ATLASS2V: ORGAN ATLAS–GUIDED PREDICTION OF 3D TISSUE DISTRIBUTIONS FROM A SINGLE SECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Three-dimensional tissue reconstruction typically relies on multiple spatially profiled sections from the same specimen, limiting its applicability to valuable datasets represented by a single section. We introduce ATLASS2V, which combines one observed 2D tissue section with a spatially registered 3D reference atlas to predict cell-density and cell-type distributions above and below the section. The atlas supplies depth-resolved anatomical structure and spatial context beyond the observed plane, while the observed section captures specimen-specific deviations from this population prior. ATLASS2V calibrates atlas abundance to the observed section and transfers its residual through a reference-trained distance gate, retaining more individual information nearby and relying increasingly on the atlas at greater distances. On an independent mouse, ATLASS2V improves reconstruction over central-copy and Atlas-only across near-to-intermediate offsets and converges toward the atlas at longer distances. Strict outer-animal validation shows consistent gains over central-copy in all four held-out mice, increasing mean composition Pearson from 0.392 to 0.629. A marker-density adaptation further improves distance-integrated correlation over each animal's best fixed expert in 58 of 60 held-out mice. Organ-specific studies demonstrate feasibility in human lung, breast, and kidney. Atlas-derived context also improves sparse-panel inference: with ten measured genes, hidden-gene correlation increases from 0.6755 to 0.7268, while 34-class cell-typing macro-F1 increases from 0.3058 with 2D atlas context to 0.4707 with correct 3D context. These results support reference-conditioned prediction as a route to extending single-section datasets into virtual 3D tissue distributions without collecting serial sections from each target specimen.

1 INTRODUCTION

Most spatial transcriptomic measurements and analyses describe two-dimensional tissue sections (Ståhl et al., 2016; Rodriques et al., 2019), whereas tissues exist and function in three dimensions. Cellular neighborhoods, anatomical boundaries, and molecular gradients extend above and below any sampled plane. Patterns that appear isolated in one section may belong to continuous 3D structures, while relationships across depth remain unobserved. A 2D section therefore captures only part of the spatial context needed to interpret local cell states and tissue organization. Moving from planar observations to 3D distributions is important for understanding how these local patterns fit within the broader anatomical and molecular organization of the surrounding tissue, as illustrated by whole-brain spatial references (Zhang et al., 2023; Yao et al., 2023).

Existing 3D approaches span several complementary directions. PASTE, GPSA, and ST-GEARS align multiple measured sections (Zeira et al., 2022; Jones et al., 2023; Xia et al., 2024); SpatialZ generates virtual sections between measured planes (Lin et al., 2026); and HoloTea and VORTEX exploit serial histology or volumetric morphology (Sanian et al., 2025; Almagro-Pérez et al., 2025). Thus, the predominant route reconstructs tissue from multiple sections, while morphology-guided alternatives obtain additional target-specific structure from imaging. These input requirements leave valuable single-section datasets without a direct route to volumetric reconstruction when neither neighboring sections nor 3D images are available. For rare or archived specimens, additional sectioning may be impractical; increasing molecular coverage within the available plane does not supply the missing depth. Extending such datasets requires information beyond the observed plane, which alone does not uniquely determine the surrounding volume.

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

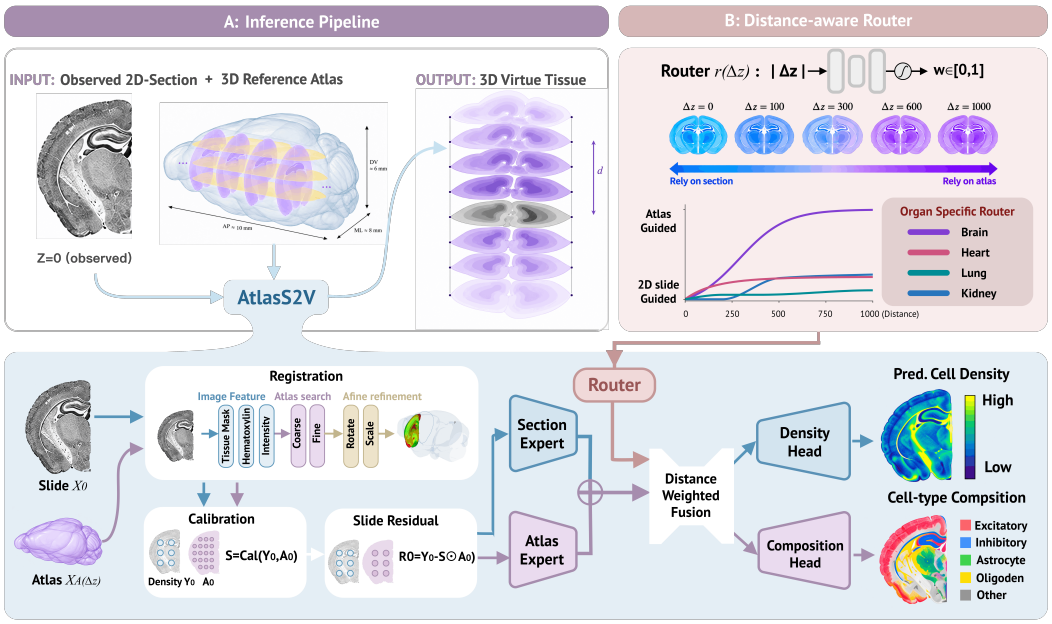


Figure 1: Overview of ATLASS2V for atlas-guided single-section-to-3D prediction.

We introduce ATLASS2V to predict 3D cell-density and cell-type distributions from one observed section and an external, spatially registered 3D reference (Fig. 1). Here, an *atlas* specifies cellular distributions in anatomical coordinates, not merely a catalogue of cell types. The reference adds spatial context, while the observed section captures specimen-specific variation. ATLASS2V calibrates reference abundance, extracts the observed residual, and transfers it through a reference-trained distance gate: nearby predictions retain more individual information, whereas distant predictions rely increasingly on reference structure. The output estimates tissue distributions rather than exact unseen cell positions.

Across held-out specimens, ATLASS2V predicts cellular and molecular organization beyond the observed plane. Independent-mouse reconstruction, cross-animal validation, and human-organ studies assess this ability under task-specific fitting. Downstream experiments then test whether the resulting spatial context improves inference when the observed section contains only a small measured gene panel. Our contributions are:

- **We propose ATLASS2V to bridge 2D sections and virtual 3D tissue.** By referencing a complete 3D organ atlas, it predicts tissue distributions beyond a single observed section. An organ-specific router adjusts the atlas contribution across depth, balancing reference anatomy with individual information from the observed section. To our knowledge, this is the first framework for *single-section-to-volume prediction* requiring neither additional target sections nor volumetric imaging.
- **We reconstruct cellular and molecular distributions in 3D.** ATLASS2V recovers cell density, cell-type composition, and gene expression on unobserved sections, capturing spatial organization across evaluated offsets up to 2 mm. Held-out evaluations show improved reconstruction over copying the observed section, with cross-animal validation and organ-specific studies supporting applicability beyond a single specimen (mean composition Pearson: $0.392 \rightarrow 0.629$ over central-copy across four held-out mice).
- **We improve downstream inference with reconstructed 3D context.** Atlas-derived spatial information enhances hidden-gene prediction and cell-type inference from sparse measured gene panels. Hidden-gene prediction improves beyond strong 2D inputs, while correct depth information further improves cell typing over 2D atlas context, demonstrating practical benefits of the bridge from single-section measurements to virtual 3D tissue.

2 Related Work

3D reconstruction from multiple tissue sections. Multi-section approaches recover volumes by aligning measured tissue planes. PASTE uses optimal transport (Zeira et al., 2022); GPSA learns shared coordinates and spatial functions (Jones et al., 2023); and ST-GEARS combines cross-section correspondences with rigid and elastic registration (Xia et al., 2024). ATLASS2V predicts unobserved planes from one target section and an external 3D atlas. The single-section constraint applies to the target; reference construction and training may use multiple sections and specimens.

Volumetric prediction from tissue morphology. HoloTea predicts expression from H&E images using context from adjacent sections (Sanian et al., 2025), while VORTEX combines volumetric morphology with limited spatial transcriptomics (Almagro-Pérez et al., 2025). Both obtain additional target-specific structure from imaging. ATLASS2V obtains depth-dependent spatial context from the reference atlas, requiring no serial histology or volumetric imaging at inference.

Spatial reference registration. Tangram maps single-cell transcriptomes onto spatial measurements, while cell2location estimates cell-type abundances at measured locations (Biancalani et al., 2021; Kleshchevnikov et al., 2022). scArches maps query datasets into learned reference representations (Lotfollahi et al., 2022). These approaches relate observed molecular data to spatial or learned references. Their outputs primarily characterize data within observed support. ATLASS2V combines a spatial atlas query with observed-section information to predict cellular distributions beyond the measured plane.

3 Method

ATLASS2V predicts surrounding tissue by combining a 3D reference with information from one observed section. Inference queries the reference at the observed and requested depths, calibrates its abundance, extracts the specimen's observed deviation, and transfers that deviation with a distance-dependent weight. Density and composition are then read from the predicted class counts (Fig. 1).

3.1 Inputs and prediction target

Let $S_0 = \{(\boldsymbol{u}_i, \boldsymbol{x}_i)\}_{i=1}^n$ denote the observed section, with in-plane coordinates $\boldsymbol{u}_i$ and measured features $\boldsymbol{x}_i$. Given a spatial atlas $A : \mathbb{R}^3 \to \mathbb{R}_+^K$ and registered section pose ϕ, we predict a hidden field $Y_d : \Omega \to \mathbb{R}_+^K$ at signed offset d:

$$\hat{Y}_d = f_\theta(S_0, A, d). \tag{1}$$

For the main reconstruction task, $\mathbf{Y}_0(\boldsymbol{u})$ is the observed class-count field and $K = 8$. Reference construction and parameter fitting exclude the test animal; its observed section supplies calibration, while hidden sections are used only for scoring.

3.2 Querying the 3D reference

The Atlas supplies the spatial structure missing from a single section. Reference animals are mapped to a common anatomical frame and smoothed into nonnegative class-density fields with explicit coverage. Let $T_\phi(\boldsymbol{u}, d)$ map the observed surface and signed offset into Atlas coordinates. We query $\mathbf{A}_0(\boldsymbol{u}) = \mathbf{A}(T_\phi(\boldsymbol{u}, 0))$ and $\mathbf{A}_d(\boldsymbol{u}) = \mathbf{A}(T_\phi(\boldsymbol{u}, d))$. The paired queries distinguish the specimen's observed deviation from anatomical changes already represented in the reference. Signed offsets select different target anatomy on either side of the section. The main benchmark uses the supplied CCF pose and interpolates along the signed AP offset, without a target-derived registration surface. Atlas construction and optional pose-uncertainty queries are detailed in Appendix A.2.

3.3 Calibrating abundance and extracting individual variation

Calibration aligns reference abundance with the specimen before transferring local deviations. On the common supported mask $\mathcal{M}_0$, define observed and reference class totals $n_k = \sum_{\boldsymbol{u} \in \mathcal{M}_0} Y_{0,k}(\boldsymbol{u})$ and $a_k = \sum_{\boldsymbol{u} \in \mathcal{M}_0} A_{0,k}(\boldsymbol{u})$, with global scale $g = \sum_k n_k / \sum_k a_k$. For supported classes, $r_k = n_k / a_k$;

otherwise $r_k = g$. Clipping and shrinkage stabilize the class-wise scale:

$$s_k = \exp\big[0.75 \log\big(\operatorname{clip}(r_k, 0.1g, 10g)\big) + 0.25 \log g\big] . \tag{2}$$

Shrinkage limits how strongly a poorly represented class can alter the reference, while preserving class-specific adjustments when the observed section provides support. We then separate the calibrated population prediction from the specimen's spatial deviation:

$$\mathbf{F}_d(\boldsymbol{u}) = \mathbf{s} \odot \mathbf{A}_d(\boldsymbol{u}), \tag{3}$$

$$\mathbf{I}_d(\boldsymbol{u}) = [\mathbf{F}_d(\boldsymbol{u}) + \mathbf{Y}_0(\boldsymbol{u}) - \mathbf{s} \odot \mathbf{A}_0(\boldsymbol{u})]_+ . \tag{4}$$

Here $[\cdot]_+$ clips negative counts. The residual $\mathbf{Y}_0 - \mathbf{s} \odot \mathbf{A}_0$ retains individual information, and $\mathbf{I}_0 = \mathbf{Y}_0$. Whereas calibration adjusts overall class abundance, the residual records where the specimen differs from that calibrated reference, retaining spatial variation rather than only section-wide totals. The population expert $\mathbf{F}_d$ is the calibrated Atlas-only comparator. Numerical safeguards appear in Appendix A.3.

3.4 Fusing predictions across distance

A deviation observed at the center need not persist at distant depths. We therefore blend the individualized and population experts with a weight that decreases with absolute distance:

$$\hat{\mathbf{C}}_d(\boldsymbol{u}) = w(|d|)\mathbf{I}_d(\boldsymbol{u}) + [1 - w(|d|)]\mathbf{F}_d(\boldsymbol{u}), \tag{5}$$

$$w(|d|) = \sigma(b - \operatorname{softplus}(a)|d|) . \tag{6}$$

Distance is measured in millimeters; all locations at a requested offset share the same weight. Nearby predictions retain more specimen-specific variation, while distant predictions approach the calibrated Atlas. Because both experts share the target-depth reference, routing controls the influence of the observed residual while retaining the Atlas's depth-dependent structure. Only the gate uses absolute distance: the Atlas query remains signed, so predictions at equal distances on opposite sides need not coincide.

The two parameters a, b are fitted on four reference animals through leave-one-reference-animal-out prediction, using the other three to construct each training Atlas. The fitted gate and final four-animal Atlas are frozen before independent-mouse evaluation. Outer-animal validation repeats fitting within the remaining references. Training distances and the weighted least-squares objective are specified in Appendix B.2. At inference, only the abundance scales and spatial residual are recomputed from the observed section; neither the Atlas nor the distance gate is updated on the test specimen.

3.5 Reading out density and composition

The predicted counts yield both outputs without a separate learned head:

$$\hat{\rho}_d(\boldsymbol{u}) = \sum_k \hat{C}_{d,k}(\boldsymbol{u}), \qquad \hat{\pi}_{d,k}(\boldsymbol{u}) = \frac{\hat{C}_{d,k}(\boldsymbol{u})}{\hat{\rho}_d(\boldsymbol{u})} \quad \text{for } \hat{\rho}_d(\boldsymbol{u}) > 0. \tag{7}$$

Zero-mass bins remain empty, and total density may change with depth. Querying several offsets produces a stack of predicted fields, each anchored to the same observed section rather than recursively using previously predicted layers. The main predictor is fully specified by the Atlas, observed-section calibration and residual, and the fitted distance gate.

Marker-density, gene-expression, and organ studies adapt their reference fields or predictors to each task. Downstream imputation and typing instead fit separate readouts that use Atlas context on the observed section (Appendix F). These experiments do not apply one unchanged mouse checkpoint to every task or organ.

Evaluation separates spatial pattern agreement from composition and abundance errors: class-wise Pearson compares count patterns, JSD compares normalized cell-type proportions, and nMAE measures absolute count error relative to target count. All methods share the same scoring grid and masks within each comparison (Appendix A.4).

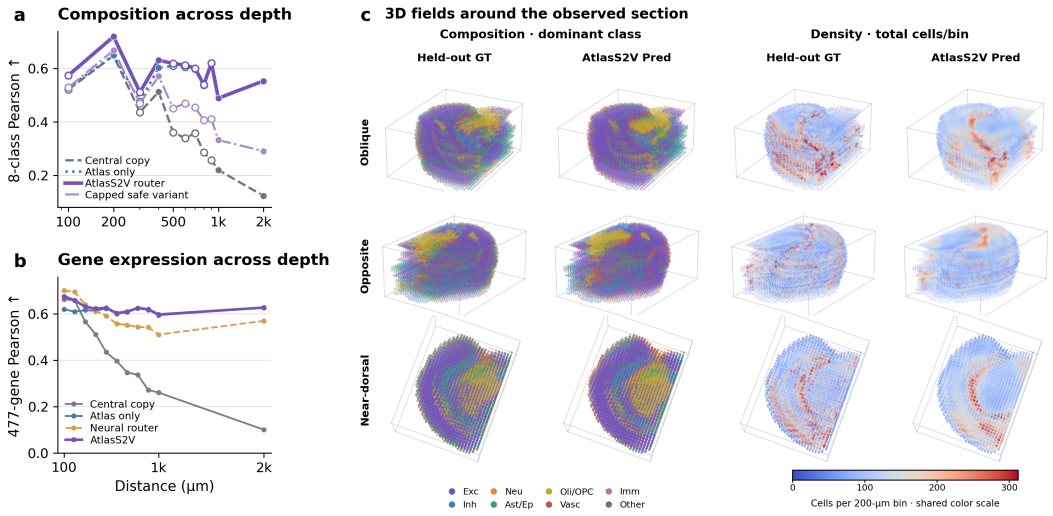


Figure 2: Single-section prediction in held-out C57BL6J-638850. (a) Distance routing improves cell-type prediction across depth. (b) Atlas guidance improves expression prediction across 477 genes. (c) Predicted 3D composition and density capture major spatial patterns.

4 Experiments

4.1 Evaluation Overview

We evaluate two questions: whether one observed section and an atlas predict hidden tissue distributions, and whether atlas context improves downstream inference under sparse molecular measurement. Hidden targets are scoring-only; reference construction and fitting follow each study's specified specimen- or section-level isolation protocol. Reconstruction uses one observed target section per prediction, with specimen-specific calibration restricted to that section. Test measurements never guide model selection. Reconstruction uses spatial Pearson, composition JSD, and normalized count or density error; downstream tasks use hidden-gene Pearson, macro-F1, and composition metrics. Uncertainty is summarized at specimen or spatial-block level. Study-specific datasets and splits are described below, with detailed protocols in Appendix A.

4.2 Single-Section-to-3D Reconstruction

We evaluate hidden-section cell-type, density, and gene-expression prediction across depth, followed by generalization across specimens and organs. Comparisons with central-copy and Atlas-only assess the contribution of reference guidance and individual observations.

4.2.1 AtlasS2V predicts 3D distributions from one section

Using an Atlas built from four reference mice, we evaluate cell-type distributions, total density, and gene expression in held-out C57BL6J-638850 at offsets of 100–2,000 μm. The experiment tests prediction fidelity and its persistence with distance from the observed section. The observed section supplies specimen-specific calibration, while hidden sections are reserved for evaluation.

Cell-type reconstruction. AtlasS2V outperforms all compared baselines at most evaluated distances in spatial correlation and composition error (Fig. 2a; Table 1). Predicted maps capture major regional cell-type patterns observed in the held-out tissue (Fig. 2c), providing spatial evidence alongside the aggregate metrics. Appendix Fig. 6 and Table 6 provide section-wise visualizations and complete scores.

Cell-density reconstruction. Summing predicted class counts yields total density. Density Pearson is 0.7516 at 200 μm, 0.3448 at 1,000 μm, and 0.5116 at 2,000 μm; at the two farther offsets, central-copy reaches 0.2224 and 0.3407. The maps capture major high- and low-density regions using a

shared count scale. Agreement varies with depth; spatial comparisons and complete distance-resolved results appear in Fig. 2c and Appendix Table 5; Appendix Fig. 7 provides enlarged views.

Gene-level reconstruction. A separately trained expression predictor extends the framework to molecular fields. Across 477 genes, mean spatial Pearson remains between 0.5977 and 0.6758 over the evaluated offsets (Fig. 2b). At 2,000 μm, it reaches 0.6273 versus 0.1011 for central-copy. The distance-balanced mean is 0.6272, compared with 0.4139 for central-copy and 0.6141 for the calibrated molecular Atlas. Table 2 shows reference-ranked variable and highly expressed genes; Appendix B.7 reports this exploratory extension's protocol and complete comparisons.

Together, these results show that one observed section and a reference Atlas can recover cell-type organization, density variation, and molecular patterns on unobserved sections at evaluated offsets up to 2 mm, supporting virtual 3D reconstruction at both cellular and molecular levels.

Table 1: Cell-type reconstruction.

Method / Distance (μm)	200	400	600	800	2,000
Class Pearson ↑					
Central-copy	0.648	0.513	0.339	0.286	0.124
Atlas-only	0.649	0.604	0.606	0.538	0.553
U-Net	0.547	0.518	0.499	0.454	0.461
Nearest ref.	0.637	0.602	0.584	0.501	0.543
ATLAS2V	**0.720**	**0.631**	**0.613**	**0.539**	**0.553**
Composition JSD ↓					
Central-copy	0.0524	0.0787	0.1047	0.1132	0.1787
Atlas-only	0.0388	0.0380	0.0430	0.0399	**0.0372**
U-Net	0.0370	0.0378	0.0429	0.0406	0.0385
Nearest ref.	0.0412	0.0411	0.0471	0.0432	0.0406
ATLAS2V	**0.0349**	**0.0370**	**0.0424**	**0.0398**	0.0372

Table 2: Gene-level reconstruction.

Gene / Distance (μm)	200	400	600	800	2,000
Highly variable					
Slc17a7	0.914	0.880	0.887	0.891	0.905
Mog	0.890	0.864	0.863	0.867	0.818
Ptk2b	0.758	0.769	0.699	0.793	0.786
Cbln1	0.792	0.737	0.746	0.788	0.741
Zic1	0.886	0.866	0.869	0.865	0.858
Mean (all 120)	**0.783**	**0.748**	**0.724**	**0.748**	**0.754**
Highly expressed					
Cldn5	0.578	0.551	0.513	0.546	0.573
Slc17a7	0.914	0.880	0.887	0.891	0.905
Nfix	0.712	0.665	0.560	0.664	0.721
Pou3f3	0.729	0.660	0.665	0.697	0.658
Sox10	0.839	0.815	0.809	0.824	0.788
Mean (all 120)	**0.753**	**0.715**	**0.694**	**0.709**	**0.713**

Both tables evaluate C57BL6J-638850 with separately fitted composition and expression predictors. **Left:** native eight-class counts at a 400-μm grid; U-Net averages three seeds. The five displayed offsets are a presentation subset: 600/800 μm are post hoc, while the prespecified endpoints remain 200/400/1,000/2,000 μm. All 11 distances are retained in Table 6. **Right:** ATLAS2V gene spatial Pearson at each distance, with centres weighted equally; each group contains the reference-ranked top 25% of the 477-gene panel. Five examples are shown, but each mean includes all 120 group members. Bold marks the best method on the left and the group means on the right.

4.2.2 CROSS-SPECIMEN GENERALIZATION: GAINS ACROSS 64 ANIMALS

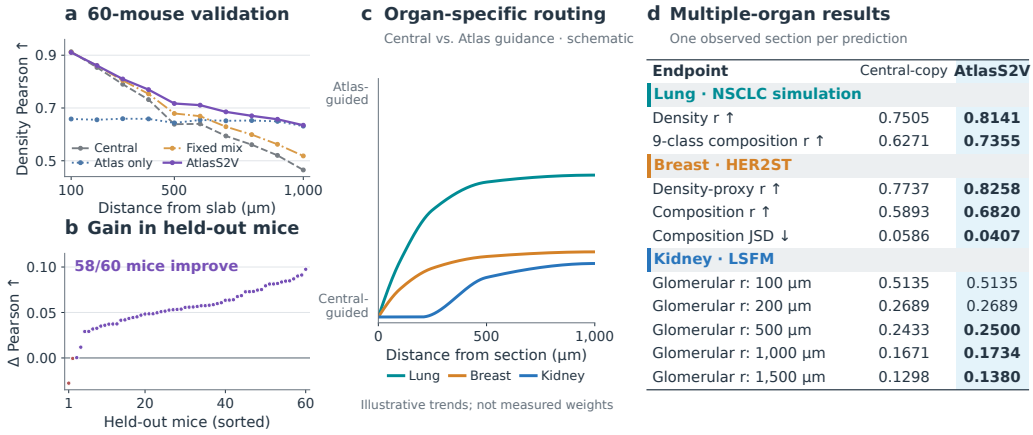
We next ask whether single-section prediction generalizes across individuals, beyond the independent mouse examined above. Two complementary cohorts test different aspects of this ability:

- **Cell-type composition:** four Zhuang-ABCA mice test whether predictions of multiclass spatial organization remain accurate across animals.
- **Cell density:** 60 BrainCAD mice (Zhao et al., 2025) spanning 20 marker-defined populations test distance-adaptive density prediction across individuals.

Both studies fit models for their respective tasks with each test animal excluded from fitting; the four-mouse study also rebuilds the Atlas without that animal.

Cell-type composition improves in all four held-out mice. ATLAS2V outperforms central-copy in spatial Pearson, composition JSD, and normalized count error in every animal when evaluated over the distances shared by all four mice. Mean Pearson increases from 0.392 to 0.629. Compared with the calibrated Atlas, spatial correlation improves in three of four animals and count error decreases in all four. These results show that the benefit of Atlas guidance extends across animals, while the observed section provides individual corrections beyond the population reference. Appendix C.1 reports the complete results and leave-one-animal-out protocol.

Distance routing improves density prediction in 58 of 60 held-out mice. The router exceeds each mouse's best fixed expert in distance-integrated density correlation for 58 of 60 animals, with a mean gain of 0.055 (Fig. 3b). The distance curves explain this advantage: center-derived information is most useful nearby, whereas Atlas guidance remains reliable farther away (Fig. 3a). By shifting their relative contributions with distance, the router combines these complementary strengths across



Endpoint	Central-copy	AtlasS2V
Lung · NSCLC simulation		
Density r ↑	0.7505	**0.8141**
9-class composition r ↑	0.6271	**0.7355**
Breast · HER2ST		
Density-proxy r ↑	0.7737	**0.8258**
Composition r ↑	0.5893	**0.6820**
Composition JSD ↓	0.0586	**0.0407**
Kidney · LSFM		
Glomerular r: 100 µm	0.5135	0.5135
Glomerular r: 200 µm	0.2689	0.2689
Glomerular r: 500 µm	0.2433	**0.2500**
Glomerular r: 1,000 µm	0.1671	**0.1734**
Glomerular r: 1,500 µm	0.1298	**0.1380**

Figure 3: Cross-animal generalization and multiple-organ reconstruction. (a) Distance-aware routing improves mean density prediction across 60 held-out mice. (b) Gains over each mouse's best fixed expert in 58/60 mice support cross-animal generalization. (c) Schematic curves illustrate organ-specific balances between observed-section information and reference guidance. (d) Lung, breast, and kidney results demonstrate reconstruction under different reference and fitting protocols.

animals and marker-defined populations. Each mouse contributes one such population, so this experiment evaluates density prediction. Full metrics, animal-level uncertainty estimates, and fitting details appear in Appendix C.2.

Together, these studies support cross-animal generalization of Atlas-guided single-section prediction for both multiclass composition and cell density under task-specific fitting. The observed section supplies individual information, while the Atlas provides spatial structure that supports prediction beyond that section.

4.2.3 Multiple-Organ Generalization: Atlas Gains Beyond the Brain

We next test whether the same single-section prediction framework extends to human organs with different tissue organization and reference availability. The schematic routing curves in Fig. 3c illustrate an organ-specific balance between observed-section information and reference guidance. Nearby predictions retain more information from the observed section, while reference guidance can contribute more strongly at greater distances. The timing and extent of this transition depend on tissue structure and reference reliability, motivating organ-specific routing policies. Figure 3d reports the measured reconstruction results below.

Lung: reference sections improve density and composition prediction. Across six hidden-section predictions within one NSCLC tumour (Pentimalli et al., 2025), the serial reference improves both endpoints over central-copy: density Pearson rises from 0.751 to 0.814 and composition Pearson from 0.627 to 0.736. Correct-depth queries also outperform reversed-depth queries, supporting the contribution of reference spatial structure to within-tumour prediction (Appendix D.1).

Breast: cross-patient guidance improves prediction in all eight patients. Leave-one-patient-out evaluation on HER2ST (Andersson et al., 2021) shows improved density-proxy correlation, composition correlation, and composition JSD in every patient. Composition Pearson increases from 0.589 to 0.682. The gains span 28 hidden sections, indicating that the reference remains useful despite patient-specific tumour organization. These results support reference-guided prediction even without a shared anatomical coordinate system. Density is an H&E-derived proxy; Appendix D.2 provides tissue-only results and controls.

Kidney: donor-independent guidance yields modest gains farther from the section. In two held-out donors (McLaughlin et al., 2025), the model retains central-copy nearby and introduces an Atlas correction at greater depths. Mean glomerular-field Pearson increases modestly from 0.265 to 0.269, with lower nMAE and JSD. Both donors improve at the farthest evaluated distance. The

correct anatomical reference also outperforms a wrong-region control at every far-distance endpoint, linking these gains to relevant reference structure (Appendix D.3).

Together, these results extend ATLASS2V to lung, breast, and kidney through organ-specific adaptation. The benefit of reference guidance depends on tissue structure and reference availability, motivating different balances between individual observations and reference information.

4.3 ABLATION: COMPONENT CONTRIBUTIONS AND ORGAN-SPECIFIC ROUTING

We examine Atlas guidance, specimen-specific residuals, and adaptive routing. Table 3 compares frozen mouse composition variants without refitting on the test animal; Table 4 evaluates routing in additional organs.

Table 3: Held-out mouse component ablations.

Method	Pearson ↑	JSD ↓	nMAE ↓
Central-copy (w/o Atlas)	0.3762	0.11086	0.7077
w/o residual	0.5737	0.03968	0.5363
w/o router (fixed 1:1)	0.5953	0.04107	0.5170
ATLASS2V	**0.5983**	**0.03845**	**0.5055**

Table 4: Organ-specific adaptive routing.

Experiment	Fixed 0.5	Full model	Δ
Zebrafish heart	0.5980	**0.6593**	**+0.0613**
Zebrafish heart	0.4361	**0.4752**	**+0.0391**
Breast	0.6516	**0.6820**	**+0.0304**

Left: held-out C57BL6J-638850; eight-class predictions share the same 400-μm grid and scoring masks. Metrics are averaged equally over 200/400/1,000/2,000 μm (193 directed section pairs). Pearson is the macro class-wise spatial correlation; nMAE is absolute class-count error normalized by ground-truth total count. **Right:** fixed 0.5 routing gives equal weights to the two experts; the full model uses the organ-specific router. All three endpoints are Pearson correlations (↑); Δ is full model minus fixed 0.5 router.

(I) Atlas. Reference guidance improves prediction beyond the observed section. The full model increases composition Pearson from 0.3762 with central-copy to 0.5983 and reduces both JSD and normalized count error.

(II) Residual. Individual spatial corrections add information beyond abundance calibration. Removing residual transfer retains the calibrated Atlas but reduces Pearson from 0.5983 to 0.5737 and increases JSD from 0.03845 to 0.03968. Thus, calibrating reference abundance alone does not capture all useful specimen-specific information.

(III) Router. Adaptive-routing gains vary across organs. In the mouse experiment, fixing both expert weights at 0.5 approaches the full model in Pearson, with a gap of only 0.0030; adaptive routing also reduces JSD and nMAE by 6.4% and 2.2%. Larger gains appear in zebrafish heart (Li et al., 2025) and breast: routing improves heart density and composition Pearson by 0.0613 and 0.0391, respectively, and breast composition Pearson by 0.0304. These results support adapting the balance between observed-section evidence and reference guidance to the organ and task. An equal-cell-budget reference-coverage ablation appears in Appendix E.1.

4.4 DOWNSTREAM TASKS: ATLAS CONTEXT IMPROVES SPARSE-PANEL INFERENCE

These experiments evaluate molecular and cellular inference within the observed section under sparse gene measurement. They complement hidden-section reconstruction by testing whether reference context supplies information missing from the measured panel and whether retaining depth improves cell typing beyond a two-dimensional Atlas. Hidden-gene prediction uses a separately fitted readout of measured expression and Atlas features, distinct from the density and composition predictor.

4.4.1 ATLAS GUIDANCE IMPROVES HIDDEN-GENE PREDICTION

Evaluation. We test whether Atlas context improves hidden-gene prediction beyond a strong 2D predictor using measured expression, spatial neighborhoods, and in-plane coordinates. Three Xenium sections from one mouse (10x Genomics, 2023) rotate through whole-section training, validation, and testing. All 198 genes outside the maximum 50-gene panel remain hidden at every panel size, with prespecified endpoints of 10, 25, and 50 measured genes.

Sparse-panel gains. With ten measured genes, mean hidden-gene Pearson increases from 0.6755 to 0.7268, with gains in all three held-out sections (Fig. 4a). The relative improvement decreases

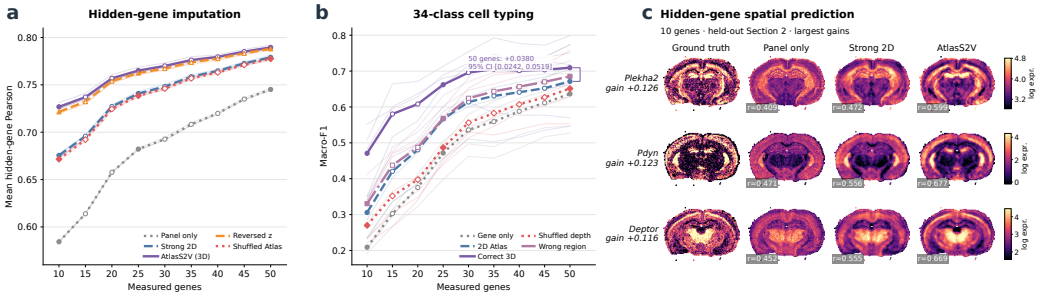


Figure 4: Atlas context improves prediction from sparse gene panels. (a) Hidden-gene prediction improves beyond strong 2D inputs. (b) Correct 3D context improves cell typing across held-out animals. (c) Selected gene maps illustrate improved spatial expression recovery.

from 7.60% at ten genes to 1.34% at fifty genes. Selected maps illustrate recovered spatial patterns (Fig. 4c); because genes were selected by mean gain across rotations, these examples visualize the observed improvement but do not independently validate prediction quality.

Interpretation. The declining gain with richer panels suggests that Atlas context complements information missing from sparse molecular observations. A depth-marginalized Atlas performs similarly to the soft 3D query, so these results support reference utility without isolating a precise-depth advantage. Appendix F.1 details the section-level protocol and reference controls.

4.4.2 Depth Information Improves Cell-Type Inference

Evaluation. We isolate the contribution of depth through leave-one-animal-out 34-class typing in four Zhuang-ABCA mice with official CCF coordinates. Each test animal is excluded from classifier fitting, Atlas construction, and model selection, allowing us to assess transfer across animals.

Cross-animal results. Correct 3D context improves macro-F1 over the depth-marginalized Atlas at ten genes ($0.3058 \rightarrow 0.4707$) and fifty genes ($0.6713 \rightarrow 0.7094$; Fig. 4b). All four animals improve at the prespecified panel sizes, with computational seeds averaged within each animal. At fifty genes, the paired macro-F1 gain is 0.0380 (animal-bootstrap 95% CI $[0.0242, 0.0519]$). Permuting only the target cells' section-normal coordinate or shifting their in-plane Atlas query by 1 mm reduces performance without altering measured expression, showing the value of matched anatomical context.

Interpretation. Together, the depth-marginalized comparison and coordinate controls support a contribution from anatomical depth under known coordinates (Appendix F.2). This benefit extends to pseudo-spot composition: with ten genes, Pearson increases from 0.6505 to 0.7859 at 55 μm and from 0.6490 to 0.7814 at 100 μm. Gains approach zero with larger panels, indicating that the value of Atlas context depends on how much molecular information is already observed (Appendix F.3).

5 Limitations and Conclusion

AtlasS2V combines spatial reference structure with specimen-specific observations to extend single sections into virtual 3D tissue. Reconstruction and held-out-animal evaluations support this approach, with downstream gains strongest under sparse molecular measurement (Figs. 2–4). These predictions remain reference-conditioned estimates: a single section cannot uniquely determine unseen cells or pathology, and incomplete references may omit rare states. Brain benchmarks assume known anatomical coordinates, organ studies require task-specific adaptation, and Xenium imputation covers only one animal. Negative transfer further shows that Atlas utility depends on reference compatibility (Appendix H.4). Future work should combine registration-aware evaluation, disease-relevant references, and calibrated uncertainty with frozen validation across specimens and platforms. Within these limits, AtlasS2V offers a practical route for extending valuable single-section datasets into virtual 3D tissue, retaining individual information from the observed section while using the Atlas to supply spatial structure missing from sparse measurements.

AI USE STATEMENT

Generative AI assisted writing and LaTeX editing, not experimental measurements. The authors reviewed all content and remain responsible for the manuscript.

ETHICS AND REPRODUCIBILITY STATEMENT

We reuse public animal and de-identified human datasets under their access conditions, with no new interventions or recruitment. Predictions are not clinical measurements. Anonymous code is available at `https://anonymous.4open.science/r/AtlasS2V-code-7DC9`. We will release dataset manifests, specimen splits, preprocessing specifications, frozen model configurations, and reproduction scripts. Appendix A details data isolation and evaluation; development cohorts remain distinct from external tests.

REFERENCES

10x Genomics. Fresh Frozen Mouse Brain Replicates. Xenium demonstration dataset, 2023. URL `https://www.10xgenomics.com/datasets/fresh-frozen-mouse-brain-replicates-1-standard`. Released January 22, 2023. Xenium Onboard Analysis 1.0.2.

Cristina Almagro-Pérez, Andrew H. Song, Luca Weishaupt, Ahrong Kim, Guillaume Jaume, Drew F. K. Williamson, Konstantin Hemker, Ming Y. Lu, Kritika Singh, Bowen Chen, et al. AI-driven 3D spatial transcriptomics. *arXiv preprint arXiv:2502.17761*, 2025. doi: 10.48550/arXiv.2502.17761.

Alma Andersson, Ludvig Larsson, Linnea Stenbeck, Fredrik Salmén, Anna Ehinger, Sunny Z. Wu, et al. Spatial deconvolution of HER2-positive breast cancer delineates tumor-associated cell type interactions. *Nature Communications*, 12:6012, 2021. doi: 10.1038/s41467-021-26271-2.

Tommaso Biancalani, Gabriele Scalia, Lorenzo Buffoni, Raghav Avasthi, Ziqing Lu, Aman Sanger, Neriman Tokcan, Charles R. Vanderburg, Åsa Segerstolpe, Meng Zhang, et al. Deep learning and alignment of spatially resolved single-cell transcriptomes with Tangram. *Nature Methods*, 18: 1352–1362, 2021. doi: 10.1038/s41592-021-01264-7.

Guillaume Jaume, Paul Doucet, Andrew H. Song, Ming Y. Lu, Cristina Almagro-Pérez, Sophia J. Wagner, Anurag J. Vaidya, Richard J. Chen, Drew F. K. Williamson, Ahrong Kim, and Faisal Mahmood. HEST-1k: A dataset for spatial transcriptomics and histology image analysis. In *Advances in Neural Information Processing Systems*, volume 37, pp. 53798–53833, 2024. doi: 10.52202/079017-1704.

Andrew Jones, F. William Townes, Didong Li, and Barbara E. Engelhardt. Alignment of spatial genomics data using deep Gaussian processes. *Nature Methods*, 20:1379–1387, 2023. doi: 10.1038/s41592-023-01972-2.

Vitalii Kleshchevnikov, Artem Shmatko, Emma Dann, Alexander Aivazidis, Hamish W. King, Tong Li, Rasa Elmentaite, Artem Lomakin, Veronika Kedlian, Adam Gayoso, et al. Cell2location maps fine-grained cell types in spatial transcriptomics. *Nature Biotechnology*, 40:661–671, 2022. doi: 10.1038/s41587-021-01139-4.

Lei Li, Meina Lu, Lidong Guo, Xuejiao Zhang, Qun Liu, Meiling Zhang, Junying Gao, Mengyang Xu, Yijian Lu, Fang Zhang, et al. An organ-wide spatiotemporal transcriptomic and cellular atlas of the regenerating zebrafish heart. *Nature Communications*, 16:3716, 2025. doi: 10.1038/s41467-025-59070-0.

Senlin Lin, Zhikang Wang, Yan Cui, Qi Zou, Chuangyi Han, Rui Yan, Zhidong Yang, Wei Zhang, Rui Gao, Jiangning Song, Michael Q. Zhang, Hanchuan Peng, Jintai Yu, Jianfeng Feng, Yi Zhao, and Zhiyuan Yuan. Bridging the dimensional gap from planar spatial transcriptomics to 3D cell atlases. *Nature Methods*, 23:360–372, 2026. doi: 10.1038/s41592-025-02969-9.

Mohammad Lotfollahi, Mohsen Naghipourfar, Malte D. Luecken, Matin Khajavi, Maren Büttner, Marco Wagenstetter, Žiga Avsec, Adam Gayoso, Nir Yosef, Marta Interlandi, et al. Mapping

single-cell data to reference atlases by transfer learning. *Nature Biotechnology*, 40:121–130, 2022. doi: 10.1038/s41587-021-01001-7.

Liam McLaughlin, Bo Zhang, Siddharth Sharma, Amanda L. Knoten, Madhurima Kaushal, Jeffrey M. Purkerson, Heidie L. Huyck, Gloria S. Pryhuber, Joseph P. Gaut, and Sanjay Jain. Three dimensional multiscalar neurovascular nephron connectivity map of the human kidney across the lifespan. *Nature Communications*, 16:5161, 2025. doi: 10.1038/s41467-025-60435-8.

Tancredi Massimo Pentimalli, Simon Schallenberg, Daniel León-Periñán, Ivano Legnini, Ilan Theurillat, Gwendolin Thomas, Anastasiya Boltengagen, Sonja Fritzsche, Jose Nimo, Lukas Ruff, et al. Combining spatial transcriptomics and ECM imaging in 3D for mapping cellular interactions in the tumor microenvironment. *Cell Systems*, 16(5):101261, 2025. doi: 10.1016/j.cels.2025.101261.

Samuel G. Rodriques, Robert R. Stickels, Aleksandrina Goeva, Carly A. Martin, Evan Murray, Charles R. Vanderburg, Joshua Welch, Linlin M. Chen, Fei Chen, and Evan Z. Macosko. Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution. *Science*, 363(6434):1463–1467, 2019. doi: 10.1126/science.aaw1219.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, volume 9351 of *Lecture Notes in Computer Science*, pp. 234–241. Springer, 2015. doi: 10.1007/978-3-319-24574-4_28.

Mohammad Vali Sanian, Arshia Hemmat, Amirhossein Vahidi, Jonas Maaskola, Jimmy Tsz Hang Lee, Stanislaw Makarchuk, Yeliz Demirci, Nana-Jane Chipampe, Muzlifah Haniffa, Omer Bayraktar, Lassi Paavolainen, and Mohammad Lotfollahi. 3D-guided scalable flow matching for generating volumetric tissue spatial transcriptomics from serial histology. *arXiv preprint arXiv:2511.14613*, 2025. doi: 10.48550/arXiv.2511.14613.

Patrik L. Ståhl, Fredrik Salmén, Sanja Vickovic, Anna Lundmark, José Fernández Navarro, Jens Magnusson, Stefania Giacomello, Michaela Asp, Jakub O. Westholm, Mikael Huss, et al. Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science*, 353(6294): 78–82, 2016. doi: 10.1126/science.aaf2403.

Quanxin Wang, Song-Lin Ding, Yang Li, Josh Royall, David Feng, Phil Lesnar, Nile Graddis, Maitham Naeemi, Benjamin Facer, Anh Ho, et al. The Allen Mouse Brain Common Coordinate Framework: A 3D reference atlas. *Cell*, 181(4):936–953.e20, 2020. doi: 10.1016/j.cell.2020.04.007.

Sui Wang, Xinyu Wang, Qiangwei Peng, and Tiejun Li. STITCH: Spatial transcriptomics imputation via flow matching with internal learning. *bioRxiv*, 2026. doi: 10.64898/2026.06.03.729557.

Tianyi Xia, Luni Hu, Lulu Zuo, Lei Cao, Yunjia Zhang, Mengyang Xu, Qin Lu, Lei Zhang, Taotao Pan, Bohan Zhang, et al. ST-GEARS: Advancing 3D downstream research through accurate spatial information recovery. *Nature Communications*, 15:7806, 2024. doi: 10.1038/s41467-024-51935-0.

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

Ron Zeira, Max Land, Alexander Strzalkowski, and Benjamin J. Raphael. Alignment and integration of spatial transcriptomics data. *Nature Methods*, 19:567–575, 2022. doi: 10.1038/s41592-022-01459-6.

Meng Zhang, Xingjie Pan, Won Jung, Aaron R. Halpern, Stephen W. Eichhorn, Zhiyun Lei, Limor Cohen, Kimberly A. Smith, Bosiljka Tasic, Zizhen Yao, et al. Molecularly defined and spatially resolved cell atlas of the whole mouse brain. *Nature*, 624:343–354, 2023. doi: 10.1038/s41586-023-06808-9.

Mengting Zhao, Jiandong Zhou, Tao Jiang, Miao Ren, Chuhao Dou, Lingyi Cai, Shengda Bao, Xueyan Jia, Zhaoyang Meng, Feifang Tang, et al. Molecularly defined cellular atlas of the entire mouse brain with isotropic single-cell resolution. *Nature Communications*, 16:10167, 2025. doi: 10.1038/s41467-025-65238-5.

Appendix

Additional Experiments and Implementation Details

Overview. The appendix follows the main evaluation sequence: implementation, single-specimen reconstruction and baselines, cross-specimen validation, organ-specific studies, ablations, and downstream tasks. Registration and qualitative deployment are separated from quantitative prediction tests. The final sections address single-section ambiguity, transfer limits, and reproducibility. Within each study, we specify the reference, fitting split, prediction inputs, and evaluation unit; additional distance and panel sweeps are distinguished from prespecified endpoints.

Appendix contents

A Evaluation protocol and implementation
B Single-specimen reconstruction, predictive baselines, and molecular fields
C Four-animal composition and 60-mouse density validation
D Lung, breast, and kidney evaluation
E Reference-coverage and field-refinement ablations
F Sparse-panel downstream tasks
G Image-to-CCF registration and single-section deployment
H Identifiability and transfer limits
I Reproducibility record

A Evaluation Protocol and Implementation Details

A.1 Leakage barriers and evaluation units

Reference construction, development, and final testing use disjoint specimens whenever the dataset permits animal-level separation. Hidden test sections are excluded from reference construction, predictive pose estimation, feature selection, training normalization, expert fitting, router calibration, and model selection. Target measurements may define scoring masks and evaluation-only alignment where explicitly stated, as in HER2ST; these transforms and target values do not enter the predictor. A prediction episode consists of an observed center section, a signed physical offset chosen without inspecting the target, and a target section opened only for scoring. Grid cells and individual cells are never treated as biological replicates; the independent unit is an animal, donor, or complete held-out section.

For an external confirmatory claim we additionally require: (i) a frozen reference and router; (ii) all prespecified distances and spatial scales; (iii) central, Atlas-only, and naive-fusion baselines; (iv) correct-depth comparisons against reversed-, wrong-, or shuffled-reference controls; and (v) animal- or donor-level uncertainty. Hyperparameters may be changed after a failed development cohort, but that cohort remains labeled development and a new specimen is required for confirmation.

A.2 Population Atlas construction and soft queries

The mouse-brain Atlas uses four Allen Zhuang-ABCA specimens (Zhang et al., 2023). Cells are mapped to the Allen Common Coordinate Framework (Wang et al., 2020), rasterized on a 100-μm grid of size $133 \times 80 \times 60$, and smoothed with a support-normalized 100-μm kernel. Empty locations outside measured support are unknown rather than biological zeros. We compared support-normalized kernel estimation with Fourier and U-Net (Ronneberger et al., 2015) fields under four-fold leave-one-animal-out evaluation; the kernel field was selected before downstream use because neither neural alternative passed the joint density, abundance, and composition gate. The final frozen artifact contains 34-class density, population standard deviation, total density, and support. The final count-field predictor uses the center-derived calibration in Eq. 2; task-specific adaptations calibrate their own density units.

At inference, the Atlas is queried at the observed center and requested target surface. Table 1 uses the supplied CCF pose with linear interpolation along AP; no target-derived registration surface is

allowed. For downstream protocols that specify pose uncertainty, let $q(\phi \mid S_0)$ be a distribution over section-to-Atlas transformations. The corresponding mean query and variation are

$$\bar{\mathbf{A}}_d(\boldsymbol{u}) = \mathbb{E}_{\phi\sim q(\phi|S_0)}[\mathbf{A}(T_\phi(\boldsymbol{u}, d))], \qquad U_d(\boldsymbol{u}) = \mathrm{Var}_{\phi\sim q(\phi|S_0)}[\mathbf{A}(T_\phi(\boldsymbol{u}, d))]. \tag{8}$$

This marginalization is used only where specified by a downstream protocol; it does not contribute to the final Table 1 count-field predictions.

A.3 Observed-section calibration safeguards

Equation 2 uses only the observed section and its center-Atlas query. The common calibration mask requires observed coverage, center-Atlas support above 0.02, and observed total mass above 10^{-6}. The global scale g is bounded below by 10^{-6} and its denominator by 10^{-8}. A class ratio $r_k = n_k/a_k$ is used only when both class masses exceed 10^{-6}; otherwise $r_k = g$. Ratios are clipped to $[0.1g, 10g]$, then combined with g using the 0.75/0.25 log-scale shrinkage in the main equation. These numerical safeguards do not use a hidden target. Outputs retain native counts and are aggregated to the 400-μm scoring grid in Table 1; predicted total mass is not constrained to equal that of the observed section.

A.4 Metrics, aggregation, and uncertainty

Density Pearson measures spatial pattern fidelity; density NMAE is the sum of absolute voxel errors divided by total target mass. Slab-count Pearson evaluates whether integrated predicted mass tracks integrated target mass across hidden slabs. Composition Pearson is computed per coarse class over spatial bins and macro-averaged; JSD is computed between normalized class distributions and is lower-is-better. Aggregation follows the sampling unit of each study: BrainCAD distance curves average three outer folds, whereas the four-animal composition study and cell-typing study weight animals equally. The expression extension averages episodes within center, centers within distance, and then distances; its full protocol is given in Appendix B.7. Animal-level effects first average a method within each mouse over all reported distances and then compare methods, preventing animals with more valid voxels from dominating. Confidence intervals in the BrainCAD study use 10,000 bootstrap resamples of the 60 animal effects.

B Single-Specimen Reconstruction and Baselines

B.1 Atlas–test isolation

The composition Atlas is constructed only from Zhuang-ABCA-1–4 and reduced from 34 taxonomy classes to eight prespecified coarse classes. The test specimen, MERFISH-C57BL6J-638850, is a separate mouse and experiment with 3,739,961 cells joined one-to-one to official CCF coordinates and 53 scoreable coronal sections. It does not enter Atlas construction or distance-gate fitting. Inference abundance calibration uses only its observed center section. Target surfaces are obtained by adding the signed physical offset to the observed center surface; target cells and target-derived surfaces are scoring-only. This is independent-animal evaluation within the Allen ABC ecosystem, not a cross-institution test.

B.2 Fitting the final distance policy

For each of the four reference folds, we construct the Atlas without the supervised animal and generate center–target episodes at 100/200/300/500/700/1,000 μm. Candidate center planes are sampled every five 100-μm AP bins, using both offset signs. Calibration and target masks must each contain at least 32 supported bins. The target mask is used only for reference-animal supervision. Counts within each fold are divided by the 90th percentile of positive class-field values in that supervised reference animal before pooling squared-error statistics.

Writing the normalized target, individualized expert, and population expert as $\tilde{\mathbf{Y}}_e, \tilde{\mathbf{I}}_e, \tilde{\mathbf{F}}_e$, define $\mathbf{D}_e = \tilde{\mathbf{I}}_e - \tilde{\mathbf{F}}_e$. At distance d, the pooled mixture optimum is

$$w_d^* = \mathrm{clip}\left(\frac{\sum_{e:|d_e|=d}\langle \tilde{\mathbf{Y}}_e - \tilde{\mathbf{F}}_e, \mathbf{D}_e\rangle}{\max(\sum_{e:|d_e|=d}\|\mathbf{D}_e\|_2^2, 10^{-12})}, 0, 1\right).$$

We fit the two parameters in Eq. 6 by weighted least squares against w_d^*, with weights proportional to the denominator energy and normalized to mean one. The optimizer starts at $(b, a) = (3, 4)$ and allows 5,000 function evaluations. Inference clips the sigmoid logit to $[-30, 30]$ for numerical stability. No independent-mouse target enters these statistics or parameter fitting.

B.3 Dense-distance characterization and interpretation

The full-distance benchmark in Table 6 evaluates the frozen final predictor every 100 μm through 1 mm and at 2 mm. To prevent adjacent nominal distances from reusing a section pair, it uses non-overlapping $\pm$50-μm windows (49.999 μm in code), yielding 470 directed episodes. The four 200/400/1,000/2,000-μm distances were specified in the original study; the additional distances and matched baseline comparison are retrospective characterization. The 100-μm bin contains only five unique pairs. Exact pairing geometry is retained in the data record. The final predictor exceeds central-copy and Atlas-only through 1 mm and becomes numerically indistinguishable from the calibrated Atlas-only comparator at 2 mm. The lighter capped curve in Fig. 2 is a constrained comparator. It limits the Atlas composition update to 0.5 and anchors total mass to the observed center section. The final predictor does not constrain predicted total mass to equal the center's mass.

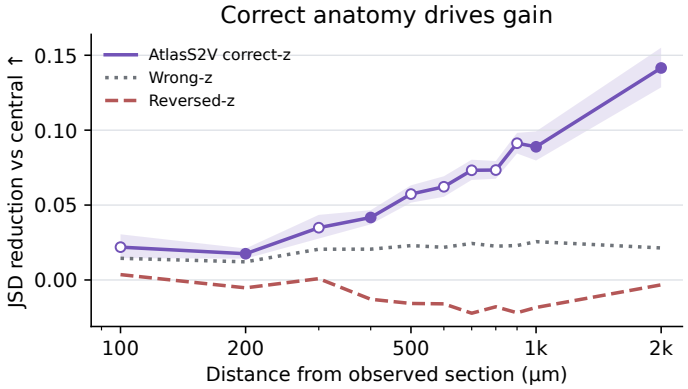


Figure 5: Depth-specific composition controls on C57BL6J-638850. Paired JSD reduction relative to central-copy is compared for correct, wrong, and reversed depth. The ribbon is the original 95% target-section block-bootstrap interval. Filled markers identify the prespecified 200/400/1,000/2,000-μm endpoints; open markers show post-hoc distances.

B.4 Total cell-density prediction

Density is the sum of the eight predicted class counts in each 400-μm in-plane bin. Table 5 uses the same frozen predictions, physical section pairs, and scoring masks as the composition evaluation. These correlations measure total-count agreement rather than class-wise spatial correlation or composition JSD. Gains are not uniform: central-copy is slightly stronger at 400 μm, nearest-reference retrieval is stronger at 1 mm, and the final predictor matches the calibrated Atlas at 2 mm.

B.5 Sixteen-layer visual comparison

Figure 6 collects section-wise comparisons for the same observed center as Fig. 2: C57BL6J-638850.37, at CCF AP 7.3356 mm. We select the eight acquired hidden sections nearest to the center on each side using physical distance alone. These 16 targets span approximately -1.66 to $+1.83$ mm, within the original depth-evaluation range. The negative-offset block appears above the positive-offset block; columns in each block increase in signed depth. Selection does not use prediction scores, and the observed center is an input rather than a hidden endpoint. The maps show excitatory-cell fractions on a fixed 200-μm in-plane grid, using one shared 0–1 color scale. Predictions use only the center section and reference Atlas; hidden measurements supply ground truth, display masks, and scoring.

To retain comparability with the original layer panel, each target's common display mask requires positive pooled target coverage and cell count and a pooled *sum* of Atlas support above 0.02. All four

Table 5: Total cell-density Pearson on held-out C57BL6J-638850. Counts are summed over the eight classes at the same 400-μm grid and on the same scoring masks as Table 6. U-Net reports the mean of three separately evaluated seeds. All 11 evaluated distances are shown.

$\lvert d\rvert$ (μm)	Central-copy	Calibrated Atlas	U-Net	Nearest ref.	ATLASS2V
100	0.4245	0.3299	0.2301	0.2676	0.4337
200	0.7414	0.5931	0.5228	0.5653	0.7516
300	0.4406	0.3869	0.2768	0.3476	0.4601
400	0.6235	0.5535	0.4768	0.5202	0.6165
500	0.4699	0.4974	0.3964	0.4632	0.5198
600	0.3621	0.4518	0.3475	0.4336	0.4644
700	0.4599	0.5329	0.4332	0.4838	0.5383
800	0.3239	0.4132	0.2991	0.3817	0.4152
900	0.4246	0.5589	0.4602	0.4791	0.5597
1,000	0.2224	0.3446	0.2390	0.4080	0.3448
2,000	0.3407	0.5116	0.3978	0.4615	0.5116

methods share that mask, which is slightly less restrictive than the pooled-mean threshold in the main 3D rendering. Pearson badges quantify spatial correlation for the displayed excitatory fraction to four decimals, not the eight-class aggregate reported in Table 1. No model refitting, target-informed tuning, or interpolation of ground-truth layers is performed. The complete 52-target intermediate inference record remains in the experiment archive; only this geometrically selected 16-layer figure is included in the paper.

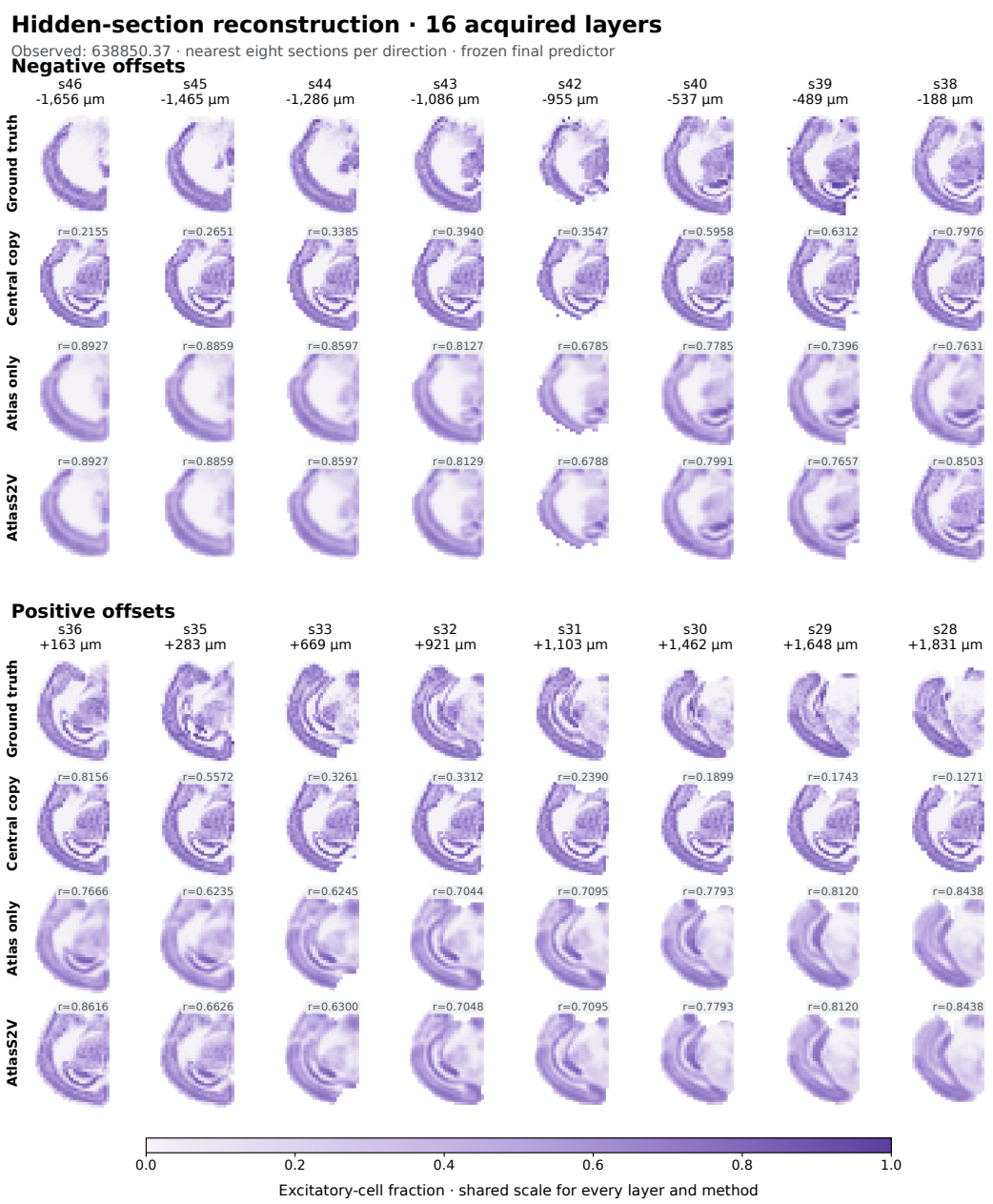


Figure 6: Sixteen acquired hidden sections around C57BL6J-638850.37. The eight nearest negative-offset targets (top) and eight nearest positive-offset targets (bottom) are sorted by signed depth, without performance-based selection. Within each block, rows compare ground truth, central-copy, calibrated Atlas-only, and the frozen final predictor. All maps share a 200-μm grid, 0–1 excitatory-fraction scale, and common per-target display mask. Gray badges show spatial Pearson for this one displayed class to four decimals, not the eight-class aggregate in Table 1.

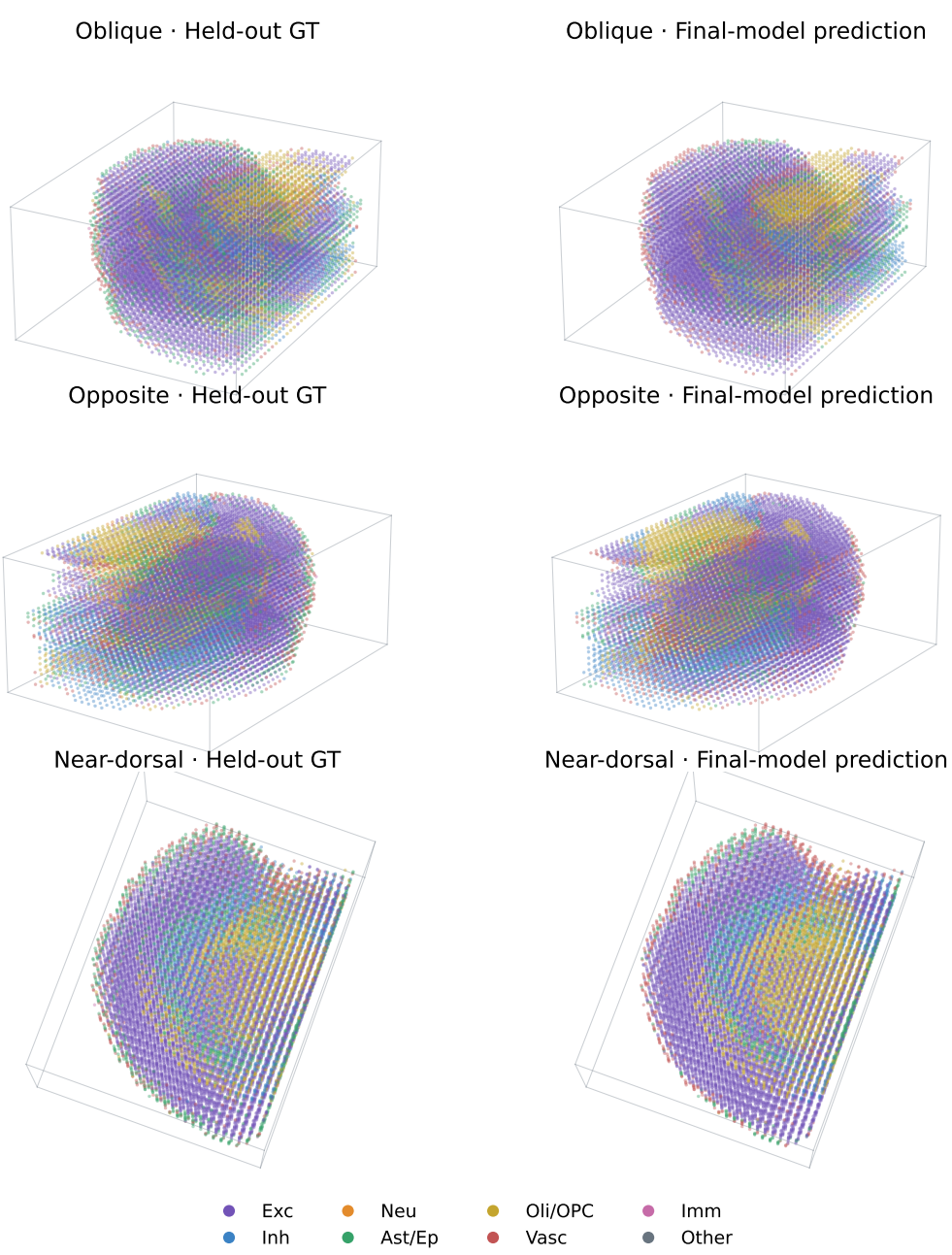


Figure 7: Three views of the same independent-mouse cell-type field. Each row uses one camera for held-out MERFISH ground truth (left) and the final frozen Table 1 predictor (right), centered on the observed section C57BL6J-638850.37 at AP 7.34 mm. The scene contains that observed center and 18 acquired sections within approximately ± 2 mm; prediction uses only the center and reference Atlas, while target sections define ground truth and the common display mask. Each translucent point is a 200-μm planar bin, colored by its dominant one of eight coarse classes, not a reconstructed single-cell position. The shared mask permits visual comparison but selects bins with available target measurements; this illustration is not an additional quantitative endpoint.

B.6 Matched Predictive Baselines

B.6.1 Shared evaluation protocol

Table 1 compares the final predictor from Section 3 with central-copy, Atlas-only, a residual U-Net, and nearest-reference retrieval. All methods use the same four reference animals, all 470 directed pairs, a 49.999-μm pairing tolerance, center-derived target query surfaces, and identical 400-μm scoring masks. Native count predictions are scored without central-occupancy projection. The final ATLASS2V policy is fitted by reference-animal cross-validation as specified in Appendix B.2, then frozen before independent-mouse evaluation.

For the learned baselines, ABCA-1–3 supply training examples and ABCA-4 selects hyperparameters. The final U-Nets are reinitialized and fitted on all four reference animals, excluding each supervised animal from its Atlas. All 11 evaluation distances, including 2 mm, are represented in baseline development and refitting; the ATLASS2V gate uses the six training distances specified in Appendix B.2. No external target enters selection. Checkpoints are serialized and independently reloaded before external scoring. This is a retrospective matched benchmark on C57BL6J-638850, not a new biological cohort.

B.6.2 Single-section baselines

Direct residual U-Net. A two-level encoder–decoder with widths 16/32/64, GroupNorm, and bilinear skip decoding receives the observed, Atlas-center, and Atlas-target eight-channel log-density maps and a signed-distance channel. It predicts a learned log residual to the target Atlas without the analytic center-minus-Atlas transfer or a routing gate. The masked loss combines log-density MSE and composition JSD. Training uses AdamW, learning rate 10^{-3}, weight decay 10^{-4}, batch size eight, and at most 30 development epochs. Seeds 20260914, 20260915, and 20260916 select 4, 6, and 3 epochs, respectively, before refitting on all reference animals.

Nearest-reference retrieval. We compare the observed section with all four reference animals at AP shifts of $-0.2, -0.1, 0, 0.1, 0.2$ mm using center-calibrated log-MSE, and transfer the selected animal's field at the corresponding target depth. Retrieval uses only the observed section for matching. An empty observed input gives an empty prediction.

B.6.3 Results and reporting

Tables 1 and 6 report the mean of three separately evaluated U-Net seeds, not an ensemble prediction. Among the reported comparators, ATLASS2V has the highest class Pearson at 100–900 μm and the lowest JSD at every distance through 1 mm. Retrieval has higher Pearson at 1 mm; distant predictions converge toward the Atlas. The independent biological sample size is one mouse. Overlapping section pairs and optimization seeds are not biological replicates. Appendix C.1 separately evaluates the final model structure with a newly fitted, test-animal-excluded predictor in each of four outer folds.

B.6.4 Publication record

The records `tables/table1_matched_benchmark_data.json` and `tables/table1_matched_benchmark_seed_summary.json` retain the reported metrics, seed variability, pairing geometry, and checkpoint identity. The table-generation scripts reconstruct the comparisons without fitting; unexecuted and privileged-input comparators are documented in Appendix I.

Table 6: Dense-distance prediction and matched baselines on independent mouse C57BL6J-638850 at a 400-μm XY grid. All methods use the same four reference animals, directed pairs, and scoring masks; outputs retain native counts. Columns show nominal target distances in μm, using non-overlapping $\pm$50-μm pairing windows. ATLASS2V is the calibrated count-field predictor defined in Section 3. U-Net entries average three seeds; baseline development and calibration use reference animals and the observed section only (Appendix B.6). Bold marks the best unrounded value at each distance within each metric; blue identifies ATLASS2V. Distances other than 200/400/1,000/2,000 μm are post hoc; the 100-μm bin contains five unique pairs.

Method / Distance (μm)	100	200	300	400	500	600	700	800	900	1,000	2,000
Class Pearson ↑. Spatial agreement of cell-type distributions.											
Central-copy	0.521	0.648	0.437	0.513	0.361	0.339	0.358	0.286	0.257	0.220	0.124
Atlas-only	0.529	0.649	0.476	0.604	0.610	0.606	0.599	0.538	0.621	0.489	0.553
U-Net	0.435	0.547	0.393	0.518	0.449	0.499	0.514	0.454	0.518	0.411	0.461
Nearest-reference	0.504	0.637	0.452	0.602	0.596	0.584	0.542	0.501	0.587	**0.514**	0.543
ATLASS2V	**0.575**	**0.720**	**0.511**	**0.631**	**0.619**	**0.613**	**0.601**	**0.539**	**0.621**	0.489	**0.553**
JSD ↓. Difference in cell-type composition.											
Central-copy	0.0635	0.0524	0.0752	0.0787	0.0957	0.1047	0.1138	0.1132	0.1303	0.1336	0.1787
Atlas-only	0.0455	0.0388	0.0423	0.0380	0.0391	0.0430	0.0408	0.0399	0.0390	0.0448	**0.0372**
U-Net	0.0453	0.0370	0.0416	0.0378	0.0393	0.0429	0.0411	0.0406	0.0397	0.0453	0.0385
Nearest-reference	0.0485	0.0412	0.0459	0.0411	0.0423	0.0471	0.0445	0.0432	0.0441	0.0493	0.0406
ATLASS2V	**0.0416**	**0.0349**	**0.0404**	**0.0370**	**0.0384**	**0.0424**	**0.0406**	**0.0398**	**0.0390**	**0.0448**	0.0372

B.7 HIDDEN-LAYER GENE EXPRESSION RECONSTRUCTION

This section documents the expression extension in Section 4.2.1. Table 7 compares 10 model/control variants. Table 8 resolves seven predictive alternatives at all 11 distances; individual-seed and unscaled-model detail remains in the complete record. Table 9 reports five methods with archived module-level scores. Figure 8 contrasts correlation and absolute error. The publication records retain all group memberships and distance-resolved values, including every computational repetition.

B.7.1 SCOPE, MEASURED TARGETS, AND ISOLATION

This exploratory extension predicts measured expression on hidden physical sections, rather than imputing genes on the observed section. It uses 477 unique measured genes shared by the four Zhuang-ABCA references and C57BL6J-638850, and eight previously frozen coexpression modules. All 477 genes are available on the single observed section; this is not a low-panel experiment. The predictor is trained from scratch for expression and does not load the composition checkpoint in Table 1. BrainCAD density fitting is also separate, with no shared fitted weights or interchangeable biological endpoints.

Development supervision uses ABCA-1, -3, and -4, with the whole ABCA-2 animal reserved for epoch and gate-scale selection. Each development Atlas excludes both the supervised animal and ABCA-2. The expression standardizer is fitted on development animals only. The four references contain respectively 129, 54, 22, and 3 native sections; the last two specimens are sagittal, and pairing respects each specimen's native section normal. Development, validation, and final four-animal leave-one-animal-out refitting contain respectively 364, 64, and 428 directed section pairs (54,799, 8,014, and 62,813 scored grid rows). In final refitting, each supervised animal is excluded from its episode's Atlas. The test animal enters neither reference construction nor fitting, epoch selection, or gate-scale selection.

The external evaluation uses the same previously examined C57BL6J-638850 specimen: 53 scoreable native sections and 470 directed center–target pairs. Thus this is an exploratory re-evaluation, not a new confirmatory animal. Targets are actual measured sections nearest to offsets of 100, 200, ..., 1,000, and 2,000 μm, with a tolerance strictly below 50 μm; both directions are retained. Ground-truth expression is not interpolated between sections. The single observed section supplies the CCF pose; the target query is the observed surface shifted by the requested offset, not a target-derived registration surface.

B.7.2 MOLECULAR ATLAS AND STRICT RESIDUAL ROUTING

Per-cell expression is normalized to 1,000 total counts across the common genes, averaged within native 400-by-400-μm in-plane bins, transformed with $\log(1+x)$, and standardized with development-only gene statistics. This grid does not turn a physical section into a 400-μm-thick target slab. Molecular Atlases use 400-μm CCF cubes, support-normalized Gaussian smoothing with standard deviation 0.75 voxel, and trilinear sampling. An observed-section per-gene mean-bias calibration is applied identically to center and target Atlas queries. Unsupported molecular queries fall back to the development mean. No hidden target expression is an input.

Let Y_0 denote observed expression, A_0 the calibrated center Atlas, and A_d its target query. The individualized branch and strict mixture are

$$I_d = \max\{A_d + Y_0 - A_0 + \min(|d|/(1\,\mathrm{mm}), 1) f_\theta(\xi_d),\ \ell\}\,, \tag{9}$$

$$w_g(d) = \exp\left[-|d|/\left(\mathrm{softplus}(\tau_g) + 10^{-4}\right)\right], \tag{10}$$

$$\widehat{Y}_d = A_d + \lambda w(d) \odot (I_d - A_d), \tag{11}$$

where distances in the gate are in millimeters, ℓ is the gene-wise standardized physical zero floor, and $\lambda \in [0, 1]$ is selected using the reference validation animal. The correction receives Y_0, A_0, A_d, a six-neighbor Atlas context at ± 400 μm, and CCF position, signed/absolute offset, section-axis, and support features. A shared 477-to-64 encoder and a two-hidden-layer, width-128 MLP produce the 477-gene correction (approximately 143,000 parameters including the gate).

Every individualized correction lies inside the gated branch. Setting the effective gate to zero returns exactly the same calibrated Atlas used as a baseline, regardless of the MLP output. Only I_d is floor-clipped: clipping the final mixture would change that Atlas endpoint. Calibrated standardized log-expression scores may consequently lie below the standardized physical floor and are not raw nonnegative transcript counts. Relative to the earlier ungated-correction comparator, both the correction's gate placement and the final clipping semantics change; the comparison is not a one-factor architectural ablation. With $\lambda < 1$, the model does not enforce exact reconstruction at zero offset; no zero-offset pair is scored here.

B.7.3 REFERENCE SELECTION, REFITTING, AND COMPLETE CONTROLS

Two computational seeds (20260921 and 20260922) are initialized from scratch, using AdamW, learning rate 10^{-3}, weight decay 10^{-4}, batch size 1,024, and at most 40 epochs. ABCA-2 selects epochs 39 and 38 by episode-equal 477-gene MSE. On their mean validation prediction, the five candidate gate scales $\{0, 0.25, 0.5, 0.75, 1\}$ yield MSEs $\{0.743604, 0.717485, 0.710875, 0.723774, 0.756183\}$, selecting $\lambda = 0.5$. Each seed is then reinitialized and refitted on four-animal leave-one-animal-out episodes for its selected number of epochs. The gate scale is held fixed and the two predictions are averaged for external scoring. These seeds are not independent biological replicates.

Including $\lambda = 0$ guarantees only that the chosen candidate is no worse than the Atlas on the selection set and selection objective. It does not guarantee improvement after refitting, on an unseen animal, at every depth, or under another metric. The unscaled strict ensemble is therefore reported alongside the selected version. Section-only and Atlas-conditioned MLPs, uncalibrated and calibrated Atlases, the earlier ungated correction, and both unscaled strict seeds are retained in Table 7.

All methods use the same scoring mask: at least 20 cells in both observed and hidden bins and the previously frozen composition-Atlas support. This evaluates expression within shared support, not the appearance of tissue outside the observed-section footprint. Pearson is computed spatially per gene or module, then averaged over the 477 genes or eight modules. Episode scores are averaged within each observed center, across centers within each distance, and finally equally across the 11 distances. The 470 overlapping directed pairs are not 470 biological replicates, and no population-level confidence interval or significance claim is assigned to this one-mouse study. Module definitions were frozen in the earlier reference analysis (ABCA-1–3), and are used only for evaluation, not the new loss or model selection; their error scale differs from standardized gene expression.

ATLAS2V improves average gene/module Pearson over the calibrated Atlas (0.6272/0.7680 versus 0.6141/0.7558). Its module RMSE is also worse than the calibrated Atlas (0.2526 versus 0.2490). Restricting aggregation to 200, 400, 1,000, and 2,000 μm gives gene/module Pearson 0.6266/0.7643

Table 7: Hidden-layer expression: distance-balanced external results on one mouse. All methods share 470 directed pairs and scoring masks. Gene errors are in development-standardized log-expression units; module errors use the frozen module units. Seeds are computational repetitions, not biological replicates.

	477 measured genes			8 frozen modules		
Method	Pearson ↑	RMSE ↓	MAE ↓	Pearson ↑	RMSE ↓	MAE ↓
Central copy	0.4139	0.8910	0.6276	0.5379	0.3767	0.2699
Uncalibrated Atlas	0.6141	0.8757	0.6965	0.7558	0.3470	0.2904
Calibrated Atlas	0.6141	0.6700	0.5112	0.7558	0.2490	0.1908
Section-only MLP	0.3989	0.9257	0.7089	0.5562	0.4703	0.3957
Atlas-conditioned MLP	0.5154	0.8689	0.6701	0.7158	0.4231	0.3614
Ungated-correction router	0.5925	0.7512	0.5774	0.7394	0.3408	0.2814
Unscaled strict, seed 1	0.6027	0.7446	0.5718	0.7384	0.3124	0.2530
Unscaled strict, seed 2	0.6044	0.7193	0.5504	0.7486	0.2785	0.2208
Unscaled strict, ensemble	0.6145	0.7113	0.5443	0.7545	0.2852	0.2282
ATLASS2V	0.6272	0.6676	0.5072	0.7680	0.2526	0.1952

for ATLASS2V and 0.6110/0.7501 for the calibrated Atlas. Neither aggregation establishes that a neural correction is necessary or that this model is uniformly best.

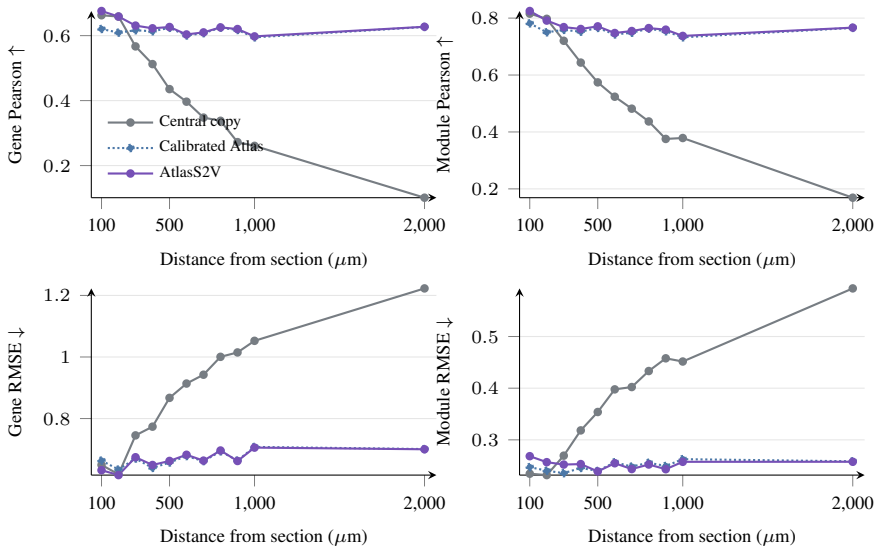


Figure 8: Expression controls across depth: central-copy, the calibrated molecular Atlas, and ATLASS2V. Pearson gains do not imply uniform gains in absolute error. Curves use the same measured targets, masks, and distance aggregation as Figure 2b. Gene and module RMSE have different units.

Table 8: Distance-resolved expression controls for seven predictive alternatives. Offsets match measured hidden sections, not interpolated ground truth. Unscaled-model and individual-seed summaries appear in Table 7; the full summaries at every distance remain in the accompanying JSON.

$\lvert d\rvert$ (μm)	Method	Gene r	RMSE	MAE	Module r	RMSE	MAE
100	Central copy	0.6627	0.6493	0.4468	0.8156	0.2345	0.1676
100	Uncalibrated Atlas	0.6205	0.8869	0.6952	0.7816	0.3488	0.2901
100	Calibrated Atlas	0.6205	0.6635	0.4979	0.7816	0.2472	0.1873
100	Section-only MLP	0.6467	0.6763	0.4879	0.8138	0.2788	0.2194
100	Atlas-conditioned MLP	0.6621	0.6730	0.4886	0.8301	0.2887	0.2317
100	Ungated-correction router	0.7011	0.5888	0.4252	0.8503	0.2286	0.1768
100	ATLASS2V	0.6758	0.6322	0.4637	0.8249	0.2684	0.2126
200	Central copy	0.6588	0.6188	0.4288	0.7975	0.2318	0.1589
200	Uncalibrated Atlas	0.6093	0.8633	0.6903	0.7501	0.3424	0.2864
200	Calibrated Atlas	0.6093	0.6344	0.4817	0.7501	0.2403	0.1826
200	Section-only MLP	0.6321	0.6590	0.4868	0.7868	0.2952	0.2364
200	Atlas-conditioned MLP	0.6554	0.6459	0.4792	0.8160	0.2918	0.2355
200	Ungated-correction router	0.6946	0.5730	0.4270	0.8330	0.2277	0.1769

$\lvert d \rvert$ (μm)	Method	Gene r	RMSE	MAE	Module r	RMSE	MAE
200	ATLASS2V	0.6588	0.6163	0.4628	0.7913	0.2568	0.2009
300	Central copy	0.5669	0.7453	0.5182	0.7203	0.2694	0.1877
300	Uncalibrated Atlas	0.6166	0.8851	0.6976	0.7587	0.3414	0.2851
300	Calibrated Atlas	0.6166	0.6700	0.5069	0.7587	0.2355	0.1785
300	Section-only MLP	0.5311	0.8009	0.5996	0.7067	0.3791	0.3165
300	Atlas-conditioned MLP	0.5838	0.7746	0.5846	0.7728	0.3707	0.3144
300	Ungated-correction router	0.6404	0.6837	0.5145	0.7861	0.2874	0.2325
300	ATLASS2V	0.6309	0.6743	0.5045	0.7684	0.2522	0.1932
400	Central copy	0.5124	0.7736	0.5349	0.6437	0.3183	0.2184
400	Uncalibrated Atlas	0.6134	0.8623	0.6902	0.7523	0.3453	0.2889
400	Calibrated Atlas	0.6134	0.6396	0.4865	0.7523	0.2456	0.1873
400	Section-only MLP	0.4718	0.8276	0.6258	0.6264	0.4118	0.3420
400	Atlas-conditioned MLP	0.5499	0.7838	0.5973	0.7398	0.3800	0.3193
400	Ungated-correction router	0.6139	0.6963	0.5332	0.7622	0.3122	0.2552
400	ATLASS2V	0.6226	0.6491	0.4918	0.7618	0.2531	0.1948
500	Central copy	0.4354	0.8672	0.6020	0.5742	0.3539	0.2472
500	Uncalibrated Atlas	0.6240	0.8678	0.6896	0.7655	0.3423	0.2874
500	Calibrated Atlas	0.6240	0.6572	0.4997	0.7655	0.2399	0.1827
500	Section-only MLP	0.3989	0.9155	0.6968	0.5721	0.4625	0.3905
500	Atlas-conditioned MLP	0.5111	0.8563	0.6583	0.7235	0.4251	0.3666
500	Ungated-correction router	0.5907	0.7477	0.5741	0.7442	0.3419	0.2844
500	ATLASS2V	0.6266	0.6623	0.5019	0.7710	0.2389	0.1823
600	Central copy	0.3967	0.9137	0.6403	0.5238	0.3977	0.2786
600	Uncalibrated Atlas	0.6004	0.8765	0.6982	0.7426	0.3543	0.2960
600	Calibrated Atlas	0.6004	0.6781	0.5192	0.7426	0.2566	0.1965
600	Section-only MLP	0.3534	0.9666	0.7445	0.5023	0.5110	0.4334
600	Atlas-conditioned MLP	0.4770	0.8976	0.6962	0.6951	0.4542	0.3897
600	Ungated-correction router	0.5581	0.7834	0.6060	0.7156	0.3655	0.3039
600	ATLASS2V	0.6037	0.6823	0.5219	0.7476	0.2547	0.1950
700	Central copy	0.3477	0.9424	0.6625	0.4818	0.4022	0.2855
700	Uncalibrated Atlas	0.6076	0.8682	0.6940	0.7486	0.3453	0.2905
700	Calibrated Atlas	0.6076	0.6617	0.5068	0.7486	0.2483	0.1910
700	Section-only MLP	0.3168	0.9917	0.7686	0.4801	0.5177	0.4409
700	Atlas-conditioned MLP	0.4658	0.9239	0.7228	0.6861	0.4627	0.4020
700	Ungated-correction router	0.5519	0.8019	0.6258	0.7070	0.3783	0.3179
700	ATLASS2V	0.6100	0.6633	0.5079	0.7546	0.2435	0.1877
800	Central copy	0.3379	1.0005	0.7008	0.4369	0.4332	0.3119
800	Uncalibrated Atlas	0.6245	0.8994	0.7130	0.7628	0.3556	0.2965
800	Calibrated Atlas	0.6245	0.6934	0.5281	0.7628	0.2557	0.1962
800	Section-only MLP	0.3078	1.0647	0.8236	0.4544	0.5652	0.4829
800	Atlas-conditioned MLP	0.4625	0.9875	0.7697	0.6582	0.4938	0.4265
800	Ungated-correction router	0.5443	0.8738	0.6794	0.6770	0.4348	0.3678
800	ATLASS2V	0.6254	0.6962	0.5305	0.7649	0.2523	0.1932
900	Central copy	0.2722	1.0148	0.7186	0.3755	0.4578	0.3303
900	Uncalibrated Atlas	0.6174	0.8632	0.6897	0.7536	0.3455	0.2903
900	Calibrated Atlas	0.6174	0.6637	0.5077	0.7536	0.2494	0.1921
900	Section-only MLP	0.2689	1.0571	0.8256	0.4159	0.5717	0.4878
900	Atlas-conditioned MLP	0.4508	0.9673	0.7604	0.6681	0.4914	0.4272
900	Ungated-correction router	0.5418	0.8224	0.6429	0.6907	0.3938	0.3312
900	ATLASS2V	0.6203	0.6621	0.5072	0.7594	0.2432	0.1873
1,000	Central copy	0.2607	1.0525	0.7558	0.3787	0.4516	0.3304
1,000	Uncalibrated Atlas	0.5947	0.8915	0.7108	0.7325	0.3517	0.2941
1,000	Calibrated Atlas	0.5947	0.7078	0.5465	0.7325	0.2626	0.2040
1,000	Section-only MLP	0.2463	1.1095	0.8703	0.3996	0.5933	0.5084
1,000	Atlas-conditioned MLP	0.4223	1.0515	0.8313	0.6245	0.5141	0.4435
1,000	Ungated-correction router	0.5110	0.8975	0.7063	0.6611	0.4204	0.3542
1,000	ATLASS2V	0.5977	0.7056	0.5450	0.7374	0.2575	0.1999
2,000	Central copy	0.1011	1.2228	0.8946	0.1690	0.5931	0.4527
2,000	Uncalibrated Atlas	0.6267	0.8689	0.6929	0.7656	0.3440	0.2887
2,000	Calibrated Atlas	0.6267	0.7009	0.5424	0.7656	0.2585	0.2005
2,000	Section-only MLP	0.2140	1.1136	0.8678	0.3597	0.5864	0.4945
2,000	Atlas-conditioned MLP	0.4292	0.9969	0.7825	0.6594	0.4810	0.4189
2,000	Ungated-correction router	0.5701	0.7943	0.6173	0.7063	0.3580	0.2942
2,000	ATLASS2V	0.6273	0.7002	0.5418	0.7665	0.2576	0.1998

B.7.4 INDIVIDUAL COEXPRESSION MODULES

The eight coexpression groups were fixed using reference data before this expression experiment. Their IDs do not imply independently validated pathways. Table 9 applies the same equal-distance aggregation to each module separately; five methods with archived module-specific results are displayed. Their distance-resolved records remain in `figures/figure2_expression_data.json`.

Table 9: Individual frozen coexpression modules: equal-distance means over all 11 hidden-layer offsets. Module IDs are fixed reference-derived groups, not named biological pathways. Five methods with archived module-level scores are shown; aggregate module endpoints for all methods appear in Table 7.

Module	Method	Pearson ↑	RMSE ↓	MAE ↓
1	Calibrated Atlas	0.6329	0.2213	0.1723
1	Unscaled strict, seed 1	0.6067	0.3123	0.2629
1	Unscaled strict, seed 2	0.6386	0.2816	0.2360
1	Unscaled strict, ensemble	0.6421	0.2907	0.2448
1	ATLASS2V	0.6568	0.2388	0.1938
2	Calibrated Atlas	0.7057	0.2403	0.1957
2	Unscaled strict, seed 1	0.5933	0.3143	0.2631
2	Unscaled strict, seed 2	0.6526	0.2919	0.2426
2	Unscaled strict, ensemble	0.6349	0.2818	0.2344
2	ATLASS2V	0.6844	0.2494	0.2053
3	Calibrated Atlas	0.7066	0.2100	0.1629
3	Unscaled strict, seed 1	0.6766	0.2875	0.2458
3	Unscaled strict, seed 2	0.6823	0.2385	0.1908
3	Unscaled strict, ensemble	0.6932	0.2575	0.2141
3	ATLASS2V	0.7130	0.2214	0.1761
4	Calibrated Atlas	0.8546	0.2322	0.1718
4	Unscaled strict, seed 1	0.8646	0.2484	0.1892
4	Unscaled strict, seed 2	0.8653	0.2656	0.2075
4	Unscaled strict, ensemble	0.8698	0.2408	0.1828
4	ATLASS2V	0.8701	0.2251	0.1679
5	Calibrated Atlas	0.7574	0.2559	0.1942
5	Unscaled strict, seed 1	0.7459	0.2998	0.2404
5	Unscaled strict, seed 2	0.7324	0.2628	0.2060
5	Unscaled strict, ensemble	0.7527	0.2673	0.2125
5	ATLASS2V	0.7685	0.2480	0.1908
6	Calibrated Atlas	0.8278	0.2179	0.1670
6	Unscaled strict, seed 1	0.8430	0.2691	0.2237
6	Unscaled strict, seed 2	0.8375	0.2249	0.1743
6	Unscaled strict, ensemble	0.8476	0.2391	0.1917
6	ATLASS2V	0.8497	0.2148	0.1682
7	Calibrated Atlas	0.8373	0.2679	0.2048
7	Unscaled strict, seed 1	0.8506	0.3281	0.2573
7	Unscaled strict, seed 2	0.8521	0.2938	0.2303
7	Unscaled strict, ensemble	0.8573	0.3077	0.2413
7	ATLASS2V	0.8565	0.2743	0.2076
8	Calibrated Atlas	0.7244	0.3468	0.2576
8	Unscaled strict, seed 1	0.7265	0.4396	0.3416
8	Unscaled strict, seed 2	0.7280	0.3693	0.2789
8	Unscaled strict, ensemble	0.7381	0.3964	0.3038
8	ATLASS2V	0.7449	0.3489	0.2516

B.7.5 REFERENCE-RANKED GENE-LEVEL RESULTS

Table 2 reports quantitative results for two reference-defined gene groups. Rankings use only ABCA-1, -3, and -4, excluding the validation animal ABCA-2 and external mouse 638850. Within each development mouse, we retain native 400-μm bins with at least 20 cells. For the highly variable group, we calculate across-bin variance of $\log(1+\text{mean CP1000})$ before gene standardization, then average these variances equally across the three mice. This is a spatial-variance ranking, not a mean-adjusted single-cell HVG algorithm. For the highly expressed group, we calculate across-bin mean CP1000 within each mouse and then average equally across mice. We define each full group as the top 25% of its independent ranking, rounding up to 120 of 477 genes and breaking ties by gene name; the five highest-ranked members are displayed as examples. Overlap between groups is allowed rather than suppressing a highly ranked gene. Neither ranking uses external predictions, errors, or improvement over a baseline. The 25% cutoff was introduced for this exploratory presentation, not selected by test performance or prospectively registered; the groups refer only to this measured panel, not all genes in the transcriptome.

Each gene row in Table 2 reports ATLASS2V spatial Pearson separately at 200, 400, 600, 800, and 2,000 μm, with episodes averaged within observed center and then centers weighted equally at each distance. The final row in each group, "Mean (all 120)", is the arithmetic mean of all 120 members' unrounded gene scores at that distance, not just the five displayed examples or a pooled correlation across expression values. Values are shown to three decimals; bold and blue shading identify group means. Complete methods and absolute-error metrics remain in Tables 7 and 8. The publication record retains all rankings and unrounded gene-level scores.

B.7.6 IMPLEMENTATION CHECKS AND PUBLICATION RECORD

For all 470 external episodes, forcing the effective gate to zero reproduces the calibrated Atlas exactly (maximum elementwise difference zero). Ten implementation tests check gate endpoints, interpolation bounds, gating of the full correction, gradients, distant fallback, and reference-only candidate selection. Reaggregated metrics agree with the frozen summaries to numerical precision; calibrated-Atlas scores agree with the preceding experiment within 9×10^{-17}. These checks establish implementation consistency, not new biological validation.

The compact publication record `figures/figure2_expression_data.json` contains all signed/absolute-distance summaries, every module's distance-resolved metrics, source hashes, selected gate scale, checkpoint manifests, and audits. The record `figures/figure2_gene_table.json` stores reference-only ranking definitions and scores, all 40 gene-level Pearson values and eight group means, aggregation rules, and source hashes. The experiment archive additionally preserves all per-gene metrics, module predictions, and maps for every frozen module at native section 638850.37. A new independent expression animal is needed to test the robustness of this exploratory extension.

C CROSS-SPECIMEN VALIDATION

The composition and marker-density studies test different outputs and use separate fitting protocols. The four-animal study refits the final count-field structure; BrainCAD evaluates a marker-specific density adaptation.

C.1 STRICT OUTER-ANIMAL COMPOSITION VALIDATION

Nested fitting and isolation. We repeat the final model structure in four outer leave-one-animal-out folds on Zhuang-ABCA-1–4. The outer test animal is excluded from the population atlas, distance-policy fitting, and parameter selection. Each of the other three animals in turn supplies reference supervision against an atlas built from the remaining two. The six training distances, support-normalized 100-μm reference fields, class calibration, and two-parameter monotone fitting rule are unchanged from Appendix B.2. We then construct the outer atlas from all three reference animals and serialize a separately fitted predictor before evaluating the test animal. The four-animal checkpoint used in Table 1 is never reused for these outer tests. Calibration constants are fixed; no test outcome selects a setting. The fit is deterministic, so optimization seeds are not introduced as extra replicates.

Native sections and target queries. Test inputs and targets are raw eight-class counts from distinct physical sections. A center section is built only from its own cells, without three-dimensional smoothing across the test animal's sections. This differs from slicing a smoothed held-out volume, which could introduce neighboring cells into the input. ABCA-1/2 contain 129/54 coronal sections, and ABCA-3/4 contain 22/3 sagittal sections. We retain the true cutting direction: AP-normal queries for coronal sections and ML-normal queries for sagittal sections. Official CCF coordinates provide registration. Each target query translates the center-derived surface by the signed difference in median section coordinates; the target's local surface, coverage, and cell composition are not predictor inputs. Target coverage enters scoring only. This evaluates composition prediction conditional on registration, not automatic image-to-CCF accuracy.

Pairing, metrics, and complete results. For every center and both directions, we select the nearest measured target to each nominal distance from 100 to 1,000 μm in 100-μm steps, plus 2,000 μm,

accepting at most 49.999 μm mismatch. Center calibration and target evaluation each require at least 32 supported 100-μm bins. All methods use the same 400-μm pooled scoring mask and the Table 1 metric implementation. Counts remain unconstrained by central occupancy. ABCA-1–4 contribute 2,451, 528, 182, and 4 directed predictions, respectively (1,331, 266, 91, and 2 unique unordered section pairs). ABCA-4 supports only the 200- and 900-μm bins under this pairing rule; absent distances remain missing. Table 10 reports all animals and nominal distances, including those gaps.

At the two distances supported by every animal, we first average distance-level metrics within animal and then average animals equally. ATLASS2V improves all three metrics over central-copy in every animal: mean Pearson increases from 0.3922 to 0.6288, composition JSD decreases from 0.1073 to 0.0399, and NMAE decreases from 0.7200 to 0.5024. Against the calibrated Atlas, Pearson improves in three of four animals and NMAE in all four. Their four-animal means change from 0.6203 to 0.6288 and from 0.5205 to 0.5024, respectively, demonstrating gains in spatial class-count prediction across the outer folds. Table 10 gives the complete Pearson, JSD, and NMAE results for all three methods. The outer folds share reference animals, the section pairs overlap, and ABCA-4 is sparsely sampled; the biological sample size is four, not 3,165. ABCA data informed earlier development, so this is retrospective cross-validation with strict computational isolation rather than a new blinded cohort. The separately held-out mouse in Table 1 remains a distinct evaluation.

Source records. The frozen publication record `tables/final_composition_loao_data.json` stores the three reported methods, all episode identities and distance metrics, nested training identities, and checkpoint/atlas/source hashes. The script `tables/make_final_composition_loao.py` regenerates the complete table and its animal-level summary without fitting. Source predictions agree exactly with the final analytic mixture, and reloading saved predictions reproduces the reported metrics.

Table 10: Complete strict outer-animal composition evaluation on native sections. N counts directed predictions and $\widetilde{d}$ is their median measured separation. Central denotes central-copy and Atlas the center-calibrated population field. Methods share the same pairs and 400-μm masks. Pearson measures class-count spatial correlation, JSD composition discrepancy, and NMAE normalized class-count error. Blue columns identify ATLASS2V. Dashes denote no eligible measured pair, not zero error. Within-animal means equally weight the common 200/900-μm distances; the final row then equally weights animals.

$\lvert d\rvert$ (μm)	N	$\widetilde{d}$ (μm)	Class Pearson ↑			Composition JSD ↓			NMAE ↓		
			Central	Atlas	ATLASS2V	Central	Atlas	ATLASS2V	Central	Atlas	ATLASS2V
ABCA-1: coronal, 129 physical sections, 2,451 directed predictions											
100	245	97	0.7363	0.7035	0.7629	0.0498	0.0393	0.0374	0.4224	0.4500	0.3957
200	240	206	0.6326	0.7007	0.7135	0.0680	0.0398	0.0423	0.5097	0.4543	0.4438
300	228	295	0.5699	0.6995	0.7028	0.0824	0.0394	0.0426	0.5742	0.4578	0.4595
400	228	406	0.5013	0.6917	0.6934	0.0936	0.0392	0.0414	0.6247	0.4554	0.4575
500	224	505	0.4461	0.6813	0.6845	0.1035	0.0395	0.0408	0.6750	0.4640	0.4646
600	224	598	0.4057	0.6759	0.6783	0.1099	0.0393	0.0397	0.7023	0.4617	0.4609
700	226	701	0.3662	0.6724	0.6740	0.1169	0.0391	0.0392	0.7363	0.4606	0.4602
800	221	805	0.3312	0.6706	0.6715	0.1224	0.0390	0.0390	0.7596	0.4542	0.4540
900	215	909	0.2855	0.6605	0.6607	0.1307	0.0402	0.0402	0.8002	0.4697	0.4697
1,000	209	1009	0.2532	0.6724	0.6724	0.1361	0.0392	0.0392	0.8144	0.4545	0.4546
2,000	191	1999	0.0961	0.6215	0.6215	0.1723	0.0388	0.0388	0.9423	0.4805	0.4805
Mean: 200/900 μm			0.4591	0.6806	0.6871	0.0994	0.0400	0.0413	0.6549	0.4620	0.4567
ABCA-2: coronal, 54 physical sections, 528 directed predictions											
100	18	138	0.5811	0.6227	0.6426	0.0811	0.0474	0.0543	0.5760	0.5412	0.5324
200	78	185	0.6048	0.6187	0.6564	0.0739	0.0413	0.0463	0.5815	0.5508	0.5349
300	44	315	0.5034	0.6411	0.6532	0.1009	0.0439	0.0480	0.6623	0.5453	0.5443
400	48	386	0.4999	0.6579	0.6668	0.0946	0.0384	0.0408	0.6448	0.4850	0.4877
500	55	493	0.4104	0.6033	0.6102	0.1132	0.0420	0.0433	0.7287	0.5777	0.5750
600	42	592	0.3340	0.6609	0.6637	0.1201	0.0384	0.0392	0.7489	0.4851	0.4860
700	55	696	0.3553	0.5942	0.5961	0.1277	0.0432	0.0435	0.7783	0.5733	0.5727
800	43	799	0.2531	0.5911	0.5919	0.1391	0.0416	0.0418	0.8305	0.5501	0.5499
900	63	906	0.3058	0.6441	0.6444	0.1273	0.0382	0.0382	0.7903	0.5020	0.5019
1,000	34	999	0.1607	0.5583	0.5584	0.1614	0.0435	0.0435	0.9073	0.6049	0.6049
2,000	48	1992	0.1182	0.6122	0.6122	0.1715	0.0402	0.0402	0.9591	0.5064	0.5064
Mean: 200/900 μm			0.4553	0.6314	0.6504	0.1006	0.0397	0.0422	0.6859	0.5264	0.5184
ABCA-3: sagittal, 22 physical sections, 182 directed predictions											
100	4	114	0.5936	0.6199	0.6704	0.0612	0.0383	0.0393	0.5501	0.5229	0.4767
200	30	196	0.5668	0.5890	0.6512	0.0578	0.0334	0.0334	0.4878	0.4829	0.4134
300	10	300	0.4417	0.5848	0.6216	0.0801	0.0364	0.0372	0.6147	0.5067	0.4706
400	22	386	0.4341	0.5859	0.6163	0.0872	0.0336	0.0344	0.6396	0.4864	0.4660
500	12	455	0.3495	0.5421	0.5625	0.1020	0.0371	0.0375	0.7245	0.5146	0.5025
600	24	580	0.3378	0.6057	0.6131	0.1108	0.0343	0.0345	0.7524	0.4801	0.4756
700	18	714	0.2907	0.5329	0.5376	0.1252	0.0369	0.0368	0.8114	0.5136	0.5115
800	16	777	0.2946	0.6343	0.6361	0.1161	0.0345	0.0345	0.7836	0.4826	0.4814
900	20	900	0.2254	0.5388	0.5393	0.1378	0.0362	0.0361	0.8618	0.5057	0.5052
1,000	14	1024	0.2177	0.5613	0.5615	0.1221	0.0346	0.0345	0.8228	0.4930	0.4928
2,000	12	2010	0.0981	0.5943	0.5943	0.1752	0.0382	0.0382	0.9603	0.5136	0.5136
Mean: 200/900 μm			0.3961	0.5639	0.5953	0.0978	0.0348	0.0348	0.6748	0.4943	0.4593
ABCA-4: sagittal, 3 physical sections, 4 directed predictions											
100	0	—	—	—	—	—	—	—	—	—	—
200	2	193	0.3590	0.5852	0.5387	0.0970	0.0391	0.0445	0.7266	0.6254	0.5777
300	0	—	—	—	—	—	—	—	—	—	—
400	0	—	—	—	—	—	—	—	—	—	—
500	0	—	—	—	—	—	—	—	—	—	—
600	0	—	—	—	—	—	—	—	—	—	—
700	0	—	—	—	—	—	—	—	—	—	—
800	0	—	—	—	—	—	—	—	—	—	—
900	2	933	0.1577	0.6257	0.6260	0.1658	0.0383	0.0384	1.0022	0.5731	0.5727
1,000	0	—	—	—	—	—	—	—	—	—	—
2,000	0	—	—	—	—	—	—	—	—	—	—
Mean: 200/900 μm			0.2583	0.6054	0.5824	0.1314	0.0387	0.0414	0.8644	0.5993	0.5752
Four-animal mean			0.3922	0.6203	0.6288	0.1073	0.0383	0.0399	0.7200	0.5205	0.5024

C.2 BrainCAD cross-animal density validation

C.2.1 Cohort and outer cross-fitting

The population field is constructed from four Allen Zhuang-ABCA specimens and fixed before BrainCAD training. BrainCAD contains 60 mice: 20 marker-defined populations with three biological replicates per marker. A mouse contains one marker population, not all 20 populations simultaneously. The three outer folds hold out one replicate per marker (20 test mice) and use only the other two replicates (40 mice) for expert fitting, normalization, marker calibration, router training, and checkpoint selection. Rotating the held-out replicate evaluates every animal exactly once. Density is voxelized into 20-μm anterior–posterior slabs and 100-μm in-plane bins. Eligible prediction episodes use both signs of offsets 100, 200, . . ., 1,000 μm; the observed slab is an input and each target slab is scoring-only.

C.2.2 Frozen experts and monotone router

The near expert is a specimen-anchored center-plus-Atlas residual model fitted at 100 and 200 μm. The far expert is an Atlas-only density model. Both experts are frozen before the gate is learned. For marker m and absolute distance $|d|$ in millimeters, the primary gate is

$$w_{\text{near}}(d, m) = \sigma(b + \delta_m - \text{softplus}(s)|d|)\,, \qquad \hat{\rho}_d = w_{\text{near}}\hat{\rho}_{\text{near}} + (1 - w_{\text{near}})\hat{\rho}_{\text{Atlas}}. \tag{12}$$

The non-negative slope makes reliance on the near expert decrease monotonically with distance. The gate is fitted across 100–1,000 μm using only outer-training animals. Figure 9 reports $1 - w_{\text{near}}$, the Atlas-expert weight.

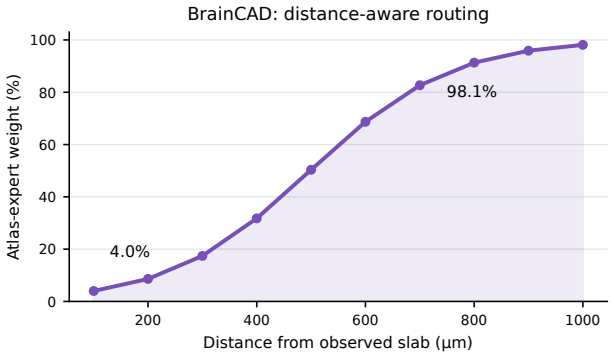


Figure 9: Frozen BrainCAD distance policy across 60 outer-held-out mice. Mean Atlas-expert weight rises from 4.0% at 100 μm to 98.1% at 1,000 μm. Each fold's policy is fitted only on outer-training animals and frozen before evaluation.

C.2.3 Complete distance-resolved results

Table 11 summarizes the distance-averaged comparison over all 60 held-out mice; Table 12 reports every evaluated distance. Central evidence is strongest at 100 μm, Atlas-only is comparatively stable through depth, and the router forms the upper envelope without using test targets for selection. It is best among the displayed non-oracle methods from 200 through 1,000 μm and is near-optimal at 100 μm.

C.2.4 Animal-level robustness

For each mouse, we compare the router with the best distance-integrated value among four fixed experts: central extrusion, central-only, Atlas-only, and fixed center-plus-Atlas. This is stricter than comparing with one globally selected baseline because the comparator may differ by mouse. The router improves 58/60 animals. The mean integrated gain is 0.0546 (median 0.0557; bootstrap 95% CI 0.0484–0.0604). Table 13 shows how this advantage emerges after the near plane.

Table 11: Primary BrainCAD result over 60 outer-held-out mice, averaged equally across 100–1,000-μm offsets. Each fixed method uses one expert at every distance; ATLASS2V uses the train-animal-fitted monotone distance router. NMAE is normalized mean absolute error.

Method	Density Pearson ↑	NMAE ↓	Slab-count r ↑
Central extrusion	0.6708	0.6437	0.9587
Central-only	0.6935	0.6554	0.9589
Atlas-only	0.6515	0.7823	0.9353
Fixed central + Atlas	0.6987	0.6406	**0.9624**
ATLASS2V distance router	**0.7426**	**0.6294**	0.9619

Table 12: Complete BrainCAD held-out distance curve. Values are three-fold mean density Pearson; NMAE belongs to the primary distance router. Atlas weight equals $1 - w_{\mathrm{near}}$.

$\lvert d\rvert$ (μm)	Central extr.	Atlas-only	Fixed central+Atlas	ATLASS2V	ATLASS2V NMAE ↓	Atlas wt.
100	**0.9132**	0.6584	0.9113	0.9106	0.3380	0.040
200	0.8537	0.6553	0.8605	**0.8608**	0.4428	0.086
300	0.7900	0.6594	0.8044	**0.8099**	0.5294	0.174
400	0.7318	0.6587	0.7535	**0.7696**	0.5951	0.318
500	0.6386	0.6435	0.6793	**0.7170**	0.6561	0.504
600	0.6397	0.6546	0.6690	**0.7108**	0.7008	0.687
700	0.5940	0.6515	0.6291	**0.6854**	0.7360	0.827
800	0.5610	0.6525	0.5991	**0.6703**	0.7569	0.913
900	0.5204	0.6494	0.5624	**0.6572**	0.7651	0.959
1,000	0.4652	0.6313	0.5184	**0.6347**	0.7741	0.981

C.2.5 EXPLORATORY ATLAS-CONFIDENCE FEATURES

An exploratory variant adds local Atlas density, relative population variation, and support to Eq. 12. Its distance-averaged Pearson is 0.7433 versus 0.7426 for the distance-only gate, an animal-level mean increment of 0.0010 (95% CI 0.0007–0.0013); NMAE is 0.6293 versus 0.6294. Because these three Atlas-derived covariates are correlated and the increment is small, we retain the simpler monotone distance gate as the primary method and do not use the exploratory variant for the headline claim.

Table 13: Per-distance mouse-level comparison with each mouse's best fixed expert at that distance. The main 58/60 statistic instead uses one best fixed expert after integrating each method across all ten distances.

$\lvert d\rvert$ (μm)	Mean Δr	Mice improved	$\lvert d\rvert$ (μm)	Mean Δr	Mice improved
100	-0.0060	14/60	600	$+0.0431$	60/60
200	$+0.0005$	46/60	700	$+0.0348$	59/60
300	$+0.0089$	56/60	800	$+0.0203$	58/60
400	$+0.0216$	59/60	900	$+0.0084$	57/60
500	$+0.0368$	57/60	1,000	$+0.0019$	56/60

D MULTIPLE-ORGAN EVALUATION

Each study uses a reference appropriate to its tissue and acquisition setting: serial sections from the same tumour for lung, other patients for breast, and independent donors for kidney. Their predictors and evaluation endpoints are specified separately.

D.1 LUNG: WITHIN-TUMOUR SERIAL REFERENCE

In the NSCLC study of Section 4.2.3, six registered CosMx sections come from one patient and one tumour (Pentimalli et al., 2025). Sections 4, 16, and 28 form the pseudo-atlas; disjoint sections 10, 22, and 34 alternate as observed center and hidden target, yielding six directed predictions at 60/120 μm on a 100-μm XY grid. The fixed rule retains half the observed section's residual around the target Atlas. It improves mean density Pearson from 0.7505 to 0.8141 and nine-class composition Pearson from 0.6271 to 0.7355. Correct-depth queries exceed reversed-depth queries by 0.0481 and 0.1369 in density and composition correlation, respectively. This is a within-tumour feasibility experiment using a fixed residual coefficient, not a fitted mouse-brain gate.

D.2 HER2ST: SINGLE-SECTION PREDICTION WITH A CROSS-PATIENT TUMOUR PRIOR

Scope and data. This exploratory extension predicts distributions on unobserved serial sections, not a same-section downstream label. HER2ST provides 36 H&E-paired ST sections from eight breast-cancer patients (Andersson et al., 2021). Patients A–D contribute six sections each and E–H contribute three. The observed sections are A3, B3, C3, D3, E2, F2, G2, and H2; all other 28 sections are targets. Unlike the brain experiments, there is no shared anatomical tumour coordinate system. The reference is a population residual model learned from other patients using central-section features, tumour-relative geometry, and absolute section distance. It is not an Allen-derived field or a standard breast 3D atlas.

Resolution, endpoints, and registration. The native array has 35×33 positions, approximately 100-μm spot diameter and 200-μm centre spacing. We retain this lattice without target pooling or Gaussian smoothing and set unmeasured positions to zero. This is a native-spot-grid evaluation, not a claim of continuous 100-μm sampling or single-cell resolution. The density endpoint is the local mean grayscale darkness of H&E around each spot, scaled using the section's 5th and 95th percentiles. It is an image-intensity proxy, not a segmented nuclear count or an absolute cell density. Composition uses the released deconvolution proportions for B cells, cancer-associated fibroblasts, endothelial cells, epithelial cells, myeloid cells, plasma cells, PVL cells, and T cells; these are inferred rather than directly measured cell-type fractions.

Target fields are aligned to the observed section by integer translations on the spot grid, selected within ± 5 positions using only H&E support and darkness. Target H&E is used for evaluation alignment, not as a predictor input; target molecular values do not choose the transformation. The registration audit found all spot centres inside their source images and identity transforms for the central sections. Mean noncentral support Dice was 0.9048 and H&E-intensity correlation was 0.4866. Shifting within a finite array can discard out-of-bounds values and introduces zero padding; sub-spot and nonrigid misalignment remain unresolved.

Supplementary Fig. 1 of the source study reports a 32-μm separation for A–D, consecutive sections for E–H, and a 16-μm section thickness (Andersson et al., 2021). We use 32 and 16 μm, respectively, as the nominal sampling steps, yielding offsets of 16, 32, 64, and 96 μm. The source does not specify whether the 32-μm separation is a centre distance or an inter-section gap; our reported offsets retain the former convention and are nominal, not independently measured coordinates. Because tumours do not share a signed anatomical axis, the predictor receives $|z|$: opposite sides at equal distance receive identical predictions. These results test distance-conditioned serial-section distributions rather than a signed 3D tumour geometry.

Patient-held-out training and predictor. Each outer fold holds out one complete patient. Within the other seven, leave-one-patient-out validation selects ridge regularization from $\{1, 10, 100, 1000\}$ and a residual gate from $\{0, 0.1, \ldots, 1\}$, independently for density and composition, using patient-balanced normalized MSE. The selected models are refit on the seven outer-training patients; the test patient's hidden endpoints never select a model or gate. Features include central support, signed distance to the support boundary, tumour-relative radius, H&E density, and (for composition) the central cell-type maps, with linear, quadratic, and log-distance features and interactions. Gaussian-blurred central features with $\sigma = 1.1$ grid positions supplement the native inputs; the *targets* are not smoothed.

For task t, the prediction is

$$\widehat{F}_t(z) = \mathrm{clip}_t\big[F_t(0) + g_t R_t\big(F(0), |z|\big)\big]\,, \tag{13}$$

where R_t is a ridge residual predictor and g_t is fixed within an outer fold. Density is clipped to $[0, 1]$ and composition is clipped below at zero. Selected density gates are 1.0 and composition gates are 0.8–0.9. Here distance enters the residual predictor, while its coefficient is constant within each fold; this breast-specific adaptation differs from the monotone distance gate used for mouse composition. Every target prediction rereads the same observed section; generated layers are not recursively fed back.

Controls. Central-copy repeats the observed map. Population-only predicts an absolute field from central support geometry and distance, plus central H&E intensity for composition; it is therefore not an unconditional atlas-only lookup. Correct distance uses the nominal target offset. Zero distance (called collapsed-distance in the result files) queries the *same trained model* at $|z| = 0$; it is not a separately retrained no-distance baseline. Wrong distance exchanges $16 \leftrightarrow 64$ and $32 \leftrightarrow 96$ μm and is not a reversed signed-z control. A matched retrained 2D-only model and a tuned smoothed-copy baseline have not been evaluated for these endpoints.

Metrics and aggregation. For density, r correlates flattened target and predicted grids. For composition, r averages the finite per-class spatial correlations, not the correlation of a spot's full gene-expression profile. Whole-grid correlations include zero-filled background; tissue-only r_{tissue} uses target density > 0.03 or summed target composition > 0.03. Normalized MAE is mean absolute error divided by mean absolute target intensity. Spatial JSD averages Jensen–Shannon divergence between normalized eight-class vectors at target-supported positions, with natural logarithms. Slice-composition scores compare the normalized class totals over a whole section and do not measure spatial recovery. Unless stated otherwise, metrics are averaged over hidden sections within each patient and then equally over the eight patients. Eight patients, not the thousands of lattice positions, define the biological sample size.

Density and composition correlations improve in all eight patients. Table 14 reports every evaluated method and metric. Correct-distance prediction improves density r by 0.0521, composition r by 0.0927, and composition JSD by 0.0179 relative to central-copy, with favourable changes in all eight patients for these three endpoints. Density and composition nMAE improve in six and five patients, respectively. Tissue-only gains are smaller: $0.4388 \rightarrow 0.4703$ for density and $0.2566 \rightarrow 0.3005$ for composition. Slice-level composition correlation and JSD do not improve over central-copy. We consequently do not describe the method as uniformly superior across metrics.

The correct-distance composition query has better whole-grid correlation and JSD than both distance controls, but zero distance has essentially equal tissue-only correlation (0.3007 versus 0.3005).

Density also changes little between correct and wrong distance. Thus the results support cross-patient distribution prediction, with suggestive distance dependence, but do not isolate depth-specific information from learned denoising and tissue-support effects. This resolution experiment followed inspection of the coarse-grid results and is exploratory.

Resolution sensitivity and layer-change diagnostic. The earlier setting averaged 2×2 spots to an 18×17 grid at approximately 400-μm pitch, then applied a Gaussian with $\sigma = 0.65$ coarse cells (260 μm; Gaussian FWHM approximately 610 μm). Table 17 includes both central-copy and model results at both resolutions. Coarsening and smoothing make the target easier and can obscure local heterogeneity. The model's blurred-feature bandwidth also changes in physical units with the grid, so this comparison is not an isolated causal ablation of target smoothing.

For the native predictions, the correlation of $\widehat{F}(z) - F(0)$ with $F(z) - F(0)$ is 0.5425 for density and 0.5664 for composition. These diagnostics stack hidden layers within each patient, average composition correlations across classes, then average across patients. Subtracting the same central map from both arguments can itself induce correlation, so these values are not independent proof of correct depth recovery. The distance-, patient-, and layer-resolved tables below report the actual endpoint scores. In particular, 16-μm targets are from patients E–H whereas longer distances are from A–D; the distance table cannot be read as a within-cohort decay curve. None of these results establishes precise spot-level gene-expression recovery.

Source records. Tables are transcribed programmatically from the native-spot run's overall, patient, distance, episode, and volume metric CSVs and the coarse run's overall CSV. A frozen copy and SHA-256 provenance are retained in `tables/her2st_source_data.json`. The per-layer table includes all 28 hidden targets; the volume table is a separately labelled aggregation, not another independent cohort.

Table 14: HER2ST native-lattice results, averaged equally over eight patients. D denotes the H&E density proxy; C denotes composition. Whole-grid, tissue-only, and whole-section composition metrics are distinct. Boldface identifies the correct-distance model, not the best value in every column.

	Density proxy			Composition					
Method	$r \uparrow$	$r_{\text{tissue}} \uparrow$	nMAE$\downarrow$	$r \uparrow$	$r_{\text{tissue}} \uparrow$	nMAE$\downarrow$	JSD$\downarrow$	Slice $r \uparrow$	Slice JSD$\downarrow$
Central-copy	0.7737	0.4388	0.5093	0.5893	0.2566	0.5229	0.0586	0.9827	0.0030
Population-only	0.7538	0.0556	0.6415	0.6411	0.0518	0.8329	0.1021	0.8334	0.0597
Zero distance	0.8229	0.4709	0.5515	0.6471	0.3007	0.5546	0.0478	0.9737	0.0065
Wrong distance	0.8246	0.4625	0.4950	0.6755	0.2852	0.5623	0.0473	0.9515	0.0106
Correct distance	0.8258	0.4703	0.4915	0.6820	0.3005	0.5214	0.0407	0.9757	0.0055

Table 15: HER2ST results by nominal absolute section offset. Each distance includes four patients: 16-μm targets are from E–H, and 32–96-μm targets from A–D. Distances use the convention stated in the protocol and do not form a within-cohort decay curve.

		Density proxy			Composition			
$\lvert z\rvert$ (μm)	Method	$r\uparrow$	$r_{\text{tissue}}\uparrow$	nMAE$\downarrow$	$r\uparrow$	$r_{\text{tissue}}\uparrow$	nMAE$\downarrow$	JSD$\downarrow$
16	Central-copy	0.7839	0.4773	0.4396	0.6342	0.2878	0.4051	0.0385
16	Population-only	0.7400	0.0191	0.5957	0.6702	0.0305	0.7334	0.0797
16	Zero distance	0.8261	0.5039	0.4601	0.6882	0.3321	0.4410	0.0312
16	Wrong distance	0.8269	0.4928	0.4305	0.7139	0.3084	0.4783	0.0352
16	Correct distance	0.8274	0.5032	0.4269	0.7177	0.3378	0.4109	0.0258
32	Central-copy	0.7686	0.4455	0.5803	0.5457	0.2545	0.6580	0.0804
32	Population-only	0.7469	0.0883	0.7131	0.6077	0.0790	0.9363	0.1263
32	Zero distance	0.8173	0.4782	0.6556	0.6022	0.2966	0.6913	0.0693
32	Wrong distance	0.8184	0.4579	0.5786	0.6385	0.2707	0.7349	0.0694
32	Correct distance	0.8200	0.4769	0.5675	0.6330	0.3006	0.6376	0.0616
64	Central-copy	0.7536	0.3603	0.5848	0.5633	0.2382	0.6086	0.0721
64	Population-only	0.7761	0.0602	0.6629	0.6401	0.0846	0.9174	0.1198
64	Zero distance	0.8131	0.3882	0.6442	0.6308	0.2908	0.6292	0.0551
64	Wrong distance	0.8154	0.3886	0.5600	0.6544	0.2972	0.5671	0.0495
64	Correct distance	0.8171	0.3802	0.5625	0.6762	0.2874	0.5965	0.0462
96	Central-copy	0.7729	0.3899	0.5641	0.5042	0.1415	0.6699	0.0888
96	Population-only	0.7921	0.1637	0.6851	0.5646	0.0382	0.9545	0.1300
96	Zero distance	0.8383	0.4565	0.6146	0.5641	0.1719	0.7002	0.0730
96	Wrong distance	0.8445	0.4684	0.5199	0.6000	0.1742	0.6280	0.0594
96	Correct distance	0.8470	0.4728	0.5210	0.6131	0.1406	0.6913	0.0627

Table 16: HER2ST patient-level results, averaged over each patient's hidden sections. C denotes central-copy and M denotes the correct-distance model. The eight patients are the biological evaluation units.

	Density proxy						Composition					
	$r\uparrow$		$r_{\text{tissue}}\uparrow$		nMAE$\downarrow$		$r\uparrow$		$r_{\text{tissue}}\uparrow$		JSD$\downarrow$	
Patient	C	M	C	M	C	M	C	M	C	M	C	M
A	0.7807	0.8517	0.3001	0.3577	0.5162	0.4778	0.4103	0.5045	0.2643	0.3225	0.0597	0.0463
B	0.6913	0.7669	0.4995	0.5585	0.7293	0.6863	0.5285	0.6489	0.1974	0.2098	0.1071	0.0773
C	0.7008	0.7721	0.1979	0.1946	0.7066	0.6438	0.5339	0.6305	0.1853	0.2139	0.0946	0.0633
D	0.8810	0.9063	0.6037	0.6388	0.3634	0.4169	0.7051	0.8013	0.2545	0.3070	0.0537	0.0357
E	0.7487	0.8140	0.2902	0.3380	0.4639	0.4103	0.6076	0.6848	0.1862	0.2254	0.0257	0.0118
F	0.7733	0.8155	0.3174	0.2976	0.3985	0.3832	0.5554	0.6712	0.1652	0.2062	0.0311	0.0201
G	0.8174	0.8623	0.5370	0.5828	0.4158	0.4393	0.6909	0.7562	0.4071	0.4618	0.0603	0.0471
H	0.7959	0.8177	0.7647	0.7946	0.4803	0.4746	0.6829	0.7585	0.3926	0.4577	0.0369	0.0243

Table 17: Resolution sensitivity, including both baselines and models at both scales. Native targets have no pooling or added smoothing; the coarse setting pools 2×2 positions and smooths with $\sigma = 0.65$ coarse cells. All scores are patient-balanced means.

Endpoint	Grid	Method	$r \uparrow$	$r_{\text{tissue}} \uparrow$	nMAE↓	JSD↓
Density proxy	400 μm, smoothed	Central-copy	0.9486	0.8932	0.2307	–
Density proxy	400 μm, smoothed	Correct distance	0.9505	0.8962	0.2347	–
Density proxy	Native spot lattice	Central-copy	0.7737	0.4388	0.5093	–
Density proxy	Native spot lattice	Correct distance	0.8258	0.4703	0.4915	–
Composition	400 μm, smoothed	Central-copy	0.8979	0.8244	0.2636	0.0115
Composition	400 μm, smoothed	Correct distance	0.8990	0.8264	0.2651	0.0129
Composition	Native spot lattice	Central-copy	0.5893	0.2566	0.5229	0.0586
Composition	Native spot lattice	Correct distance	0.6820	0.3005	0.5214	0.0407

D.2.1 Layer-resolved and alternative volume aggregation

The following tables retain all 28 hidden targets and the alternative concatenated-volume aggregation. They use the same predictions as the summaries above and are not additional cohorts.

Table 18: All 28 hidden-layer predictions of the correct-distance model on the native lattice. D denotes the density proxy and C denotes composition. Signed offsets identify the targets, but the predictor receives only the absolute distance. The observed central layer is excluded.

Input→target	Nominal z (μm)	D: r	D: r_{tissue}	D: nMAE	C: r	C: r_{tissue}	C: nMAE	C: JSD
A3→A1	-64	0.8396	0.2815	0.4898	0.5543	0.3394	0.4878	0.0406
A3→A2	-32	0.8410	0.3901	0.5157	0.5015	0.3689	0.4839	0.0504
A3→A4	+32	0.8645	0.4241	0.4484	0.5212	0.3714	0.3837	0.0278
A3→A5	+64	0.8443	0.3482	0.5123	0.5603	0.4019	0.4948	0.0364
A3→A6	+96	0.8688	0.3449	0.4228	0.3852	0.1309	0.6606	0.0763
B3→B1	-64	0.7118	0.4498	0.7242	0.6232	0.2152	0.7508	0.0591
B3→B2	-32	0.7625	0.5646	0.7750	0.6276	0.2042	0.8478	0.0948
B3→B4	+32	0.8348	0.6741	0.5804	0.6601	0.2777	0.8016	0.0935
B3→B5	+64	0.7174	0.4620	0.7100	0.6429	0.1813	0.7805	0.0535
B3→B6	+96	0.8079	0.6422	0.6417	0.6908	0.1704	0.9062	0.0858
C3→C1	-64	0.7990	0.1117	0.6327	0.7141	0.2811	0.6508	0.0538
C3→C2	-32	0.5792	0.0346	0.9082	0.4175	0.1219	1.0432	0.1091
C3→C4	+32	0.8530	0.3542	0.5078	0.7250	0.3306	0.6300	0.0475
C3→C5	+64	0.8153	0.1148	0.5599	0.7189	0.2697	0.5918	0.0455
C3→C6	+96	0.8138	0.3578	0.6105	0.5770	0.0663	0.7334	0.0605
D3→D1	-64	0.8964	0.5874	0.4471	0.8064	0.3108	0.5704	0.0545
D3→D2	-32	0.9042	0.7055	0.4071	0.7963	0.3707	0.4832	0.0373
D3→D4	+32	0.9207	0.6683	0.3973	0.8151	0.3592	0.4278	0.0328
D3→D5	+64	0.9127	0.6865	0.4240	0.7894	0.2994	0.4451	0.0260
D3→D6	+96	0.8974	0.5465	0.4088	0.7995	0.1950	0.4650	0.0279
E2→E1	-16	0.8244	0.3208	0.3982	0.6919	0.2126	0.3054	0.0134
E2→E3	+16	0.8035	0.3551	0.4224	0.6776	0.2381	0.3309	0.0103
F2→F1	-16	0.8109	0.3177	0.3993	0.6797	0.2227	0.3409	0.0192
F2→F3	+16	0.8200	0.2774	0.3672	0.6627	0.1898	0.3629	0.0210
G2→G1	-16	0.8548	0.5861	0.4581	0.7384	0.4171	0.5326	0.0438
G2→G3	+16	0.8697	0.5795	0.4205	0.7740	0.5066	0.5376	0.0504
H2→H1	-16	0.9366	0.8664	0.2849	0.8140	0.4769	0.3190	0.0205
H2→H3	+16	0.6989	0.7228	0.6644	0.7029	0.4384	0.5583	0.0281

Table 19: Hidden-volume scores: concatenate hidden layers within each patient before computing spatial metrics, then average across eight patients. The observed central layer is excluded. This aggregation differs from the per-layer-then-patient averages used in the main text and is not an independent experiment.

Endpoint	Method	$r \uparrow$	$r_{\text{tissue}} \uparrow$	nMAE↓	JSD↓
Density proxy	Central-copy	0.7730	0.4360	0.5057	–
Density proxy	Population-only	0.7531	0.0532	0.6374	–
Density proxy	Zero distance	0.8224	0.4653	0.5471	–
Density proxy	Wrong distance	0.8237	0.4550	0.4913	–
Density proxy	Correct distance	0.8254	0.4651	0.4877	–
Composition	Central-copy	0.5822	0.2495	0.5216	0.0586
Composition	Population-only	0.6306	0.0496	0.8314	0.1020
Composition	Zero distance	0.6392	0.2919	0.5532	0.0478
Composition	Wrong distance	0.6643	0.2712	0.5609	0.0472
Composition	Correct distance	0.6706	0.2905	0.5203	0.0407

D.3 Human-kidney distance-resolved feasibility

The study uses glomerulus masks from nine complete human-kidney LSFM volumes spanning six donors (McLaughlin et al., 2025): three donors construct the reference, one fits the distance router, and two are held out, with no donor overlap between roles. Each prediction observes one approximately 10-μm section and targets a 100-μm hidden slab. An internal four-donor cross-fit also reproduces the direction of improvement in all four held-out folds. The fixed two-donor evaluation uses a common set of 94 directed episodes at each distance. The candidate equals central-copy at 100/200 μm and uses the mass-preserving Atlas-residual router at 500–1,500 μm. The 1,500-μm endpoint is a post-hoc distance-support extension with this near/far rule held fixed. Table 20 reports all three metrics through 1,500 μm. The gains increase toward 1,500 μm, where both held-out donors improve in Pearson and nMAE. The correct-region field also outperforms the wrong-region control across the far-distance endpoints, consistent with transfer of coarse cortical organization.

Table 20: Human-kidney glomerular-field prediction in two held-out donors. Targets are 100-μm hidden slabs; values are equal-donor means over the common-anchor episodes.

	Pearson ↑		nMAE ↓		JSD ↓	
$\lvert d \rvert$ (μm)	Central	Model	Central	Model	Central	Model
100	0.5135	0.5135	1.0933	1.0933	0.2420	0.2420
200	0.2689	0.2689	1.4342	1.4342	0.4079	0.4079
500	0.2433	**0.2500**	1.4996	**1.4817**	0.4630	**0.4512**
1,000	0.1671	**0.1734**	1.5961	**1.5757**	0.4866	**0.4726**
1,500	0.1298	**0.1380**	1.7852	**1.7621**	0.4980	**0.4836**

E Ablation and Model Alternatives

The main component comparisons isolate the reference, residual, and gate. The mouse ablations examine reference coverage and learned refinements under the same independent-mouse evaluation; an additional routing study uses uninjured zebrafish hearts.

E.1 Equal-budget atlas-completeness ablation

We additionally test whether reference depth coverage contributes beyond cell count, using the same frozen composition predictor as Table 1, with no weight or gate refitting. The four references contain 5,305,932 valid CCF-registered cells. We take the floor of 10% of each animal's cells, giving budgets 256,336, 103,568, 154,796, and 15,891 for Zhuang-ABCA-1–4, respectively, totaling 530,591 cells. Complete coverage samples approximately proportionally within every nonempty 100-μm CCF anterior–posterior slab, retaining at least one cell per nonempty slab. Sparse coverage retains one such slab per 500 μm, with the five repetitions covering all five phases. Gap coverage removes 16 consecutive slabs; gap locations are set from reference-cell AP-mass quantiles 0.20, 0.35, 0.50,

0.65, and 0.80, without consulting test performance. The resulting intervals are 3.6–5.2, 5.1–6.7, 6.4–8.0, 7.9–9.5, and 9.5–11.1 mm. Both incomplete conditions refill the fixed per-animal budget from retained regions. Because reference specimens have different physical section orientations, these are CCF-depth interventions, not deletion of identically oriented native sections.

At the four previously specified offsets (200/400/1,000/2,000 μm), complete coverage reaches mean class Pearson 0.5692 ± 0.0047 and JSD 0.04204 ± 0.00013. Equal-budget sparse coverage gives 0.5549 ± 0.0141 and 0.04441 ± 0.00043; the contiguous-gap condition gives 0.5526 ± 0.0105 and 0.04573 ± 0.00108. These are means and sample standard deviations across five layouts, not five biological replicates. The original 5.31-million-cell complete atlas remains stronger overall (Pearson 0.5983; JSD 0.03845), so the result does not imply that fewer cells are preferable.

Rebuilding the full atlas reproduces the original benchmark exactly. Holding total cell count fixed while reducing coverage increases sampling density in retained layers. Two auxiliary controls therefore restrict the actual complete-coverage sample to the sparse or gap layout without refilling: retained cell identities are identical, but total budgets are lower (99,517–111,430 for sparse and 429,884–462,507 for gap coverage). All atlases use the original 34-class, 100-μm support-normalized Gaussian density construction and four-animal population aggregation, followed by the same eight-class reduction. Missing AP planes use per-location linear interpolation from retained supported planes and nearest-plane extension at the ends; locations with no retained support remain unsupported. Deleted full-atlas values are never used. All 25 ablated atlases and their sampled cell identities are fixed before this evaluation reads test maps. Rebuilding the original full atlas reproduces its density, variation, and support arrays exactly; rescoring the original predictor reproduces the composition benchmark at all four summary distances with zero numerical discrepancy.

Missing depth reduces fidelity even with identical retained cells. Every condition uses the original 470 directed episodes, 11 distances, 400-μm scoring grid, and full-reference scoring masks. Covered and missing regions are defined by the queried CCF-AP coordinates for each layout, identically across methods. A scoring bin belongs to a pure region only when all 16 native sub-bins belong to that region; mixed boundary bins remain in the overall scores. Each layout is summarized by an equal average over 200/400/1,000/2,000 μm, followed by the five-layout mean and sample standard deviation. Complete coverage improves overall Pearson in four of five sparse comparisons and all five gap comparisons. Within the removed interval, complete versus gap coverage gives Pearson 0.5629 versus 0.4139 and JSD 0.03690 versus 0.06128; in retained regions, the denser gap atlas is slightly better (Pearson 0.5718 versus 0.5681). Under equal budgets, its gap-region Pearson gain is 0.1491 ± 0.0142 and JSD reduction is 0.02438 ± 0.00938. With identical retained cells, the corresponding missing-region Pearson gain is 0.1535 ± 0.0167, versus -0.0001 ± 0.0008 in retained regions. These results distinguish the missing-depth penalty from changes in retained-region density. This is a retrospective inference-reference intervention on one test mouse with a predictor previously fitted using the complete reference, not independent-animal replication or a matched end-to-end training-budget comparison.

E.2 Conditional-Field Refinement Architecture Comparison

Question and scope. Does learning a more expressive spatial correction improve the final AtlasS2V composition field? We compare a pointwise MLP, a 2D Transformer, and a sparse 3D Transformer as conditional count-field refiners. The reference prediction is the calibrated expert mixture with the frozen distance gate defined in Section 3. All three refiners receive identical conditions and supervision, but relative to the final predictor they add a midpoint Atlas query and a learned field correction. Thus this is a controlled comparison among *refinement architectures*, not evidence that a neural refiner is required by the final method.

Conditional input and output. Each model receives the observed eight-class field and center-calibrated Atlas fields at offsets 0, $d/2$, and d, with signed relative-depth information. The observed field is repeated across the three planes; no hidden specimen section is an input. Counts are transformed by $\log(1 + Y/u)$, where u is the 90th percentile of positive observed per-bin total counts. A learned correction is added to the log-transformed frozen fidelity prediction, followed by a nonnegative clamp and inverse transform. The final correction layer is initialized to zero, so all variants initially reproduce the same reference prediction.

Table 21: Conditional-field refinement on C57BL6J-638850. Metrics average the four fixed summary distances; refiners report mean ± sample SD over three optimization seeds. Parameter counts refer only to the added refiner, not the frozen policy or Atlas. All predictions retain native counts, without occupancy projection.

Method	Parameters	Class Pearson ↑	JSD ↓
Central-copy	–	0.3762	0.11086
Atlas-only	–	0.5737	0.03968
Frozen fidelity policy	–	0.5983	0.03845
+ Pointwise MLP	149,234	0.5861 ± 0.0046	0.03978 ± 0.00040
+ 2D Transformer	288,264	0.5410 ± 0.0115	0.04070 ± 0.00132
+ Sparse 3D Transformer	149,000	0.5314 ± 0.0036	0.04229 ± 0.00145

The pointwise MLP mixes the 51 stacked conditional channels using two hidden layers of width 357, without exchanging information between XY positions. Both Transformers use hidden width 64, two attention blocks, four heads, feed-forward width 128, no dropout, and 8×8 spatial patches. The 2D variant stacks the planes as channels before XY attention; the sparse 3D variant retains separate depth tokens and attends jointly across XY and the three Atlas planes. Learned positional embeddings encode XY and relative depth. Target-plane features are bilinearly upsampled and decoded with the original conditional channels. The output is one queried plane, not a generated whole-organ volume. The MLP and 3D Transformer have 149,234 and 149,000 parameters; the 2D variant has 288,264 because of its larger input patch projection and is not parameter-matched.

Training, freezing, and evaluation. ABCA-1–3 supply 1,436 training episodes and ABCA-4 supplies 417 development episodes for epoch selection. Each final model is reinitialized and refitted on 1,854 episodes from all four reference animals, excluding the supervised animal from its own Atlas. All variants minimize the same masked log-count MSE plus composition JSD with AdamW, learning rate 10^{-3}, weight decay 10^{-4}, batch size eight, gradient-norm clipping at five, and at most 20 epochs. Seeds 20260914/15/16 use matched minibatch ordering across architectures. Selected epoch counts are 15/14/19 for the MLP, 7/3/9 for 2D attention, and 7/8/8 for 3D attention. No external outcome is used to choose an epoch or seed.

All nine checkpoints are saved and independently reloaded before opening the external maps in this run. Evaluation uses the same C57BL6J-638850 mouse, 470 directed pairs, 49.999-μm pairing tolerance, 400-μm scoring grid, and masks as Table 1. The frozen-policy metrics reproduce the matched benchmark exactly. The summary endpoints (200/400/1,000/2,000 μm) were fixed before this run; we average endpoints equally within each seed, then report the seed mean and sample SD. This is a retrospective study on a previously evaluated mouse, not a fresh blinded cohort. Overlapping section pairs and three optimization seeds do not constitute independent biological replicates.

Neural refiners do not improve composition fidelity under the tested protocol. Under this protocol, none of the new refiners improves the primary composition summaries (Table 21; Fig. 10). Sparse 3D attention changes class Pearson from 0.5983 to 0.5314 and JSD from 0.03845 to 0.04229; both worsen in all three seeds. It also underperforms the parameter-matched pointwise MLP. Not every auxiliary metric declines: the MLP increases Pearson for the sum of its eight class-count fields from 0.5561 to 0.5663, while class-mass NMAE worsens from 0.5055 to 0.5224. This total-count metric is not a new BrainCAD density-head result.

Training loss can continue falling while development loss rebounds, suggesting overfitting risk, but this comparison does not isolate sample size, distribution shift, loss choice, or architecture as the cause. It supports retaining the frozen policy under the tested settings, not a general claim that Transformers are unsuitable or that reference quality universally outweighs model capacity. This experiment does not reevaluate the BrainCAD density adaptation, and conditional diffusion was not evaluated.

Reproducibility. The compact artifact `tables/conditional_field_ablation_data.json` retains the protocol, nine checkpoint hashes, every distance and seed's aggregate metrics, exact pair identities, and the source-report hash. The self-contained script `figures/make_`

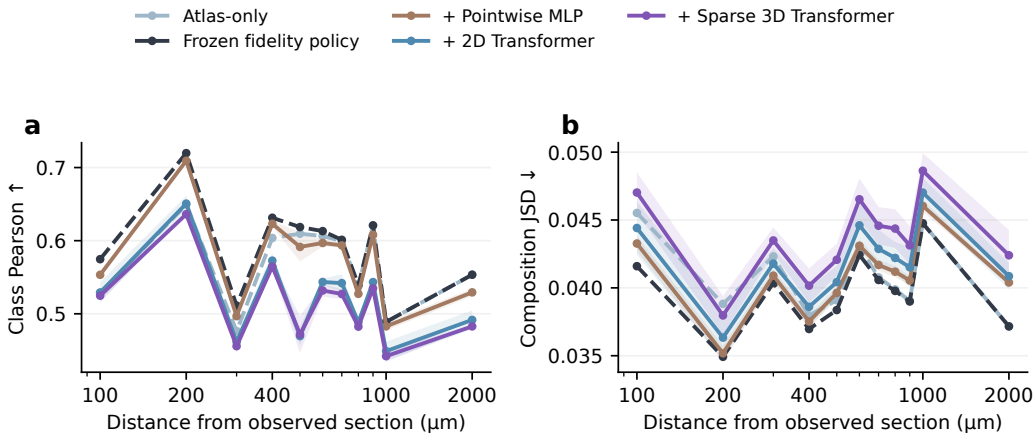


Figure 10: Conditional-field refinement architectures on the independent mouse. Curves show class Pearson (left; higher is better) and composition JSD (right; lower is better) at all 11 evaluated distances. Solid curves are means of three separately evaluated models, not ensemble predictions; bands show optimization-seed SD, not animal-level confidence intervals. Dashed curves are the frozen fidelity policy and Atlas-only. Central-copy is retained in Table 21 and omitted here for readability. All methods share the scoring pairs and masks; only the three new refiners share the augmented conditional input.

`conditional_field_ablation.py` regenerates the table and vector/raster curves without training. Complete per-pair outputs and weights remain in the experiment archive identified by that record; no best-performing seed is substituted for the reported mean.

E.3 Zebrafish-heart routing data

The zebrafish-heart routing comparisons in Table 4 use the uninjured subset of the Stereo-seq atlas from Li et al. (2025). We construct the reference from 119 sections of one heart and evaluate predictions on six other uninjured hearts. The analysis reads `Stereo-seq-3D.meta.txt` and `Stereo-seq-regeneration.meta.txt` from the processed-data release on Zenodo (DOI: 10.5281/zenodo.14991776). The original sequencing data are deposited under NCBI BioProject PRJNA1233465 and CNSA `CNP0005245`.

F Additional Downstream Details

F.1 Xenium hidden-gene prediction

We use the *Fresh Frozen Mouse Brain Replicates* dataset released by 10x Genomics on January 22, 2023 (10x Genomics, 2023), analyzed with Xenium Onboard Analysis 1.0.2. The three adjacent 10-μm sections come from one C57BL/6 mouse and were placed on the same slide. The samples `Xenium_V1_FF_Mouse_Brain_MultiSection_1`, `Xenium_V1_FF_Mouse_Brain_MultiSection_2`, and `Xenium_V1_FF_Mouse_Brain_MultiSection_3` contain 162,033, 154,654, and 158,047 cells, respectively.

The downstream probe receives measured panel expression, a local 2D expression summary, in-plane coordinates, and the mean Atlas query $\bar{\mathbf{A}}_0$ and variation U_0 defined in Eq. 8. It predicts hidden genes under whole-section training, validation, and test splits. This is a separately fitted readout using Atlas context, not the class-count predictor's density/composition output.

Together, the three sections contain 474,734 cells, aggregated into 15,069 non-overlapping 100-μm pseudo-spots. The strong 2D model receives measured panel genes, a 151-μm local expression neighborhood, and in-plane coordinates. AtlasS2V additionally receives a soft Allen query with a 500-μm pose uncertainty scale. At the prespecified 10/25/50-gene panels, relative gains over the

Table 22: Complete nine-panel Xenium hidden-gene results. Entries are mean per-gene Pearson across the three held-out whole-section rotations. "Only" columns use Atlas features without measured expression; all other Atlas columns augment the strong 2D model. Bold panel sizes are the frozen endpoints; intermediate rows are post-hoc characterization.

Measured genes	Panel only	Strong 2D	Atlas 2D only	Atlas 3D only	+ 2D marginal	+ 3D point	+ soft 3D	+ reversed 3D	+ shuffled 3D
10	0.5843	0.6755	0.6006	0.5976	0.7268	0.7066	0.7268	0.7215	0.6715
15	0.6139	0.6958	0.6006	0.5976	0.7371	0.7203	0.7371	0.7323	0.6922
20	0.6576	0.7272	0.6006	0.5976	0.7577	0.7457	0.7569	0.7539	0.7245
25	0.6822	0.7408	0.6006	0.5976	0.7655	0.7558	0.7651	0.7625	0.7385
30	0.6926	0.7482	0.6006	0.5976	0.7699	0.7614	0.7700	0.7673	0.7461
35	0.7083	0.7587	0.6006	0.5976	0.7765	0.7692	0.7762	0.7738	0.7569
40	0.7199	0.7648	0.6006	0.5976	0.7803	0.7736	0.7795	0.7777	0.7631
45	0.7349	0.7727	0.6006	0.5976	0.7859	0.7801	0.7851	0.7834	0.7711
50	0.7452	0.7790	0.6006	0.5976	0.7902	0.7854	0.7895	0.7881	0.7776

Table 23: Section-resolved Xenium primary comparison. Each pair gives strong 2D / strong 2D plus soft 3D Atlas Pearson on one completely held-out section. The final column averages the paired 3D-minus-2D difference over sections.

Measured genes	Section 1		Section 2		Section 3		Mean gain
	2D	+3D	2D	+3D	2D	+3D	
10	0.6769	0.7313	0.6732	0.7245	0.6764	0.7247	+0.0513
15	0.6975	0.7411	0.6938	0.7354	0.6961	0.7347	+0.0413
20	0.7290	0.7608	0.7256	0.7551	0.7272	0.7548	+0.0297
25	0.7431	0.7687	0.7393	0.7633	0.7401	0.7633	+0.0242
30	0.7506	0.7731	0.7464	0.7682	0.7477	0.7687	+0.0218
35	0.7605	0.7797	0.7571	0.7736	0.7584	0.7752	+0.0175
40	0.7664	0.7817	0.7633	0.7778	0.7647	0.7790	+0.0147
45	0.7746	0.7875	0.7712	0.7834	0.7722	0.7843	+0.0124
50	0.7811	0.7920	0.7776	0.7879	0.7785	0.7887	+0.0105

strong 2D model are 7.60%, 3.27%, and 1.34%; every held-out section improves at all nine panel sizes in the expanded sweep. Figure 4c displays held-out Section 2 for the three genes with the largest mean AtlasS2V-over-strong-2D gain across the frozen ten-gene rotations. Within each row, ground truth and all model maps share a color scale; badges report spatial Pearson. These selected examples illustrate performance rather than constituting independent evidence. In Fig. 4a, thin lines show the three section rotations and thick lines their means. Filled markers in panels a and b denote the prespecified endpoints, while open markers denote the post-hoc intermediate panel sizes.

The three sections are rotated through whole-section train, validation, and test roles. The 10/25/50-gene panels are the frozen endpoints; 15/20/30/35/40/45 genes are post-hoc characterization along the same nested maximum-50 ranking. All 198 genes outside the maximum panel remain hidden at every panel size. At ten measured genes, the correct soft Atlas reaches 0.7268, compared with 0.6755 for the strong 2D model, 0.7215 for reversed depth, and 0.6715 after within-section Atlas shuffling. A depth-marginalized Atlas is close to the correct soft query; the experiment therefore establishes Atlas utility under sparse measurement but is not used as independent evidence of precise anterior–posterior recovery.

F.2 STANDARD-CCF SINGLE-CELL TYPING

The four specimens provide 1,122-gene expression profiles and 34-class labels. Gene panels are nested prefixes of a maximum 50-gene ranking selected without the test animal. Figure 4b shows animal-level thin curves after averaging computational seeds; thick curves show the mean across animals. At fifty genes, the correct-3D-minus-2D macro-F1 gain is 0.0380 (5.67% relative; animal-bootstrap 95% CI $[0.0242, 0.0519]$), and the gain over gene-only is 0.0728. Correct 3D is 8.91% above shuffled depth. These results do not establish the same gain after H&E/DAPI-to-CCF registration or on an external platform.

Table 24: Complete nine-panel single-cell-typing results. Entries are 34-class macro-F1 averaged first over three seeds within each held-out animal and then over the four animals. Atlas-only columns omit gene-expression scores; the other Atlas columns augment the same gene classifier. Bold panel sizes are confirmatory; intermediate rows are post-hoc characterization.

Genes	Gene only	2D Atlas only	3D Atlas only	Gene + 2D Atlas	Gene + correct 3D	Gene + shuffled depth	Gene + wrong region
10	0.2090	0.0633	0.2047	0.3058	**0.4707**	0.2699	0.3304
15	0.3026	0.0633	0.2047	0.4205	**0.5804**	0.3519	0.4377
20	0.3754	0.0633	0.2047	0.4805	**0.6082**	0.3976	0.4873
25	0.4716	0.0633	0.2047	0.5665	**0.6620**	0.4869	0.5683
30	0.5362	0.0633	0.2047	0.6138	**0.6959**	0.5562	0.6242
35	0.5592	0.0633	0.2047	0.6307	**0.7055**	0.5828	0.6435
40	0.5879	0.0633	0.2047	0.6410	**0.7014**	0.6073	0.6566
45	0.6116	0.0633	0.2047	0.6523	**0.7035**	0.6268	0.6700
50	0.6366	0.0633	0.2047	0.6713	**0.7094**	0.6513	0.6853

Table 25: Animal-resolved depth-specific cell-typing effect. Animal columns are correct-3D minus depth-marginalized-2D macro-F1 after averaging the three seeds within each animal. Confidence intervals bootstrap the four animals. The last column counts animal-by-seed runs in which correct 3D exceeds every displayed gene-conditioned control.

Genes	Animal 1	Animal 2	Animal 3	Animal 4	Mean	Animal-bootstrap 95% CI	All-control wins
10	+0.2041	+0.1970	+0.1556	+0.1031	+0.1649	[0.1266, 0.2006]	12/12
15	+0.1682	+0.1985	+0.1474	+0.1253	+0.1599	[0.1360, 0.1857]	12/12
20	+0.1277	+0.1638	+0.1105	+0.1086	+0.1276	[0.1095, 0.1505]	12/12
25	+0.1018	+0.1177	+0.0540	+0.1085	+0.0955	[0.0676, 0.1137]	12/12
30	+0.0939	+0.0936	+0.0386	+0.1021	+0.0821	[0.0524, 0.1000]	12/12
35	+0.0862	+0.0849	+0.0470	+0.0809	+0.0748	[0.0564, 0.0856]	12/12
40	+0.0624	+0.0765	+0.0415	+0.0612	+0.0604	[0.0467, 0.0726]	12/12
45	+0.0541	+0.0764	+0.0323	+0.0421	+0.0512	[0.0372, 0.0679]	12/12
50	+0.0473	+0.0565	+0.0250	+0.0234	+0.0380	[0.0242, 0.0519]	12/12

The four official Zhuang-ABCA expression matrices are joined one-to-one to CCF metadata for 5,402,051 cells. In each outer fold, one animal is test-only, one complete animal is used for inner validation, and two animals select genes with a class-balanced χ^2 score, fit the expression classifier, and construct the inner Atlas. After the 2D and 3D fusion weights are selected by inner-validation macro-F1, the classifier and 250-μm Gaussian-smoothed categorical CCF field are refit using all three outer-training animals. The held-out animal is evaluated on at most 150,000 representative cells. Three fixed seeds vary classifier fitting and sampling but reuse the frozen panel-selection protocol; biological N is therefore four.

The classifier is a class-balanced linear logistic model. Its class scores are combined with support-gated Atlas log-likelihood ratios. The 2D control constructs the same class prior after marginalizing the known section-normal axis. The shuffled-depth control permutes only the target cells' normal coordinate, and the wrong-region control displaces the in-plane query by 1 mm. For all 10/25/50-gene panels, every seed passes the prespecified requirement that correct 3D improve on gene-only, 2D, and shuffled-depth controls and that at least three of four held-out animals improve. Correct 3D in fact exceeds every displayed control in all 12 animal-by-seed runs within each panel. These controls establish depth-specific CCF information under known coordinates; they do not model rotation, nonlinear warping, segmentation error, or image-to-CCF registration.

The post-hoc intermediate sweep uses prefixes of exactly the same frozen 50-gene ranking. Across all nine panel sizes, correct 3D exceeds every displayed control in all 12 animal-by-seed runs, and its absolute advantage over the depth-marginalized 2D Atlas decreases monotonically from 0.1649 at ten genes to 0.0380 at 50 genes. Tables 24 and 25 report the full means and animal-level paired effects; the intermediate rows are descriptive rather than additional confirmatory tests.

F.3 Low-panel composition deconvolution

Allen cells are aggregated into 55- and 100-μm pseudo-spots under four-animal leave-one-animal-out evaluation with three panel seeds. With ten genes, correct 3D Atlas features raise composition Pearson from 0.6505 to 0.7859 at 55 μm and from 0.6490 to 0.7814 at 100 μm; JSD changes from 0.264 to 0.186 and from 0.232 to 0.147. The method saturates with 25–50 measured genes, so this result identifies a low-information regime rather than a universal advantage.

G Image Registration and Single-Section Deployment

These examples illustrate deployment from an image-derived pose. They are distinct from the registered mouse benchmarks and do not supply hidden-section accuracy measurements.

G.1 Image-to-CCF registration and anatomical localization

We place each observed H&E section in the Allen CCF before querying the 3D Atlas. This stage estimates the section's anterior–posterior (AP) position and an in-plane coordinate transform from image appearance; it is separate from predicting cell-type composition. Figure 11 shows the four frozen registrations used in the single-section deployment experiment.

Inputs and anatomical priors. Registration uses the public HEST (Jaume et al., 2024) H&E thumbnail, image dimensions, and estimated pixel size, together with the Allen CCFv3 average template and tissue mask on a 25-μm grid. All four sections are treated as coronal, single-hemisphere inputs. The coronal orientation comes from the public titles for TENX43/TENX52 and from the visible hippocampal anatomy for TENX131/TENX181. We match a canonical CCF hemisphere and enumerate in-plane rotations and reflections; no point landmarks are supplied. Expression, cell-type assignments, and predictor outputs are excluded from pose selection.

Image preparation. We identify tissue by saturation greater than 0.075 and mean RGB intensity below 242 on an 8-bit scale, close small gaps, remove small connected components, and fill holes. The tissue bounding box is cropped and resampled at its metadata-derived physical scale to 25 μm per pixel. Color deconvolution supplies the hematoxylin signal. Gaussian-smoothed intensity and Sobel-gradient magnitude at 25, 50, and 100 μm capture internal anatomy in addition to the outer tissue outline. The reference channels use the CCF template, not an expression-derived cell-density map.

Coarse-to-fine pose search. We scan supported coronal CCF planes at 100-μm intervals, considering four right-angle rotations and their mirrored counterparts. Each candidate is initialized by matching tissue centroids and applying an area-based isotropic scale clipped to $[0.80, 1.30]$. All coarse candidates are ranked using both outline and internal structure. The best 48 plane–orientation candidates are expanded to neighboring planes within ± 75 μm at 25-μm spacing; the best 24 expanded candidates undergo constrained affine refinement. Powell optimization uses at most 30 iterations, allowing an additional rotation of $\pm 10^\circ$, independent in-plane scale factors of $[0.92, 1.08]$, and translations of ± 600 μm relative to initialization. The search does not fit tilt away from the coronal plane or a nonlinear deformation.

The alignment score is

$$S = 0.45D + 0.15\bar{r}_1 + 0.15\bar{r}_2 + 0.15I + 0.10\bar{r}_g - 0.05P_s - 0.08P_a - 0.50(1 - c), \qquad (14)$$

where D is tissue-outline Dice, $\bar{r} = (r+1)/2$ rescales a correlation, and r_1, r_2, r_g average correlations across three scales for the two intensity channel groups and gradient magnitude, respectively. Both source intensity groups use hematoxylin at 25/50/100 μm; the corresponding template groups use 20/50/100 and 25/50/100 μm. Correlations are evaluated within the overlapping tissue. I is mean normalized mutual information for the second intensity group, estimated with 32 bins after percentile normalization. For singular values s_1, s_2 of the in-plane transform, $P_s = \sum_j (\log s_j)^2$ and $P_a = |\log(s_1/s_2)|$ discourage excessive scaling and anisotropy. The coverage fraction c penalizes tissue falling outside the target canvas. We retain the highest-scoring refined pose.

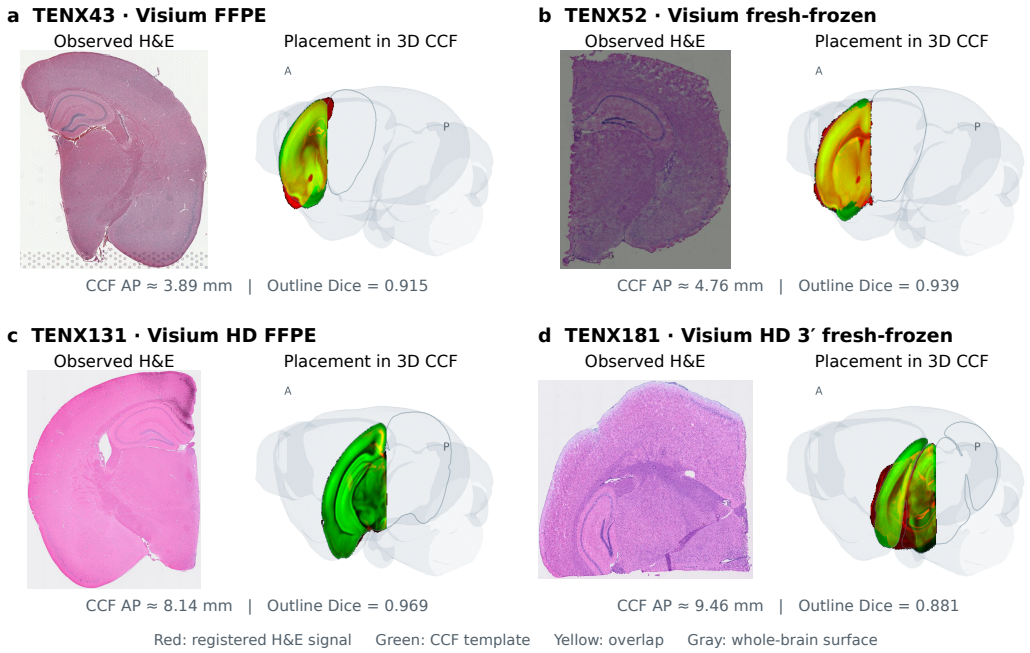


Figure 11: Image-to-CCF registration of four public HEST sections. Each panel pairs the observed H&E image with its frozen registration inside the Allen CCF: (a) TENX43, Visium FFPE; (b) TENX52, Visium fresh-frozen; (c) TENX131, Visium HD FFPE; and (d) TENX181, Visium HD 3′ fresh-frozen. Red denotes the registered image-derived hematoxylin signal, green the CCF template, and yellow their overlap. The translucent gray surface supplies whole-brain context; A/P indicates anterior/posterior. AP position and outline Dice are shown below each pair. H&E crops preserve aspect ratio and are fitted independently for display; the four CCF renderings share a physical scale. Registration uses image appearance and the stated anatomical priors, without expression-based pose selection.

Coordinate transfer and Atlas access. The frozen crop/resampling and affine transforms map each original spot coordinate to an in-plane CCF location (u, v). If k is the selected coronal plane index, its physical CCF coordinate is

$$(x_{\mathrm{AP}}, x_{\mathrm{DV}}, x_{\mathrm{ML}}) = 0.025\,(k + \tfrac{1}{2},\ v + \tfrac{1}{2},\ u + \tfrac{1}{2})\ \mathrm{mm}. \tag{15}$$

The downstream adapter maps these coordinates to its 100-μm Atlas grid, using the same canonical-hemisphere convention. Thus anatomical localization precedes Atlas querying: a named brain region, when needed, is obtained by looking up the CCF annotation at the registered coordinate, not by a separate H&E region classifier. This deployment experiment uses the spatial Atlas fields rather than scoring discrete region labels. The selected plane and transform are frozen before observed expression is loaded; the deployment uses this selected pose rather than averaging over the retained registration candidates.

Registration diagnostics. For TENX43, TENX52, TENX131, and TENX181, the estimated AP positions are 3.8875, 4.7625, 8.1375, and 9.4625 mm, with outline Dice scores of 0.915, 0.939, 0.969, and 0.881. These are coordinates within the CCF volume, not bregma coordinates. Figure 11 overlays the registered hematoxylin signal in red and the template in green, with yellow indicating overlap, inside the CCF whole-brain surface. All four CCF views share the same camera and physical scale. Outline Dice measures tissue overlap; the internal overlays provide complementary qualitative inspection, since these specimens lack landmark-based localization ground truth.

G.2 SINGLE-SECTION PREDICTION WITH THE FINAL MODEL

The four public HEST entries contain one physical section each: Visium FFPE (TENX43), Visium fresh-frozen (TENX52), Visium HD FFPE (TENX131), and Visium HD 3′ fresh-frozen (TENX181). Their image-derived registrations are documented in Appendix G.1. These entries come from separate public dataset pages, but their metadata do not establish four distinct donors.

Figure 12 uses the same frozen predictor checkpoint as Table 1 (SHA-256 prefix `14e3b525`), without retraining on HEST. A fixed nonnegative least-squares adapter estimates eight-class composition from observed expression using 384 reference-selected Allen signature genes. We retain the original input adapter and image-derived poses. The input field combines these proportions with normalized, smoothed spot occupancy; its mass is a sampling proxy, not a measured cell count. Atlas queries use the fractional registered AP coordinate plus each signed offset, followed by the final class-wise calibration, residual transfer, and distance mixture. Predicted mass is not constrained to equal the center's mass. We display only estimated composition at 200-μm resolution. These four single-section entries have no same-specimen hidden-section ground truth, so this is a qualitative deployment example rather than an accuracy evaluation.

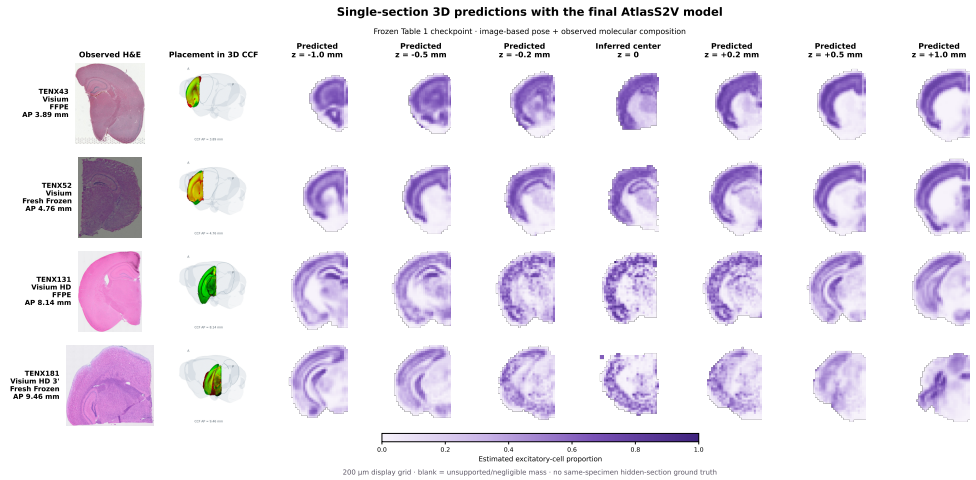


Figure 12: Single-section 3D composition predictions from the final AtlasS2V checkpoint. Rows show four public HEST inputs, their frozen image-to-CCF placement, and estimated excitatory-cell proportions at the center and offsets ± 0.2, ± 0.5, and ± 1 mm. The center is inferred from observed molecular expression, not from H&E alone; the six noncentral fields are model predictions. All maps share a 0–1 color scale and 200-μm display grid; blank locations indicate unsupported queries or negligible mass. H&E crops are fitted independently, while CCF renderings share a physical scale. The AP positions are approximately 3.89, 4.76, 8.14, and 9.46 mm. Same-specimen hidden-section ground truth is unavailable. Registration is documented separately in Fig. 11.

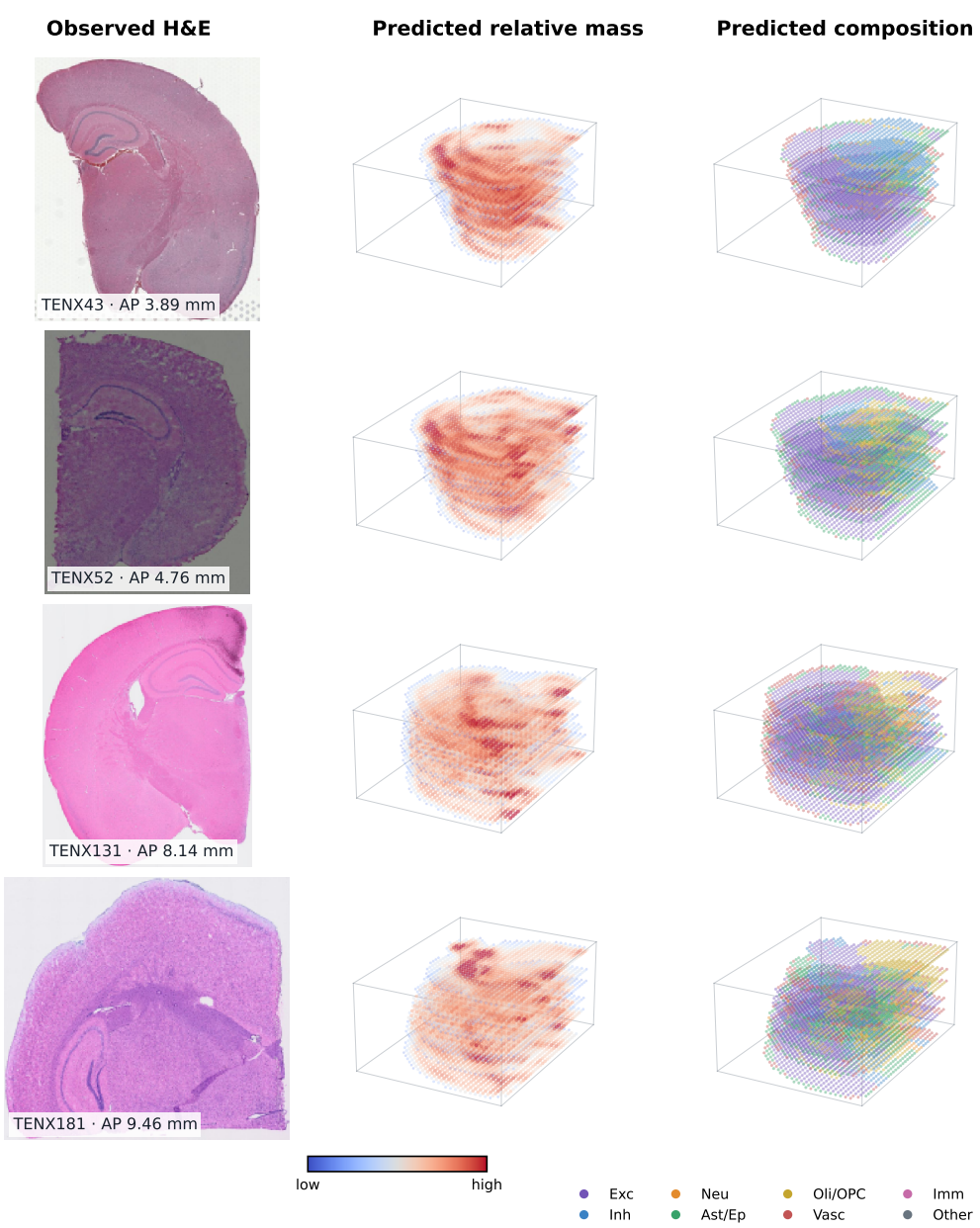


Figure 13: Qualitative virtual 3D fields for the four registered HEST single-section examples. Each row pairs the observed H&E input (left) with six frozen final-model predictions at signed offsets ± 0.2, ± 0.5, and ± 1 mm: relative total field mass (middle; blue low, red high) and dominant eight-class composition (right; colors as in Fig. 7). Points represent 200-μm display bins, not individual cells; the depth axis is expanded for legibility. The mass is derived from normalized spot occupancy and molecular-expression-based central composition, not measured cell density or an H&E-only prediction. Its blue–red scale is normalized separately within each sample and is not comparable across rows. No same-specimen hidden-layer ground truth exists for these examples, so the panels do not establish prediction accuracy.

H IDENTIFIABILITY AND TRANSFER LIMITS

H.1 SINGLE-SECTION AMBIGUITY BOUND

Let two volumes produce the same observed input S_0 and target vectors $\boldsymbol{y}$ and $\boldsymbol{y}'$. A deterministic section-only predictor returns a common value $\hat{\boldsymbol{y}}$. By the parallelogram identity,

$$\frac{1}{2}\left(\|\hat{\boldsymbol{y}}-\boldsymbol{y}\|_2^2+\|\hat{\boldsymbol{y}}-\boldsymbol{y}'\|_2^2\right)=\left\|\hat{\boldsymbol{y}}-\frac{\boldsymbol{y}+\boldsymbol{y}'}{2}\right\|_2^2+\frac{1}{4}\|\boldsymbol{y}-\boldsymbol{y}'\|_2^2. \tag{16}$$

Thus the average error over the two indistinguishable volumes is at least $\frac{1}{4}\|\boldsymbol{y}-\boldsymbol{y}'\|_2^2$. An Atlas cannot guarantee removal of this bound, but it can reduce Bayes risk if it provides information about which volume is more plausible conditional on pose and organ context.

H.2 CONTINUOUS-VOLUME INFORMATION AUDIT

For each target cell, we compare its true 32-nearest-neighbor set in 3D with an equal-count context sampled from a 10-μm section. Boundary-crossing neighborhoods are excluded. Sampling-noise controls compare two independent 32-cell samples from the nearest 64 three-dimensional candidates. Confidence intervals use spatial-block bootstrap. The section misses 0.824/0.826 of true neighbors in the two volumes; composition JSD is 0.1102/0.1280 versus 0.0598/0.0709 under sampling noise, and gene-program error is 0.0322/0.0354 versus 0.0163/0.0175.

H.3 PLATFORM SHIFT

Development experiments evaluated Atlas-conditioned residual transfer on 101 Slide-seq sections and 75 Ortiz spatial-transcriptomics sections. On Slide-seq, correct-depth composition reaches 0.237 at 1 mm versus 0.183 for central-copy and 0.126 for wrong depth, but a shared scalar gate degrades total mapped mass. On Ortiz, Atlas density is helpful at 2 mm (Pearson 0.407 versus 0.308 for central-copy at a 400-μm grid) but harms performance at 400 μm and 1 mm. These mixed development results motivate task-specific calibration; they are not an external evaluation of the final Table 1 checkpoint.

H.4 ATLAS CONTEXT DOES NOT TRANSFER UNIVERSALLY

A frozen external ageing task selects zero Atlas shrinkage, and a DAPI-native image encoder does not add information beyond a coarsely positioned Atlas. These negative results limit the scope of the positive downstream findings: they do not establish a universal transferable representation. Target-relevant molecular features and a plausible anatomical pose are presently more reliable than morphology-only transfer across acquisition domains.

I REPRODUCIBILITY RECORD

Model identity and inputs. The final count-field predictor is identified by checkpoint SHA-256 prefix `14e3b525`. Table 1 and the independent-mouse visualizations use this calibrated expert mixture and distance policy. The outer-animal composition study refits the same structure within each training fold. Marker density, molecular fields, and organ-specific studies use the adaptations specified in their respective protocols; downstream imputation and typing fit separate readouts.

Figure 1 illustrates the workflow. Its organ-routing curves and tissue maps are schematic, not measured weights or experimental outputs. Registration, abundance calibration, and residual transfer precede deterministic density and composition readouts; network-like symbols represent this analytic mixture. Main mouse benchmarks use supplied CCF coordinates, whereas HEST deployment uses image-derived registration.

Reconstructing figures and tables. The compact publication files preserve frozen inputs, source hashes, scoring pairs, and all reported aggregates:

- **Single-mouse fields and baselines.** `figures/make_figure3_3d_fields.py` rebuilds Fig. 2 from the composition NPZ, expression JSON, and final-model 3D

fields. `tables/make_table1_matched_benchmark.py` rebuilds the main and full-distance comparisons. Their records retain the checkpoint identity, pairing geometry, and U-Net seed results.

- **Cross-specimen and organ results.** `figures/make_figure3_generalization_organs.py` rebuilds Fig. 3 from the BrainCAD and multiple-organ records. `tables/make_final_composition_loao.py` reconstructs all outer-animal composition tables, including missing distance bins. The HER2ST source record contains all patient, distance, layer, resolution, and volume metrics.
- **Expression and downstream results.** `figures/make_figure2_expression_braincad.py` rebuilds the expression comparisons from `figures/figure2_expression_data.json`. The full record includes computational repetitions omitted from the compact distance table. The downstream JSON and NPZ files listed in `figures/README.md` retain section- and animal-resolved curves, bootstrap effects, displayed maps, and the map-selection rule.
- **Deployment and implementation checks.** The HEST publication record `figures/appendix_single_section_virtual_fields_data.json` identifies the input adapter, frozen poses, Atlas, checkpoint, and prediction hashes. Its NPZ stores all displayed fields. All 24 virtual-layer predictions agree with the final analytic mixture; these are implementation checks, not accuracy measurements. `figures/make_appendix_3d_fields.py` rebuilds the independent-mouse and HEST 3D views.

Benchmark availability. SpatialZ requires two observed flanking sections, so it is not part of the single-section ranking. No matched final-predictor result is reported for that privileged-input protocol. STITCH (Wang et al., 2026) was not run: the repository snapshot inspected on September 14, 2026 (commit `d203288`, `https://github.com/WANG-SUI/STITCH`) did not contain an executable implementation. Neither method is represented by a substitute baseline.

Release contents. The release will include dataset manifests, specimen splits, frozen configurations and checkpoints, preprocessing specifications, and commands for regenerating the tables and figures. Source records retain the distinction between biological replicates, computational seeds, and post-hoc characterization. Machine-readable detail complements, rather than replaces, the reported aggregate comparisons and controls.