

MP-MNCA: Morphology-Prior Masked Neighbor Cross-Attention for Efficient Spatial Domain Identification

Sidharthan S C*
IIT Madras

be21b039@smail.iitm.ac.in

Sanjeev Manivannan*
IIT Madras

sanjeev.manivannan@gmail.com

Abstract

*Spatial transcriptomics jointly captures molecular state, tissue morphology, and spatial organization, yet spatial-domain methods are typically compared only at late training checkpoints, obscuring differences in convergence efficiency. We introduce MP-MNCA, a morphology-prior masked neighbor cross-attention framework that preserves the full 3,000-dimensional highly-variable-gene representation while selectively routing molecular information across spatial neighbors, with histological similarity incorporated as a continuous attention prior under a self-supervised masked neighbor-contrastive objective. Across 12 heterogeneous human dorsolateral prefrontal cortex (DLPFC) sections, MP-MNCA is evaluated against established spatial-domain baselines using both convergence trajectories and epoch-wise spatial-domain visualizations. Our results show that MP-MNCA (i) reaches coherent spatial-domain structure in substantially fewer training epochs, as reflected quantitatively by convergence curves and qualitatively by progressive spatial grids, and (ii) maintains strong clustering quality, achieving a mean peak refined ARI of **0.546** versus **0.549** for the strongest baseline (STAIG), while attaining the highest mean refined NMI of **0.657** versus **0.640**. Moreover, MP-MNCA outperforms STAIG on **7 of 12** sections in ARI and **11 of 12** in NMI, demonstrating faster convergence without sacrificing spatial-domain quality.*

1. Introduction

Spatial transcriptomics (ST) jointly measures transcriptome-wide molecular activity, tissue morphology, and physical position, making it possible to study tissue organization without dissociating cells from their native context [1, 6, 13]. Modern ST platforms provide spatial coordinates with histological images, allowing molecular, spatial, and morphological information to be analyzed jointly [7, 8]. A central computational problem in ST is

therefore *spatial domain identification*: grouping measurement locations into coherent anatomical tissue regions without requiring domain labels during representation learning [3, 5, 17, 19]. Such domains capture organized biological structure and form the basis for downstream analysis of tissue architecture, regional gene-expression programs, and disease-associated spatial variation [1, 6, 8].

ST datasets such as the human dorsolateral prefrontal cortex (DLPFC) dataset can be viewed hierarchically as *donors* $\rightarrow$ *tissue sections* $\rightarrow$ *spatial capture spots* [7, 10]. DLPFC contains 12 independently profiled sections from three donors, with four sections per donor [7, 10]. Each section contains approximately 4,000 spatially localized Visium spots, where every spot is associated with a high-dimensional gene-expression profile, a two-dimensional coordinate, and a corresponding H&E image region [7, 12]. For evaluation, spots were retrospectively annotated into cortical layers L1–L6 and white matter (WM) using cytoarchitecture and molecular markers [7]. This donor–section–spot hierarchy is representative of ST datasets in which spatially arranged molecular observations are embedded within a shared histological context [6, 8, 13]. Despite this structured organization, recovering biologically meaningful spatial domains without using anatomical annotations during training remains difficult due to four major challenges:

The *first major challenge* is inter-donor and inter-section variability under label-free learning. DLPFC provides cortical-layer annotations for retrospective evaluation, but spatial-domain methods must recover these structures without using the labels during representation learning [5, 7, 17]. Individual sections differ in sequencing coverage, tissue geometry, molecular composition, and preparation, while additional donor- and batch-specific variation further shifts their distributions [5, 7, 16]. Accordingly, methods such as GraphST, MuCoST, and STAIG, discussed in Sec. 2, learn from unlabeled molecular and spatial structure rather than anatomical supervision [5, 16, 18]. Spatial-domain identification is therefore fundamentally a representation-learning problem: the learned features must remain discriminative across section- and subject-specific variation [5, 16, 18].

*Equal contribution.

The *second major challenge* is that a Visium spot is a mixed local observation rather than an isolated semantic object. Each spot can contain transcripts from multiple cells and cell types, while its paired histological observation is likewise a local H&E patch [5, 16]. As a result, spot-level expression can vary substantially even within the same anatomical region because of cellular composition, measurement sparsity, and local biological heterogeneity [5, 15]. Spatial context is therefore essential: methods that exploit neighboring relationships recover more coherent tissue domains than purely pointwise clustering [3, 5, 17]. This motivates neighborhood-aware representations in which nearby spots contribute contextual information to the spot representation [3, 5, 16, 18].

The *third major challenge* is that spatial proximity does not imply equal neighbor informativeness. Coordinate-based neighborhoods provide useful candidate interactions, but they can also connect spots across biological boundaries [5, 16]. STAIG addresses such cross-domain edges through morphology-guided edge dropping, while MuCoST supplements spatial adjacency with a separate co-expression graph [16, 18]. GraphST aggregates coordinate-defined neighbors with a GCN encoder [5], MuCoST learns from spatial and co-expression graphs with a shared GCN autoencoder [18], and STAIG propagates information over morphology-augmented graphs [16]. Thus, although these methods adapt neighborhood structure in different ways, information exchange is still primarily through graph construction or propagation rather than an explicit center-conditioned molecular comparison with each neighbor [3, 5, 16, 18]. The key challenge is therefore not only determining *which spots are adjacent*, but *how informative each neighbor is for the specific center spot* [16, 18].

The *fourth major challenge* is that *histological similarity* is informative but not sufficient by itself. For example, DLPFC layers were annotated using both cytoarchitecture and molecular markers (e.g., *MBP*, *PCP4*), rather than H&E morphology alone [7]. Conversely, gene expressions are also imperfect, with several cortical marker genes exhibiting noisy spot-level expression [15]. Histology therefore provides complementary structural evidence, as demonstrated by morphology-aware methods such as STAIG [16], but image-derived cues can vary with staining and appearance [16]. Thus, neither morphology nor gene expression alone is a sufficient semantic signal; the challenge is to use histology as complementary evidence while preserving molecular information as the primary representation of spot identity [7, 15, 16].

Our Work: To address these challenges, we introduce MP-MNCA, a morphology-prior masked neighbor cross-attention framework that represents each spot as its intrinsic molecular identity together with a context-dependent neighborhood correction. We retain the full $G = 3000$ HVG

representation and use each center spot as a query over its $k = 6$ coordinate-defined spatial neighbors, whose expression profiles provide keys and values for center-specific molecular compatibility. Histological similarity and relative spatial position enter the attention logits as soft priors, while masked two-view self-supervision learns the representation without anatomical domain labels. The resulting neighborhood context is added residually to the original spot profile,

$$\hat{\mathbf{x}}_i = \text{LN}(\mathbf{x}_i + \Delta\mathbf{x}_i^{\text{context}}), \quad (1)$$

which we term *contextual molecular augmentation*. This preserves original molecular dimensionality while enabling selective, morphology-guided information exchange between spatial neighbors. Our *main contributions* are:

- **Full-dimensional morphology-prior cross-attention.** We introduce a neighbor cross-attention encoder that preserves the 3000-dimensional HVG-representation and learns center-specific information transfer between spatial neighbors. Histological similarity and relative position act as soft attention priors instead of replacing molecular similarity or defining a fixed propagation operator as in Eq. 1.

- **Masked self-supervised neighborhood learning.** We train MP-MNCA independently on each tissue section using two independently masked gene-expression views and a spatial neighbor-contrastive objective, without using anatomical domain labels during representation learning. The resulting embedding preserves intrinsic spot identity through residual contextual molecular augmentation.

- **Convergence-focused spatial-domain evaluation.** Beyond final-checkpoint performance, we evaluate how spatial-domain structure emerges throughout optimization using checkpoint-wise ARI/NMI trajectories and epoch-wise spatial-domain visualizations. Across the 12 DLPFC sections, MP-MNCA reaches coherent spatial organization substantially earlier than established spatial-domain baselines while retaining competitive peak clustering quality.

Experiments: We evaluate MP-MNCA on all 12 DLPFC sections against SpaGCN, GraphST, MuCoST, and STAIG [3, 5, 16, 18]. Beyond late-stage performance, we track refined adjusted Rand index (ARI) [4] and normalized mutual information (NMI) [14] throughout training and examine epoch-wise spatial domain visualizations to assess both numerical and qualitative convergence. We further report section-wise peak performance and ablations of molecular cross-attention, morphology priors, and positional information. MP-MNCA reaches competitive spatial-domain structure in substantially fewer epochs, with mean peak refined ARI of **0.546** versus **0.549** for the strongest baseline and the highest mean refined NMI of **0.657** versus **0.640**. It further outperforms STAIG on **7/12** sections in ARI and in **11/12** sections in NMI, demonstrating faster convergence without sacrificing spatial-domain quality.

2. Related Work

Graph-based spatial representation learning: A common formulation in spatial transcriptomics represents capture spots as graph nodes, associates each node with its gene-expression profile, and uses graph structure to determine which spots exchange information. The principal distinction between existing methods lies in how these inter-spot relationships are constructed and exploited. SpaGCN incorporates spatial coordinates and histological information into its graph and applies graph convolution to aggregate expression from neighboring spots [3]. GraphST instead constructs a spatial neighborhood graph directly from spot coordinates and employs a GCN encoder with self-supervised contrastive learning to produce spatially informed latent representations [5]. MuCoST extends this formulation through multiple graph views: a spatial-adjacency graph captures local proximity, a co-expression graph connects spots with similar molecular profiles, and a shuffled graph provides a repulsive contrastive view; these graphs are processed using a shared GCN autoencoder [18]. STAIG similarly begins with a coordinate-derived spatial graph in which gene-expression profiles constitute the node information, but uses self-supervised histological embeddings to guide graph augmentation by controlling edge-removal probabilities and filtering potentially homologous negative pairs during contrastive learning [16].

Despite their different graph-construction and training strategies, these approaches principally encode neighborhood information through graph propagation or graph augmentation, typically producing a compressed latent representation of each spot. MP-MNCA retains the useful spatial-neighborhood assumption but changes the mechanism by which molecular information is exchanged. Spatial coordinates define only a small candidate neighborhood; they do not prescribe how strongly every neighbor should contribute. Instead, the center spot queries the full HVG representations of its neighbors through learned cross-attention, producing center-dependent molecular routing. Histology is introduced as a continuous morphological prior on the corresponding attention logits, so that morphological evidence modulates *how much molecular information should be trusted from each neighbor* rather than serving as another feature stream to be fused or merely determining a graph augmentation. Thus, genes carry transmitted content, spatial proximity determines candidate interactions, and morphology modulates reliability of those interactions.

Self-supervised visual features: We use frozen BYOL embeddings for morphology. BYOL learns visual representations by matching online and exponentially averaged target networks without explicit negative samples [2]. In our framework, these image embeddings are not further trained by the ST objective; only their local similarities enter the attention prior.

3. Method

Overview: Let a tissue section contain n spatial transcriptomics spots. For spot i , we denote its gene-expression vector, spatial coordinate, and local histology patch by

$$\mathbf{x}_i \in \mathbb{R}^G, \quad \mathbf{p}_i \in \mathbb{R}^2, \quad \mathcal{I}_i, \quad (2)$$

respectively. We use the top $G = 3000$ highly variable genes (HVGs) and learn a full-dimensional representation

$$f_\theta : (X, M, P) \rightarrow Z, \quad Z \in \mathbb{R}^{n \times G}. \quad (3)$$

Our framework consists of two stages. First, a *multi-view BYOL encoder* learns a morphology representation from the histology surrounding each spot. Second, *MP-MNCA* uses the resulting morphology similarity as a soft prior for center-conditioned molecular cross-attention over spatial neighbors. The morphology representation controls *where molecular information is routed*; the transmitted content itself always comes from gene expression.

3.1. Stage I: Multi-View Morphology Encoding

For each spot i , we extract a local histological patch $\mathcal{I}_i$ centered at its spatial coordinate. To learn morphology feature *representation* that are invariant to nuisance appearance variation, we employ a BYOL encoder with multiple stochastic views of the same patch. Specifically, two transformations $t_1, t_2 \sim \mathcal{T}$ produce

$$\mathcal{I}_i^{(1)} = t_1(\mathcal{I}_i), \quad \mathcal{I}_i^{(2)} = t_2(\mathcal{I}_i). \quad (4)$$

The online and target BYOL branches are optimized to produce consistent representations for these two views without requiring tissue annotations. After pretraining, the image encoder is frozen and a $d_m = 128$ dimensional embedding $\mathbf{m}_i \in \mathbb{R}^{128}$ is extracted for every spot.

The morphology encoder is not jointly optimized with the molecular model. Instead, image embeddings are standardized independently within each tissue section and reduced to 16 dimensions using Principal Component Analysis (PCA) [11] solely for computing the morphology prior:

$$\bar{M} = \text{PCA}_{16}(\text{Standardize}(M)) \quad (5)$$

Importantly, image embeddings are never concatenated with gene expression and are never used as molecular value vectors. For a center spot i and neighbor j , we define their morphology agreement using cosine similarity,

$$s_{ij}^{\text{img}} = \frac{\langle \bar{\mathbf{m}}_i, \bar{\mathbf{m}}_j \rangle}{\|\bar{\mathbf{m}}_i\|_2 \|\bar{\mathbf{m}}_j\|_2}, \quad r_{ij} = \frac{1 + s_{ij}^{\text{img}}}{2} \in [0, 1]. \quad (6)$$

Thus, morphology provides a continuous measure of whether a neighboring spot is visually compatible with the center rather than defining a hard image-derived graph or cluster assignment.

MP-MNCA: Morphology-Prior Masked Neighbour Cross-Attention

Spatial domain identification in DLPFC

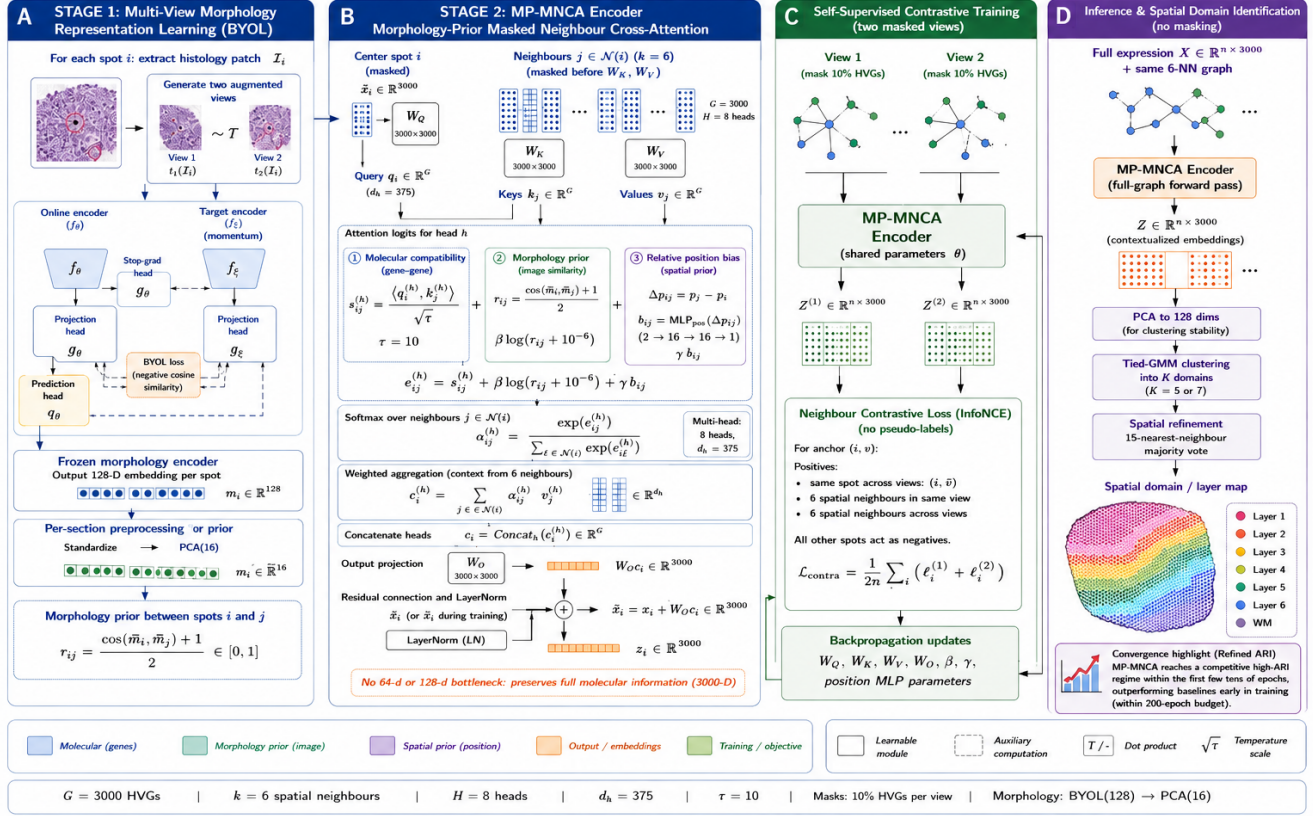


Figure 1. **MP-MNCA architecture:** (A) Multi-view BYOL learns frozen morphology embeddings used to construct a continuous image-similarity prior. (B) MP-MNCA performs full-dimensional center-to-neighbor molecular cross-attention over six spatial neighbors, combining gene compatibility, morphology prior, and relative-position bias before residual contextualization. (C) Two independently masked molecular views are optimized with a self-supervised neighbor-contrastive loss without pseudo-labels or layer supervision. (D) Inference uses unmasked expression followed by PCA-assisted tied-GMM clustering and spatial refinement.

3.2. Stage II: Morphology-Prior Masked Neighbor Cross-Attention

Spatial candidate neighborhood: For each center spot i , we define a local seven-spot receptive field consisting of the center and its six nearest spatial neighbors. Let

$$\mathcal{N}(i) = \text{kNN}_{j \neq i, 6}(\mathbf{p}_i), \quad \mathcal{S}(i) = \{i\} \cup \mathcal{N}(i), \quad (7)$$

where $|\mathcal{N}(i)| = 6$ and $|\mathcal{S}(i)| = 7$. Consequently, each local receptive field contains seven molecular profiles: the center $\mathbf{x}_i$ and six neighboring profiles $\{\mathbf{x}_j : j \in \mathcal{N}(i)\}$. The coordinate graph only determines which spots are eligible to exchange information, not their aggregation weights.

Masked molecular views: To learn representations robust to partial molecular corruption, two independent views of the complete HVG matrix are generated during training.

For $v \in \{1, 2\}$,

$$\tilde{\mathbf{x}}_\ell^{(v)} = \mathbf{m}^{(v)} \odot \mathbf{x}_\ell, \quad (8)$$

where $\mathbf{m}^{(v)} \in \{0, 1\}^G$ masks approximately 10% of gene dimensions. The selected gene columns are shared across all spots within a view, while the two views are masked independently. Spatial neighborhoods and image-derived morphology priors remain unchanged.

Center-conditioned molecular cross-attention: For each center spot, the center molecular representation generates the query, while its six spatial neighbors generate keys and values:

$$\mathbf{q}_i = W_Q \tilde{\mathbf{x}}_i, \quad \mathbf{k}_j = W_K \tilde{\mathbf{x}}_j, \quad \mathbf{v}_j = W_V \tilde{\mathbf{x}}_j, \quad (9)$$

where $W_Q, W_K, W_V \in \mathbb{R}^{G \times G}$ and $j \in \mathcal{N}(i)$. The model therefore retains the full $G = 3000$ dimensional molecular state rather than introducing a learned low-dimensional bottleneck.

We use $H = 8$ attention heads, giving $d_h = \frac{G}{H} = 375$. For head h , molecular compatibility between center i and neighbor j is

$$s_{ij}^{\text{gene},(h)} = \frac{\langle \mathbf{q}_i^{(h)}, \mathbf{k}_j^{(h)} \rangle}{\sqrt{\tau}}, \quad (10)$$

where $\tau = 10$ is the configured attention scale.

Morphology-prior routing: Molecular compatibility alone may assign high attention to a transcriptionally similar neighbor whose local tissue morphology is inconsistent with the center. We therefore inject the image-derived similarity from Eq. (6) directly into attention logits. Spatial orientation is further encoded using the relative coordinate

$$\Delta \mathbf{p}_{ij} = \mathbf{p}_j - \mathbf{p}_i, \quad (11)$$

and a positional network $\text{MLP}_{\text{pos}} : \mathbb{R}^2 \rightarrow \mathbb{R}^{16} \rightarrow \mathbb{R}$,

$$b_{ij} = \text{MLP}_{\text{pos}}(\Delta \mathbf{p}_{ij}), \quad (12)$$

with GELU non linearity. The resulting scalar b_{ij} allows the attention mechanism to learn direction- and displacement-dependent spatial preferences rather than treating all six neighboring positions equivalently. The complete attention energy for head h is

$$e_{ij}^{(h)} = \underbrace{\frac{\langle \mathbf{q}_i^{(h)}, \mathbf{k}_j^{(h)} \rangle}{\sqrt{\tau}}}_{\text{molecular compatibility}} + \underbrace{\beta \log(r_{ij} + \epsilon)}_{\text{morphology prior}} + \underbrace{\gamma b_{ij}}_{\text{pos. bias}} \quad (13)$$

with $\epsilon = 10^{-6}$. The scalars β and γ are learnable and are initialized to 1.0 and 0.5, respectively. Attention is normalized only over the six eligible neighbors,

$$\alpha_{ij}^{(h)} = \frac{\exp(e_{ij}^{(h)})}{\sum_{\ell \in \mathcal{N}(i)} \exp(e_{i\ell}^{(h)})}. \quad (14)$$

Equivalently, $\alpha_{ij}^{(h)} \propto \exp(s_{ij}^{\text{gene},(h)}) r_{ij}^\beta \exp(\gamma b_{ij})$, showing that morphology acts as a multiplicative routing prior. It changes the relative probability that a neighboring molecular message is transmitted, without replacing that message with image information.

Contextual molecular augmentation: Each attention head computes a center-specific molecular context

$$\Delta \mathbf{c}_i^{(h)} \equiv \mathbf{c}_i^{(h)} = \sum_{j \in \mathcal{N}(i)} \alpha_{ij}^{(h)} \mathbf{v}_j^{(h)} \in \mathbb{R}^{d_h}. \quad (15)$$

The head-wise contexts are concatenated and projected back to the original HVG space:

$$\mathbf{c}_i = W_O \text{Concat}_{h=1}^H (\mathbf{c}_i^{(h)}), W_O \in \mathbb{R}^{G \times G}. \quad (16)$$

Crucially, the neighboring context *augments* the center spot rather than replacing it. The **final representation** is

$$\mathbf{z}_i = \text{LN}(\mathbf{x}_i + \mathbf{c}_i) \quad (17)$$

for an unmasked input. During masked-view training, $\mathbf{x}_i$ in Eq. (17) is replaced by the respective masked center view.

Self-supervised neighbor contrastive learning: The same MP-MNCA parameters process the two independently masked molecular views, producing

$$\mathbf{Z}^{(1)}, \mathbf{Z}^{(2)} \in \mathbb{R}^{n \times G}. \quad (18)$$

Let

$$\bar{\mathbf{z}}_i^{(v)} = \frac{\mathbf{z}_i^{(v)}}{\|\mathbf{z}_i^{(v)}\|_2} \quad (19)$$

denote the normalized representation. For an anchor (i, v) , the positive set contains

$$\mathcal{P}_i^{(v)} = \{(i, \bar{v})\} \cup \{(j, v) : j \in \mathcal{N}(i)\} \cup \{(j, \bar{v}) : j \in \mathcal{N}(i)\}, \quad (20)$$

where $\bar{v}$ denotes the opposite masked view. Thus, each anchor has the same spot in the other view together with its six spatial neighbors in both views as positives.

For any pair of normalized embeddings, define

$$\text{sim}(\bar{\mathbf{z}}_i^{(v)}, \bar{\mathbf{z}}_j^{(u)}) = \exp\left(\frac{\langle \bar{\mathbf{z}}_i^{(v)}, \bar{\mathbf{z}}_j^{(u)} \rangle}{\tau_c}\right), \quad (21)$$

where τ_c is the contrastive temperature. Let $\mathcal{A}_i^{(v)}$ contain all valid non-anchor embeddings from both views. The directional multi-positive contrastive loss is

$$\ell_i^{(v)} = -\log \frac{\sum_{(j,u) \in \mathcal{P}_i^{(v)}} \text{sim}(\bar{\mathbf{z}}_i^{(v)}, \bar{\mathbf{z}}_j^{(u)})}{\sum_{(j,u) \in \mathcal{A}_i^{(v)}} \text{sim}(\bar{\mathbf{z}}_i^{(v)}, \bar{\mathbf{z}}_j^{(u)})}. \quad (22)$$

The final objective is symmetric across the two views,

$$\mathcal{L}_{\text{contra}} = \frac{1}{2n} \sum_{i=1}^n (\ell_i^{(1)} + \ell_i^{(2)}). \quad (23)$$

All remaining spot representations contribute through the contrastive denominator. In particular, **no image KMeans assignments, gene pseudo-labels, anatomical labels, or pseudo-label-based negative masking are used during MP-MNCA training**. Backpropagation therefore learns W_Q, W_K, W_V, W_O , the positional MLP, and the routing calibration parameters β and γ purely from masked-view consistency, spatial locality, gene-expression compatibility, and the continuous morphology prior.

Table 1. **Peak-checkpoint spatial-domain identification on the 12 DLPFC sections.** We report the maximum refined ARI observed within the same 200-epoch optimization budget and the corresponding refined NMI at that checkpoint, with the selected epoch shown in parentheses. This complementary analysis assesses whether MP-MNCA’s faster convergence is achieved without sacrificing peak clustering quality. Higher is better, and the best result for each section and metric is shown in bold.

Section	STAIG		GraphST		MuCoST		MP-MNCA	
	ARI (ep.) $\uparrow$	NMI $\uparrow$	ARI (ep.) $\uparrow$	NMI $\uparrow$	ARI (ep.) $\uparrow$	NMI $\uparrow$	ARI (ep.) $\uparrow$	NMI $\uparrow$
151507	0.5786 (90)	0.6458	0.4848 (200)	0.6507	0.5556 (100)	0.6638	0.5734 (15)	0.6884
151508	0.4938 (200)	0.6334	0.5133 (1)	0.5565	0.4154 (115)	0.5341	0.5459 (195)	0.6894
151509	0.6310 (160)	0.6548	0.4915 (195)	0.5406	0.5799 (120)	0.6893	0.5420 (50)	0.6803
151510	0.5702 (125)	0.6525	0.4185 (155)	0.5853	0.5480 (110)	0.6499	0.5704 (5)	0.6600
151669	0.4444 (200)	0.5008	0.4417 (190)	0.4967	0.4908 (185)	0.5494	0.4549 (10)	0.5038
151670	0.5488 (190)	0.5974	0.2498 (130)	0.4209	0.5803 (20)	0.5239	0.4221 (10)	0.4473
151671	0.5953 (45)	0.6864	0.8064 (200)	0.7332	0.7464 (195)	0.7232	0.6054 (15)	0.7207
151672	0.5762 (40)	0.6574	0.6756 (10)	0.6352	0.7151 (180)	0.7145	0.6064 (150)	0.7245
151673	0.5556 (40)	0.6556	0.3673 (1)	0.5392	0.4791 (50)	0.5901	0.5287 (45)	0.6898
151674	0.5248 (115)	0.6503	0.4284 (15)	0.5412	0.3442 (1)	0.5205	0.5663 (90)	0.6902
151675	0.5151 (185)	0.6715	0.3700 (5)	0.5504	0.3216 (5)	0.5142	0.5989 (40)	0.7100
151676	0.5587 (165)	0.6686	0.3822 (5)	0.5075	0.2133 (50)	0.3974	0.5312 (85)	0.6730
Mean $\pm$ SD	0.5494 $\pm$ 0.0472	0.6395 $\pm$ 0.0468	0.4691 $\pm$ 0.1414	0.5631 $\pm$ 0.0773	0.4991 $\pm$ 0.1503	0.5892 $\pm$ 0.0953	0.5455 $\pm$ 0.0547	0.6565 $\pm$ 0.0836
Median [IQR]	0.5572 [0.0544]	0.6537 [0.0175]	0.4351 [0.1178]	0.5458 [0.0665]	0.5194 [0.1824]	0.5698 [0.1471]	0.5561 [0.0492]	0.6889 [0.0254]

4. Experiments

4.1. Dataset and Evaluation Protocol

We evaluate on the 12 human dorsolateral prefrontal cortex (DLPFC) sections introduced by Maynard *et al.* [7]. The sections span three donors and contain approximately 3,500–4,800 spatial spots each, with annotated cortical layers and white matter providing the reference spatial domains. Performance is evaluated using Adjusted Rand Index (ARI) and Normalized Mutual Information (NMI).

Our evaluation is designed primarily to compare the *convergence behavior of the model architectures*. To minimize differences arising from initialization, all methods use the same section-wise preprocessing, spatial neighborhood construction, cluster count, and common pre-model initialization protocol. Thus, for a given section, each method starts from the same initialized molecular/spatial state before being passed through its respective architecture; only the subsequent representation-learning and message-routing mechanisms differ between models. Architecture-specific trainable parameters necessarily retain their respective parameterizations, but are initialized under the same deterministic experimental protocol.

Following this common starting point, MP-MNCA, STAIG, GraphST, and MuCoST are optimized and evaluated over the same 200-epoch budget. Refined ARI is measured throughout training and aggregated across the 12 sections to characterize how rapidly each architecture develops spatially discriminative representations and whether this quality is maintained as optimization proceeds. We additionally report the best ARI with NMI at the corresponding checkpoint, to compare attainable clustering quality alongside convergence speed.

4.2. Implementation Details

Gene expression is represented using 3,000 log1p-normalized HVGs with six coordinate-defined neighbors per spot. Morphology is encoded using frozen 128-dimensional BYOL features, standardized per section and reduced to 16 dimensions by PCA. MP-MNCA uses eight attention heads with $d_k = 375$, preserving the full 3,000-dimensional molecular representation. During training, 10% of HVGs are independently masked in each view before the $Q/K/V$ projections; β and γ are initialized to 1 and 0.5, respectively.

We use attention dropout 0.1 and Adam with learning rate 3×10^{-4} and weight decay 10^{-5} . For convergence evaluation, MP-MNCA, STAIG, GraphST, and MuCoST are trained for 200 epochs under the matched protocol described above, while retaining method-specific repository-supported optimization settings. We compare epoch-wise convergence rather than wall-clock runtime, as per-epoch computational cost differs across architectures.

4.3. Ablation Protocol

The ablation study isolates the principal components of the proposed routing mechanism. Removing the morphology prior ($\beta = 0$) tests whether histological similarity contributes beyond molecular compatibility, while removing the position term ($\gamma = 0$) evaluates the additional role of relative spatial geometry. Setting $\beta = \gamma = 0$ isolates molecular cross-attention, and replacing learned attention coefficients with uniform $1/k$ weights tests adaptive routing against fixed neighborhood averaging. Low-dimensional variants further test whether the full-HVG representation is important for the observed behavior.

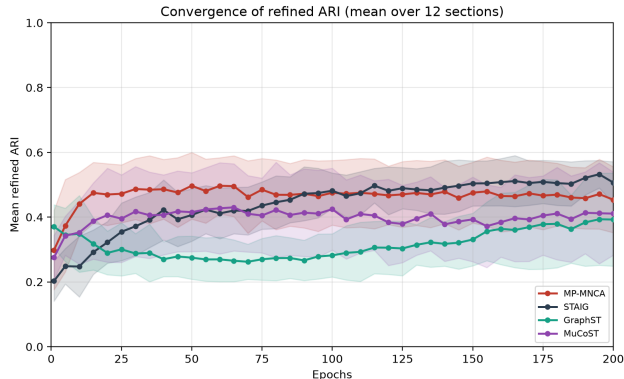


Figure 2. **Convergence of refined ARI across the 12 DLPFC sections.** Mean refined ARI over 200 epochs with $\pm$ one standard deviation across sections. MP-MNCA reaches a competitive regime substantially earlier than the baselines.

5. Results

5.1. Convergence Across DLPFC Sections

Figure 8 shows the evolution of refined ARI averaged over all 12 DLPFC sections. MP-MNCA reaches a competitive high-ARI regime within the first 10–20 epochs and remains comparatively stable thereafter, whereas the baselines require substantially longer optimization or stabilize at lower mean performance. Although STAIG eventually becomes comparable and occasionally exceeds MP-MNCA at later checkpoints, MP-MNCA reaches similar clustering quality much earlier, highlighting its stronger convergence behavior. The overlap between the section-wise variance bands at later epochs further motivates evaluating peak attainable quality separately rather than claiming uniform dominance throughout optimization.

5.2. Peak Attainable Clustering Quality

Table 1 shows that faster convergence does not cost peak quality. Within the same 200-epoch budget, MP-MNCA attains a mean peak ARI of 0.5455 ± 0.0547 , essentially matching STAIG (0.5494 ± 0.0472) and exceeding GraphST and MuCoST, while achieving the highest mean NMI of 0.6565 ± 0.0836 versus 0.6395 ± 0.0468 for STAIG, with the best NMI on 8 of 12 sections against 3 for STAIG, with median peak epochs of 42.5 and 142.5.

5.3. Qualitative Analysis and Ablations

The same dynamics are visible in the predicted partitions. Figure 3 shows section 151507 at epochs 10, 20, 50, 100, and 200: MP-MNCA recovers the dominant laminar organization within 10–20 epochs and keeps coherent boundaries thereafter, whereas STAIG shows substantial early lo-

cal mixing before organizing, and GraphST and MuCoST show greater boundary mixing or deformation at intermediate checkpoints. MP-MNCA peaks at ARI 0.5734 (epoch 15) against 0.5786 for STAIG (epoch 90), with GraphST and MuCoST peaking at epochs 200 and 100, and attains the highest NMI on this section (0.6884).

Ablations support the routing design. Compressing the molecular representation to 128 dimensions drops mean refined ARI to 0.272, 0.243, and 0.234 for the contrastive, gene-cosine, and BYOL variants, confirming the value of the full HVG representation. The $\beta = 0$, $\gamma = 0$, $\beta = \gamma = 0$, and uniform-neighbor controls isolate the contributions of morphology, position, molecular compatibility, and adaptive neighbor weighting under the same six-neighbor support and evaluation pipeline.

5.4. Attention Routing and Early Convergence

The observed convergence profile is consistent with the inductive structure of the attention block:

$$\alpha_{ij} \propto \exp\left(\frac{q_i^\top k_j}{\sqrt{d}} + \beta m_{ij} + \gamma p_{ij}\right), \quad (24)$$

where $j \in \mathcal{N}_6(i)$, and the six-neighbor graph only defines candidate support, while molecular compatibility, morphology, and relative position determine routing strength. This selects informative local context directly in the full HVG space, while the residual preserves the center representation. The resulting inductive bias is consistent with the rapid early ARI increase in Figure 8, though causal attribution requires component-wise convergence ablations.

6. Discussion and Conclusion

MP-MNCA preserves the full molecular state and learns neighbor-specific routing before compression; the degradation of the 128-dimensional controls shows the gains are not from low-dimensional fusion or spatial smoothing alone. The architecture separates topology, prior, and content: coordinates define candidate neighbors, molecular compatibility sets contribution, and histology modulates reliability through an additive prior that acts multiplicatively after exponentiation-discounting an unreliable molecular message without fusing morphology into the representation. This matters most near tissue boundaries, where morphology and molecular identity diverge.

Evaluation is restricted to DLPFC, but the benchmark spans multiple sections and heterogeneous subjects, and the consistency across sections suggests the gains are not driven by a single tissue instance; additional datasets would nevertheless establish cross-dataset generality. The four 3000×3000 projections introduce roughly 36M parameters, motivating low-rank or shared projections.

Because morphology only reweights message passing, the two modalities never share a coordinate system.

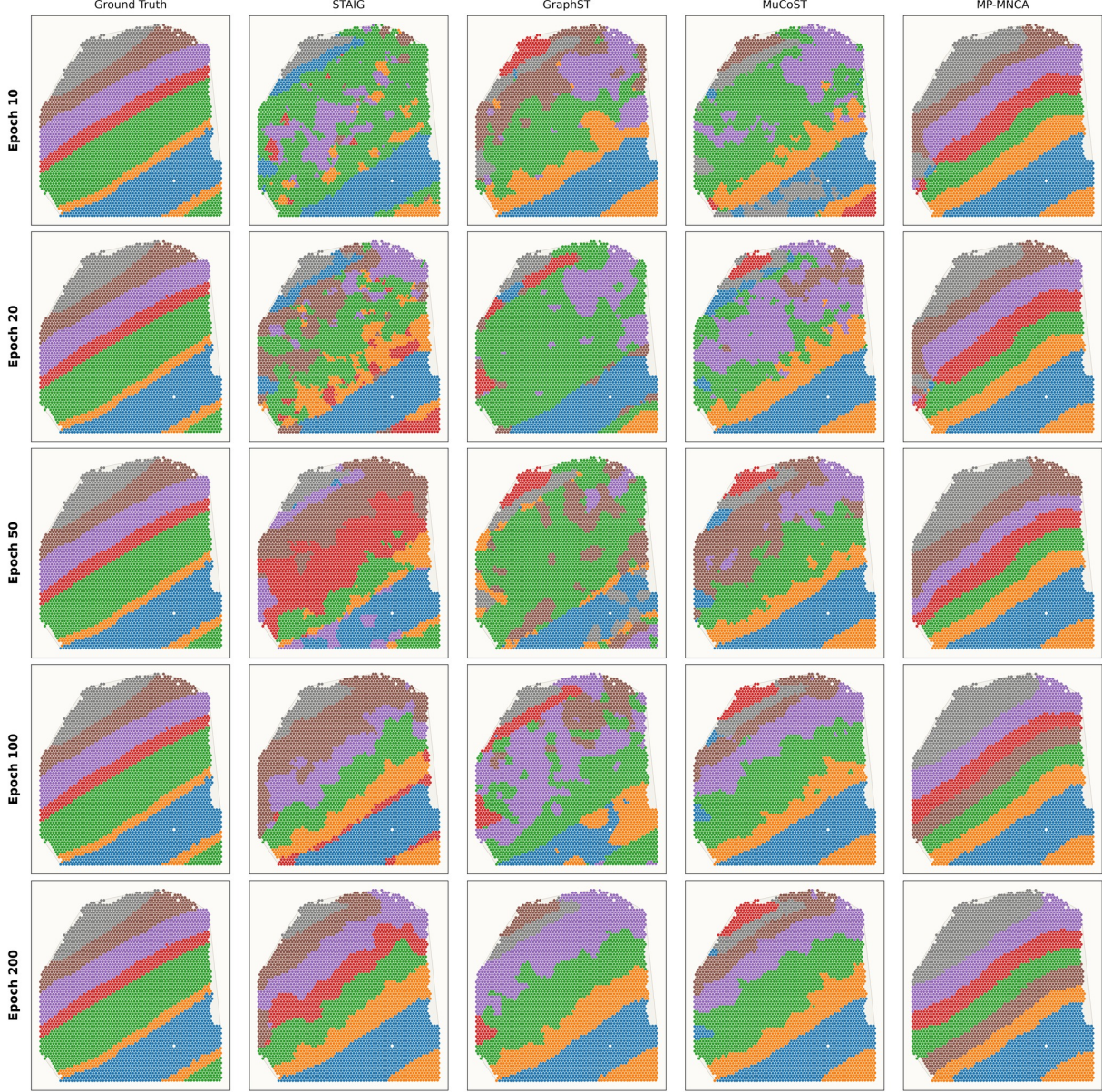


Figure 3. **Qualitative convergence on DLPFC section 151507.** Spatial-domain predictions at epochs 10, 20, 50, 100, and 200 for STAIG, GraphST, MuCoST, and MP-MNCA. MP-MNCA recovers coherent laminar structure earlier, while competing methods show slower organization or greater domain mixing. Ground truth is shown for reference; evaluation uses permutation-invariant ARI and NMI.

We therefore also investigated **SPARC-align**, in which concept-aligned sparse autoencoders [9] encode expression and histology into a shared dictionary with consistent per-dimension meaning across modalities, passing the resulting concept codes to the cross-attention stage of Sec. 8. Two adaptations proved important: smoothing each spot over its histology-weighted neighborhood before encoding, since sparse autoencoders assume i.i.d. samples while capture spots are spatially organized; and giving each modality private dictionary capacity alongside the shared block,

since expression and morphology are asymmetrically predictive. With both, SPARC-align approaches established baselines while yielding an interpretable cross-modal concept space that none of the compared methods provide, making concept-aligned sparse coding a promising direction for multimodal spatial transcriptomics. Full formulation, per-section results, convergence, ablations, and limitations appear in the supplementary material.

References

- [1] Michaela Asp, Joseph Bergenstr hle, and Joakim Lundberg. Spatially resolved transcriptomes—next generation tools for tissue exploration. *BioEssays*, 42(10): 1900221, 2020. doi: 10.1002/bies.201900221. [1](#)
- [2] Jean-Bastien Grill, Florian Strub, Florent Altch , et al. Bootstrap your own latent: A new approach to self-supervised learning. In *Advances in Neural Information Processing Systems*, volume 33, 2020. [3](#)
- [3] Jian Hu, Xiangjie Li, Kyle Coleman, Amelia Schroeder, Nan Ma, David J. Irwin, Edward B. Lee, Russell T. Shinohara, and Mingyao Li. SpaGCN: Integrating gene expression, spatial location and histology to identify spatial domains and spatially variable genes by graph convolutional network. *Nature Methods*, 18(11):1342–1351, 2021. doi: 10.1038/s41592-021-01255-8. [1](#), [2](#), [3](#)
- [4] Lawrence Hubert and Phipps Arabie. Comparing partitions. *Journal of Classification*, 2(1):193–218, 1985. doi: 10.1007/BF01908075. [2](#)
- [5] 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*, 14:1155, 2023. doi: 10.1038/s41467-023-36796-3. [1](#), [2](#), [3](#)
- [6] Sophia K. Longo, Margaret G. Guo, Andrew L. Ji, and Paul A. Khavari. Integrating single-cell and spatial transcriptomics to elucidate intercellular tissue dynamics. *Nature Reviews Genetics*, 22:627–644, 2021. doi: 10.1038/s41576-021-00370-8. [1](#)
- [7] Kristen R. Maynard, Leonardo Collado-Torres, Lukas M. Weber, Cedric Uytingco, Brianna K. Barry, Stephen R. Williams, Joseph L. Catallini II, Matthew N. Tran, Zachary Besich, Madhavi Tippani, Jennifer Chew, Yifeng Yin, Joel E. Kleinman, Thomas M. Hyde, Nikhil Rao, Stephanie C. Hicks, Keri Martinowich, and Andrew E. Jaffe. Transcriptome-scale spatial gene expression in the human dorsolateral prefrontal cortex. *Nature Neuroscience*, 24:425–436, 2021. doi: 10.1038/s41593-020-00787-0. [1](#), [2](#), [6](#)
- [8] Lambda Moses and Lior Pachter. Museum of spatial transcriptomics. *Nature Methods*, 19:534–546, 2022. doi: 10.1038/s41592-022-01409-2. [1](#)
- [9] Ali Nasiri-Sarvi, Hassan Rivaz, and Mahdi S. Hosseini. SPARC: Concept-aligned sparse autoencoders for cross-model and cross-modal interpretability. *arXiv preprint arXiv:2507.06265*, 2025. [8](#), [11](#)
- [10] Brenda Pardo, Abby Spangler, Lukas M. Weber, Stephanie C. Page, Stephanie C. Hicks, Andrew E. Jaffe, Keri Martinowich, Kristen R. Maynard, and Leonardo Collado-Torres. spatialLIBD: an R/Bioconductor package to visualize spatially-resolved transcriptomics data. *BMC Genomics*, 23(1): 434, 2022. doi: 10.1186/s12864-022-08601-w. [1](#)
- [11] Karl Pearson. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572, 1901. doi: 10.1080/14786440109462720. [3](#)
- [12] Lizhong Shang and Xiang Zhou. Spatially aware dimension reduction for spatial transcriptomics. *Nature Communications*, 13:7203, 2022. doi: 10.1038/s41467-022-34879-1. [1](#)
- [13] Patrik L. St hl, Fredrik Salm n, Sanja Vickovic, Anna Lundmark, Jos  Fern ndez Navarro, Jens Magnusson, Stefania Giacomello, Michaela Asp, Jakub O. Westholm, Mikael Huss, Annelie Mollbrink, Sten Linnarsson, Simone Codeluppi,  ke Borg, Fredrik Pont n, Paul Igor Costea, Pelin Sahl n, Jan Mulder, Olaf Bergmann, Joakim Lundberg, and Jonas Fris n. Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science*, 353(6294):78–82, 2016. doi: 10.1126/science.aaf2403. [1](#)
- [14] Alexander Strehl and Joydeep Ghosh. Cluster ensembles—a knowledge reuse framework for combining multiple partitions. *Journal of Machine Learning Research*, 3:583–617, 2002. [2](#)
- [15] Zhen Tang, Zheng Li, Ting Hou, Tao Zhang, Bo Yang, Jie Su, and Qing Song. SiGra: Single-cell spatial elucidation through an image-augmented graph transformer. *Nature Communications*, 14:5618, 2023. [2](#)
- [16] Yitao Yang, Yang Cui, Xin Zeng, Yubo Zhang, Martin Loza, Sung-Joon Park, et al. STAIG: Spatial transcriptomics analysis via image-aided graph contrastive learning for domain exploration and alignment-free integration. *Nature Communications*, 16:1067, 2025. doi: 10.1038/s41467-025-56276-0. [1](#), [2](#), [3](#)
- [17] Zhiyuan Yuan, Fangyuan Zhao, Senlin Lin, Yu Zhao, Jianhua Yao, Yan Cui, Xiao-Yong Zhang, and Yi Zhao. Benchmarking spatial clustering methods with spatially resolved transcriptomics data. *Nature Methods*, 21:712–722, 2024. doi: 10.1038/s41592-024-02215-8. [1](#), [2](#)
- [18] Lei Zhang, Shu Liang, and Lin Wan. A multi-view graph contrastive learning framework for deciphering spatially resolved transcriptomics data. *Briefings in Bioinformatics*, 25(4):bbae255, 2024. doi: 10.1093/bib/bbae255. [1](#), [2](#), [3](#)
- [19] Edward Zhao, Matthew R. Stone, Xing Ren, Jamie Guenthoer, Kimberly S. Smythe, Thomas Pulliam,

Stephen R. Williams, Cedric R. Uytingco, Sarah E. B. Taylor, Paul Nghiem, Jason H. Bielas, and Raphael Gottardo. Spatial transcriptomics at subspot resolution with BayesSpace. *Nature Biotechnology*, 39:1375–1384, 2021. doi: 10.1038/s41587-021-00935-2. [1](#)

This supplementary includes **SPARC-align**, the concept-aligned cross-modal sparse-coding variant referenced. We describe the motivation (Sec. 7), the method (Sec. 8), results on all twelve DLPFC sections (Sec. 9), and the limitations that determine where we would take the approach next (Sec. 10). All numbers are produced under the evaluation protocol of the main paper, and all training is fully unsupervised.

7. From MP-MNCA to SPARC-align

MP-MNCA routes molecular information across spatial neighbors using histological similarity as a continuous attention prior. Morphology therefore enters as a *weighting* on message passing: it decides how strongly a neighbor contributes, but the representation being learned remains a refinement of gene expression. The two modalities are never placed in a common coordinate system, so one cannot ask which morphological structure a given molecular program corresponds to.

That question motivates a different formulation. If gene expression and histology were encoded into a *shared* latent space in which an individual dimension carried the same meaning for both, the resulting code would be interpretable across modalities rather than merely informed by one of them. Sparse autoencoders let us do the same via dictionary learning: SPARC [9] learns per-stream encoders into a common dictionary and enforces alignment by selecting a single shared set of active dimensions from the summed pre-activations, so both streams must explain a sample using the same concepts.

We therefore built **SPARC-align**: a two-phase pipeline in which a concept-aligned sparse autoencoder replaces MP-MNCA’s implicit fusion, followed by MP-MNCA’s neighbor cross-attention operating on the resulting concept codes. Applying SPARC to spatial transcriptomics is not, however, a direct transfer. SPARC assumes i.i.d. samples, whereas capture spots are spatially organized: in DLPFC, a spot and its six nearest neighbors share a cortical layer 92.3% of the time. Sec. 8 describes two modifications that follow from this—spatially smoothing the inputs before encoding, and relaxing the shared support so each modality retains private dictionary capacity. These were the only changes, among many evaluated, that improved clustering.

8. Method

8.1. Streams and notation

Each section is processed independently. For section with n spots we take two frozen streams: gene expression $\mathbf{x}_i^g \in \mathbb{R}^{3000}$ (the same highly-variable-gene matrix used by MP-MNCA) and histology $\mathbf{x}_i^h \in \mathbb{R}^{128}$ (the frozen BYOL embedding of the main paper). Both are standardized per feature within a section; this is required because the two streams differ in scale by an order of magnitude, and an unstandardized shared selection would be decided by histology alone. Let $\mathcal{N}(i)$ denote the six spatial nearest neighbors of spot i .

8.2. Global TopK concept alignment

Each stream $s \in \{g, h\}$ has an affine encoder into a common L -dimensional dictionary, producing pre-activations

$$\mathbf{h}_i^s = W_E^s (\mathbf{x}_i^s - \mathbf{b}_{\text{pre}}^s) + \mathbf{b}_{\text{lat}}^s. \quad (25)$$

Alignment is enforced structurally. Rather than sparsifying each stream separately, the active set is chosen *once* from the summed pre-activations,

$$\mathcal{I}_i = \text{TopK}\left(\sum_s \mathbf{h}_i^s, k\right), \quad \mathbf{z}_i^s = \text{ReLU}(\text{gather}(\mathbf{h}_i^s, \mathcal{I}_i)), \quad (26)$$

so both streams activate identical dimensions for a given spot while retaining stream-specific magnitudes. Per-stream affine decoders reconstruct each input, and the objective combines self- and cross-reconstruction in normalized MSE,

$$\mathcal{L} = \sum_s \mathcal{L}_{\text{NMSE}}(\mathbf{x}^s, D_s(\mathbf{z}^s)) + \lambda \sum_{s \neq t} \mathcal{L}_{\text{NMSE}}(\mathbf{x}^t, D_t(\mathbf{z}^s)), \quad (27)$$

with an auxiliary term reviving dormant dictionary atoms. Cross-reconstruction is what makes the shared support *semantically* aligned rather than merely coincident.

8.3. Spatial k -NN smoothing before alignment

SPARC treats samples independently, which discards the strongest prior available in spatial transcriptomics. We inject it *before* encoding, replacing each spot’s features with a morphology-weighted blend of itself and its neighborhood:

$$\tilde{\mathbf{x}}_i^s = (1 - \alpha) \mathbf{x}_i^s + \alpha \sum_{j \in \mathcal{N}(i)} w_{ij} \mathbf{x}_j^s, \quad (28)$$

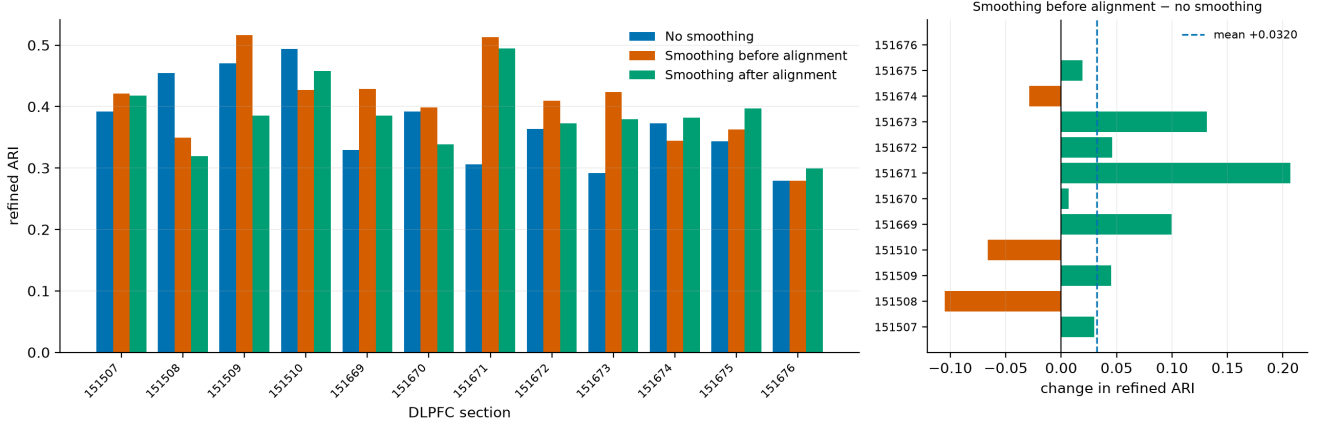


Figure 4. Effect of spatial smoothing across the twelve DLPFC sections. **Left:** refined ARI per section for no smoothing, smoothing applied before alignment, and smoothing applied to the latents after alignment. **Right:** per-section change from smoothing before alignment. The gain is concentrated in sections whose sparse codes were least stable (151671, +0.207; 151673, +0.132), and four sections regress.

$$w_{ij} = \text{softmax}_j(\beta \log s_{ij}), \quad s_{ij} = \frac{1}{2}(\cos(\mathbf{m}_i, \mathbf{m}_j) + 1), \quad (29)$$

where s_{ij} is the histological similarity prior of the main paper. Weighting by s_{ij} rather than uniformly means morphologically dissimilar neighbors contribute less, so laminar boundaries blur less than under a plain neighborhood mean. We use $\alpha = 0.5$, giving the spot and its neighborhood equal weight; this value was set by analogy to a normalized adjacency with self-loops and was *not* tuned.

The design rationale is asymmetry between the streams. Histology reconstructs well (self-NMSE ≈ 0.28) whereas expression does not (≈ 0.88), so the shared support is dominated by histology and the concept space carries little laminar information. Smoothing raises the signal-to-noise of the weaker stream so that it competes in the shared selection. Consistent with this account, smoothing the *latents* after alignment instead—which cannot change what the selection sees—helps substantially less (Sec. 9).

8.4. Partitioned dictionary

A shared support is well motivated when the streams are symmetric views of the same content, as in the vision–language setting SPARC was designed for. Spatial transcriptomics is not symmetric: expression predicts morphology well (cross-reconstruction NMSE 0.04–0.21) whereas morphology predicts expression poorly (0.89–0.96). Forcing every dimension to be shared therefore obliges molecular programs with no morphological counterpart to compete for the same atoms as programs that do have one.

We relax the constraint by partitioning the dictionary into a shared block and one private block per stream. Global TopK selects k_{shared} atoms from the shared block using the summed pre-activations, while each stream additionally selects k_{priv} atoms from its own private block:

$$\mathcal{I}_i^s = \text{TopK}\left(\sum_t \mathbf{h}_i^t|_{\mathcal{S}}, k_{\text{shared}}\right) \cup \text{TopK}\left(\mathbf{h}_i^s|_{\mathcal{P}_s}, k_{\text{priv}}\right). \quad (30)$$

The private budget is carved out of k rather than added to it, so each stream remains exactly k -sparse and the comparison is capacity-matched. We use an equal split of both the dictionary and the sparsity budget. Concepts that are genuinely bimodal remain shared and interpretable across streams; modality-specific structure is no longer forced into that shared vocabulary.

8.5. Phase 2 and evaluation

The concept codes are passed to MP-MNCA’s masked neighbor cross-attention with the neighbor-contrastive objective, trained for 20 epochs with pseudo-labels from k -means on histology—never from anatomical annotations. Evaluation follows the main paper exactly: PCA to 128 dimensions, a tied-covariance Gaussian mixture with the component count set by the observed number of layers, 15-nearest-neighbor spatial refinement, and permutation-invariant ARI/NMI. Phase 1 uses $L = 1024$ dictionary atoms with $k = 128$ active, 50 epochs, batch size 256, Adam at 10^{-4} .

9. Results

9.1. Spatial smoothing before alignment

Table 2 reports per-section ARI and NMI for both support mechanisms. Fig. 4 compares the three placements of the smoothing operator, holding the shared support fixed. Smoothing *before* alignment reaches 0.4061 ± 0.0676 mean refined ARI, against 0.3740 without smoothing and 0.3856 when smoothing is applied to the latents afterwards—supporting the account in Sec. 8 that the intervention acts by changing what the shared selection sees.

We report this as a promising rather than an established effect. The $+0.0320$ improvement wins 8 of 12 sections but does not reach significance when paired across sections (Wilcoxon $p = 0.233$; paired t -test $p = 0.218$), and four sections regress, the largest by -0.105 . Donor structure is visible in the variance: three of the four weakest sections belong to donor 3.

We additionally evaluated per-stream quotas and reciprocal rank fusion for the shared selection, nonlinear encoders, per-stream logit normalization, dictionary capacity, batch size, masked denoising, and a single dictionary pooled across all sections. All were neutral or harmful. Only spatial smoothing and the partitioned dictionary (Sec. 8.4) improved clustering.

9.2. Effect of relaxing the shared support

Table 2 compares the two support mechanisms. The partitioned dictionary attains 0.4228 ± 0.0533 mean refined ARI and 0.5285 ± 0.0496 NMI, against 0.4061 ± 0.0676 and 0.5091 ± 0.0501 under a fully shared support. The mean difference of $+0.0167$ is not significant across the twelve sections (Wilcoxon $p = 0.519$), and the variant wins six of twelve.

Two aspects are more informative than the mean. First, the between-section standard deviation falls to 0.0533, the lowest of any configuration we evaluated, and the gains are concentrated in the sections that previously performed worst—151676 moves from 0.2789 to 0.4074 and 151674 from 0.3441 to 0.4565. This is the behaviour the asymmetry argument predicts: sections whose molecular structure is least reflected in morphology benefit most from private capacity. Second, NMI improves on ten of twelve sections, a more consistent pattern than ARI alone suggests.

Fig. 6 shows the corresponding convergence on section 151507, which is stable across epoch budgets (0.404 at 10 epochs to 0.416 at 200).

9.3. Convergence and stability

Fig. 7 and Fig. 8 show refined ARI as a function of Phase-1 epochs. Two observations stand out. First, the mean varies by only 0.031 across a twenty-fold range of training, and every epoch budget tested is optimal for at least one section, so no transferable epoch setting exists. Second, the two-phase pipeline is markedly more stable than Phase 1 alone: across all (section, epoch) checkpoints it attains higher mean ARI (0.364 vs. 0.300) at half the within-section volatility (0.031 vs. 0.063). Phase 2 therefore contributes stabilization as much as accuracy, smoothing a sparse code that is otherwise erratic across checkpoints.

Fig. 9 shows the batch-size sweep. Across a sixteen-fold range the total spread is 0.040 ARI, and the nominal optimum is not significant (Wilcoxon $p = 0.30$). Which section is being processed matters roughly twice as much as which batch size is chosen.

9.4. Qualitative behavior

Fig. 5 shows predicted domains on section 151507 across training, with and without smoothing. Smoothing changes the character of the failure. Without it, predictions are spatially incoherent blobs; with it, the model recovers recognizable laminar bands that follow the ground-truth orientation and sharpen with training. The residual errors are systematic rather than random: Layer 4—the thinnest cortical layer—is consistently rendered too thick, and white matter remains entangled with Layer 6. Smoothing therefore corrects the *geometry* of the prediction without resolving layer *identity* at boundaries.

10. Discussion, limitations and future work

A modality gap bounds the approach. SPARC-align improves substantially on the published SPARC configuration ($0.2662 \rightarrow 0.4228$ mean refined ARI) but does not reach MP-MNCA. The limitation appears structural rather than one of tuning. Cross-reconstruction from histology to expression never falls below ≈ 0.89 NMSE in any configuration we tested, meaning morphology cannot predict expression in this tissue; the shared concept space consequently remains largely a histological code, while laminar identity resides in the molecular channel. A concept-agreement diagnostic makes the same point: neighboring spots share four to five times more active concepts than random pairs (Jaccard 0.29 vs. 0.06), yet same-layer and cross-layer neighbors differ by only ≈ 0.02 (AUC ≈ 0.56). The learned concepts capture spatial structure without capturing layer boundaries—precisely the behavior visible in Fig. 5.

Section	Donor	Shared support		Partitioned	
		ARI	NMI	ARI	NMI
151507	1	0.4214	0.5216	0.4125	0.5344
151508	1	0.3491	0.4635	0.4451	0.5450
151509	1	0.5159	0.5710	0.5082	0.5816
151510	1	0.4272	0.5267	0.4558	0.5095
151669	2	0.4287	0.5202	0.3678	0.4968
151670	2	0.3990	0.4546	0.3124	0.3965
151671	2	0.5131	0.5942	0.4772	0.5869
151672	2	0.4093	0.5112	0.4397	0.5588
151673	3	0.4235	0.5549	0.4162	0.5469
151674	3	0.3441	0.4616	0.4565	0.5371
151675	3	0.3626	0.5010	0.3745	0.5107
151676	3	0.2789	0.4289	0.4074	0.5383
Mean $\pm$ std		0.4061 ± 0.0676	0.5091 ± 0.0501	0.4228 ± 0.0533	0.5285 ± 0.0496
Median		0.4154	0.5157	0.4279	0.5377

Table 2. SPARC-align with spatial smoothing on the twelve DLPFC sections, under a shared concept support (Sec. 8) and under the partitioned dictionary (Sec. 8.4). Standard deviation uses ddof = 1, treating sections as the biological replicates. Bold marks the larger of the two where the difference exceeds 0.05.

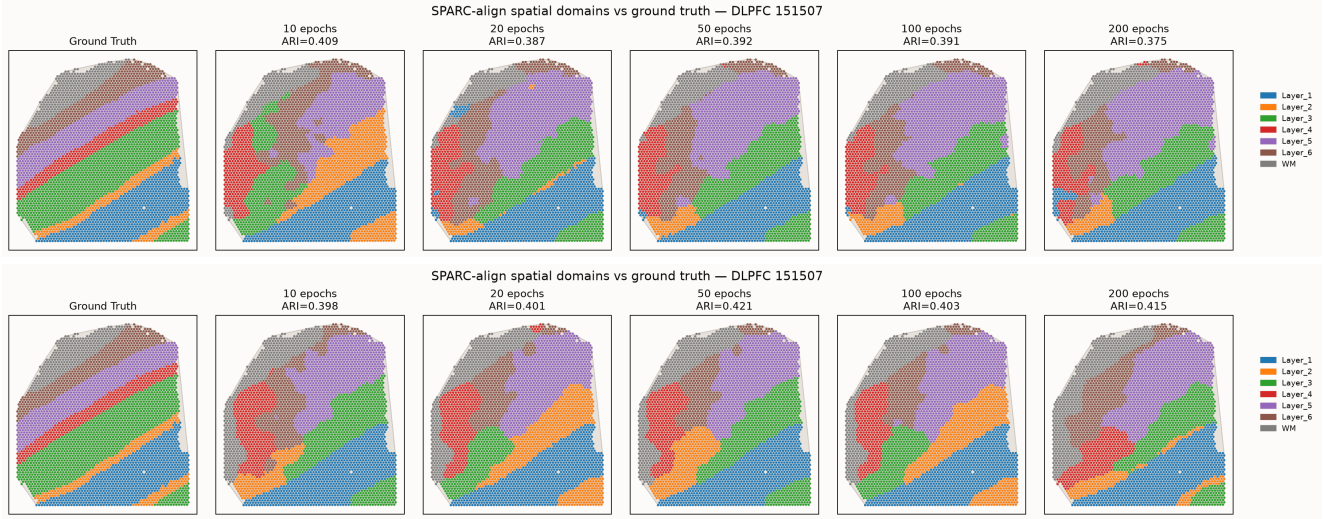


Figure 5. Predicted spatial domains on DLPFC section 151507 at 10, 20, 50, 100 and 200 Phase-1 epochs, without smoothing (**top**) and with smoothing before alignment (**bottom**); ground truth at left. Smoothing converts spatially incoherent predictions into laminar bands, but Layer 4 remains too thick and white matter stays entangled with Layer 6 at every budget.

Reconstruction is not the binding constraint. It is natural to attribute the gap to poor reconstruction of the expression stream. We find this is not the case. Across the epoch range, gene self-NMSE improves by 15.4% while ARI improves by 8.9%; pooling a single dictionary over all twelve sections raises data per atom elevenfold yet moves self-NMSE only from 0.880 to 0.872; and enlarging the encoder degrades reconstruction under a fixed budget, indicating undertraining rather than insufficient capacity. Improving reconstruction fidelity is therefore an inefficient route to better spatial domains.

Relaxing the shared support helps, and points forward. Of every modification we evaluated, only two improved clustering: spatial smoothing and the partitioned dictionary. Both act on the same underlying issue—that a histology-dominated shared code carries limited laminar information—and the partitioned variant additionally reduces between-section variance

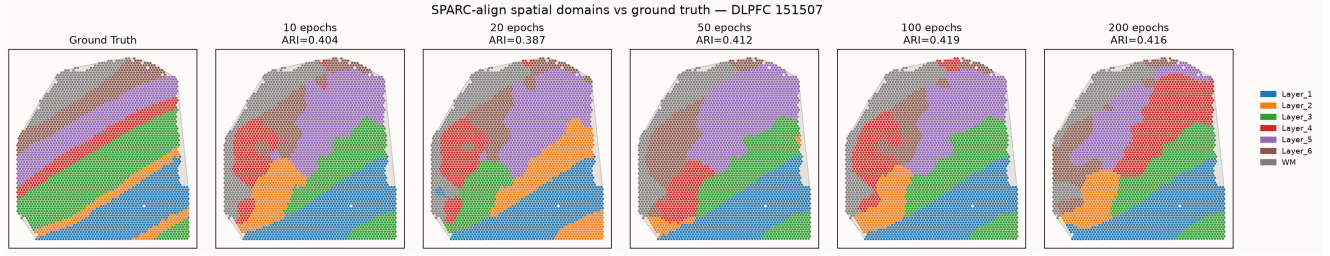


Figure 6. Predicted spatial domains on section 151507 under the partitioned dictionary, at 10, 20, 50, 100 and 200 Phase-1 epochs, with ground truth at left. Laminar organization is established by the earliest checkpoint and remains stable thereafter.

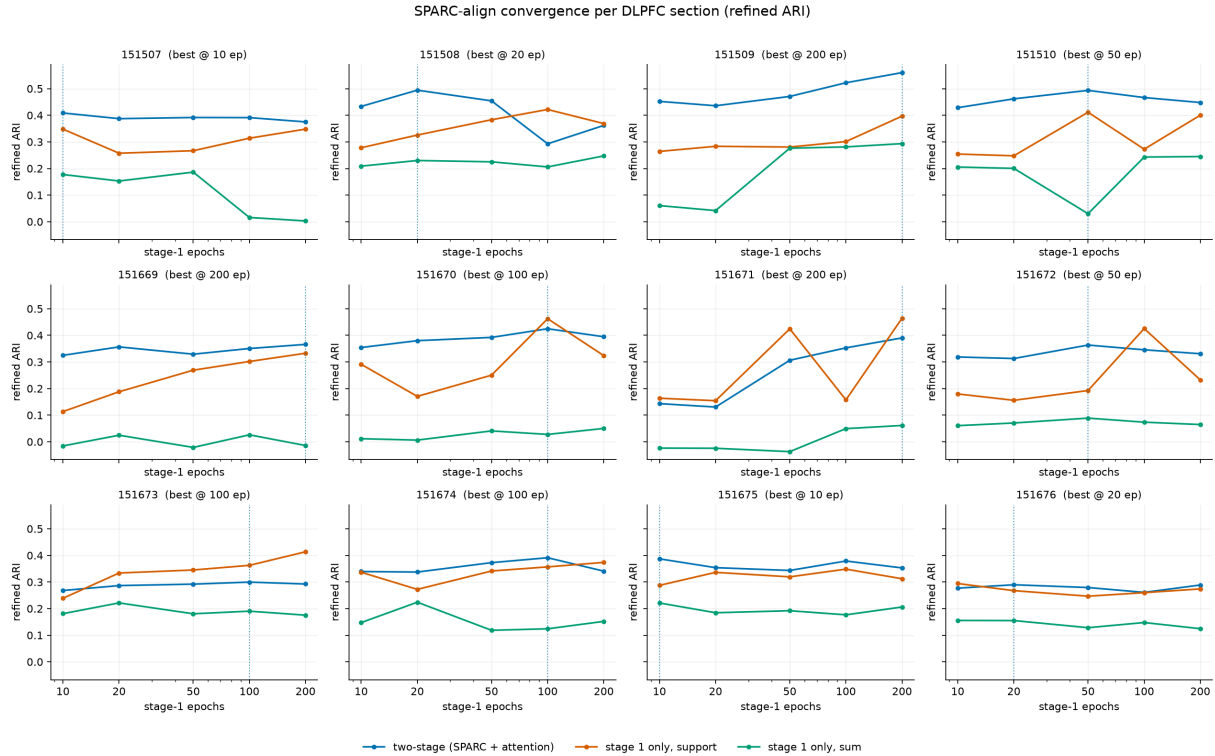


Figure 7. Refined ARI against Phase-1 epochs for each section, with Phase 2 fixed at 20 epochs. Blue is the full two-phase pipeline; orange and green are Phase 1 alone under two cluster-input choices. The dotted line marks each section's best epoch. All checkpoints come from a single 200-epoch run per section, so the curves are trajectories rather than independent runs.

while raising NMI on ten of twelve sections. Neither reaches significance individually at $n = 12$, but they are consistent in direction and mechanism, which suggests that the productive axis is how much representational independence the two modalities are granted, rather than how the shared selection is weighted. A control that weighted smoothing by expression similarity instead of histology was significantly *worse* (Wilcoxon $p = 0.021$), indicating that the benefit of the morphology prior derives from its independence from the stream being smoothed.

Limitations. (i) Neither principal effect is statistically established at $n = 12$ ($p = 0.233$ for smoothing, $p = 0.519$ for partitioning). With twelve sections and differences of this magnitude, the achievable power is low; the pipeline is bit-reproducible at a fixed seed, so additional seeds would reduce per-section measurement noise but would not increase the number of biological replicates. Confirmation requires further sections or a second tissue. (ii) $\alpha = 0.5$ and $\beta = 1$ were set by analogy, not tuned. (iii) Smoothing alters the reconstruction target, so NMSE is not comparable between smoothed and unsmoothed variants; only clustering metrics are. (iv) Results are per-section transductive on a single tissue type;

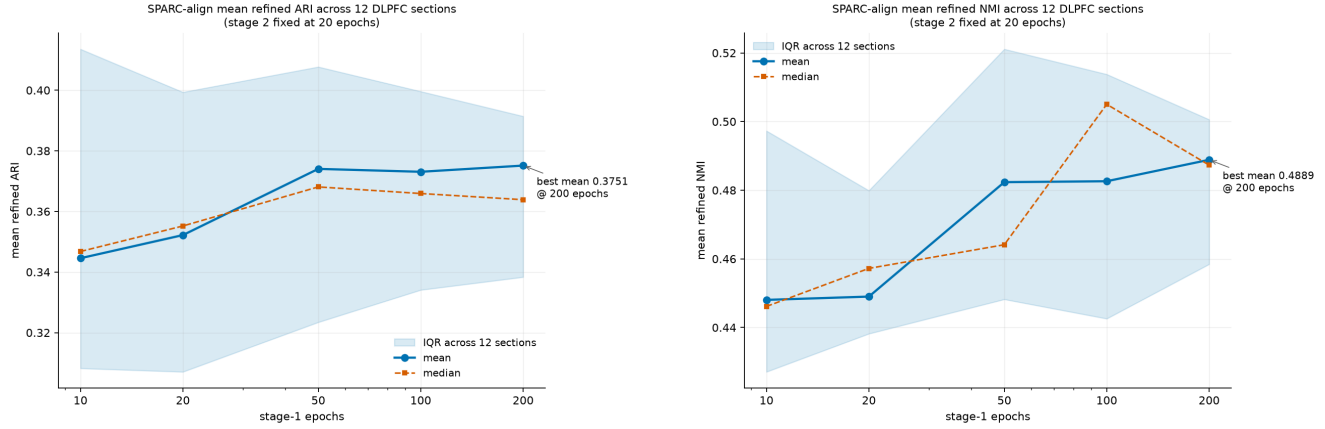


Figure 8. Mean refined ARI (left) and NMI (right) across the twelve sections against Phase-1 epochs, with an interquartile band. The mean changes by 0.031 ARI over a twenty-fold range of training.

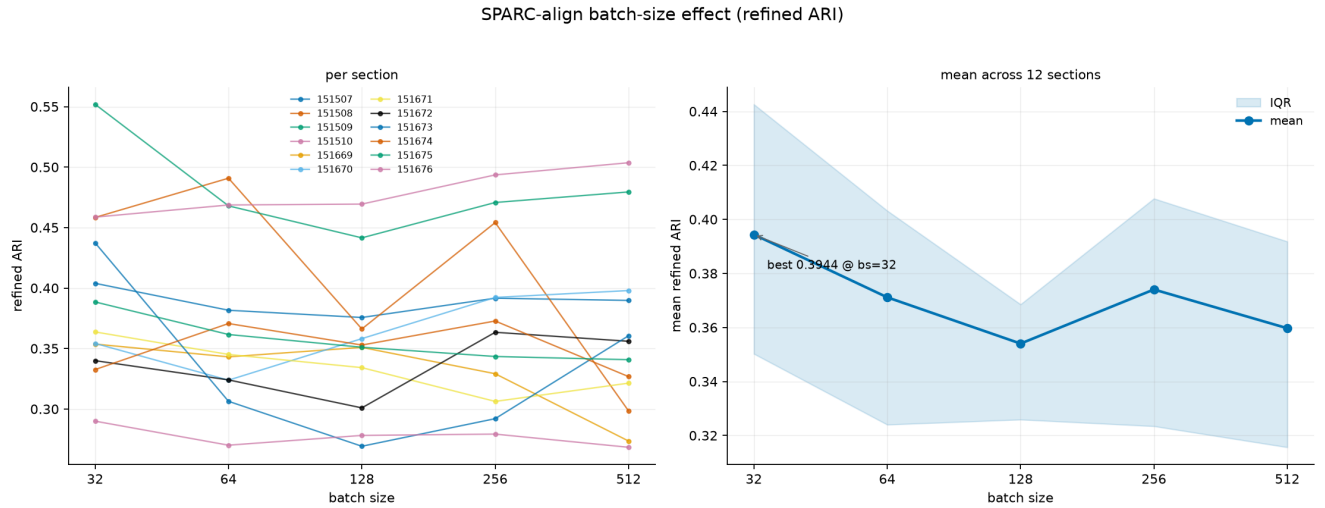


Figure 9. Batch-size sweep at a fixed epoch budget. Left: per section. Right: mean with interquartile band. The spread over a sixteen-fold range is 0.040 ARI and the nominal optimum is not significant.

generalization to held-out donors or other tissues is untested. (v) The component count is set from the observed number of annotated layers, so model selection is label-count-informed, as in the main paper.

Future work. The most informative next step is evaluation on additional tissue, since the effects we observe are not resolvable at twelve sections regardless of how many seeds are averaged (Limitations, (i)). Beyond that, the diagnostics point away from the sparse-reconstruction formulation itself. A reconstruction objective allocates capacity to variance, and laminar identity is evidently not where the variance lies; an objective that is discriminative in the concept space appears more promising than further refinement of the shared-support mechanism. The partitioned dictionary suggests a related direction: its shared and private blocks were split evenly without tuning, and the balance between them is a natural axis to explore. Validating dictionary atoms against published laminar marker genes would also test concept quality directly, without relying on the annotations used for evaluation.

11. Section-wise Convergence Analysis

To examine optimization behavior beyond a single final checkpoint, we track the refined ARI of all methods throughout the common 200-epoch training budget on each of the 12 DLPFC sections. Figure 10 shows the complete section-wise

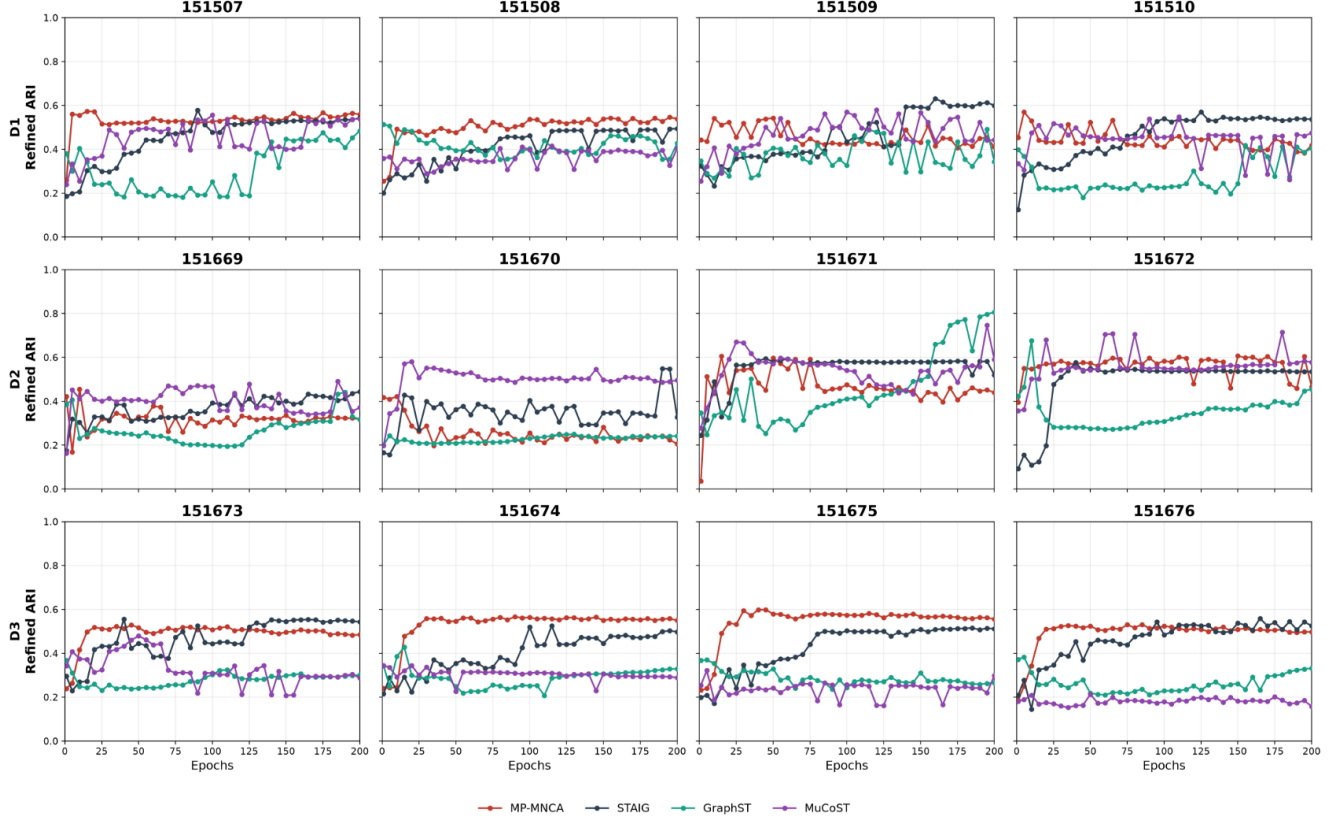


Figure 10. **Section-wise convergence of refined ARI on the 12 DLPFC sections.** Each panel reports refined ARI over the common 200-epoch optimization budget for MP-MNCA, STAIG, GraphST, and MuCoST. MP-MNCA typically reaches a competitive regime early in training, whereas the competing methods exhibit more section-dependent convergence behavior.

trajectories. The results reveal substantial variation in optimization dynamics across tissue sections. MP-MNCA generally enters a competitive high-ARI regime within the early stages of training and then remains comparatively stable. In contrast, STAIG often improves more gradually and reaches competitive performance later in training, while GraphST and MuCoST exhibit larger section-dependent fluctuations. These trajectories therefore distinguish convergence behavior that would be obscured by reporting only a final or best checkpoint.

To summarize these section-level trajectories, we additionally aggregate performance across all 12 sections at every epoch. Figure 11 reports the mean refined ARI together with the corresponding standard deviation across sections. The aggregate view confirms that MP-MNCA reaches a competitive performance regime within the first few tens of epochs, whereas STAIG improves more progressively and becomes comparable later in training. The shaded regions further capture cross-section variability, complementing the individual trajectories in Figure 10. Importantly, this comparison measures *epoch-wise optimization dynamics* under an identical training budget rather than wall-clock computational efficiency.

12. Epoch-wise Spatial-Domain Evolution Across DLPFC Sections

To complement the quantitative convergence analysis, we visualize the evolution of predicted spatial domains throughout training on all 12 DLPFC sections. For each section, rows correspond to checkpoints at epochs 10, 20, 50, 100, and 200, while columns show the ground-truth annotation and predictions from STAIG, GraphST, MuCoST, and MP-MNCA. These visualizations expose differences in optimization dynamics that are not captured by a single final-checkpoint metric, including the emergence of anatomically coherent laminar structure, boundary stabilization, fragmentation, and local spatial inconsistencies.

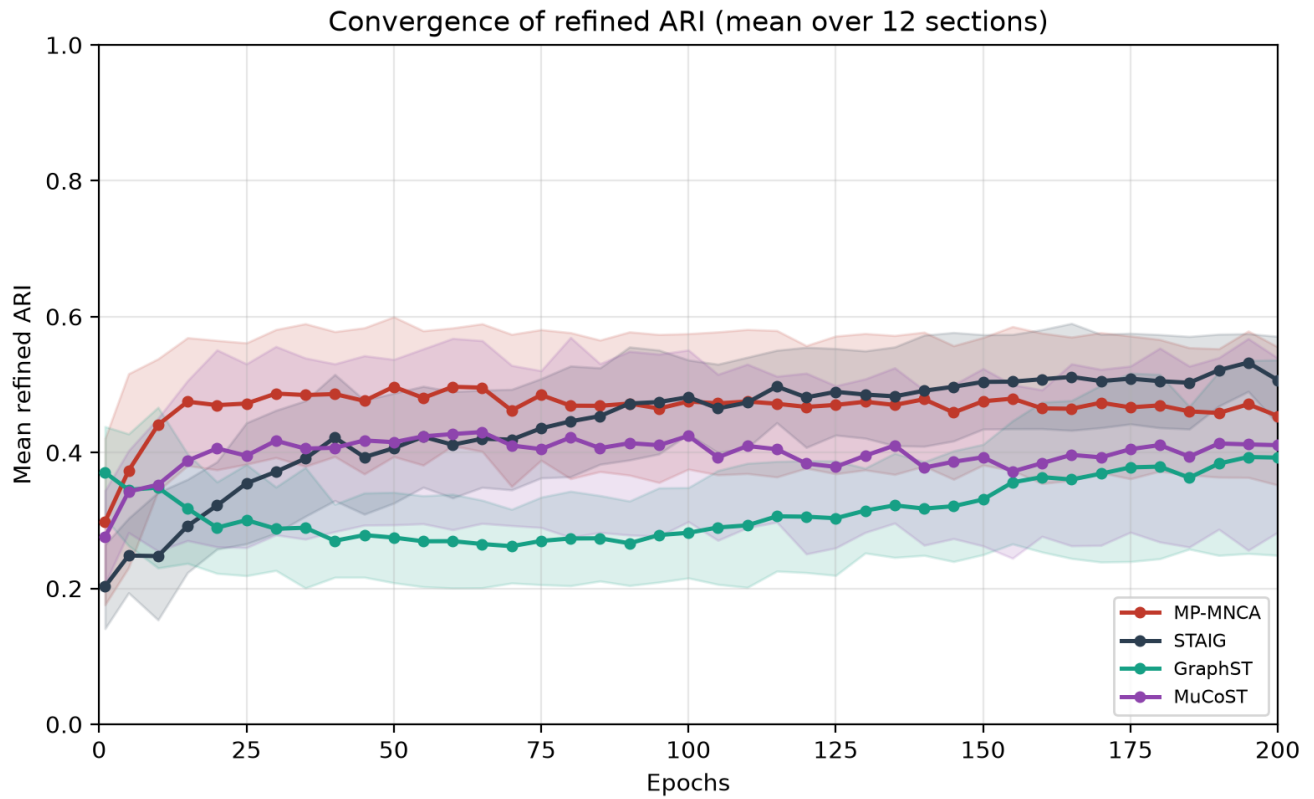


Figure 11. **Mean convergence of refined ARI across the 12 DLPFC sections.** Curves denote the mean refined ARI at each epoch and shaded regions denote $\pm$ one standard deviation across sections. MP-MNCA reaches a competitive high-ARI regime early in optimization, while STAIG improves more gradually and becomes comparable later in training.

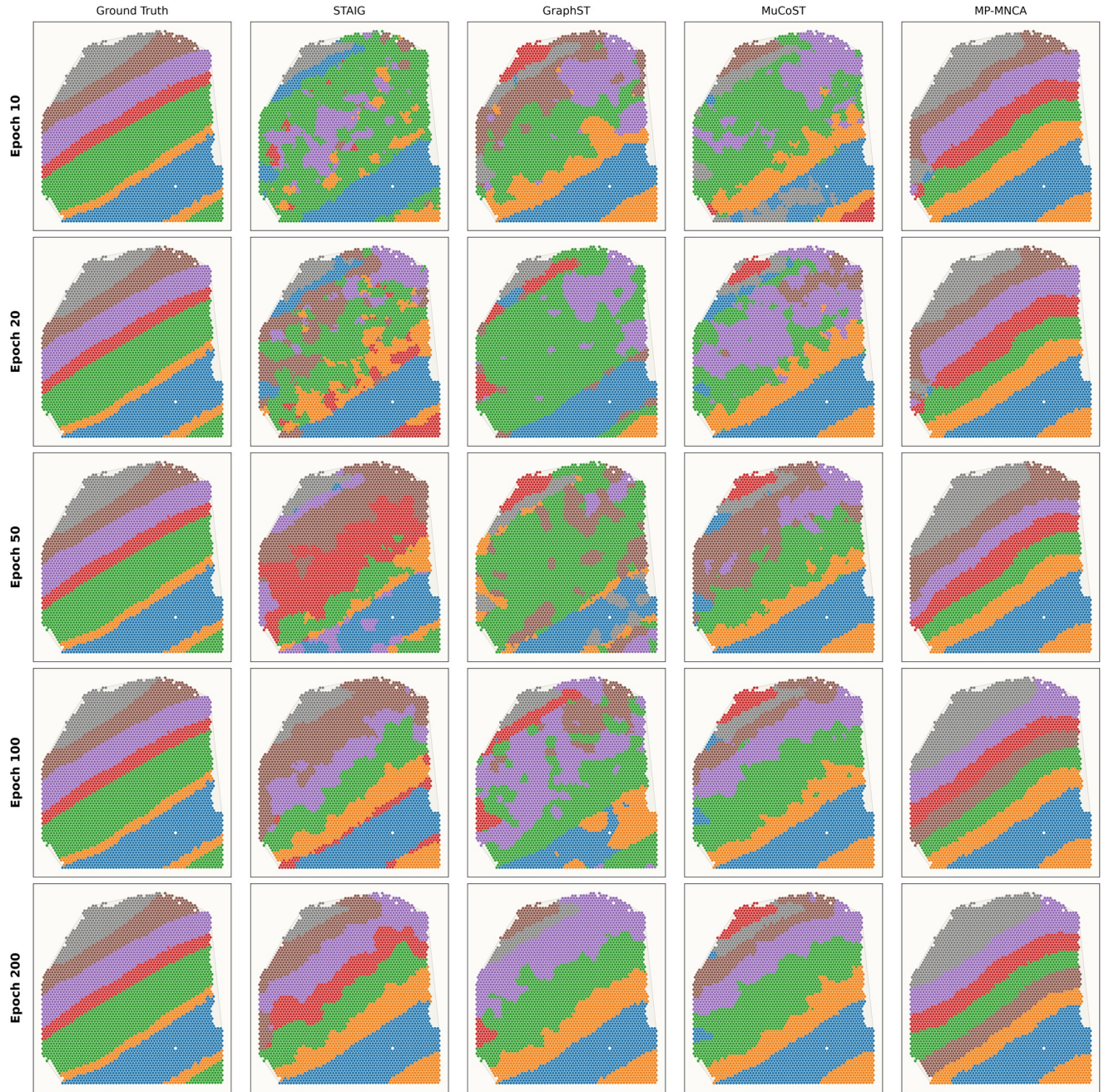


Figure 12. **Epoch-wise spatial-domain evolution on DLPFC section 151507.** Predictions are shown at epochs 10, 20, 50, 100, and 200 for STAIG, GraphST, MuCoST, and MP-MNCA, together with the ground-truth cortical layers. The sequence highlights how the spatial organization and cortical boundaries produced by each method evolve throughout optimization.

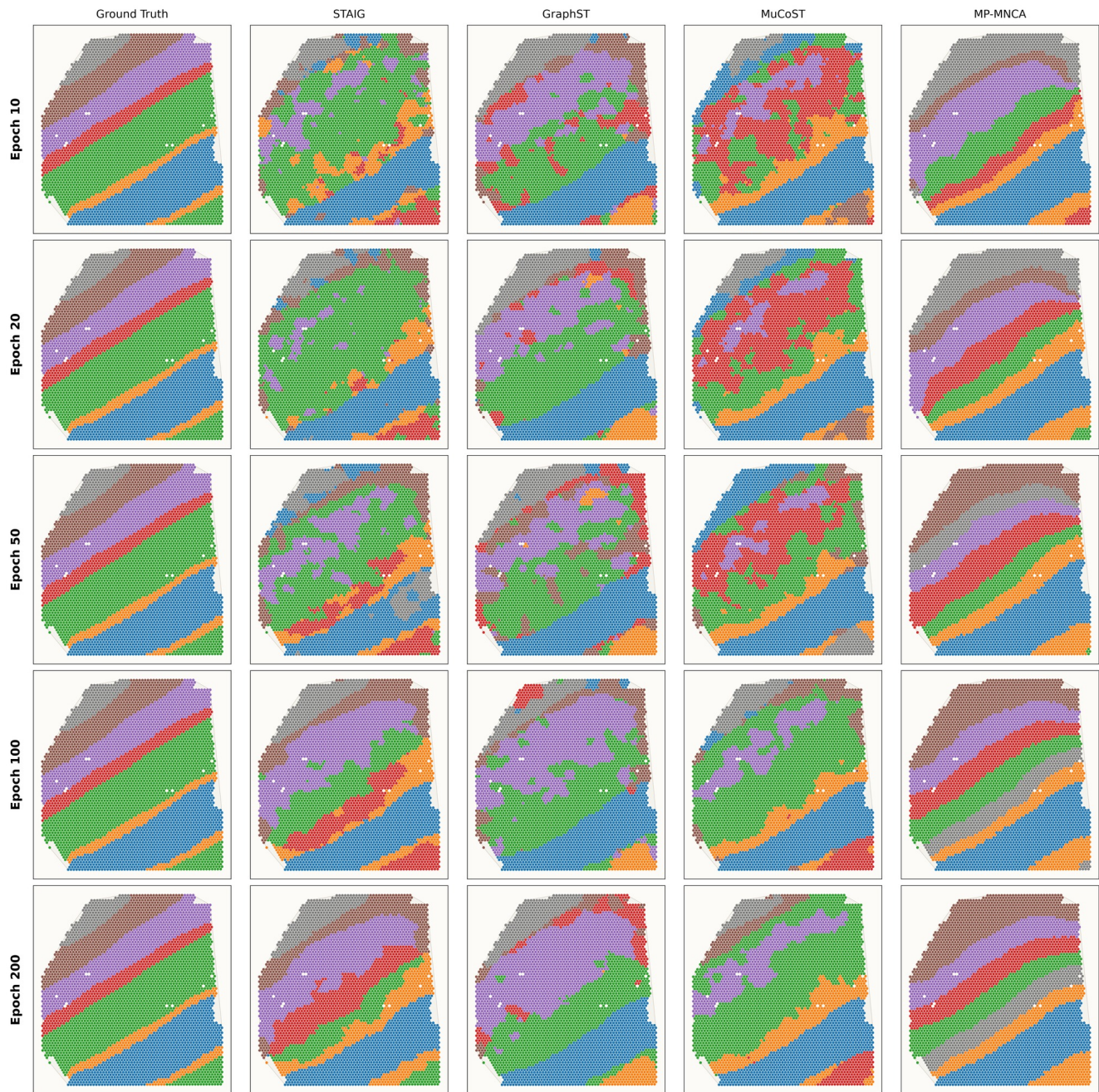


Figure 13. **Epoch-wise spatial-domain evolution on DLPFC section 151508.** Spatial-domain predictions at epochs 10, 20, 50, 100, and 200 illustrate the progression from early optimization to the later training regime for STAIG, GraphST, MuCoST, and MP-MNCA.

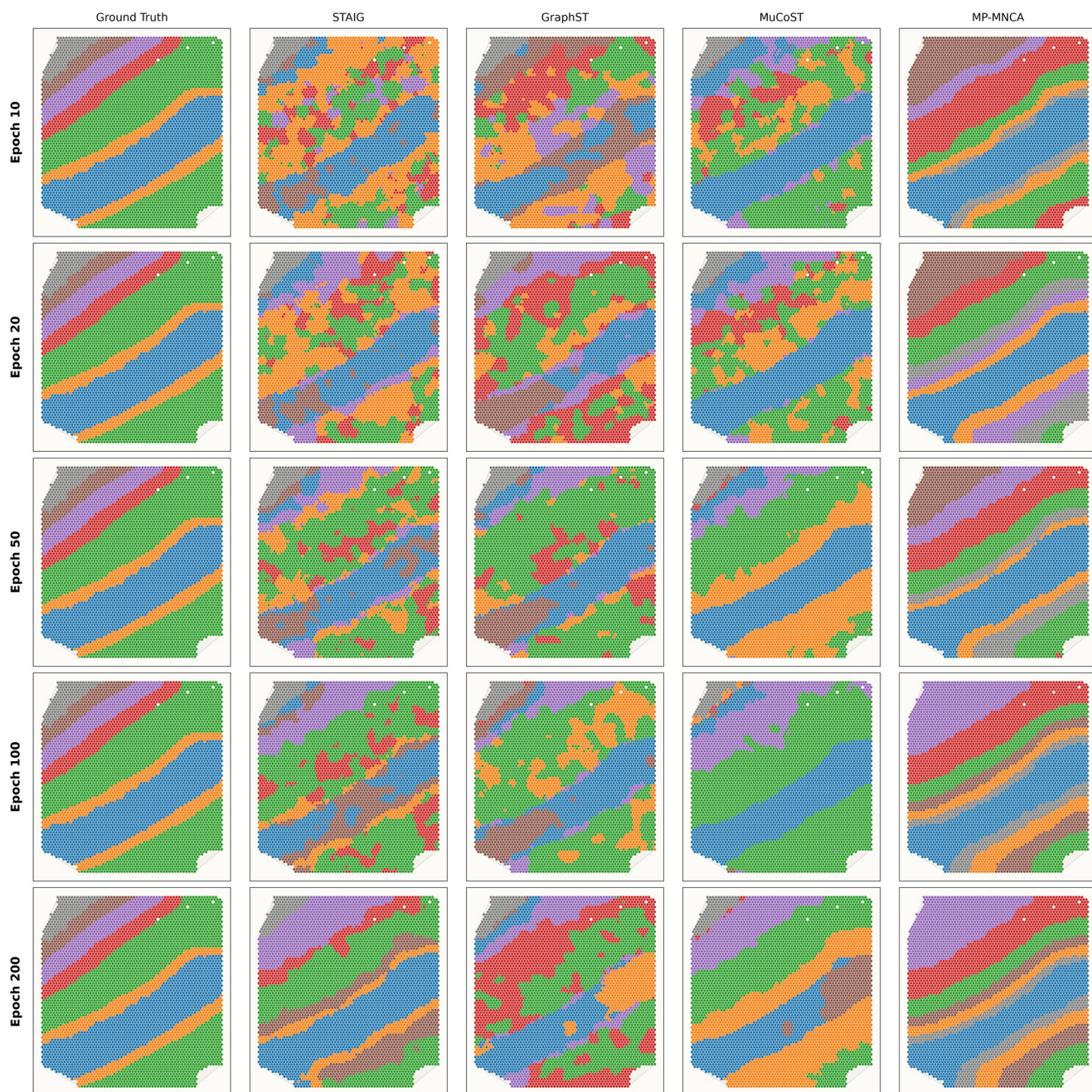


Figure 14. **Epoch-wise spatial-domain evolution on DLPFC section 151509.** This comparatively challenging section illustrates that convergence behavior is not uniform across tissue sections. The maps show the evolution of layer continuity, boundary placement, and local fragmentation for each method over the common 200-epoch budget.

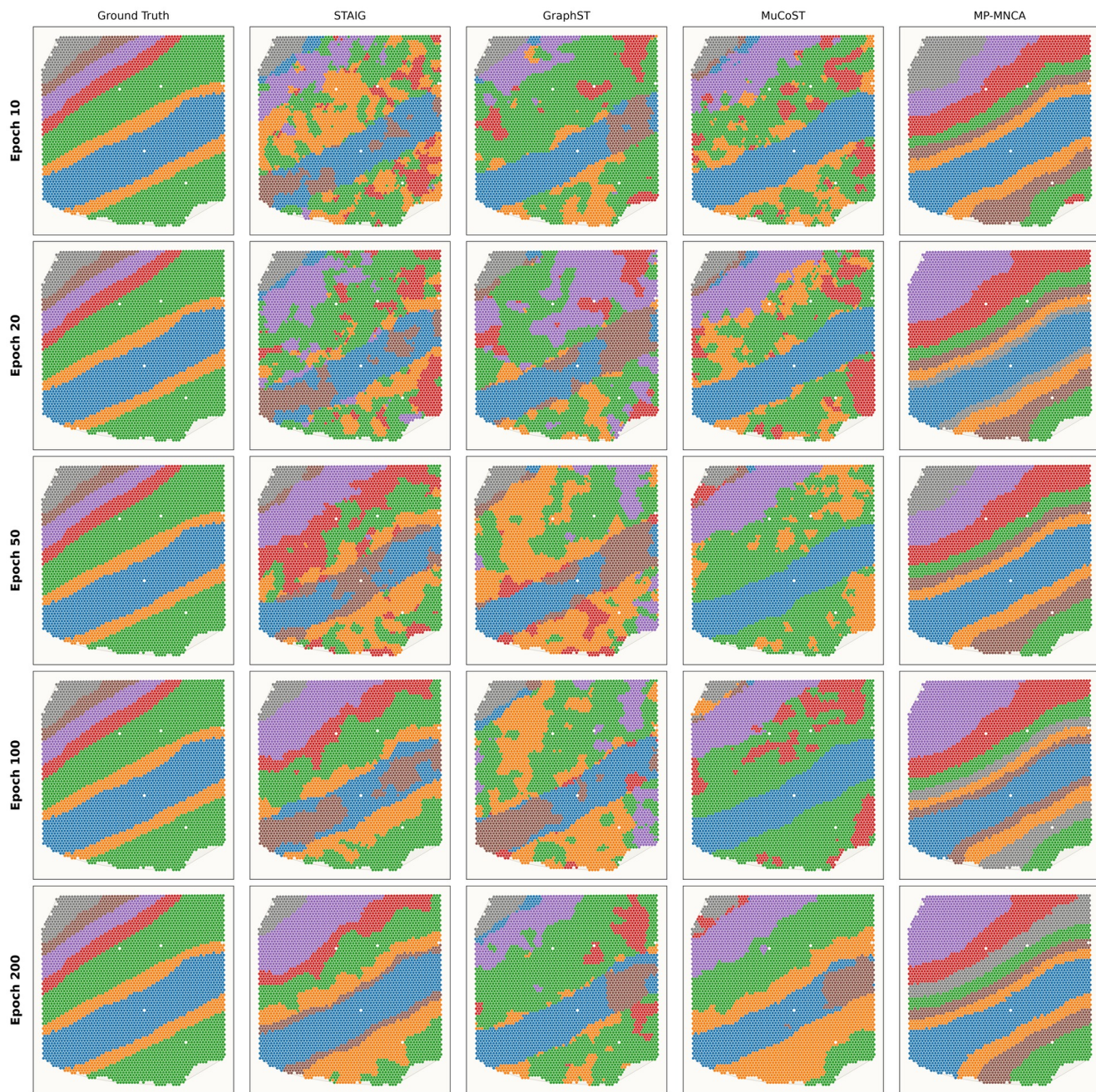


Figure 15. **Epoch-wise spatial-domain evolution on DLPFC section 151510.** Predictions across the five training checkpoints provide a qualitative view of how cortical-layer organization develops and stabilizes under STAIG, GraphST, MuCoST, and MP-MNCA.

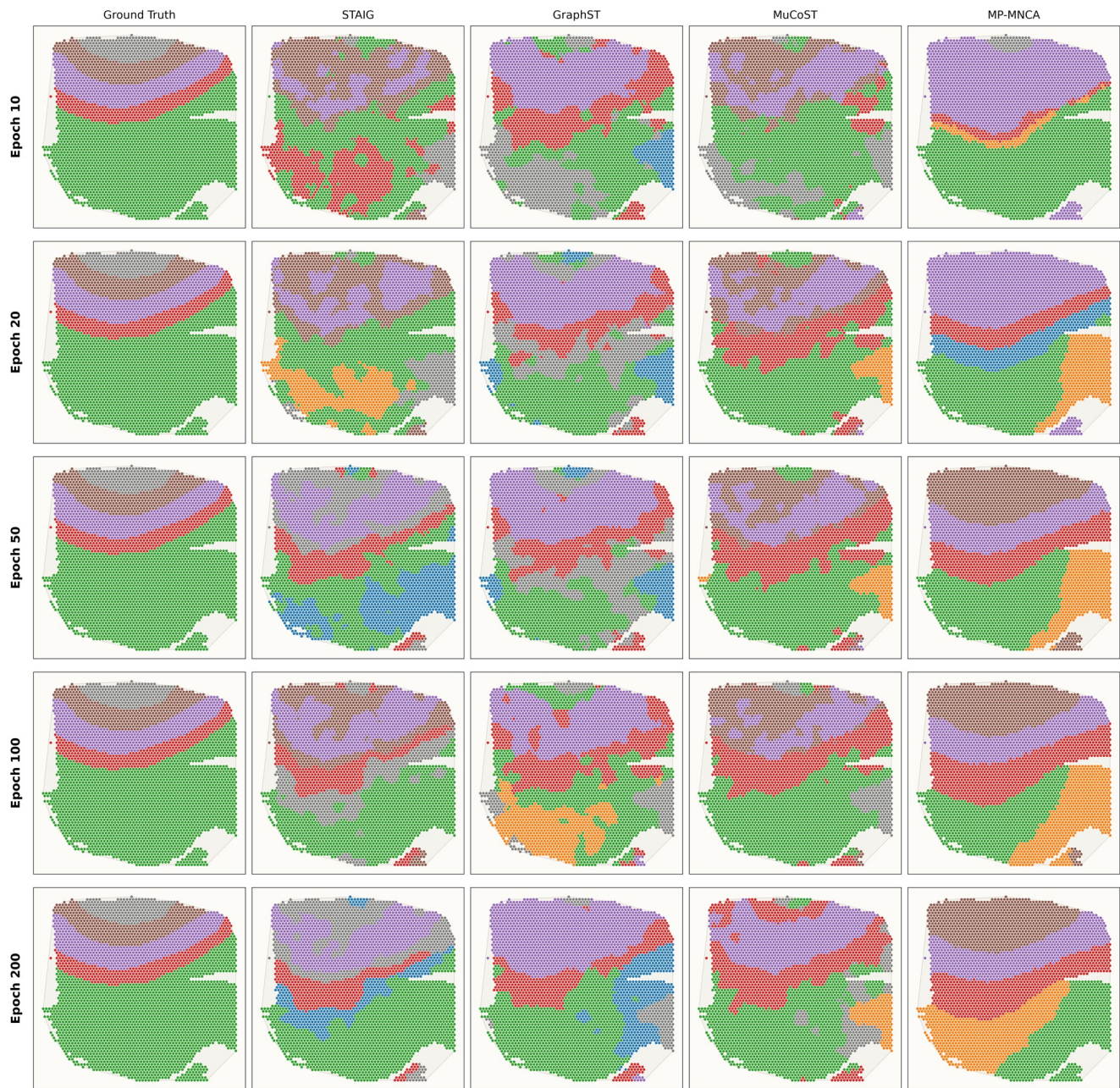


Figure 16. **Epoch-wise spatial-domain evolution on DLPFC section 151669.** The checkpoint sequence compares the formation and stabilization of spatial domains across competing methods, with the ground-truth laminar organization shown for reference.

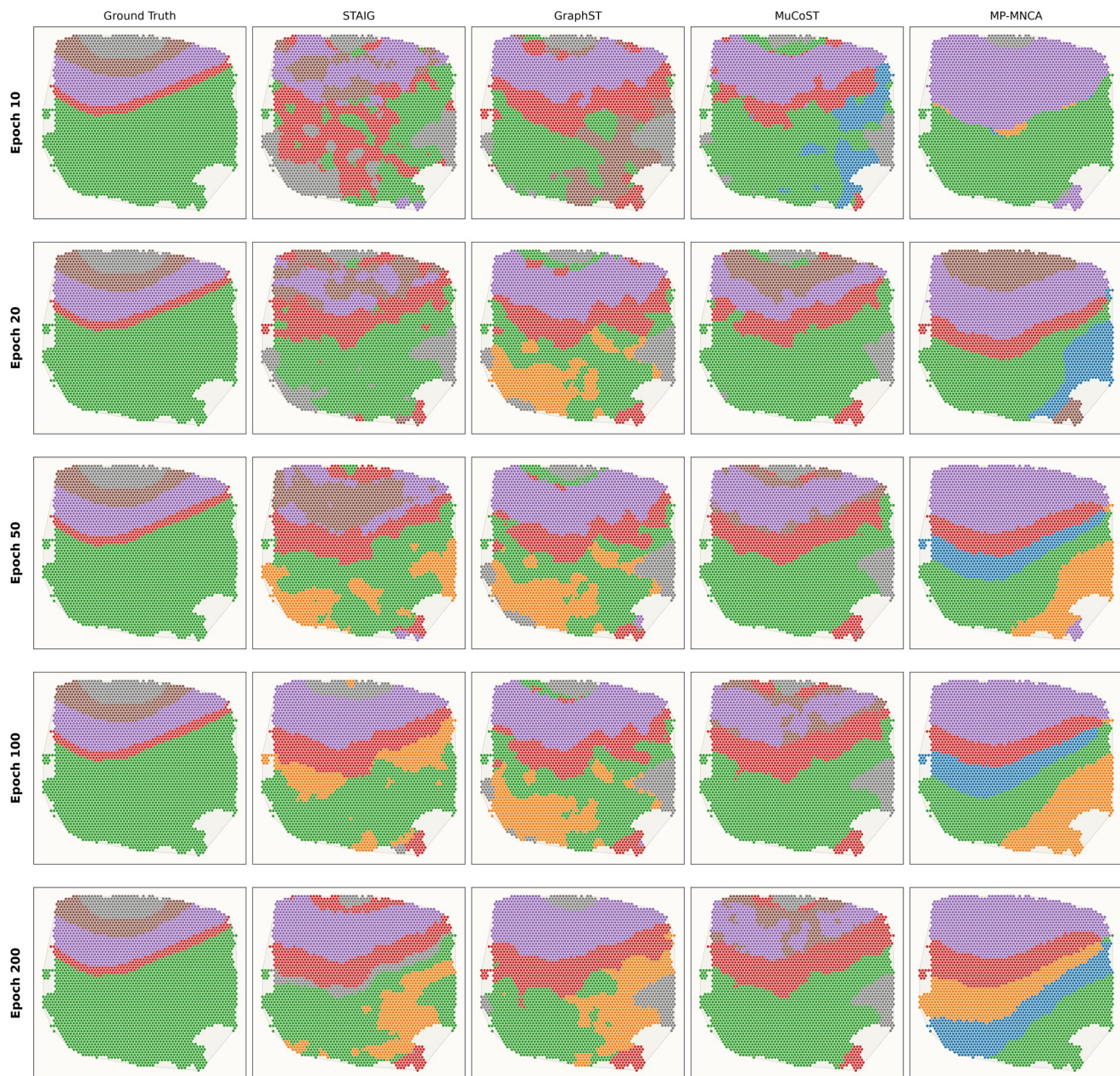


Figure 17. **Epoch-wise spatial-domain evolution on DLPFC section 151670.** Spatial-domain maps at epochs 10–200 reveal the temporal evolution of cortical boundaries and within-layer consistency for STAIG, GraphST, MuCoST, and MP-MNCA.

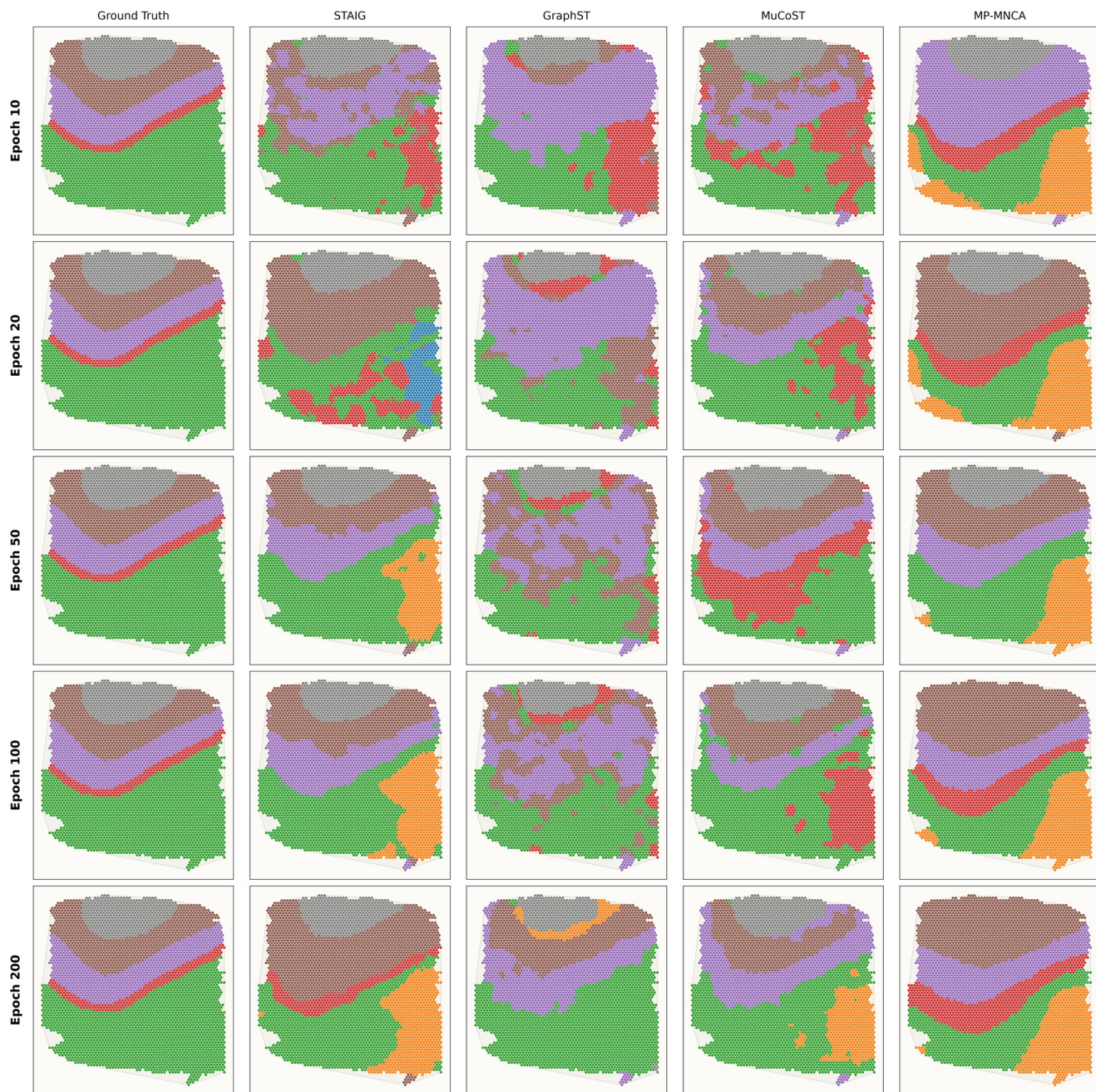


Figure 18. **Epoch-wise spatial-domain evolution on DLPFC section 151671.** The visualization tracks changes in predicted cortical organization throughout the shared training budget and complements the corresponding quantitative convergence measurements.

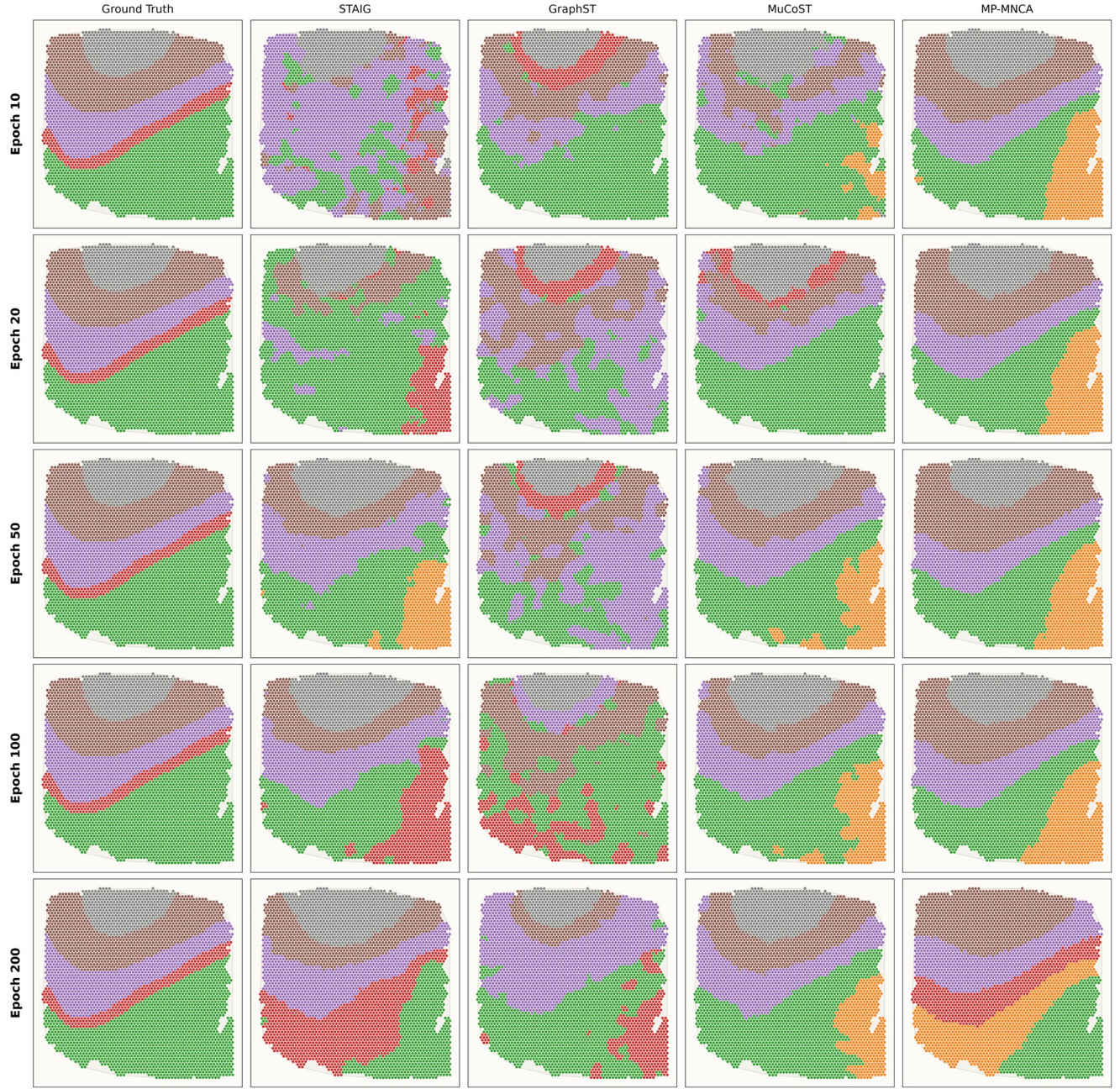


Figure 19. **Epoch-wise spatial-domain evolution on DLPFC section 151672.** Predictions from successive checkpoints illustrate differences between methods in the development of coherent laminar structure and the stabilization of spatial-domain boundaries.

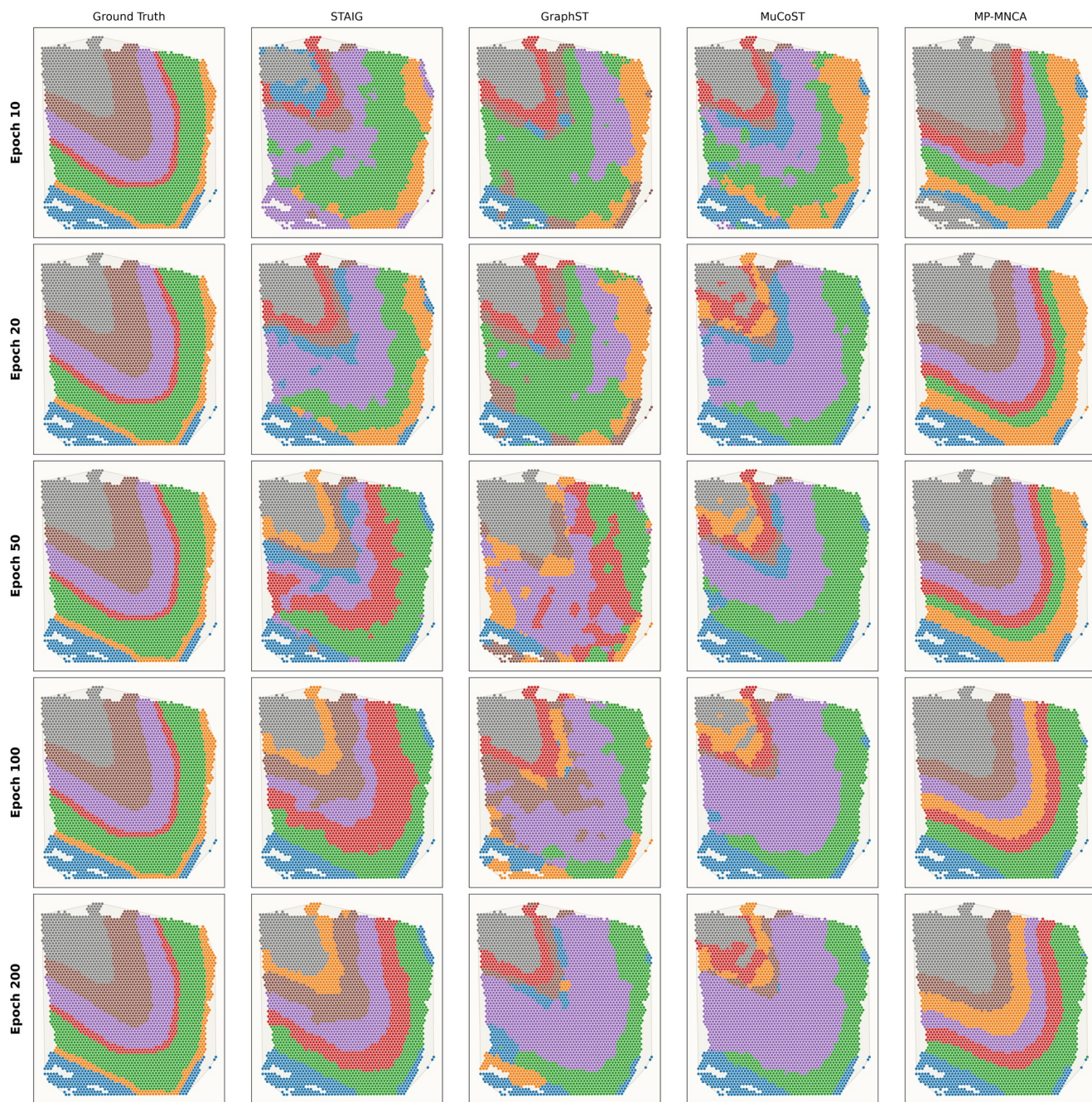


Figure 20. **Epoch-wise spatial-domain evolution on DLPFC section 151673.** The five checkpoints show how spatial-domain assignments evolve from early training through epoch 200, exposing differences in boundary consistency and regional fragmentation across methods.

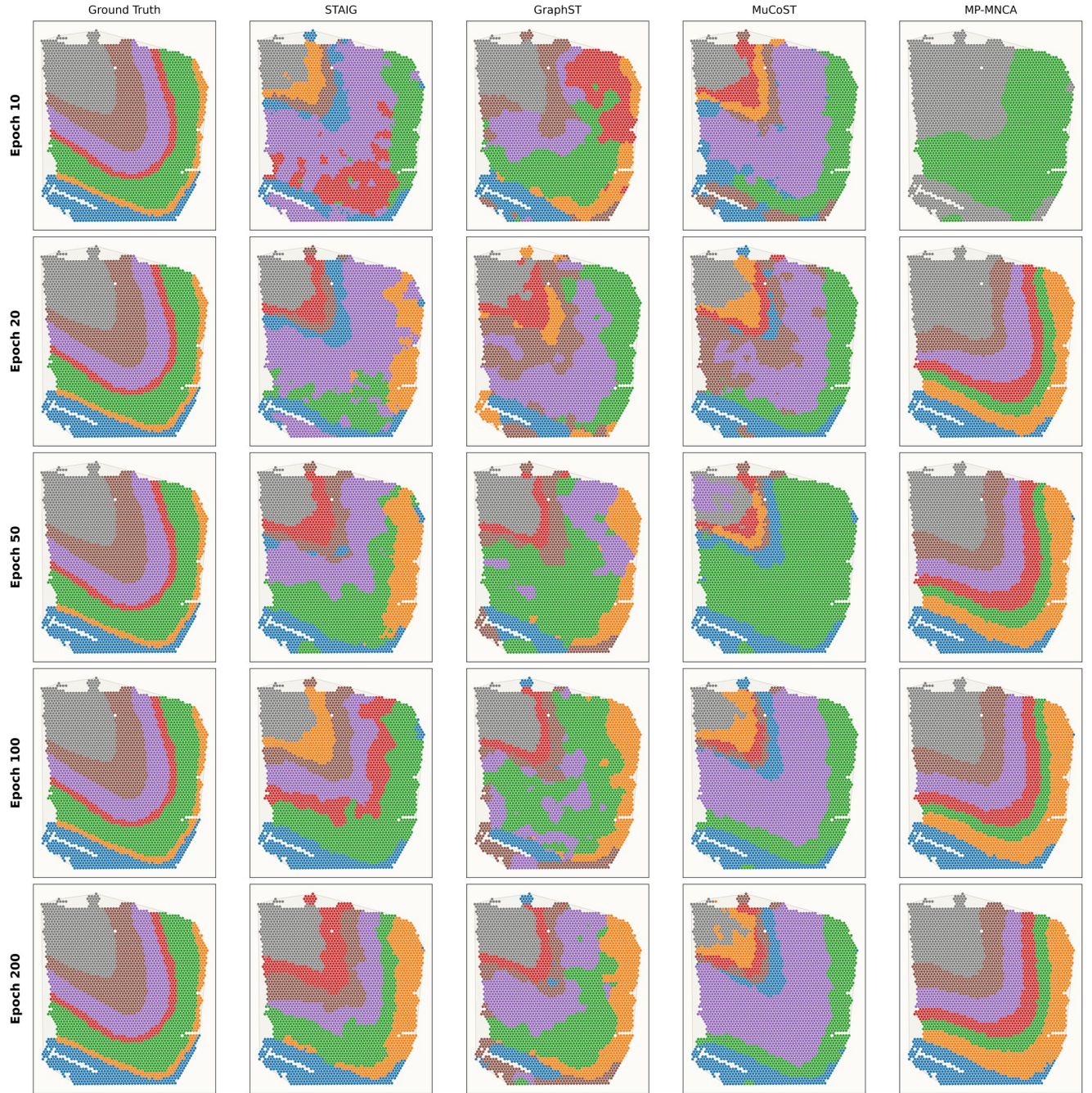


Figure 21. **Epoch-wise spatial-domain evolution on DLPFC section 151674.** The progression across checkpoints provides a qualitative comparison of optimization dynamics and the emergence of cortical-layer structure for the four evaluated methods.

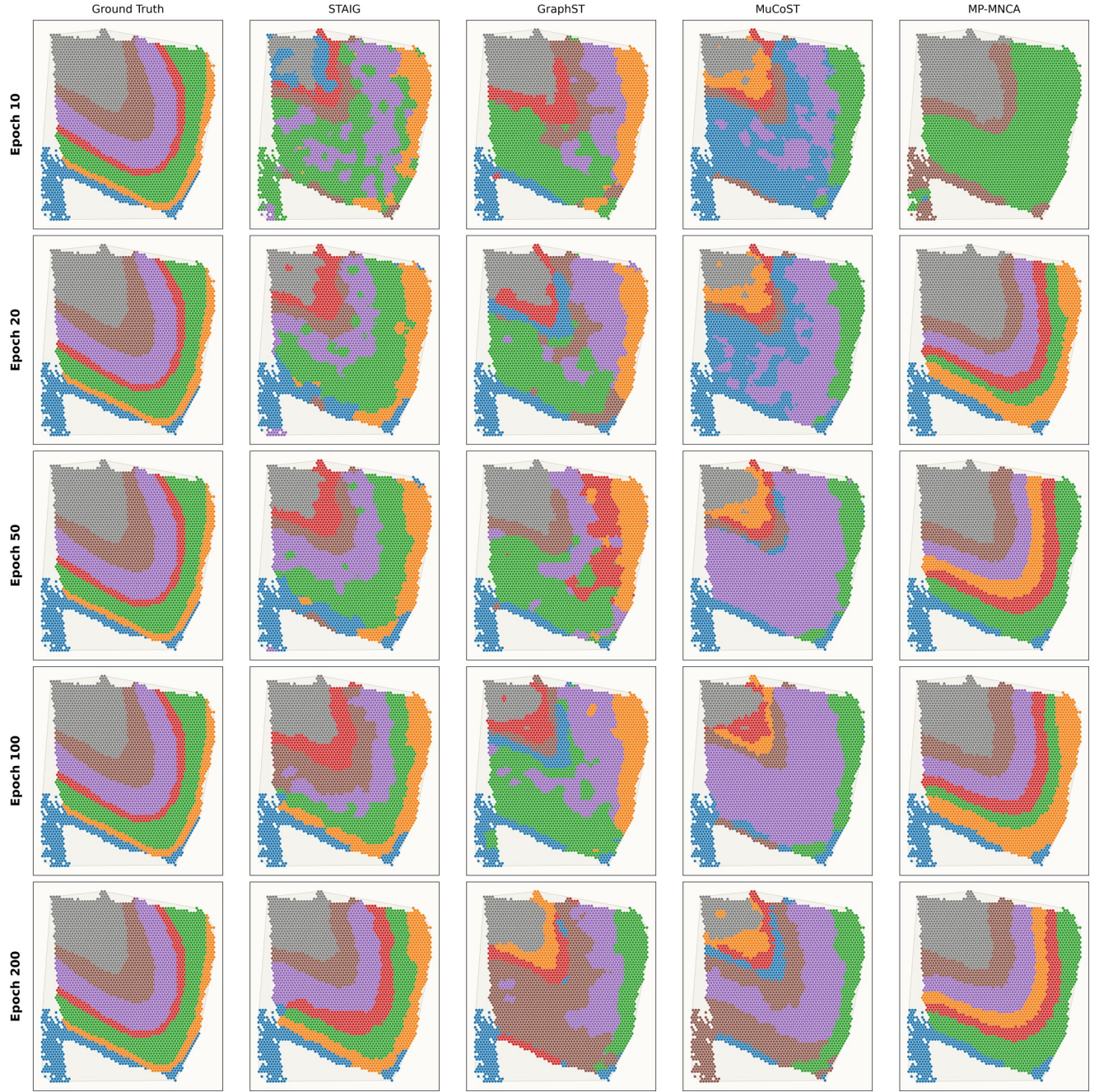


Figure 22. **Epoch-wise spatial-domain evolution on DLPFC section 151675.** This challenging section provides an important counterexample to uniform improvement across sections. The checkpoint maps reveal differences in layer fragmentation, boundary placement, and optimization behavior over the complete 200-epoch training budget.

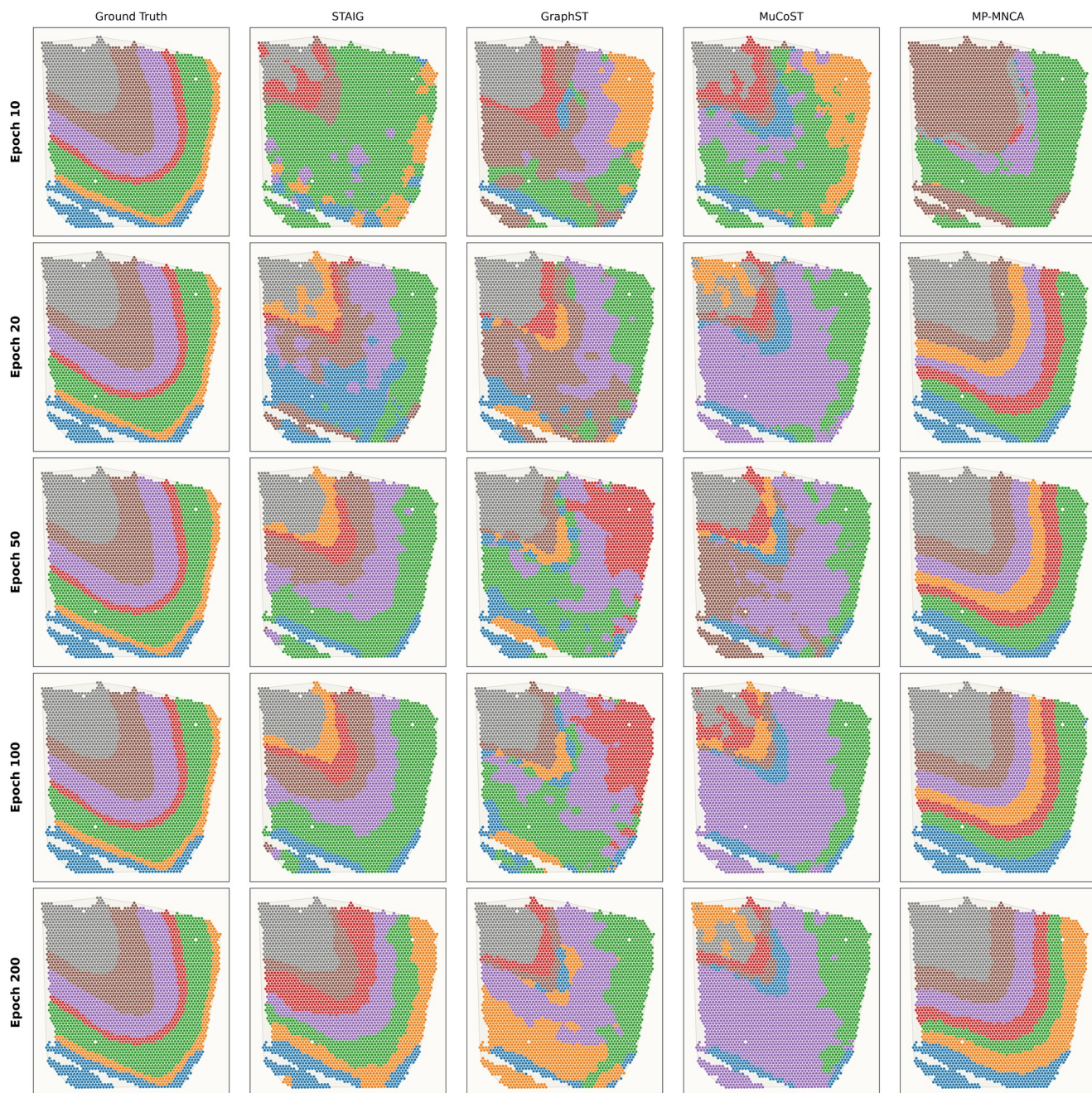


Figure 23. **Epoch-wise spatial-domain evolution on DLPFC section 151676.** Predictions at epochs 10, 20, 50, 100, and 200 complete the qualitative comparison across all 12 DLPFC sections and illustrate the evolution of cortical-domain structure under each optimization procedure.