STACK-JEPA: MOLECULAR GRAPH DISTILLATION FOR 3D SPATIAL TRANSCRIPTOMICS

Anonymous authors
Paper under double-blind review

ABSTRACT

Spatial transcriptomics (ST) observes isolated two-dimensional sections, cutting anatomy and molecular neighborhoods at one arbitrary plane. Three-dimensional (3D) ST lifts that limit, but profiling a section by ST costs far more than imaging it by histology, so volumes are built by imaging every section with hematoxylin and eosin (H&E), profiling a sparse subset, and imputing the rest. Existing imputers are supervised in expression space: they regress the expression of one fixed gene panel spot by spot from paired sections, so the relations between neighboring spots are never a training target, a new panel forces a refit, and accuracy is capped by an inter-section alignment computed once before the model runs. We introduce STACK-JEPA, a masked molecular graph-distillation framework for learning molecularly informed representations throughout a serial histology stack. On a profiled section, an exponential-moving-average (EMA) teacher defines both per-location latent states and their local molecular relations. A stack student is trained to recover the held-out section's molecular representation and relation graph from a profiled context section, the target H&E image, in-plane coordinates, and tissue depth. Frozen TERRA and UNI2-h encoders feed a transformer with 3D rotary position embeddings, while node- and relation-level distillation transfer molecular structure without placing the gene-panel dimension in the backbone. After pretraining, the backbone is frozen and only a lightweight panel-specific count decoder is fitted. The same learned descriptors can replace raw expression in partial fused Gromov-Wasserstein transport, including when the target section has no measured transcriptome. Across three tasks on public 3D ST benchmarks, STACK-JEPA outperforms the supervised and generative state of the art on imputation in both per-cell and per-gene correlation, recovers the spatial neighborhood graph of held-out sections more accurately than any expression-based representation, and improves on PASTE2 alignment under its subslice protocol while also aligning sections for which no transcriptome was measured, which expression-based aligners cannot handle at all.

1 INTRODUCTION

Spatial transcriptomics (ST) measures gene expression while preserving the spatial coordinates of each measured spot or cell (Ståhl et al., 2016), linking molecular state to the tissue morphology in which it arises. Paired histology and expression are now available at the scale of thousands of samples across both spot-array and cell-resolution platforms (Marx, 2021; Bressan et al., 2023; Jaume et al., 2024). However, almost all of that data is read one section at a time. A single plane truncates every structure that crosses it, so a tumor boundary, vessel, cortical layer, or multicellular niche is observed only as its intersection with the cut, and the neighborhood of each location stops at the section face instead of continuing into the often nearer material above and below (Lin et al., 2026; Sui et al., 2025).

Serial-section datasets now make that organization measurable, on spot arrays (Andersson et al., 2021; He et al., 2020; Maynard et al., 2021) and at cell resolution (Schott et al., 2024; Sui et al., 2025) alike. They do, however, expose an acquisition asymmetry that will not disappear: a 2D ST acquisition costs substantially more than a whole-slide histology image, so a ten-section stack costs an order of magnitude more again (Zhu et al., 2025a). Sections are therefore profiled sparsely along depth, one every few tens to hundreds of microns although each is only 10 to 20 microns

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.

thick (Bressan et al., 2023; Zhang et al., 2026), while H&E can be collected on every section at comparatively low cost. Morphology is therefore observed throughout the stack and transcriptome on a subset of sections, and the task is to infer the rest using, rather than collapsing, the depth axis.

Recent work come in two forms. One predicts expression from histology one section at a time, differing in how dependence within the plane is modeled, from spot-wise regression to slide-level generative refinement (He et al., 2020; Pang et al., 2021; Zeng et al., 2022; Xie et al., 2023; Min et al., 2024; Zhu et al., 2025b; Huang et al., 2025). The cut is the unit of inference: tissue above and below it is unavailable, and a profiled section cannot anchor its neighbors. The other conditions unprofiled sections on profiled ones through inter-layer correspondences (Zhu et al., 2025a; Sanian et al., 2025; Li et al., 2025), so axial context reaches only the few layers those correspondences link. Both types of methodologies are supervised in expression space, where the model regresses counts directly, so whatever correspondence it learns between histology and transcriptome lives in the output layer of one panel. Both train and evaluate on a fixed panel of fifty to a few hundred genes chosen by variability or mean expression (Huang et al., 2025; Zhu et al., 2025a; Sanian et al., 2025), so a new panel changes the output dimension and forces a refit. Such evaluations measure the accuracy of one expression matrix, not whether a 3D representation has been learned that transfers across panels and downstream tasks.

Our approach. We introduce STACK-JEPA, a JEPA for serial tissue sections that predicts latent molecular state along tissue depth instead of reconstructing counts (Assran et al., 2023; Bardes et al., 2024). Each cell or spot is one token, tissue depth plays the role of V-JEPA's temporal axis (Assran et al., 2025; Mur-Labadia et al., 2026), and masking runs along that axis: an entire section is hidden rather than a patch within a frame. This carries latent prediction from fixed token grids, and from the unordered geometry of point-cloud JEPAs (Saito et al., 2025; Hu et al., 2024a), to ordered sequences of variable-size spot sets. During training one profiled section from the same specimen, drawn with exponential decay in section distance, supplies context. Frozen TERRA (Birk et al., 2026) and UNI2-h (Chen et al., 2024) encode molecular state and local morphology, a trainable transformer models their joint context, 3D rotary embeddings expose in-plane location and section depth, and the predictor receives only the target coordinates and target H&E. The prediction target is the teacher's own embedding of the hidden section, so pretraining is self-supervised and the gene dimension never enters the backbone. Expression readout is a separate, cheap step: at inference the encoder-predictor stays frozen and a panel-specific count decoder is fitted from held-in sections. With an auxiliary graph-reconstruction loss (Sec. 2.3), the learned embeddings also replace expression as the feature cost of partial fused Gromov–Wasserstein alignment (Sec. 3.3), extending it to sections whose transcriptome was never measured.

Contributions.

- We adapt V-JEPA to variable-size spot sets through section-relative in-plane position and physical or nominal depth, distance-decayed context sampling, target-location histology, and dense hierarchical supervision over both the masked section and the visible context.
- We decouple self-supervised 3D representation learning from supervised expression readout: within a dataset, one frozen backbone serves spatial panels of different widths through separately fitted decoders.
- We frame masked latent prediction as 3D graph distillation that aligns histology with transcriptomic state, and strengthen that alignment with a spatial-graph head and a fused Gromov-Wasserstein objective over the relational structure of the two representations.
- Against ASIGN, STFlow, HoloTea, and relevant baselines, STACK-JEPA attains the highest mean per-cell and per-gene correlation in all six dataset-panel settings across three public serial-section datasets.

2 METHOD

STACK-JEPA casts missing-section imputation as masked latent prediction along the depth axis of a serial-section stack. Every hyperparameter and the optimization setup are in Table 5.

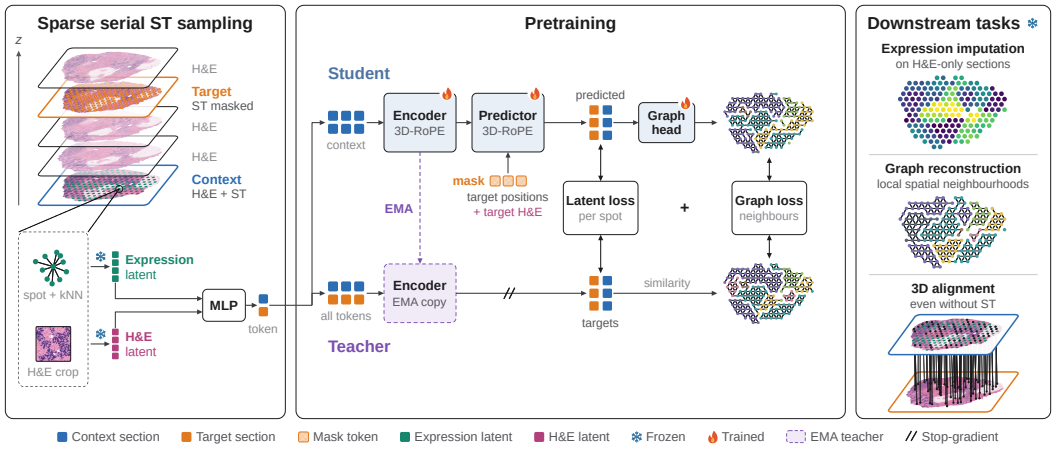


Figure 1: **Overview of STACK-JEPA. Left:** every serial section of a specimen carries H&E, but only a sparse subset is profiled by ST. Concordantly, training pairs a profiled context section with a target section whose expression is masked. Each spot becomes one token, fusing through an MLP a frozen TERRA latent of the spot and its in-plane kNN neighbors with a frozen UNI2-h latent of its H&E crop. **Middle:** the student encoder reads the context tokens, and the predictor fills mask tokens placed at the target positions and conditioned on target H&E, both under 3D-RoPE. The predictions regress the multi-level features of an EMA teacher that encodes all tokens without gradient (latent loss, Eq. 2). A graph head recovers the representational similarity of adjacent target spots from the predicted states (graph loss, Eq. 3). **Right:** with the backbone frozen, the learned states support expression imputation on H&E-only sections, reconstruction of local spatial neighborhoods, and 3D alignment of adjacent sections, including sections with no ST.

2.1 SETUP AND TOKENIZATION

Problem formulation and masking strategy. A specimen is a stack of sections $\{S_k\}_{k=1}^{N}$ at physical or nominal axial positions z_k. Section S_k contains n_k cells or spots, and each one is a token: location i has in-plane coordinates $c_i^k \in \mathbb{R}^2$, an H&E crop I_i^k, and a count vector $g_i^k \in \mathbb{N}^G$. Each training sample is a section pair (S_t, S_c) of a target and a context section from the same specimen. Every molecular input from S_t is withheld from the online path; only its H&E and coordinates remain visible, which is exactly what an unprofiled section supplies at deployment, so training and inference share one information boundary. The model predicts a fixed-width latent state $r_i^t \in \mathbb{R}^d$ at each target location, keeping the gene dimension out of the 3D backbone: one encoder-predictor serves every panel of a dataset through separately fitted readouts (Sec. 2.4), and one backbone is trained per fold under the protocol of Sec. 3.1.

Frozen molecular and morphological features. Molecular state is encoded by a frozen TERRA cell-graph model (Birk et al., 2026), which summarizes a given cell from its own highest-ranked genes and those of its nearest in-plane neighbors. Morphology is encoded by a frozen UNI2-h model (Chen et al., 2024) applied to the section's own H&E image. The two are normalized, concatenated, and condensed to the token width d by a two-layer MLP, $x_i^k = W_2\,\sigma(W_1[\mathrm{LN}(e_i^{\mathrm{TERRA}}); \mathrm{LN}(e_i^{\mathrm{UNI}})])$ with LN a per-token LayerNorm (Ba et al., 2016) and GELU nonlinearity σ.

Three-dimensional rotary position. Every attention block uses rotary position embeddings (Su et al., 2024) in the 3D form of V-JEPA 2 (Assran et al., 2025), which partitions per-head channels across z, y, and x and rotates each segment by its own axis. In-plane coordinates are taken relative to the section-pair minimum in a fixed physical unit; depth is divided by the smallest section spacing in the dataset, preserving irregular source gaps.

2.2 MASKED-SECTION LATENT PREDICTION

We draw a target section t uniformly from a specimen, choose an available side of it uniformly, and sample the context section at rank distance r along that side with

$$p(r) \propto \exp[-(r-1)/\tau]\,, \tag{1}$$

for a decay constant τ. Every target cell is masked and exactly one complete context section is visible. The decay favors adjacent context while exposing the model to the longer gaps seen at evaluation. Rank governs sampling only, where axial coordinates, with gaps included, enter separately through RoPE. Let f_θ be the online encoder and $f_{\bar\theta}$ its exponential moving average, updated as $\bar\theta \leftarrow m\bar\theta + (1-m)\theta$ at a constant momentum m; both are transformers of width d. Following the deep self-supervision of V-JEPA 2.1 (Mur-Labadia et al., 2026), we read normalized features after a fixed set of intermediate blocks $\mathcal{K}$. The target encoder processes the complete section pair without gradient, including S_t's TERRA and UNI2-h features; the online encoder sees only S_c. Its $|\mathcal{K}|$ levels are concatenated, projected to predictor width, and consumed by a predictor q_ϕ of the same depth. At each target location the predictor appends a learned mask token placed at (x_i^t, y_i^t, z_t) and conditioned on target morphology through a zero-initialized projection $m_i^t = m_0 + W_{\mathrm{H\&E}}\, e_i^{\mathrm{UNI}}$ with $W_{\mathrm{H\&E}} \in \mathbb{R}^{d \times d_{\mathrm{UNI}}}$. The distillation target is therefore the EMA transformation of fused TERRA-UNI input.

Pretraining objective. The predictor estimates all levels of $\mathcal{K}$. With $h_{T,\ell}$ the target-encoder features at target positions and $\hat h_{T,\ell}$ their predictions, the masked-section loss and the pretraining objective are

$$\mathcal{L}_{\mathrm{pred}} = \frac{1}{|\mathcal{K}|} \sum_{\ell \in \mathcal{K}} \big\langle \big|\hat h_{T,\ell} - \mathrm{LN}(h_{T,\ell})\big| \big\rangle_T, \qquad \mathcal{L}_{\mathrm{dense}} = \mathcal{L}_{\mathrm{pred}} + \lambda(s)\, \mathcal{L}_{\mathrm{ctx}}. \tag{2}$$

Here LN carries no affine parameters and applies to the target side alone, so the predictor estimates a layer-normed state without emitting one. As in the dense predictive loss of V-JEPA 2.1, the predictor also estimates target-encoder features at visible context positions, weighting context cell i by $w_i = 1/\sqrt{\max(\delta_i, 1)}$ for δ_i its model-coordinate distance to the nearest target cell. Averaging the same multi-level L_1 error over context cells gives $\mathcal{L}_{\mathrm{ctx}}$. $\lambda(s)$ is held at zero early in training and then ramped linearly to its full value, allowing the masked-section term to shape the representation first.

2.3 SPATIAL GRAPH RECONSTRUCTION

The objective in Eq. 2 supervises target representations pointwise, but does not explicitly constrain their local organization as tissue, e.g. which spots form coherent structures and where boundaries lie. We therefore introduce a graph reconstruction loss that recovers the masked section's local molecular similarity structure over its neighbourhood graph from the predicted states.

Let $G^k = (V^k, A^k)$ denote the weighted spatial graph of section k, where nodes V^k are spot locations and edges A^k connect spatial neighbors. For the target section S_t, we define

$$A^t = \{(i,j) : i < j,\ \|c_i^t - c_j^t\| \le r\},$$

with r set to 1.1 times the platform pitch to capture the immediate neighbors of each spot. The edge set is fixed by position and is not predicted. The model reconstructs the edge weight s_{ij}, defined as the cosine similarity of frozen target embeddings.

Graph reconstruction objective. Let $\hat z_i^t \in \mathbb{R}^d$ denote the predicted embedding at target location i. We project these embeddings with a two-layer MLP and apply L layers of message passing over the radius graph to incorporate local neighbourhood information, yielding updated representations h_i. For each spatially adjacent pair $(i,j) \in \mathcal{A}^t$, the graph head predicts its molecular similarity as $\hat s_{ij} = \cos(h_i, h_j)$.

The head receives no coordinates, edge lengths, or relative displacements, so the predicted similarity must be inferred from the predicted molecular states and their local context. The loss is the squared error, averaged within each section before averaging over the batch,

$$\mathcal{L}_{\mathrm{graph}} = \frac{1}{B} \sum_t \frac{1}{|\mathcal{A}^t|} \sum_{(i,j) \in \mathcal{A}^t} \big(\hat s_{ij} - s_{ij}\big)^2, \tag{3}$$

because $|\mathcal{A}^t|$ grows with section size and would otherwise let the largest section of a batch set the weight of the term. Reconstructing edges and relative geometry keeps the imputed volume spatially self-consistent across sections. Target positions stay visible under our information contract, so the head performs graph completion.

The pre-training objective becomes

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{dense}} + \beta(t)\mathcal{L}_{\text{graph}} \tag{4}$$

with $\beta(t)$ ramped linearly from zero to β at the start of training.

2.4 Task 1. Expression readout

After pretraining, f_θ and q_ϕ are frozen and only a count decoder is fitted, once per (fold, panel). Its training field is built from predicted states, matching the states it is applied to: each held-in section is masked in turn, predicted from every other held-in section of the specimen, and the predictions averaged per cell. Evaluation uses the same two-section structure, pairing each target with its nearest held-in section. Per-token LayerNorm on both predicted and encoded states places them in a common space before the decoder. The readout is an scVI (Lopez et al., 2018) decoder with a single hidden layer, a softmax mean-fraction head, and a negative-binomial likelihood with global per-gene dispersion. It is trained on raw held-in counts with the observed panel library as offset, and applied at a fixed library size. Predictions are the NB mean μ, scored with every other method in the shared expression space of Appendix A so that metrics are comparable across methods.

2.5 Task 2. 2D Graph reconstruction

This task evaluates whether frozen representations preserve sufficient spatial information to recover local tissue organization without further task-specific training. Specifically, successful reconstruction indicates that the representations encode fine-grained spatial continuity and neighbourhood structure, rather than merely broad tissue identity or global spatial proximity. Importantly, the head is never trained on the ground-truth spatial graph A^k: its objective (Eq. 3) regresses expression similarity within the ring, whereas Task 2 evaluates binary spatial adjacency within a wider local ball. The task therefore tests whether the geometry of the learned representations captures local tissue topology rather than how well the head reproduces its training target.

2.6 Task 3. 3D alignment

Task 3 aligns adjacent sections from the same specimen. Given a source section A with n_A spots and a target section B with n_B spots, the aligner estimates a partial transport plan $\pi \in \mathcal{P}(g^A, g^B, s)$ where π_{ij} represents the mass transported between spot i in A and spot j in B, and s controls the fraction of mass to be matched. In the sparse-profiling setting, target expression may be unobserved, making expression-based alignment inapplicable.

Transport with learned embeddings. We use the partial fused Gromov-Wasserstein objective of PASTE2 (Liu et al., 2023), replacing only its expression-based feature cost:

$$\min_{\pi \in \mathcal{P}(g^A, g^B, s)} (1-\alpha) \sum_{i,j} M_{ij}\pi_{ij} + \alpha \sum_{i,j,k,l} \left(D^A_{ik} - D^B_{jl}\right)^2 \pi_{ij}\pi_{kl}, \qquad M_{ij} = 1 - \cos\left(u^A_i, u^B_j\right), \tag{5}$$

where D^A and D^B are within-section spatial distance matrices. PASTE2 computes M from gene-expression features of both sections; we instead compute it from Stack-JEPA embeddings u. When the target expression is observed, u^A and u^B are teacher embeddings of measured sections.

3 Experiments

3.1 Datasets and evaluation protocol

Datasets and splits. We evaluate on three public 3D spot-based spatial transcriptomics datasets (Table 6): HER2+ breast tumour (Andersson et al., 2021), ST-data breast tumour (He et al., 2020), and human dorsolateral prefrontal cortex (DLPFC) (Maynard et al., 2021), with 36, 48, and 12

sections from 8, 16, and 3 specimens and 13,467, 23,040, and 45,960 spots after per-section quality control, respectively. The native spot coordinates across each section and dataset is retained. The task is to complete partially profiled stacks under specimen-level cross-validation. In each fold, the first two sections of each held-out specimen are held in ('profiled') for representation training and decoder fitting, and its remaining sections are targets. All other specimens are fully available for training. Unless stated otherwise, every number is the mean over cross-validation folds $\pm$ the fold standard deviation. Fold construction and spacing are detailed in Table 6.

Gene panels. We report three gene panels. SVG-250 and SVG-500 are the 250 and 500 genes with the highest spatial autocorrelation, the standard criterion for spatially variable genes (Svensson et al., 2018). It is measured with Moran's I on each specimen's 3D Delaunay graph and averaged across specimens. Inspired by HEST (Jaume et al., 2024), we restrict candidates to genes with $\geq 10\%$ detection across sections. HEG-250, the 250 genes with the highest mean expression, is reported in Table 7.

3.2 Evaluation metrics and baselines

Expression readout task. We evaluate with PCC_{gene}, PCC_{cell} and MSE. PCC_{gene} is the Pearson correlation across cells, averaged over genes, and measures how well the spatial variation of each gene is recovered, while PCC_{cell} is the Pearson correlation across genes, averaged over cells, and measures how well each cell's expression profile is recovered. MSE measures the squared error of the predicted expression values.

For baselines, we include approaches that isolate the contributions of 3D proximity and local morphology. *Overlap*, inspired by ASIGN (Zhu et al., 2025a), assigns each target cell the expression of the nearest provided cell in registered 3D space. *UNI* applies the same count decoder as STACK-JEPA to frozen UNI2-h features from each target cell's own tile, providing a morphology-only baseline without cross-section information. We further compare against ASIGN (Zhu et al., 2025a), a supervised 3D imputation method; STFlow (Huang et al., 2025), a whole-slide 2D flow-matching model; and HoloTea (Sanian et al., 2025), its 3D-aware extension. All methods use the authors' published code and default settings without tuning on our data, and are evaluated in the shared space defined in Appendix A.

2D graph reconstruction task. We rank the fixed candidate pairs defined above and evaluate edge AUPRC over the candidate pairs compared to the ground-truth spatial graph. Ground truth is each held-out section's spatial κ-nearest-neighbour graph, symmetrized, within a radius of R pitched, so that no edge jumps across a tissue hole or a QC-removed regions. For a representation u, candidate pairs are scored by cosine similarity after section-wise centring, using the same centring as for the training target; for the graph head, we use $\hat{s}_{ij}$ directly.

For baselines, we compare methods only within the same target-expression setting. When target expression is unobserved, STACK-JEPA predicts from the nearest measured section using the target H&E and coordinates, and is compared with UNI2-h H&E cosine. When expression is observed, STACK-JEPA uses its EMA-teacher encoding and is compared with expression cosine, TERRA cosine, and LUNA (Yu et al., 2025), whose predicted coordinates are affinely fitted to the true layout. Random ranking provides the chance level.

3D alignment task. To assess whether spot embeddings learned by STACK-JEPA support cross-section alignment of corresponding spatial domains along the z-axis, we compare against PASTE2 (Liu et al., 2023) using the same evaluation protocol. Since PASTE2 requires measured expression in both sections, we compare it with STACK-JEPA only when target expression is observed. Full details of the evaluation protocol are provided in Appendix C.

3.3 Benchmark analysis

Expression readout task. STACK-JEPA attains the highest PCC_{gene} and PCC_{cell} of all published methods and baselines on every dataset and all three gene panels (Table 1; HEG-250 in Table 7), and the lowest MSE on SVG-250 and SVG-500. We include HEG-250 results in Table 7), to demonstrate that previously published values on the same methods and datasets have been reproduced, even under

Table 1: **Expression readout on held-out sections.** SVG-250 and SVG-500 are the top 250 and top 500 genes of the same spatial ranking among those clearing the detection floor of Sec. 3.1. All methods use the shared log-composition space of Appendix A. Best per column in bold.

		HER2+			ST-data			DLPFC		
Panel	Method	$PCC_{gene}\uparrow$	$PCC_{cell}\uparrow$	MSE↓	$PCC_{gene}\uparrow$	$PCC_{cell}\uparrow$	MSE↓	$PCC_{gene}\uparrow$	$PCC_{cell}\uparrow$	MSE↓
SVG-250	Overlap	0.170 ±0.134	0.646 ±0.083	0.747	0.294 ±0.047	0.693 ±0.038	0.602	0.085 ±0.085	0.745 ±0.059	0.529
	UNI (Chen et al., 2024)	0.343 ±0.098	0.738 ±0.059	0.531	0.391 ±0.034	0.776 ±0.032	0.458	0.244 ±0.085	0.844 ±0.029	0.335
	ASIGN (Zhu et al., 2025a)	0.201 ±0.197	0.743 ±0.059	0.615	0.419 ±0.081	0.793 ±0.036	0.450	0.065 ±0.066	0.835 ±0.033	0.343
	STFlow (Huang et al., 2025)	0.359 ±0.117	0.747 ±0.055	0.522	0.442 ±0.046	0.794 ±0.020	0.419	0.220 ±0.111	0.843 ±0.031	0.331
	HoloTea (Sanian et al., 2025)	0.384 ±0.096	0.758 ±0.052	0.500	0.398 ±0.149	0.763 ±0.065	0.482	0.235 ±0.101	0.842 ±0.031	0.336
	STACK-JEPA (ours)	**0.406 ±0.096**	**0.777 ±0.060**	**0.479**	**0.468 ±0.027**	**0.808 ±0.025**	**0.400**	**0.257 ±0.083**	**0.848 ±0.030**	**0.322**
SVG-500	Overlap	0.119 ±0.096	0.585 ±0.084	0.672	0.209 ±0.045	0.650 ±0.036	0.576	0.051 ±0.054	0.671 ±0.061	0.497
	UNI (Chen et al., 2024)	0.281 ±0.075	0.702 ±0.065	0.475	0.326 ±0.030	0.756 ±0.028	0.416	0.195 ±0.059	0.799 ±0.034	0.312
	ASIGN (Zhu et al., 2025a)	0.154 ±0.152	0.709 ±0.064	0.528	0.321 ±0.088	0.770 ±0.033	0.415	0.039 ±0.044	0.790 ±0.036	0.318
	STFlow (Huang et al., 2025)	0.279 ±0.113	0.696 ±0.077	0.488	0.355 ±0.048	0.765 ±0.028	0.396	0.168 ±0.098	0.790 ±0.030	0.324
	HoloTea (Sanian et al., 2025)	0.284 ±0.106	0.687 ±0.078	0.478	0.324 ±0.128	0.739 ±0.050	0.439	0.158 ±0.105	0.787 ±0.034	0.322
	STACK-JEPA (ours)	**0.332 ±0.081**	**0.735 ±0.064**	**0.443**	**0.390 ±0.027**	**0.783 ±0.023**	**0.376**	**0.201 ±0.063**	**0.802 ±0.034**	**0.306**

Table 2: **2D graph reconstruction on held-out sections** (edge AUPRC). *Unobserved*: STACK-JEPA predicts the target from the nearest profiled section; *observed*: it uses its spot encoding. Best in bold.

Method	HER2+	ST-data	DLPFC
Target expression unobserved			
Random	0.391 ±0.009	0.388 ±0.008	0.372 ±0.003
UNI cosine	0.493 ±0.007	0.488 ±0.019	**0.467 ±0.002**
STACK-JEPA	**0.588 ±0.014**	**0.630 ±0.016**	0.455 ±0.023
Target expression observed			
LUNA	0.411 ±0.014	0.399 ±0.008	0.378 ±0.003
Expression cosine	0.455 ±0.018	0.441 ±0.009	0.386 ±0.009
TERRA cosine	0.468 ±0.018	0.444 ±0.013	0.408 ±0.020
STACK-JEPA	**0.618 ±0.035**	**0.608 ±0.013**	**0.452 ±0.020**

Table 3: **3D alignment of adjacent held-out sections** (LTARI under PASTE2 protocol).Feature cost from measured expression (PASTE2) or STACK-JEPA embeddings, imputed for unobserved targets (—: not applicable). Best in bold.

Method	ST-data	DLPFC
Target expression unobserved		
PASTE2	—	—
STACK-JEPA	**0.145 ±0.027**	**0.204 ±0.078**
Target expression observed		
PASTE2	0.148 ±0.023	0.367 ±0.029
STACK-JEPA	**0.165 ±0.023**	**0.408 ±0.039**

the shared-space scoring. On SVG-500 STACK-JEPA exceeds the strongest published method on each dataset by an average of 15.5% in PCC_{gene} and 2.2% in PCC_{cell} (per-fold scores for HER2+ in Figure 4). The larger margin on PCC_{gene} means that STACK-JEPA recovers where each gene varies across the target section, not only the expression profile of each cell, which is the signal that spatial analyses of the completed volume read. Across the six dataset-panel settings, STACK-JEPA also has the lowest average fold standard deviation of any method on both correlations, 0.063 in PCC_{gene} and 0.039 in PCC_{cell}.

2D graph reconstruction task. Table 2 reports edge AUPRC on held-out sections. With the target transcriptome unobserved, STACK-JEPA clearly exceeds UNI2-h cosine on both breast-tumour datasets (0.588 vs. 0.493 on HER2+, 0.630 vs. 0.488 on ST-data), so the predicted states carry molecular continuity that morphology alone does not. On DLPFC the two are within one standard deviation: cortical layers are broad bands that morphology already resolves, and the $300\,\mu\mathrm{m}$ B–C gap is the longest extrapolation in the benchmark. With the target observed, STACK-JEPA outperforms every expression-based ranking, including its own TERRA input (0.618 vs. 0.468, 0.608 vs. 0.444, 0.452 vs. 0.408), as well as LUNA, whose predicted coordinates capture global layout rather than the spot-level adjacency scored here.

3D alignment task. Table 3 reports LTARI under both regimes, with the overlap fraction s selected by the solver rather than given. With both sections profiled, replacing PASTE2's expression cost with STACK-JEPA embeddings raises LTARI from 0.148 to 0.165 on ST-data and from 0.367 to 0.408 on DLPFC. Without a target transcriptome PASTE2 cannot run, whereas STACK-JEPA aligns from imputed states and retains 88% of its profiled-regime score on ST-data (0.145) and 50% on DLPFC (0.204), where the $300\,\mu\mathrm{m}$ B–C pairs dominate the average and, as in Task 2, impute worst. On ST-data the imputed regime still matches PASTE2 with measured expression on both sides.

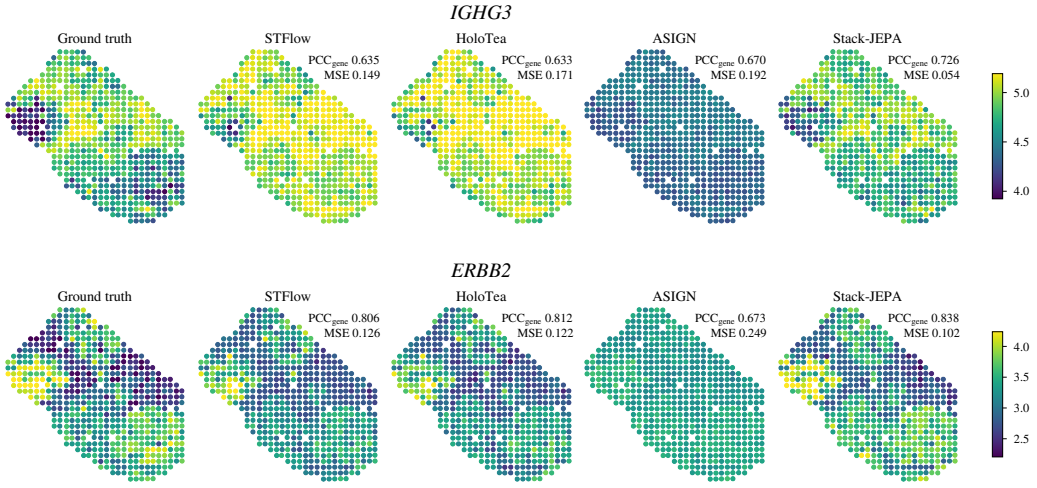


Figure 2: *IGHG3* (top) and *ERBB2* (bottom) on held-out HER2+ section E3. In each row, predictions share the ground truth's color scale; PCC_{gene} and MSE are computed over the section's spots.

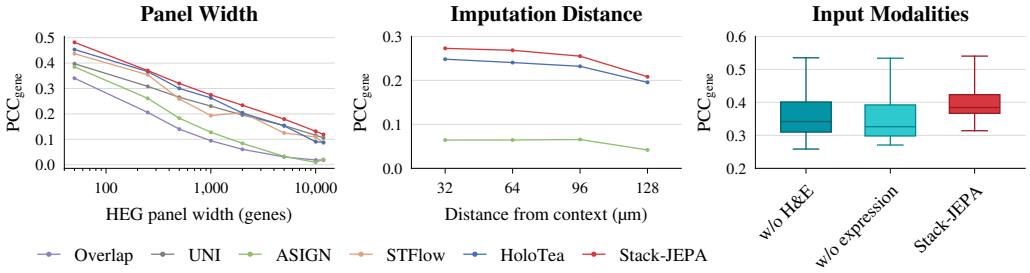


Figure 3: Ablations on HER2+. **Left:** PCC_{gene} of all methods on nested HEG panels, cross-validation means, from the 50 most highly expressed genes to the full set of 11,871 (rightmost point). **Middle:** PCC_{gene} of the three 3D methods on SVG-250, on the target sections of specimens A-D by distance from the nearer profiled section, averaged over the four specimens, one per fold. **Right:** PCC_{gene} on SVG-250 of STACK-JEPA and its two single-modality variants, *w/o H&E* (TERRA only) and *w/o expression* (UNI2-h only), each retrained with the other modality removed and decoded and scored as in Table 1. Each box summarizes the four folds (line: median, box: interquartile range, whiskers: lowest and highest fold).

4 ANALYSIS

Recovery of spatially relevant genes. STACK-JEPA recovers the spatial expression of key HER2+ genes more faithfully than the published methods. On a held-out section, the immunoglobulin gene *IGHG3*, a marker of the tumor's tertiary lymphoid-like structures (Andersson et al., 2021), is depleted where *ERBB2*, the amplified oncogene that defines the HER2+ subtype (Swain et al., 2023), peaks (Figure 2). The lead of STACK-JEPA in Table 1 holds for these disease genes: it attains the highest PCC_{gene} and lowest MSE on both, reproducing their complementary layout, their expression level and the sharp boundaries of the *ERBB2*-high domains. STFlow and HoloTea overpredict *IGHG3* across the section and recover the *ERBB2*-high domains, but at reduced contrast, whereas ASIGN predicts a nearly uniform field for both genes.

Panel width. Figure 3 (left) widens the panel along the mean-expression ranking that defines HEG-250, from 50 genes to all 11,871, so each panel contains the previous one. The published methods are retrained at every width; STACK-JEPA keeps its frozen representation and fits only a new count decoder. PCC_{gene} decreases with width for every method as the panel admits lower-

Table 4: **Effect of graph reconstruction loss** $\mathcal{L}_{\text{graph}}$ on 3D alignment under the PASTE2 subslice protocol (LTARI, mean $\pm$ s.d. over section pairs and folds). Best per column in bold.

Model	ST-data	DLPFC
w/o $\mathcal{L}_{\text{graph}}$	0.151 $\pm$0.025	0.243 $\pm$0.046
STACK-JEPA	**0.165** $\pm$**0.023**	**0.408** $\pm$**0.039**

expressed genes, and STACK-JEPA leads at all eight widths. Its lead over the strongest published method is 1-6% up to 1,000 genes and widens to 14-35% from 2,000 genes on, reaching 0.119 against HoloTea's 0.088 on the full set. From 5,000 genes on, the own-H&E baseline also exceeds all three published methods, while Overlap and ASIGN fall to 0.02.

Distance from profiled sections. Each target is predicted from the nearer of its specimen's two profiled sections, so its distance from that section sets how far STACK-JEPA must extrapolate in depth. In HER2+ specimens A-D, sections are 32 μm apart and the four targets lie 32-128 μm beyond the second profiled section; E-H contribute a single target each and are left out. Figure 3 (middle) scores each target section separately on SVG-250, with PCC_{gene} computed within the section. STACK-JEPA leads HoloTea by 0.023-0.028 up to 96 μm and by 0.013 at 128 μm, and ASIGN by 0.17-0.21 throughout. Up to 96 μm, STACK-JEPA and HoloTea retain at least 93% of their 32 μm PCC_{gene} and ASIGN does not decline. All three drop at 128 μm, which is also the last section of every stack: STACK-JEPA retains 76%, HoloTea 79% and ASIGN 65%. STACK-JEPA shows the same profile on SVG-500 and HEG-250, retaining at least 93% up to 96 μm and 75-76% at 128 μm (Figure 5).

Input modalities. Figure 3 (right) retrains STACK-JEPA on HER2+ with one component removed at a time. *w/o H&E* tokenizes each location by its TERRA embedding alone and *w/o expression* by its UNI2-h embedding alone. Both also drop the H&E conditioning $W_{\text{H\&E}}$ of the mask token, so the first sees no histology and the second no expression. Removing either modality costs 0.037-0.042 PCC_{gene} on the SVG-250 panel. Table 8 lists all three metrics on SVG-250 and repeats the ablation on SVG-500 and HEG-250 with the same result.

Graph loss. The graph loss consistently improves 3D alignment under the PASTE2 subslice protocol (Table 4). Removing $\mathcal{L}_{\text{graph}}$ reduces alignment performance from 0.165 to 0.151 on ST-data and from 0.408 to 0.243 on DLPFC, indicating that explicitly preserving local molecular structure improves cross-section alignment. For the graph reconstruction task, the effect of $\mathcal{L}_{\text{graph}}$ can be read from Table 2 by comparing the *TERRA cosine* and STACK-JEPA rows: because STACK-JEPA takes TERRA embeddings as its molecular input and distils a TERRA-based teacher state, the gap between the two rows (0.618 vs. 0.468 on HER2+, 0.608 vs. 0.444 on ST-data, 0.452 vs. 0.408 on DLPFC) isolates what training on the local similarity graph adds beyond the input representation itself.

5 CONCLUSION

We presented STACK-JEPA, a masked molecular graph-distillation framework that learns molecularly informed representations throughout a serial histology stack. A frozen molecular teacher defines per-location states and their local relations on profiled sections, and a stack student recovers both for a held-out section from one profiled context section, the target H&E, and 3D position. A single frozen representation serves every gene panel, each read out by a lightweight count decoder. On three public serial-section datasets, STACK-JEPA attains the highest PCC_{gene} and PCC_{cell} in all nine dataset-panel settings, exceeding the strongest published method by an average of 15.5% in PCC_{gene} on SVG-500, and on HER2+ it leads at every panel width up to the full transcriptome. The same states preserve local tissue topology and, substituted for expression in partial fused Gromov-Wasserstein transport, raise label-transfer ARI over PASTE2 and register sections whose transcriptome was never measured. Molecular profiling can therefore stay sparse along depth: wherever H&E is imaged, STACK-JEPA can supply a molecular representation that transfers across gene panels and downstream tasks.

AI USAGE CLARIFICATION

Large language models (LLMs) were used as assistive tools in preparing this work. They helped draft and revise the text, check grammar, and organize the manuscript into sections. They were also used to generate code under a controlled workflow in which the authors specified each task and reviewed every change. All AI-generated outputs (text and code) were manually checked for correctness. LLMs were not used to generate research ideas, to design the method, or to interpret results The authors take full responsibility for the content of this paper.

ETHICS STATEMENT

This work complies with the ICLR Code of Ethics. All experiments use publicly available, de-identified human tissue datasets released by their original studies under their respective ethical approvals.

REPRODUCIBILITY STATEMENT

All results reported in this paper are reproducible. The anonymized source code for data preparation, training, and evaluation is provided with the submission. Additionally, the architecture, pretraining, and expression readout hyperparameters are listed in Table 5. All three datasets are publicly available; quality control, gene-panel construction, and the specimen-level cross-validation splits are described in Sec. 3.1 and Table 6, and the full preparation and split details are included with the code.

REFERENCES

Alma Andersson, Ludvig Larsson, Linnea Stenbeck, Fredrik Salmén, Anna Ehinger, Sunny Z Wu, Ghamdan Al-Eryani, Daniel Roden, Alex Swarbrick, Åke Borg, et al. Spatial deconvolution of HER2-positive breast cancer delineates tumor-associated cell type interactions. *Nature communications*, 12(1):6012, 2021.

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15619–15629. IEEE, 2023.

Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, et al. V-JEPA 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

Adrien Bardes, Quentin Garrido, Jean Ponce, Xinlei Chen, Michael Rabbat, Yann LeCun, Mahmoud Assran, and Nicolas Ballas. Revisiting feature prediction for learning visual representations from video. *arXiv preprint arXiv:2404.08471*, 2024.

Sebastian Birk, Mohammad Vali Sanian, Amirhossein Vahidi, Samuel Ogden, Daniyal J Jafree, Adib Miraki Feriz, Carlo Leonardi, Arpit Merchant, Zichen He, Lloyd Steele, et al. Multi-scale modeling of human tissues from spatial transcriptomics with TERRA. *BioRxiv*, pp. 2026–07, 2026.

Dario Bressan, Giorgia Battistoni, and Gregory J Hannon. The dawn of spatial omics. *Science*, 381 (6657):eabq4964, 2023.

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.

Bryan He, Ludvig Bergenstråhle, Linnea Stenbeck, Abubakar Abid, Alma Andersson, Åke Borg, Jonas Maaskola, Joakim Lundeberg, and James Zou. Integrating spatial gene expression and breast tumour morphology via deep learning. *Nature biomedical engineering*, 4(8):827–834, 2020.

Naiwen Hu, Haozhe Cheng, Yifan Xie, Shiqi Li, and Jihua Zhu. 3D-JEPA: A joint embedding predictive architecture for 3D self-supervised representation learning. *arXiv preprint arXiv:2409.15803*, 2024a.

Yunfei Hu, Manfei Xie, Yikang Li, Mingxing Rao, Wenjun Shen, Can Luo, Haoran Qin, Jihoon Baek, and Xin Maizie Zhou. Benchmarking clustering, alignment, and integration methods for spatial transcriptomics. *Genome biology*, 25(1):212, 2024b.

Tinglin Huang, Tianyu Liu, Mehrtash Babadi, Wengong Jin, and Rex Ying. Scalable generation of spatial transcriptomics from histology images via whole-slide flow matching. *arXiv preprint arXiv:2506.05361*, 2025.

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

Bohan Li, Feng Bao, Yuhang Sun, Tianyun Wang, Fengji Li, Hongjue Li, Qionghai Dai, and Yue Deng. Three-dimensional spatial transcriptomics at isotropic resolution enabled by generative deep learning. *bioRxiv*, pp. 2025–08, 2025.

Senlin Lin, Zhikang Wang, Yan Cui, Qi Zou, Chuangyi Han, Rui Yan, Zhidong Yang, Wei Zhang, Rui Gao, Jiangning Song, et al. Bridging the dimensional gap from planar spatial transcriptomics to 3D cell atlases. *Nature Methods*, 23(2):360–372, 2026.

Xinhao Liu, Ron Zeira, and Benjamin J Raphael. Partial alignment of multislice spatially resolved transcriptomics data. *Genome research*, 33(7):1124–1132, 2023.

Romain Lopez, Jeffrey Regier, Michael B Cole, Michael I Jordan, and Nir Yosef. Deep generative modeling for single-cell transcriptomics. *Nature methods*, 15(12):1053–1058, 2018.

Vivien Marx. Method of the year: spatially resolved transcriptomics. *Nature methods*, 18(1):9–14, 2021.

Kristen R Maynard, Leonardo Collado-Torres, Lukas M Weber, Cedric Uytingco, Brianna K Barry, Stephen R Williams, Joseph L Catallini, Matthew N Tran, Zachary Besich, Madhavi Tippani, et al. Transcriptome-scale spatial gene expression in the human dorsolateral prefrontal cortex. *Nature neuroscience*, 24(3):425–436, 2021.

Wenwen Min, Zhiceng Shi, Jun Zhang, Jun Wan, and Changmiao Wang. Multimodal contrastive learning for spatial gene expression prediction using histology images. *Briefings in Bioinformatics*, 25(6):bbae551, 2024.

Lorenzo Mur-Labadia, Matthew Muckley, Amir Bar, Mido Assran, Koustuv Sinha, Mike Rabbat, Yann LeCun, and Nicolas Ballas. V-JEPA 2.1: Unlocking dense features in video self-supervised learning. In *European Conference on Computer Vision*, pp. 671–689. Springer, 2026.

Minxing Pang, Kenong Su, and Mingyao Li. Leveraging information in spatial transcriptomics to predict super-resolution gene expression from histology images in tumors. *BioRxiv*, pp. 2021–11, 2021.

Ayumu Saito, Prachi Kudeshia, and Jiju Poovvancheri. Point-JEPA: A joint embedding predictive architecture for self-supervised learning on point cloud. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 7348–7357. IEEE, 2025.

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

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, 2024.

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.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. RoFormer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

Xin Sui, Jennifer A Lo, Shuchen Luo, Yichun He, Zefang Tang, Zuwan Lin, Dániel L Barabási, Yiming Zhou, Wendy Xueyi Wang, Jia Liu, et al. Scalable spatial single-cell transcriptomics and translatomics in 3D thick tissue blocks. *Nature Methods*, 22(12):2574–2584, 2025.

Valentine Svensson, Sarah A Teichmann, and Oliver Stegle. SpatialDE: identification of spatially variable genes. *Nature methods*, 15(5):343–346, 2018.

Sandra M Swain, Mythili Shastry, and Erika Hamilton. Targeting her2-positive breast cancer: advances and future directions. *Nature reviews Drug discovery*, 22(2):101–126, 2023.

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.

Yunzhi Yan, Tianyi Gu, Chengcheng Sun, Yinghao Zhang, Yan Cui, Senlin Lin, Qi Zou, Yixuan Du, Chuangyi Han, Kairan Kang, et al. Benchmarking alignment methods for spatial transcriptomics data. *Nature Computational Science*, pp. 1–18, 2026.

Tingyang Yu, Chanakya Ekbote, Nikita Morozov, Jiashuo Fan, Pascal Frossard, Stéphane d'Ascoli, and Maria Brbić. Tissue reassembly with generative AI. *bioRxiv*, pp. 2025–02, 2025.

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.

Tianjiao Zhang, Shenghe Li, Hongfei Zhang, Ruolan Zhang, Zhongqian Zhao, Ruihan Wang, Xiaopeng Teng, Long Wan, Yucai Jiang, Jianyi Lyu, et al. SINTER3D: continuous 3D reconstruction of spatial transcriptomics via implicit neural representations. *Genome Biology*, 27(1):274, 2026.

Junchao Zhu, Ruining Deng, Tianyuan Yao, Juming Xiong, Chongyu Qu, Junlin Guo, Siqi Lu, Mengmeng Yin, Yu Wang, Shilin Zhao, et al. ASIGN: an anatomy-aware spatial imputation graphic network for 3D spatial transcriptomics. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 30829–30838. IEEE, 2025a.

Sichen Zhu, Yuchen Zhu, Molei Tao, and Peng Qiu. Diffusion generative modeling for spatially resolved gene expression inference from histology images. In *International Conference on Learning Representations*, volume 2025, pp. 19666–19690, 2025b.

Table 5: Architecture, pretraining, and count-readout settings of STACK-JEPA. Pretraining is early-stopped when applicable.

Architecture	
TERRA genes per location	256
TERRA in-plane neighbors	10
TERRA width	384
UNI2-h H&E tile	112 μm at 224 $\times$ 224 px
UNI2-h width	1536
STACK-JEPA token width d	384
encoder blocks / heads / MLP ratio	12 / 6 / 4
predictor blocks / heads / width	12 / 12 / 384
Pretraining	
context decay τ (Eq. 1)	2
context weight λ	0.5, held at 0 for 5% of updates, linear ramp to 10%
EMA momentum m	0.99925
optimizer	AdamW (β_1=0.9, β_2=0.999)
weight decay	0.04
learning rate / warmup fraction	10^{-4} to 6×10^{-4} / 0.04
batch size (section pairs)	256 (HER2+, ST-data), 64 (DLPFC)
epochs	50
updates	12,500 (250 per epoch)
Count readout	
likelihood	NB
hidden width	128
optimizer / learning rate	AdamW / 10^{-3}
batch size	128
epochs	75

Table 6: Dataset and split statistics. Quality control filters each section at ±5 median absolute deviations (MAD) on log total counts and log genes detected, and at $+3$ MAD on mitochondrial fraction.

	HER2+	ST-data	DLPFC
folds	4	4	3
specimens / sections	8 / 36	16 / 48	3 / 12
section spacing	32 μm (A-D), 16 μm (E-H)	16 μm	10/290 μm
platform	legacy ST	legacy ST	10x Visium
locations before QC	13,620	23,176	47,681
locations after QC	13,467	23,040	45,960
genes measured	11,871	12,279	33,538

A SCORING AND METRIC DEFINITIONS FOR EXPRESSION READOUT TASK (TASK 1)

Shared prediction space. We score every method on expression composition over the G evaluated genes. Count outputs are normalized by their row total; log-normalized outputs $p_g = \log[(x_g + 1)/\sum_j(x_j + 1)]$ are first mapped to counts as $\max\{(S + G)\exp(p_g) - 1, 0\}$. Rows with zero predicted mass remain zero. The resulting panel fraction f is compared with the target after

$$y = \log(1 + Sf), \tag{6}$$

where S is the mean panel total over held-in cells of the fold, shared by every method in a (fold, panel) comparison. S uses no target-section counts, so predicted library sizes, including the fixed 10,000-count decoder library, do not enter the score. This shared scale is essential for fairly comparing MSE across methods as MSE depends on the magnitude of y.

Metric definitions. Let $P, T \in \mathbb{R}^{n \times G}$ be prediction and target in the shared space, and $\widetilde{P}, \widetilde{T}$ their row-centered versions. PCC_{cell} averages over cells

$$\frac{\sum_g \widetilde{P}_{ig} \widetilde{T}_{ig}}{\sqrt{\left(\sum_g \widetilde{P}_{ig}^2\right)\left(\sum_g \widetilde{T}_{ig}^2\right) + 10^{-8}}}, \tag{7}$$

so a constant row scores zero. PCC_{gene} is the column-wise analogue, excluding genes with near-zero variance. MSE averages over all nG entries; R^2 and cosine similarity are computed over the full matrix. Per-specimen PCC_{cell} averages cells within each held-out specimen, then weights specimens equally. Cross-validation averages weight folds equally.

B Scoring and metric definitions for 2D graph reconstruction task (Task 2)

Ground truth. For each held-out section we build the spatial κ-nearest-neighbour graph on the measured spot coordinates ($\kappa = 6$), symmetrise it, and remove every edge longer than $R = 2.5$ pitches. The radius cap prevents edges from bridging tissue holes or regions removed by quality control, where κ-NN alone would connect spots that are not tissue neighbours. The resulting undirected edge set E^k is the positive class.

Candidate set. Scoring all $\binom{n_k}{2}$ pairs would let global proximity dominate: almost every distant pair is a trivial negative. We therefore score a local candidate set. For every true edge $(i, j) \in E^k$ we draw five non-edges uniformly from the pairs $(i, l) \notin E^k$ with $\|c_i - c_l\| \leq 2.5$ pitches, giving a positive rate of $1/6$ and a random-ranking AUPRC near that value (the *Random* row of Table 2; it sits slightly above $1/6$ because sections near tissue edges have fewer eligible negatives). The candidate set is sampled once per section with a fixed seed and reused for every method.

Scores. A representation u scores a candidate pair by $\cos(\tilde{u}_i, \tilde{u}_j)$, where $\tilde{u}$ is u centred by its section mean, the same centring applied to the training target s_{ij} of Eq. 3. This holds for measured expression (log-normalised counts of the SVG-500 panel), TERRA, UNI2-h and the STACK-JEPA backbone. For the graph head we use $\hat{s}_{ij}$ directly. LUNA (Yu et al., 2025) outputs coordinates; we fit an affine map from predicted to true coordinates on the section and score pairs by negative Euclidean distance in the fitted frame.

Regimes. In the *unobserved* regime the target transcriptome is withheld; STACK-JEPA predicts the section from its nearest held-in section with the target H&E and coordinates, exactly as in Task 1, and the only baseline with the same inputs is UNI2-h cosine on the target tiles. In the *observed* regime the target is measured, STACK-JEPA uses the EMA teacher encoding of the full section pair, and expression-based baselines apply.

Metric. Edge AUPRC is the area under the precision–recall curve of the candidate ranking against E^k, computed per section, averaged over the target sections of a fold, then over folds; the reported $\pm$ is the fold standard deviation.

C Scoring and metric definitions for 3D alignment task (Task 3)

We evaluate alignment following the PASTE2 protocol (Liu et al., 2023), which constructs ground-truth correspondence without manual landmarks. We evaluate alignment on ST-data and DLPFC, the only two of the three datasets that contain annotations for multiple sections within a fold. The DLPFC dataset (Maynard et al., 2021) contains four Visium sections (A-D) from each of three donors, with A-B and C-D separated by $10\,\mu\text{m}$ and B-C by $300\,\mu\text{m}$. We used cortical layers annotations for DLPFC. For ST-data, we use tumour labels. Following PASTE2, we crop complementary subslices with approximately 70% overlap in both horizontal and vertical directions, making the true correspondence known by construction. We follow the same evaluation convention used in subsequent ST alignment benchmarks (Hu et al., 2024b; Yan et al., 2026) with s selected automatically rather than provided directly to the alignment method.

Table 7: Comparison on HEG-250, the 250 most highly expressed genes. Protocol, scoring and ± as in Table 1. Best per column in bold.

	HER2+			ST-data			DLPFC		
Method	PCC_{gene}↑	PCC_{cell}↑	MSE↓	PCC_{gene}↑	PCC_{cell}↑	MSE↓	PCC_{gene}↑	PCC_{cell}↑	MSE↓
Overlap	0.206 ±0.134	0.597 ±0.094	0.807	0.314 ±0.044	0.648 ±0.023	0.623	0.054 ±0.051	0.628 ±0.071	0.574
UNI (Chen et al., 2024)	0.308 ±0.128	0.678 ±0.081	0.631	0.380 ±0.036	0.725 ±0.024	0.499	0.181 ±0.057	0.769 ±0.044	0.378
ASIGN (Zhu et al., 2025a)	0.261 ±0.180	0.687 ±0.065	0.685	0.445 ±0.063	0.751 ±0.027	0.480	0.046 ±0.045	0.760 ±0.048	0.375
STFlow (Huang et al., 2025)	0.354 ±0.124	0.715 ±0.066	0.584	0.438 ±0.041	0.752 ±0.018	0.450	0.147 ±0.090	0.756 ±0.045	0.386
HoloTea (Sanian et al., 2025)	0.367 ±0.111	0.715 ±0.064	0.575	0.465 ±0.027	0.768 ±0.015	**0.423**	0.133 ±0.113	0.754 ±0.051	0.393
STACK-JEPA (ours)	**0.370** ±0.124	**0.736** ±0.072	**0.564**	**0.471** ±0.027	**0.775** ±0.017	0.425	**0.194** ±0.053	**0.773** ±0.044	**0.365**

Table 8: Component ablations on HER2+ on the three gene panels; Figure 3 (right) plots the SVG-250 PCC_{gene} per fold. Each variant is retrained with one component of STACK-JEPA removed and decoded and scored as in Table 1; means over the four folds, with ± the fold standard deviation on the two correlations. Rows are ordered by PCC_{cell} within a panel. Best per column within a panel in bold.

Panel	Variant	PCC_{gene}↑	PCC_{cell}↑	MSE↓
SVG-250	w/o H&E (TERRA only)	0.369 ±0.118	0.766 ±0.062	0.502
	w/o expression (UNI2-h only)	0.364 ±0.117	0.767 ±0.063	0.497
	STACK-JEPA (full)	**0.406** ±0.096	**0.777** ±0.060	**0.479**
SVG-500	w/o H&E (TERRA only)	0.302 ±0.096	0.725 ±0.065	0.456
	w/o expression (UNI2-h only)	0.301 ±0.095	0.726 ±0.067	0.456
	STACK-JEPA (full)	**0.332** ±0.081	**0.735** ±0.064	**0.443**
HEG-250	w/o H&E (TERRA only)	0.349 ±0.141	0.723 ±0.074	0.577
	w/o expression (UNI2-h only)	0.347 ±0.142	0.725 ±0.076	0.573
	STACK-JEPA (full)	**0.370** ±0.124	**0.736** ±0.072	**0.564**

Alignment is evaluated by Label Transfer ARI (LTARI). Each location in the second section inherits the cortical-layer label of its most probable counterpart in the first, and the transferred labels are compared with the manual annotations using adjusted Rand index. We report LTARI for each adjacent pair (AB, BC, CD) and donor under two regimes. In the *profiled* regime, both sections have measured ST uses STACK-JEPAdescriptors, with α fixed to the PASTE2 setting. In the *imputed* regime, ST in the second section is withheld and its representation is imputed from the available inputs; expression-based baselines are therefore not applicable.

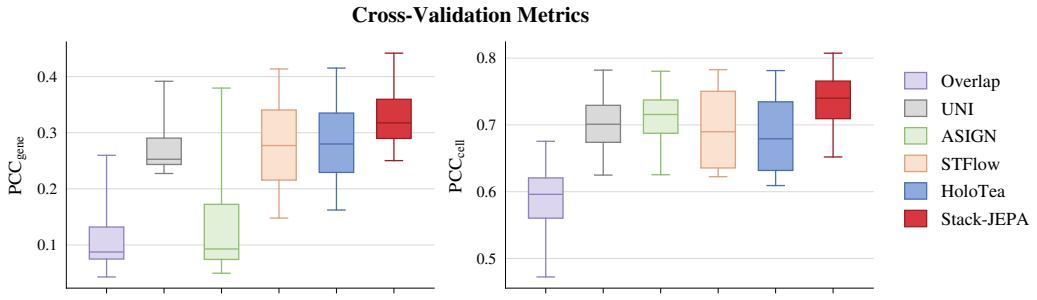


Figure 4: Cross-fold variability of each method on HER2+ with the SVG-500 panel. Each box shows how one method's PCC_{gene} (left) or PCC_{cell} (right) varies across the four cross-validation folds of Table 1, each of which holds out different specimens: the line marks the median fold, the box the interquartile range, and the whiskers the lowest and highest fold. A narrower spread means the method's accuracy depends less on which specimens are held out.

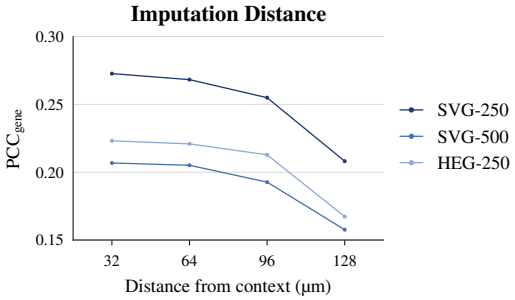


Figure 5: STACK-JEPA PCC_{gene} on HER2+ on SVG-250, SVG-500 and HEG-250, on the target sections of specimens A-D by distance from the nearer profiled section, averaged over the four specimens.