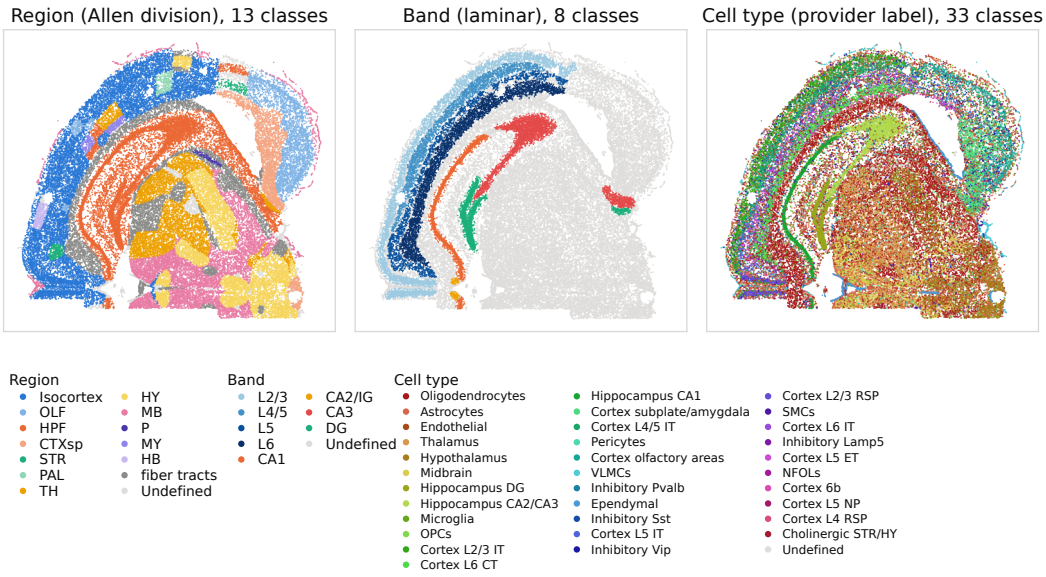



Figure A.2: The three label levels of the multi-scale clustering benchmark on one section. Region is the Allen brain division of the registered position, Band the laminar band read from the layer-specific neurons around each cell, and Cell type the clustering label of the data provider. Cells without a label at a level are grey.

distribution through a code-to-type table built outside the window on the same section, and accuracy is read as an expectation.

**Immune-recruitment prediction.** From the 345 single-perturbation clones of three sections we keep those whose perturbation has at least two clones with consistent measured recruitment, a coefficient of variation across these clones of at most 0.35, so that the measured recruitment is reproducible enough to serve as ground truth, whose 50-$\mu$m band holds at least 20 immune cells, so that the immune fraction is estimable, and whose clone and band together hold at most 1,000 cells, the neighborhood capacity of the model, giving 125 questions. We mask all cells within a 50-$\mu$m band around the clone, give the model only the cells of the clone, generate the states of the cells in the band, and read out the fraction of immune cells among the generated non-tumor cells. The model input carries no perturbation identity and the prediction depends only on spatial context. The readout goes through the codebook: on the fine-tuning sections we estimate the conditional distribution $P(y \mid c)$ from each two-level code to annotated type; writing $I$ for the set of immune types, $y_{\mathrm{tum}}$ for the tumor type and $B$ for the cells to be generated in the band, and $p_\theta(c_j \mid \text{clone})$ for the distribution of cell $j$ at the step it is committed, the predicted immune fraction of a question is the expectation

$$\hat{f} = \frac{\sum_{j \in B} \sum_c p_\theta(c_j = c \mid \text{clone})\, P(y \in I \mid c)}{\sum_{j \in B} \sum_c p_\theta(c_j = c \mid \text{clone})\, P(y \neq y_{\mathrm{tum}} \mid c)}. \tag{12}$$

The metrics are the Spearman correlation between $\hat{f}$ and the measured fraction over the clones of each section and the rank error of each clone, as defined in Section 4.3.

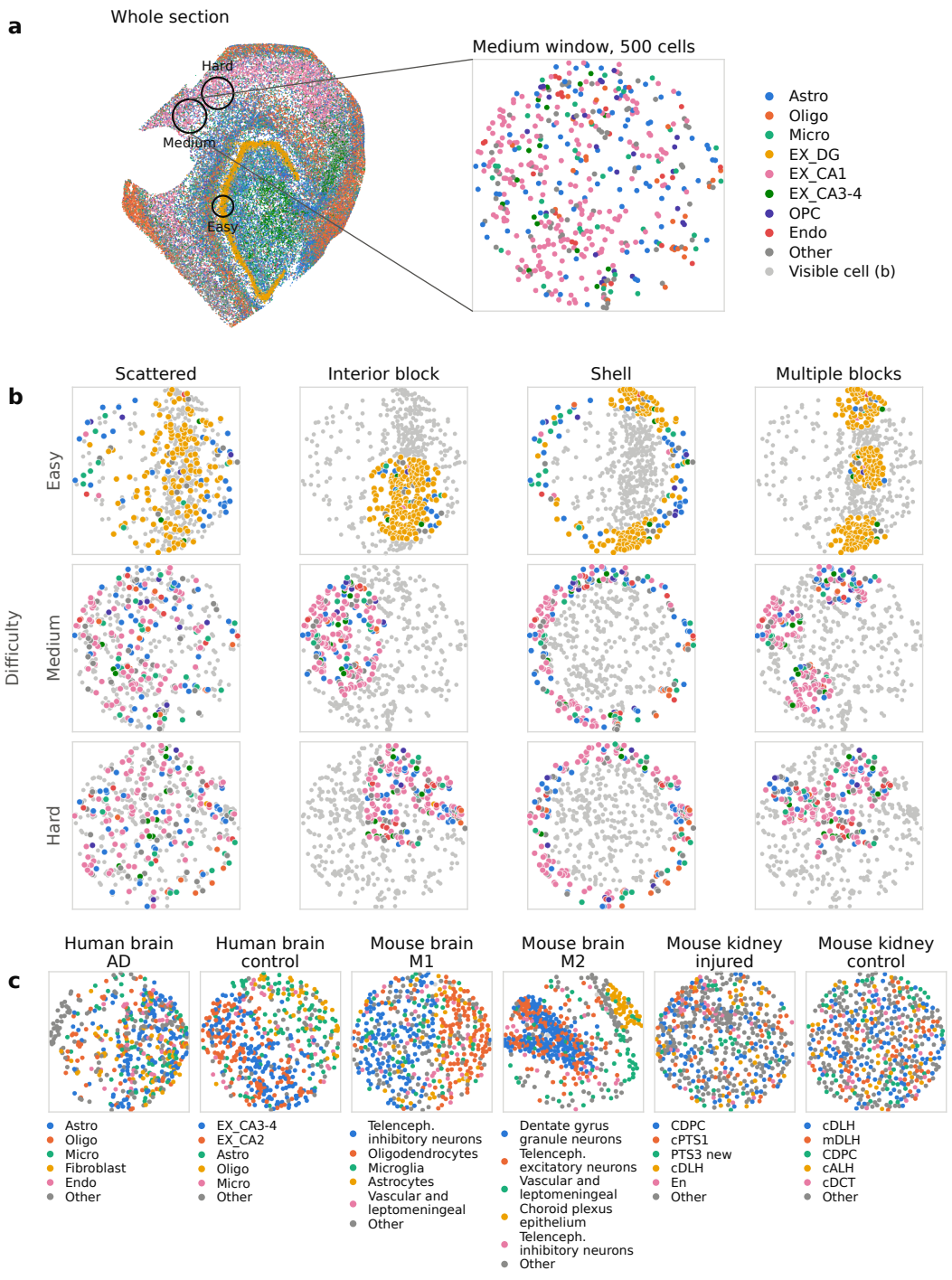



Figure A.3: Spatial generation questions. (a) A human control brain section with an easy, a medium and a hard window of 500 cells circled, the medium one magnified, all cells colored by true type. (b) The four missing geometries (columns, in the order of Table 4) at missing fraction 0.4 on the three windows (rows). Grey dots are the visible cells given to the model, colored dots the cells to be generated, colored by their true type. (c) One window per tissue group with all cells colored by type. The windows in (b) are the most type-diverse of their tier on one section, and those in (c) the most type-diverse of their group.

# E BASELINES

**Baselines.** Every external model is run from its pre-trained weights with its default settings and read out at the same neighborhood radius as SUDOKU. GeST, CellPLM, Nicheformer and Novae, whose training code is available, are further fine-tuned on the benchmark sections with their own self-supervised objectives and no benchmark label; scGPT-spatial, SToFM and CellWorld release inference code only, and the self-supervised adaptation of TERRA exceeds our compute budget, so these four are used as released. Where a metric needs a type, a model's generated cells are typed through a table built on the same reference cells as our code-to-type table. Human sections are matched to each model's vocabulary by Ensembl identifier or gene symbol, and the mouse sections of the multi-scale benchmark through the one-to-one ortholog table, with genes outside a vocabulary dropped. *CellPLM* (Wen et al., 2024) starts from the released 85M checkpoint, is fine-tuned with its own objective on the benchmark sections, and is read out through its 512-dimensional cell embedding. *Nicheformer* (Tejada-Lapuerta et al., 2025) uses the released weights and is read out as the mean of the last layer over the gene tokens; its output head is not released and is trained for one epoch of its masked gene modeling on the benchmark sections, and HESTA is given a reserved platform token since Stereo-seq is not among its platforms. *scGPT-spatial* (Wang et al., 2025) and *SToFM* (Zhao et al., 2025) use their released weights and cell embeddings; SToFM tiles each section into sub-slices of about 1,000 cells as in its own pipeline. *CellWorld* (Liu et al., 2026) uses the released base model on the 512 most expressed genes of each cell, with the organ token set to brain for the brain sections and to unknown for HESTA, and is read out at each cell's own row of its context encoder. *TERRA* (Birk et al., 2026) uses the released 112M model on each cell with its ten nearest neighbors and is read out through its neighborhood embedding; on the multi-scale benchmark the cell type level uses its cell embedding. *Novae* (Blampey et al., 2025) uses the released human weights on HESTA and the released mouse weights on the multi-scale benchmark, each fine-tuned with its own objective on the benchmark sections. *GeST* (Hao et al., 2025) starts from a checkpoint pre-trained on the same corpus as SUDOKU and is fine-tuned with its own objective on each benchmark. Downstream tasks use its default settings. *QueST* (Chen et al., 2024), *STAGATE* and *BANKSY* (Dong & Zhang, 2022; Singhal et al., 2024) are task-specific methods without pre-training, designed for niche retrieval and spatial domain detection on ST data. We use their default settings. The training-free baselines are $k$-NN on spatial generation, which assigns each masked cell the majority type of its $k$ nearest visible cells, and expression interpolation on SPAC-seq, which interpolates the expression of each cell in the band from the visible cells and types it by the nearest type centroid.