ZOOMST: PRETRAINING SPATIAL TRANSCRIPTOMIC REPRESENTATIONS WITH LEARNABLE CROSS-GRANULARITY TRANSFORMATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Tissue function depends on spatial organization across scales, from cellular neighborhoods to anatomical regions. Expanding spatial context around a focal location yields related but distinct representations of tissue organization at different granularities. Although existing spatial transcriptomics pretraining models incorporate multiscale context for representation learning, they do not explicitly learn how representations change as spatial context expands. We introduce ZoomST, a self-supervised framework that jointly learns granularity-specific, cross-scale-compatible representations and transformations among them. ZoomST combines molecular-spatial dual-view alignment within each granularity with a soft distributional equivariance constraint across granularities. Across held-out mouse sections spanning all sampled developmental stages, frozen ZoomST representations outperform the evaluated pretrained baselines and remain competitive with target-fitted methods for tissue mapping. Anatomical analysis in a mouse developmental atlas further illustrates the complementary information captured by fine and coarse spatial context. Continuous transformations learned only from discrete training granularities predict representations at unseen intermediate spatial scales and reveal relative differences in context sensitivity across tissue locations. By making spatial granularity adjustable after pretraining, ZoomST turns changes in spatial context into both a source of self-supervision and an axis for tissue analysis.

1 INTRODUCTION

Tissue function depends not only on local cellular composition, but also on how cells and their neighborhoods are organized within larger anatomical structures (Schürch et al., 2020). Spatial transcriptomics measures molecular profiles together with spatial position, making it possible to study these levels of tissue organization within the same specimen (Chen et al., 2022; Pan et al., 2026). Importantly, tissue characteristics depend on the spatial scale at which they are considered: local neighborhoods capture cellular composition and interactions, whereas broader spatial contexts reveal regional organization and anatomical structure. Relating these scales is therefore essential for understanding how local molecular environments are embedded within larger tissue organization.

Spatial representation methods typically characterize tissue context within a predefined neighborhood or spatial graph (Long et al., 2023; Singhal et al., 2024). Multiscale methods extend this view by learning representations at several spatial scales or combining information across them (Yuan, 2024; Zhao et al., 2025; Lin et al., 2024). However, these approaches largely provide representations at a discrete set of predefined scales and treat them as parallel sources of information to be combined. Tissue organization can change systematically with scale. **Local structures may persist as broader spatial context introduces new regional characteristics and reorganizes their spatial relationships.** Capturing such scale-dependent organization therefore requires modeling not only representations at individual scales, but also how they transform across scales.

Existing pretraining objectives provide limited supervision for learning these transformations. Single-cell pretraining commonly predicts masked genes or expression values from molecular context within a cell (Theodoris et al., 2023; Cui et al., 2024), while spatial predictive objectives incorporate neighboring cells to recover molecular information or latent cell states (Zhao et al., 2025;

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.

Liu et al., 2026). These objectives primarily learn dependencies within predefined spatial contexts, providing little direct supervision for how representations change across scales and leaving scale-dependent structural variation largely unused as a pretraining signal.

We introduce ZoomST, a self-supervised pretraining framework for spatial transcriptomics that learns molecular–spatial dual-view representations of tissue context across spatial scales. We design a soft distributional equivariance constraint to learn structured transformations between scale-specific representations within a shared latent space while preserving scale-dependent tissue characteristics. This yields a structured cross-scale representation space that extends beyond a discrete set of embeddings at predefined spatial granularities, enabling tissue context to be characterized continuously across spatial granularities.

We evaluate both the representations and their transformations in developmental spatial transcriptomic atlases. Our main contributions are:

- **Molecular–spatial dual-view pretraining for tissue representation.** We develop a self-supervised pretraining framework that jointly learns molecular and spatial views of tissue context across multiple spatial scales, and provides scale-specific representations in a shared latent space.
- **Soft distributional equivariance for structured cross-scale learning.** We introduce a soft distributional equivariance constraint that imposes cross-scale consistency within a shared latent space while preserving scale-dependent tissue characteristics. This enables ZoomST to learn structured transformations between representations rather than treating different scales as independent embeddings.
- **Continuous cross-scale transformation and scale sensitivity.** The learned transformations interpolate tissue representations at unseen intermediate spatial scales despite supervision at only a discrete set of scales. They further provide a means to characterize how strongly tissue representations respond to changes in spatial context, revealing spatial variation in scale sensitivity.
- **Generalization across developmental spatial transcriptomic atlases.** We validate ZoomST on human and mouse developmental atlases. Frozen representations generalize to held-out tissue sections and support downstream tasks including tissue mapping and developmental organization, outperforming the evaluated pretrained baselines on tissue mapping while remaining competitive with target-fitted methods.

2 Related Work

Spatial context and multiscale representations. Spatial graph encoders combine expression with local connectivity (Dong & Zhang, 2022; Long et al., 2023), while neighborhood-based methods construct context-augmented features for tissue domains and niches (Singhal et al., 2024; Varrone et al., 2024). Novae extends contextual graph representations to zero-shot domain inference and hierarchical spatial domains (Blampey et al., 2025). Multiscale approaches organize broader context in different ways: MENDER summarizes cell-state composition across multiple ranges, SToFM supplies tissue-level summaries alongside local cells, and ST-Align aligns multimodal features within and across spot and niche levels (Yuan, 2024; Zhao et al., 2025; Lin et al., 2024). These approaches aggregate or align information across contexts. In contrast, ZoomST learns how a fixed location's representation transforms between specified granularities, while retaining the representations associated with each granularity.

Predictive self-supervision. Predictive objectives differ in both the context they use and the targets they infer. Nicheformer applies masked-token pretraining to expression sequences and supports downstream spatial-label and neighborhood-composition prediction (Tejada-Lapuerta et al., 2025). SToFM reconstructs masked cell embeddings and recovers pairwise distances after coordinate perturbation (Zhao et al., 2025). Representation-level prediction is also central to I-JEPA, which predicts target-region representations from image context (Assran et al., 2023). In spatial transcriptomics, TERRA predicts masked gene-token representations, whereas CellWorld predicts masked cell representations from visible context and partial-expression hints (Birk et al., 2026; Liu et al., 2026).

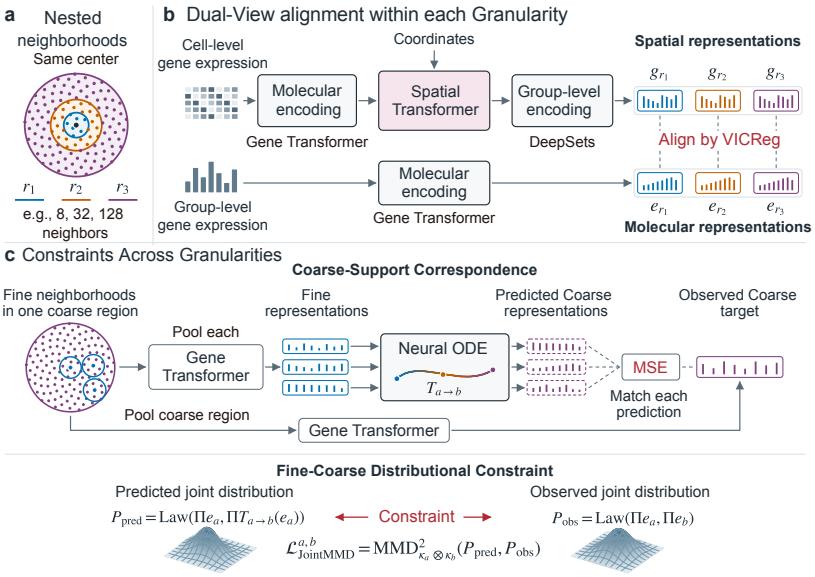

Figure 1: **ZoomST overview. a**, Nested neighborhoods centered at the same position define granularities $r = 8, 32, 128$. **b**, Molecular and spatial representations e_r and g_r are aligned by VICReg within each granularity. **c**, A shared neural ODE learns cross-granularity transformations of molecular representations using coarse-support MSE and JointMMD.

ZoomST likewise uses representation-level targets, but predicts a description at a specified observation granularity rather than masked content within a given context.

Equivariance and learned transformations. Learned equivariance maps relate input transformations to changes in representation space (Lenc & Vedaldi, 2015). E-SSL encourages sensitivity to selected augmentations through transformation prediction (Dangovski et al., 2022), while deep scale-spaces develops scale-equivariant operators for images (Worrall & Welling, 2019). ZoomST indexes representation transformations by the extent of tissue context around a fixed location and learns their statistical relation from nested observations.

3 METHOD

In this section, we present the core architecture of ZoomST, illustrated in Figure 1. ZoomST is pretrained on nested neighborhoods constructed at multiple spatial scales from spatial transcriptomic data. It combines molecular–spatial representation learning at each scale with structured cross-scale transformations.

3.1 MULTI-GRANULARITY INPUT FROM NESTED NEIGHBORHOODS

A spatial transcriptomic tissue section contains N measured positions indexed by $i \in \{1, \ldots, N\}$. Each position i is associated with a position-level molecular profile $x_i \in \mathbb{R}^G$, containing expression measurements over G genes, and a spatial coordinate $p_i \in \mathbb{R}^2$. Depending on the spatial transcriptomics platform, a measured position may correspond to a cell, spot, or spatial bin.

For each focal position i, we construct a nested neighborhood $S_{i,r}$ containing i and its $r-1$ nearest spatial neighbors, where neighborhood size r defines the spatial granularity of the tissue context, with larger r corresponding to broader spatial scales. We use a discrete set of granularities $r \in \mathcal{R} = \{8, 32, 128\}$, yielding nested tissue contexts centered at the same focal position. The corresponding granularity-level molecular profile is obtained by averaging the preprocessed expression profiles within each neighborhood, $m_{i,r} = \frac{1}{r} \sum_{j \in S_{i,r}} \bar{x}_j$. Both position-level and granularity-level molecular profiles are subsequently tokenized into fixed-length sequences of 512 gene tokens, with

each token encoding a gene identity and its discretized expression rank, before being passed to the ZoomST encoders.

3.2 Molecular–Spatial Dual-View Representation Learning

At each spatial granularity r, ZoomST learns complementary spatial and molecular representations of the same tissue context. For the spatial view, a position-level molecular Transformer first encodes each molecular profile x_j, $j \in S_{i,r}$, into a latent state $h_j = E_{\text{pos}}(\text{Tokenize}(x_j))$. A shared spatial Transformer then contextualizes these states within $S_{i,r}$ using their relative spatial coordinates and a granularity identifier. The contextualized states are subsequently aggregated by a shared DeepSets encoder (Zaheer et al., 2017) to obtain the spatial representation $g_{i,r} \in \mathbb{R}^d$.

For the molecular view, the granularity-level molecular profile $m_{i,r}$ is directly encoded by a separate molecular Transformer, $e_{i,r} = E_{\text{gran}}(\text{Tokenize}(m_{i,r})) \in \mathbb{R}^d$. Both representations use $d = 256$. The position-level and granularity-level molecular Transformers have separate parameters while sharing gene and expression-rank embeddings.

The dual-view representations are jointly pretrained with masked expression reconstruction and within-granularity VICReg alignment (Bardes et al., 2022), which aligns the corresponding representations from the two views at the same focal position and granularity. We define the base pretraining objective as

$$\mathcal{L}_{\text{base}} = \mathcal{L}_{\text{rec}} + \lambda_{\text{V}} \mathcal{L}_{\text{VICReg}}. \tag{1}$$

Objective details are provided in Appendix B.

3.3 Continuous Cross-Granularity Transformation Learning

To model how molecular representations change across spatial granularities, we parameterize a shared continuous transformation with a neural ODE (Chen et al., 2018). We define a normalized log-granularity coordinate $u(r) = \frac{\log_2(r/8)}{4}$, which maps the training granularities $r \in \{8, 32, 128\}$ to 0, $\frac{1}{2}$, and 1, respectively, so that equal multiplicative changes in neighborhood size correspond to equal intervals.

We define the continuous transformation from the molecular representation $e_{i,a}$ at source granularity a to target granularity b by integrating a learned vector field f_θ over the granularity coordinate,

$$\frac{dz(u)}{du} = f_\theta(z(u), u), \qquad z(u(a)) = e_{i,a}, \qquad T_{a\to b}(e_{i,a}) = z(u(b)). \tag{2}$$

The shared vector field defines a continuous family of cross-granularity transformations rather than separate mappings for individual granularity pairs. It is implemented as an MLP with a zero-initialized output layer, so training starts from identity transformations. We use fourth-order Runge–Kutta integration, with implementation details provided in Appendix B.

Nested neighborhoods provide local correspondence for learning these transformations. For each focal position i and target granularity b, we use the directly encoded coarse-granularity representation at each training iteration as a stop-gradient supervision target, $y_{i,b} = \text{sg}(e_{i,b})$, where sg denotes stop-gradient. For a batch of B focal positions and granularity pairs $\mathcal{P} = \{(8, 32), (8, 128), (32, 128)\}$, we define the coarse-support correspondence loss as

$$\mathcal{L}_{\text{CS}} = \frac{1}{|\mathcal{P}|B} \sum_{(a,b)\in\mathcal{P}} \sum_{i=1}^{B} \frac{1}{bd} \sum_{j\in S_{i,b}} \|T_{a\to b}(e_{j,a}) - y_{i,b}\|_2^2 . \tag{3}$$

This objective encourages fine-scale representations within the same coarse spatial support to transform toward a shared coarse context.

3.4 Soft Distributional Equivariance Across Granularities

The coarse-support correspondence provides local supervision for learning cross-granularity transformations, but does not explicitly constrain the joint structure of representations across scales. Since broader contexts contain additional scale-specific information, we impose a soft distributional

equivariance constraint on both the learned representations and transformations rather than requiring exact pointwise agreement.

For a granularity pair (a, b), let I index a sampled focal location. We define the joint distributions of the transformation and direct-encoding routes as

$$\begin{aligned} P_{ab}^{T} &= \mathrm{Law}\big(e_{I,a}, T_{a\to b}(e_{I,a})\big), \\ P_{ab} &= \mathrm{Law}\big(e_{I,a}, e_{I,b}\big). \end{aligned} \tag{4}$$

We impose soft distributional equivariance by encouraging $P_{ab}^{T} \approx P_{ab}$.

In practice, we compare the joint source–target distributions across matched focal locations. For a fixed orthogonal projection $\Pi : \mathbb{R}^{256} \to \mathbb{R}^{16}$, let $\widehat{P}_{ab}^{T,\Pi}$ and $\widehat{P}_{ab}^{\Pi}$ denote the empirical distributions of $(\Pi e_{i,a},\ \Pi T_{a\to b}(e_{i,a}))$ and $(\Pi e_{i,a},\ \Pi e_{i,b})$, respectively. We minimize their squared maximum mean discrepancy (MMD; Gretton et al., 2012),

$$\mathcal{L}_{\mathrm{JointMMD}} = \frac{1}{|\mathcal{P}|} \sum_{(a,b)\in\mathcal{P}} \mathrm{MMD}^2_{\kappa_a\otimes\kappa_b}\left(\widehat{P}_{ab}^{T,\Pi}, \widehat{P}_{ab}^{\Pi}\right). \tag{5}$$

The product kernel compares source and target representations jointly, making the objective sensitive to cross-granularity associations beyond their marginal distributions. Implementation details are provided in Appendix B.

3.5 Joint Pretraining Objective

The two cross-granularity losses are jointly pretrained with the base objective defined in Section 3.2,

$$\mathcal{L} = \mathcal{L}_{\mathrm{base}} + \lambda_{\mathrm{CS}}\mathcal{L}_{\mathrm{CS}} + \lambda_{\mathrm{J}}\mathcal{L}_{\mathrm{JointMMD}}. \tag{6}$$

Weights λ_{CS} and λ_{J} are gradually warmed up to one during training. Both cross-granularity objectives use unmasked molecular representations, with directly encoded coarse states as stop-gradient targets. Cross-granularity supervision is applied only at the training granularities.

3.6 Continuous Granularity Queries

After pretraining, ZoomST allows representations to be queried (interpolated) at any target granularity r within the modeled range through the learned transformation, $\hat{e}_i(r \mid a) = T_{a\to r}(e_{i,a})$, including intermediate granularities not explicitly used for supervision.

To quantify how strongly the representation of a tissue location responds to changes in spatial context, we define normalized response,

$$R_i(a, b \mid a_0) = \frac{\|\hat{e}_i(b \mid a_0) - \hat{e}_i(a \mid a_0)\|_2}{\sqrt{d}\,\log_2(b/a)}, \tag{7}$$

where a_0 is a predefined common source granularity. $R_i(a, b \mid a_0)$ measures the dimension-normalized representation change per doubling of neighborhood size, enabling comparison of relative context sensitivity across tissue locations.

4 Evaluation design

Data and evaluation protocol. We pretrain separate models on MOSTA mouse embryos (Chen et al., 2022) and HESTA human embryos (Pan et al., 2026), using 45 and 68 sections, respectively, and reserve 8 MOSTA sections spanning E9.5–E16.5 and 7 HESTA sections spanning CS12–13 to CS23 for held-out evaluation. We pretrain with 3 discrete granularities, $r \in \{8, 32, 128\}$. For all held-out evaluations, the pretrained encoders and granularity transformation are frozen; preprocessing and task-specific readouts follow Appendices A, C, and G.

Representation evaluation. We evaluate ZoomST against pretrained foundation models and non-pretrained task-specific methods on four complementary tasks: anatomical clustering, spatial-neighborhood preservation, developmental trajectory recovery, and developmental-stage prediction.

For the pretrained baselines, we use their released pretrained checkpoints and keep their encoders frozen during evaluation. Anatomical clustering assesses the recovery of annotated tissue organization, while spatial-neighborhood preservation measures the retention of the 32 nearest neighbors defined in physical space. Developmental information is further evaluated through diffusion-pseudotime–stage correlation and leave-one-section-out linear stage prediction.

We compare 3 ZoomST representation variants: **Molecular**, **Spatial**, and **Joint**, where Joint concatenates the molecular and spatial representations at each granularity. Unless otherwise stated, each variant is formed by concatenating its representations across all 3 training granularities. Ablations use the same total number of training updates while removing cross-granularity supervision (No Cross) or, on MOSTA, retaining only coarse-support supervision (CS).

Granularity transformation and response evaluation. On held-out MOSTA sections, we evaluate the generalization of the learned transformation by predicting molecular representations at both trained and unseen target granularities from e_8. Predictions are compared with representations directly encoded at the corresponding granularities. We further assess context sensitivity by comparing predicted responses with changes in directly encoded representations as the neighborhood expands (Appendices D–E).

5 Results

5.1 Pretrained ZoomST representations recover anatomical and developmental organization

We first evaluate representations inferred by pretrained ZoomST on held-out tissue sections through anatomical clustering, spatial-neighborhood preservation, and developmental trajectory recovery. Molecular, Spatial, and Joint representations are compared with pretrained foundation models and non-pretrained task-specific methods under the shared evaluation protocols.

Table 1: **Tissue mapping and neighbor retention on eight held-out MOSTA sections. (a)** AMI; Avg.: mean across sections. **(b)** NR: mean physical 32-neighbor retention. Bold marks the top three scores per column. Protocols: Appendix C.2.

		(a) AMI ↑									(b) R_{32} ↑
Group	Method	E9.5	E10.5	E11.5	E12.5	E13.5	E14.5	E15.5	E16.5	Avg.	NR
ZoomST views	**Joint**	**0.547**	**0.549**	**0.628**	**0.613**	**0.578**	**0.563**	**0.626**	0.513	**0.577**	**0.527**
	Molecular	0.486	0.516	**0.608**	**0.614**	0.541	0.556	**0.625**	0.497	0.555	**0.502**
	Spatial	**0.556**	0.534	**0.615**	**0.607**	**0.579**	0.558	**0.614**	**0.532**	**0.574**	**0.523**
Frozen pretrained	Novae	0.371	0.390	0.444	0.447	0.432	0.450	0.494	0.449	0.435	0.313
	SToFM	0.211	0.262	0.291	0.307	0.334	0.407	0.405	0.331	0.319	0.033
Task-Specific	GraphST	0.538	0.534	0.502	0.509	0.386	0.453	0.422	0.403	0.468	0.268
	BANKSY	0.540	**0.539**	0.573	0.583	**0.593**	**0.572**	0.614	**0.526**	**0.567**	0.483
	MENDER	0.530	0.531	0.561	0.582	0.573	0.535	0.600	0.509	0.553	0.354
	STAGATE	**0.556**	**0.544**	0.559	0.585	0.544	**0.579**	0.604	**0.540**	0.564	0.354

Across all eight held-out MOSTA sections, ZoomST Joint representations achieve the highest mean AMI among all compared methods. In particular, ZoomST outperforms the evaluated pretrained models, Novae and SToFM, on every individual section (Table 1a; Figure C2). All three ZoomST representation variants also show strong performance in the spatial-neighborhood preservation task, with Joint representations achieving the highest mean neighbor retention score (Table 1b; Figure C4). Molecular and Spatial representations provide complementary information, while their combination in the Joint representation reduces both extra and missed assignments and improves anatomical delineation (Figure C3).

An independently pretrained ZoomST model on the human HESTA dataset shows consistent generalization across seven held-out sections, outperforming the other evaluated methods in both anatomical clustering and spatial-neighborhood preservation tasks (Table 2).

Table 2: **Tissue mapping and neighbor retention on seven held-out HESTA sections.** Readouts, abbreviations, and highlighting follow Table 1. Protocols: Appendix C.2.

		(a) AMI ↑								(b) R_{32} ↑
Group	Method	CS12-13	CS14-15	CS17	CS18	CS19	CS20	CS23	Avg.	NR
ZoomST views	**Joint**	**0.556**	**0.583**	**0.517**	**0.569**	**0.536**	**0.593**	**0.550**	**0.558**	**0.495**
	Molecular	**0.550**	**0.556**	**0.495**	**0.553**	**0.524**	**0.553**	**0.527**	**0.537**	**0.458**
	Spatial	**0.556**	**0.565**	**0.523**	**0.563**	**0.548**	**0.574**	**0.533**	**0.552**	**0.494**
Frozen pretrained	Novae	0.338	0.331	0.376	0.283	0.382	0.419	0.423	0.365	0.215
	SToFM	0.170	0.205	0.270	0.175	0.289	0.257	0.320	0.241	0.017
	scGPT-spatial	0.342	0.384	0.416	0.366	0.447	0.438	0.490	0.412	0.040
	TERRA Cell	0.228	0.279	0.358	0.267	0.382	0.338	0.393	0.321	0.021
	TERRA Nbh.	0.327	0.410	0.417	0.389	0.464	0.438	0.478	0.418	0.226
	TERRA S-cell	0.318	0.416	0.417	0.411	0.471	0.457	0.486	0.425	0.223
	CellWorld	0.285	0.228	0.322	0.252	0.289	0.321	0.376	0.296	0.030
Task-Specific	DECIPHER Ctr.	0.402	0.416	0.476	0.331	0.483	0.465	0.469	0.434	0.037
	DECIPHER Ctx.	0.475	0.523	0.414	0.455	0.343	0.424	0.422	0.436	0.352

We further assess developmental trajectory recovery within Brain by relating diffusion pseudotime to the observed stage order. Joint representations achieve the highest DPT–stage correlation among methods evaluated across all stages in both MOSTA and HESTA (Figure 2).

Together, these results show that pretrained ZoomST recovers anatomical and local spatial organization within held-out sections and developmental organization across stages.

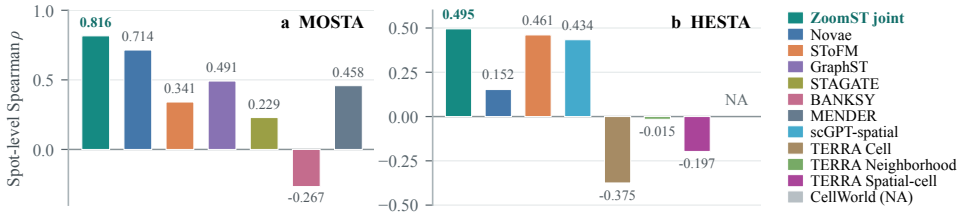

Figure 2: **Developmental ordering on MOSTA and HESTA Brain.** Spot-level DPT–stage Spearman correlations in (a) MOSTA and (b) HESTA. NA: no complete-stage readout. Protocols: Appendix C.3.

5.2 Cross-granularity supervision enriches developmental information

Having established the generalization of pretrained ZoomST, we next isolate the contribution of cross-granularity supervision. We compare ZoomST with a No Cross model trained for the same number of updates but without cross-granularity supervision, and evaluate Joint representations separately at $r \in \{8, 32, 128\}$ on held-out MOSTA and HESTA sections. Cross-granularity supervision preserves anatomical clustering across granularities and improves fine-granularity AMI in seven of eight MOSTA and all seven HESTA sections (Table 3a; Table C3).

We next examine whether the developmental organization observed above depends on cross-granularity supervision, and whether it also supports developmental-stage prediction. Relative to No Cross, ZoomST shows stronger developmental organization particularly at broader granularities. In MOSTA, DPT–stage correlation at $r = 128$ increases from -0.239 to 0.354, while HESTA shows higher correlations at all three granularities. For developmental-stage prediction, ZoomST achieves lower leave-one-section-out MAE at every granularity in both atlases, with the largest reductions at $r = 128$ (Table 3b).

On MOSTA, adding JointMMD to coarse-support supervision further improves DPT–stage correlation and reduces developmental-stage prediction MAE at all three granularities (Appendix F). Together, these results show that cross-granularity supervision enriches developmental information, particularly in broad-context representations, while retaining anatomical utility.

Table 3: **Cross-granularity learning across mouse and human atlases.** **(a)** Mean Joint AMI across eight MOSTA or seven HESTA held-out sections. **(b)** Brain DPT–stage correlation and stage-prediction MAE. Bold marks better paired values, including ties. Protocols: Appendix C.2–C.3.

		(a) Anatomical clustering			(b) Developmental information					
		Mean AMI ↑			Spot-level ρ ↑			Stage MAE ↓		
Atlas	Model	8	32	128	8	32	128	8	32	128
MOSTA	ZoomST	**0.585**	**0.571**	0.538	0.868	**0.822**	**0.354**	**0.803**	**0.715**	**0.734**
	No Cross	0.570	0.562	**0.539**	**0.903**	0.738	-0.239	0.936	0.853	0.929
HESTA	ZoomST	**0.554**	**0.558**	**0.538**	**0.500**	**0.461**	**0.420**	**1.505**	**1.424**	**1.445**
	No Cross	0.536	0.550	0.530	0.426	0.439	0.335	1.629	1.512	1.583

5.3 Continuous cross-granularity transformations reveal tissue-context dependence

ZoomST learns how representations of the same tissue location change as spatial context expands. We test whether these transformations link representations across granularities, extend to unseen granularities, and reveal relative tissue-context sensitivity.

Granularity-dependent anatomical organization. Fine and coarse spatial contexts resolve different anatomical structures. Across held-out MOSTA sections, broader context favors locally homogeneous regions, whereas finer context favors boundary-rich regions, where local neighborhoods more often cross anatomical boundaries (Figure 3; Appendix C.1). Lung and Smooth muscle illustrate these two patterns (Figure C1).

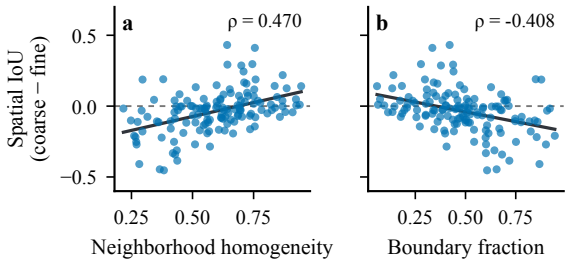

Figure 3: **Granularity preferences depend on tissue structure.** Points are annotated regions in held-out sections. Intersection over union (IoU) measures overlap between recovered and annotated regions; positive Spatial $\mathrm{IoU}_{128} - \mathrm{IoU}_{8}$ favors broader context. (a) Mean fraction of neighbors sharing the focal label. (b) Fraction of positions adjacent to other labels. Solid lines: linear fits; dashed lines: equal IoU. Definitions: Appendix C.1.

Continuous transformations generalize representation changes to unseen granularities. On held-out MOSTA sections, applying the learned transformation reduces the discrepancy between source and directly encoded target representations for all three granularity relations used during pretraining (Figure 4a; Appendix D), showing that granularity-specific representations are linked within a shared representation space. This relationship extends beyond the training granularities: although ZoomST is trained only at $r \in \{8, 32, 128\}$, it generates intermediate representations at six unseen granularities $r \in \{12, 16, 24, 48, 64, 96\}$ that are closer to direct encodings in all 48 section–granularity conditions, reducing target discrepancy by 21.96% on average relative to the untransformed source representation (Figure 4b; Appendix E.2). We further find that directly encoded representations at these unseen granularities contain more developmental information and show better predictive power to developmental stage than the e_8 representation. Transforming e_8 to these target granularities reproduces this improvement, reducing mean prediction MAE by 5.9% across the six targets (Figure G1; Appendix G). Additional comparisons with an alternative transformation parameterization further support the neural ODE design used in ZoomST (Figure E1; Appendix E.1).

Cross-granularity response reflects tissue-context sensitivity. Because the learned transformations encode representation changes across granularities, we use them to quantify relative tissue-

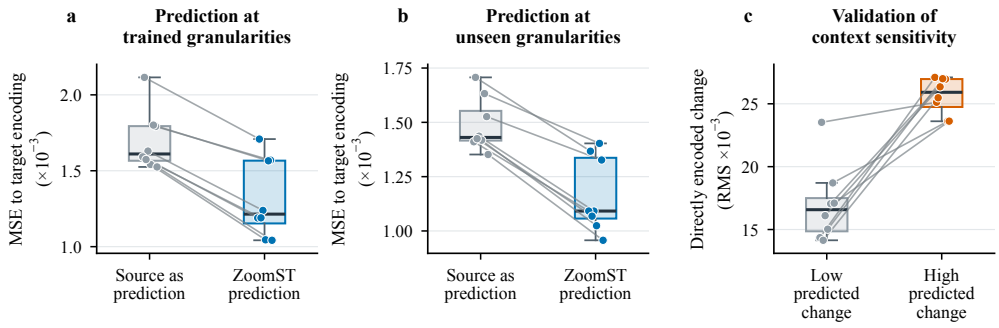

Figure 4: **Zero-shot prediction and context sensitivity.** (a,b) MSE to directly encoded target representations for the source used as prediction and ZoomST: (a) trained pairs; (b) unseen targets. (c) Directly encoded changes in the lowest/highest predicted-change quartiles. Connected points show the same held-out section, averaged over pairs (a), targets (b), or intervals (c); 8 sections are shown. Protocols: Appendices D–E.

context sensitivity. We define the cross-granularity response as the normalized representation change per doubling of neighborhood size (Equation 7). Across all eight held-out sections, high-response groups show larger mean directly encoded changes than low-response groups after averaging four doubling intervals (Figure 4c; Appendix E.3). In training-section examples, higher responses track changes in surrounding tissue composition in Facial nerve; in Brain, they track directly encoded changes even when neighborhoods remain entirely Brain-labeled (Figure 5; Appendix E.3).

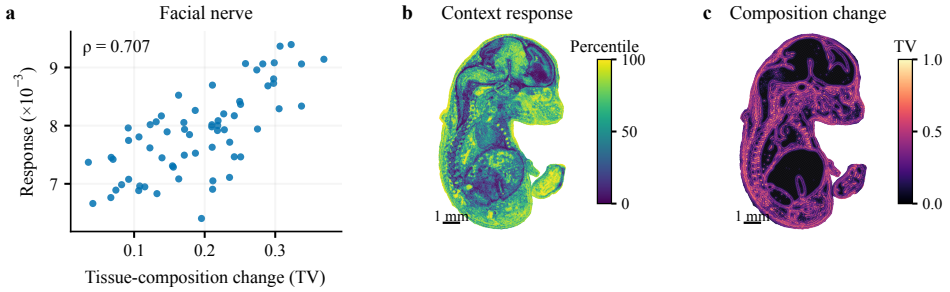

Figure 5: **Context sensitivity and tissue composition.** (a) Facial nerve response versus tissue-composition change (64 → 128; 64 locations, E14.5 E1S1). Composition change is total variation (TV) between tissue-label proportions in the two neighborhoods. (b,c) E16.5 E1S1 maps of path-averaged response and composition change over 8–128. Scale bars: 1 mm. Definitions: Appendix E.3.

6 Discussion and Conclusion

ZoomST treats relationships across spatial granularities as a learnable component of spatial transcriptomic pretraining. Molecular–spatial learning provides granularity-specific representations, while soft distributional equivariance links them through structured cross-granularity transformations. Across mouse and human developmental atlases, pretrained representations generalize to held-out sections for anatomical and developmental organization, while fine and coarse spatial contexts retain complementary anatomical characteristics. The learned continuous transformations extend beyond the discrete training granularities to unseen intermediate scales, and the resulting cross-granularity responses characterize relative tissue-context sensitivity associated with changes in surrounding tissue composition and representation changes within anatomical compartments. By exploiting nested neighborhoods around the same tissue location as a self-supervised signal, ZoomST makes spatial granularity adjustable after pretraining and provides an axis for analyzing tissue-context dependence.

Reproducibility Statement

The appendices specify data splits, preprocessing, model architecture, optimization, and evaluation protocols. We provide an anonymized code archive as supplementary material, including the ZoomST implementation, preprocessing and training scripts, configuration files, evaluation utilities, and usage instructions. Reported figures and tables are derived from saved model outputs using the stated readouts. Downstream K-means clustering is repeated with three seeds: 42, 43, and 44.

Ethics Statement

This work uses published MOSTA mouse embryo and HESTA human embryo spatial-transcriptomics data and their associated anatomical labels, following the original data access and use conditions. The model is intended for biological research.

AI-Use Disclosure

Generative AI tools assisted with language polishing, translation, and consistency checks across the text, figures, tables, and reported results. They also supported literature searches and code preparation. The authors retain responsibility for the final manuscript and all reported findings.

References

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15619–15629, 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Assran_Self-Supervised_Learning_From_Images_With_a_Joint-Embedding_Predictive_Architecture_CVPR_2023_paper.html.

Adrien Bardes, Jean Ponce, and Yann LeCun. VICReg: Variance-invariance-covariance regularization for self-supervised learning. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=xm6YD62D1Ub.

Sebastian Birk, Mohammad Vali Sanian, Amirhossein Vahidi, Samuel Ogden, Daniyal J. Jafree, Adib Miraki Feriz, Carlo Leonardi, Arpit Merchant, Zichen He, Lloyd Steele, Adam Boxall, Marta Rosa Sallese, Jonas Maaskola, Ciro Ramirez-Suastegui, Sevda Ogut, Koen Rademaker, MS Vijayabaskar, Batuhan Cakir, Sergio Marco Salas, Valentina Lorenzi, Josephine M. Bryant, Cecilia Kyany'a, James O. Jones, Raheleh Rahbari, Arti Mala Raghubar, Grant D. Stewart, Benjamin Rumney, Catherine Tudor, Minal Patel, Jasmine Halliwell, Hon Man Chan, Tong Li, Heather Stanley, April Rose Foster, Fani Memi, Kenny Roberts, Andrew L. Trinh, Elizabeth Tuck, Tannia Gracia, Shreya Rai, David J. Adams, Simone Webb, Martin Prete, Roser Vento-Tormo, Mats Nilsson, Heiko Lickert, Fabian J. Theis, Amin Ardestani, James J. Sun, Foad J. Rouhani, Lassi Paavolainen, Menna R. Clatworthy, Omer Ali Bayraktar, Muzlifah Haniffa, Thomas Mitchell, Mostafa Bakhti, and Mo Lotfollahi. Multi-scale modeling of human tissues from spatial transcriptomics with TERRA. *bioRxiv*, 2026. doi: 10.64898/2026.07.29.741565. URL https://www.biorxiv.org/content/10.64898/2026.07.29.741565v1. 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(12):2539–2550, 2025. doi: 10.1038/s41592-025-02899-6.

Ao Chen, Sha Liao, Mengnan Cheng, Kailong Ma, Liang Wu, Yiwei Lai, Xiaojie Qiu, Jin Yang, Jiangshan Xu, Shijie Hao, Xin Wang, Huifang Lu, Xi Chen, Xing Liu, Xin Huang, Zhao Li, Yan Hong, Yujia Jiang, Jian Peng, Shuai Liu, Mengzhe Shen, Chuanyu Liu, Quanshui Li, Yue Yuan, Xiaoyu Wei, Huiwen Zheng, Weimin Feng, Zhifeng Wang, Yang Liu, Zhaohui Wang, Yunzhi Yang, Haitao Xiang, Lei Han, Baoming Qin, Pengcheng Guo, Guangyao Lai, Pura Muñoz-Cánoves, Patrick H. Maxwell, Jean Paul Thiery, Qing-Feng Wu, Fuxiang Zhao, Bichao Chen,

Mei Li, Xi Dai, Shuai Wang, Haoyan Kuang, Junhou Hui, Liqun Wang, Ji-Feng Fei, Ou Wang, Xiaofeng Wei, Haorong Lu, Bo Wang, Shiping Liu, Ying Gu, Ming Ni, Wenwei Zhang, Feng Mu, Ye Yin, Huanming Yang, Michael Lisby, Richard J. Cornall, Jan Mulder, Mathias Uhlén, Miguel A. Esteban, Yuxiang Li, Longqi Liu, Xun Xu, and Jian Wang. Spatiotemporal transcriptomic atlas of mouse organogenesis using DNA nanoball-patterned arrays. *Cell*, 185(10): 1777–1792.e21, 2022. doi: 10.1016/j.cell.2022.04.003.

Ricky T. Q. Chen, Yulia Rubanova, Jesse Bettencourt, and David K. Duvenaud. Neural ordinary differential equations. In *Advances in Neural Information Processing Systems*, volume 31, 2018. URL https://papers.nips.cc/paper_files/paper/2018/hash/69386f6bb1dfed68692a24c8686939b9-Abstract.html.

Haotian Cui, Chloe Wang, Hassaan Maan, Kuan Pang, Fengning Luo, Nan Duan, and Bo Wang. scGPT: toward building a foundation model for single-cell multi-omics using generative AI. *Nature Methods*, 21(8):1470–1480, 2024. doi: 10.1038/s41592-024-02201-0.

Rumen Dangovski, Li Jing, Charlotte Loh, Seungwook Han, Akash Srivastava, Brian Cheung, Pulkit Agrawal, and Marin Soljačić. Equivariant contrastive learning. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=gKLAAfiytI.

Kangning Dong and Shihua Zhang. Deciphering spatial domains from spatially resolved transcriptomics with an adaptive graph attention auto-encoder. *Nature Communications*, 13(1):1739, 2022. doi: 10.1038/s41467-022-29439-6.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(25):723–773, 2012. URL https://www.jmlr.org/papers/v13/gretton12a.html.

Laleh Haghverdi, Maren Büttner, F. Alexander Wolf, Florian Buettner, and Fabian J. Theis. Diffusion pseudotime robustly reconstructs lineage branching. *Nature Methods*, 13(10):845–848, 2016. doi: 10.1038/nmeth.3971.

Karel Lenc and Andrea Vedaldi. Understanding image representations by measuring their equivariance and equivalence. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 991–999, 2015. URL https://openaccess.thecvf.com/content_cvpr_2015/html/Lenc_Understanding_Image_Representations_2015_CVPR_paper.html.

Yuxiang Lin, Ling Luo, Ying Chen, Xushi Zhang, Zihui Wang, Wenxian Yang, Mengsha Tong, and Rongshan Yu. ST-Align: A multimodal foundation model for image-gene alignment in spatial transcriptomics. *arXiv preprint arXiv:2411.16793*, 2024. doi: 10.48550/arXiv.2411.16793. URL https://arxiv.org/abs/2411.16793.

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. URL https://arxiv.org/abs/2608.06659.

Yahui Long, Kok Siong Ang, Mengwei Li, Kian Long Kelvin Chong, Raman Sethi, Chengwei Zhong, Hang Xu, Zhiwei Ong, Karishma Sachaphibulkij, Ao Chen, Li Zeng, Huazhu Fu, Min Wu, Lina Hsiu Kim Lim, Longqi Liu, and Jinmiao Chen. Spatially informed clustering, integration, and deconvolution of spatial transcriptomics with GraphST. *Nature Communications*, 14(1):1155, 2023. doi: 10.1038/s41467-023-36796-3.

Jiexue Pan, Yuejiao Li, Zhongliang Lin, Qing Lan, Ying Zhang, Huixi Chen, Shengwei Sui, Man Zhai, Gaochen Zhang, Yi Cheng, Yunhui Tang, Qingchen Wang, Yue Xu, Guoling Li, Chunyan Xu, Haoqi Yan, Yiting Mao, Xingxia Wang, Hao Li, Yiping Zhu, Qinfang Chen, Yichun Guan, Nan Meng, Chang Wang, Haiqian Lu, Xiangjuan Li, Tingting Zheng, Xiaoying Yao, Tingyu Yang, Xuan Chen, Qiuyu Qin, Bin Jiang, Yuxing Ren, Xinmei Liu, Yuxin Zhang, Minghui Yu, Lifang Wang, Yanrong Wei, Meiqi Luo, Ji Nancuo, Fuhe Ma, Ziwei Wang, Zhihua Ou, Ying Lei, Xin Jin, Jianzhong Sheng, Congjian Xu, Yanting Wu, Chenming Xu, Lijian Zhao, Hongbo Yang, Ya Gao,

Guolian Ding, Xun Xu, and Hefeng Huang. Spatiotemporal transcriptome atlas of human embryos after gastrulation. *Nature*, 654(8119):751–761, 2026. doi: 10.1038/s41586-026-10545-0.

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, Shigemi Kinoshita, Nikolay Samusik, Yury Goltsev, and Garry P. Nolan. 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.

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. doi: 10.1038/s41588-024-01664-3.

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(12):2525–2538, 2025. doi: 10.1038/s41592-025-02814-z.

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. doi: 10.1038/s41586-023-06139-9.

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, 2024. doi: 10.1038/s41588-023-01588-4.

Daniel E. Worrall and Max Welling. Deep scale-spaces: Equivariance over scale. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://papers.nips.cc/paper_files/paper/2019/hash/f04cd7399b2b0128970efb6d20b5c551-Abstract.html.

Chen-Rui Xia, Zhi-Jie Cao, and Ge Gao. DECIPHER for learning disentangled cellular embeddings in large-scale heterogeneous spatial omics data. *Nature Communications*, 16(1):7991, 2025. doi: 10.1038/s41467-025-63140-8. URL https://www.nature.com/articles/s41467-025-63140-8.

Zhiyuan Yuan. MENDER: fast and scalable tissue structure identification in spatial omics data. *Nature Communications*, 15(1):207, 2024. doi: 10.1038/s41467-023-44367-9.

Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabás Póczos, Ruslan Salakhutdinov, and Alexander J. Smola. Deep Sets. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://papers.nips.cc/paper_files/paper/2017/hash/f22e4747da1aa27e363d86d40ff442fe-Abstract.html.

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*, volume 267 of *Proceedings of Machine Learning Research*, pp. 77589–77607. PMLR, 2025. URL https://proceedings.mlr.press/v267/zhao25p.html.

A DATA AND PREPROCESSING

Data split. MOSTA supplies 53 mouse embryo sections. We reserved one section from each developmental stage from E9.5 to E16.5 in advance for evaluation and used the remaining 45 sections for pretraining. HESTA supplies 75 sections with raw counts: 68 for independent human-model pretraining and seven held out for evaluation. Table A1 lists both held-out sets; Table A2 summarizes data use and aggregation for each analysis.

Table A1: **Held-out section inventory.** Every listed section is excluded from both pretraining stages.

Stage	Held-out section
MOSTA	
E9.5	E9.5 E2S4
E10.5	E10.5 E2S1
E11.5	E11.5 E1S4
E12.5	E12.5 E2S1
E13.5	E13.5 E1S4
E14.5	E14.5 E2S2
E15.5	E15.5 E2S1
E16.5	E16.5 E2S13
HESTA	
CS12–13	CS12–13 E1S3
CS14–15	CS14–15 E2S1
CS17	CS17 E1S9
CS18	CS18 E1S2
CS19	CS19 E1S1
CS20	CS20 E1S6
CS23	CS23 E1S3

Input preprocessing. MOSTA inputs use the original `count` layer and a 30,235-entry gene vocabulary. Existing center tokens are inherited from the prepared source data. Neighborhood profiles average normalized expression over the full gene space before selecting up to 512 tokens and assigning 31 nonzero expression bins, with bin zero reserved for padding. The inherited preprocessing uses library normalization to 10^4, $\log(1 + x)$, and section-specific nonzero gene medians; genes with fewer than ten nonzero observations use a fallback median. Target sections supply their own gene statistics and use the inherited vocabulary. Whole-section readout standardization, PCA, and clustering are fitted on target-section positions; method-specific settings are given in Appendix C.2.

HESTA preprocessing starts from `layers["counts"]`. Within each section, genes present at fewer than three positions are filtered, counts are normalized to 10^4 per position and log1p-transformed, then scaled by nonzero gene medians. Genes with fewer than ten nonzero values use the median of the stable gene medians as fallback. The 57,350-entry vocabulary, including padding, is constructed from gene names across the 75 count-bearing sections. Center and neighborhood tokens are built with the same top-512 and expression-bin scheme; neighborhood expression is averaged before token selection. Each held-out section supplies its own preprocessing statistics. Mouse and human models use separate vocabularies and weights.

Relative coordinates are centered at the focal position and scaled by the maximum distance within its 128-neighbor window, then sliced to smaller supports. Granularity counts observations; its physical extent depends on the sampled tissue. Maps preserve the measured lattice and leave missing positions empty, without label interpolation.

External pretrained models. Novae uses the public mouse-0 weights, and SToFM uses its public checkpoint with the specified mouse-to-human gene mapping. The MOSTA sections held out from ZoomST training appear in the SToCorpus-88M inventory, so SToFM may have been exposed to our zero-shot evaluation data.

Table A2: **Data split and evaluation protocol.** These conditions apply to the corresponding evaluations throughout the paper.

Evaluation	Data use and aggregation
Held-out sections	Table A1 lists eight MOSTA and seven HESTA sections evaluated with frozen pretrained encoders. Input preprocessing and target-fitted readouts are specified above and in Appendix C.2.
Anatomical analyses	Regional IoU and neighborhood geometry use all 43 regions in the eight held-out sections, with annotation-balanced clustering fits and full-section scoring (Appendix C.1). E14.5 E1S1 (granularity and Facial nerve cases) and E16.5 E1S1 (state and context-response cases) belong to the training set. The E12.5 E2S1 whole-section map and E16.5 E2S13 Brain mapping display use held-out sections.
Developmental ordering	Stage-balanced Brain observations from the held-out sections evaluate pseudotime–stage agreement. Sampling, shared baseline fits, and spot-level correlation are specified in Appendix C.3.
Stage prediction	Frozen Brain representations enter a linear readout. Each fold fits feature standardization and regression on the other sections, using their stage labels, then evaluates one entire section. MAE averages sections equally (Appendix C.3).
Predicted-state readout	Predicted molecular states at six unseen granularities enter a shared age readout fitted on directly encoded states at trained granularities from the other seven MOSTA sections. Evaluation uses fixed Brain positions and equal section weights (Appendix G).
Response summaries	High–low response comparisons use the eight held-out sections. Anatomical cases use the training sections identified above; directly encoded changes and neighborhood composition provide evaluation references.
Objective ablation	MOSTA No Cross, CS, and ZoomST share initialization, training data, and training budget, differing only in their loss terms. Frozen evaluation follows the held-out protocols in Section 5.2 and Appendix F.
Aggregation	Section summaries receive equal weight. Position-level distributions describe within-section variation; queries, projections, shuffles and batches specify sampling and numerical averaging.

B Architecture and optimization

B.1 Backbone and within-granularity learning

Both representation views have width 256. The position-wise and pseudobulk encoders each use two Transformer layers and eight attention heads, with separate Transformer parameters and shared gene and expression-bin embeddings. The spatial-context Transformer has four layers and eight attention heads and is shared across granularities. The DeepSets maps ϕ and ρ each use layer normalization and a two-layer multilayer perceptron with hidden width 512, GELU activation, and dropout 0.1. The same maps are applied at each granularity.

For clarity, the pooled profile and spatial set readout are

$$m_{i,r} = r^{-1} \sum_{j \in S_{i,r}} \bar{x}_j, \qquad g_{i,r} = \rho\left(r^{-1} \sum_{j \in S_{i,r}} \phi(H_{ij,r}) \right). \tag{8}$$

The base objective combines reconstruction with same-support view alignment:

$$\mathcal{L}_{\text{base}} = \mathcal{L}_{\text{rec}} + \frac{\lambda_{\text{V}}}{|\mathcal{R}|} \sum_{r \in \mathcal{R}} \mathcal{L}_{\text{VICReg}}(e_r, g_r). \tag{9}$$

Here e_r, g_r collect the sampled centers at granularity r, and $\mathcal{L}_{\text{VICReg}}$ in Equation 1 denotes the average of these per-granularity losses.

The molecular view averages preprocessed expression over the neighborhood before selecting up to 512 gene tokens and assigning expression bins. There are 31 nonzero expression bins and one padding bin. Reconstruction masks gene and value identities together at 30% of visible nonzero token positions, with at most 154 queries per profile. A shared gene-bin decoder predicts the expression-bin targets, and only masked nonzero targets contribute to the reconstruction loss. With

branch weights one,

$$\mathcal{L}_{\mathrm{rec}} = \mathcal{L}_{\mathrm{center}} + \frac{1}{|\mathcal{R}|} \sum_{r \in \mathcal{R}} \left(\mathcal{L}^{\mathrm{PB}}_{g,r} + \mathcal{L}^{\mathrm{PB}}_{e,r} \right). \tag{10}$$

The neighborhood losses reconstruct the pooled expression profile from the spatial and molecular states, respectively. Within-granularity VICReg is applied directly to these states, with invariance, variance, and covariance weights $(1, 1, 0.04)$ and outer weight $\lambda_{\mathrm{V}} = 0.1$ in Equation 9.

B.2 Shared field and numerical integration

For state $z \in \mathbb{R}^{256}$ and granularity coordinate u, the field is

$$f_\theta(z, u) = W_2 \, \mathrm{GELU}(W_1[\mathrm{LN}(z); u] + b_1) + b_2, \tag{11}$$

where $W_1 \in \mathbb{R}^{128 \times 257}$ and $W_2 \in \mathbb{R}^{256 \times 128}$. The field has 66,560 trainable parameters, including layer normalization. Both W_2 and b_2 are initialized to zero. The transition is evaluated in single precision, and gradients are propagated through the numerical integration steps.

For $u_a = u(a)$ and $u_b = u(b)$, integration uses uniform steps within each query interval, with $n = \max(1, \lceil |u_b - u_a| / 0.25 \rceil)$ and step size $h = (u_b - u_a)/n$. A zero-length interval returns its input. A fourth-order Runge–Kutta step from (u, z) is

$$\begin{aligned} k_1 &= f_\theta(z, u), \\ k_2 &= f_\theta(z + hk_1/2, u + h/2), \\ k_3 &= f_\theta(z + hk_2/2, u + h/2), \\ k_4 &= f_\theta(z + hk_3, u + h), \\ z^+ &= z + \frac{h}{6}(k_1 + 2k_2 + 2k_3 + k_4). \end{aligned} \tag{12}$$

The training intervals $8 \to 32$, $32 \to 128$, and $8 \to 128$ use two, two, and four steps, respectively.

The coordinate satisfies $u = \log_2(r/8)/4$, so the field per unit $\log_2 r$ is $f_\theta/4$. Equation 7 uses the corresponding finite-interval normalization to compare response magnitudes across intervals of different widths.

B.3 Coarse-support regression

For a focal center i and pair (a, b), all b positions in $S_{i,b}$ contribute a source state. Each state $e_{j,a}$ is encoded from the fine neighborhood centered at j. The observed target is $y = \mathrm{sg}(e_{i,b})$, the current encoder's clean coarse state. Write $q_j = T_{a \to b}(e_{j,a})$ and $\bar{q} = b^{-1} \sum_{j \in S_{i,b}} q_j$. The loss for this support decomposes as

$$\frac{1}{bd} \sum_{j \in S_{i,b}} \|q_j - y\|_2^2 = \frac{\|\bar{q} - y\|_2^2}{d} + \frac{1}{bd} \sum_{j \in S_{i,b}} \|q_j - \bar{q}\|_2^2. \tag{13}$$

The cross term vanishes because $\sum_j (q_j - \bar{q}) = 0$. Thus, the objective combines mean-to-target alignment and within-support prediction consensus. Every source is transformed before its error is computed. Losses are averaged within each support and then equally across focal centers and the three granularity pairs. The target is detached, while the predicted states carry gradients to the field and source encoder.

B.4 Projected product-kernel MMD

A fixed projection with orthonormal rows maps each 256-dimensional state to 16 dimensions and is stored with the model. For one granularity pair, define

$$s_i = \Pi \, \mathrm{sg}(e_{i,a}), \qquad q_i = \Pi T_{a \to b}(e_{i,a}), \qquad t_i = \Pi \, \mathrm{sg}(e_{i,b}). \tag{14}$$

With the stop-gradient notation of the main text, the empirical joint measures are

$$\widehat{P}_{ab}^{T,\Pi} = \frac{1}{B}\sum_{i=1}^{B} \delta_{(\Pi\,\mathrm{sg}(e_{i,a}),\,\Pi T_{a\to b}(e_{i,a}))},$$
$$\widehat{P}_{ab}^{\Pi} = \frac{1}{B}\sum_{i=1}^{B} \delta_{(\Pi\,\mathrm{sg}(e_{i,a}),\,\Pi y_{i,b})}, \tag{15}$$

where δ denotes a point mass. With B focal-center samples, the biased empirical statistic used in Equation 5 is

$$\widehat{\mathrm{MMD}}^{\,2}_{ab} = \frac{1}{B^2}\sum_{i,j=1}^{B} \kappa_a(s_i,s_j)\left[\kappa_b(q_i,q_j)+\kappa_b(t_i,t_j)-2\kappa_b(q_i,t_j)\right]. \tag{16}$$

We use multibandwidth radial basis function kernels

$$\kappa_r(v,w) = \frac{1}{3}\sum_{c\in\{0.5,1,2\}} \exp\left(-\frac{\|v-w\|_2^2}{2c^2\sigma_r^2}\right). \tag{17}$$

The source bandwidth σ_a^2 is the median of positive pairwise squared distances among s_i, and the destination bandwidth σ_b^2 is computed from t_i. Bandwidths are detached, and the three granularity-pair statistics are averaged equally.

The source coordinate in the product kernel is held fixed during each gradient calculation. The prediction is computed from the full 256-dimensional source state before projection, so its gradients update the source encoder as well as the transition. Observed coarse states are clean, same-step outputs of the current encoder, detached as regression and MMD targets. The projected statistic is the empirical penalty used to optimize the joint relationship in Equation 4.

B.5 Joint optimization

The reconstruction and view-alignment objective initializes the representations and remains active during cross-granularity training. Both relation coefficients are increased to one during warmup. All three source–target granularity pairs receive equal final weight. Encoder and transition parameters are optimized together; clean encodings supply the relation losses, while reconstruction is evaluated on masked inputs. Pretraining uses molecular measurements and spatial coordinates without anatomical labels.

For MOSTA, the base-pretrained encoder is followed by two joint-training epochs over 45 source sections, totaling 46,400 updates with a global batch of 160 (four devices, 40 centers per device). AdamW uses learning rates 2×10^{-4} for molecular and embedding parameters and 10^{-4} for the spatial context parameters, with respective weight decays 0.02, 0.01, and 0.05. Learning rates and the two relation weights warm up over 1,200 updates; cosine decay has a final learning-rate ratio of 0.1. Gradient norm is clipped at five. Backbone computation uses bfloat16 mixed precision and the ODE uses float32. Evaluation uses the final two-epoch checkpoint.

The No Cross ablation continues training from the epoch-10 base-pretrained initialization for 46,400 updates using reconstruction and within-granularity view alignment without cross-granularity supervision. Including the 161,810 base-pretraining updates, both No Cross and the full model receive 208,210 updates in total. Evaluation uses the completed continuation for each model.

For HESTA, all compared variants receive ten training epochs on the same 68 source sections: eight epochs with the base objective, followed by two epochs with each variant's objective. ZoomST adds coarse-support regression and JointMMD during the final two epochs; No Cross continues with reconstruction and within-granularity view alignment. Evaluation uses each variant's final ten-epoch model.

The HESTA ZoomST continuation uses a global batch of 672 (eight devices, 84 centers per device). AdamW uses learning rates 2×10^{-4} for molecular and embedding parameters and 10^{-4} for spatial context parameters, with respective weight decays 0.02, 0.01, and 0.05; normalization and bias terms receive no weight decay. Learning rates and the two relation weights warm up over 500 updates, followed by cosine learning-rate decay to a final ratio of 0.1. Gradient norm is clipped at five. Backbone computation uses bfloat16 mixed precision and the ODE uses float32.

C TISSUE REPRESENTATION EVALUATION

C.1 REGIONAL SCALE PREFERENCES AND ANATOMICAL CASES

Regional readout. The regional analysis uses all annotated regions in each of the eight held-out sections: 43 distinct regions and 137 section–region observations. ZoomST Spatial at granularities 8, 32, and 128 and the six external methods use the representations and full-section feature transforms specified in Appendix C.2. Each ZoomST Spatial representation is standardized per feature dimension and projected separately to 32 principal components. External methods retain their original feature preprocessing, including BANKSY's weighted representation with $\lambda = 0.8$ and no second feature standardization. These feature coordinates remain fixed during the regional readout.

For each section and clustering seed, we sample without replacement the same number of positions from every annotated class, using the size of the smallest class. All methods and granularities share these sampled indices and their shuffled order. MiniBatchKMeans fits these balanced positions with the number of clusters set to the section's annotation count, k-means++ initialization, ten initializations, batch size 4,096, at most 300 iterations, and no-improvement limit ten. It then predicts all original positions in the section, including the fitting positions. A global one-to-one Hungarian assignment maximizes cluster–annotation overlap on the full section, and region-wise IoU is computed from all positions. Sampling and clustering are repeated with the three downstream seeds specified in Appendix C.2.

We first average the three seeds within each section–region observation, then report the mean and sample SD across the sections containing that region. A region present in one section has one reported value. Table C1 presents selected examples of coarse- and fine-context preferences.

Neighborhood geometry and scale preference. Let y_i be the anatomical label at measured position i, and $N_k(i)$ its k nearest other positions by Euclidean distance in the section's spatial coordinates. Neighbor searches use all measured positions in the section. For the positions R belonging to one annotated region, we define

$$H_{128}(R) = \frac{1}{|R|} \sum_{i \in R} \frac{1}{128} \sum_{j \in N_{128}(i)} \mathbf{1}[y_j = y_i], \tag{18}$$

$$B_8(R) = \frac{1}{|R|} \sum_{i \in R} \mathbf{1}[\exists j \in N_8(i) : y_j \neq y_i]\,. \tag{19}$$

H_{128} measures local annotation homogeneity; B_8 is the fraction of positions touching a different annotated region within their eight-neighbor window. The geometry is computed before any clustering-fit subsampling. For each section–region observation, the coarse-scale gain is $\Delta\mathrm{IoU} = \mathrm{IoU}_{128} - \mathrm{IoU}_8$, using the seed-averaged Spatial scores. Figure 3 shows all 137 observations and pooled least-squares trend lines. Spearman correlations use average ranks for tied values. We report both the pooled correlations (0.470 for H_{128} and -0.408 for B_8) and the equal-weight mean of the eight within-section correlations (0.409 and -0.337). Positive homogeneity correlations and negative boundary correlations each occur in seven of eight sections.

Source-section anatomical illustration. The Lung and Smooth muscle maps in Figure C1 illustrate how context interacts with regional shape. In their fixed visceral domain, Lung IoU increases from 0.894 at granularity 8 to 0.955 at 128. Smooth muscle's narrow bands and rings are better separated at granularity 8, with IoU 0.779 compared with 0.580 at 128.

The anatomical granularity analysis uses E14.5 E1S1 and the MOSTA model defined in Section 4. Its 102,519 measured positions supply neighborhood context; scoring uses the fixed union of Heart (2,952), Lung (2,011), Liver (6,015), GI tract (1,260), Smooth muscle (1,737), and Pancreas (425). Each output is standardized and projected by full-SVD PCA 32 on these 14,400 positions. MiniBatchKMeans uses six clusters, batch size 2,048, three initializations, at most 200 iterations, no-improvement limit 30, and seed 42. Atlas annotations define the domain, cluster count, post-hoc maximum-overlap one-to-one assignment, and scores. They do not fit the feature transform or clustering.

Table C1: **Distinct anatomical structures favor coarse or fine Spatial context.** Selected examples from the 43-region benchmark. The upper group favors granularity 128; the lower group favors granularity 8. In each example, the preferred ZoomST granularity also attains the highest IoU among the compared methods. Values are means with sample SD below across n sections, after averaging three clustering seeds per section; single-section regions have one value. H_{128} is neighborhood homogeneity. Bold marks the highest IoU across columns; underlining marks the best ZoomST granularity.

			ZoomST Spatial			Frozen pretrained		Target-fitted			
Region	n	H_{128}	8	32	128	Novae	SToFM	GraphST	BANKSY	MENDER	STAGATE
Coarse-context examples											
Spinal cord	4	0.780	0.390 ±0.108	0.404 ±0.065	**<u>0.489</u>** ±0.117	0.311 ±0.172	0.198 ±0.082	0.436 ±0.166	0.398 ±0.119	0.402 ±0.089	0.426 ±0.154
Kidney	1	0.755	0.286	0.425	**<u>0.697</u>**	0.045	0.136	0.639	0.007	0.007	0.112
Olfactory epithelium	2	0.750	0.540 ±0.372	0.572 ±0.196	**<u>0.616</u>** ±0.061	0.412 ±0.027	0.040 ±0.020	0.528 ±0.301	0.552 ±0.299	0.463 ±0.433	0.194 ±0.266
Fine-context examples											
Mucosal epithelium	4	0.518	**<u>0.302</u>** ±0.178	0.260 ±0.112	0.188 ±0.122	0.221 ±0.174	0.076 ±0.022	0.270 ±0.189	0.196 ±0.153	0.212 ±0.160	0.290 ±0.242
Blood vessel	2	0.439	**<u>0.347</u>** ±0.254	0.306 ±0.110	0.138 ±0.193	0.127 ±0.039	0.182 ±0.227	0.001 ±0.001	0.134 ±0.153	0.218 ±0.231	0.296 ±0.307
Surface ectoderm	2	0.428	**<u>0.399</u>** ±0.059	0.304 ±0.063	0.312 ±0.029	0.154 ±0.054	0.174 ±0.016	0.256 ±0.133	0.301 ±0.051	0.226 ±0.157	0.342 ±0.068

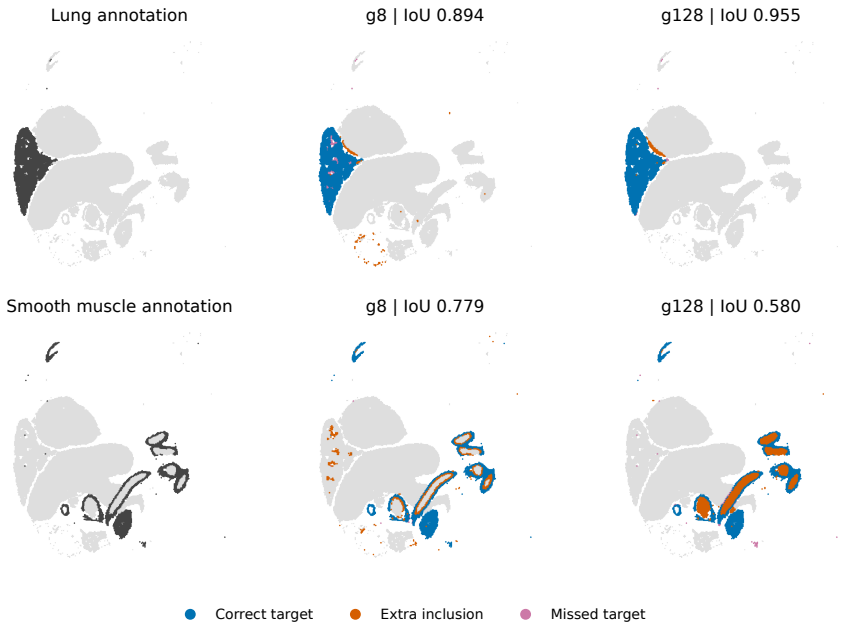

Figure C1: **Broad Lung regions and narrow Smooth muscle favor different granularities.** Annotation, g_8, and g_{128} assignments use identical positions and readouts in MOSTA E14.5 E1S1, a training-section illustration. Blue marks correct assignments, orange extra assignments, and magenta missed targets. Both granularities use the same 14,400-position visceral domain and clustering readout.

Scores are computed over all positions in the fixed evaluation domain. The maps preserve measured coordinates without smoothing or interpolation. Coarse and fine maps share viewing limits and the same readout settings.

For class c, predicted and annotated position sets P_c, T_c within the fixed scoring domain define

$$\mathrm{IoU}_c = \frac{|P_c \cap T_c|}{|P_c \cup T_c|} = \frac{\mathrm{TP}_c}{\mathrm{TP}_c + \mathrm{FP}_c + \mathrm{FN}_c}. \tag{20}$$

TP counts correct assignments, FP extra assignments, and FN missed targets.

Adjusted mutual information (AMI) measures the information shared by cluster and annotation assignments after subtracting chance mutual information and dividing by mean marginal entropy minus that same chance expectation. The adjusted Rand index (ARI) measures chance-corrected agreement in whether pairs of positions share a class. Both are invariant to cluster renaming.

C.2 Whole-section tissue mapping and neighbor retention

The tissue-mapping evaluation uses the MOSTA and HESTA held-out sections listed in Appendix A. Both atlases include cross-method comparisons and the granularity-specific No Cross comparison. Target-fitted methods are identified separately.

The equal-budget controlled ablation compares the final ZoomST and No Cross checkpoints after the same total number of training updates within each atlas; No Cross omits cross-granularity supervision. Both models use the same Joint readout. Optimization and update budgets are specified in Appendix B.

The displayed Molecular View, Spatial View, and Joint refer to three readouts of the same ZoomST model. Molecular features concatenate the three 256-dimensional outputs e_8, e_{32}, e_{128}; spatial features concatenate g_8, g_{32}, g_{128}. Joint features concatenate all six outputs. For MOSTA, Novae uses its public mouse-0 checkpoint; SToFM uses its public checkpoint with the declared mouse-to-human gene mapping. Each ZoomST or public pretrained candidate is independently standardized and projected to 32 dimensions using randomized PCA with seed 42, 32 oversamples, five power iterations, and no whitening. Standardization and PCA fit all eligible target-section positions without annotation values. MiniBatchKMeans uses K clusters, where K is the number of classes in the section's primary MOSTA `annotation` field or HESTA coarse `celltype` field, with k-means++ initialization, ten initializations, batch size 4,096, at most 300 iterations, a ten-step no-improvement limit, and seeds 42, 43, and 44.

Scores use the same eligible positions for every method within each benchmark: all positions for MOSTA and the common valid cohort specified below for the HESTA cross-method comparison. Adjusted mutual information (AMI) and adjusted Rand index (ARI) measure chance-adjusted agreement between the cluster partition and the atlas annotation. One-to-one maximum-overlap Hungarian matching assigns anatomical labels to clusters for the maps and class-specific IoU. For class c, $\mathrm{IoU}_c = \mathrm{TP}_c/(\mathrm{TP}_c + \mathrm{FP}_c + \mathrm{FN}_c)$. AMI and ARI are invariant to the matching. Section-level scores are averaged with equal section weights.

Target-fitted spatial methods. GraphST, BANKSY, MENDER, and STAGATE (Long et al., 2023; Singhal et al., 2024; Yuan, 2024; Dong & Zhang, 2022) use each complete target section independently, with expression and coordinates but no anatomical training labels. Each fixed representation uses the shared clustering protocol above. Tables group these target-fitted methods separately from pretrained models.

GraphST and STAGATE select 3,000 highly variable genes by Seurat-v3 on raw counts, followed by counts-per-10,000 normalization and log1p. GraphST additionally applies noncentered gene scaling clipped at 10. GraphST uses its Stereo-seq three-neighbor graph and a sparse implementation of the native objective, with 600 epochs, learning rate 0.001, and hidden dimension 64. The readout uses its normalized reconstructed-gene output. STAGATE uses the official PyG model, a symmetrized six-neighbor graph with self loops, hidden dimensions 512 and 30, 1,000 epochs, learning rate 0.001, weight decay 0.0001, and gradient clipping at 5. It uses the full section graph without an annotation-derived auxiliary graph.

BANKSY and MENDER normalize all genes to counts per 10,000, apply log1p, and select 3,000 genes by Seurat dispersion. BANKSY uses neighborhood statistics with 15 and 30 neighbors for orders zero and one and scaled-Gaussian decay. The fixed primary weight is $\lambda = 0.8$, the official default for tissue-domain segmentation (Singhal et al., 2024); the documentation explicitly includes Stereo-seq among the technologies for which this setting is recommended.[1] BANKSY's native channel normalization and weighting are preserved, without a second StandardScaler before PCA32. MENDER derives 32 unsupervised expression states using StandardScaler, PCA32, and MiniBatchKMeans with ten initializations; it then aggregates these states over six rings with ring parameter four, including the center, producing 192 features. MENDER's state labels are independent of the anatomical annotation.

The shared randomized-PCA and clustering settings above apply to these features. STAGATE retains its native 30 dimensions; the other methods use PCA32. The comparison shares positions, labels, and downstream algorithms while retaining these method-specific feature definitions.

Human cross-method benchmark. Table 2 uses the independently pretrained human ZoomST model and seven HESTA sections. All methods share 1,129,281 positions with valid representations, excluding six positions outside the common cohort. Novae uses its public human-0 checkpoint with native count normalization, log1p, and unlabeled per-section gene selection and scaling. SToFM uses human gene names directly. scGPT-spatial normalizes raw expression by the whole-section matched nonzero mean and pretrained gene means. TERRA (Birk et al., 2026) provides Cell, Neighborhood, and Spatial-cell outputs using its native shifted-log input, ten-neighbor context, and 2,816-token limit. These encoders use their public weights without target-section updates.

CellWorld-base (Liu et al., 2026) receives raw counts normalized to 10,000 per position, followed by log1p and selection of 300 highly variable genes per section. We use Scanpy's Seurat-dispersion selection with 20 bins and no batch key. Its native tokenizer and normalized-coordinate Hilbert ordering form contexts of 4,096 positions; unknown organ and platform identifiers are set to zero. All native feature construction uses unlabeled section data; the common cohort and shared standardization, PCA32, and clustering readout are then applied. Public models retain their original pretraining corpora; exposure to individual HESTA target sections is not established here.

DECIPHER (Xia et al., 2025) is fitted independently on each target section without anatomical labels, yielding its native 32-dimensional Center and Context representations. Version 0.3.2 selects 2,000 Seurat-v3 highly variable genes from counts, normalizes to 10,000, applies log1p, and scales with clipping at 10. The minimum gene count per position is zero to retain the evaluation rows; genes require three expressing positions. The native spatial graph uses 20 neighbors. Both training stages use batch size 256, and a limit of five epochs or 10,000 updates, whichever occurs first. The molecular encoder uses a 256-unit hidden layer and learning rate 0.001; the spatial encoder uses three single-head transformer layers and learning rate 0.0001. Both use weight decay 10^{-5}. Standardization and PCA32 precede the shared readouts.

The compact table labels Nbh., S-cell, Ctr., and Ctx. denote Neighborhood, Spatial-cell, Center, and Context, respectively. Molecular and Spatial denote the two ZoomST views defined above.

Whole-section anatomical results. Tables 1 and 2 report mean AMI across three downstream clustering initializations for eight MOSTA and seven HESTA sections, respectively. Table C2 reports MOSTA ARI from the same runs, including sample standard deviations (denominator $3 - 1$). For each atlas's summary, sections are averaged equally within each clustering run before averaging the runs. Reported standard deviations quantify clustering variability with fixed representations and sections. Bold marks the top three unrounded values in each comparison, including exact third-place ties. Joint ZoomST features have the highest mean MOSTA ARI (0.345).

Granularity-specific anatomical ablation. At each $k \in \{8, 32, 128\}$, Joint concatenates e_k and g_k and uses the whole-section standardization, PCA32, and three-seed clustering protocol above. HESTA uses its coarse anatomical labels and all 1,129,287 positions across the seven sections. Both HESTA models' exported features pass through float16 and then float32 before feature preprocessing, giving them common stored precision. Table 3a reports the mean of equal-section means across

[1] `https://prabhakarlab.github.io/Banksy/`, accessed 15 September 2026.

Table C2: **ARI on all eight held-out sections.** Methods, positions, class sets, and clustering seeds match Table 1. Values are mean ± sample SD over downstream clustering initializations; the last row first averages sections equally within each seed. Bold marks the top three unrounded values per row, including third-place ties.

	ZoomST views			Frozen pretrained		Target-fitted			
Stage	Joint	Molecular View	Spatial View	Novae	SToFM	GraphST	BANKSY	MENDER	STAGATE
E9.5	**0.391**	0.316	**0.400**	0.233	0.107	**0.418**	0.353	0.367	0.391
	±0.026	±0.012	±0.024	±0.016	±0.006	±0.025	±0.031	±0.016	±0.039
E10.5	0.290	0.257	0.270	0.193	0.127	**0.293**	0.285	**0.297**	**0.314**
	±0.015	±0.017	±0.009	±0.003	±0.004	±0.027	±0.005	±0.006	±0.033
E11.5	**0.376**	**0.364**	**0.356**	0.270	0.147	0.302	0.325	0.282	0.329
	±0.007	±0.004	±0.015	±0.012	±0.008	±0.026	±0.029	±0.001	±0.022
E12.5	**0.374**	**0.423**	0.363	0.281	0.172	0.272	**0.365**	0.331	0.351
	±0.023	±0.009	±0.011	±0.006	±0.020	±0.027	±0.019	±0.018	±0.034
E13.5	**0.334**	0.316	0.329	0.254	0.158	0.228	**0.368**	**0.344**	0.313
	±0.022	±0.015	±0.006	±0.009	±0.013	±0.052	±0.030	±0.012	±0.016
E14.5	0.324	**0.339**	0.288	0.274	0.231	0.278	**0.331**	0.285	**0.345**
	±0.025	±0.030	±0.001	±0.015	±0.008	±0.035	±0.016	±0.026	±0.043
E15.5	0.387	**0.391**	0.368	0.332	0.202	0.212	**0.391**	**0.394**	0.377
	±0.007	±0.037	±0.017	±0.002	±0.002	±0.004	±0.017	±0.019	±0.005
E16.5	0.283	0.298	**0.304**	0.249	0.193	0.219	**0.315**	0.297	**0.321**
	±0.025	±0.013	±0.004	±0.008	±0.004	±0.013	±0.013	±0.044	±0.014
Mean ARI	**0.345**	0.338	0.335	0.261	0.167	0.278	**0.342**	0.325	**0.343**
	±0.006	±0.009	±0.002	±0.001	±0.0005	±0.005	±0.008	±0.006	±0.004

the three downstream clustering runs. Section-level gains compare the models after averaging these runs within each held-out section. Table C3 gives these paired section means.

Table C3: **Joint AMI on every held-out section and granularity.** Each value averages the three downstream clustering runs; bold marks the higher unrounded value within each section–granularity pair.

Dataset	Section (AMI)	8		32		128	
		No Cross	ZoomST	No Cross	ZoomST	No Cross	ZoomST
MOSTA	E9.5 E2S4	0.542	**0.565**	**0.528**	0.524	**0.469**	0.427
	E10.5 E2S1	0.528	**0.552**	0.509	**0.536**	**0.516**	0.510
	E11.5 E1S4	0.587	**0.605**	0.602	**0.615**	0.601	**0.612**
	E12.5 E2S1	0.607	**0.609**	0.595	**0.617**	0.601	**0.608**
	E13.5 E1S4	0.546	**0.577**	0.567	**0.582**	0.537	**0.558**
	E14.5 E2S2	0.577	**0.582**	0.562	**0.573**	**0.542**	0.540
	E15.5 E2S1	0.628	**0.649**	0.609	**0.617**	0.578	**0.584**
	E16.5 E2S13	**0.541**	0.541	**0.521**	0.505	**0.468**	0.464
HESTA	CS12-13 E1S3	0.518	**0.533**	0.535	**0.546**	0.509	**0.526**
	CS14-15 E2S1	0.570	**0.572**	0.574	**0.580**	0.518	**0.529**
	CS17 E1S9	0.498	**0.513**	**0.534**	0.520	0.508	**0.512**
	CS18 E1S2	0.533	**0.569**	0.559	**0.569**	0.553	**0.557**
	CS19 E1S1	0.537	**0.552**	0.534	**0.553**	**0.532**	0.528
	CS20 E1S6	0.564	**0.581**	0.577	**0.584**	0.564	**0.579**
	CS23 E1S3	0.530	**0.561**	0.538	**0.550**	0.527	**0.531**

Neighbor retention. Neighbor retention compares neighborhood membership in physical and representation spaces, describing local geometric correspondence alongside anatomical agreement. Let

$N_k^p(i)$ and $N_k^z(i)$ denote the k nearest neighbors of query i in physical coordinates and the corresponding PCA representation space, respectively. We report

$$R_k = \frac{1}{|Q|} \sum_{i \in Q} \frac{|N_k^p(i) \cap N_k^z(i)|}{k}. \tag{21}$$

Each section supplies 2,048 queries equally spaced among eligible positions in original row order, with every eligible position as a candidate and the query excluded. Exact squared-Euclidean search breaks distance ties by source index. Integer-grid coordinates preserve raw spatial distances up to isometry where available; E15.5 retains the original spatial coordinates. Tables 1 and 2 use $k = 32$ and the same standardized PCA features as each method's anatomical readout. ZoomST features use the view concatenations defined above. Figure C4 displays all eight MOSTA sections, with colored lines connecting the same section across feature sets.

Whole-section maps. Both views contain molecular information: e encodes aggregated expression over a spatial support, while g also uses relative arrangement. Joint features exceed the molecular view on seven of eight section-level AMIs; their benefit over the spatial view varies by section (Table 1).

The E12.5 E2S1 display uses all 23,640 positions and the same 18-class `annotation` field and randomized-PCA readout as Table 1. Fixed seed 42 predictions are shown with one-to-one matched anatomical colors. Integer-grid coordinates are related to the raw coordinates by a distance-preserving rotation/reflection. Joint-feature IoUs for Brain and Choroid plexus are 0.402 and 0.228, respectively.

The E16.5 E2S13 Brain display uses all positions in the section, without smoothing or interpolation. Molecular features produce 504 extra and 74 missed Brain assignments; spatial features produce 60 extra and 373 missed assignments; joint features produce 77 extra and 97 missed assignments. These maps use the seed-42 run included in the three-seed quantitative evaluation.

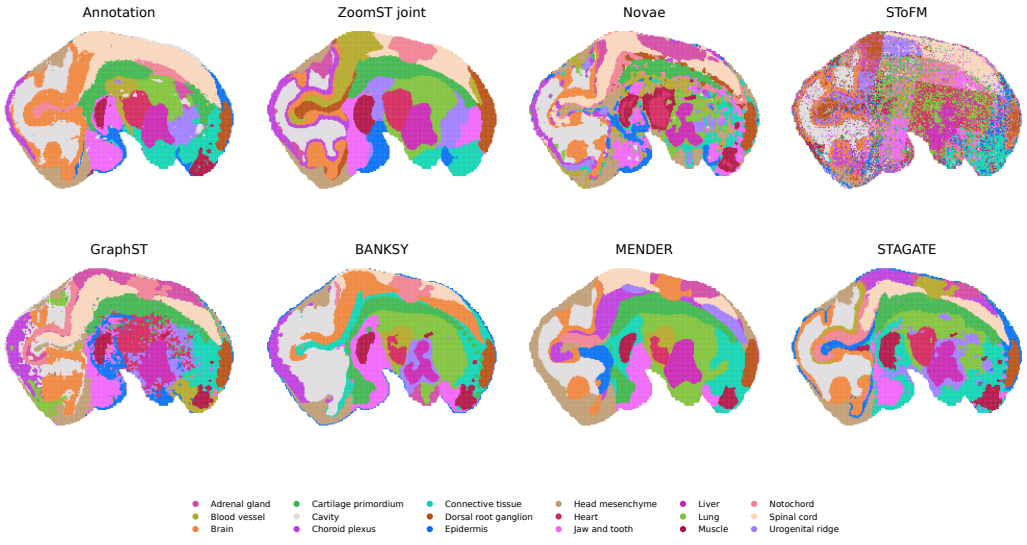

Figure C2: **Example case: whole-section tissue mapping in E12.5 E2S1.** The annotation and seven-method predictions share all 23,640 positions and 18 primary classes. Maximum-overlap one-to-one matching assigns anatomical colors. The top row shows annotations and predictions from pretrained models; the lower row contains target-fitted methods. Maps use no smoothing or label interpolation.

C.3 Developmental ordering and stage prediction within Brain

MOSTA Brain is selected by coverage across all eight stages and the minimum available stage count. DPT uses the same 589 Brain observations from each held-out section, for 4,712 observations in

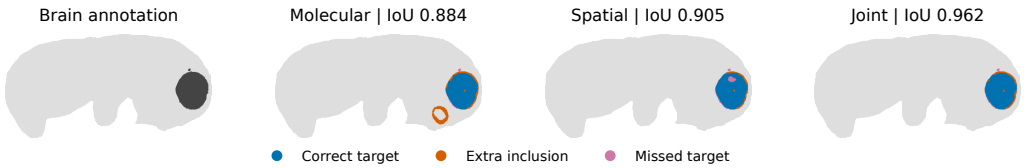

Figure C3: **Complementary molecular and spatial errors on a held-out section.** Annotation, molecular, spatial, and joint Brain assignments in E16.5 E2S13. Blue: correct assignments; orange: extra assignments; magenta: missed Brain.

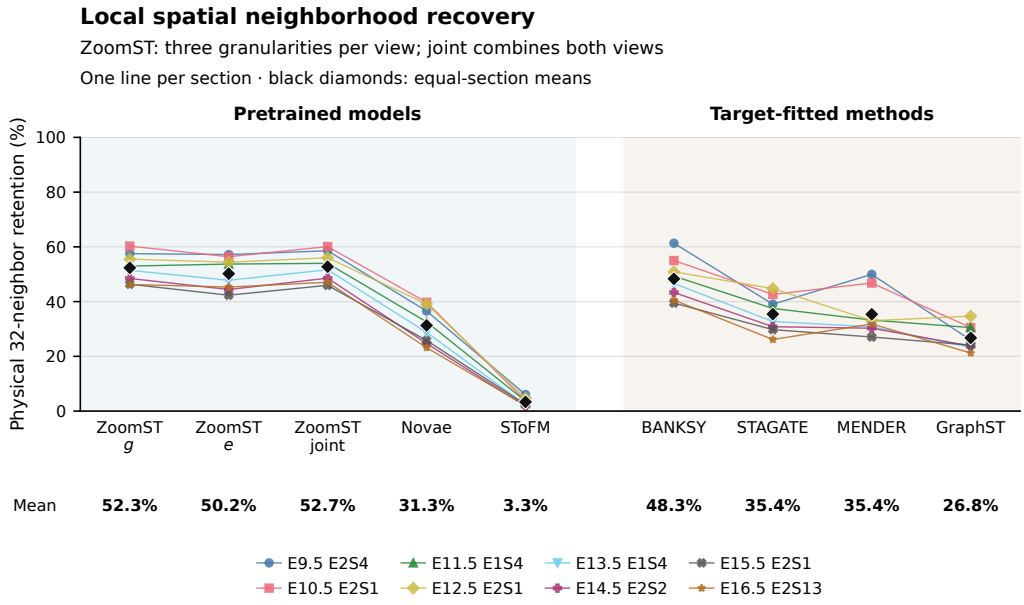

Figure C4: **Local neighbor retention in the tissue-mapping representations.** Markers show physical 32-neighbor recovery over 2,048 fixed queries per section; all positions are candidates and self is excluded. Colored lines connect the same section within each method group, and black diamonds mark equal-section means. ZoomST g and e each concatenate granularities 8, 32, and 128; joint concatenates both views. All three use the same standardized PCA features as Table 1. Pretrained models and target-fitted spatial methods are grouped separately.

total. HESTA Brain has sufficient coverage at CS17, CS18, CS19, CS20, and CS23; we fix 512 observations from each section, for 2,560 in total. Models share the same positions within each atlas. ZoomST and No Cross use the models and Joint representations defined in Appendix C.2; the cross-method MOSTA comparison additionally uses Novae and SToFM's representations and the shared-fit baselines below.

Table 3b evaluates Joint separately at each context size $k \in \{8, 32, 128\}$. Within each atlas, all six model–scale combinations use the same balanced Brain cohort, 30 neighbors, and a fixed root. Each developmental stage is represented by one held-out section. Table C4 additionally retains the MOSTA whole-section anatomical and physical-neighbor readouts alongside DPT.

Shared representations across sections. For the cross-method MOSTA comparison, GraphST and STAGATE each fit one encoder jointly across all eight target sections with shared gene selection and separate within-section spatial graphs. BANKSY uses a common 3,000-gene panel, and MENDER uses a common codebook of 32 unsupervised expression states. These fits use the complete target sections before extracting the balanced Brain observations. Other feature-construction and model-training settings follow Appendix C.2.

Pseudotime. External baselines in both atlases share featurewise z-score standardization over the pooled Brain cohort, followed by PCA32 (STAGATE retains its native 30 dimensions). DPT (Haghverdi et al., 2016) uses Euclidean neighbor graphs and Scanpy's default distance definition with ten diffusion components.[2] The default neighborhood size is 30. To address DPT computation failures, we expand the graph to 50 neighbors for STAGATE, BANKSY, and MENDER on MOSTA. Other DPT settings are shared. The MOSTA root is the E9.5 observation nearest the spatial centroid of the sampled E9.5 Brain domain. Sampling and numerical routines use seed 42. Root distances are normalized by the largest finite distance within each method; all selected MOSTA observations have finite pseudotime under the reported settings.

Pseudotime–stage agreement. For observation i, let t_i denote pseudotime and s_i its measured developmental stage. Spot-level Spearman correlation is $\mathrm{Spearman}(t_i, s_i)$ over all selected observations in an atlas, using unrounded pseudotimes and average ranks for ties. The observations span eight MOSTA stages or five HESTA stages. Figure 2 reports both atlases; Table C5 retains the combined-granularity MOSTA values.

For HESTA, the root is the first observation in the sorted sampled CS17 indices, fixed before evaluating the scores; its distances are normalized by the largest finite distance. ZoomST and No Cross have finite pseudotime for all selected Brain observations. Carnegie stage numbers specify the ordered target. PCA and graph construction pool the unlabeled Brain observations within each atlas; stage labels anchor the earliest-stage root and evaluate the resulting order.

HESTA cross-method comparison. The HESTA panel uses the same 2,560 Brain observations and CS17 root. External pretrained representations comprise Novae, SToFM, scGPT-spatial, the Cell, Neighborhood, and Spatial-cell outputs of TERRA, and CellWorld. They follow the standardization and DPT protocol above, with 30 neighbors for every method. CellWorld uses its CP10K/log1p and per-section Seurat-HVG300 export. Its 512 CS18 observations are disconnected from the common root, so the figure marks its complete-stage readout as NA. The remaining methods cover all five stages. The existing independently fitted per-section DECIPHER embeddings are not pooled into a shared cross-stage coordinate system.

Table C4: **MOSTA Joint readouts by granularity.** AMI means follow Table 3a; SD is the sample SD across three downstream clustering runs. Neighbor retention is equal-section physical 32-neighbor recovery. Spot-level correlation uses the 4,712-observation Brain cohort. Bold marks the higher value in each pair, including ties.

	AMI ↑			Neighbor retention (%) ↑			Spot-level ρ ↑		
Model	8	32	128	8	32	128	8	32	128
ZoomST	**0.585 ± 0.006**	**0.571 ± 0.003**	0.538 ± 0.002	**35.1**	**52.7**	**62.5**	0.868	**0.822**	**0.354**
No Cross	0.570 ± 0.001	0.562 ± 0.005	**0.539 ± 0.0003**	34.2	50.5	61.4	**0.903**	0.738	-0.239

Table C5: **Spot-level pseudotime–stage Spearman correlation within Brain.** Correlation is computed over all 4,712 observations spanning eight MOSTA stages. Bold marks the highest value.

Representation	Spot-level Spearman ρ
ZoomST joint	**0.816**
Novae	0.714
SToFM	0.341
ZoomST no cross	0.651
GraphST	0.491
STAGATE	0.229
BANKSY	-0.267
MENDER	0.458

[2] https://scanpy.readthedocs.io/en/stable/generated/scanpy.tl.dpt.html

Developmental-stage prediction. The linear probe uses the original Joint features without PCA. Each leave-one-section-out fold fits StandardScaler and a Ridge regressor only on Brain observations from the remaining sections. The intercept is unpenalized and the Ridge coefficient is $\alpha = n_{\text{train}}\lambda$, with fixed $\lambda = 0.1$. MOSTA training folds use 512 Brain observations per section and evaluate every Brain observation in the test section. HESTA uses the same fixed 512-observation-per-section cohort for training and testing across folds. Targets are embryonic days for MOSTA and Carnegie stages for HESTA.

For section s with n_s evaluated observations, $\mathrm{MAE}_s = n_s^{-1} \sum_i |\hat{s}_i - s|$; reported MAE averages the included sections equally. Relative reduction is $(\overline{\mathrm{MAE}}_{\text{No Cross}} - \overline{\mathrm{MAE}}_{\text{ZoomST}})/\overline{\mathrm{MAE}}_{\text{No Cross}}$.

D CROSS-GRANULARITY PREDICTION

Representation states. The final MOSTA model specified in Section 4 supplies native 256-dimensional molecular states without standardization. Data use follows Table A2.

Same-location state prediction. For every position in each section, define $m_i^T = d^{-1}\|T_{a\to b}(e_{i,a}) - e_{i,b}\|_2^2$ and $m_i^0 = d^{-1}\|e_{i,a} - e_{i,b}\|_2^2$. The unchanged source is the no-transformation reference; the target is directly encoded from the actual broader neighborhood at the same location. MSE averages over all positions in each section. Across all 24 held-out section–direction conditions, transformed states have lower target MSE than unchanged sources.

Figures 4a and D1a show the corresponding absolute MSEs. Each paired point is one section's mean over the three trained relations, with each relation first averaged over all positions; the eight section means form each box.

Projected joint association. Each section supplies the same fixed 2,048 positions, partitioned into four disjoint batches of 512. For each batch, correct predicted pairings and 20 shuffled predictions are compared with directly encoded fine–coarse pairs. Each shuffle has no fixed points and permutes only predicted destinations, preserving their marginal distribution. The product kernel uses source and target bandwidths from the corresponding observed states, shared across controls, with the settings in Appendix B.

We evaluate the checkpoint's training projection and eight fixed orthogonal projections. Correct source–prediction pairings yield lower joint discrepancy than shuffled pairings across all evaluated batch–projection–direction conditions. Figure D1b compares absolute MMDs, first averaged over batches and the eight fresh projections within each section and direction, then equally over the three trained directions within each section.

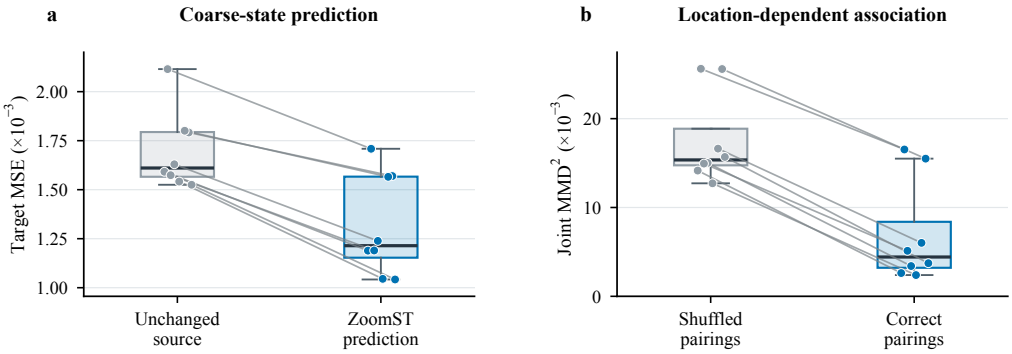

Figure D1: **Predictive relationships across held-out tissue sections. a**, Target MSE for unchanged sources and ZoomST predictions. **b**, Joint MMD for shuffled and correct predicted pairings. Points are means over the three trained granularity pairs for each of eight sections; lines connect the same section. Boxes show medians, interquartile ranges and 1.5-IQR whiskers.

E UNSEEN GRANULARITIES AND CONTEXT RESPONSE

Figure 4: evaluation and aggregation. All three panels use eight sections excluded from pretraining. In panels (a,b), target MSE compares unchanged source states and ZoomST predictions against the same directly encoded targets. Panel (a) averages the trained pairs $8 \to 32$, $8 \to 128$ and $32 \to 128$ within each section, using all section positions. Panel (b) averages the six unseen targets $12, 16, 24, 48, 64, 96$ queried from e_8 at 1,024 fixed positions per section. Panel (c) selects the lowest and highest predicted-response quarters separately within each section and doubling interval ($8 \to 16$, $16 \to 32$, $32 \to 64$, $64 \to 128$), with 256 locations per quarter, then averages their directly encoded changes over the four intervals. Each box therefore contains eight section means. The highest-response quarter shows larger encoded changes than the lowest quarter in 31 of 32 section–interval conditions (Figure 4c).

E.1 TRANSITION HEADS ON A SHARED ENCODER

Shared-encoder comparison. The neural ODE and granularity-conditioned residual MLP are trained at 8, 32 and 128 on a shared frozen encoder, then evaluated against the same directly encoded targets. Averaging MSE over six unseen granularities, the ODE reduces error on every held-out section, with a mean paired reduction of 10.86% (Figure E1).

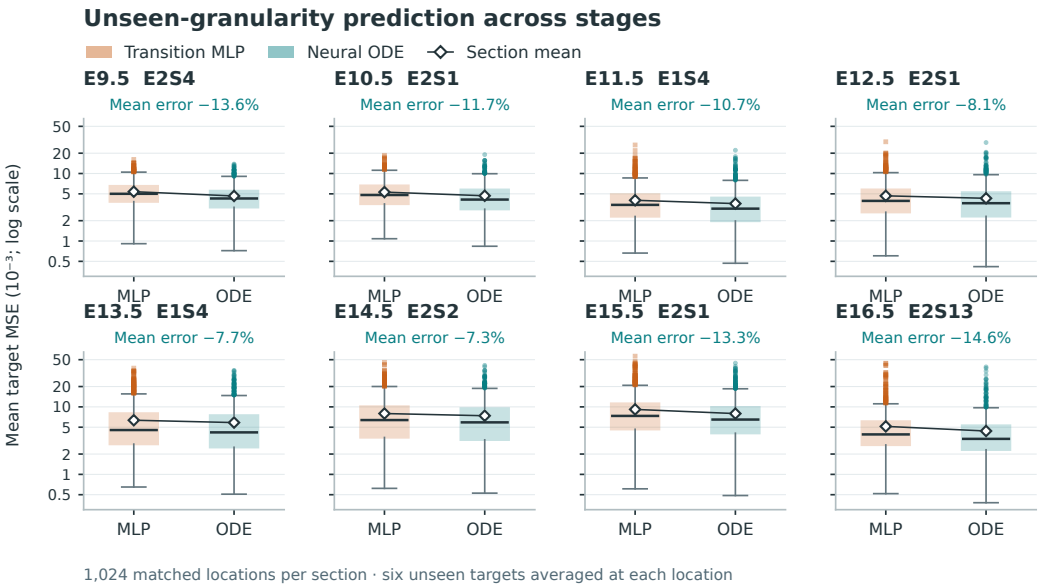

Figure E1: **Lower mean unseen-granularity error with neural ODE.** Each panel compares the granularity-conditioned residual MLP and ODE on one held-out section. Both heads use the same encoder, fine-state inputs and directly encoded targets. Each box contains 1,024 location-level MSEs, each averaged equally over $k \in \{12, 16, 24, 48, 64, 96\}$. Boxes show medians and interquartile ranges; whiskers extend to 1.5 interquartile ranges, with all outliers shown. Connected diamonds mark section means; percentages give their paired relative reductions. All panels share the same logarithmic axis.

We compare the ODE with the granularity-conditioned residual MLP,

$$T^{\mathrm{MLP}}_{a\to b}(z) = z + \mathrm{MLP}([\mathrm{LN}(z), h(a), h(b), u(b) - u(a)]) .$$

The MLP has dimensions $321 \to 128 \to 256$ with GELU activation and 32-dimensional scale embeddings; the ODE field has dimensions $257 \to 128 \to 256$. For a query between adjacent trained coordinates u_j, u_{j+1}, the MLP uses $h(k) = (1-w)h_j + wh_{j+1}$ with $w = (u(k) - u_j)/(u_{j+1} - u_j)$. The heads contain 74,848 and 66,560 trainable parameters, respectively.

Both heads are initialized afresh and trained on representations from the shared encoder using all 45 source sections for two epochs (7,224 updates; batch 1,024). AdamW uses learning rate 2×10^{-4},

188 warmup steps and cosine decay to one tenth of the peak rate; gradient norms are clipped at five. Weight decay is 0.02 for weight matrices, 0.01 for scale embeddings and zero for normalization and bias terms. The ODE uses RK4 with maximum step 0.25 in u. Evaluation uses the same 1,024 positions per held-out section and the same directly encoded targets for both heads.

For section s, let M_s^m be its MSE averaged equally over the six unseen targets for head m. The mean paired reduction is $100\,\mathrm{mean}_s(1-M_s^{\mathrm{ODE}}/M_s^{\mathrm{MLP}}) = 10.86\%$, ranging from 7.28% to 14.56% across sections. Table E1 gives the section summaries underlying Figure E1. These are separately fitted transition heads on a common reference encoder.

Table E1: **Matched transition heads on a shared frozen encoder.** MSE is averaged over all six unseen targets and 1,024 locations per section. The mean reduction averages the eight paired section reductions.

Section	MLP MSE ($\times 10^{-3}$)	ODE MSE ($\times 10^{-3}$)	Reduction (%)
E9.5 E2S4	5.381	4.650	13.57
E10.5 E2S1	5.303	4.682	11.73
E11.5 E1S4	4.014	3.585	10.69
E12.5 E2S1	4.673	4.295	8.08
E13.5 E1S4	6.342	5.853	7.71
E14.5 E2S2	7.971	7.391	7.28
E15.5 E2S1	9.174	7.957	13.26
E16.5 E2S13	5.140	4.392	14.56
Mean	6.000	5.350	10.86

E.2 Agreement with unseen neighborhood encodings

Prediction protocol. The MOSTA encoder and ODE are jointly trained at 8, 32, and 128. We use 1,024 fixed positions on each of the eight held-out sections and directly encode neighborhoods at $\mathcal{K} = \{8, 12, 16, 24, 32, 48, 64, 96, 128\}$. The molecular encoding procedure and neighborhood ordering remain unchanged. All predictions start from the same $e_{i,8}$; the encoded intermediate states are scoring references and do not enter these predictions.

Let $\widehat{e}_i(k) = T_{8\to k}(e_{i,8})$ and $\mathcal{U} = \mathcal{K} \setminus \{8, 32, 128\}$. For section s, target MSE averages $d^{-1}\|\widehat{e}_i(k) - e_i(k)\|^2$ over positions. The section's relative gain is one minus its MSE averaged over $k \in \mathcal{U}$ divided by the corresponding unchanged-e_8 MSE. Gains are then averaged over eight sections, yielding 21.96% (Table E2). Normalized RMSE divides RMSE by the target's root-mean-square variation around its section mean, so $R^2 = 1 - \mathrm{NRMSE}^2$ within each section. Prediction improves over unchanged e_8 in all 48 section–target conditions. Figure 4b plots absolute MSE averaged over the six targets within each section, giving eight paired section means.

Table E2: **Unseen-granularity MSE by section.** MSE averages the six unseen targets at 1,024 positions per section. Reductions compare ZoomST with the source used as prediction; the last row averages the eight section reductions.

	MSE ($\times 10^{-3}$)		
Section	Source as prediction	ZoomST prediction	Reduction (%)
E9.5 E2S4	1.7070	1.3672	19.91
E10.5 E2S1	1.5263	1.3266	13.09
E11.5 E1S4	1.4363	1.0915	24.00
E12.5 E2S1	1.6323	1.4031	14.05
E13.5 E1S4	1.4106	1.0917	22.61
E14.5 E2S2	1.4178	1.0241	27.77
E15.5 E2S1	1.4248	1.0680	25.04
E16.5 E2S13	1.3519	0.9570	29.21
Mean reduction	—	—	**21.96**

Distance to a common directly encoded target. For the direct-distance comparison in Figure E2, define

$$d_i^{\text{start}}(k) = \frac{\|e_{i,8} - e_i(k)\|_2}{\sqrt{d}}, \qquad d_i^{\text{pred}}(k) = \frac{\|\widehat{e}_i(k) - e_i(k)\|_2}{\sqrt{d}}.$$

At each target, we average these per-position RMS distances within each section. Sections receive equal weight in the summary. Averaging $1 - \overline{d}_s^{\text{pred}}(k)/\overline{d}_s^{\text{start}}(k)$ equally over the six unseen targets and eight sections gives a 10.38% reduction, ranging from 3.04% to 20.55% across the 48 conditions. Every condition has a lower mean and median prediction distance. Within individual conditions, 81.15–99.71% of positions are closer to the target. This RMS-distance statistic differs from the reduction in mean squared error reported above.

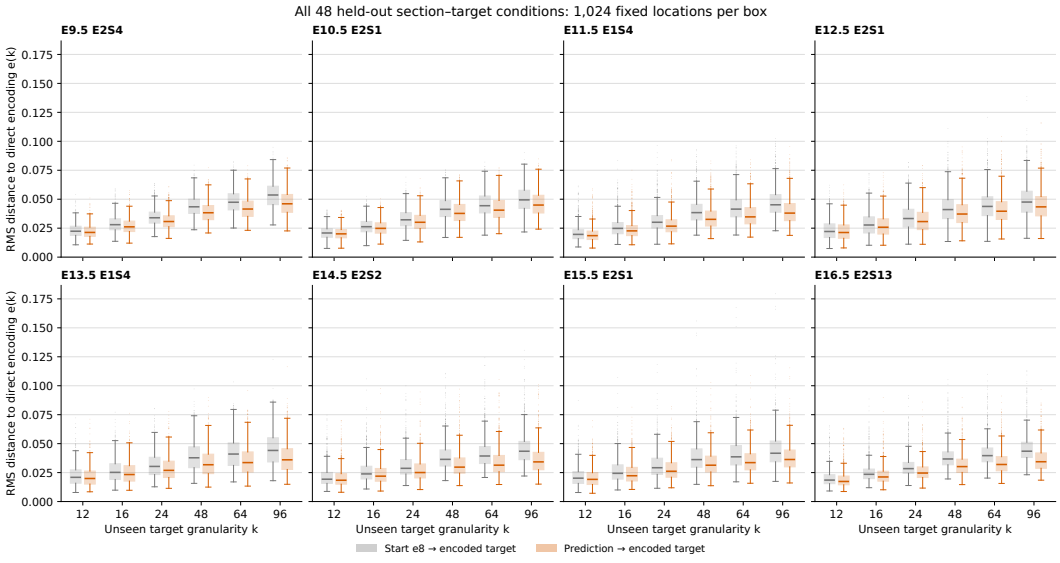

Figure E2: **Within-section distributions of distance to directly encoded unseen states.** The panels cover eight held-out sections and six unseen target granularities. Each pair compares distances from unchanged e_8 and from the prediction to the same target at the same 1,024 fixed positions. Boxes show medians and interquartile ranges, whiskers extend to 1.5 interquartile ranges, and outliers are retained. Position distributions describe each section; section means are the units in the main comparison.

Shared-space visualization and full-state error. The E16.5 E1S1 analysis uses 1,024 fixed positions. We fit a center-only, full-SVD PCA on the directly encoded states at all nine granularities, then apply the same mean and components to predictions. All positions are shown with common axis limits. The first two PCs explain 20.25% and 17.02% of encoded-state variance. Table E3 reports prediction accuracy in the full 256-dimensional space.

Table E3: **Full-dimensional accuracy underlying the case PCA.** MSE ratio compares prediction with unchanged e_8 against directly encoded targets; NRMSE and R^2 use the same full-dimensional targets.

Target k	MSE ratio	NRMSE	R^2
12	0.884	0.206	0.958
16	0.810	0.243	0.941
24	0.739	0.280	0.922
48	0.664	0.333	0.889
64	0.643	0.352	0.876
96	0.621	0.377	0.858

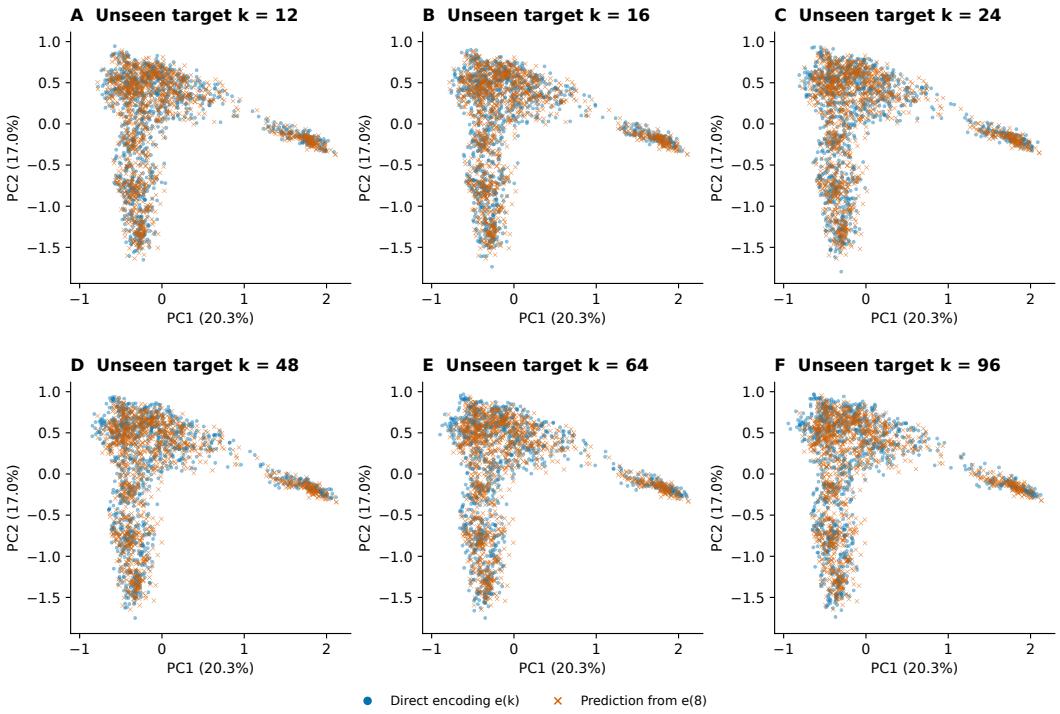

Figure E3: **All six unseen target states in the E16.5 E1S1 case.** Every panel uses the same PCA basis and the same 1,024 positions. Metrics are computed in 256 dimensions. Targets and predictions are paired at each location.

E.3 Context-response evaluation across observation ranges

Matched response definitions and sampling. For $\widehat{e}_i(k) = T_{8\to k}(e_{i,8})$, Figure 4c, Figure 5a and the grouping in Figure E4 use

$$D_i^{\mathrm{pred}}(a,b) = \frac{\|\widehat{e}_i(b) - \widehat{e}_i(a)\|_2}{\sqrt{d}\,\log_2(b/a)}, \qquad D_i^{\mathrm{enc}}(a,b) = \frac{\|e_i(b) - e_i(a)\|_2}{\sqrt{d}\,\log_2(b/a)}.$$

Both operands in the predicted displacement belong to queries from the same fixed e_8. Encoded states are evaluation references, and all displayed doubling intervals have log width one.

Held-out comparisons use 1,024 fixed locations per section. Anatomical cases use the training sections identified in Table A2, with up to 128 locations per annotation selected by a fixed hash of observation identity; smaller domains retain every location.

Held-out location ordering. For each held-out section and each doubling interval, we sort the 1,024 locations by predicted response and split them into four equal groups; ties preserve sample order. We compare the directly encoded changes in the lowest and highest quarters. Figure 4c averages these group means over the four intervals within each section, giving eight paired section means.

Actual neighborhood composition. Let $q_{ic}(k)$ be the proportion of annotation c among the $k-1$ neighbors after excluding the focal center. For the same interval as the response,

$$C_i(a,b) = \tfrac{1}{2}\sum_c |q_{ic}(b) - q_{ic}(a)|.$$

Support ordering is identical to the model input, and unknown-label mass is retained. The Facial nerve composition bars (Figure E4) follow the same 16 locations in each extreme predicted-response quarter. Their mean encoded changes are 0.0191 and 0.0343 in the low and high groups, respectively

(ratio 1.799). Corresponding mean TV values are 0.134 and 0.284. Composition TV quantifies the change in annotated tissue proportions.

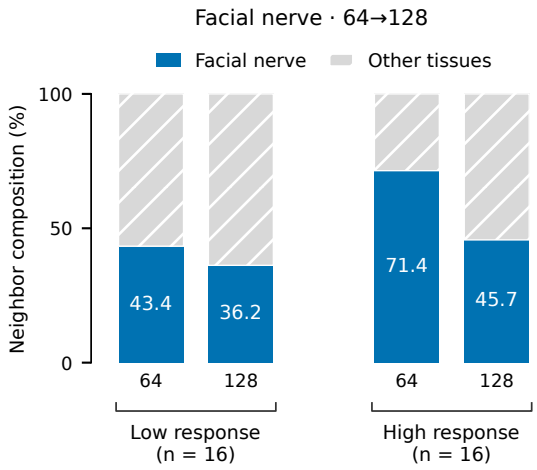

Figure E4: **Neighborhood composition in low- and high-response Facial nerve locations.** The groups are the lowest and highest predicted-response quarters among all 64 Facial nerve locations in the E14.5 E1S1 training section, using the $64 \rightarrow 128$ response in Figure 5a. Each pair of bars follows the same 16 locations at both granularities. Blue shows the mean Facial nerve fraction; gray shows all other labels. Predictions start from fixed e_8.

Within-Brain changes under a constant annotation. Among sampled Brain locations whose neighborhoods remain entirely Brain-labeled at both $r = 8$ and $r = 16$, predicted responses correlate with directly encoded changes (Spearman $\rho = 0.645$), showing molecular-context variation within a constant anatomical annotation.

Whole-section spatial overview. Figure 5b averages the instantaneous response $s_i(2^t) = \|f_\theta(z_i((t-3)/4), (t-3)/4)\|_2/(4\sqrt{d})$ over 64 equally spaced midpoints in $t = \log_2 k \in [3, 7]$. This path average differs from the finite response used in panel (a) and to define the groups in Figure E4. Colors use an empirical percentile reference pooled over the full-range mean and the four doubling-range means, with all 121,767 locations included. Panel (c) shows observed composition TV from 8 to 128. The maps show spatial distributions without smoothing; their whole-section Spearman correlation is 0.174. Both maps are rotated together by $180°$ relative to the source coordinates for display, preserving distances and the common orientation. Spatial scale bars use the nominal $25\,\mu$m pitch of the bin50 matrices (Chen et al., 2022); response remains normalized by $\log_2 k$.

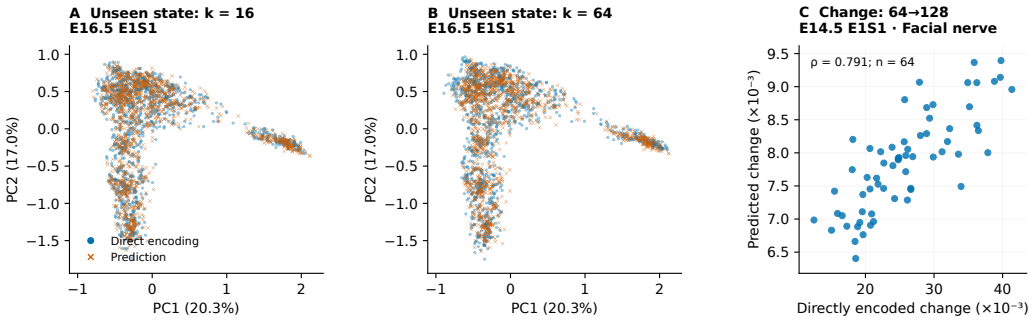

Figure E5: **Illustrative unseen-state predictions and response correlation in source-section cases. A,B**, Direct encodings and predictions at unseen granularities 16 and 64 for 1,024 matched E16.5 E1S1 positions, using a shared PCA basis fitted on direct encodings at nine granularities. Full-dimensional R^2 is 0.941 and 0.876. **C**, Predicted versus directly encoded change over $64 \rightarrow 128$ in E14.5 E1S1 Facial nerve. Points show all 64 matched locations; axes report RMS representation change. Both sections belong to the training set. All predictions start from e_8; 16 and 64 are absent from granularity supervision. Held-out prediction and response summaries are shown in Figure 4.

F OBJECTIVE ABLATION ON MOSTA

We compare No Cross, CS, and ZoomST using a common initialization, the same training data and training budget, and identical training settings apart from the loss terms. No Cross uses within-granularity learning; CS adds coarse-support supervision; ZoomST further adds JointMMD. Evaluation uses the held-out MOSTA split from Section 5.2, with identical evaluation observations and readout settings across the three models. All readouts use Joint representations at granularities 8, 32, and 128.

Table F1 follows the anatomical clustering, Brain developmental ordering, and stage-prediction protocols in Appendix C.2–C.3.

Adding JointMMD to CS reduces stage-prediction MAE and increases spot-level DPT correlation at all three granularities, with the largest gains at granularity 128. AMI increases at granularities 8 and 32 and remains similar at granularity 128.

Table F1: **Controlled objective ablation on MOSTA.** Joint AMI, Brain spot-level DPT–stage Spearman correlation, and Brain stage-prediction MAE. AMI averages the eight held-out sections equally; $\pm$ denotes SD over downstream clustering runs. DPT and stage MAE use the protocols and aggregation of Table 3b. Bold marks the best unrounded mean in each column.

	Joint AMI ↑			Brain DPT ρ ↑			Brain MAE (days) ↓		
Model	8	32	128	8	32	128	8	32	128
ZoomST	**0.585** ±0.006	**0.571** ±0.003	0.538 ±0.002	0.868	**0.822**	**0.354**	**0.803**	**0.715**	**0.734**
CS	0.575 ±0.003	0.564 ±0.001	**0.539** ±0.009	0.857	0.737	0.143	0.818	0.733	0.802
No Cross	0.570 ±0.001	0.562 ±0.005	0.539 ±0.0003	**0.903**	0.738	-0.239	0.936	0.853	0.929

G DEVELOPMENTAL READOUT OF PREDICTED UNSEEN-GRANULARITY STATES

We evaluate whether predicted molecular states retain developmental information under a shared age readout, using all eight held-out MOSTA sections and six unseen granularities.

G.1 EVALUATION PROTOCOL

The eight sections span E9.5–E16.5 and were excluded from ZoomST pretraining. In each leave-one-section-out fold, feature standardization, PCA32 without whitening, and ridge regression are fitted only on directly encoded molecular states at trained granularities $r \in \{8, 32, 128\}$ from the other seven sections. Each training section contributes 512 Brain positions, yielding $N = 7 \times 512 \times 3$ readout examples. The primary ridge penalty is $\alpha = N\lambda$ with $\lambda = 0.1$, and the intercept is unpenalized. Target-section labels and directly encoded unseen-granularity states do not enter fitting.

The primary evaluation uses 512 Brain positions per test section. Predictions $\hat{e}_i(r) = T_{8\to r}(e_{i,8})$ and all comparison inputs pass through the same fitted readout within each fold. MAE is measured in embryonic days and averaged equally across sections; the six-target summary first averages targets within each section.

G.2 READOUT ACROSS UNSEEN GRANULARITIES

Averaged across all eight sections and six unseen targets, predicted states reduce MAE from 1.0300 to 0.9689 days (5.9%) relative to unchanged e_8. Every target has a positive section-mean gain, with improvement in six of eight sections at each target (Table G1). Figure G1 shows the mean errors across all six targets.

Table G1: **Developmental-age readout at unseen granularities.** Values are section-equal mean MAE in days on the primary positions. Direct target encodings read the actual target neighborhoods and serve as references. The uniform-shift control is defined in Appendix G.3. Gain is the relative reduction in section-mean MAE compared with unchanged e_8.

Target r	Unchanged e_8	Uniform shift	ODE query	Direct reference	Gain vs. e_8
12	1.0300	1.0157	1.0089	0.9697	2.0%
16	1.0300	1.0071	0.9956	0.9447	3.3%
24	1.0300	0.9965	0.9787	0.9242	5.0%
48	1.0300	0.9820	0.9535	0.9167	7.4%
64	1.0300	0.9771	0.9443	0.9193	8.3%
96	1.0300	0.9710	0.9323	0.9292	9.5%

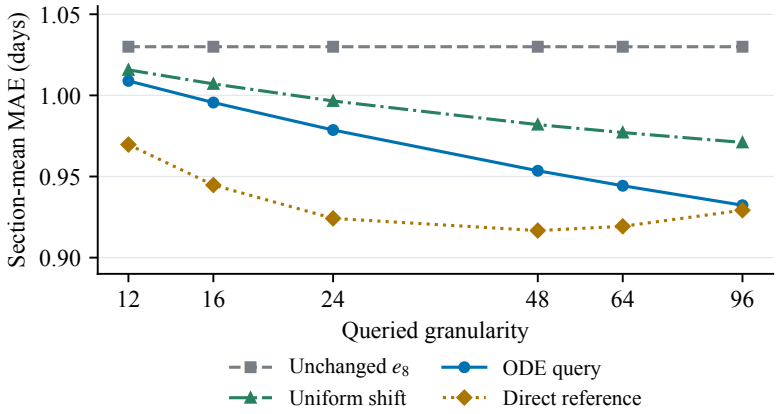

Figure G1: **Section-averaged developmental readout across unseen granularities.** Each point is the mean MAE over all eight held-out sections on the primary positions. The horizontal axis uses log-granularity spacing. Sections are the evaluation units; target granularities are repeated queries on the same sections.

G.3 Age-readout gain over unchanged source states

To test whether a uniform shift explains the readout gain, we average ODE displacements over Brain positions in the seven readout-training sections. For held-out section s and target r,

$$\bar{\Delta}_{-s,r} = \frac{1}{7 \times 512} \sum_{j \in \mathcal{I}_{-s}} \left[T_{8 \to r}(e_{j,8}) - e_{j,8} \right], \qquad e_{i,r}^{\text{uniform}} = e_{i,8} + \bar{\Delta}_{-s,r}.$$

No target-section observations or age labels are used to estimate this displacement. The control uses the same readout as the full ODE query. At $r = 96$, ODE queries reduce section-mean age-prediction MAE from 1.0300 to 0.9323 days, approaching the performance of directly encoded target states under the same readout (0.9292 days; Table G2).

Table G2: **Age-readout gain over unchanged source states at $r = 96$.** Values are section-equal mean MAE in days across all eight sections. The first three rows require only e_8 from each test location; the final row reads the actual target neighborhood. All rows use the same fitted readout within each fold. Alternative positions are disjoint from the primary set and are described in Appendix G.4.

Input to the shared readout	Primary positions	Alternative positions
Unchanged e_8	1.0300	1.0303
Uniform ODE displacement	0.9710	0.9647
ODE query $\hat{e}_{96}$	0.9323	0.9266
Direct target encoding e_{96} (reference)	0.9292	0.9316

G.4 Readout sensitivity

We repeat evaluation without refitting the readout on alternative Brain positions disjoint from the primary set: 259 at E9.5, 77 at E10.5, and 512 in each remaining section. At $r = 96$, MAE decreases from 1.0303 to 0.9266 days (10.1%), with improvement over unchanged e_8 in six of eight sections (Table G2).

Keeping standardization and PCA32 fixed, we vary $\lambda \in \{0.01, 0.1, 1\}$ under $\alpha = N\lambda$ and $\alpha = N\lambda v$, where v is the mean training-feature variance after preprocessing. Across all six settings and both position sets, ODE queries have lower section-mean MAE than unchanged e_8 and the uniform-shift control. Table G3 gives all six settings on the primary positions.

Table G3: **Regularization sensitivity at $r = 96$.** Values are section-equal mean MAE in days on the primary positions. All eight sections are included in every row.

Penalty α	λ	Unchanged e_8	Uniform shift	ODE query	Direct reference
$N\lambda$	0.01	1.0401	0.9801	0.9394	0.9426
$N\lambda$	0.1	1.0300	0.9710	0.9323	0.9292
$N\lambda$	1.0	0.9890	0.9378	0.9088	0.8717
$N\lambda v$	0.01	1.0324	0.9732	0.9340	0.9324
$N\lambda v$	0.1	0.9938	0.9412	0.9106	0.8788
$N\lambda v$	1.0	1.0406	1.0040	0.9892	0.9244

Without PCA, retaining training-only standardization and $\lambda = 0.1$, MAE decreases from 1.1061 to 1.0441 days on primary positions and from 1.1091 to 1.0404 days on alternative positions relative to unchanged e_8. These checks assess readout and position-sampling sensitivity using the same model and held-out sections.