DEPTH-RESOLVED GENE EXPRESSION ESTIMATION FROM TISSUE MORPHOLOGY AND SPARSE SPATIAL TRANSCRIPTOMICS

Anonymous authors
Paper under double-blind review

ABSTRACT

Most spatial transcriptomics assays measure gene expression on a limited number of two-dimensional tissue sections, leaving molecular variation across depth largely unobserved. We investigate whether volumetric tissue imaging can provide the structural context needed to estimate this missing molecular information from sparse 2D measurements. We introduce VISTA (Volumetric Inference from Sparse Transcriptomic Anchors), a structure-guided framework for estimating depth-resolved gene-expression fields from 3D tissue imaging using 2D spatial transcriptomics measurements as molecular anchors. VISTA combines a frozen pathology foundation encoder with multi-scale structural encoding, depth-aware attention, and 3D graph propagation to integrate local and long-range tissue context, while a Schrödinger-bridge objective along depth encourages coherent predictions that remain consistent with observed ST measurements and preserve 3D structural organization. A single formulation covers three settings that differ in how directly depth-resolved prediction can be checked. On registered serial sections, predictions are checked against held-out measured sections, under supervision as thin as a single revealed plane. On a serial volume in which every section is sequenced, depth accuracy is measured directly against held-out expression and reported against a split-half noise ceiling that bounds what is recoverable; VISTA attains roughly half of it, against a few percent for interpolation. On continuous volumes, where only one plane is sequenced and no method can report per-depth accuracy, we isolate the contribution of depth with a control that replaces the volume by its own projection. Across settings, VISTA improves agreement with measured ST and produces more coherent and depth-sensitive molecular estimates than existing baselines. These results support structure-guided volumetric modeling as a practical approach for extending sparse sectional transcriptomic measurements into depth-resolved molecular representations.

1 INTRODUCTION

Spatial transcriptomics enables molecular profiling within tissue context by measuring gene expression together with spatial coordinates (Stahl et al., 2016). Recent imaging-based and sequencing-based technologies, including Visium, Slide-seq, MERFISH, and seqFISH, have enabled the study of spatial domains, cellular organization, and tissue microenvironments beyond dissociated single-cell assays (Rodriques et al., 2019; Chen et al., 2015; Eng et al., 2019). However, most ST technologies remain fundamentally two-dimensional, profiling only thin tissue sections despite tissues being inherently three-dimensional. This limitation is particularly important in heterogeneous systems such as tumors, developing organs, and structured epithelia, where cellular interactions and molecular programs vary substantially across depth (Davidson, 2012; Gómez-Gálvez et al., 2021).

Recovering molecular organization in 3D is critical for understanding tissue architecture and disease progression (Zhu et al., 2025; Almagro-Pérez et al., 2025). Continuous 3D gene expression maps can reveal how tumor niches, immune infiltration, stromal organization, and developmental gradients extend throughout tissue volume rather than within a single plane. However, dense 3D ST acquisition remains expensive, destructive, and technically challenging. Serial sectioning provides sparse measurements across depth but does not produce a continuous molecular landscape

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.

(Schott et al., 2024a). These limitations motivate computational approaches that infer volumetric gene expression from sparse molecular supervision and dense tissue morphology.

We introduce VISTA (**V**olumetric **I**nference from **S**parse **T**ranscriptomic **A**nchors), a morphology-guided framework for learning depth-resolved 3D gene-expression fields from volumetric tissue data under limited 2D ST supervision. VISTA uses dense volumetric morphology as the primary inference substrate and predicts gene expression across all spatial locations and depth slices. Rather than only matching measured ST sections, the framework learns a spatially coherent molecular field throughout the tissue volume. To achieve this, VISTA integrates a frozen pathology foundation encoder, multi-scale morphological encoding, volumetric attention, graph-based spatial propagation, and Schrödinger-bridge regularization along depth within a unified architecture. Unlike methods that align or interpolate a small number of measured ST sections, VISTA directly models continuous volumetric molecular organization under sparse supervision. The same framework applies to stacked serial-section ST data, where registered 2D sections are treated as sparse measured planes across depth: the encoder, 3D context modeling, and training objective are shared, and only the supervision available at each depth differs. This lets us evaluate a single model on continuous volumetric data, on sparse serial sections, and on Open-ST, a serial volume in which every section is sequenced and held-out depths therefore carry measured ground truth. Our main contributions are as follows:

(i) We propose VISTA, a morphology-guided framework that predicts depth-resolved gene expression from tissue morphology under limited ST supervision, combining a frozen pathology foundation encoder with multi-scale 3D encoding, depth-aware attention, multi-level 3D graph propagation, and attention-based pooling to the measured support. One formulation serves both densely imaged volumes and registered serial sections; only the supervision available along depth differs.

(ii) We verify a single claim that depth-varying molecular signal can be extracted from morphology from three directions of decreasing verification directness. On registered HER2+ serial sections, predictions are checked against held-out measured sections, including a protocol in which a single section is revealed and five are predicted.

(iii) Using Open-ST, where every serial section is sequenced, we check depth-resolved prediction *directly* against measured expression at withheld depths, and report it against a split-half noise ceiling that bounds what is recoverable at all. To our knowledge this is the first such measurement for this family of models; it converts depth-resolved reconstruction from an assumption into a quantity.

(iv) On continuous 3D HT volumes, where only one plane is sequenced and no method can report per-depth accuracy, we isolate the contribution of the depth dimension with a control that replaces the volume by its own maximum-intensity projection while holding architecture, objective, and splits fixed. We further report volumetric structure criteria—spatial coherence, morphology coupling, boundary enrichment—as a practical evaluation strategy for settings that lack dense molecular ground truth.

2 Preliminaries and Motivation

We study the problem of predicting a depth-resolved molecular field from a 3D tissue volume when gene expression is measured on only one section. The volume contains Z depth slices of spatial size $H \times W$ and is divided into tiles. The goal is to predict a gene-expression vector $\widehat{y}_{t,z} \in \mathbb{R}^G$ for every tile–slice pair (t, z), where G is the number of target genes. The collection of all tiles' predictions forms a continuous 3D molecular field.

Problem statement. For each measured ST spot s, let $\mathcal{T}_s$ be the set of tiles whose aligned centers fall inside that spot. The observed target is the spot-level expression vector $y_s \in \mathbb{R}^G$. Because depth-resolved expression is never directly observed, supervision is available only after aggregating tile–slice predictions to the spot level. The model must learn a mapping such that the aggregated prediction $\widehat{y}_s$ agrees with y_s, representing a weakly supervised inverse problem in which the model predicts a dense 3D field from sparse 2D labels. This is underdetermined: many 3D fields can produce the same 2D measurement after aggregation. A useful solution must remain consistent with observed ST while producing depth-resolved predictions that are spatially coherent and morphology-guided rather than being constant across depth.

Stacked serial-ST setting. The same problem arises when the volume is sampled by serial sectioning rather than imaged densely. Here we observe L aligned ST sections at depths z_ℓ ($\ell = 1, \ldots, L$), each with measured expression $y_{i,\ell}$. Since these sections may be separated by large physical gaps, depth is represented by an irregular-depth graph over the observed sections rather than by uniform slice indices, enabling cross-section propagation without imposing volumetric continuity. Supervision is otherwise the same: a subset of planes is revealed and the remainder withheld, which reduces to the single-section case when only one plane is measured.

A common unit across settings. The three assays we study name their spatial unit differently. Throughout, we call the unit at which predictions are compared with measurements a *site*: a Visium spot in the continuous 3D setting, a registered ST spot on stacked serial sections, and a $100\,\mu$m bin on Open-ST. Site correlation therefore means the same quantity everywhere, computed over whichever unit the assay provides. The *tile* remains distinct from the site: on dense volumes many tiles fall within one site and are pooled to it, whereas on serial sections and Open-ST the prediction and measurement units coincide.

3 METHODOLOGY

Building on Section 2, VISTA infers 3D gene expression from limited 2D spatial transcriptomics and volumetric tissue morphology by combining morphology-aware encoding, attention-based spatial integration, graph propagation, and depth regularization using the pipeline depicted in Figure 1. We describe one formulation, used unchanged across settings; only the supervision available along depth differs. Frequently used symbols are summarized in Table 1; additional notation is defined where first introduced.

Table 1: Key symbols used in methodology.

Sym.	Meaning	Sym.	Meaning
s	Measurement site (spot or bin)	D	Latent embedding dimension
t	Tissue-volume tile	B, K	Batch size, max tiles per spot
z	Depth slice index ($1 \ldots Z$)	X_t	Tile volume $\in \mathbb{R}^{C \times Z \times H \times W}$
u	Token-grid position ($1 \ldots U$)	y_s	Measured expression $\in \mathbb{R}^G$
g	Gene index ($1 \ldots G$)	$\mathcal{T}_s$	Valid tiles assigned to spot s

3.1 PREPROCESSING AND SPOT-TILE CONSTRUCTION

Tile centers are transformed into the Visium coordinate system via the available 3D-to-2D registration. A tile is assigned to the spot whose disk contains its aligned center; unmatched spots and tiles are removed. The Visium count matrix is restricted to the selected gene panel and standardized using log transform. Training proceeds in spot-level batches. Each batch is a padded tensor $X \in \mathbb{R}^{B \times K \times C \times Z \times H \times W}$ accompanied by a binary tile mask $m \in \{0, 1\}^{B \times K}$ that prevents padded entries from affecting graph construction, pooling, or losses. Further preprocessing details are in Appendix A.1.

3.2 HYBRID MULTI-SCALE 3D TISSUE-VOLUME ENCODER

Displayed in the top left panel of Figure 1, the encoder converts each tile volume X_t into per-slice token representations that capture fine in-plane detail, coarse low-frequency morphology, and local volumetric context. Each slice is first embedded by **UNI2-h** (Chen et al., 2024), a pathology foundation model that stays frozen throughout, whose patch tokens are pooled to the working token grid. A lightweight 2D convolutional stem runs alongside it, and the concatenated slice features feed two token branches. The *fine* branch tokenizes the full-resolution feature map into a grid of local tokens; the *coarse* branch downsamples before tokenization and upsamples the resulting tokens back to fine resolution, providing a larger effective receptive field. In parallel, a shallow 3D convolutional stem processes the full tile volume, preserving Z while downsampling XY to the same token grid. Let $v^{\text{fine}}_{t,z,u}$, $v^{\text{coarse}}_{t,z,u}$, and $v^{\text{3D}}_{t,z,u}$ denote the three branch token vectors. VISTA fuses them via a learned per-token softmax gate:

$$r_{t,z,u} = \sum_{q \in \{\text{fine},\text{coarse},\text{3D}\}} \pi^{(q)}_{t,z,u}\, v^{(q)}_{t,z,u}, \qquad \sum_q \pi^{(q)}_{t,z,u} = 1, \tag{1}$$

where the coefficients $\pi^{(q)}_{t,z,u}$ are produced by a small MLP from the concatenated branch tokens, allowing the model to adaptively weight fine, coarse, and volumetric cues at each position. Because the frozen encoder already carries a strong morphological prior, the trainable parameters are spent

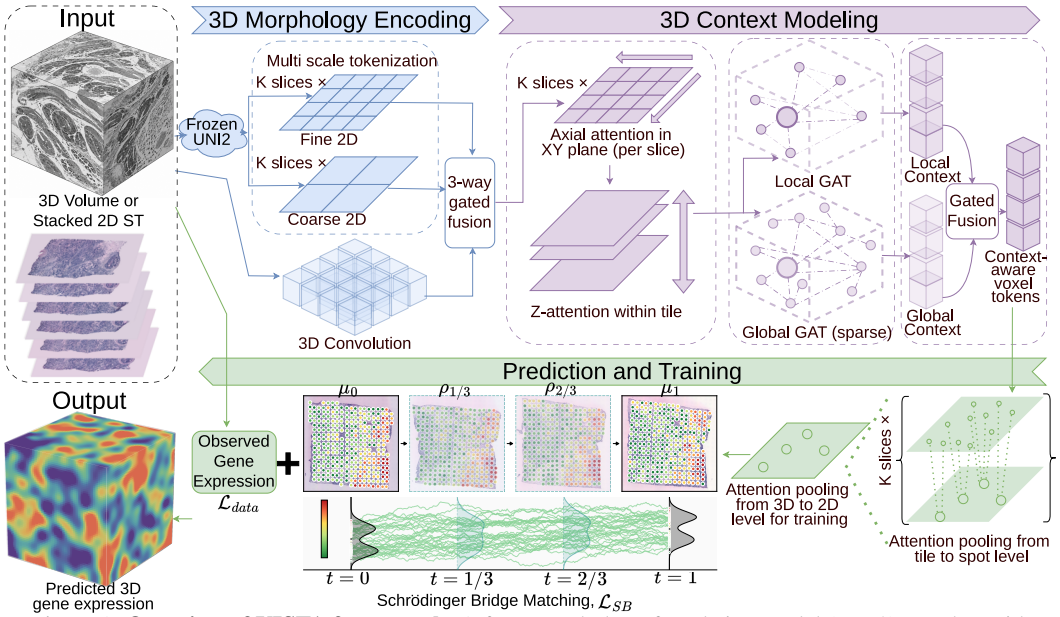


Figure 1: **Overview of VISTA framework.** A frozen pathology foundation model (UNI2) together with multi-scale 2D and shallow 3D encoders process the unstained 3D tissue volume (top left); depth-aware attention and local/global 3D graph attention incorporate 3D context (top right); attention pooling maps tile-level predictions to spots and 2D (bottom right); parts of the objective function including Schrödinger Bridge matching (bottom middle); and predicted 3D gene expression (bottom left)

on depth context and spatial propagation rather than on relearning tissue appearance from scratch. Architectural details including stem configuration and patch sizes in Appendix A.2.

3.3 In-plane axial attention and within-tile z-attention

As depicted in the top central panel of Figure 1, 3D context modeling starts with refining the fused tokens $r_{t,z,u}$ by XY axial attention, which applies self-attention separately along the horizontal and vertical axes of the token grid. This captures long-range in-plane context at lower cost than full 2D attention, followed by a feed-forward block with residual connections and layer normalization.

Tokens within each slice are then pooled to a slice embedding $h_{t,z} \in \mathbb{R}^D$. A lightweight z-attention module processes $\{h_{t,z}\}_{z=1}^{Z}$ within each tile, modeling depth dependencies without cross-slice token-level attention. Its output $\widetilde{h}_{t,z}$ serves as the node feature for graph propagation. Layer configurations and complexity analysis are in Appendix A.3.

3.4 Sparse local and global graph attention

Each tile–slice pair (t, z) defines a graph node with feature $\widetilde{h}_{t,z}$ and 3D coordinate $p_{t,z} = (x_t, y_t, \eta_z z)$, where η_z controls the relative importance of depth versus XY distance. The local graph $\mathcal{G}_{\text{loc}}$ connects k-nearest neighbors in this space; the global graph $\mathcal{G}_{\text{glob}}$ connects all valid tiles within the same spot and z-plane. Both are processed by multi-head graph attention (Veličković et al., 2018), yielding embeddings $h_{t,z}^{\text{loc}}$ and $h_{t,z}^{\text{glob}} \in \mathbb{R}^D$. Attention coefficients and layer configuration are in Appendix A.9.

3.5 Context fusion and slice-level gene prediction

The z-attended embedding, local output, and global output are combined via a learned sigmoid gate $g_{t,z} \in (0, 1)^D$:

$$h_{t,z}^{\text{fuse}} = g_{t,z} \odot \widetilde{h}_{t,z} + (1 - g_{t,z}) \odot \phi\Big([\widetilde{h}_{t,z}; h_{t,z}^{\text{loc}}; h_{t,z}^{\text{glob}}]\Big) \tag{2}$$

where ϕ is a learned linear map and $\odot$ is elementwise multiplication. This gated residual form preserves pre-graph slice identity while incorporating graph-propagated context. Slice-level expression is predicted as $\widehat{y}_{t,z} = \psi(h_{t,z}^{\text{fuse}}) \in \mathbb{R}^G$, where ψ is an MLP gene head. The gate and its texture-dispersion input are defined in Appendix A.4.

3.6 ATTENTION POOLING FOR WEAK SPOT-LEVEL SUPERVISION

Demonstrated in the top right panel of Figure 1, VISTA aggregates slice predictions to the spot level via two-stage attention pooling $\widehat{y}_t = \sum_{z=1}^{Z} \alpha_{t,z}\, \widehat{y}_{t,z}$ and $\widehat{y}_s = \sum_{t \in \mathcal{T}_s} \beta_{s,t}\, \widehat{y}_t$ where $\alpha_{t,z}$ and $\beta_{s,t}$ are learned attention weights satisfying $\sum_z \alpha_{t,z} = 1$ and $\sum_t \beta_{s,t} = 1$. Padded tiles receive zero contribution via the mask. This links slice-level predictions to the measured spot-level target y_s. Softmax formulation details with learned logits and temperature parameters in Appendix A.6.

3.7 TRAINING OBJECTIVE

The total loss combines spot-level fidelity with a Schrödinger-bridge matching term along depth: $\mathcal{L} = \mathcal{L}_{\text{data}} + \lambda_{\text{SB}} \mathcal{L}_{\text{SB}}$, where $\mathcal{L}_{\text{data}}$ applies Smooth L1 between $\widehat{y}_s$ and y_s. $\mathcal{L}_{\text{SB}}$ treats the revealed planes as the endpoints of a diffusion bridge over depth and regresses the predicted increment onto the Brownian-bridge target between them, constraining how expression may change between measured planes without forcing it to change smoothly. Relational structure across views, adjacent depths, and graph neighbours enters through the positive and contrastive channels of Section 3.2, which shape the representation directly. Full definitions are in Appendix A.7 and A.8.

3.8 SUPERVISION ACROSS SETTINGS

The encoder, axial and depth attention, graph propagation, and training objective described above are used unchanged in every setting; what differs is how predictions are compared with the measurement. When a dense volume is paired with one matched Visium section, the measurement integrates the whole depth column, so predictions are compared after the attention pooling of Section 3.7. When registered serial sections are measured, each section is a plane, so predictions are compared at that plane and depth enters through the physical-depth graph: nodes carry coordinates (x, y, z) in micrometres, so irregular gaps between sections are represented as measured rather than as uniform slice indices. Training uses masked-plane reconstruction in both cases: a subset of planes is revealed and the remainder withheld, which reduces to single-section supervision when only one plane is measured. Details in Appendix A.5.

4 EXPERIMENT

Tissue morphology carries recoverable depth-varying molecular signal, and VISTA recovers it. We verify this from three directions, which differ in how directly the verification can be made rather than in what is claimed. On registered HER2+ serial sections (Section 4.1), the check is against held-out measured sections. On Open-ST (Section 4.2), where every section of the volume is sequenced, it is a direct check against measured expression at withheld depths, quantified against a split-half noise ceiling that bounds what is recoverable at all. On continuous holotomography volumes (Section 4.3), where only one plane is measured, per-depth accuracy cannot be computed for any method, so the check is whether the volumetric input contributes information its own 2D projection does not, together with volumetric structure criteria. An ablation over every architectural and loss component closes the section.

4.1 APPLICATION TO PSEUDO-3D VOLUMES FROM STACKED 2D ST SECTIONS

We begin where predictions can be checked against measured sections. We evaluate on pseudo-3D volumes constructed by registering (details in Appendix E) and stacking 2D ST sections from the HER2+ breast cancer dataset (Andersson et al., 2021), containing four samples with six sections and four samples with three sections. These data consist of sparse serial sections with measured gene expression available at each section, so held-out sections provide direct supervision and test whether graph-based cross-section modeling improves prediction across discrete depth planes. We evaluate the model in two settings: a six-section setting where samples A and B are used for training, C for validation, and D for testing, and a three-section setting where E and F are used for training, G for validation, and H for testing. Performance is measured using per-site Pearson correlation on the held-out sample. The baselines used for comparison are explained in Appendix G.

Appendix D.1 additionally reports a strict one-section-in, five-sections-out protocol, in which VISTA improves the benchmark mean from 0.545 for the best baseline to 0.631. Table 2 shows that VISTA achieves the strongest learned-model performance on both stacked-ST settings. On Sample D, where spatial heterogeneity is more pronounced, VISTA obtains the best average correlation and leads across most individual sections, indicating that graph-based cross-section propagation ef-

Table 2: VISTA on pseudo 3D volume (site correlation ↑). **Best** in bold, second best underlined.

		Sample D							Sample H			
	Model	D1	D2	D3	D4	D5	D6	Avg	H1	H2	H3	Avg
Naive	kNN-IDW	0.277.001	0.291.002	0.342.002	0.294.002	0.291.002	0.313.002	0.301.002	0.476.002	0.419.002	0.458.002	0.451.002
2D Methods	Hist2ST	0.398.002	0.462.002	0.455.002	0.309.002	0.326.003	0.322.003	0.379.002	0.474.002	0.494.002	0.459.002	0.475.002
	THItoGene	0.462.002	0.500.002	0.495.002	0.201.001	0.134.003	0.131.002	0.321.002	0.475.002	0.485.002	0.460.002	0.473.002
	STGAT	0.465.002	0.469.002	0.466.002	0.462.002	0.458.002	0.481.002	0.467.002	0.437.002	0.452.003	0.426.002	0.438.002
	BLEEP	0.456.002	**0.572**.002	0.455.002	0.446.002	0.440.003	0.530.002	0.483.002	0.330.003	0.350.002	0.348.002	0.343.002
	TRIPLEX	0.386.002	0.365.003	0.450.002	0.324.002	0.355.002	0.448.002	0.388.002	0.418.002	0.431.002	0.398.002	0.416.002
3D Methods	X-pression	0.226.001	0.204.002	0.223.002	0.197.002	0.214.002	0.223.002	0.215.002	0.486.002	0.495.002	0.467.002	0.483.002
	ASIGN	0.482.002	0.497.002	0.487.002	0.489.002	0.478.002	0.513.002	0.491.002	0.482.002	0.492.002	0.467.002	0.480.002
	VISTA	**0.533**.002	0.547.002	**0.547**.002	**0.548**.002	**0.554**.002	**0.572**.002	**0.550**.002	**0.545**.002	**0.555**.002	**0.524**.002	**0.541**.002

fectively exploits spatial correspondence between aligned serial sections. Despite constituting lower spatial heterogeneity in case of sample H, VISTA demonstrated consistent performance surpassing all the baselines showing the generalization power of the model across varying scenarios.

Figure 2 provides a qualitative assessment on test Sample D using two breast cancer associated biomarkers- *MYL6* has been linked to tumor aggressiveness (Deguchi et al., 2024), while *GPX4* has been associated with tumor progression and therapeutic resistance in triple-negative breast cancer (Gu et al., 2025). The predicted heatmaps broadly agree with the measured ground truth expression patterns, and they also show more visual alignment and clear spatial separation between tumor and non-tumor regions when compared with the pathologist-annotated histology. This supports the ability of VISTA to learn biologically meaningful spatial molecular patterns from stacked ST sections.

These results establish that the framework predicts unmeasured sections better than published alternatives, and that it does so under supervision as thin as a single measured plane, including on an unseen test patient, where VISTA leads in every sample but one (Table 3). They do not yet establish that the predicted variation *along depth* is correct, because each location is measured at only a few widely separated depths. Section 4.2 supplies that check.

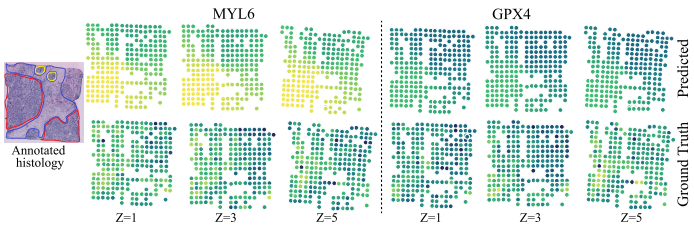


Figure 2: **Stacked 2D sections**: predicted site-level expression for MYL6 and GPX4 on test sample D of the HER2+ dataset, against measured ground truth and pathologist-annotated histology.

Table 3: **Unseen test patient**: site correlation, mean over all sections (Appendix D.2).

Model	A	B	C	D
VISTA	**0.642**	0.368	**0.685**	**0.557**
BLEEP	0.617	**0.380**	0.671	0.543
THItoGene	0.628	**0.380**	0.670	0.524
Hist2ST	0.582	0.366	0.583	0.522
TRIPLEX	0.611	0.230	0.660	0.532
X-pression	0.536	0.325	0.630	0.520
STGAT	0.505	0.322	0.627	0.506
ASIGN	0.478	0.333	0.625	0.490
kNN-IDW	0.525	0.336	0.589	0.464

4.2 Depth-resolved evaluation on Open-ST

The setting above checks predictions at held-out planes, but cannot verify a predicted depth *profile*: each location is measured at only a few depths, separated by tens of micrometres. Open-ST (Schott et al., 2024b) removes that restriction, sequencing every section of a serial volume and providing measured expression at 19 depths on a shared 100 μm bin grid. We adopt a two-anchor protocol: the first and last sections are revealed and the 17 sections between them are withheld, matching the case where only the outer faces of a tissue block are assayed. The framework is applied unchanged, with the same encoder, 3D context modeling, and objective used in the other settings.

We report depth correlation: for each site trajectory, the Pearson correlation between predicted and measured expression across depth, centred within each section so that in-plane structure cannot contribute. Because this quantity is bounded by measurement noise, we also estimate a split-half noise ceiling by dividing the cells within each bin, scoring the halves against each other, and applying a Spearman–Brown correction. The ceiling is 0.795, so depth correlations should be read against that value rather than against 1.0. Protocol and trajectory-support details are in Appendix B.

Table 4: VISTA on the Open-ST serial volume under the two-anchor protocol. Depth correlation is reported against the split-half noise ceiling of 0.795 measured on the same support. **Best** in bold, second best underlined.

		Similarity with measured ST			Depth	
	Model	Site corr. ↑	Gene corr. ↑	MSE ↓	Depth corr. ↑	% of ceiling ↑
Naive	kNN-IDW	0.6462	0.2949	0.7617	0.0460	5.8
2D	BLEEP	0.7131	0.3753	0.5497	0.2196	27.6
2D	TRIPLEX	0.7465	0.4504	0.4781	0.2713	34.1
3D	X-pression	0.6728	0.3122	0.6339	0.0971	12.2
3D	ASIGN	0.6403	0.0447	0.7589	0.0116	1.5
3D	VISTA	**0.8261**	**0.5933**	**0.3499**	**0.4223**	**53.1**

Table 4 shows that VISTA recovers 53% of the attainable depth signal, ahead of TRIPLEX at 34% and BLEEP at 28%, while X-pression reaches 12% and kNN-IDW and ASIGN only 5.8% and 1.5%. Two observations follow. First, the 2D histology-to-expression models recover a substantial share of the depth signal from per-section morphology alone, without any 3D context modeling; this supports the premise that tissue morphology carries depth-varying molecular information, and makes them considerably stronger baselines here than interpolation. Second, the ranking on depth does not follow the ranking on in-plane agreement alone: kNN-IDW attains a site correlation of 0.646 while carrying almost no depth information, because interpolation between two revealed faces can only produce a monotone profile along depth. The two 3D baselines are also furthest from their design points here, since both were built to impute a stack from a single measured ST slide rather than to extrapolate across seventeen withheld planes; X-pression additionally predicts expression *patterns* rather than raw expression, which its adaptation to this task must invert. This confirms, against measured ground truth, that the depth variation VISTA predicts is genuine rather than an artifact of smoothing. It also establishes a point we rely on in Section 4.3: agreement with measured ST in-plane is a weak proxy for depth fidelity, so a method can look competitive on site-level correlation while carrying no depth information at all.

Degradation with distance from the revealed section. Prediction quality should fall as the target plane moves away from the nearest measured section. Table 5 reports this for HER2+ and Table 6 for Open-ST, where the gap is measured to the nearer of the two revealed faces.

Table 5: HER2+: site correlation by physical distance from the revealed section.

Gap	VISTA	BLEEP	TRIPLEX	X-pr.	ASIGN
32 μm	**0.661**	0.555	0.544	0.526	0.356
64 μm	**0.648**	0.550	0.545	0.523	0.343
96 μm	**0.617**	0.545	0.543	0.520	0.339
128 μm	**0.600**	0.541	0.539	0.521	0.338
160 μm	**0.565**	0.523	0.550	0.496	0.325

Table 6: Open-ST: site correlation by distance to the nearer revealed face.

Gap	VISTA	BLEEP	TRIPLEX	X-pr.	ASIGN
$\leq$40 μm	**0.830**	0.705	0.746	0.680	0.660
41–80 μm	**0.827**	0.734	0.773	0.699	0.587
81–120 μm	**0.836**	0.734	0.764	0.660	0.616
121–160 μm	**0.838**	0.712	0.726	0.680	0.663
$>$160 μm	**0.792**	0.644	0.700	0.606	0.660

On HER2+, VISTA degrades from 0.661 at 32 μm to 0.565 at 160 μm while retaining the largest margin over every baseline at every gap. On Open-ST the degradation is markedly milder, from 0.830 to 0.792 across a wider span, which is consistent with Open-ST's denser 100 μm bin grid and with both faces of the block being revealed rather than one.

4.3 Application to continuous 3D volumes (HT)

Sections 4.1 and 4.2 checked the claim where expression is measured at several depths. We now apply the same framework to continuous volumes, where morphology is dense but only one plane is sequenced. Here no method can report per-depth accuracy, so we ask instead whether the volumetric input contributes information its own 2D projection does not, and whether the predicted field satisfies volumetric structure criteria. These experiments use two gastric-cancer samples, RN49 and HL55, each containing a 65-slice holotomography (HT) volume registered to a matched 2D Visium section.[1] The HT volume supplies dense 3D morphology but no molecular measurement; Visium supplies site-level expression on one plane. After registration with STalign (Clifton et al., 2023), HT tiles are mapped to Visium sites and the model predicts expression for every tile at every depth; registration quality is quantified in Appendix E. Both samples contain regions tumor0 and tumor1 and share 370 genes; RN49 has $>$700 sites and $\approx$15,000 mapped tiles, HL55 $\approx$900 and $>$19,000.

[1]The dataset is not publicly available at submission time because it is governed by institutional data-use agreements. We will release the data upon publication, subject to institutional approval.

Table 7: Application of VISTA on the continuous 3D HT volume. The intra-tissue row shows the 20% held-out fraction; 40% and 80% are in Appendix D.3. Best per setting in **bold**.

		3D Structure Quality				Similarity with 2D ST	
	Model	Moran's $I \uparrow$	Geary's $C \downarrow$	Morphology*	BES $\uparrow$	Site corr. $\uparrow$	MSE $\downarrow$
Intra-tissue	kNN-IDW	$0.6918_{\pm 0.0023}$	$0.3083_{\pm 0.0017}$	$0.0408_{\pm 0.0012}$	–	$0.5640_{\pm 0.0025}$	$\mathbf{0.1431}_{\pm 0.0014}$
	ASIGN	$0.2867_{\pm 0.0014}$	$0.7188_{\pm 0.0021}$	$0.0408_{\pm 0.0011}$	$1.05_{\pm 0.02}$	$0.5388_{\pm 0.0022}$	$0.1660_{\pm 0.0019}$
	VISTA	$\mathbf{0.8143}_{\pm 0.0018}$	$\mathbf{0.1858}_{\pm 0.0015}$	$\mathbf{0.3130}_{\pm 0.0024}$	$\mathbf{1.54}_{\pm 0.03}$	$\mathbf{0.5648}_{\pm 0.0019}$	$0.1646_{\pm 0.0021}$
Cross-tissue	kNN-IDW	$0.6891_{\pm 0.0024}$	$0.3109_{\pm 0.0019}$	$0.0544_{\pm 0.0013}$	–	$0.3375_{\pm 0.0017}$	$0.1502_{\pm 0.0020}$
	ASIGN	$0.1417_{\pm 0.0015}$	$0.8665_{\pm 0.0026}$	$0.0721_{\pm 0.0014}$	$1.09_{\pm 0.02}$	$0.3902_{\pm 0.0021}$	$0.1151_{\pm 0.0016}$
	VISTA	$\mathbf{0.7450}_{\pm 0.0019}$	$\mathbf{0.2518}_{\pm 0.0023}$	$\mathbf{0.2942}_{\pm 0.0018}$	$\mathbf{1.51}_{\pm 0.03}$	$\mathbf{0.4457}_{\pm 0.0025}$	$\mathbf{0.1079}_{\pm 0.0014}$
Cross-sample	kNN-IDW	$0.6076_{\pm 0.0016}$	$0.3918_{\pm 0.0023}$	$0.0396_{\pm 0.0012}$	–	$0.2813_{\pm 0.0020}$	$0.2659_{\pm 0.0022}$
	ASIGN	$0.1140_{\pm 0.0018}$	$0.8605_{\pm 0.0022}$	$0.0984_{\pm 0.0016}$	$1.14_{\pm 0.02}$	$0.1471_{\pm 0.0024}$	$0.2124_{\pm 0.0019}$
	VISTA	$\mathbf{0.8506}_{\pm 0.0021}$	$\mathbf{0.1549}_{\pm 0.0015}$	$\mathbf{0.1755}_{\pm 0.0025}$	$\mathbf{1.34}_{\pm 0.04}$	$\mathbf{0.4843}_{\pm 0.0018}$	$\mathbf{0.1975}_{\pm 0.0015}$

*Optimal value is near average; values near zero or above average indicate poor predictions Chelebian et al., 2025 (see Section 4).

Experiment setup. We evaluate under increasing generalization difficulty: *intra-tissue* holds out sites within a region (20%, 40% and 80%); *cross-tissue* trains on 'tumor0' and tests on 'tumor1' within RN49; *cross-sample* trains on HL55 and tests on RN49. Together they ask whether the model moves beyond within-region interpolation to relationships that transfer.

HT carries no dense molecular ground truth, so we score 3D structure (Moran's I, Geary's C, morphology coupling, boundary enrichment) alongside agreement with the measured plane; definitions are in Appendix H. We compare against kNN-IDW and ASIGN; VORTEX and X-pression cannot be applied faithfully to continuous volumes, for the reasons given with the baselines in Appendix G.

Benchmark results. Table 7 shows that VISTA produces the strongest 3D structure quality across intra-tissue, cross-tissue, and cross-sample settings, with higher Moran's I, lower Geary's C, stronger morphology coupling, and higher BES than ASIGN and kNN-IDW. The gap is largest in morphology coupling, where VISTA exceeds the next-best baseline by 3–8×, so its predicted gradients follow tissue morphology rather than being spatially smooth artifacts. Its BES indicates expression changes 50–60% more strongly at tissue boundaries than in the interior, whereas ASIGN remains near 1.0.

The 2D similarity metrics illustrate here the limitation that Section 4.2 demonstrated directly against ground truth. kNN-IDW obtains competitive correlation or MSE after projection to Visium spots despite producing depth-invariant volumes, because per-site Pearson correlation rewards predicting the correct gene-wise *mean* more than it rewards spatial specificity; on Open-ST the same baseline reached a site correlation of 0.646 while recovering only 5.8% of the attainable depth signal. VISTA instead holds both, and retains coherent 3D structure and competitive 2D agreement under the cross-tissue and cross-sample shifts, indicating that its volumetric representations transfer across samples.

The volumetric input carries information its own projection does not. Per-depth accuracy is not computable on HT, but the contribution of depth is. We run the same model and losses on a 2D maximum-intensity projection of the volume (VISTA-2D), removing only the 3D convolution and z-attention, against eight baselines on identical splits. With architecture, objective, and splits held fixed, any difference between VISTA-2D and VISTA-3D is attributable to the volumetric input rather than to the inductive biases, and is measured against real held-out ST rather than a structural proxy. VISTA-2D already outperforms the eight baselines, so the architecture and regularizers are themselves competitive. The 3D version improves further, and most where the problem is hardest: +13% cross-tissue and +27% cross-sample over the strongest 2D competitor. Depth therefore carries signal on HT as well – the same claim verified on Open-ST, established here by the one route the data permit.

Biological plausibility. To assess whether the predicted 3D molecular fields are biologically meaningful, we examine both morphology coupling and depth-resolved marker-gene patterns, discussed further in Appendix C. Figure 3 shows the per-gene distribution of morphology coupling across all 370 genes. VISTA exhibits consistently stronger morphology–expression correspondence across intra-tissue, cross-tissue, and cross-sample settings, indicating that the improvement is not driven by a small number of outlier genes. Figure 4 visualizes predicted 3D expression for three gastric-cancer-relevant genes: *AQP8* (poor overall survival (Thapa et al., 2018)), *RTKN2* (tumor invasion (Zhao et al., 2022)), and *CCL27* (immune cell migration (Terefe et al., 2022)). All three

Table 8: 2D control on HT (correlation ↑). VISTA-2D carries identical inductive biases and regularizers to VISTA-3D but sees only a maximum-intensity projection. The 40% and 80% held-out fractions are in Appendix D.3.

Method	Intra-tissue	Cross-tissue	Cross-sample
kNN-IDW	0.545	0.219	0.185
BLEEP	0.548	0.384	0.367
Hist2ST	0.501	0.359	0.353
STGAT	0.329	0.209	0.378
THItoGene	0.524	0.365	0.338
TRIPLEX	0.515	0.369	0.333
ASIGN	0.390	0.295	0.293
X-pression	0.523	0.331	0.378
VISTA-2D	0.549	0.394	0.382
VISTA-3D	**0.565**	**0.446**	**0.484**

Table 9: Ablation study on continuous 3D HT volume (intra-tissue, 20% test). Superscripts show % change from full model.

		3D Structure			2D ST Sim.	
	Variant	Moran ↑	Geary ↓	Morph*	Corr ↑	MSE ↓
Enc.	w/o UNI2	$0.8065^{-1.0}$	$0.1774^{+4.5}$	$0.2493^{-21.0}$	$0.5749^{+1.8}$	$0.1461^{+12.0}$
	w/o fine 2D	$0.8752^{+7.5}$	$0.1531^{+17.7}$	$0.3076^{-2.2}$	$0.5527^{-2.1}$	$0.1728^{-4.2}$
	w/o coarse 2D	$0.7675^{-5.7}$	$0.2338^{-25.8}$	$0.3425^{+8.9}$	$0.5531^{-2.1}$	$0.1723^{-3.6}$
	w/o all 2D	$0.7526^{-7.5}$	$0.2277^{-22.6}$	$0.3684^{+16.8}$	$0.5451^{-3.5}$	$0.1816^{-9.6}$
	w/o 3D conv	$0.8363^{+2.7}$	$0.1552^{+16.7}$	$0.2211^{-29.8}$	$0.5454^{-3.5}$	$0.1829^{-10.2}$
Context	w/o axial	$0.7853^{-3.6}$	$0.2146^{-15.6}$	$0.3832^{+21.5}$	0.5650^{0}	$0.1629^{+1.8}$
	w/o z-attn	$0.7666^{-5.8}$	$0.2321^{-24.7}$	$0.3354^{+6.3}$	$0.5621^{-0.5}$	0.1658^{0}
	w/o local GAT	$0.7423^{-8.8}$	$0.2572^{-38.2}$	0.3151^{0}	$0.5341^{-5.5}$	$0.1746^{-5.4}$
	w/o global GAT	$0.7983^{-2.0}$	$0.2027^{-9.1}$	$0.3136^{-0.3}$	$0.5546^{-1.8}$	$0.1743^{-4.8}$
	w/o all graph	$0.7304^{-10.3}$	$0.2668^{-43.5}$	$0.3025^{-3.8}$	$0.5287^{-6.4}$	$0.1750^{-5.4}$
Pool	w/o attn pool	$0.7541^{-7.4}$	$0.2471^{-32.8}$	$0.3619^{+14.9}$	$0.5641^{-0.2}$	$0.1651^{+0.6}$
Loss	$\mathcal{L}_{\text{data}}$ only	$0.6604^{-18.9}$	$0.3487^{-87.6}$	$0.3046^{-3.2}$	$0.5121^{-9.4}$	$0.1998^{-20.5}$
	Full	**0.8143**	**0.1858**	**0.3154**	**0.5648**	**0.1660**

*Optimal value is near average; values near zero or above average indicate poor predictions Chelebian et al., 2025 (see Section 4).

show depth-varying expression with significant morphology coupling (0.33, 0.19, 0.38; $p \leq 0.005$) and boundary enrichment (1.54, 1.29, 1.72; $p \leq 0.005$; spatial permutation, $n=200$), indicating that predicted gradients align with tissue architecture rather than spatially arbitrary variation.

4.4 Ablation Study

Table 9 isolates the contribution of each component on the continuous 3D setting (intra-tissue, 20% test). We highlight three key findings: i) **Graph attention is the most critical component.** Removing attention causes the largest degradation across both metric groups: Moran's I falls by 10.3%, Geary's C worsens by 43.5%, and 2D correlation drops by 6.4%. Local graph attention contributes more than global, consistent with the intuition that nearest-neighbor spatial context in the tissue volume is essential for coherent 3D prediction. ii) **3D convolution anchors morphology coupling.** Removing it drops morphology coupling into the low range (≈ 0.22), indicating that the predicted expression gradients no longer follow tissue structure. The shallow volumetric stem therefore supplies the morphological grounding that keeps predicted gradients aligned with tissue architecture, rather than merely smoothing the volume. iii) **No single component is sufficient.** Training with data loss alone yields the weakest result on every metric. Each additional architectural component and the depth regularization term improve different aspects of the reconstruction, confirming that the full model's balance of morphology-aware encoding, graph-based spatial context, and depth regularization is necessary to recover coherent 3D gene-expression fields from limited 2D supervision.

5 Discussion and Conclusion

VISTA learns depth-resolved gene-expression fields from tissue morphology under limited ST supervision, combining multi-scale morphology encoding, depth-aware attention, 3D graph learning, and a depth-directed objective into fields that stay anchored to measured ST. The three settings differ in how directly that claim can be checked, and it is worth being explicit about what each measures. On Open-ST, depth accuracy is measured against held-out RNA at 19 depths and read against a split-half noise ceiling: VISTA reaches 53% of what is recoverable, against 5.8% for interpolation between the same two measured faces. On HER2+, accuracy is measured against held-out sections under supervision as thin as a single measured plane. On HT only one plane is sequenced, so neither VISTA nor any baseline can report per-depth accuracy; what is measurable is whether the depth dimension contributes, and it does, by +13% to +27% over the model's own 2D projection. Column-aggregated supervision is therefore a limit on what the available metrics resolve on HT, not a weaker statement about the field: many 3D fields produce the same aggregated measurement, so agreement with the measured plane cannot by itself distinguish among them. That is why we report volumetric criteria alongside it, and why a dataset like Open-ST matters, since it is where the mechanism can be verified rather than assumed; Appendix I takes up the evaluation and registration issues this raises. Across registered HER2+ serial sections, the Open-ST serial volume, and continuous HT volumes, VISTA improves structure quality, morphology alignment, and gene-expression prediction over existing approaches, a step toward bridging planar molecular measurements with volumetric tissue biology.

AI USE STATEMENT

In this work, we used generative AI tools to assist with software implementation, debugging, and execution of computational experiments. AI-assisted code and experimental procedures were reviewed, verified, and tested by the authors before being used to produce reported results.

Generative AI tools were not responsible for the formulation of the research questions, selection of the scientific methodology, interpretation of experimental results, or derivation of the conclusions presented in this work. These research decisions and analyses were performed by the authors. All reported results were independently examined and validated by the authors, and conclusions were drawn from those results by the authors.

We have reviewed all AI-assisted work and take full responsibility for the scientific content, methodology, results, claims, and conclusions of this work.

REPRODUCIBILITY STATEMENT

A zipped folder containing all the scripts for the VISTA implementation is uploaded with this submission on OpenReview as supplementary material. Every component of the framework and both terms of the objective are specified in Appendix A, with the encoder architecture in Appendix A.2 and hyperparameters, optimizer settings, and hardware in Appendix F. The evaluation protocols are given in Appendix B for Open-ST and Appendix D for the serial-section settings; preprocessing and registration are described in Appendix E, and every metric, including the split-half noise ceiling against which depth correlation is reported, is defined in Appendix H. Appendix I records the evaluation pitfalls that motivate reporting volumetric criteria alongside in-plane agreement.

REFERENCES

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, Long Phi Le, Alexander S. Baras, Sizun Jiang, Ali Bashashati, Jonathan T. C. Liu, and Faisal Mahmood. AI-driven 3D Spatial Transcriptomics, 2025.

Alma Andersson, Ludvig Larsson, Linnea Stenbeck, Fredrik Salmén, Anna Ehinger, Simon Z. Wu, Ghamdan Al-Eryani, Daniel Roden, Alexander Swarbrick, Ake Borg, and Joakim Lundeberg. 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.

Sudipto Baul, Khandakar Tanvir Ahmed, Qibing Jiang, Guangyu Wang, Qian Li, Jeongsik Yong, and Wei Zhang. Integrating spatial transcriptomics and bulk RNA-seq: predicting gene expression with enhanced resolution through graph attention networks. *Briefings in Bioinformatics*, 25(4): bbae316, 2024.

John A Bogovic, Philipp Hanslovsky, Allan Wong, and Stephan Saalfeld. Robust registration of calcium images by learned contrast synthesis. In *2016 IEEE 13th international symposium on biomedical imaging (ISBI)*, pp. 1123–1126. IEEE, 2016.

Eduard Chelebian, Christophe Avenel, and Carolina Wählby. Combining spatial transcriptomics with tissue morphology. *Nature Communications*, 16(1):4452, 2025.

Kok Hao Chen, Alistair N. Boettiger, Jeffrey R. Moffitt, Siyuan Wang, and Xiaowei Zhuang. Spatially resolved, highly multiplexed RNA profiling in single cells. *Science*, 348(6233):aaa6090, 2015. doi: 10.1126/science.aaa6090.

Richard J Chen, Tong Ding, Ming Y Lu, Drew FK Williamson, Guillaume Jaume, Andrew H Song, Bowen Chen, Andrew Zhang, Daniel Shao, Muhammad Shaban, et al. Towards a general-purpose foundation model for computational pathology. *Nature medicine*, 30(3):850–862, 2024.

Youngmin Chung, Ji Hun Ha, Kyeong Chan Im, and Joo Sang Lee. Accurate Spatial Gene Expression Prediction by Integrating Multi-Resolution Features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11591–11601, 2024. doi: 10.48550/arXiv.2403.07592.

Kalen Clifton, Manjari Anant, Gohta Aihara, Lyla Atta, Osagie K Aimiuwu, Justus M Kebschull, Michael I Miller, Daniel Tward, and Jean Fan. STalign: Alignment of spatial transcriptomics data using diffeomorphic metric mapping. *Nature communications*, 14(1):8123, 2023.

Lance A Davidson. Epithelial machines that shape the embryo. *Trends in Cell Biology*, 22(2):82–87, 2012.

Yui Deguchi, Chie Kikutake, and Mikita Suyama. Subtype-specific alternative splicing events in breast cancer identified by large-scale data analysis. *Scientific reports*, 14(1):14158, 2024.

Chee-Huat Linus Eng, Michael Lawson, Qian Zhu, Ruben Dries, Na Koulena, Yodai Takei, Jun Yun, Christopher Cronin, Chris Karp, Guo-Cheng Yuan, and Long Cai. Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH+. *Nature*, 568:235–239, 2019. doi: 10.1038/s41586-019-1049-y.

Pedro Gómez-Gálvez, Pablo Vicente-Munuera, Samira Anbari, Javier Buceta, and Luis M Escudero. The complex three-dimensional organization of epithelial tissues. *Development*, 148(1): dev195669, 2021.

Tingyu Gu, Kun Wang, Xiao Yuan, Haowen Tang, Shouchao Wang, and Zhihong Zhao. The dual role of GPX4 in breast cancer: mechanisms of therapeutic resistance and potential for novel targeted therapies. *Cancer Gene Therapy*, 32(9):913–922, 2025.

J. D. Hunter. Matplotlib: A 2D graphics environment. *Computing in Science & Engineering*, 9(3): 90–95, 2007. doi: 10.1109/MCSE.2007.55.

Yuran Jia, Junliang Liu, Li Chen, Tianyi Zhao, and Yadong Wang. THItoGene: a deep learning method for predicting spatial transcriptomics from histological images. *Briefings in Bioinformatics*, 25(1):bbad464, 2024.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

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.

Johannes Schindelin, Ignacio Arganda-Carreras, Erwin Frise, Verena Kaynig, Mark Longair, Tobias Pietzsch, Stephan Preibisch, Curtis Rueden, Stephan Saalfeld, Benjamin Schmid, et al. Fiji: an open-source platform for biological-image analysis. *Nature methods*, 9(7):676–682, 2012.

Marie Schott, Daniel León-Periñán, Elena Splendiani, Leon Strenger, Jan Robin Licha, Tancredi Massimo Pentimalli, Simon Schallenberg, Jonathan Alles, Sarah Samut Tagliaferro, Anastasiya Boltengagen, et al. Open-ST: High-resolution spatial transcriptomics in 3D. *Cell*, 187(15): 3953–3972, 2024a.

Marie Schott, Daniel León-Periñán, Elena Splendiani, Leon Strenger, Jan Robin Licha, Tancredi Massimo Pentimalli, Simon Schallenberg, Jonathan Alles, Sarah Samut Tagliaferro, Anastasiya Boltengagen, et al. Open-ST: High-resolution spatial transcriptomics in 3D. *Cell*, 187(15): 3953–3972, 2024b.

Nicholas Sofroniew, Talley Lambert, Grzegorz Bokota, Juan Nunez-Iglesias, Peter Sobolewski, Andrew Sweet, Lorenzo Gaifas, Kira Evans, Alister Burt, Draga Doncila Pop, et al. napari: a multi-dimensional image viewer for Python. *Zenodo*, 2022.

Patrik L. Stahl, Fredrik Salmén, Sanja Vickovic, Anna Lundmark, José Fernández Navarro, Jens Magnusson, Stefania Giacomello, Michaela Asp, Jakub O. Westholm, Mikael Huss, Annelie Mollbrink, Sten Linnarsson, Simone Codeluppi, Ake Borg, Fredrik Pontén, Paul Igor Costea, Pelin Sahlén, Jan Mulder, Olaf Bergmann, Joakim Lundeberg, and Jonas Frisén. Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science*, 353(6294): 78–82, 2016. doi: 10.1126/science.aaf2403.

Ermias Mergia Terefe, Maria Jade Catalan Opulencia, Amir Rakhshani, Mohammad Javed Ansari, Sergushina Elena Sergeevna, Sura A Awadh, Djamila Sh Polatova, Adnan Hashim Abdulkadhim, Yasser Fakri Mustafa, Hamzah H Kzar, et al. Roles of CCR10/CCL27–CCL28 axis in tumour development: mechanisms, diagnostic and therapeutic approaches, and perspectives. *Expert Reviews in Molecular Medicine*, 24:e37, 2022.

Saroj Thapa, Mandika Chetry, Kaiyu Huang, Yangpei Peng, Jinsheng Wang, Jiaoni Wang, Yingying Zhou, Yigen Shen, Yangjing Xue, and Kangting Ji. Significance of aquaporins' expression in the prognosis of gastric cancer. *Bioscience Reports*, 38(3):BSR20171687, 2018.

Demeter Túrós, Lollija Gladiseva, Marius Botos, Chang He, G. Tuba Barut, Inês Berenguer Veiga, Nadine Ebert, Anne Bonnin, Astrid Chanfon, Llorenç Grau-Roma, Alberto Valdeolivas, Sven Rottenberg, Volker Thiel, and Etori Aguiar Moreira. Deep Learning-Based 3D Spatial Transcriptomics with X-Pression, 2025.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph Attention Networks. *International Conference on Learning Representations*, 2018.

Ronald Xie, Kuan Pang, Sai Chung, Catia Perciani, Sonya MacParland, Bo Wang, and Gary Bader. Spatially resolved gene expression prediction from histology images via bi-modal contrastive learning. *Advances in Neural Information Processing Systems*, 36:70626–70637, 2023.

Yuansong Zeng, Zhuoyi Wei, Weijiang Yu, Rui Yin, Yuchen Yuan, Bingling Li, Zhonghui Tang, Yutong Lu, and Yuedong Yang. Spatial transcriptomics prediction from histology jointly through transformer and graph neural networks. *Briefings in Bioinformatics*, 23(5):bbac297, 2022.

HG Zhao, JJ Yin, X Chen, J Wu, W Wang, and LW Tang. RTKN2 enhances radioresistance in gastric cancer through regulating the Wnt/β-Catenin signalling pathway. *Folia Biol*, 68(1):33–39, 2022.

Junchao Zhu, Ruining Deng, Tianyuan Yao, Juming Xiong, Chongyu Qu, Junlin Guo, Siqi Lu, Mengmeng Yin, Yu Wang, Shilin Zhao, Haichun Yang, and Yuankai Huo. ASIGN: An Anatomy-aware Spatial Imputation Graph Network for 3D Spatial Transcriptomics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. doi: 10.48550/arXiv.2412.03026.

A Detailed objective functions and mathematical definitions

This appendix defines the mathematical objectives used by VISTA. The notation follows Section 3. The index s denotes a Visium spot, t denotes a tissue-volume tile assigned to spot s, z denotes a depth slice, $u \in \{1, \ldots, U\}$ denotes a token-grid position, and g denotes a gene. The number of selected genes is G, and $\mathcal{T}_s$ is the set of valid tiles assigned to spot s. The slice-level prediction is $\widehat{y}_{t,z} \in \mathbb{R}^G$, the tile-level prediction is $\widehat{y}_t \in \mathbb{R}^G$, and the aggregated spot-level prediction is $\widehat{y}_s \in \mathbb{R}^G$. The measured Visium expression for spot s is $y_s \in \mathbb{R}^G$.

A.1 Preprocessing details

The Visium count matrix is restricted to a curated gene panel. Preprocessing applies gene filtering to remove low-variance or unexpressed genes, log-transformation via $\log(1 + x)$, and per-gene z-scoring using the training-set mean and standard deviation. The same fitted transform (training-set statistics) is applied identically to validation and test samples to prevent information leakage. Tile volumes are independently normalized to zero mean and unit variance per tile before being fed to the encoder.

A.2 Encoder architecture

Frozen foundation encoder. Each slice is passed through UNI2-h (Chen et al., 2024), a ViT-giant pathology foundation model with 681M parameters, which remains frozen and is never updated. Its patch tokens are pooled with adaptive average pooling to the working token grid, which matters because the transformer's token grid and the convolutional grid need not agree: for a 128×128 tile the convolutional path yields a 16×16 grid against the encoder's 8×8, and pooling to $\sqrt{N_{\text{tok}}}$ reconciles them. Section 4.2 reports the effect of this choice, which is the single largest determinant of depth accuracy we measured.

2D convolutional stem. In parallel, each raw slice $I_{t,z}$ is processed by a lightweight 2D CNN stem consisting of three convolutional layers with group normalization and GELU activations. The stem preserves spatial resolution (stride 1) and outputs a feature map $f_{t,z} \in \mathbb{R}^{C_{\text{stem}} \times H \times W}$. The frozen encoder tokens and the stem feature map are concatenated and shared by both 2D token branches.

Fine and coarse tokenization. The fine branch applies a non-overlapping $P \times P$ patch projection to $f_{t,z}$, producing a token grid of size $H_p \times W_p$ with $H_p = H/P$, $W_p = W/P$. The coarse branch first spatially downsamples $f_{t,z}$ by a factor of 2 via bilinear interpolation, applies the same patch projection (yielding a smaller token grid), and then bilinearly upsamples the token grid back to $H_p \times W_p$. This gives the coarse tokens a larger effective receptive field than fine tokens at the same grid resolution.

Shallow 3D stem. The full tile volume $X_t \in \mathbb{R}^{1 \times Z \times H \times W}$ is processed by a shallow 3D CNN (two layers, kernel size $3 \times P \times P$ with stride $1 \times P \times P$ along (z, h, w)). This preserves the number of depth slices while downsampling XY to match the 2D token grid, producing volumetric tokens $v^{\text{3D}}_{t,z,u} \in \mathbb{R}^D$.

Three-way gated fusion. The gate coefficients $\pi^{(q)}_{t,z,u}$ for $q \in \{\text{fine}, \text{coarse}, \text{3D}\}$ are computed by a two-layer MLP applied to the concatenation $[v^{\text{fine}}_{t,z,u};\ v^{\text{coarse}}_{t,z,u};\ v^{\text{3D}}_{t,z,u}] \in \mathbb{R}^{3D}$, followed by a softmax over the three branches. The fused token is $r_{t,z,u} = \sum_q \pi^{(q)}_{t,z,u}\, v^{(q)}_{t,z,u}$, after which layer normalization and optional dropout are applied.

A.3 Axial attention and z-attention

XY axial attention. For each slice, tokens $\{r_{t,z,u}\}_{u=1}^{U}$ are arranged in a 2D grid of size $H_p \times W_p$. Self-attention is applied first along the W_p axis (each row independently) and then along the H_p axis (each column independently), each followed by a residual connection. A feed-forward block with GELU activation, layer normalization, and residual connection completes the axial-attention layer. This factorization reduces the per-slice attention cost from $O(U^2)$ to $O(U\sqrt{U})$.

Slice pooling. After axial attention, tokens within each slice are averaged to produce the slice embedding $h_{t,z} \in \mathbb{R}^D$.

Z-attention. The sequence $\{h_{t,z}\}_{z=1}^{Z}$ within each tile is processed by a standard multi-head self-attention block (with key padding mask to handle padded tiles) followed by a feed-forward residual block. The output $\widetilde{h}_{t,z}$ captures within-tile depth dependencies. Two z-attention layers are used.

A.4 Gated context fusion

The gate $g_{t,z} \in (0,1)^D$ is produced by applying a sigmoid activation to a linear projection of the concatenated inputs $[\widetilde{h}_{t,z};\, h^{\mathrm{loc}}_{t,z};\, h^{\mathrm{glob}}_{t,z}] \in \mathbb{R}^{3D}$. The linear mixing function ϕ is a separate learned projection from $\mathbb{R}^{3D}$ to $\mathbb{R}^D$. The fused representation is:

$$h^{\mathrm{fuse}}_{t,z} = g_{t,z} \odot \widetilde{h}_{t,z} + (1 - g_{t,z}) \odot \phi\big([\widetilde{h}_{t,z};\, h^{\mathrm{loc}}_{t,z};\, h^{\mathrm{glob}}_{t,z}]\big). \tag{3}$$

When $g_{t,z} \approx 1$, the node retains its pre-graph embedding; when $g_{t,z} \approx 0$, it adopts the graph-propagated mixture. This prevents over-smoothing while allowing informative graph context to update the representation.

A.5 Depth representation across settings

The encoder, axial attention, within-tile depth attention, and graph propagation are used unchanged for stacked serial sections. The single adjustment concerns how depth is represented. In the continuous setting the depth index is a slice number and neighboring slices are physically adjacent. In the stacked setting the sections are separated by large and irregular physical gaps, so each node carries its measured depth z in micrometres and the depth graph is built over those physical coordinates. Cross-section interaction is therefore handled by geometry-aware graph propagation over aligned 3D coordinates, which shares information across sections without assuming that consecutive measured planes are adjacent.

A.6 Attention pooling

For each tile t, slice-level predictions are pooled into a tile-level prediction using nonnegative attention weights $\alpha_{t,z}$:

$$\widehat{y}_{t,g} = \sum_{z=1}^{Z} \alpha_{t,z} \widehat{y}_{t,z,g}, \qquad \alpha_{t,z} = \frac{\exp(\ell^z_{t,z}/\tau_z)}{\sum_{z'=1}^{Z} \exp(\ell^z_{t,z'}/\tau_z)}. \tag{4}$$

Here, $\widehat{y}_{t,z,g}$ is the predicted expression of gene g for tile t and slice z, $\ell^z_{t,z}$ is a learned slice-pooling logit, and $\tau_z > 0$ is an optional temperature controlling the sharpness of slice pooling. The weights satisfy $\sum_z \alpha_{t,z} = 1$.

Tile-level predictions are pooled into spot-level predictions using valid tiles only:

$$\widehat{y}_{s,g} = \sum_{t \in \mathcal{T}_s} \beta_{s,t} \widehat{y}_{t,g}, \qquad \beta_{s,t} = \frac{\exp(\ell^t_{s,t}/\tau_t)}{\sum_{t' \in \mathcal{T}_s} \exp(\ell^t_{s,t'}/\tau_t)}. \tag{5}$$

Here, $\ell^t_{s,t}$ is a learned tile-pooling logit and $\tau_t > 0$ is an optional tile-pooling temperature. Padded tiles are excluded by the tile mask.

A.7 Smooth L1 data fidelity loss

The supervised data-fidelity term compares the aggregated prediction $\widehat{y}_s$ to the measured Visium vector y_s. For residual $r_{s,g} = \widehat{y}_{s,g} - y_{s,g}$ and Smooth L1 transition parameter $\delta_{\mathrm{SL1}} > 0$, define

$$\rho^{\mathrm{SL1}}_{\delta_{\mathrm{SL1}}}(r) = \begin{cases} \frac{1}{2} r^2 / \delta_{\mathrm{SL1}}, & |r| < \delta_{\mathrm{SL1}}, \\ |r| - \frac{1}{2}\delta_{\mathrm{SL1}}, & |r| \geq \delta_{\mathrm{SL1}}. \end{cases} \tag{6}$$

The data loss is

$$\mathcal{L}_{\text{data}} = \frac{1}{|\mathcal{B}|G} \sum_{s \in \mathcal{B}} \sum_{g=1}^{G} \rho^{\text{SL1}}_{\delta_{\text{SL1}}}(\widehat{y}_{s,g} - y_{s,g}), \tag{7}$$

where $\mathcal{B}$ is the set of spots in the mini-batch. The Smooth L1 penalty is used because it behaves quadratically near zero residuals and linearly for larger residuals, making spot-level supervision less sensitive to outliers than mean squared error while remaining stable for noisy spatial transcriptomics measurements. The parameter δ_{SL1} determines the transition between the quadratic and linear regimes.

A.8 Schrödinger-bridge matching along depth

Let $z_a < z_b$ be two revealed planes at a location, with measured expression x_a and x_b. For a withheld plane at depth z, write $\tau = (z - z_a)/(z_b - z_a) \in (0, 1)$. The reference Brownian bridge between the endpoints has conditional drift

$$b^\star(x, \tau) = \frac{x_b - x}{1 - \tau + \varepsilon}, \tag{8}$$

where $\varepsilon > 0$ prevents the target from diverging as $\tau \to 1$; we use $\varepsilon = 0.1$ and exclude the far endpoint from the supervised depths. Writing b_θ for the increment predicted by the gene head between consecutive depths, the matching loss is

$$\mathcal{L}_{\text{SB}} = \frac{1}{|\mathcal{P}|} \sum_{(i,z) \in \mathcal{P}} \| b_\theta(\widehat{y}_{i,z}, \tau_z) - b^\star(x_{i,z}, \tau_z) \|_2^2, \tag{9}$$

over the set $\mathcal{P}$ of location–depth pairs lying strictly between two revealed planes. This corresponds to a single iteration of iterative Markovian fitting: the bridge is matched once against the reference process rather than alternated to convergence.

A.9 Graph attention

Let $\mathcal{G} = (\mathcal{U}, \mathcal{E})$ be a graph over tile–slice nodes, where $\mathcal{U}$ is the node set. For node i, let $\mathcal{N}(i)$ denote its incoming neighbors and let $h_i \in \mathbb{R}^D$ be its input feature. A graph attention head computes

$$e_{ij} = \text{LeakyReLU}\left(\mathbf{a}_{\text{gat}}^\top [W h_i; W h_j]\right), \qquad j \in \mathcal{N}(i), \tag{10}$$

where W is a learned linear projection and $\mathbf{a}_{\text{gat}}$ is a learned attention vector. The normalized edge coefficient is

$$\eta_{ij} = \frac{\exp(e_{ij})}{\sum_{\ell \in \mathcal{N}(i)} \exp(e_{i\ell})}. \tag{11}$$

The output of the head is

$$h_i' = \sum_{j \in \mathcal{N}(i)} \eta_{ij} W h_j. \tag{12}$$

Multiple heads are averaged or concatenated, followed by residual and feed-forward layers. VISTA uses one graph for local 3D neighborhoods and one graph for global spot-level context.

B Open-ST protocol details

Data and protocol. Open-ST (Schott et al., 2024b) sequences every section of a serial volume. We aggregate the released barcodes onto a shared $100\,\mu$m bin grid, retain the 19 sections passing quality control, and normalize expression as $\log(1 + 10^4 c/\ell)$ for count c and library size ℓ. The two-anchor protocol reveals the first and last sections and withholds the 17 sections between them. Section depths span 20–360 μm and are irregularly spaced, so node coordinates carry the measured depth in micrometres.

Baselines. kNN-IDW fits every site measured on the two revealed sections as a reference point at its physical (x, y, z) and interpolates the withheld depths, using the same k and inverse-distance power as in the other experiments. BLEEP, TRIPLEX, X-pression, and ASIGN are run from their released implementations with their published default hyperparameters, extended only so that two sections can be revealed rather than one. All are trained on the spots of the revealed sections and predict the withheld planes, and all are scored with the metric and support described below, identical to VISTA. Because the Open-ST tiles are 64×64 pixels, every model receives images at that resolution.

Trajectory support. Depth correlation is computed on site trajectories across the withheld sections. Requiring a site to be measured in *every* withheld section leaves only 133 sites, because a few sections are sparsely covered; this biases the metric toward the densest tissue core. We therefore require coverage of at least 10 of the 17 withheld sections, giving 1,088 trajectories. The ordering of methods is unchanged between the two supports.

Noise ceiling. Cells within each bin are split into halves, each half normalized independently, and the halves scored against each other with the same depth-correlation function on the same trajectories. The Spearman–Brown correction converts half-sample reliability r to full-sample reliability $\rho = 2r/(1 + r)$, and the ceiling on correlation against a noisy target is $\sqrt{\rho}$. This yields a ceiling of 0.795 on the ragged support (reliability 0.588).

C QUALITATIVE RESULTS

In the continuous 3D setting no dense molecular ground truth exists, so the predicted volume cannot be scored depth by depth against a measurement. What can be asked is whether the field is organised the way tissue is organised. The two figures in this appendix ask that question at two different granularities: Figure 3 gene by gene over the whole panel, and Figure 4 location by location for three marker genes.

Per-gene morphology coupling (Figure 3). Morphology gradient correlation (defined in Appendix H) correlates the gradient field of the predicted 3D expression volume with the gradient field of the volumetric morphology, so a high value means predicted expression changes where tissue structure changes. The figure summarises the distribution of that score computed *per gene* over all 370 panel genes, for each method and separately for the intra-tissue, cross-tissue, and cross-sample settings. Two features of the figure carry the argument. First, VISTA's distribution is shifted above the baselines across its whole range rather than only in its upper tail, so the aggregate advantage in Table 7 is not produced by a small number of well-fitted genes. Second, the ordering between methods is preserved as the evaluation moves from intra-tissue to cross-sample, where the test tissue is never seen during training, which is what one expects if the morphology–expression relationship transfers rather than being fitted per sample. The interpolation baselines concentrate at much lower coupling, consistent with the 3–8$\times$ gap reported in Section 4.3: a depth-constant or smoothly interpolated volume has little gradient structure available to align with morphology.

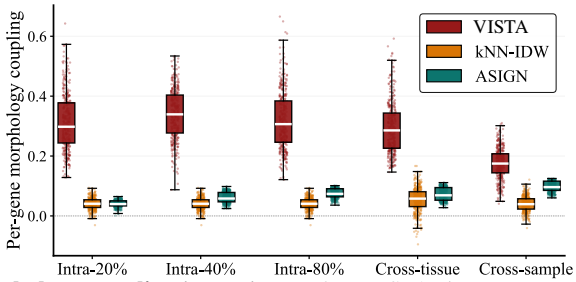


Figure 3: **Per-gene morphology coupling** in continuous 3D. VISTA shows stronger morphology-aligned expression gradients across genes and settings.

Depth-resolved marker expression (Figure 4). The figure shows VISTA's predicted expression for *AQP8*, *RTKN2*, and *CCL27* on the `tumor0` region of gastric cancer sample RN49, rendered as a 3D volume together with in-plane heatmaps at several depth indices Z. These three genes were chosen because each has an established role in gastric cancer – *AQP8* with poor overall survival

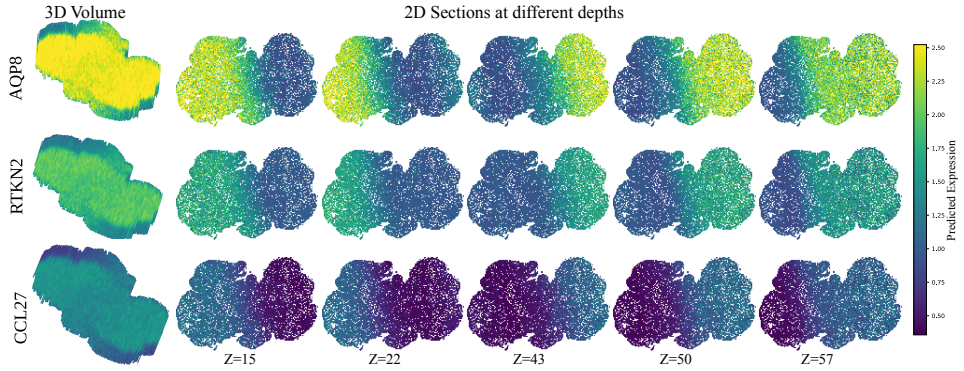


Figure 4: **3D gene expression heatmap** of three biomarker genes- AQP8, RTKN2, and CCL27 predicted by VISTA on 'tumor0' tissue region of RN49 gastric cancer sample. It also displays the heatmap at different depths of the tissue volume denoted by Z indices.

(Thapa et al., 2018), *RTKN2* with tumor invasion (Zhao et al., 2022), and *CCL27* with immune cell migration (Terefe et al., 2022) – so their predicted arrangement can be read against tissue architecture and not only against a held-out number. The panels across Z are the informative part: the maps vary with depth instead of reproducing the single sequenced plane at every slice, which is the characteristic failure of projection-based interpolation, and the variation is morphology-aligned rather than diffuse. One limitation should be stated plainly. Because only one plane of this volume was sequenced, these maps are not validated depth by depth here; the depth-resolved accuracy of the same framework is measured against withheld planes on Open-ST in Section 4.2.

D Additional Experiments

D.1 Single-section supervision on stacked serial ST

The stacked protocol of Section 4 reveals several measured sections. To test the limited-supervision claim in its strictest form, we also run a one-section-in, five-sections-out protocol on the HER2+ cohort (samples A–D, six registered sections each at physical depths $0/32/64/96/128/160\,\mu$m). A single section's expression is revealed during training and all remaining sections are predicted, rotating the revealed section through every position, giving six independent folds per sample. Held-out expression is never used for optimization, gene-panel selection, or early stopping, and is opened only for final evaluation; all baselines follow the same label firewall.

Sections are not treated as adjacent voxels. XY is converted to microns, z is the measured physical position, neighbours are formed in physical XYZ, and cross-section message passing is weighted by $\exp(-\Delta z/\ell)$ with a learnable length scale ℓ, so the irregular physical gaps between sections enter propagation directly rather than through a slice index.

Table 10: Single-section supervision on HER2+ serial sections: site correlation, mean of six reveal-one folds. **Best** in bold.

Model	A	B	C	D	Mean ↑	MSE ↓
VISTA	**0.799**	**0.404**	**0.723**	**0.598**	**0.631**	**0.347**
BLEEP	0.776	0.214	0.687	0.503	0.545	0.432
TRIPLEX	0.705	0.334	0.613	0.512	0.541	0.427
kNN-IDW	0.725	0.332	0.641	0.463	0.540	0.454
X-pression	0.779	0.287	0.483	0.526	0.519	0.474
THItoGene	0.684	0.301	0.591	0.476	0.513	0.447
Hist2ST	0.625	0.308	0.519	0.441	0.473	0.488
ASIGN	0.498	0.337	0.191	0.347	0.343	0.696
STGAT	0.419	0.200	0.344	0.292	0.314	0.589

VISTA has the highest mean correlation on every patient and improves the benchmark mean from 0.545 for the best baseline to 0.631. Table 11 reports fold-level variability on test sample D, giving the range of correlations over the five hidden target sections for each choice of revealed section.

Table 11: Fold-level variability on HER2+ sample D: range of site correlation over the five hidden sections, for each revealed section.

Revealed section	kNN-IDW	ASIGN	VISTA
D1	0.435–0.467	0.376–0.403	**0.503–0.591**
D2	0.448–0.467	0.282–0.324	**0.592–0.639**
D3	0.464–0.475	0.373–0.410	**0.603–0.639**
D4	0.440–0.484	0.267–0.331	**0.587–0.655**
D5	0.448–0.485	0.400–0.433	**0.596–0.667**
D6	0.436–0.488	0.299–0.317	**0.486–0.629**

D.2 Unseen test patient

The protocols in Section 4.1 hold out sections within a cohort that the model has partly seen: samples A and B are used for training, so the test sample shares a tissue type, a staining batch and a section-thickness regime with the training data. The protocol reported here removes that shared context. The test patient is unseen, and each training patient contributes a single measured section, so the model must transfer a morphology–expression relationship across patients rather than interpolate within one.

This is the strictest protocol in the paper, and every method loses ground relative to the within-patient setting, which is the expected direction: patient-to-patient variation in tissue composition and staining is larger than section-to-section variation within a block. What distinguishes the methods is how much they lose. VISTA leads on samples A, C and D and is level with the strongest baseline on B, where every method degrades sharply and the ordering is not meaningful.

One fold overlaps the setting reported in Table 2, which allows a direct consistency check: patient D scores 0.557 here against 0.550 there, under considerably less supervision. That the two agree to within a point, despite the large difference in supervision, suggests the within-patient result is not carried by cohort-specific cues.

Results are in Table 3.

D.3 Intra-tissue held-out fractions

Table 7 reports the 20 per cent held-out fraction, the least demanding of the three. Table 12 gives the other two. The comparison worth making is not the absolute level but its stability: quadrupling the held-out fraction from 20 to 80 per cent leaves VISTA's structural metrics essentially unchanged (Moran's I 0.8143 to 0.7996, Geary's C 0.1858 to 0.1897) and costs 4.6% of site correlation. An interpolator behaves differently, because its predictions depend on how close the nearest measured site happens to be: kNN-IDW loses ground steadily as the held-out fraction grows (0.5640 to 0.5205). Stability under thinning supervision is therefore evidence that the model has learned a morphology–expression relationship rather than an interpolation rule over nearby measured sites.

Table 12: Intra-tissue held-out fractions of 40% and 80% on the continuous 3D HT volume. Best per setting in **bold**.

		3D Structure Quality				Similarity with 2D ST	
Test %	Model	Moran's I ↑	Geary's C ↓	Morphology*	BES ↑	Site corr. ↑	MSE ↓
40%	kNN-IDW	$0.6621_{\pm 0.0018}$	$0.3381_{\pm 0.0021}$	$0.0408_{\pm 0.0013}$	–	$0.5510_{\pm 0.0019}$	$\mathbf{0.1444}_{\pm 0.0017}$
	ASIGN	$0.2854_{\pm 0.0019}$	$0.7249_{\pm 0.0024}$	$0.0616_{\pm 0.0015}$	$1.06_{\pm 0.01}$	$0.5364_{\pm 0.0021}$	$0.1613_{\pm 0.0018}$
	VISTA	$\mathbf{0.7763}_{\pm 0.0025}$	$\mathbf{0.2225}_{\pm 0.0016}$	$\mathbf{0.3393}_{\pm 0.0018}$	$\mathbf{1.62}_{\pm 0.04}$	$\mathbf{0.5633}_{\pm 0.0023}$	$0.1583_{\pm 0.0015}$
80%	kNN-IDW	$0.6482_{\pm 0.0020}$	$0.3521_{\pm 0.0018}$	$0.0408_{\pm 0.0011}$	–	$0.5205_{\pm 0.0023}$	$\mathbf{0.1538}_{\pm 0.0015}$
	ASIGN	$0.2522_{\pm 0.0022}$	$0.7578_{\pm 0.0015}$	$0.0757_{\pm 0.0019}$	$1.10_{\pm 0.02}$	$0.5157_{\pm 0.0025}$	$0.1638_{\pm 0.0020}$
	VISTA	$\mathbf{0.7996}_{\pm 0.0017}$	$\mathbf{0.1897}_{\pm 0.0024}$	$\mathbf{0.3189}_{\pm 0.0021}$	$\mathbf{1.62}_{\pm 0.03}$	$\mathbf{0.5388}_{\pm 0.0016}$	$0.1651_{\pm 0.0022}$

*Optimal value is near average; values near zero or above average indicate poor predictions Chelebian et al., 2025.

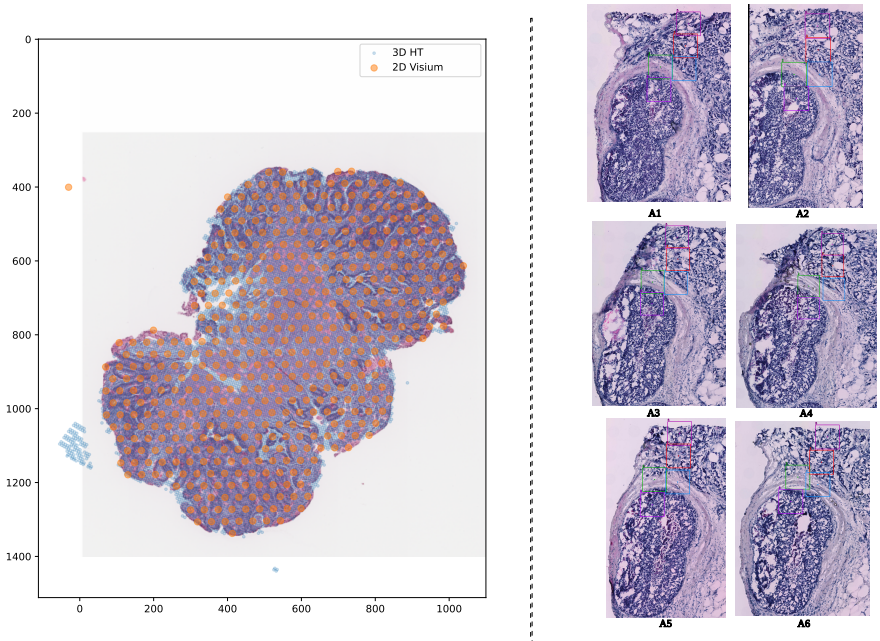
The same split applies to the 2D control of Table 8, whose 40% and 80% held-out columns are given in Table 13. The ordering of methods is unchanged across fractions, so the conclusion drawn at 20% does not depend on how much of the tissue is held out.

Table 13: 2D control on HT at the 40% and 80% held-out fractions (correlation ↑).

Method	40%	80%
kNN-IDW	0.542	0.497
BLEEP	0.546	0.533
Hist2ST	0.501	0.497
STGAT	0.331	0.319
THItoGene	0.521	0.502
TRIPLEX	0.513	0.495
ASIGN	0.389	0.369
X-pression	0.491	0.203
VISTA-2D	<u>0.548</u>	<u>0.532</u>
VISTA-3D	**0.563**	**0.539**

E DATA ALIGNMENT AND REGISTRATION

In case of continuous 3D volume experiment in Section, the two different types of data- HT and Visium, vary in resolution, shape and structure due to the distinctiveness of each technology. In preparation for multimodal integration, they need to undergo preprocessing steps that harmonize their spatial coordinate systems and resolution scales, establishing a unified space for downstream cross-modal analysis and predictive modeling. We leveraged the STalign Clifton et al. (2023) framework to align the positional coordinates of the HT tiles and Visium spots of the same tissue sample. As depicted in Figure 5a), most of the tiles get aligned perfectly to corresponding spots.



a) Registration between Visium and HT

b) Registration between 2D ST sections

Figure 5: **Data alignment**. a) Registration between HT and Visium data displayed on top b) Registration between 2D ST sections.

Table 14 quantifies the cross-modal alignment, computed from the saved alignment and the measured spot coordinates with the physical scale fixed by the $100\,\mu$m Visium spot pitch. Dice is the overlap between the aligned HT-tile footprint and the Visium spot footprint; values above 0.8 are conventionally regarded as good agreement.

Table 14: Cross-modal registration quality between HT tiles and Visium spots.

Sample	Dice ↑	IoU ↑	Spots receiving aligned tiles
RN49 tumor0	0.934	0.877	98.4%
RN49 tumor1	0.849	0.738	90.5%
HL55 tumor0	0.843	0.728	82.4%
HL55 tumor1	0.888	0.798	93.0%

The serial 2D ST sections are also required to be registered so the corresponding spots can be mapped to each other. We used BigWarp Bogovic et al. (2016) tools from Fiji (ImageJ) (Schindelin et al., 2012) to align the 2D H&E stained whole slide images (WSI) available for the Visium data. Figure 5b) displays perfect harmonization among five registered spots bounded by different colored square boxes across six 2D ST sections for sample A.

Registration places all sections in a common frame, so the dispersion of a correspondence group about its centroid is the post-registration residual: the median is 42–$68\,\mu$m over all eight samples, or 0.21–0.34 of the $200\,\mu$m spot pitch. Registration error is therefore below spot spacing, so spot-to-spot correspondence is unambiguous, which is the property the pipeline requires. An independent estimator agrees: leave-one-out landmark target registration error gives $\mathrm{rTRE} = 1.05\%$ of the image diagonal (68.7 px) against 69.0 px from the residual metric on the same sample. From the section images, tissue Dice against the reference rises from 0.595 to 0.792 on every eligible sample, and the centroid offset falls from roughly 10% to 1.5% of image width. Every metric degrades monotonically under deliberate misalignment, peaking at the achieved registration.

F Implementation details

Slice features come from UNI2-h (Chen et al., 2024), held frozen, with its patch tokens adaptively pooled to the working token grid. The VISTA model was trained using the AdamW optimizer with a learning rate of 1×10^{-4}. The Smooth L1 loss was used with $\beta = 1$. The Schrödinger-bridge matching term was weighted by $\lambda_{\mathrm{SB}} = 0.10$, with bridge regularizer $\varepsilon = 0.1$. For tile selection, the top $k = 65$ tiles were retained per spot, with a spot diameter of 32 pixels and a maximum of 16 tiles sampled per spot. Training was performed with a spot-level batch size of 1. The 3D neighborhood size was set to $k_{3\mathrm{D}} = 8$.

All experiments were implemented using `Pytorch` (Paszke et al., 2019) and executed on a server with an `AMD Ryzen 9` processor, `124Gigabyes` RAM and an `Nvidia RTX A4500` GPU with `20BB` memory (CUDA 11.3), running Ubuntu 22.04 operating system. For the registration of different modalities of image, we used BigWarp (Bogovic et al., 2016) tools from Fiji (ImageJ) (Schindelin et al., 2012). STalign (Clifton et al., 2023) was used for coordinate alignment. Furthermore, we used Napari (Sofroniew et al., 2022) software for visualization of 3D image and Matplotlib (Hunter, 2007) for displaying 2D image sections.

G Baseline descriptions

We compare VISTA against two naive baselines, five 2D histology-to-expression methods, and two 3D spatial transcriptomics methods. Below we briefly describe each baseline and provide its source. All learned baselines are trained using the hyperparameters and training procedures recommended in their respective papers and public codebases.

G.1 Naive baseline

kNN-IDW (k-nearest neighbor inverse-distance weighting). For each test spot s, we identify the k nearest training spots $\{s_1, \ldots, s_k\} \subset \mathcal{S}_{\mathrm{train}}$ by Euclidean distance in the 2D Visium coordinate plane and predict

$$\widehat{y}_s = \frac{\sum_{i=1}^{k} w_i \, y_{s_i}}{\sum_{i=1}^{k} w_i}, \qquad w_i = \frac{1}{d(s,\, s_i)}, \tag{13}$$

where $y_{s_i} \in \mathbb{R}^G$ is the measured expression of training spot s_i and $d(s, s_i)$ is the Euclidean distance between spots in the Visium coordinate space. This produces a spatially smooth 2D expression field. For volumetric evaluation in the continuous 3D setting, the 2D prediction is replicated across all Z depth slices, yielding a volume with spatial coherence but no genuine depth structure. We use k=10 in all experiments.

G.2 2D HISTOLOGY-TO-EXPRESSION METHODS

Hist2ST (Zeng et al., 2022). A deep learning model that predicts spatial gene expression from H&E histology by combining three modules: a Convmixer for local 2D vision features around each spot, a Transformer for global spatial dependencies across the whole image, and a graph neural network for local spatial dependencies between neighboring patches. Predictions are modeled with a zero-inflated negative binomial (ZINB) distribution, and a self-distillation mechanism is used during training. Code: `https://github.com/biomed-AI/Hist2ST`.

THItoGene (Jia et al., 2024). A hybrid neural network that uses dynamic convolutions and capsule networks to extract multi-view neighborhood features from H&E images, combined with a transformer module for spatial context. The capsule network captures relationships between high-resolution pathology phenotypes and gene expression. Code: `https://github.com/yrjia1015/THItoGene`.

STGAT (Baul et al., 2024). A Spatial Transcriptomics Graph Attention Network that uses Graph Attention Networks (GAT) over Visium spots to capture spatial dependencies between neighboring spots while predicting per-spot gene expression. STGAT was originally developed for predicting spot-level expression from whole-slide images and bulk RNA-seq in TCGA-style settings. Code: `https://github.com/compbiolabucf/STGAT`.

BLEEP (Xie et al., 2023). A bi-modal contrastive embedding framework that aligns H&E image patches and gene-expression profiles in a shared low-dimensional joint embedding space. At inference, gene expression for a query patch is predicted by averaging the expression profiles of its k nearest neighbors in the joint embedding space (query–reference imputation). Code: `https://github.com/bowang-lab/BLEEP`.

TRIPLEX (Chung et al., 2024). A multi-resolution framework that concurrently extracts features at three scales: the target spot, the local neighborhood around the spot, and the global tissue organization across the whole-slide image. The three feature streams are fused and passed through a linear head to predict spot-level gene expression. Code: `https://github.com/NEXGEM/TRIPLEX`.

G.3 3D SPATIAL TRANSCRIPTOMICS METHODS

X-pression (Túrós et al., 2025). A 3D convolutional neural network framework that reconstructs gene expression programs from volumetric micro-CT tissue data. X-pression is trained on a single 2D ST section paired with a co-registered 3D micro-CT volume, and predicts 2D expression patterns (rather than raw gene expression) for 3D volumetric patches centered on captured spots. Hence, although it could be used as a baseline for pseudo-3D experiment which had sparse ST sections with ground truth information, the model could not be faithfully applied on the continuous 3D volumetric tissue data. Code: `https://github.com/rockdeme/x-pression`.

ASIGN (Zhu et al., 2025). The Anatomy-aware Spatial Imputation Graph Network performs 3D ST imputation from a stack of 3D whole-slide histology sections combined with a single 2D ST slide. ASIGN extends 2D spatial relationships into 3D through cross-layer overlap and similarity-based expansion, and uses a multi-level spatial attention graph network to integrate features across data sources. Code: `https://github.com/hrlblab/ASIGN`.

H EVALUATION METRICS

Because dense 3D gene-expression ground truth is unavailable in the continuous-volume setting, we evaluate VISTA using complementary metrics that measure both volumetric structure and agreement with observed 2D ST measurements.

Moran's I. Moran's I measures spatial autocorrelation in the predicted 3D gene-expression field. It evaluates whether nearby locations have similar expression values. Its value typically ranges from -1 to 1, where values near 1 indicate strong positive spatial organization, values near 0 indicate weak or random spatial structure, and negative values indicate alternating or checkerboard-like patterns. In this study, a higher Moran's I is better because meaningful gene-expression fields are expected to show spatially coherent structure rather than random variation.

Geary's C. Geary's C measures local spatial dissimilarity between neighboring locations. Unlike Moran's I, which captures global spatial autocorrelation, Geary's C is more sensitive to local differences. Its expected value under spatial randomness is approximately 1. Values below 1 indicate positive spatial autocorrelation, meaning neighboring locations are similar, while values above 1 indicate local dissimilarity or noisy spatial variation. Therefore, lower Geary's C is better in our setting, as it reflects smoother and more locally coherent 3D molecular predictions.

Morphology gradient correlation. Morphology gradient correlation measures whether changes in predicted gene expression align with changes in the underlying tissue morphology. We compute gradients of the predicted 3D expression field and compare them with gradients of the volumetric tissue image. The score is a Pearson correlation and therefore ranges from -1 to 1. Values near 0 indicate that predicted expression changes are largely unrelated to morphology, while moderate positive values indicate that molecular transitions are aligned with anatomical structure. Extremely high values may suggest that the model is copying morphology too directly rather than learning molecular structure Chelebian et al., 2025. In this study, moderate positive values are preferred.

Site-level correlation. Site-level correlation measures agreement between predicted and measured ST expression at the unit the assay provides: a Visium spot, an ST spot, or a $100\,\mu$m bin. On dense volumes it is computed after aggregating the predicted 3D field to the observed spot support. We use Pearson correlation, which ranges from -1 to 1. Higher values indicate better agreement with the measured ST profiles. This metric verifies that the predicted 3D field remains anchored to observed molecular measurements. However, it should not be used alone, because a model can achieve reasonable site-level correlation while producing a depth-collapsed or spatially uninformative 3D volume.

Mean squared error. Mean squared error (MSE) measures the average squared difference between aggregated predicted expression and measured ST expression. It is non-negative, with 0 indicating perfect agreement. Lower MSE is better. Like site-level correlation, MSE evaluates consistency with observed ST measurements, but it does not directly assess whether the predicted 3D field is biologically plausible or spatially coherent across depth.

Boundary enrichment score. The boundary enrichment score (BES) measures whether predicted expression changes are concentrated near tissue-structure boundaries. We first identify tissue boundaries using high-gradient regions in the volumetric morphology image, and then compare the magnitude of predicted expression gradients near these boundaries with the magnitude in interior regions. A BES of approximately 1 indicates no enrichment, meaning expression changes are no stronger at tissue boundaries than elsewhere. Values greater than 1 indicate that predicted molecular transitions are enriched near morphological boundaries, while values below 1 suggest weaker boundary association. In this study, higher BES is better because biologically meaningful molecular changes are expected to align with structural transitions in the tissue. However, as with morphology gradient correlation, extremely high values should be interpreted carefully, since they may indicate that the model is overly driven by morphology rather than learning gene-specific molecular structure.

Together, these metrics separate two aspects of performance: 3D structural quality and 2D molecular consistency. Moran's I, Geary's C, morphology gradient correlation, and BES evaluate whether the

predicted expression forms a coherent and morphology-aligned 3D molecular landscape, while site-level correlation and MSE evaluate whether the prediction remains consistent with measured ST data.

I EVALUATION AND REGISTRATION CONSIDERATIONS

Evaluation and registration remain critical challenges for volumetric expression prediction, and both shaped how the results in Section 4 are reported.

Standard in-plane metrics such as site-level correlation and MSE can be misleading on their own, because a simple interpolation baseline may perform competitively after projection while producing a constant or depth-collapsed volume. This is measured rather than hypothetical: on HT, kNN-IDW reaches a site correlation within a few points of VISTA (Table 7), yet on Open-ST, where depth is measured, the same baseline recovers only 5.8% of the attainable depth signal (Table 4). Projection discards exactly the structure the task is about, so a method can score well on the projected measurement while failing the underlying problem. 3D prediction should therefore be assessed with volumetric criteria — spatial coherence, depth non-collapse, morphology coupling, and boundary enrichment — reported alongside in-plane agreement rather than in place of it, which is the convention followed throughout this paper.

VISTA also depends on accurate registration between the 3D tissue volume and the 2D ST section. Misalignment assigns tiles to the wrong spots, and that error propagates into training rather than averaging out, because the supervision signal itself is displaced. The effect is most pronounced near tissue boundaries and in deformed regions, where the local correspondence changes fastest and a small spatial error therefore carries the largest change in expression. Appendix E reports the alignment quality achieved on each sample, and the residual misregistration there is a floor on the accuracy attainable in the continuous 3D setting.