MELON: Multimodal Learning Framework for Spatial Multi-Omics Data Integration

Anonymous Authors[1]

Abstract

Spatial multi-omics technologies measure multiple molecular modalities on the same tissue section, but existing integration methods optimize for shared structure and rarely preserve the cross-modal synergistic signal that emerges only from joint observation across modalities. We present MELON, a representation-learning framework built around a partial-information-decomposition (PID)-guided contrastive objective that explicitly preserves redundant, unique, and synergistic cross-modal information, combined with a learned neighborhood-aware spatial bias that respects local tissue structure. On a controlled simulation isolating cross-modal synergy, MELON recovers a synergy-only label well above chance while seven multi-omics baselines remain at chance. On three real-data benchmarks spanning RNA–ATAC and RNA–protein settings, MELON achieves consistently higher agreement with anatomical reference labels than seven established baselines and produces more spatially contiguous domains with sharper boundaries; a tri-modal tonsil extension confirms that these gains transfer beyond the bi-modal setting.

1. Introduction

Spatial multi-omics technologies measure multiple molecular and morphological modalities on the same tissue section while preserving its spatial layout, enabling the study of cellular identity, regulatory state, and microenvironmental organization in situ. Representative settings include transcriptome and proteome measurements in lymphoid tissue, transcriptome and epigenome assays such as spatial ATAC plus RNA-seq, and transcriptome with histology integration that combines molecular profiles with tissue morphology (Long et al., 2024; Zhang et al., 2023; Jiang et al., 2023; Coleman et al., 2025). A key analysis task is spatial domain identification, which groups spatial locations into biologically meaningful regions or cell-state neighborhoods, and is naturally framed as a multi-modality representation-learning problem in which each modality contributes complementary information: gene expression captures transcriptional output, while paired modalities such as chromatin accessibility, protein abundance, or hematoxylin and eosin (H&E)-stained histology images add regulatory, phenotypic, or structural context.

Existing approaches range from spatial-agnostic factor analyses to dedicated spatial multi-omics models with graph or smoothness priors (§2), and they share a common objective: capture what the modalities have in common. This is a substantial limitation: spots with similar transcriptomes can differ in chromatin accessibility, protein abundance, or tissue morphology, reflecting distinct regulatory programs, signaling states, or microenvironmental contexts. Accessible enhancers may indicate lineage priming before RNA changes are detectable, protein levels may reflect post-transcriptional regulation not captured by gene expression, and histology may reveal stromal or immune structure only weakly correlated with molecular profiles. Methods that prioritize consensus structure therefore collapse biologically distinct spots into the same cluster, since alignment-, reconstruction-, and shared-embedding objectives all emphasize redundant information by construction.

To address this, we propose MELON (Multi-modal lEarning framework for spatial muLti-Omics integratioN), a spatially aware framework that preserves shared, modality-specific, and synergistic cross-modal information. MELON encodes each modality with a dedicated encoder, then injects a learned neighborhood-aware spatial bias from a graph-attention module into a fusion transformer that models higher-order cross-modal interactions. The whole stack is trained with a partial-information-decomposition (PID)-guided (Williams & Beer, 2010; Bertschinger et al., 2014) multi-modality contrastive objective (Tian et al., 2020; Dufumier et al., 2024) that contrasts each single-modality embedding against the joint embedding, encouraging the learned representation to retain redundant, unique, and synergistic signal rather than collapsing to what the modalities share.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Submitted to the AI for Science workshop (ICML 2026).

Figure 1 summarizes the framework.

We evaluate MELON on paired spatial multi-omics datasets spanning molecular–molecular and molecular–protein settings, plus a tri-modal RNA, protein, and H&E extension. Our contributions are:

- **Problem framing.** We formulate spatial multi-omics integration as preserving shared, modality-specific, *and* synergistic cross-modal information, rather than only what the modalities share.
- **Architecture.** We propose MELON, which injects a learned NeighborhoodGAT spatial bias directly into the multi-modality fusion-transformer attention of a PID-guided contrastive backbone.
- **Empirical synergy preservation.** On a controlled simulation isolating cross-modal synergy, MELON recovers a synergy-only label well above chance while every baseline remains at chance.
- **Empirical utility.** On three paired spatial multi-omics benchmarks plus a tri-modal tonsil extension, MELON produces more accurate, spatially coherent, and biologically interpretable spot embeddings than seven established baselines.

2. Related Work

Existing spatial multi-omics integration methods learn a unified embedding from heterogeneous modalities but typically optimize for shared structure: Seurat (Hao et al., 2021), MOFA+ (Argelaguet et al., 2020), TotalVI (Gayoso et al., 2021), and MultiVI (Ashuach et al., 2023) treat spots as i.i.d. and ignore spatial neighborhoods; MEFISTO (Velten et al., 2022) adds Gaussian-process spatial priors but its reconstruction loss emphasizes redundancy; SpatialGlue (Long et al., 2024) couples graph convolutions with a VAE per modality, yet still relies on reconstruction and can oversmooth boundaries. MultiGate (Miao et al., 2025) uses a two-level graph attention autoencoder that couples spatial context with cross-modality feature relationships, and MISO (Coleman et al., 2025) integrates spatial molecular measurements with H&E morphology in a deep multimodal framework. None target cross-modal synergistic signal explicitly.

Contrastive representation learning offers a complementary lens. InfoNCE (van den Oord et al., 2018) underpins SimCLR (Chen et al., 2020), CLIP (Radford et al., 2021), and the multi-view CMC framework of Tian et al. (2020); standard CLIP-style two-modality losses provably capture only the redundant component R of the partial information decomposition (PID, Williams & Beer (2010); Bertschinger et al. (2014)). CoMM (Dufumier et al., 2024) introduced a modality-masking InfoNCE that aligns with all four PID components (R, U_1, U_2, synergy S); MELON adopts this objective and applies it to spatial multi-omics tokens.

Two families inject spatial structure into representation learning for spatially resolved data. Graph-neural-network approaches building on graph attention (Veličković et al., 2018) do explicit message passing over a spatial graph (e.g., STAGATE (Dong & Zhang, 2022) as a graph attention autoencoder, GraphST (Long et al., 2023) for graph contrastive learning, and SpaceFlow (Ren et al., 2022)), and are designed for a single modality. Attention-bias approaches instead modulate Transformer attention with positional priors, e.g. ALiBi (Press et al., 2022) and relative position biases (Raffel et al., 2020; Shaw et al., 2018). MELON's NeighborhoodGAT belongs to the second family: it outputs an additive log-space attention bias from a k-NN spatial prior plus a feature-aware GAT score, without performing message passing.

3. Methods

MELON consists of an architecture (§3.2–§3.4) that maps a multimodal spatial input to a spot embedding through modality-specific encoders, a NeighborhoodGAT spatial bias, and a fusion transformer, and a training procedure (§3.5–§3.6) based on label-preserving augmentations, modality masking, and a PID-guided contrastive loss. Preprocessing, token layouts, and other architectural details are deferred to Appendix B.

3.1. Problem Formulation

We consider N spatial spots, each observed with n aligned modalities and a 2D coordinate. Spot i is represented by per-modality feature vectors $x_i^{(m)} \in \mathbb{R}^{d_m}$ for $m = 1, \ldots, n$ and a coordinate $s_i \in \mathbb{R}^2$, giving matrices $X^{(m)} \in \mathbb{R}^{N \times d_m}$ and $S \in \mathbb{R}^{N \times 2}$. We write the full input as $X = \left(X^{(1)}, \ldots, X^{(n)}, S\right)$. The framework is modality-agnostic: each $X^{(m)}$ can be a count matrix, a continuous feature matrix, or a low-rank reduction of either, and the same architecture applies for any $n \geq 2$ (we instantiate $n = 2$ in the bi-modal experiments and $n = 3$ in the tonsil tri-modal extension). The goal is to learn spot embeddings $\left(x_i^{(1)}, \ldots, x_i^{(n)}, s_i\right) \mapsto z_i \in \mathbb{R}^d$ that support spatial domain clustering while preserving shared, modality-specific, and synergistic cross-modal information.

3.2. Modality Specific Encoders

Each modality m is first mapped by a dedicated encoder f_m to an intermediate hidden-feature tensor $H^{(m)}$ in its native feature space, and then projected by a lightweight converter ϕ_m to a per-modality token tensor $T^{(m)}$ in a shared

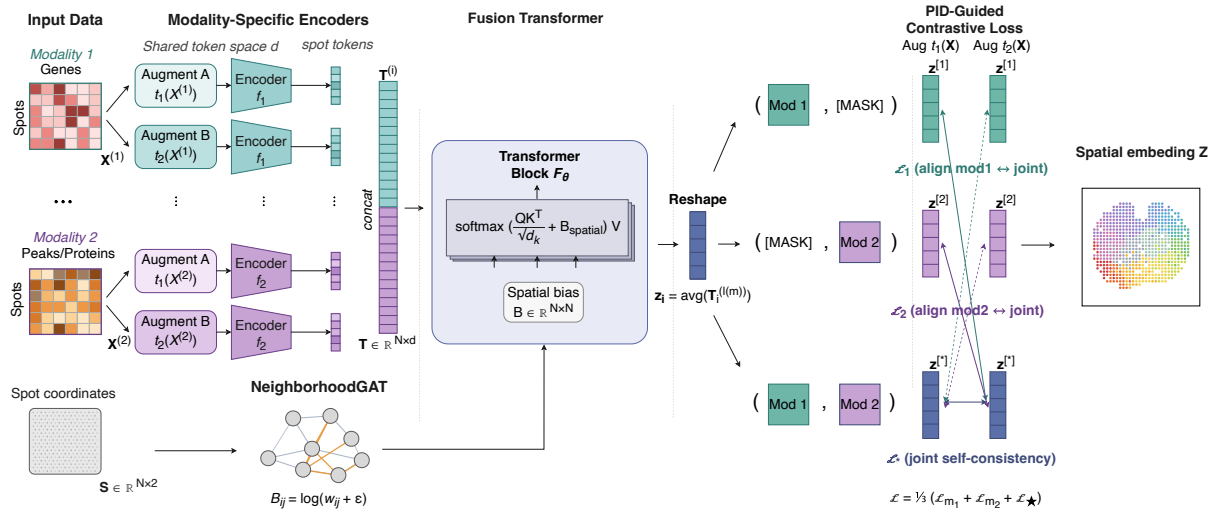


Figure 1. Overview of the MELON framework. The input is a pair of modality matrices (e.g., genes and ATAC peaks or proteins) together with spot coordinates. Each modality is passed through two label-preserving augmentations and its own encoder into a shared token space. The per-modality tokens are then fused by a transformer whose attention is biased by a learned spatial term produced by NeighborhoodGAT from the k-NN spatial graph. Per-spot embeddings are computed under three modality masks (Mod 1 only, Mod 2 only, and joint) and trained with a multi-modality contrastive loss that aligns each single-modality view with the joint view, targeting the redundancy, modality-specific uniqueness, and synergy components of PID. The resulting spatial embedding supports downstream clustering.

d-dimensional token space:

$$H^{(m)} = f_m(X^{(m)}) \in \mathbb{R}^{N \times L_m \times d_e}, \quad T^{(m)} = \phi_m(H^{(m)}) \in \mathbb{R}^{N \times L'_m \times d}, \quad m \in \{1, \dots, n\}. \tag{1}$$

Here N is the number of spots, L_m and L'_m are modality-specific token-sequence lengths (equal to 1 under the spot-level tokenization of Appendix B.2), d_e is the encoder's native hidden width, and d is the shared token width used by the fusion transformer. The same encoders are shared across both augmentations and all $n+1$ modality masks (a total of $2(n+1)$ training embeddings per spot). Separate encoders preserve modality-specific structure, while the converters ϕ_m align all modalities into a common token space of dimension d for downstream fusion. Default architectures and tokenization details are given in Appendix B.2.

3.3. Neighborhood-Aware Graph Attention

To encode local tissue structure, we combine a k-nearest-neighbor spatial prior with a feature-aware graph attention layer (NeighborhoodGAT) that produces a learned attention bias for the fusion transformer. Let $\{s_i\}_{i=1}^N$ be the 2D coordinates and $x_i \in \mathbb{R}^{d_x}$ the per-spot node features obtained by concatenating the per-modality token streams. From the coordinates we build a binary kNN mask $M \in \{0,1\}^{N \times N}$ and a Gaussian distance prior

$$w_{ij}^{\text{spatial}} = \exp\left(-\frac{\|s_i - s_j\|_2^2}{2\sigma^2}\right) M_{ij}, \tag{2}$$

with $\sigma = \text{median}\big(\{\|s_i - s_j\|_2 : M_{ij} = 1\}\big)$, which depends on the spatial coordinates alone. For each of n_h attention heads we project node features through a learned matrix $W^{(h)} \in \mathbb{R}^{d_h \times d_x}$, $h_i^{(h)} = W^{(h)} x_i$, and score each ordered pair (i, j) with two per-head learned vectors $a_{\text{src}}^{(h)}, a_{\text{dst}}^{(h)} \in \mathbb{R}^{d_h}$ in additive GAT form:

$$\alpha_{ij}^{(h)} = \text{LeakyReLU}\Big(a_{\text{src}}^{(h)\top} h_i^{(h)} + a_{\text{dst}}^{(h)\top} h_j^{(h)}\Big). \tag{3}$$

We add the log spatial prior to the edge logit, mask non-kNN entries to $-\infty$, and softmax over neighbors,

$$w_{ij}^{(h)} = \frac{\exp\big(\alpha_{ij}^{(h)} + \log w_{ij}^{\text{spatial}}\big)\, M_{ij}}{\sum_{j' :\, M_{ij'}=1} \exp\big(\alpha_{ij'}^{(h)} + \log w_{ij'}^{\text{spatial}}\big)}, \tag{4}$$

and the final bias is the head-averaged log weight,

$$w_{ij} = \frac{1}{n_h} \sum_{h=1}^{n_h} w_{ij}^{(h)}, \qquad B_{ij} = \log(w_{ij} + \varepsilon), \tag{5}$$

where $\varepsilon > 0$ is for numerical stability. The spatial prior w^{spatial} and the kNN mask M are deterministic functions of the spot coordinates and act as a structural constraint that restricts attention to local neighborhoods; the GAT scores $\alpha^{(h)}$ are feature-dependent and let the bias adapt to local content. Bias tiling and masked-embedding subsetting are described in Appendix B.3.

3.4. Fusion Transformer

Given the per-modality tokens and the spatial bias B, the fusion module stacks per-modality, per-spot tokens from all

N spots in the batch into a single token sequence,

$$T = \mathrm{concat}\big(T^{(1)}, T^{(2)}, \ldots, T^{(n)}\big) \in \mathbb{R}^{nN \times d}, \quad (6)$$

where each $T^{(m)} \in \mathbb{R}^{N \times d}$ contributes one token per spot. Transformer attention (Vaswani et al., 2017) then runs jointly across all nN spot–modality tokens, with the spot-level bias B from NeighborhoodGAT tiled into the $n \times n$ modality blocks of an $nN \times nN$ matrix B_{spatial} (Appendix B.3), so the same spot-level bias governs both within-modality and cross-modal attention. Adding a learned or geometry-derived bias to the attention logits follows the attention-bias paradigm used by ALiBi (Press et al., 2022) and T5/Shaw-style relative position biases (Raffel et al., 2020; Shaw et al., 2018); here the bias is derived from the spatial k-NN graph via NeighborhoodGAT rather than from token indices:

$$\mathrm{Attn}(Q, K, V) = \mathrm{softmax}\left(\frac{QK^\top}{\sqrt{d_k}} + B_{\text{spatial}}\right) V. \quad (7)$$

After fusion, the output tokens $\tilde{T} = F_\theta(T; B_{\text{spatial}})$ are reshaped back to modality-by-spot blocks and averaged across modalities to recover one embedding per spot:

$$z_i = \frac{1}{n} \sum_{m=1}^{n} \tilde{T}_i^{(m)}. \quad (8)$$

Exact token layouts and recovery operations are deferred to Appendix B.2 and Appendix B.3. We write the full encoder pipeline (modality encoders + NeighborhoodGAT + fusion + spot pooling) as f_θ in the remainder of this section.

3.5. Multimodal Augmentations and Masked Embeddings

Following standard contrastive self-supervised practice (Chen et al., 2020; Tian et al., 2020), for each batch we draw two label-preserving augmentations $t_1, t_2 \sim \mathcal{T}$ and form two augmented multimodal inputs

$$X' = t_1(X), \qquad X'' = t_2(X). \quad (9)$$

The augmentation family $\mathcal{T}$ is chosen per assay rather than fixed by the architecture; the concrete instantiations used in our experiments (e.g. Gaussian noise on continuous features, Bernoulli dropout on sparse binary features, aligned cross-spot mixing) are described in Appendix B.1. Each augmented input is passed through f_θ under $n+1$ modality masks: n *single-modality* masks $\mathcal{M}^{[m]}$ that keep only modality m and replace every other modality slot with a learned [MSK] token, and one *joint* mask $\mathcal{M}^{[\star]}$ that keeps all modalities, yielding embeddings

$$z'^{[k]} = f_\theta(X', \mathcal{M}^{[k]}), \quad z''^{[k]} = f_\theta(X'', \mathcal{M}^{[k]}), \quad k \in \{1, \ldots, n, \star\}, \quad (10)$$

or $2(n+1)$ embeddings per spot in total. The n single-modality embeddings probe what each modality contributes in isolation; the joint embedding $z^{[\star]}$ is what all single-modality embeddings are aligned against in the contrastive loss of §3.6.

3.6. PID-Guided Contrastive Learning

We train MELON with a multi-modality contrastive objective whose components target the four terms of the partial information decomposition (PID) of Williams & Beer (2010); Bertschinger et al. (2014): redundancy (R), modality-specific uniqueness (U_m), and synergy (S). The objective extends the multi-modality contrastive paradigm of Tian et al. (2020) via the modality-masked, PID-aware decomposition of Dufumier et al. (2024), instantiated on the per-modality token streams produced by our spatially biased fusion transformer. Given two batches of normalized embeddings $Z, Z' \in \mathbb{R}^{N \times d}$, with rows (z_i, z'_i) forming positive pairs, we use the InfoNCE estimator (van den Oord et al., 2018)

$$\widehat{I}_{\mathrm{NCE}}(Z, Z') = \mathbb{E}_{(z, z'_{\text{pos}})} \mathbb{E}_{z'_{\text{neg}}} \left[\log \frac{\exp(\mathrm{sim}(z, z'_{\text{pos}})/\tau)}{\sum_{z'_{\text{neg}}} \exp(\mathrm{sim}(z, z'_{\text{neg}})/\tau)} \right], \quad (11)$$

where $\mathrm{sim}(\cdot, \cdot)$ is the cosine similarity and $\tau > 0$ is a temperature. Expectations are approximated with in-batch positives and negatives.

Recall from §3.5 that for each spot we have $2(n+1)$ embeddings: a joint multimodal embedding $z^{[\star]}$ and n single-modality embeddings $z^{[m]}$, each computed under both augmentations $A = t_1(X)$ and $B = t_2(X)$. We denote the corresponding batch matrices by $Z'^{[k]}$ and $Z''^{[k]}$ for $k \in \{1, \ldots, n, \star\}$.

For each mask k, the per-mask loss pairs the mask-k embedding under one augmentation against the joint embedding under the *other* augmentation, in a symmetrized form:

$$\mathcal{L}_k = -\frac{1}{2}\Big(\widehat{I}_{\mathrm{NCE}}\big(Z'^{[k]}, Z''^{[\star]}\big) + \widehat{I}_{\mathrm{NCE}}\big(Z''^{[k]}, Z'^{[\star]}\big)\Big). \quad (12)$$

The cross-augmentation pairing (anchor from A, joint embedding from B, and vice versa) prevents the same augmentation from appearing on both sides of an InfoNCE term. The total MELON loss averages over all $n+1$ masks, including the joint mask itself, which acts as a self-consistency term for the joint embedding:

$$\mathcal{L}_{\mathrm{MELON}} = \frac{1}{n+1}\Big(\mathcal{L}_\star + \sum_{m=1}^{n} \mathcal{L}_m\Big). \quad (13)$$

Concretely, $\mathcal{L}_\star$ aligns the joint embedding under both augmentations (encouraging augmentation invariance of

the multimodal representation), while each $\mathcal{L}_m$ aligns the modality-m-only embedding with the joint embedding (encouraging the joint embedding to remain consistent with what each modality contributes in isolation). For $n = 2$ this reduces to a three-term objective $\mathcal{L} = \frac{1}{3}(\mathcal{L}_{m_1} + \mathcal{L}_{m_2} + \mathcal{L}_\star)$. The synergy-identification argument is inherited from the modality-masked InfoNCE of Dufumier et al. (2024).

4. Experiments

4.1. Simulation

Real-data clustering metrics (ARI, NMI) cannot isolate *which* aspects of representation quality drive a method's success: spatial-domain preservation, modality-specific attribute preservation, and cross-modal information retention all contribute to spatially coherent domains. To separate these effects, we evaluate MELON and six external integration methods on a controlled spatial multi-omics simulation with four ground-truth targets: a coarse spatial domain R, modality-specific attributes U (RNA-side) and C (ATAC-side), and a binary cross-modal matching label Y generated from a hidden permutation π that ties U to C. Each dataset is a paired RNA/ATAC simulated section, generated under three independent seeds. We freeze each method's embedding and probe its information about R, U, C, and Y separately, reporting mean $\pm$ std across the three seeds. Full data-generating process, probe specifications, hyperparameters, and validation are in Appendix C.

Method comparison. Table 1 reports balanced accuracy on all four targets. MELON is the only method that recovers the synergy label Y clearly above chance; every external baseline remains near chance on Y, even those that preserve the modality-specific attributes U and C at high accuracy. Preserving U and C alone is therefore not sufficient: the embedding must additionally retain the cross-modal binding between them. The simulation is best read as a controlled cross-modal synergy assay, with the headline claim specifically about Y.

Component ablation. Within MELON, we ablate two design axes on Y recovery: data augmentation and the NeighborhoodGAT spatial bias (Table 2). Removing augmentation produces the largest single drop in Y recovery, identifying it as the dominant MELON-internal contributor to cross-modal information retention. Removing the spatial bias also reduces both Y and the spatial-domain target R, so spatial conditioning helps both spatial-domain and cross-modal recovery. As a unimodal sanity check, retraining MELON on RNA-only or ATAC-only inputs collapses Y to chance, confirming that Y recovery requires joint observation of both modalities rather than memorisation of either modality-specific attribute alone (full six-variant ablation including double-knockouts and unimodal controls in Appendix C.7).

4.2. Main Performance

We evaluate MELON on three paired spatial multi-omics datasets and compare against the same seven baselines as §4.1, dispatching TotalVI for the RNA–protein lymph node and MultiVI for the RNA–ATAC datasets: (i) a Mouse E15.5 brain MISAR-seq (Jiang et al., 2023) RNA–ATAC sagittal cryosection (Fig. 2); (ii) a Human lymph node 10x Visium RNA–ADT co-profile (Appendix Fig. 5); and (iii) a Mouse E13 embryo whole-section RNA–ATAC profile (Appendix Fig. 6). Dataset statistics and curated anatomical labels are in Appendix A; metric definitions in Appendix D.2.

Performance across datasets and metrics. Across the eleven-metric panel (full Appendix Table 13; eight-metric subset in Table 3), MELON attains the best median (median $\pm$ IQR over swept cluster numbers k) on *all 11 metrics* for the Mouse E15.5 brain, on *9 of 11* for the Human lymph node (with SpatialGlue leading homogeneity and MISO leading FMI), and on *10 of 11* for the Mouse E13 embryo (with Seurat WNN winning only the internal-compactness CHI). The spatial-aware SpARI mirrors the ARI lead, consistent with sharper anatomical boundaries at the DPallm/DPallv, subpallium/thalamus, and midbrain/hindbrain interfaces in the brain (Fig. 2), the follicle/T-cell-zone and medulla/capsule interfaces in the lymph node (Appendix Fig. 5), and the neural/mesenchymal contacts in the embryo (Appendix Fig. 6).

4.3. Scalability on Triple-Modal Integration

Beyond the bi-modal benchmarks, MELON's fusion architecture extends naturally to three or more modalities, and its modality-agnostic encoder design also accommodates image-derived features as an additional input modality alongside molecular profiles. We demonstrate both on a public 10x Visium CytAssist FFPE human tonsil sample with paired RNA, antibody capture, and H&E modalities (full preprocessing in Appendix D.7). We reuse the bi-modal baselines from §4.2 that support tri-modal inputs (MISO, MOFA+, MEFISTO, Seurat WNN, SpatialGlue) and add the tri-modal-specific GROVER (Xiao et al., 2025). With no manual annotation for this slide, we evaluate cluster recovery against six canonical tonsil compartments (germinal center, mantle, T-zone, plasma, myeloid, epithelium) defined by IHC marker rules on the antibody panel; per-compartment best-cluster F1 against the marker pseudo-mask is reported, with the cross-compartment mean as the tissue-wide score.

MELON achieves the highest mean F1 (Table 4) and ranks first on the three largest compartments (mantle, T-zone, plasma; Appendix Fig. 13). Modality ablation reduces mean F1 substantially when any single modality is dropped

Table 1. Method comparison on the controlled simulation. Balanced accuracy and macro F1 are reported as mean ± std across three seeds; AUROC is reported for the binary target Y. Best per column in **bold**.

	R		U		C		Y		
Method	BalAcc	F1	BalAcc	F1	BalAcc	F1	BalAcc	F1	AUROC
MELON	0.930 ± 0.019	0.931 ± 0.016	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{0.718 \pm 0.044}$	$\mathbf{0.768 \pm 0.055}$	$\mathbf{0.881 \pm 0.081}$
SpatialGlue	$\mathbf{0.980 \pm 0.012}$	$\mathbf{0.979 \pm 0.012}$	0.964 ± 0.009	0.964 ± 0.008	0.939 ± 0.010	0.938 ± 0.011	0.512 ± 0.011	0.496 ± 0.012	0.604 ± 0.023
Seurat WNN	0.972 ± 0.012	0.972 ± 0.013	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	0.499 ± 0.001	0.473 ± 0.004	0.534 ± 0.016
MISO	0.673 ± 0.041	0.662 ± 0.045	0.877 ± 0.064	0.876 ± 0.064	0.966 ± 0.018	0.965 ± 0.018	0.503 ± 0.003	0.481 ± 0.005	0.552 ± 0.023
MultiVI	0.229 ± 0.011	0.222 ± 0.010	0.990 ± 0.014	0.990 ± 0.014	0.198 ± 0.043	0.183 ± 0.045	0.501 ± 0.001	0.477 ± 0.002	0.496 ± 0.007
MEFISTO	0.212 ± 0.004	0.192 ± 0.006	0.937 ± 0.022	0.935 ± 0.023	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	0.504 ± 0.006	0.484 ± 0.013	0.529 ± 0.018
MOFA+	0.208 ± 0.003	0.190 ± 0.006	0.928 ± 0.027	0.924 ± 0.031	0.999 ± 0.001	0.999 ± 0.001	0.500 ± 0.005	0.478 ± 0.005	0.515 ± 0.022

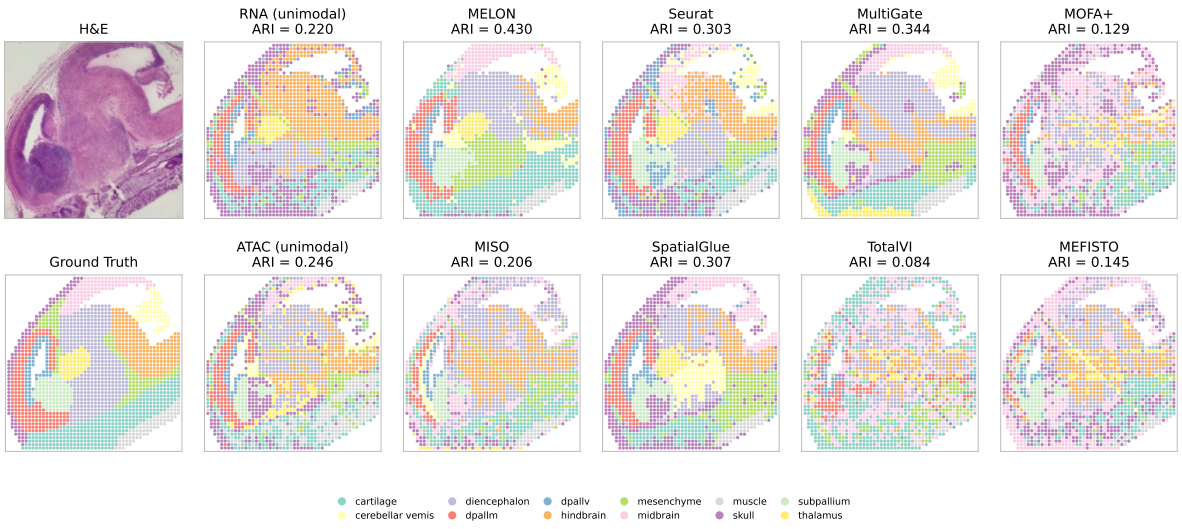


Figure 2. Mouse E15.5 brain results: H&E tissue image, ground-truth annotation, and spatial domain assignments from MELON and seven multimodal baselines (ARI labeled per panel; methods ordered consistently across datasets). Quantitative clustering metrics across all methods are reported in Table 3.

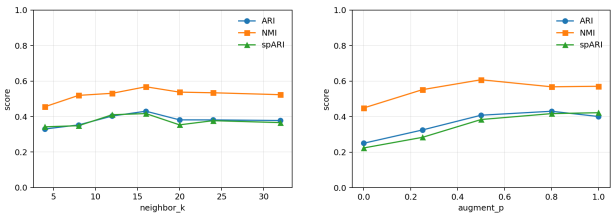
Table 2. MELON variant ablation on the simulation. Balanced accuracy (mean ± std across three seeds) on each of the four targets. Best per column in **bold**. Macro-F1 and AUROC for Y are reported in Appendix Table 11.

Variant	R	U	C	Y
MELON (full)	$\mathbf{0.930 \pm 0.019}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{0.718 \pm 0.044}$
− spatial bias	0.870 ± 0.045	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	0.592 ± 0.033
− augmentation	0.912 ± 0.042	$\mathbf{1.000 \pm 0.000}$	0.997 ± 0.001	0.516 ± 0.018

(Table 4, lower rows), confirming that each modality contributes substantively to the integrated performance. Per-compartment AUROC, H&E overlay, and per-method spatial maps are in Appendix D.7.

4.4. Sensitivity Analysis and Ablation Study

We sweep MELON's two main hyperparameters on the Mouse E15.5 brain dataset: the spatial neighborhood size k_{nn} used by NeighborhoodGAT, and the augmentation probability p_{aug} used during contrastive training (Fig. 3). MELON is robust to the spatial graph size with graceful degradation in both directions around the default, indicating that the spatial bias does not require careful tuning. Augmentation is the dominant contributor to real-data performance: ARI rises sharply once augmentation is enabled and then saturates with little dependence on its exact frequency. Removing either augmentation or the spatial bias substantially degrades every clustering metric on the Mouse E15.5 brain, with ARI and SpARI roughly halving (Table 5; full eight-metric panel in Appendix D.5), confirming that the simulation-level component conclusions transfer to real spatial multi-omics tissue.



(a) Sensitivity of k_{nn}. *(b)* Sensitivity of p_{aug}.

Figure 3. Parameter sensitivity analysis on the Mouse E15.5 brain.

Table 3. Main performance: clustering metrics across the three benchmarked datasets (median ± IQR over swept cluster numbers k; $k \in \{6, 8, 10, 12, 14\}$ for Mouse E15.5 Brain and Human Lymph Node, $k \in \{10, 12, 14, 16, 18\}$ for Mouse E13 Embryo). Best per (dataset, metric) in **bold**. Full 11-metric panel in Appendix Table 13.

Dataset	Method	ARI	NMI	AMI	Homog.	SpARI	FMI	Jaccard	Purity
Mouse E15.5 Brain	**MELON**	**0.480 ± 0.022**	**0.569 ± 0.005**	**0.561 ± 0.006**	**0.532 ± 0.028**	**0.532 ± 0.030**	**0.567 ± 0.016**	**0.384 ± 0.016**	**0.678 ± 0.032**
	MISO	0.216 ± 0.011	0.372 ± 0.027	0.364 ± 0.026	0.355 ± 0.046	0.176 ± 0.014	0.324 ± 0.007	0.193 ± 0.003	0.485 ± 0.034
	Seurat WNN	0.291 ± 0.017	0.483 ± 0.007	0.475 ± 0.010	0.491 ± 0.041	0.224 ± 0.026	0.376 ± 0.016	0.226 ± 0.008	0.560 ± 0.041
	SpatialGlue	0.285 ± 0.024	0.515 ± 0.019	0.508 ± 0.016	0.524 ± 0.063	0.249 ± 0.021	0.381 ± 0.026	0.235 ± 0.024	0.625 ± 0.055
	MultiGate	0.344 ± 0.025	0.515 ± 0.006	0.508 ± 0.004	0.516 ± 0.050	0.321 ± 0.110	0.425 ± 0.061	0.270 ± 0.038	0.596 ± 0.027
	MultiVI	0.086 ± 0.019	0.212 ± 0.059	0.200 ± 0.056	0.199 ± 0.068	0.001 ± 0.014	0.211 ± 0.001	0.117 ± 0.002	0.372 ± 0.051
	MOFA+	0.130 ± 0.018	0.300 ± 0.017	0.289 ± 0.014	0.273 ± 0.033	0.085 ± 0.021	0.265 ± 0.024	0.152 ± 0.013	0.436 ± 0.021
	MEFISTO	0.151 ± 0.009	0.301 ± 0.018	0.292 ± 0.017	0.281 ± 0.031	0.087 ± 0.018	0.269 ± 0.019	0.155 ± 0.009	0.443 ± 0.021
Human Lymph Node	**MELON**	**0.298 ± 0.028**	**0.379 ± 0.002**	**0.374 ± 0.001**	0.401 ± 0.033	**0.307 ± 0.038**	0.454 ± 0.018	**0.293 ± 0.015**	**0.643 ± 0.009**
	MISO	0.241 ± 0.046	0.305 ± 0.035	0.301 ± 0.032	0.263 ± 0.071	0.281 ± 0.017	**0.455 ± 0.003**	0.285 ± 0.010	0.576 ± 0.065
	Seurat WNN	0.182 ± 0.000	0.260 ± 0.006	0.256 ± 0.005	0.236 ± 0.017	0.185 ± 0.014	0.389 ± 0.018	0.240 ± 0.012	0.529 ± 0.004
	SpatialGlue	0.227 ± 0.049	0.375 ± 0.023	0.371 ± 0.024	**0.423 ± 0.023**	0.195 ± 0.084	0.371 ± 0.065	0.220 ± 0.061	0.624 ± 0.008
	MultiGate	0.148 ± 0.011	0.203 ± 0.013	0.199 ± 0.012	0.221 ± 0.038	0.147 ± 0.017	0.325 ± 0.022	0.192 ± 0.019	0.484 ± 0.001
	TotalVI	0.181 ± 0.006	0.258 ± 0.001	0.253 ± 0.000	0.232 ± 0.001	0.187 ± 0.011	0.393 ± 0.010	0.242 ± 0.007	0.525 ± 0.004
	MOFA+	0.212 ± 0.032	0.321 ± 0.033	0.318 ± 0.031	0.306 ± 0.103	0.229 ± 0.025	0.415 ± 0.036	0.261 ± 0.020	0.589 ± 0.073
	MEFISTO	0.200 ± 0.011	0.307 ± 0.007	0.302 ± 0.005	0.310 ± 0.045	0.221 ± 0.050	0.400 ± 0.044	0.249 ± 0.036	0.594 ± 0.046
Mouse E13 Embryo	**MELON**	**0.385 ± 0.007**	**0.528 ± 0.010**	**0.519 ± 0.007**	**0.546 ± 0.039**	**0.362 ± 0.026**	**0.452 ± 0.022**	**0.290 ± 0.023**	**0.655 ± 0.008**
	MISO	0.227 ± 0.014	0.441 ± 0.031	0.431 ± 0.028	0.417 ± 0.046	0.251 ± 0.005	0.337 ± 0.002	0.197 ± 0.002	0.521 ± 0.053
	Seurat WNN	0.266 ± 0.043	0.454 ± 0.017	0.443 ± 0.020	0.487 ± 0.017	0.223 ± 0.073	0.341 ± 0.047	0.201 ± 0.041	0.571 ± 0.005
	SpatialGlue	0.280 ± 0.014	0.460 ± 0.022	0.451 ± 0.018	0.472 ± 0.063	0.238 ± 0.043	0.355 ± 0.018	0.215 ± 0.016	0.600 ± 0.061
	MultiGate	0.170 ± 0.016	0.380 ± 0.008	0.369 ± 0.007	0.392 ± 0.031	0.168 ± 0.056	0.263 ± 0.036	0.151 ± 0.025	0.438 ± 0.021
	MultiVI	0.092 ± 0.003	0.224 ± 0.033	0.212 ± 0.029	0.164 ± 0.030	0.140 ± 0.005	0.321 ± 0.000	0.149 ± 0.000	0.307 ± 0.006
	MOFA+	0.201 ± 0.012	0.389 ± 0.012	0.377 ± 0.009	0.383 ± 0.040	0.199 ± 0.033	0.295 ± 0.017	0.173 ± 0.010	0.491 ± 0.030
	MEFISTO	0.207 ± 0.015	0.391 ± 0.003	0.381 ± 0.007	0.392 ± 0.018	0.189 ± 0.050	0.302 ± 0.032	0.178 ± 0.021	0.496 ± 0.017

Table 4. Tri-modal extension on human tonsil. Mean best-cluster F1 across six canonical tonsil compartments (germinal center, mantle, T-zone, plasma, myeloid, epithelium) defined by IHC marker rules on the antibody panel; best in **bold**. Per-compartment breakdown in Appendix Fig. 13.

Baseline	F1 (↑)	MELON variant	F1 (↑)
MISO (Coleman et al., 2025)	0.346 ± 0.008	**MELON (full)**	**0.415 ± 0.009**
GROVER (Xiao et al., 2025)	0.351 ± 0.008	− image	0.355 ± 0.008
MOFA+ (Argelaguet et al., 2020)	0.339 ± 0.010	− protein	0.322 ± 0.008
MEFISTO (Velten et al., 2022)	0.337 ± 0.010	− RNA	0.322 ± 0.008
Seurat WNN (Hao et al., 2021)	0.370 ± 0.010		
SpatialGlue (Long et al., 2024)	0.286 ± 0.008		

Table 5. MELON component ablation on the Mouse E15.5 brain. Median ± IQR over the swept cluster numbers k used in Table 3. Best per metric in **bold**.

Variant	ARI	NMI	SpARI	FMI
MELON full	**0.383 ± 0.024**	**0.553 ± 0.017**	**0.376 ± 0.028**	**0.466 ± 0.020**
− augmentation	0.204 ± 0.017	0.368 ± 0.004	0.174 ± 0.047	0.320 ± 0.038
− spatial bias	0.215 ± 0.017	0.355 ± 0.017	0.190 ± 0.038	0.334 ± 0.026

4.5. Biological Significance

To probe biological interpretability, we train a small joint pair-decoder that reconstructs both modalities from MELON's frozen embedding through a shared low-rank bottleneck, so each factor acts as a joint cross-modal program and high-scoring gene–peak (or gene–protein) pairs are those whose two channels co-vary on the same factor (Appendix D.6). Per-factor spatial usage broadly aligns with the anatomical reference labels (Appendix Figs. 7, 8, 9), with each high-variance factor concentrating in a distinct tissue region.

Selected high-support pairs surface canonical anatomical markers across all three datasets: neural-crest *Cdh19-Pmpcb* (Simões-Costa & Bronner, 2015) and progenitor *Mak-Pou3f1* (Zhu et al., 2014) pairs in the E13 embryo localize to the developing neural tube and surrounding mesenchyme (Fig. 12); cortical and lineage genes (*Myl4*, *Abi3bp*, *Fgfr3*) coupled to chromosomal-coordinate ATAC peaks in the E15.5 brain (Fig. 10); and gene-protein pairs (*ANXA1-FCGR3A*/CD16 (Bruhns, 2012); *EPCAM-PAX5* (Nutt et al., 1999); *CRLF1-CEACAM8*) in canonical immune compartments of the lymph node (Fig. 11). For these selected pairs, per-spot pair activity often shows sharper anatomical boundaries than either marginal channel alone, supporting investigation of spatially specific regulatory programs in spatial multi-omics tissues.

5. Conclusion

We presented MELON, a spatial multimodal representation-learning framework combining modality-specific encoding, neighborhood-aware spatial attention, transformer-based fusion, and a PID-guided contrastive objective that preserves redundant, modality-specific, and synergistic information. On a controlled simulation isolating cross-modal synergy, MELON is the only method to recover the synergy-only label above chance (BalAcc $= 0.718$) while seven baselines remain at chance (≤ 0.51). Across three paired spatial multi-omics benchmarks MELON yields spatially coherent domains with higher agreement to anatomical annotations than seven established baselines on the Mouse E15.5 brain (ARI 0.480 vs. next-best 0.344), human lymph node (ARI 0.298 vs. 0.241), and Mouse E13 embryo (ARI 0.385 vs. 0.280). The framework extends naturally to a tri-modal RNA, protein, and H&E tonsil setting, where MELON's full model achieves the highest mean compartment-recovery F1 (0.415 vs. best baseline 0.370) and where ablating any

single modality drops F1 by ≥ 0.06, confirming that each modality contributes complementary signal. Together, these results demonstrate that PID-guided contrastive learning combined with a learned spatial attention bias can preserve cross-modal information that shared-representation methods systematically miss, enabling the recovery of biologically meaningful regulatory programs that emerge only from joint modality observation. The framework is also general beyond spatial multi-omics: any setting where a multimodal latent structure is organized over a known geometric prior (single-cell trajectories, longitudinal imaging, multimodal time series) can plausibly benefit from injecting a learned bias from that prior into a PID-guided contrastive backbone.

MELON has been benchmarked with up to three modalities on standard 2D sections; behavior under more modalities, substantially larger tissues, or 3D coordinates remains untested. The fusion transformer's modest capacity means the recovered PID components are a lower bound rather than an exact decomposition. On the controlled simulation, MELON excels at the synergy label Y but is not the best on the redundancy and spatial-domain target R, reflecting an inherent trade-off in modality-masked contrastive objectives; more recently developed masking strategies are a natural direction for closing this gap. The augmentation setting may need re-tuning for substantially noisier modalities, and the k-NN spatial graph does not account for physical barriers such as tissue boundaries or folds; future work could replace the Euclidean prior with a tissue-aware graph constructed from H&E segmentation or anatomy-derived adjacency.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning, with a specific application to computational biology and spatial multi-omics data integration. The method is designed for foundational research on tissue-level biology and poses no foreseeable negative societal impact; potential positive impacts include improving the resolution and interpretability of spatial-omics analyses used in biomedical research. All datasets used in this work are publicly available, and no human-subjects data are collected.

References

Ainslie, J., Lee-Thorp, J., de Jong, M., Zemlyanskiy, Y., Lebrón, F., and Sanghai, S. GQA: Training generalized multi-query transformer models from multi-head checkpoints. *arXiv preprint arXiv:2305.13245*, 2023.

Argelaguet, R., Arnol, D., Bredikhin, D., Deloro, Y., Velten, B., Marioni, J. C., and Stegle, O. MOFA+: a statistical framework for comprehensive integration of multi-modal single-cell data. *Genome Biology*, 21(1):111, 2020.

Ashuach, T., Gabitto, M. I., Koodli, R. V., Saldi, G.-A., Jordan, M. I., and Yosef, N. MultiVI: deep generative model for the integration of multimodal data. *Nature Methods*, 20(8):1222–1231, 2023.

Bertschinger, N., Rauh, J., Olbrich, E., Jost, J., and Ay, N. Quantifying unique information. *Entropy*, 16(4):2161–2183, 2014.

Bruhns, P. Properties of mouse and human IgG receptors and their contribution to disease models. *Blood*, 119(24): 5640–5649, 2012.

Caliński, T. and Harabasz, J. A dendrite method for cluster analysis. *Communications in Statistics*, 3(1):1–27, 1974.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning (ICML)*, pp. 1597–1607. PMLR, 2020.

Coleman, K., Schroeder, A., Loth, M., Zhang, D., Park, J. H., Sung, J.-Y., Blank, N., Cowan, A. J., Qian, X., Chen, J., Jiang, J., Yan, H., Samarah, L. Z., Clemenceau, J. R., Jang, I., Kim, M., Barnfather, I., Rabinowitz, J. D., Deng, Y., Lee, E. B., Lazar, A., Gao, J., Furth, E. E., Hwang, T. H., Wang, L., Thaiss, C., Hu, J., and Li, M. Resolving tissue complexity by multimodal spatial omics modeling with MISO. *Nature Methods*, 22:530–538, 2025. doi: 10.1038/s41592-024-02574-2.

Cusanovich, D. A., Hill, A. J., Aghamirzaie, D., Daza, R. M., Pliner, H. A., Berletch, J. B., Filippova, G. N., Huang, X., Christiansen, L., DeWitt, W. S., et al. A single-cell atlas of in vivo mammalian chromatin accessibility. *Cell*, 174 (5):1309–1324, 2018.

Dong, K. and Zhang, S. Deciphering spatial domains from spatially resolved transcriptomics with an adaptive graph attention auto-encoder. *Nature Communications*, 13(1): 1739, 2022.

Dufumier, B., Castillo-Navarro, J., Tuia, D., and Thiran, J.-P. What to align in multimodal contrastive learning? *arXiv preprint arXiv:2409.07402*, 2024.

Fowlkes, E. B. and Mallows, C. L. A method for comparing two hierarchical clusterings. *Journal of the American Statistical Association*, 78(383):553–569, 1983.

Gayoso, A., Steier, Z., Lopez, R., Regier, J., Nazor, K. L., Streets, A., and Yosef, N. Joint probabilistic modeling of single-cell multi-omic data with totalVI. *Nature Methods*, 18(3):272–282, 2021.

Hao, Y., Hao, S., Andersen-Nissen, E., Mauck III, W. M., Zheng, S., Butler, A., et al. Integrated analysis of multimodal single-cell data. *Cell*, 184(13):3573–3587, 2021.

Hubert, L. and Arabie, P. Comparing partitions. *Journal of Classification*, 2(1):193–218, 1985.

Jiang, F., Zhou, X., Qian, Y., Zhu, M., Wang, L., Li, Z., Shen, Q., Wang, M., Qu, F., Cui, G., Chen, K., and Peng, G. Simultaneous profiling of spatial gene expression and chromatin accessibility during mouse brain development. *Nature Methods*, 20(7):1048–1057, 2023.

Long, Y., Ang, K. S., Li, M., Chong, K. L. K., Sethi, R., Zhong, C., Xu, H., Ong, Z., Sachaphibulkij, K., Chen, A., et al. Spatially informed clustering, integration, and deconvolution of spatial transcriptomics with GraphST. *Nature Communications*, 14(1):1155, 2023.

Long, Y., Ang, K. S., Sethi, R., Liao, S., Heng, Y., van Olst, L., Ye, S., Zhong, C., Xu, H., Zhang, D., Kwok, I., Husna, N., Jian, M., Ng, L. G., Chen, A., Gascoigne, N. R. J., Gate, D., Fan, R., Xu, X., and Chen, J. Deciphering spatial domains from spatial multi-omics with SpatialGlue. *Nature Methods*, 21(9):1658–1667, 2024.

Miao, J., Li, J., Xin, J., Tu, J., Ge, M., Qi, J., Zhou, X., Zhu, Y., Yang, C., and Lin, Z. Multigate: integrative analysis and regulatory inference in spatial multi-omics data via graph representation learning. *Nature Communications*, 16:9403, 2025.

Nutt, S. L., Heavey, B., Rolink, A. G., and Busslinger, M. Commitment to the B-lymphoid lineage depends on the transcription factor Pax5. *Nature*, 401(6753):556–562, 1999.

Press, O., Smith, N. A., and Lewis, M. Train short, test long: Attention with linear biases enables input length generalization. In *International Conference on Learning Representations (ICLR)*, 2022.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, pp. 8748–8763. PMLR, 2021.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21 (140):1–67, 2020.

Ren, H., Walker, B. L., Cang, Z., and Nie, Q. Identifying multicellular spatiotemporal organization of cells with SpaceFlow. *Nature Communications*, 13(1):4076, 2022.

Rosenberg, A. and Hirschberg, J. V-Measure: A conditional entropy-based external cluster evaluation measure. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pp. 410–420, 2007.

Shaw, P., Uszkoreit, J., and Vaswani, A. Self-attention with relative position representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pp. 464–468, 2018.

Simões-Costa, M. and Bronner, M. E. Establishing neural crest identity: a gene regulatory recipe. *Development*, 142(2):242–257, 2015.

Strehl, A. and Ghosh, J. Cluster ensembles—a knowledge reuse framework for combining multiple partitions. In *Journal of Machine Learning Research*, volume 3, pp. 583–617, 2002.

Tian, Y., Krishnan, D., and Isola, P. Contrastive multiview coding. In *European Conference on Computer Vision (ECCV)*, 2020.

van den Oord, A., Li, Y., and Vinyals, O. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., and Bengio, Y. Graph attention networks. In *International Conference on Learning Representations*, 2018.

Velten, B., Braunger, J. M., Argelaguet, R., Arnol, D., Wirbel, J., Bredikhin, D., Zeller, G., and Stegle, O. Identifying temporal and spatial patterns of variation from multimodal data using MEFISTO. *Nature Methods*, 19 (2):179–186, 2022.

Vinh, N. X., Epps, J., and Bailey, J. Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *Journal of Machine Learning Research*, 11:2837–2854, 2010.

Wen, L., Dai, Q., Liu, J., Zheng, J., Dai, Y., Wang, D., Kang, Z., Wang, J., Xu, Z., and Duan, J. InfMasking: Unleashing synergistic information by contrastive multimodal interactions. *arXiv preprint arXiv:2509.25270*, 2025.

Williams, P. L. and Beer, R. D. Nonnegative decomposition of multivariate information. *arXiv preprint arXiv:1004.2515*, 2010.

Xiao, Y., Meng, D., Huang, X., Liu, Y., Ruan, S., Qiao, Z., and Zheng, X. GROVER: Graph-guided representation of omics and vision with expert regulation for adaptive spatial multi-omics fusion. *arXiv preprint arXiv:2511.11730*, 2025.

Yan, Y., Feng, X., and Luo, X. Spatially aware adjusted Rand index for evaluating spatial transcriptomics clustering. *Biometrics*, 81(3):ujaf127, 2025. doi: 10.1093/biomtc/ujaf127.

Zhang, D., Deng, Y., Kukanja, P., Agirre, E., Bartosovic, M., Dong, M., Ma, C., Ma, S., Su, G., Bao, S., Liu, Y., Xiao, Y., Rosoklija, G. B., Dwork, A. J., Mann, J. J., Leong, K. W., Boldrini, M., Wang, L., Haeussler, M., Raphael, B. J., Kluger, Y., Castelo-Branco, G., and Fan, R. Spatial epigenome–transcriptome co-profiling of mammalian tissues. *Nature*, 616(7955):113–122, 2023.

Zhu, Q., Song, L., Peng, G., Sun, N., Chen, J., Zhang, T., Sheng, N., Tang, W., Qian, C., Qiao, Y., Tang, K., Han, J.-D. J., Li, J., and Jing, N. The transcription factor Pou3f1 promotes neural fate commitment via activation of neural lineage genes and inhibition of external signaling pathways. *eLife*, 3:e02224, 2014.

A. Datasets and Preprocessing

We evaluate MELON on three paired spatial multimodal datasets that span embryonic brain, embryonic whole-embryo, and human lymphoid tissue settings. Together they cover both RNA–ATAC and RNA–protein integration, different tissue architectures, and a range of spatial domain granularities.

Table 6. Dataset statistics used in experiments. For ATAC modalities, the raw peak matrices are reduced to 50 components via TF–IDF + LSI before modeling (Appendix B.1).

Dataset	#Spots	Modality 1	Modality 2	#Domains
Mouse E15.5 Brain	1,949	2,144 genes	47,287 peaks	12
Mouse E13 Embryo	2,187	17,058 genes	24,017 peaks	14
Human Lymph Node	3,484	18,085 genes	31 proteins	10

Mouse E15.5 Brain (MISAR-seq). This dataset profiles an embryonic day 15.5 mouse brain sagittal section using MISAR-seq (Microfluidic Indexing-based Spatial ATAC and RNA-seq), which simultaneously measures chromatin accessibility and gene expression at 50 μm grid resolution ($\sim$15–30 cells per spot) (Jiang et al., 2023). The dataset contains 1,949 spots with 2,144 RNA features and 47,287 ATAC peaks. Ground-truth labels comprise 12 manually annotated neuroanatomical regions (diencephalon, cartilage, hindbrain, dorsal pallium medial/ventral, mesenchyme, midbrain, skull, subpallium, muscle, cerebellar vermis, thalamus), validated against the Allen Mouse Brain Atlas. RNA counts are library-size normalized to a target sum of 10^4 per spot, followed by $\log(1+x)$ transformation. ATAC peaks are transformed with the standard scATAC dimensionality reduction (Cusanovich et al., 2018): term-frequency–inverse-document-frequency (TF–IDF) weighting (peaks as terms, spots as documents) followed by Latent Semantic Indexing (LSI; truncated SVD with 50 components), yielding a 50-dimensional input per spot. Raw RNA and ATAC fastq files are available at NCBI SRA accession SRP491963; processed E15.5 data are released through the official MISAR-seq repository at `https://github.com/gpenglab/MISAR-seq`.

Human Lymph Node (Visium RNA–ADT). This dataset profiles two sequential 5 μm FFPE sections of a human lymph node using the 10x Genomics Visium CytAssist platform (Long et al., 2024). Sections were stained with H&E following the CytAssist FFPE protocol (CG000658, 10x Genomics) including deparaffinization, staining, imaging, and decrosslinking, with imaging on an EVOS M7000 microscope ($\times$20 objective). Spatial gene expression libraries were prepared with the Visium Human Transcriptome Probe Set v2.0 and spatial protein expression libraries with the Human FFPE Immune Profiling Panel (31 antibody targets + 4 isotype controls = 35-plex; antibodies from BioLegend and Abcam, signals normalized to isotype controls). Libraries were sequenced on an Illumina NovaSeq S2 PE50 ($\sim$2,000 M reads per lane) and processed with Space Ranger v2.1.0 against GRCh38 (GENCODE v32 / Ensembl 98). The dataset contains 3,484 spots with 18,085 genes and 31 proteins. Reference labels comprise 10 manually annotated lymphoid compartments (medulla cords, medulla sinuses, cortex, pericapsular adipose tissue, capsule, subcapsular sinus, follicle, medulla vessels, hilum, trabeculae) curated from H&E images in Loupe Browser. RNA counts are library-size normalized to a target sum of 10^4 per spot, followed by $\log(1+x)$ transformation; ADT counts are transformed via centered log-ratio (CLR) normalization. Data are available at `https://zenodo.org/records/10362607`.

Mouse E13 Embryo (spatial-ATAC–RNA-seq). This dataset profiles a whole embryonic day 13 mouse section with paired RNA and ATAC (open chromatin / Tn5 accessibility peaks) using spatial-ATAC–RNA-seq, which co-profiles the epigenome and transcriptome at 50 μm pixel resolution (Zhang et al., 2023). It contains 2,187 spatial locations, 17,058 genes, and 24,017 ATAC peaks. Ground-truth annotations comprise 14 embryonic regions. Preprocessing follows the same recipe as the E15.5 dataset: RNA counts are library-size normalized to a target sum of 10^4 per spot followed by $\log(1+x)$ transformation, and ATAC peaks are transformed via TF–IDF weighting followed by Latent Semantic Indexing (truncated SVD with 50 components). This benchmark complements the E15.5 brain dataset by testing MELON on a larger developmental system with broader anatomical diversity. Data are available at NCBI GEO accession `GSE205055`.

B. Implementation Details

B.1. Preprocessing and Augmentations

MELON operates on modality-specific input matrices after standard preprocessing for each assay. The framework is agnostic to the choice of preprocessing (continuous, normalized, or low-rank-reduced features all work as drop-in inputs to the per-modality encoder f_m); the concrete recipes used for the assays in our experiments (RNA, ATAC, ADT) are described in Appendix A. During training, augmentations are applied before the encoder so that the model never observes a clean batch in the contrastive forward pass.

For each batch, MELON samples two independent augmented multimodal inputs by independently applying three stochastic transformations: Gaussian noise, feature dropout, and aligned cross-spot mixing (the latter shares its permutation and mixing coefficient across all modalities of the same spot so that cross-modal correspondences are preserved). The functional form of each augmentation and the exact hyperparameters used in our experiments (σ, dropout fraction, mixing distribution, and activation probability) are listed in Appendix B.4.

B.2. Spot Tokenization and Token Recovery

Let N denote the batch size and n the number of modalities ($n = 2$ in the bi-modal experiments and $n = 3$ in the tonsil tri-modal extension). After the modality-specific encoders and token-space projection, MELON has one embedding per spot and per modality. With spot tokenization enabled, the batch dimension is reinterpreted as the token dimension, so that each modality contributes a length-N token sequence:

$$T^{(m)} \in \mathbb{R}^{1\times N\times d}, \qquad m \in \{1,\dots,n\}. \tag{14}$$

Concatenating modalities yields a single token sequence

$$T = \mathrm{concat}\big(T^{(1)},\dots,T^{(n)}\big) \in \mathbb{R}^{1\times (nN)\times d}. \tag{15}$$

Thus, token i corresponds to spot i in modality 1, token $N + i$ to the same spot in modality 2, and so on. The transformer therefore attends jointly across all spot–modality pairs in the batch.

After the fusion transformer outputs all active tokens, the resulting tensor is reshaped back into modality-by-spot blocks and pooled across modalities to recover one fused embedding per spot:

$$\tilde{T} \in \mathbb{R}^{1\times (nN)\times d} \to \mathbb{R}^{n\times N\times d}, \qquad z_i = \frac{1}{n}\sum_{m=1}^{n} \tilde{t}_i^{(m)}. \tag{16}$$

This tokenization scheme is the default because it allows the transformer to model both cross-modal and cross-spot interactions within a single attention layer.

B.3. Spatial Bias Tiling and Modality Masking

NeighborhoodGAT produces a spot-level spatial bias matrix $B \in \mathbb{R}^{N\times N}$, where N is the batch size as defined in Appendix B.2. Under spot tokenization, however, the transformer sees nN tokens rather than N spots. To align the spatial prior with the token layout, we tile the spot-level bias across modality blocks into the matrix B_{spatial} used in the fusion-transformer attention (§3.4). For n modalities, B_{spatial} is an $n \times n$ block matrix in which every block equals B; we write the $n = 2$ case explicitly here for concreteness:

$$B_{\text{spatial}} = \begin{bmatrix} B & B \\ B & B \end{bmatrix} \in \mathbb{R}^{2N\times 2N}. \tag{17}$$

This tiling enforces that the spatial relationship between spot i and spot j is modality-invariant: the same bias is used whether attention connects within-modality token pairs (modality m to modality m) or cross-modal token pairs (modality m to modality m') for those two spots. For n modalities the construction generalizes by replacing the 2×2 block matrix above with an $n \times n$ block matrix whose every block is B.

MELON then evaluates the fusion module under $n + 1$ modality masks: n single-modality passes and one joint multimodal pass. When only modality m is active, the token sequence is physically subsetted and the tiled bias is sliced to the diagonal

block corresponding to m, which by construction equals the underlying spot-level bias B. For the multimodal pass, the full tiled matrix is used. This design guarantees that single-modality and multimodal forward passes share the same underlying spot-level spatial bias.

B.4. Hyperparameters

MELON is implemented in PyTorch with PyTorch Lightning. Defaults are listed in Table 7 and broadly apply across the benchmarked datasets, with light per-dataset adjustment of two parameters: the augmentation strength is slightly tuned to the noise characteristics of each modality, and the batch size is scaled to the number of spots in the dataset. Sensitivity to the two main spatial/augmentation hyperparameters is examined in §4.4. Augmentation functional forms are described in §3.5; the values in the table refer to the parameters of those functions. After training, the joint embedding is clustered with Leiden clustering; the implementation also supports Louvain, Walktrap, spectral, and GMM for downstream exploration.

Table 7. MELON default hyperparameters used across the three benchmarked datasets. Augmentation strength is lightly adjusted to each dataset's noise characteristics, and batch size is scaled to the number of spots.

Group	Hyperparameter	Value
Encoder	Per-modality encoder	2-layer MLP, $d_m \to 128 \to 128$, ReLU
	Embedding dimension	128
	Projection head	3-layer MLP, $128 \to 256 \to 256 \to 128$, SyncBatchNorm
	Output normalisation	L2
Fusion transformer	Layers	1
	Attention	Grouped Query Attention, 8 heads
	MLP expansion / activation	$4\times$ / QuickGELU
	Spatial bias	additive (NeighborhoodGAT, see below)
NeighborhoodGAT	Layers / heads	1 / 8
	Hidden dim per head	64
	Attention dropout	0.1
	LeakyReLU slope	0.2
	Spatial k-NN graph	$k = 16$, self-loops included
Optimisation	Optimiser	Adam, $(\beta_1, \beta_2) = (0.9, 0.999)$
	Learning rate	1×10^{-3}
	Weight decay	1×10^{-5}
	Schedule	none
	Epochs	100
	Batch size	2,048
Loss	Temperature τ	0.1
	Per-mask weighting	equal across $n{+}1$ masks (Eq. 13)
Augmentations	Activation probability	$p = 0.8$ per augmentation
	Gaussian noise	$\sigma = 0.05$
	Feature dropout	fraction 0.2
	Neighbour mix	$\lambda \sim \mathrm{Beta}(0.3, 0.3)$, shared π, λ across modalities
Reproducibility	Random seed	42 (PyTorch, NumPy, Python `random`)
	Hardware	CPU (Apple Silicon / MPS) or single CUDA GPU
	Framework	PyTorch + PyTorch Lightning

B.5. Compute Usage

All MELON runs in this paper used a single CUDA GPU with the defaults of Table 7 (100 epochs, batch 2,048). A single MELON training on the Mouse E15.5 brain (1,949 spots, two modalities) takes about 7 minutes wall-clock (≈ 400 s training plus ≈ 18 s downstream Leiden clustering). The largest single MELON run is the tri-modal tonsil training (§D.7; 4,194 spots, three modalities), which took 36.7 minutes wall-clock with 16.4 GiB resident memory and 28.7 GiB peak memory footprint.

C. Simulation Study Details

C.1. Setup and Synergy Label

We simulate N=1,936 spots on a 44×44 spatial grid with two paired modalities M_1 and M_2 designed to mimic RNA and ATAC respectively. The simulator first generates structured Gaussian latent vectors of dimension 600 per modality, then emits sparse molecular observations: M_1 as overdispersed counts (negative-binomial, mimicking RNA) and M_2 as binary peak accessibility (Bernoulli, mimicking ATAC). Both matrices are $\log(1+x)$-transformed before model fitting; the dense latent matrices are not used directly as benchmark inputs (emission details in Appendix C.3). Each spot carries four ground-truth latent variables: a coarse spatial region $R \in \{1, \ldots, 5\}$, an M_1-side attribute $U_i \in \{1, \ldots, 10\}$, an M_2-side attribute $C_i \in \{1, \ldots, 10\}$, and a binary cross-modal matching label $Y_i \in \{0, 1\}$. A fixed random permutation $\pi : \{1, \ldots, 10\} \to \{1, \ldots, 10\}$ binds the two modality-specific attributes, and

$$Y_i \sim \text{Bernoulli}(p_i), \qquad p_i = \begin{cases} 0.90 & \text{if } C_i = \pi(U_i), \\ 0.01 & \text{otherwise.} \end{cases} \tag{18}$$

By construction, Y is a nonlinear matching function of U and C: a method can recover Y from its embedding only if the embedding preserves enough information about both modality-specific attributes for a downstream nonlinear classifier to infer the matching rule.

Three independent dataset seeds. We benchmark across three independently generated datasets, with seeds 123, 124, and 125. Each seed re-samples region centers, attribute distributions, the permutation π, and all noise. The positive class prevalence of Y is intentionally low and varies modestly across seeds (Table 8).

Table 8. Positive prevalence of the cross-modal matching label Y across the three simulation seeds.

Seed	Spots	Positive Y spots	Positive rate
123	1,936	170	0.0878
124	1,936	217	0.1121
125	1,936	186	0.0961

C.2. Feature Block Architecture

For each spot, the M_1 (RNA-mimicking) and M_2 (ATAC-mimicking) latent vectors are constructed by concatenating five structured blocks (Table 9). The signal scales are asymmetric across modalities by design: the M_2 private attribute C is given a stronger amplitude (1.2) than the M_1 private attribute U (0.6), reflecting the typically sharper region-specific accessibility patterns observed in real spatial ATAC data. The mixed blocks are the key cross-modal carriers: each combines a region signal, a weak (down-weighted) modality-specific signal, and a pair-code vector that is shared across modalities only when $C_i = \pi(U_i)$.

Table 9. Feature block architecture for each modality. Each modality has 5 blocks summing to 600 latent dimensions, with per-block per-spot Gaussian noise standard deviation σ_{noise}. Class-prototype centroids are sampled from $\mathcal{N}(\mathbf{0}, \sigma^2_{\text{cent}}\mathbf{I})$ once per class; the signal scale α multiplies the centroid before adding noise.

Modality	Block	Dim	α	σ_{noise}	Encodes
M_1 (RNA)	Shared main	150	1.0	1.0	$A^{(1)}$ (region-derived)
	Unique task	150	0.6	1.0	U (M_1-specific)
	Nuisance	100	1.0	0.7	8-class patch attribute (M_1)
	Mixed	100	—	0.7	R (1.0) + weak U (0.2) + pair-code (s=30.0)
	Noise	100	—	1.0	$\mathcal{N}(0, 1)$ i.i.d.
M_2 (ATAC)	Shared main	150	1.0	1.0	$A^{(1)}$ (region-derived)
	Nuisance	150	1.0	0.7	8-class patch attribute (M_2)
	Synergy attr	100	1.2	0.5	C (M_2-specific)
	Mixed	100	—	0.7	R (1.0) + weak C (0.2) + pair-code (s=30.0)
	Noise	100	—	1.0	$\mathcal{N}(0, 1)$ i.i.d.

Feature generation. Each non-mixed signal block is generated as $\alpha\,\boldsymbol{\mu}_k + \boldsymbol{\epsilon}$, where $\boldsymbol{\mu}_k \sim \mathcal{N}(\mathbf{0}, \sigma^2_{\text{cent}}\mathbf{I})$ is a latent-class centroid sampled once per class and $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2_{\text{noise}}\mathbf{I})$ is independent per-spot Gaussian noise. The mixed block for each modality combines region structure, a down-weighted modality-specific signal, and a cross-modal pair code:

$$\mathbf{x}_i^{(M_1,\text{mixed})} = 1.0\,\boldsymbol{\mu}^R_{R_i} + 0.2\,\boldsymbol{\mu}^U_{U_i} + s\cdot\boldsymbol{\mu}^{\text{pair}}_{U_i} + \boldsymbol{\epsilon}_i, \tag{19}$$

$$\mathbf{x}_i^{(M_2,\text{mixed})} = 1.0\,\boldsymbol{\mu}^{R'}_{R_i} + 0.2\,\boldsymbol{\mu}^C_{C_i} + s\cdot\boldsymbol{\mu}^{\text{pair}}_{\pi^{-1}(C_i)} + \boldsymbol{\epsilon}'_i, \tag{20}$$

where $s = 30.0$ is the pair-code scale used for the headline comparison (§4.1), $\boldsymbol{\mu}^{\text{pair}}_k \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ is sampled once per attribute code k, and $\boldsymbol{\epsilon}_i, \boldsymbol{\epsilon}'_i \sim \mathcal{N}(\mathbf{0}, 0.7^2\mathbf{I})$. When $C_i = \pi(U_i)$, both modalities carry the *same* pair-code vector $\boldsymbol{\mu}^{\text{pair}}_{U_i}$; otherwise, the two modalities carry discordant pair-codes. Combined with the sparse NB / Bernoulli emission step (Appendix C.3), this is the intended feature-level carrier of the cross-modal matching signal.

Latent variable sampling. Region labels R_i are assigned by Voronoi tessellation of K_R=5 seed points sampled uniformly on the grid, followed by T=3 rounds of spatial majority smoothing (8-connected neighborhood) and stochastic boundary flipping with probability p_{flip}=0.05. The shared region-derived attribute $A^{(1)}$ is a deterministic function of R followed by per-modality local label flips with probability 0.05 in each modality, producing a strong but imperfect shared spatial signal. The nuisance attributes have 8 classes per modality with local flip probability 0.08. The modality-specific attributes U_i and C_i are sampled per region from a region-biased categorical distribution. Within region r, the sampling probability over K=10 attribute codes is

$$\mathbf{p}_r \;=\; (1-\beta)\,\mathbf{u} \;+\; \beta\,\mathbf{w}_r, \qquad \mathbf{w}_r \;\sim\; \text{Dir}(\mathbf{1}_K), \qquad \mathbf{u} \;=\; \tfrac{1}{K}\mathbf{1}_K, \tag{21}$$

where $\beta = 0.12$ is the region-bias weight (identical for U and C), $\mathbf{w}_r$ is a region-specific Dirichlet draw, and $\mathbf{u}$ is the uniform prior over the K attribute codes. This produces fine-grained within-region heterogeneity that is only weakly correlated with spatial location, so that neither modality alone nor spatial coordinates are sufficient to recover the matching label Y defined in Eq. (18).

C.3. Count and Accessibility Emission

The dense latent matrices in Eqs. (19)–(20) are not used directly as benchmark inputs. Instead, the simulator emits sparse, count-like paired observations.

M_1 emission (negative-binomial counts, mimicking RNA). Each M_1 feature g (interpretable as a gene) has a baseline log expression $\log \mu_g \sim \mathcal{N}(-1.5, 1.5^2)$, and each spot i has a library-size log factor $\ell_i \sim \mathcal{N}(0, 0.4^2)$. The latent M_1 signal $\mathbf{x}^{(M_1)}_{i,g}$ is multiplied by 0.6 and added on the log-rate scale, $\log \lambda_{i,g} = 0.6\,\mathbf{x}^{(M_1)}_{i,g} + \log \mu_g + \ell_i$. Counts are sampled from a Gamma–Poisson representation of a negative-binomial with dispersion $\theta = 10.0$.

M_2 emission (Bernoulli peak accessibility, mimicking ATAC). Each M_2 feature p (interpretable as a peak) has a baseline logit $\eta_p \sim \mathcal{N}(-2.5, 1.0^2)$, and each spot i has a logit-scale accessibility factor $a_i \sim \mathcal{N}(0, 0.4^2)$. The latent M_2 signal is multiplied by 0.8 and added on the logit scale, $\text{logit}\, q_{i,p} = 0.8\,\mathbf{x}^{(M_2)}_{i,p} + \eta_p + a_i$. Peak accessibility is sampled as Bernoulli with probability $q_{i,p}$.

For benchmarking, both the M_1 count matrix and the M_2 accessibility matrix are transformed with $\log(1+x)$ before model fitting. This yields sparse, count-like paired inputs while retaining the known latent ground truth.

C.4. Simulation Hyperparameters

Table 10 lists the complete set of hyperparameters used to instantiate the controlled simulation in §4.1, covering the spatial grid, latent attribute structure, synergy generation, per-modality signal and noise scales, and the three independent dataset seeds. All values are held fixed across seeds; only the random-number generator state varies. The pair-code multiplier s=30 corresponds to the headline synergy-isolating operating point reported in the main text.

C.5. Spatial Structure of Simulated Data

Figure 4 visualizes the spatial distribution of all latent variables on the 44×44 grid. Region labels R form coarse spatial domains from Voronoi tessellation. The modality-specific attributes U and C exhibit finer within-region heterogeneity due

Table 10. Complete simulation hyperparameters. Three independent dataset seeds are evaluated: 123, 124, 125.

Category	Parameter	Value
Grid	Size	44×44 (N=1,936)
Latent	Regions K_R	5
	M_1-specific levels K_U	10
	M_2-specific levels K_C	10
	Nuisance levels (per modality)	8
	Region bias β (for U and C)	0.12
Synergy	Match prob p_{match}	0.90
	Non-match prob p_{non}	0.01
	Pair-code prototype scale	1.0
	Pair-code multiplier s (headline)	30.0
Signal scales	M_1 shared main α	1.0
	M_1 private U α	0.6
	M_2 shared main α	1.0
	M_2 private C α	1.2
	Nuisance α (each modality)	1.0
	Mixed: region component	1.0
	Mixed: weak U/C component	0.2
Noise	σ_{shared}, $\sigma_{\text{nuis}}^{M_1\text{-}U}$	1.0
	$\sigma_{M_2\text{-}C}$ (synergy attr.)	0.5
	σ_{nuisance}, σ_{mixed}	0.7
	σ_{noise} (pure noise blocks)	1.0
Spatial	Smoothing iterations T	3
	Shared attribute flip prob	0.05
	Attribute flip prob (U, C)	0.05
	Nuisance flip prob	0.08
M_1 emission (RNA-like)	Feature baseline log μ_g	$\mathcal{N}(-1.5, 1.5^2)$
	Library-size log factor	$\mathcal{N}(0, 0.4^2)$
	Latent multiplier on log-rate	0.6
	NB dispersion θ	10.0
M_2 emission (ATAC-like)	Feature baseline logit η_p	$\mathcal{N}(-2.5, 1.0^2)$
	Per-spot logit-scale factor	$\mathcal{N}(0, 0.4^2)$
	Latent multiplier on logit	0.8
	Random seeds	$\{123, 124, 125\}$

to weak region bias ($\beta = 0.12$). The synergy label Y shows no obvious spatial pattern, confirming that it cannot be trivially predicted from location.

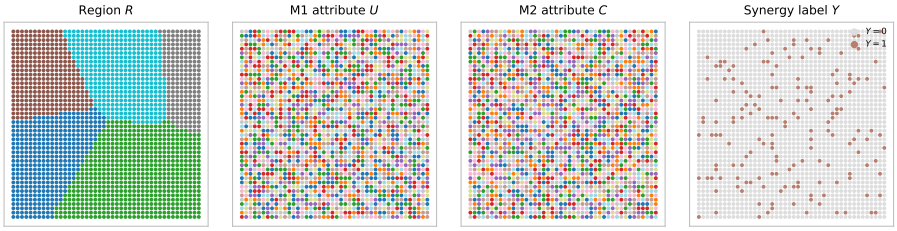


Figure 4. Spatial distribution of latent variables on the 44×44 simulation grid. Region label R forms coarse Voronoi domains; the modality-specific attributes U and C are fine-grained with weak region bias; the synergy label Y shows no apparent spatial structure (orange = Y=1).

C.6. Cross-Target Probe Comparison

Table 1 reports balanced accuracy across all methods and the four targets R, U, C, Y. The cross-target view highlights three distinct strengths. (i) Spatial-domain R recovery is led by SpatialGlue and Seurat WNN with MELON third; non-spatial linear factor methods (MultiVI, MEFISTO, MOFA+) collapse on R. (ii) Modality-specific attribute recovery: MELON and Seurat WNN saturate both U and C; MOFA+ and MEFISTO retain both at high accuracy; SpatialGlue retains both at ≥ 0.93; MISO retains C at 0.97 with U at 0.88; MultiVI retains U but collapses on C. (iii) Cross-modal matching Y at pair-code scale $= 30$ separates MELON (0.718) cleanly from every external baseline (≤ 0.512), even though several baselines preserve both U and C. The Y endpoint should therefore be interpreted as a controlled cross-modal information-recovery assay: at this synergy-isolating operating point, preserving U and C alone is insufficient: the embedding must additionally retain the cross-modal binding between them.

C.7. Full Ablation and Paired Deltas

Table 11. Full MELON ablation across the three simulation seeds (pair-code scale = 30; mean ± std). Linear SVM probes for R, U, C; nonlinear sklearn MLP probe for Y (with macro-F1 and AUROC reported alongside BalAcc). Single-knockouts isolate spatial-bias vs. augmentation contributions; the double knockout (no spatial bias and no augmentation) and unimodal controls (RNA only, ATAC only) act as sanity checks: removing both design axes or any modality collapses Y recovery toward chance, while U/C behave as expected modality-specific controls (RNA only retains U and loses C; ATAC only retains C and loses U).

Variant	R BalAcc	U BalAcc	C BalAcc	Y BalAcc	Y Macro-F1	Y AUROC
MELON (full)	$\mathbf{0.930 \pm 0.019}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{0.718 \pm 0.044}$	$\mathbf{0.768 \pm 0.055}$	$\mathbf{0.881 \pm 0.081}$
− spatial bias	0.870 ± 0.045	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	0.592 ± 0.033	0.620 ± 0.042	0.827 ± 0.023
− augmentation	0.912 ± 0.042	$\mathbf{1.000 \pm 0.000}$	0.997 ± 0.001	0.516 ± 0.018	0.503 ± 0.031	0.619 ± 0.080
− spatial bias, − aug	0.892 ± 0.035	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	0.509 ± 0.008	0.494 ± 0.012	0.619 ± 0.044
RNA only	0.850 ± 0.049	$\mathbf{1.000 \pm 0.000}$	0.103 ± 0.005	0.500 ± 0.000	0.474 ± 0.003	0.485 ± 0.030
ATAC only	0.716 ± 0.045	0.097 ± 0.012	$\mathbf{1.000 \pm 0.000}$	0.500 ± 0.001	0.475 ± 0.001	0.504 ± 0.025

Paired ablation deltas. Computed per-seed as Y score for the full model minus the ablated variant, then averaged across the three seeds (Table 12). Removing augmentation accounts for 0.203 BalAcc, removing the spatial bias for 0.126, and the two effects together for 0.210 (close to additive). Replacing both modalities with either unimodal input drives the largest paired drop (≈ 0.219), confirming that the cross-modal Y signal cannot be recovered from a single modality regardless of model design.

Table 12. Paired ablation deltas on the sklearn MLP Y probe (full MELON minus ablated variant; mean ± std across seeds 123, 124, 125). Larger values indicate a larger drop when the corresponding component or modality is removed.

Ablation	Δ Y BalAcc	Δ Y Macro-F1	Δ Y AUROC
− spatial bias	0.126 ± 0.071	0.148 ± 0.091	0.054 ± 0.101
− augmentation	0.203 ± 0.042	0.265 ± 0.050	0.263 ± 0.039
− spatial bias & augmentation	0.210 ± 0.048	0.274 ± 0.061	0.263 ± 0.087
RNA-only input	0.218 ± 0.044	0.294 ± 0.052	0.397 ± 0.068
ATAC-only input	0.219 ± 0.045	0.293 ± 0.054	0.378 ± 0.075

D. Additional Real-Data Experiments

D.1. Compared Baselines

We compare MELON against eight established multi-omics integration methods. TotalVI and MultiVI are dispatched per dataset based on modality combination (TotalVI for RNA–protein, MultiVI for RNA–ATAC), so any individual dataset sees seven external baselines. For each baseline we report the joint embedding produced by the published implementation under default settings, then evaluate it under the same downstream clustering and metric protocol described in Appendix D.2.

- **Seurat v4** (Hao et al., 2021): performs weighted nearest-neighbor (WNN) integration of multiple modalities via anchor-based canonical correlation analysis, followed by graph-based Louvain clustering. Spatial coordinates are not used.

- **MOFA+** (Argelaguet et al., 2020): multi-omics factor analysis with group and modality structure. Learns linear shared and private latent factors via variational inference. No spatial information.
- **TotalVI** (Gayoso et al., 2021): a deep generative variational autoencoder (VAE) for joint modeling of RNA and protein (CITE-seq) counts. Used for the RNA–ADT human lymph node dataset. Assumes i.i.d. cells and does not exploit spatial neighborhoods.
- **MultiVI** (Ashuach et al., 2023): a sister VAE in the `scvi-tools` family that jointly models paired RNA (NB) and ATAC (Bernoulli peak) counts with batch and modality embeddings. Used for the RNA–ATAC Mouse E15.5 brain and Mouse E13 embryo datasets in place of TotalVI; also assumes i.i.d. cells.
- **SpatialGlue** (Long et al., 2024): a graph neural network with a dual-attention mechanism that builds per-modality spatial neighbor graphs and fuses representations via cross-attention within a VAE framework. Explicitly uses spatial coordinates.
- **MEFISTO** (Velten et al., 2022): extends MOFA+ with Gaussian-process priors over spatial (or temporal) covariates to encourage smooth latent factors, partially capturing spatial structure but at the cost of over-smoothing domain boundaries.
- **MISO** (Coleman et al., 2025): a deep multimodal integration framework designed to combine spatial molecular measurements with H&E morphology, evaluated here under the same per-dataset modality configuration as the other baselines.
- **MultiGate** (Miao et al., 2025): a two-level graph attention autoencoder that jointly models spatial context and cross-modality feature relationships through stacked GAT layers.

For downstream clustering of all learned embeddings (MELON and baselines) we apply Leiden clustering. All methods are evaluated at matching numbers of clusters k.

D.2. Evaluation Metrics

We evaluate spatial domain identification using eleven clustering quality metrics that together capture cluster purity, mutual dependence, pairwise agreement, internal compactness, and spatial coherence (eight reported in the main-text Table 3; the full panel appears in Table 13). Throughout, $Y \in [C]^N$ denotes the ground-truth label of N spots, $\hat{Y} \in [K]^N$ the predicted cluster assignment, $H(\cdot)$ the Shannon entropy, and $I(Y;\hat{Y}) = H(Y) + H(\hat{Y}) - H(Y,\hat{Y})$ the mutual information. For pair-counting metrics let a (d) be the number of spot pairs assigned to the same (different) cluster in both Y and $\hat{Y}$, and b (c) the number disagreeing in Y vs. $\hat{Y}$.

Pair-counting metrics. The **Rand Index (RI)** is the fraction of consistent pairs $\mathrm{RI} = (a+d)/\binom{N}{2}$. The **Adjusted Rand Index (ARI)** (Hubert & Arabie, 1985) corrects for chance:

$$\mathrm{ARI} \;=\; \frac{\mathrm{RI} - \mathbb{E}[\mathrm{RI}]}{\max(\mathrm{RI}) - \mathbb{E}[\mathrm{RI}]}. \tag{22}$$

The **Fowlkes–Mallows index (FMI)** (Fowlkes & Mallows, 1983) is the geometric mean of pairwise precision and recall,

$$\mathrm{FMI} \;=\; \sqrt{\frac{a}{a+b} \cdot \frac{a}{a+c}}. \tag{23}$$

The **Jaccard index** on co-clustered pairs is

$$\mathrm{Jaccard} \;=\; \frac{a}{a+b+c}. \tag{24}$$

Information-theoretic metrics. The **Mutual Information (MI)** between Y and $\hat{Y}$ is $\mathrm{MI} = I(Y;\hat{Y})$. The **Normalized Mutual Information (NMI)** (Strehl & Ghosh, 2002) divides by the arithmetic mean of entropies:

$$\mathrm{NMI} \;=\; \frac{2\, I(Y;\hat{Y})}{H(Y) + H(\hat{Y})}. \tag{25}$$

The **Adjusted Mutual Information (AMI)** (Vinh et al., 2010) subtracts the expectation under random labelings:

$$\mathrm{AMI} \;=\; \frac{I(Y;\hat{Y}) - \mathbb{E}[I(Y;\hat{Y})]}{\max\big(H(Y), H(\hat{Y})\big) - \mathbb{E}[I(Y;\hat{Y})]}. \tag{26}$$

Homogeneity (Rosenberg & Hirschberg, 2007) measures whether each predicted cluster contains spots of a single ground-truth class:

$$\mathrm{Homog.} \;=\; 1 - \frac{H(Y \mid \hat{Y})}{H(Y)}, \qquad \mathrm{Comp.} \;=\; 1 - \frac{H(\hat{Y} \mid Y)}{H(\hat{Y})}, \tag{27}$$

where Completeness (Comp.) is its dual. The **V-measure** (Rosenberg & Hirschberg, 2007) is their harmonic mean,

$$\text{V-Meas.} \;=\; \frac{2 \cdot \mathrm{Homog.} \cdot \mathrm{Comp.}}{\mathrm{Homog.} + \mathrm{Comp.}}. \tag{28}$$

Hard-assignment metric. **Purity** assigns each predicted cluster to its dominant ground-truth class and reports the fraction of spots correctly covered:

$$\mathrm{Purity} \;=\; \frac{1}{N} \sum_{k=1}^{K} \max_{c \in [C]} \big|\{i : \hat{Y}_i = k \,\wedge\, Y_i = c\}\big|. \tag{29}$$

Internal-compactness metric. The **Calinski–Harabasz index (CHI)** (Caliński & Harabasz, 1974) measures between-cluster separation versus within-cluster dispersion in the embedding space, using only the embeddings $\{z_i\}_{i=1}^{N}$ and the predicted assignment $\hat{Y}$:

$$\mathrm{CHI} \;=\; \frac{\mathrm{tr}(B_K)/(K-1)}{\mathrm{tr}(W_K)/(N-K)}, \tag{30}$$

where $B_K = \sum_k n_k(\mu_k - \bar{\mu})(\mu_k - \bar{\mu})^\top$ is the between-cluster scatter and $W_K = \sum_k \sum_{i:\hat{Y}_i=k}(z_i - \mu_k)(z_i - \mu_k)^\top$ the within-cluster scatter, with μ_k the centroid of predicted cluster k and n_k its size. Because CHI scales with the embedding magnitude, its values are not directly comparable across methods that use different embedding spaces.

Spatial metrics. **Spatial Rand (SpRI)** and **Spatial Adjusted Rand (SpARI)** (Yan et al., 2025) extend RI and ARI by weighting each pair (i, j) by a spatial-proximity kernel $\phi(d_{ij})$ on the 2D coordinates, so that misassignments between spatially distant pairs are penalized more than those between neighboring pairs:

$$\mathrm{SpRI} \;=\; \frac{\sum_{i<j} \phi(d_{ij})\, \mathbb{1}[\mathbb{1}\{Y_i = Y_j\} = \mathbb{1}\{\hat{Y}_i = \hat{Y}_j\}\,]}{\sum_{i<j} \phi(d_{ij})}, \qquad \mathrm{SpARI} \;=\; \frac{\mathrm{SpRI} - \mathbb{E}[\mathrm{SpRI}]}{\max(\mathrm{SpRI}) - \mathbb{E}[\mathrm{SpRI}]}. \tag{31}$$

SpARI is the only metric in our suite that explicitly rewards spatially contiguous predictions; we use the implementation of Yan et al. (2025) with its default Gaussian kernel.

D.3. Additional Dataset Results

This subsection shows the full per-method spatial domain panels for the human lymph node (Fig. 5) and mouse E13 embryo (Fig. 6) datasets, complementing the Mouse E15.5 brain results in the main text (Fig. 2). Each figure reports the H&E reference, ground-truth annotation, and per-panel ARI for the eight benchmarked methods.

D.4. Sensitivity Analysis

We sweep MELON's two spatial-bias and augmentation hyperparameters one at a time on the Mouse E15.5 brain, retraining each value while holding all other settings at the defaults of Table 7; downstream Leiden clustering is applied identically to every variant. The corresponding plots appear as Fig. 3 in §4.4: MELON is robust to the spatial-graph size k_{nn} with graceful degradation outside the local window, while augmentation is the dominant lever and saturates once enabled, consistent with the simulation Y-recovery saturation (Table 2).

Table 13. Full 11-metric panel across the three benchmarked datasets, aggregated across cluster numbers k (median $\pm$ IQR; same k ranges as Table 3). Best median per (dataset, metric) in **bold**. Columns added relative to the main-text table: **V-Meas.** (Rosenberg & Hirschberg, 2007) (V-measure), **SpRI** (spatial Rand index, the un-adjusted variant of SpARI), and **CHI** (Calinski–Harabasz index, an internal cluster-compactness score).

Dataset	Method	ARI	NMI	AMI	Homog.	V-Meas.	SpRI	SpARI	FMI	Jaccard	CHI	Purity
Mouse E15.5 Brain	**MELON**	**0.480 ± 0.022**	**0.569 ± 0.005**	**0.561 ± 0.006**	**0.532 ± 0.028**	**0.569 ± 0.005**	**0.933 ± 0.003**	**0.532 ± 0.030**	**0.567 ± 0.016**	**0.384 ± 0.016**	**8927 ± 2007**	**0.678 ± 0.032**
	MISO	0.216 ± 0.011	0.372 ± 0.027	0.364 ± 0.026	0.355 ± 0.046	0.372 ± 0.027	0.888 ± 0.004	0.176 ± 0.014	0.324 ± 0.007	0.193 ± 0.003	156 ± 28	0.485 ± 0.034
	Seurat WNN	0.291 ± 0.017	0.483 ± 0.007	0.475 ± 0.010	0.491 ± 0.041	0.483 ± 0.007	0.905 ± 0.003	0.224 ± 0.026	0.376 ± 0.016	0.226 ± 0.008	3056 ± 373	0.560 ± 0.041
	SpatialGlue	0.285 ± 0.024	0.515 ± 0.019	0.508 ± 0.016	0.524 ± 0.063	0.515 ± 0.019	0.904 ± 0.006	0.249 ± 0.021	0.381 ± 0.026	0.235 ± 0.024	212 ± 17	0.625 ± 0.055
	MultiGate	0.344 ± 0.025	0.515 ± 0.006	0.508 ± 0.004	0.516 ± 0.050	0.515 ± 0.006	0.912 ± 0.001	0.321 ± 0.110	0.425 ± 0.061	0.270 ± 0.038	420 ± 60	0.596 ± 0.027
	MultiVI	0.086 ± 0.019	0.212 ± 0.059	0.200 ± 0.056	0.199 ± 0.068	0.212 ± 0.059	0.860 ± 0.011	0.001 ± 0.014	0.211 ± 0.001	0.117 ± 0.002	409 ± 101	0.372 ± 0.051
	MOFA+	0.130 ± 0.018	0.300 ± 0.017	0.289 ± 0.014	0.273 ± 0.033	0.300 ± 0.017	0.869 ± 0.006	0.085 ± 0.021	0.265 ± 0.024	0.152 ± 0.013	631 ± 13	0.436 ± 0.021
	MEFISTO	0.151 ± 0.009	0.301 ± 0.018	0.292 ± 0.017	0.281 ± 0.031	0.301 ± 0.018	0.873 ± 0.004	0.087 ± 0.018	0.269 ± 0.019	0.155 ± 0.009	530 ± 29	0.443 ± 0.021
Human Lymph Node	**MELON**	**0.298 ± 0.028**	**0.379 ± 0.002**	**0.374 ± 0.001**	0.401 ± 0.033	**0.379 ± 0.002**	**0.858 ± 0.004**	**0.307 ± 0.038**	0.454 ± 0.018	**0.293 ± 0.015**	**7265 ± 576**	**0.643 ± 0.009**
	MISO	0.241 ± 0.046	0.305 ± 0.035	0.301 ± 0.032	0.263 ± 0.071	0.305 ± 0.035	0.839 ± 0.012	0.281 ± 0.017	**0.455 ± 0.003**	0.285 ± 0.010	241 ± 47	0.576 ± 0.065
	Seurat WNN	0.182 ± 0.000	0.260 ± 0.006	0.256 ± 0.005	0.236 ± 0.017	0.260 ± 0.006	0.824 ± 0.000	0.185 ± 0.014	0.389 ± 0.018	0.240 ± 0.012	597 ± 126	0.529 ± 0.004
	SpatialGlue	0.227 ± 0.049	0.375 ± 0.023	0.371 ± 0.024	**0.423 ± 0.023**	0.375 ± 0.023	0.837 ± 0.012	0.195 ± 0.084	0.371 ± 0.065	0.220 ± 0.061	375 ± 51	0.624 ± 0.008
	MultiGate	0.148 ± 0.011	0.203 ± 0.013	0.199 ± 0.012	0.221 ± 0.038	0.203 ± 0.013	0.823 ± 0.002	0.147 ± 0.017	0.325 ± 0.022	0.192 ± 0.019	1289 ± 133	0.484 ± 0.001
	TotalVI	0.181 ± 0.006	0.258 ± 0.001	0.253 ± 0.000	0.232 ± 0.001	0.258 ± 0.001	0.823 ± 0.001	0.187 ± 0.011	0.393 ± 0.010	0.242 ± 0.007	418 ± 105	0.525 ± 0.004
	MOFA+	0.212 ± 0.032	0.321 ± 0.033	0.318 ± 0.031	0.306 ± 0.103	0.321 ± 0.033	0.834 ± 0.009	0.229 ± 0.025	0.415 ± 0.036	0.261 ± 0.020	1042 ± 81	0.589 ± 0.073
	MEFISTO	0.200 ± 0.011	0.307 ± 0.007	0.302 ± 0.005	0.310 ± 0.045	0.307 ± 0.007	0.832 ± 0.002	0.221 ± 0.050	0.400 ± 0.044	0.249 ± 0.036	854 ± 82	0.594 ± 0.046
Mouse E13 Embryo	**MELON**	**0.385 ± 0.007**	**0.528 ± 0.010**	**0.519 ± 0.007**	**0.546 ± 0.039**	**0.528 ± 0.010**	**0.925 ± 0.001**	**0.362 ± 0.026**	**0.452 ± 0.022**	**0.290 ± 0.023**	191 ± 40	**0.655 ± 0.008**
	MISO	0.227 ± 0.014	0.441 ± 0.031	0.431 ± 0.028	0.417 ± 0.046	0.441 ± 0.031	0.899 ± 0.004	0.251 ± 0.005	0.337 ± 0.002	0.197 ± 0.002	204 ± 29	0.521 ± 0.053
	Seurat WNN	0.266 ± 0.043	0.454 ± 0.017	0.443 ± 0.020	0.487 ± 0.017	0.454 ± 0.017	0.910 ± 0.002	0.223 ± 0.073	0.341 ± 0.047	0.201 ± 0.041	**5180 ± 318**	0.571 ± 0.005
	SpatialGlue	0.280 ± 0.014	0.460 ± 0.022	0.451 ± 0.018	0.472 ± 0.063	0.460 ± 0.022	0.910 ± 0.001	0.238 ± 0.043	0.355 ± 0.018	0.215 ± 0.016	306 ± 35	0.600 ± 0.061
	MultiGate	0.170 ± 0.016	0.380 ± 0.008	0.369 ± 0.007	0.392 ± 0.031	0.380 ± 0.008	0.899 ± 0.003	0.168 ± 0.056	0.263 ± 0.036	0.151 ± 0.025	1342 ± 133	0.438 ± 0.021
	MultiVI	0.092 ± 0.003	0.224 ± 0.033	0.212 ± 0.029	0.164 ± 0.030	0.224 ± 0.033	0.819 ± 0.002	0.140 ± 0.005	0.321 ± 0.000	0.149 ± 0.000	195 ± 22	0.307 ± 0.006
	MOFA+	0.201 ± 0.012	0.389 ± 0.012	0.377 ± 0.009	0.383 ± 0.040	0.389 ± 0.012	0.897 ± 0.008	0.199 ± 0.033	0.295 ± 0.017	0.173 ± 0.010	1203 ± 49	0.491 ± 0.030
	MEFISTO	0.207 ± 0.015	0.391 ± 0.003	0.381 ± 0.007	0.392 ± 0.018	0.391 ± 0.003	0.900 ± 0.003	0.189 ± 0.050	0.302 ± 0.032	0.178 ± 0.021	767 ± 29	0.496 ± 0.017

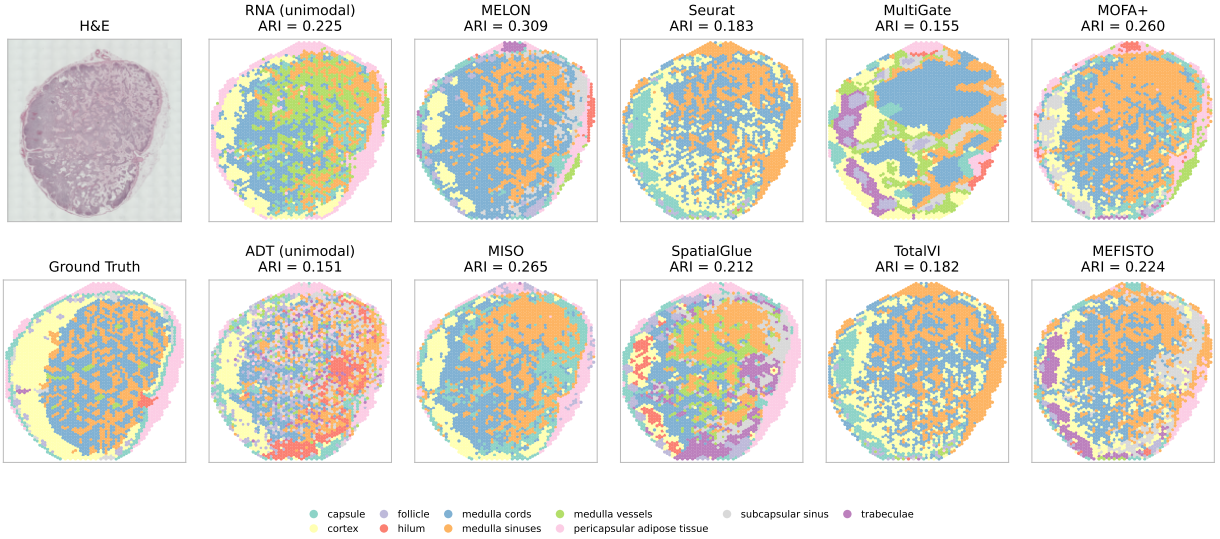


Figure 5. Human lymph node results: H&E tissue image, ground-truth annotation, and spatial domain assignments from MELON and seven multimodal baselines (per-panel ARI; methods ordered consistently across datasets). Quantitative clustering metrics across all methods are reported in Table 3.

D.5. Component Ablation on Mouse E15.5 Brain

The simulation in §4.1 ablates MELON's two main design axes (data augmentation and the neighborhood-aware spatial bias) on a controlled cross-modal target. Here we replicate that ablation on the Mouse E15.5 brain MISAR-seq benchmark, which is the largest paired RNA-ATAC dataset in our evaluation and the one with the strongest anatomical reference labels. We compare three MELON variants (FULL, no augmentation, no spatial bias) on the same cluster-number sweep used in Table 3 ($k \in \{6, 8, 10, 12, 14, 16\}$), with downstream Leiden clustering applied identically to all variants. Table 14 reports median $\pm$ IQR over the swept k. Removing either component substantially degrades every metric, with ARI and SpARI roughly halving (e.g., ARI drops from 0.383 to 0.204 without augmentation and to 0.215 without the spatial bias), confirming that the simulation-level conclusions transfer to real spatial multi-omics tissue.

D.6. Biological Interpretation

We probe MELON's frozen embedding with a joint pair-decoder and ask whether the resulting factors recover biologically meaningful, spatially organized RNA–ATAC (or RNA–protein) programs.

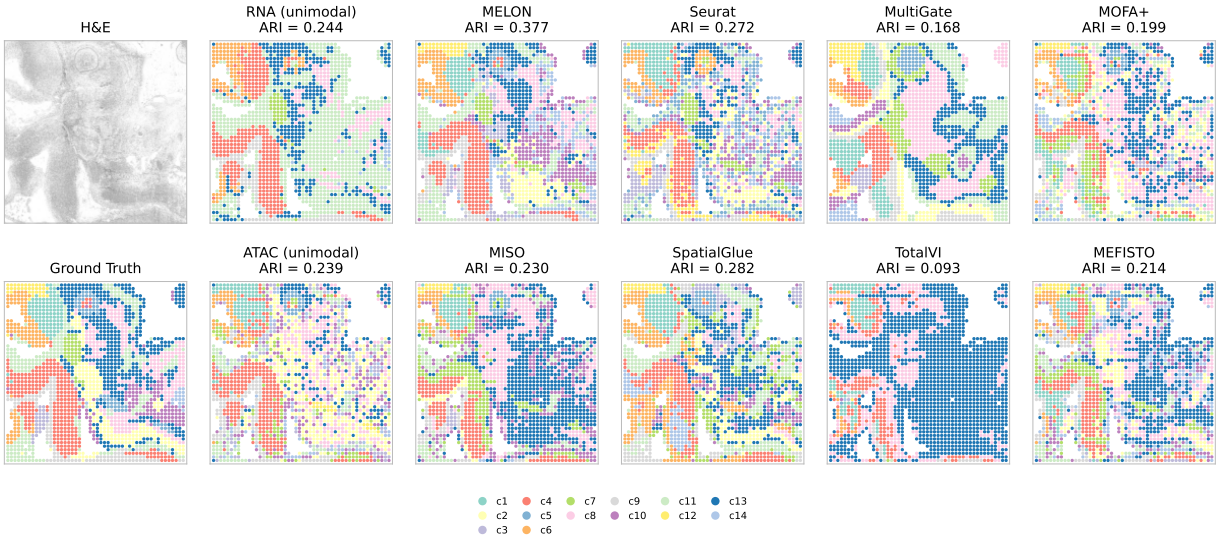


Figure 6. Mouse E13 embryo results: H&E tissue image, ground-truth annotation, and spatial domain assignments from MELON and seven multimodal baselines (per-panel ARI; methods ordered consistently across datasets). Quantitative clustering metrics across all methods are reported in Table 3.

Table 14. MELON component ablation on the Mouse E15.5 brain MISAR-seq dataset. Median ± IQR over swept cluster numbers $k \in \{6, 8, 10, 12, 14, 16\}$ (same protocol as Table 3). Best per metric in **bold**.

Variant	ARI	NMI	AMI	Homog.	SpARI	FMI	Jaccard	Purity
MELON full	$\mathbf{0.383 \pm 0.024}$	$\mathbf{0.553 \pm 0.017}$	$\mathbf{0.545 \pm 0.017}$	$\mathbf{0.534 \pm 0.040}$	$\mathbf{0.376 \pm 0.028}$	$\mathbf{0.466 \pm 0.020}$	$\mathbf{0.302 \pm 0.017}$	$\mathbf{0.652 \pm 0.026}$
− augmentation	0.204 ± 0.017	0.368 ± 0.004	0.357 ± 0.004	0.355 ± 0.015	0.174 ± 0.047	0.320 ± 0.038	0.187 ± 0.022	0.480 ± 0.005
− spatial bias	0.215 ± 0.017	0.355 ± 0.017	0.345 ± 0.017	0.334 ± 0.031	0.190 ± 0.038	0.334 ± 0.026	0.192 ± 0.014	0.485 ± 0.017

Decoder formulation. Let $\mathbf{z}_i \in \mathbb{R}^d$ denote the frozen MELON fused embedding of spot i, and let $\mathbf{x}_i^R \in \mathbb{R}^G$ and $\mathbf{x}_i^A \in \{0,1\}^P$ be its observed RNA expression and binarised ATAC (or protein) profile, where G and P are the number of genes and peaks respectively. We map each embedding through a small MLP ϕ with a low-rank bottleneck of width K (we use K=24 for the two RNA–ATAC datasets and K=16 for the lymph node):

$$\mathbf{u}_i \;=\; \mathrm{softplus}\big(\phi(\mathbf{z}_i)\big) \;\in\; \mathbb{R}^K_{\geq 0}, \tag{32}$$

so that $u_{i,k}$ is the non-negative usage of factor k at spot i. Two linear heads decode this bottleneck back into the two modalities,

$$\widehat{\mathbf{x}}_i^R \;=\; \mathbf{W}^R\,\mathbf{u}_i \;+\; \mathbf{b}^R, \qquad \widehat{\mathbf{x}}_i^A \;=\; \sigma\big(\mathbf{W}^A\,\mathbf{u}_i \;+\; \mathbf{b}^A\big), \tag{33}$$

with $\mathbf{W}^R \in \mathbb{R}^{G\times K}$, $\mathbf{W}^A \in \mathbb{R}^{P\times K}$ and σ the elementwise sigmoid. The two-headed reconstruction loss couples both modalities through the same factor activations,

$$\mathcal{L}_{\text{dec}} \;=\; \frac{1}{N}\sum_{i=1}^{N}\Big[\big\|\mathbf{x}_i^R - \widehat{\mathbf{x}}_i^R\big\|_2^2 \;+\; \lambda\,\mathrm{BCE}\big(\mathbf{x}_i^A,\, \widehat{\mathbf{x}}_i^A\big)\Big], \tag{34}$$

where $\mathrm{BCE}(\cdot,\cdot)$ is the elementwise binary cross-entropy averaged over peaks (or proteins) and λ balances the two heads (we use λ=1). Only the decoder parameters $\{\phi, \mathbf{W}^R, \mathbf{b}^R, \mathbf{W}^A, \mathbf{b}^A\}$ are trained; MELON is frozen. We optimize with Adam for 20 epochs on a 90/10 train–validation split.

Per-factor feature ranking. After training, factor k is summarized by three quantities: (i) its *spatial usage* $\{u_{i,k}\}_{i=1}^N$, plotted on tissue coordinates; (ii) its *top gene* and *top peak* (or top protein), defined as the columns of $\mathbf{W}^R$ and $\mathbf{W}^A$ with the largest loading on factor k,

$$g_k^\star \;=\; \arg\max_{g\in[G]} W^R_{g,k}, \qquad p_k^\star \;=\; \arg\max_{p\in[P]} W^A_{p,k}; \tag{35}$$

and (iii) its *top gene–peak pair score*, defined as the product of the two loadings,

$$s_{g,p,k} \;=\; W^{R}_{g,k} \cdot W^{A}_{p,k}, \qquad (g^{\dagger}_{k}, p^{\dagger}_{k}) \;=\; \arg\max_{(g,p)} s_{g,p,k}. \tag{36}$$

This product score promotes pairs in which the gene and the peak load on the *same* factor with high magnitude, so that a high score identifies cross-modal feature pairs jointly explained by a shared latent program. We rank all $G \times P$ pairs per factor by (36) and report the top entries in the supplementary CSVs.

Decoder factor program maps. For each dataset we show four high-variance factors with their top gene and top peak (or protein) label as Figures 7, 8, and 9. The mouse E13 embryo factors (0, 14, 8, 19) have biologically named top pairs including *Cdh19–Pmpcb* (factor 0; *Cdh19* is a cadherin involved in neural crest development) and *Mak–Pou3f1* (factor 14; *Pou3f1* marks early neural progenitors). The mouse E15.5 brain factors carry chromosomal-coordinate ATAC peaks that we abbreviate to `chrN:Mb` for display. The human lymph node decoder collapses every factor's top protein loading to the dominant cytokeratin marker KRT5 (which has prevalence ≈ 0.76 across the section), but the per-spot factor usage still clearly separates follicular and cortical-medullary territories. The remaining factors and full feature loadings are reported in the supplementary CSVs (`factor_top_genes.csv`, `factor_top_peaks.csv`, `factor_top_pairs.csv`) for each decoder.

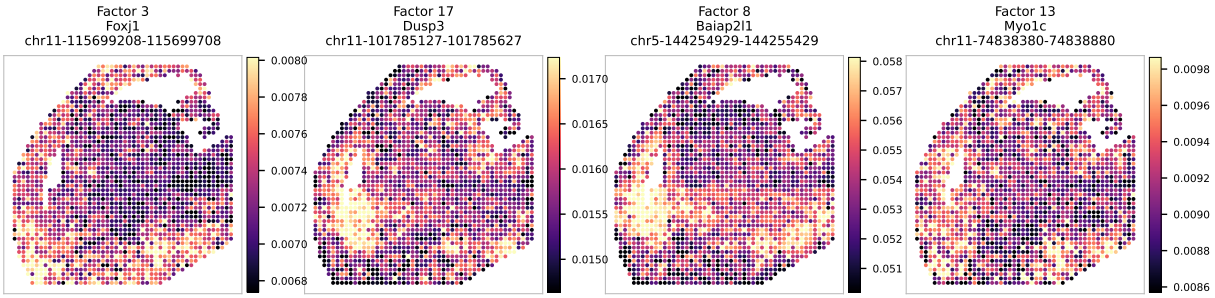


Figure 7. Mouse E15.5 brain. Spatial usage of four MELON decoder factors with the top gene and top ATAC peak (chromosomal coordinate, abbreviated to `chrN:Mb`) shown above each panel. Color is per-factor robust-clipped (5–95%) usage from the 24-factor pair decoder.

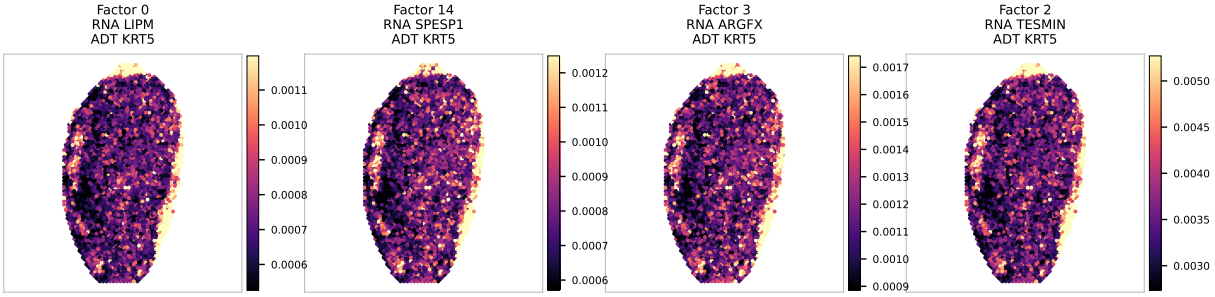


Figure 8. Human lymph node. Spatial usage of four MELON decoder factors, with the top gene for each factor shown. The dominant protein loading collapses to KRT5 across all factors, but the per-spot factor usage still cleanly separates anatomical compartments.

High-support gene–peak pair maps. As an additional perspective we select four gene–peak (or gene–protein for the lymph node) pairs per dataset from the decoder's top-pair list and visualize each pair's RNA expression, ATAC accessibility (or protein abundance), and decoder pair-activity on the tissue coordinates side by side (Figures 10, 11, and 12). The unfiltered top-pair list is dominated by ubiquitous peaks or markers (peak or protein prevalence ≈ 1.0 for several entries), so for the mouse E15.5 brain and E13 embryo we select pairs by spatial RNA-ATAC colocalization with binarized ATAC, and for the lymph node we select pairs whose RNA channel shows clear spatial pattern. Several of these pairs show concentrated activity patterns that are not explained by either modality in isolation, with per-spot pair activity having a sharper boundary than the marginal RNA or ATAC channel alone, consistent with the claim that MELON's embedding preserves cross-modal

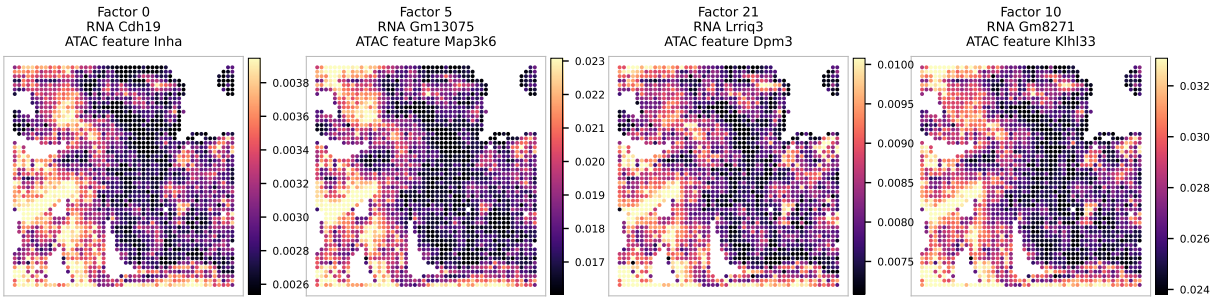


Figure 9. Mouse E13 embryo. Spatial usage of four MELON decoder factors with the top gene–peak pair shown above each panel. Each factor concentrates in a distinct anatomical region, demonstrating that the frozen embedding can be decoded into spatially coherent joint RNA-ATAC programs.

correspondence rather than collapsing to a shared consensus. The factor-based analysis above is the primary summary; the per-pair maps are included for completeness.

D.7. Tri-Modal Tonsil Extension: Additional Material

This appendix supplements §4.3. The tonsil sample is the public 10x Genomics Visium CytAssist FFPE Protein Expression human tonsil dataset (processed with Space Ranger v2.1.0): 4,194 in-tissue Visium spots; 18,085 genes (Visium Human Transcriptome Probe Set v2.0); 35-plex antibody capture (31 antibody markers + 4 isotype controls); and a matched H&E full-resolution image (31,967 $\times$ 39,132 pixels) summarized to 2,048-dimensional per-spot features via a frozen ImageNet-pretrained InceptionV3 applied to spot-centered crops at the published `spot_diameter_fullres`. The six compartment masks are defined as: *germinal center* = $\{\text{CR2} \geq q_{70}\} \cap \{\text{PCNA} \geq q_{70}\} \cap \{\text{BCL2} \leq q_{30}\} \cap \{\text{PDCD1} \geq q_{70}\}$; *mantle* = $\{\text{PAX5} \geq q_{70}\} \cap \{\text{BCL2} \geq q_{70}\} \cap \{\text{CD19} \geq q_{70}\} \cap \{\text{PCNA} \leq q_{30}\}$; *T-zone* = $\{\text{CD8A} \geq q_{70}\} \cap \{\text{HLA-DRA} \geq q_{70}\}$; *plasma* = $\{\text{SDC1} \geq q_{90}\}$; *myeloid* = $\{\text{CD68} \geq q_{70}\} \cap \{\text{CD163} \geq q_{70}\}$; *epithelium* = $\{\text{KRT5} \geq q_{90}\}$. All seven methods are clustered with $k = 10$ KMeans on their published embeddings, and per-compartment scores use the best-F1 cluster (size $\geq$ 30). MELON is trained with the same hyperparameters as the bi-modal experiments (Appendix B.4), with the third modality added as an additional input; ablation variants retrain from scratch with the corresponding modality removed. The bi-modal baselines listed in Appendix D.1 are reused here, augmented with the tri-modal-specific method **GROVER** (Xiao et al., 2025), a graph-guided spatial multi-omics fusion framework that jointly integrates omics and vision modalities through a mixture-of-experts regulator over per-modality graph encoders. Tri-modal MELON wall-clock and memory cost are reported in Appendix B.5.

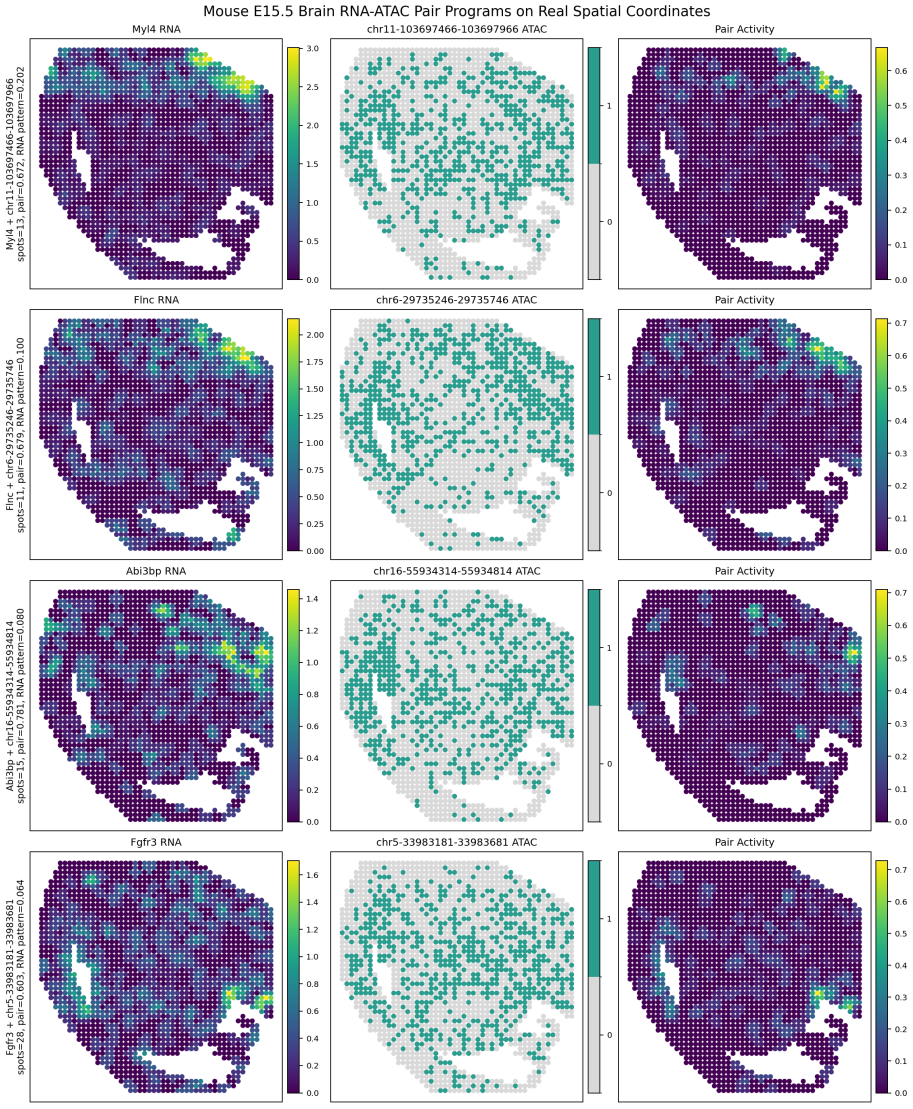


Figure 10. Mouse E15.5 brain. Spatial maps for four decoder gene–peak pairs selected by RNA-ATAC colocalization with binarized ATAC, including *Myl4*, *Abi3bp*, and *Fgfr3*. Each row is one pair; columns show (left) RNA expression of the gene, (centre) binarized ATAC accessibility of the peak (high vs. low), and (right) the decoder's joint pair activity.

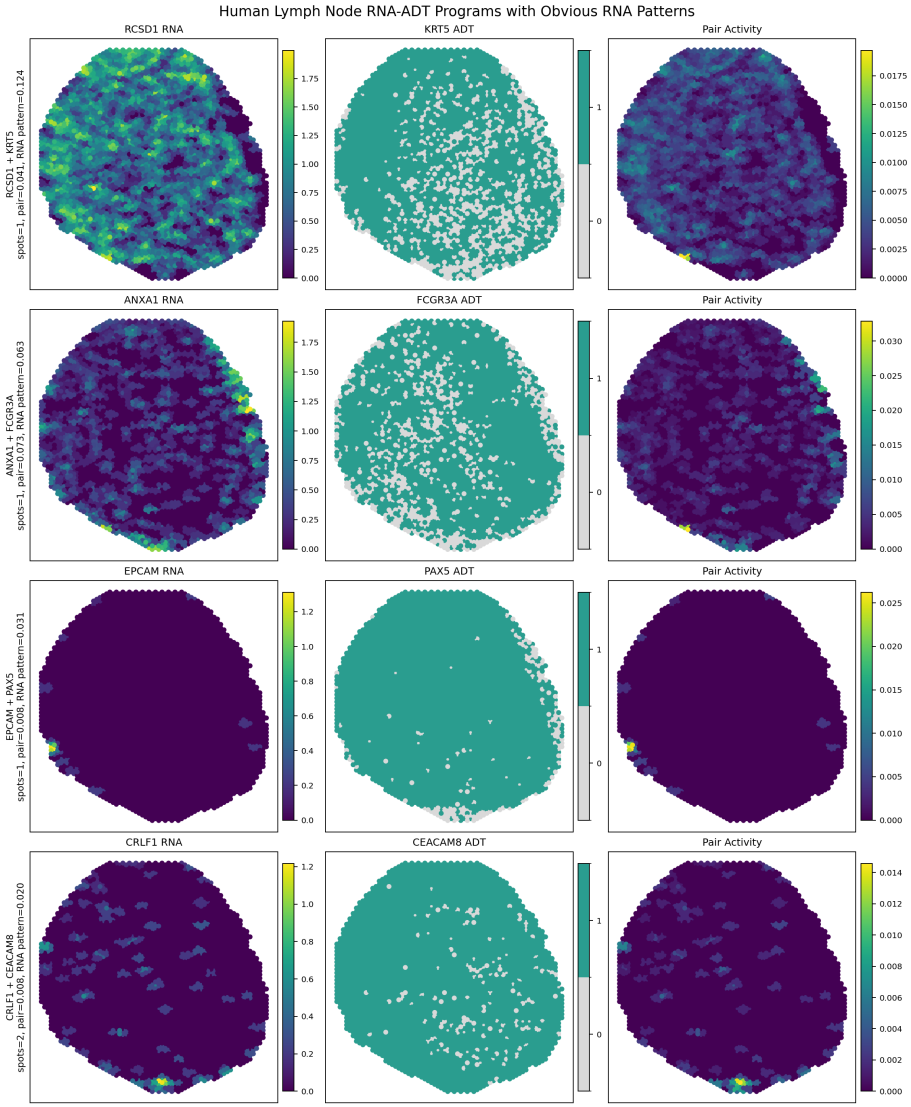


Figure 11. Human lymph node. Spatial maps for four decoder gene–protein pairs selected by spatial RNA pattern (*RCSD1–KRT5*, *ANXA1–FCGR3A*, *EPCAM–PAX5*, *CRLF1–CEACAM8*). Each row is one pair; columns show (left) RNA expression of the gene, (centre) ADT abundance of the protein marker, and (right) the decoder's joint pair activity. Pairs include canonical immune markers *FCGR3A* (CD16), *PAX5* (B-cell lineage), and *CEACAM8* (granulocyte).

Figure 12. Mouse E13 embryo. Spatial maps for four decoder gene–peak pairs selected by RNA-ATAC colocalization (*Il17re–Gm13057*, *Tap1–Gm13057*, *Nkx2.1–Smok3b*, *Gm24366–Smok3b*). Each row is one pair; columns show (left) RNA expression of the gene, (centre) binarized ATAC accessibility of the peak (high vs. low), and (right) the decoder's joint pair activity across all spots.

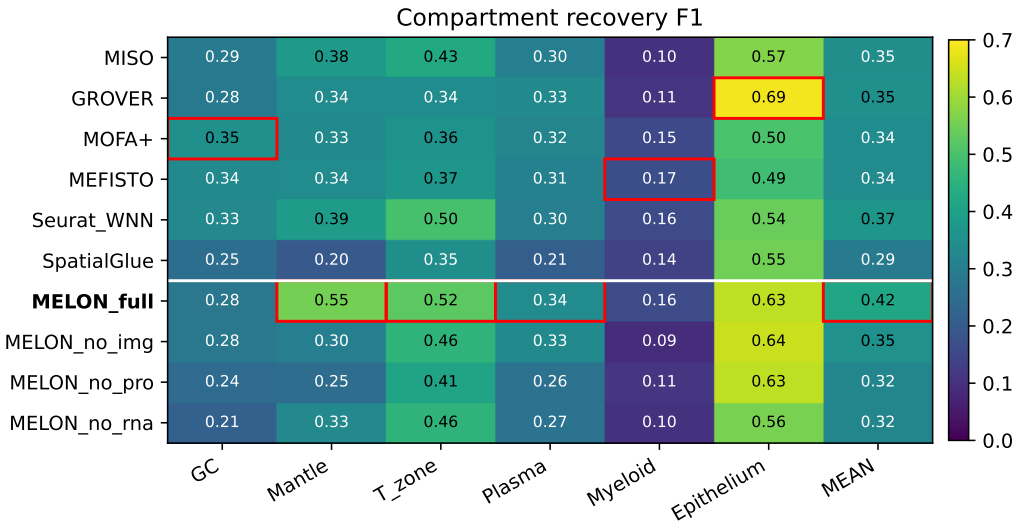


Figure 13. **Per-compartment best-cluster F1 on the human tonsil tri-modal sample** (six canonical IHC-defined compartments + the column-wise mean). Red boxes mark the column maximum; the white horizontal divider separates baselines (top) from MELON variants (bottom). MELON ranks first on the aggregate (MEAN) and on the three largest compartments (mantle, T-zone, plasma). Per-compartment AUROC variant in Fig. 14; per-method spatial maps in Fig. 15.

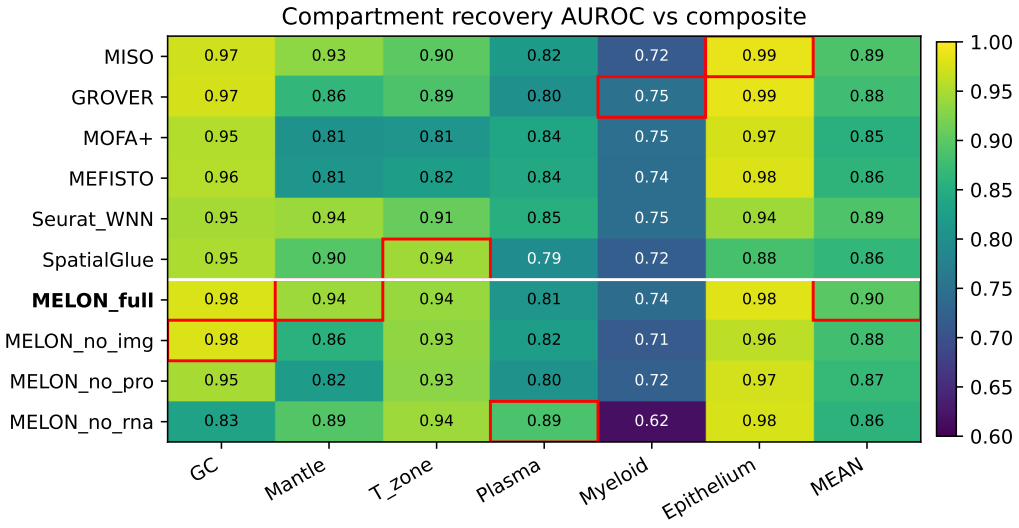


Figure 14. **Tonsil compartment recovery AUROC (continuous variant of Fig. 13).** For each compartment, AUROC of the best cluster's membership against the per-spot continuous compartment composite z-score (sum of marker z-scores with the canonical sign per marker). Red boxes mark column maxima. MELON ranks first on the aggregate (MEAN) and on the mantle compartment.

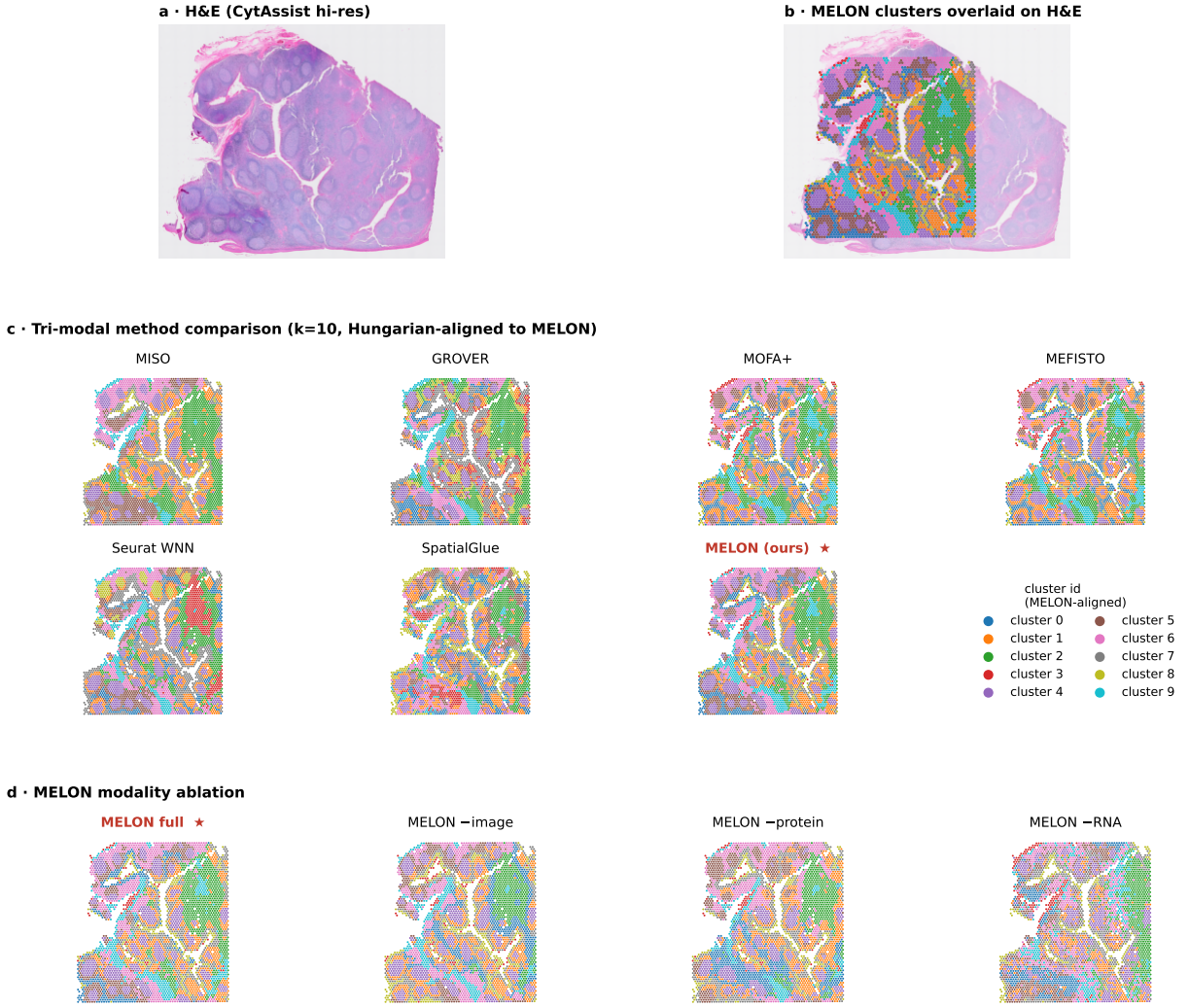


Figure 15. **Tonsil spatial cluster maps (combined panel).** (a) Hi-resolution H&E reference image. (b) MELON's $k = 10$ KMeans clusters overlaid on the H&E. (c) Side-by-side comparison of all seven tri-modal methods, with cluster IDs Hungarian-aligned to MELON's color scheme (legend at right). (d) MELON modality ablation: the full tri-modal model versus removing each modality during training. The visible degradation in (d), particularly for −protein and −RNA, mirrors the quantitative drops in Table 4.