sMCC40M: A Context-Aware Dataset and Benchmark for Single-Cell Spatial Transcriptomics

Xi Li[1] **Yaqi Hu**[1] **Ziheng Duan**[1] **Xinyi Wang**[1]
Simon Dongbo Sun[1] **Diptanshu Sikdar**[2] **Yang Liu**[3] **Jing Zhang**[1]

[1] Donald Bren School of Information and Computer Sciences
University of California, Irvine, Irvine, CA, USA

[2] Department of Computer Science and Engineering
University of California San Diego, La Jolla, CA, USA

[3] Department of Pathology, Yale School of Medicine
New Haven, CT, USA

Abstract

Spatial transcriptomics enables testing whether tissue morphology can predict molecular state, but existing benchmarks are constrained by limited scale, spot-level supervision, and incomplete biological context. We present sMCC40M, a multimodal resource comprising 38.53 million gene-positive cell/nucleus observations from 145 samples across 23 tissue categories and seven species. Visium HD contributes 13.66 million observations and Xenium 24.81 million, complemented by 53,989 profiles from 21 in-house samples. H&E-pairing eligibility is recorded separately. Three data-centric axes assess generalization beyond local interpolation. First, sMCC40M provides scale: broad organ and study coverage enables controlled encoder benchmarking and reveals a scaling trend in which larger pathology foundation models improve morphology–gene correspondence under matched evaluation. Second, sMCC40M provides resolution: by decomposing assays into aligned cell-level records, our framework converts spot-level histology–omics pipelines into single-cell predictors and evaluates them under strict in-domain and cross-acquisition splits. Third, sMCC40M provides rich context: each cell is paired with spatial, molecular, and sample-level metadata, allowing analyses such as age-band shift in ovarian samples, where age-mismatched transfer sharply reduces prediction quality despite misleading global-distance summaries. Together, sMCC40M establishes a practical framework for single-cell histology-to-gene prediction while showing that robust generalization still depends on scale, cellular resolution, and explicit biological context.

1 Introduction

Spatial transcriptomics (ST) technologies measure gene expression while preserving tissue coordinates, making it possible to ask whether histology can recover molecular state in situ. H&E images provide a high-resolution structural view of the same tissue, capturing cellular morphology, tissue architecture, and microenvironmental organization. The central modeling question is therefore no longer whether histology contains molecular signal in principle, but whether current datasets and benchmarks can measure that signal under realistic biological variation.

Existing histology–omics resources have substantially advanced the field, especially HEST-1K and STimage-1K4M for paired H&E–ST evaluation [1, 2]. Their remaining limitations are best understood along three axes. First, scale is still limited relative to the diversity of organs, patients, studies, and acquisition conditions encountered in real tissue data. Second, most benchmarks remain

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

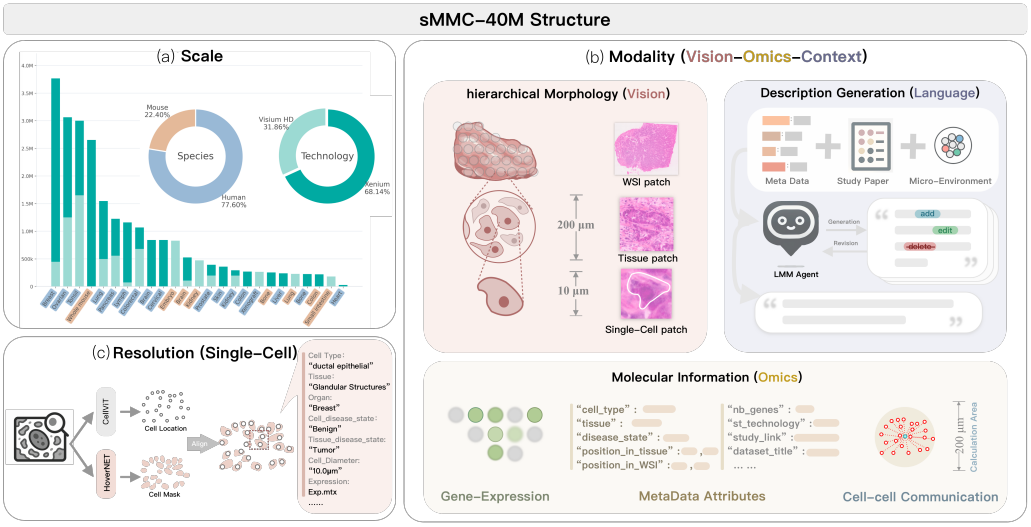


Figure 1: Overview of sMCC40M. The frozen resource contains 38.53 million gene-positive cell/nucleus observations across 145 samples. The diagram illustrates cell-aligned molecular and image context; image pairing has explicit eligibility flags, and auxiliary annotation coverage is specified per artifact.

spot-based, so each target averages multiple cells and can hide cell-level heterogeneity. Third, evaluating biological generalization requires consistent annotations and explicit shift tests across age, disease state, neighborhood organization, and cell–cell communication. Strong benchmark scores may otherwise reflect within-cohort interpolation rather than biologically meaningful transfer.

We introduce sMCC40M as a context-aware resource designed to boost ST research along three axes: **scale**, **resolution**, and **rich context**. The frozen release contains 38.53 million gene-positive observations from 53 Visium HD and 71 Xenium acquisitions across 23 harmonized tissue categories. It distinguishes segmented cells from nuclei and preserves each observation's spatial origin and measured gene panel. For histology modeling, 32.68 million cell observations meet the H&E-pairing criteria. Sample-level biological attributes and their source evidence support context-aware analyses, with missing or conflicting information explicitly recorded.

We therefore organize the experimental section around three claims. **Scale** is evaluated through encoder benchmarking and scaling-law behavior. **Resolution** is evaluated through within-acquisition spatial holdouts and cross-acquisition image-to-cell expression prediction, with matched test cells and genes where comparisons are made. **Rich context** is evaluated through a new ovarian age-effect analysis showing that sample-level biological context can dominate apparent generalization: an age-mismatched target can cause large drops in F1, gene-wise correlation, cell-wise correlation, and predicted nonzero rate, while global distribution distances alone remain misleading. Quality audits, complete current benchmark results, downstream cell-type and context analyses, caption diagnostics, and release details are reported in the appendix.

sMCC40M makes three contributions. First, it provides a large ST resource with identified H&E–expression pairs across diverse tissues, species, and two assay platforms. Second, it defines a data-centric framework for lifting spot-level image-to-gene prediction into the single-cell regime by replacing spot crops and spot-averaged targets with cell-centered histology and cell-level molecular targets. Third, it shows empirically that benchmark performance depends not only on model architecture, but also on the scale, resolution, and biological context made available by the dataset.

2 Related Resources and Benchmarks

Histology–Omics Resources. sMCC40M builds on three strands of prior work. First, spatial profiling evolved from Spatial Transcriptomics [3], in situ sequencing [4], MERFISH [5], Slide-seq [6], and multiplexed imaging [7, 8] to higher-resolution or higher-throughput platforms such as HDST [9], Slide-seqV2 [10], DBiT-seq [11], Visium [12], high-definition commercial systems [13,

Table 1: Benchmark-oriented comparison with representative resources from four categories: single-cell-only atlases, pathology image–text corpora, spot-level histology–omics resources, and sMCC40M. Entries are summarized using *Yes*, *No*, or *Partial*. "Platform transfer" indicates whether the resource natively supports evaluation across distinct acquisition technologies or assay families.

Dataset	Cell-level	Multi-organ	Study split	Platform transfer	Text	Benchmark intent
Single-cell-only atlases						
CELLxGENE Census	Yes	Yes	Yes	No	No	Single-cell atlas (no paired histology)
Pathology image–text corpora						
OpenPath	No	Partial	No	No	Yes	Pathology image–text pretraining
Quilt-1M	No	Partial	No	No	Yes	Patch-level image–text pretraining
PathGen-1.6M	No	Partial	No	No	Yes	Synthetic pathology image–text pretraining
ST + histology resources						
HEST-1K	No	Yes	Yes	Partial	No	Spot-level ST + histology benchmarking
STimage-1K4M	No	Yes	Partial	Partial	No	Spot-level image–gene pairing from ST slides
sMCC40M	**Yes**	**Yes**	**Partial**	**Yes**	**Partial**	**Single-cell evaluation**

14], and spatiotemporal profiling [15]. Second, spatial analysis and integration tools, including Giotto [16], Squidpy [17], SpaGCN [18], BayesSpace [19], Tangram [20], cell2location [21], related spatial integration and cell-type mapping methods [22, 23, 24, 25], Impeller [26], iMIRACLE [27], DISCO [28], MUSE [29], Seurat integration [30, 31], Scanpy [32], together with reference initiatives including the Human Cell Atlas [33], Tabula Sapiens [34], and the Human Tumor Atlas Network [35], provide the computational and single-cell reference backbone. Third, benchmark resources such as HEST-1K [1] and STimage-1K4M [2] connect paired histology–omics data with evaluation; sMCC40M extends this direction with cell-level alignment, multi-organ coverage, and transfer-aware splits, as summarized in Table 1.

Histology-to-Gene Prediction. HE2RNA showed bulk transcriptomic prediction from digital pathology [36], while Hist2ST [37], BLEEP [38], iSTAR [39], and THItoGene [40] established spot- or region-level histology-to-gene prediction. These methods are supported by a broader pathology modeling stack: weakly supervised WSI and MIL-style learning [41, 42, 43, 44], self-supervised histopathology representation learning [45], biomedical and nuclear segmentation [46, 47, 48], and studies and reviews on diagnostic and molecular pathology [49, 50, 51, 52]. Our benchmark differs by targeting cell-aligned image-to-gene prediction rather than bulk, slide-level, or spot-averaged prediction, under controlled in-domain and transfer splits.

Pathology Foundation Models and Multimodal Learning. Pathology and biomedical models and resources include QuiltNet [53], CTransPath [54], UNI and UNI2-h [55, 56], Phikon-v2 [57], GigaPath [58], Virchow [59], Virchow2 [60], H-optimus-0 [61], CONCH [62], PLIP [63], PathCLIP [64], BiomedCLIP [65], and PathGen-style corpora [66]. Related work includes robust medical-image pretraining [67], visual corpora [68, 69], CNN and transformer backbones [70, 71, 72, 73, 74, 75, 76, 77], self-supervised and masked objectives [78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88], multimodal and language models [89, 90, 91, 92], and foundation-model analyses [93]. WILDS [94] provides evaluation resources for distribution shifts. Complementary resources include PanNuke nucleus annotations [95] and TCGA cancer cohorts [96]. Our A/B samples come from 10x Visium HD lung-cancer resources [97, 98]. Text annotations provide auxiliary context for future multimodal evaluation; gene expression remains the prediction target.

3 The sMCC40M Resource

3.1 Design Objectives

sMCC40M has three design requirements: retain cell- or nucleus-aligned molecular measurements; preserve their spatial and image context; and make source identity, biological context, and technical eligibility auditable. The release distinguishes acquisition-level diversity from verified donor independence and model-derived segmentation from biological ground truth. See Figure 1.

Table 2: **Spatial measurements and expression profiles by data source.**

Source	How expression is localized	Samples	Profiles
Visium HD	2 μm bins assigned to segmented cell or nuclear masks	53	13,660,756
Xenium	Transcripts assigned to cells, with nuclear boundaries where available	71	24,811,436
In-house data	Single-cell RNA profiles paired with cell morphology	21	53,989
Total		**145**	**38,526,181**

We add in-house samples as our lab completes new experiments.

3.2 Scale, Coverage, and Sources

Scale. The complete resource contains **38,526,181 expression profiles from 145 samples**, comprising 53 Visium HD, 71 Xenium, and 21 in-house samples. The public Visium HD and Xenium collection contains **38,472,192 gene-positive observations**: **13,660,756 from Visium HD** and **24,811,436 from Xenium** (Table 2). Its 124 acquisitions cover seven species and 23 reporting categories, not all anatomical organs. Acquisitions are sections, capture areas, or ROIs, not independent donors or studies. About a third of our acquisitions appear in the HEST-1k publication cohort [1] or STimage-1K4M [2] (43 of 124, Table S1). Most matches appear in HEST (42 in HEST, 4 in STimage, and 3 in both).

Resolution. Each molecular observation is anchored to a segmented cell, a nucleus, or a platform-defined cell and retains its original coordinates and measured feature axis. Native 2 μm Visium HD bins are aggregated to these footprints; Xenium retains native cell assignments. HD H&E observations use cell masks, while the four HD IF acquisitions use nuclear masks. Xenium DAPI images provide nuclear image context for *cell*-level expression and are not reclassified as nuclear RNA observations.

Rich modality and context information. Records link raw expression, coordinates, image and mask references, and sample metadata with upstream URLs and field-level evidence. Tissue categories are harmonized, while diagnosis, age, sex, and other clinical attributes retain their availability and conflict status. Such attributes describe the sampled material, not every cell's identity or malignant state. HD cell-type predictions include evidence levels and unresolved labels; Xenium annotations retain their original status. Captions and communication summaries are auxiliary analyses, not uniformly available annotations across the frozen resource.

The 21 in-house samples span cell lines of mouse embryonic, cervical, and lung origin. Paired morphology and RNA profiles link cell shape to expression in each of the 53,989 cells.

The Visium HD and Xenium collection retains 46,284,626 native or segmented candidate objects. Of these, 38,486,654 have nonzero stored features, including 14,462 Xenium observations with only non-gene features. The headline count excludes these rows and measured-zero or RNA-unavailable candidates; all remain distinguishable in the metadata. Xenium panels and HD gene axes are preserved separately, without filling unmeasured genes with zeros.

Image eligibility is assessed separately. There are **32,682,585 H&E-paired cell observations** with positive gene counts, an in-image centroid, and a visible mask; **32,682,304** also support a complete native 224-pixel crop. Flags for 150/224/256-pixel and 50/100 μm windows distinguish boundary truncation from absent pairing. These checks establish technical availability, not segmentation accuracy. Release-wide QC is documented in Section B.4.

3.3 Cell-Aligned Multimodal Structure

Each record uses a stable observation identifier to link expression, coordinates, segmentation labels, and acquisition metadata. Full-resolution images and masks retain their sample-specific physical pixel sizes; cell-centered and larger-context crops can be extracted on demand at declared pixel or physical scales. Per-observation flags distinguish H&E, IF, and DAPI context, cell versus nuclear units, expression availability, and crop completeness.

This design distinguishes sMCC40M from spot-level histology–omics resources. Instead of asking whether a model can recover expression averaged over a mixed local region, the benchmark asks

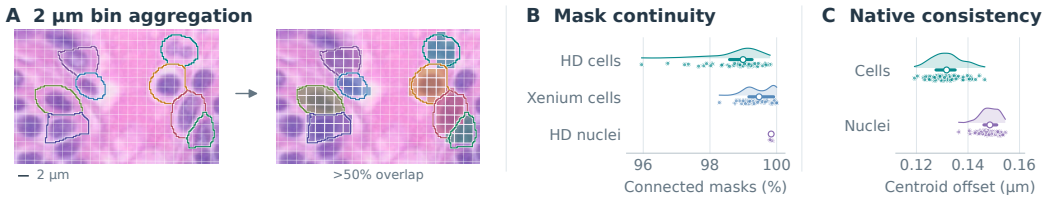


Figure 2: **Cell aggregation and technical quality control.** (A) HD-010 lung H&E with cell masks and native 2 μm bin assignments; scale bar, 2 μm. (B) Single-component masks: 49 HD cell, 71 Xenium cell, and four HD nuclear acquisitions. (C) Native polygon–summary centroid offsets for cells and nucleus references in 71 Xenium acquisitions. Dots summarize acquisitions; circles and thick segments denote medians and interquartile ranges. Half-densities require at least ten acquisitions.

whether morphology and local tissue context are sufficient to predict fine-grained molecular variation for individual cells. The same cell-aligned schema provides a common input–target format for image-to-expression prediction.

Comparing a gene across Visium HD and Xenium first requires checking that both acquisitions measured it. A gene absent from the Xenium panel has no observed value, whereas a measured gene can have a count of zero. For image-based prediction, each RNA target needs a usable histology crop. The sample gene lists and crop eligibility flags let users apply both conditions before training.

3.4 Alignment Pipeline

Native filtered 2 μm Visium HD integer UMI counts are mapped into the segmentation frame using recorded transforms. Each bin is assigned at most once, only when a cell or nuclear footprint covers **strictly more than half of the complete bin area** in physical capture-grid coordinates; assigned counts are summed (Figure 2A). Holes and disconnected components are preserved. Unassigned bins and RNA-unavailable candidates retain explicit status. Xenium retains native cell counts and maps boundaries to the selected image. Expression is retained when image or crop checks fail.

Quality control. The release-wide audit covers 124 acquisitions, checking count matrices, observation and feature indices, coordinate transforms, delivered masks, and stored HD bin assignments. Distances are expressed in μm and areas in μm^2, with image coordinates converted using per-axis pixel sizes. Median acquisition-level fractions of labels forming a single four-connected component are high (98.99% for HD cells, 99.47% for Xenium cells, and 99.84% for HD nuclei; Figure 2B). Across 71 Xenium acquisitions, median acquisition-level centroid offsets between native polygons and their summaries are submicrometer (0.132 μm for cells and 0.148 μm for nucleus references; Figure 2C). These diagnostics quantify label connectivity and internal geometric consistency; they do not establish biological boundary accuracy or independent histology-registration accuracy. Image-pairing eligibility remains a separate check. Full metrics, units, and audit provenance appear in Appendix B.4.

The experiments below evaluate encoder scaling, cell-aligned prediction within and across acquisitions, and context-associated variation in transfer performance.

4 How sMCC40M Boosts Existing ST Research

4.1 Scale: Large Coverage Enables Encoder Scaling Laws

The expanded release spans 38.53 million cell/nucleus observations across 23 harmonized reporting categories. The encoder analyses below retain their original cohorts and protocols, distinct from the expanded acquisition inventory in Appendix B and Table S2. Resource scale and the observations evaluated in each experiment are therefore reported separately.

Motivated by frozen-encoder benchmarking in HEST-1K [1], we evaluated seven histopathology encoders under a unified frozen-embedding retrieval and prediction framework on two representative tasks: mouse brain and human lung cancer. As shown in Figure 3a, pathology foundation models consistently outperformed generic backbones. H-Optimus-0 achieved the highest average Pearson correlation on both datasets (0.567 for mouse brain and 0.449 for human lung cancer). For mouse

a

Encoder	Mouse brain			Human lung cancer		
	Avg. Pearson	Sensitivity	Accuracy	Avg. Pearson	Sensitivity	Accuracy
ResNet50	0.4994 ± 0.0675	0.7927 ± 0.0140	0.6847 ± 0.0119	0.3629 ± 0.0860	0.6376 ± 0.0057	0.6222 ± 0.0004
CTransPath	0.5058 ± 0.0611	0.7933 ± 0.0127	0.6865 ± 0.0097	0.3157 ± 0.1332	0.6287 ± 0.0146	0.6166 ± 0.0060
Phikon	0.4961 ± 0.0708	0.7918 ± 0.0149	0.6828 ± 0.0138	0.3690 ± 0.0799	0.6351 ± 0.0082	0.6187 ± 0.0039
CONCH	0.5145 ± 0.0524	0.7994 ± 0.0073	0.6905 ± 0.0061	0.3784 ± 0.0705	0.6361 ± 0.0072	0.6177 ± 0.0049
UNI	0.5545 ± 0.0124	**0.8067** ± 0.0000	**0.6966** ± 0.0000	0.4352 ± 0.0137	0.6425 ± 0.0008	0.6223 ± 0.0003
UNI2-h	0.5549 ± 0.0120	0.8051 ± 0.0016	0.6953 ± 0.0013	0.4360 ± 0.0129	0.6416 ± 0.0017	0.6218 ± 0.0008
H-Optimus-0	**0.5689** ± 0.0000	0.8060 ± 0.0007	0.6962 ± 0.0004	**0.4489** ± 0.0000	**0.6433** ± 0.0000	**0.6226** ± 0.0000

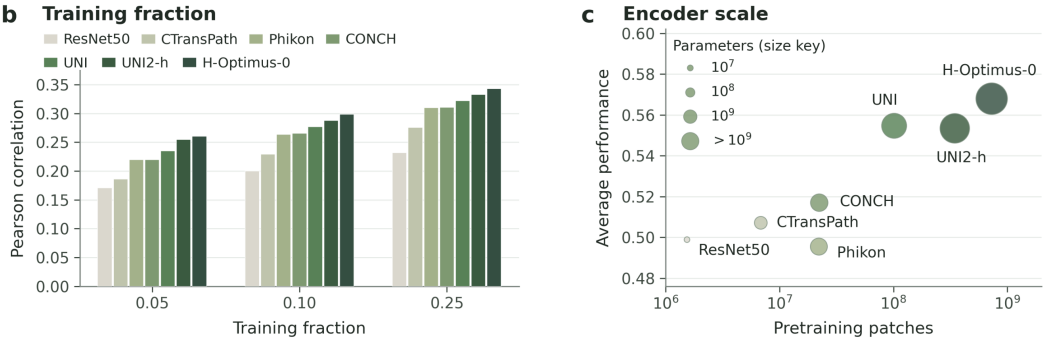


Figure 3: **Encoder benchmarking and scaling behavior on sMCC40M.** (a) Seven pathology encoders are evaluated on mouse brain and human lung cancer; H-Optimus-0 leads average Pearson, while UNI leads mouse-brain sensitivity and accuracy. (b) In-patient prediction of the top 50 variable genes improves with training size, with H-Optimus-0 strongest across ratios. (c) Scaling analysis shows that larger models and broader pretraining improve performance, with UNI, UNI2-h, and H-Optimus-0 forming the leading tier. Panel (a) preserves the printed values; panels (b–c) are approximate digitizations of the original figure, including bubble areas.

brain, UNI obtained the strongest sensitivity and accuracy (0.807 and 0.697), while for human lung cancer, H-Optimus-0 achieved the best sensitivity and accuracy (0.643 and 0.623).

We next analyzed within-patient prediction using frozen encoder embeddings and the top 50 highly variable genes. As shown in Figure 3b, performance increased with the training ratio (0.05 to 0.25), indicating that additional in-domain training samples improve morphology–gene prediction across encoders. H-Optimus-0, UNI2-h, and UNI formed the strongest group. These results suggest that large pathology-pretrained encoders provide informative morphology representations for spatial gene prediction under matched preprocessing and evaluation.

Finally, Figure 3c summarizes an empirical encoder-scale analysis. Performance generally improved with larger model capacity and broader pathology pretraining, with ResNet50, CTransPath, and Phikon forming a lower-performing tier, CONCH showing intermediate performance, and UNI, UNI2-h, and H-Optimus-0 defining the leading group. Overall, these analyses support the utility of frozen pathology foundation models for single-cell morphology–gene benchmarking.

4.2 Resolution: From Spot-Level Prediction to Single-Cell Prediction

Spot averages can obscure differences between neighboring cells. We predict expression from cell-centered H&E crops using 2 μm majority-area aggregation for Visium HD and native cell assignments for Xenium. Across 9,000 CellViT polygons, a 1 μm shift preserved relative gene expression among supported polygons (cosine similarity 0.936–0.994). Total UMI counts were more sensitive (median absolute change 6.0–13.7%), so boundary placement affects the magnitude of the prediction target (Appendix B.4.1).

Single-cell gene expression prediction benchmark

a In-domain benchmark

Category	Gene P	Coord P	KNN P	Gene S	Cell P	F1	R^2	RMSE
Brain	0.505	0.076	0.072	0.492	0.471	0.617	0.238	0.575
Breast	0.395	0.074	0.057	0.378	0.613	0.503	0.160	0.605
Colon	0.424	0.110	0.116	0.400	0.526	0.472	0.042	0.505
Embryo	0.320	0.041	0.060	0.273	0.519	0.418	−0.303	0.496
Head	0.088	−0.012	−0.001	0.057	0.131	0.023	−1.891	0.138
Heart	0.372	0.111	0.081	0.312	0.446	0.287	0.097	0.403
Kidney	0.368	0.061	0.045	0.347	0.545	0.374	0.127	0.461
Liver	0.477	0.008	0.023	0.466	0.627	0.781	0.226	0.622
Lung	0.432	0.009	0.058	0.370	0.497	0.479	0.140	0.494
Lymph Node	0.151	0.015	0.018	0.140	0.334	0.036	−0.001	0.232
Ovary	0.393	0.021	0.036	0.385	0.550	0.517	0.130	0.524
Pancreas	0.403	0.101	0.092	0.368	0.724	0.528	0.117	0.544
Arabidopsis	0.388	−0.033	−0.013	0.376	0.233	0.519	−1.051	0.716
Soybean	0.142	0.057	0.033	0.128	0.251	0.026	−0.133	0.218
Prostate	0.260	0.033	0.065	0.236	0.292	0.090	−0.174	0.257
Small Intestine	0.509	0.047	0.024	0.481	0.537	0.622	0.246	0.526
Tonsil	0.282	0.023	0.040	0.252	0.510	0.382	0.047	0.530
Xenograft	0.655	0.171	0.198	0.538	0.196	0.731	0.364	0.402
Bone†	0.417	0.058	0.089	0.380	0.524	0.480	0.157	0.538
Bone Marrow†	0.116	0.009	0.017	0.101	0.249	0.022	−0.012	0.221
Cervix†	0.308	0.042	0.041	0.268	0.368	0.231	0.072	0.329
Skin†	0.402	0.060	0.102	0.347	0.476	0.464	−0.460	0.515
Whole Organism†	0.274	0.038	0.056	0.218	0.477	0.249	−0.517	0.393
HD18 mean	0.365	0.051	0.056	0.333	0.445	0.411	−0.090	0.458
Xenium5 mean	0.303	0.042	0.061	0.263	0.419	0.289	−0.152	0.399
23-category mean	**0.351**	**0.049**	**0.057**	**0.318**	**0.439**	**0.385**	**−0.103**	**0.445**

b Cross-sample benchmark

Category	Whole acq.	Matched within	Matched cross	Δ
Brain (M)	0.462	0.532	0.448	−0.084
Breast (H)	0.252	0.365	0.246	−0.119
Colon (H)	0.317	0.379	0.302	−0.077
Embryo (M)	0.226	0.303	0.201	−0.103
Kidney (M)	0.233	0.368	0.209	−0.159
Lung (H)‡	0.424	0.441	0.410	−0.030
Lymph Node (H)	0.064	0.119	0.051	−0.068
Ovary (H)‡	0.343	0.402	0.340	−0.062
Pancreas (H)	0.234	0.355	0.243	−0.112
Tonsil (H)	0.048	0.277	0.032	−0.245
10-stratum mean	**0.260**	**0.354**	**0.248**	**−0.106**

c Matched within- vs cross-sample

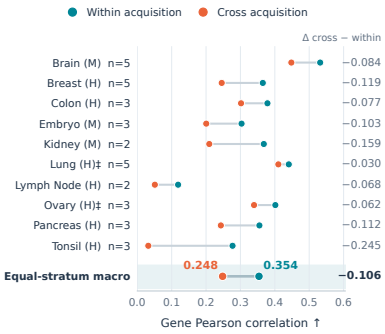


Figure 4: **Single-cell gene expression prediction benchmark. (a)** Spatial holdouts span 23 categories and 61 acquisitions. CONCH–ridge metrics include gene-wise Pearson/Spearman (Gene P/S), cell-wise Pearson (Cell P), and RMSE on $\log(1 + x)$ counts. Coord P and KNN P report gene-wise Pearson for spatial controls. † marks Xenium-only rows. **(b)** HD transfer spans 34 targets in ten tissue/species strata (H/M: human/mouse). Whole-target tests use all sampled cells. Matched fits use the same edge-test cells and externally selected genes. Both report gene-wise Pearson. **(c)** Δ is cross minus within before rounding; n counts acquisitions. Training sizes differ. Six fits in the ‡-marked lung and ovary strata involve related material, limiting independent-donor interpretation. Appendix C gives full results and available HD conditional intervals.

4.2.1 Single-cell gene expression prediction benchmark

Task and cohort. We predict cell-aligned log1p gene counts from cell-centered H&E crops, within and across acquisitions. The benchmark covers 720,433 sampled cells from 61 acquisitions: 48 Visium HD acquisitions across 18 categories and 13 Xenium acquisitions in five additional categories (Bone, Bone Marrow, Cervix, Skin, and Whole Organism), marked *Xenium data only*. These 23 reporting categories include non-organ specimens. Each acquisition contributes at most 12,000 cells; this cohort is distinct from the full release. HD targets use 2 μm majority-area aggregation, while Xenium retains native cell assignments.

Protocol and predictors. Four directional spatial holdouts each reserve 20% of cells for testing, with a 5%-axis-span training buffer; test regions can overlap. Fifty HVGs are selected from training expression over measured-gene intersections, with panel-specific selection for Skin. Frozen CONCH v1 [62] encodes 150-by-150 native-pixel crops through its 448-pixel transform. The 512-dimensional features undergo standardization, 128-component PCA, and ridge regression ($\alpha = 10$). Fitting uses only training cells; expression is not an inference input. Controls use quadratic coordinate–ridge, eight-neighbor spatial KNN, and training-mean expression. Metrics include gene Pearson/Spearman, cell Pearson, detection F1, R^2, and RMSE. Fold scores are averaged within acquisitions, acquisitions within categories, and categories equally. Details and panel exceptions appear in Appendix C.

Within-acquisition results. CONCH–ridge exceeds both spatial controls in gene Pearson across all 23 categories (macro: 0.351 versus 0.049 for coordinate regression and 0.057 for KNN; Figure 4a).

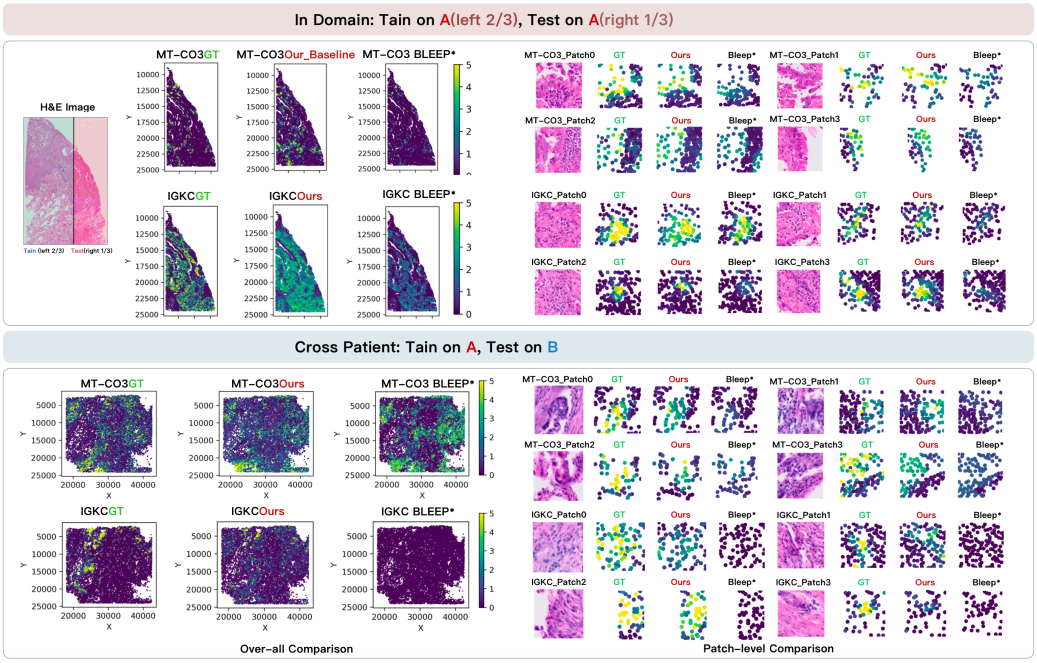


Figure 5: **Qualitative expression-prediction visualization.** The figure presents in-domain prediction in the held-out right third of A (top) and cross-sample prediction from A to B (bottom), with whole-region overviews on the left and corresponding H&E patches and cell-level expression maps on the right. MT-CO3 and IGKC are shown.

However, correlation does not ensure calibrated expression (macro $R^2 = -0.103$). Because platforms, gene panels, species, and target construction differ, this combined macro summarizes coverage rather than assay superiority.

Cross-sample results. Leave-one-acquisition-out evaluation covers 34 HD targets in ten tissue/species strata, with HVGs and all preprocessing fitted using other acquisitions. Whole-acquisition gene Pearson is positive on average (0.260; Figure 4b); this differs in cohort and targets from the 23-category score. A matched comparison instead uses identical edge-test cells and externally selected genes for within- and cross-acquisition prediction (Figure 4b,c). Gene Pearson decreases in all ten strata (macro: 0.354 to 0.248); cell Pearson falls (0.563 to 0.349) and RMSE rises (0.480 to 0.676). The matched decline shows weaker transfer with test cells and genes held fixed. Training-set sizes also differ, so acquisition shift is not the only changing factor. Negative within-acquisition macro R^2 indicates that relative expression patterns can be recovered without accurate expression magnitudes. The training-mean control achieves higher cross-acquisition cell Pearson (0.365), motivating joint assessment of gene-wise variation, calibration, and simple reference predictors. Full results appear in Appendix C.2.

Scope and uncertainty. Acquisitions are not verified independent donors: six holdout fits contain known related material (four lung and two ovary). Excluding the four linked lung acquisitions together reduces lung gene Pearson (0.424 to 0.228). Supported HD spatial-block intervals are conditional on observed samples and fixed predictions, not new-donor generalization; no new intervals are claimed for the Xenium-only rows or combined macro. Exact cell IDs, genes, splits, and software are retained in the reproducibility package; uncertainty assumptions and cohort details are documented in Appendix C. These benchmarks assess predictability, not segmentation validity or a causal benefit of cellular over spot resolution.

4.2.2 Qualitative visualization of expression predictions

Figure 5 shows observed and predicted MT-CO3 and IGKC expression in human lung cancer tissue. Both panels use the left two thirds of sample A for training. The upper panel tests on the right third

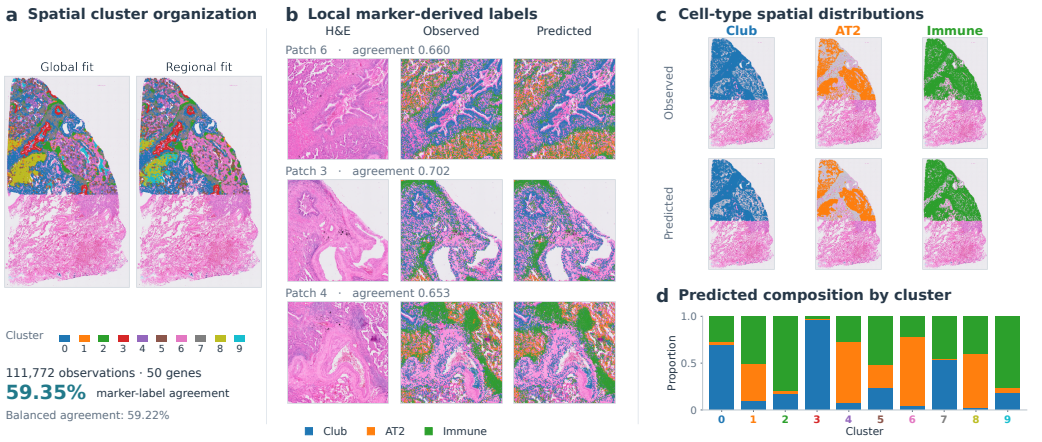


Figure 6: **Cell-type prediction as a downstream example.** A coarse-to-fine run on Visium HD lung tissue, using 111,772 observations. (a) Global and region-fitted clusters of predicted expression, aligned by cluster overlap. (b) The three patches from the original appendix analysis, comparing H&E with marker-derived labels from observed and predicted expression. (c) Spatial distributions of the three retained classes. (d) Predicted class composition by global cluster. Reference labels are marker-derived; agreement summarizes the full analysis cohort.

of A and the lower panel on sample B. Tissue overviews show spatial patterns, while matched H&E patches reveal differences between observed and predicted expression.

Sample A is the Visium HD post-Xenium FFPE control and B is a fixed-frozen lung cancer acquisition [97, 98]. The CONCH–ridge model uses training-only preprocessing and 50 training-selected genes. The patch locations were selected using expression and patch performance and are not an unbiased sample of tissue regions. A and B differ in preparation, and donor independence is not established; this illustrates cross-acquisition transfer rather than an isolated patient effect.

4.3 Supporting Diverse Downstream Analyses

Aligned histology, expression, and rich biological context support downstream analyses of cellular organization and model generalization. We illustrate these uses through marker-derived cell-type prediction and context-aware analysis of cross-sample bias.

4.3.1 Cell-Type Prediction from Inferred Gene Expression

Figure 6 presents a coarse-to-fine expression-prediction analysis on a representative Visium HD lung acquisition. Its 111,772 observations and 50-gene target subset support three coarse marker-derived classes: *Club*, *AT2*, and *Immune*. Labels derived from observed and predicted expression show partial agreement (59.35% overall; 59.22% balanced); all 10 clusters retain the same dominant class. Whole-slide maps, selected patches, and cluster compositions expose preserved structure and local disagreement. Agreement is evaluated over the full analysis cohort using expression-derived reference labels. Metrics appear in Appendix D.2.

Agreement varied even among clusters that retained their dominant class. In cluster 3, the labels assigned from predicted expression matched those assigned from observed expression for most spatial observations (93.4%). In clusters 1, 5, and 8, fewer than half of the labels agreed. A map colored by the most frequent class can therefore look consistent across an entire region while obscuring disagreements within it. The paired patch views make these local differences visible.

4.3.2 Context-Aware Analysis of Cross-Sample Bias

Rich context enables evaluation stratified by sample attributes such as age, disease state, and tissue preparation. The ovarian transfer example in Table 3 shows lower predictive agreement on the younger target: F1 falls from 0.289 to 0.072 (75.0%), and both gene- and cell-wise Spearman correlations decrease. The predicted nonzero fraction falls from 0.166 to 0.027.

Table 3: **Context-stratified ovarian transfer.** Both targets use the same AK source; the younger target shows lower prediction scores and a lower predicted nonzero fraction.

Transfer	F1	Gene Spearman	Cell Spearman	Predicted nonzero rate
AK(60+) → AD(60+)	0.289	0.036	0.206	0.166
AK(60+) → AL(< 40)	0.072	0.006	0.095	0.027

These descriptive examples use different target gene sets and do not isolate age from other sample differences. Preserving biological context makes potential sources of unequal model performance explicit and supports matched follow-up evaluations. Appendix D.3 reports the protocol and complete eight-metric visualization, including sparsity and global distribution distances.

A follow-up comparison could evaluate the same measured genes in both ovarian targets and report detection separately for each gene. This would show whether the drop in predicted nonzero expression is concentrated in a few genes or extends across the shared panel. Replicate acquisitions within each age band would then allow variation between samples of similar age to be compared with variation across bands.

5 Release, Reproducibility, and Responsible Use

The processed data are available at https://huggingface.co/datasets/zhang-lab/sMCC40M. Users can download individual acquisitions with their expression matrices, images, masks, and metadata. The release includes checksums, source attribution, and frozen benchmark splits. Code, evaluation instructions, and results are available at https://github.com/aicb-ZhangLabs/sMCC40M.

6 Discussion and Limitations

HD expression profiles are derived by aligning and aggregating spatial measurements at the cell level. We conducted a detailed audit of this process, although segmentation and registration errors can still affect which transcripts are assigned to each cell. The number of studies remains modest because few experiments are currently available, and coverage varies across tissues, species, and platforms. Some acquisitions share source material, so acquisition-level holdouts alone cannot establish transfer to independent donors. We will continue to expand and update the dataset as new studies become available, including a growing collection of in-house data that will broaden coverage and support evaluation across independent samples.

7 Conclusion

We presented sMCC40M, a spatial transcriptomics resource that connects cell-aligned expression with histology, spatial coordinates, and biological context across Visium HD and Xenium. Its scale supports encoder benchmarking, while cellular resolution enables expression prediction to be evaluated within and across acquisitions. The 23-category in-domain benchmark and matched cross-sample comparisons reveal both useful morphology–expression associations and persistent transfer gaps. Downstream examples illustrate how predicted expression can describe coarse cellular organization and how contextual metadata help interpret variation in transfer performance. Together, the resource and evaluation protocols provide a foundation for developing and comparing single-cell histology-to-molecular models. We envision sMCC40M supporting spatial perturbation research and the progression from virtual cell to virtual tissue modeling, while contributing to a shared data foundation for AI-driven biology.

Acknowledgments and Disclosure of Funding

This work was supported by the National Institutes of Health under Award Numbers R35GM150838, R01HG012572, R01NS128523, and R01DA063316.

References

[1] Guillaume Jaume, Paul Doucet, Andrew H. Song, Ming Y. Lu, Cristina Almagro-Pérez, Sophia J. Wagner, Anurag J. Vaidya, Richard J. Chen, Drew F.K. Williamson, Ahrong Kim, and Faisal Mahmood. HEST-1k: A Dataset For Spatial Transcriptomics and Histology Image Analysis. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 53798–53833. Curran Associates, Inc., 2024.

[2] Jiawen Chen, Muqing Zhou, Wenrong Wu, Jinwei Zhang, Yun Li, and Didong Li. STimage-1K4M: A histopathology image-gene expression dataset for spatial transcriptomics. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 35796–35823. Curran Associates, Inc., 2024.

[3] Patrik L. Ståhl, 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, Åke 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.

[4] Rongqin Ke, Marco Mignardi, Alexandra Pacureanu, Jessica Svedlund, Johan Botling, Carolina Wählby, and Mats Nilsson. In situ sequencing for RNA analysis in preserved tissue and cells. *Nature Methods*, 10(9):857–860, 2013.

[5] Jeffrey R. Moffitt, Junjie Hao, Guiping Wang, Kok Hao Chen, Hazen P. Babcock, and Xiaowei Zhuang. High-throughput single-cell gene-expression profiling with multiplexed error-robust fluorescence in situ hybridization. *Proceedings of the National Academy of Sciences*, 113(39):11046–11051, 2016.

[6] 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.

[7] Guiping Wang, Jeffrey R. Moffitt, and Xiaowei Zhuang. Multiplexed imaging of high-density libraries of RNAs with MERFISH and expansion microscopy. *Scientific Reports*, 8(1):4847, 2018.

[8] Chee-Huat Linus Eng, Michael Lawson, Qian Zhu, Ruben Dries, Noushin Koulena, Yodai Takei, Jina Yun, Christopher Cronin, Christoph Karp, Guo-Cheng Yuan, and Long Cai. Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH+. *Nature*, 568(7751):235–239, 2019.

[9] Sanja Vickovic, Gökcen Eraslan, Fredrik Salmén, Johanna Klughammer, Linnea Stenbeck, Denis Schapiro, Tarmo Äijö, Richard Bonneau, Ludvig Bergenstråhle, José Fernandéz Navarro, Joshua Gould, Gabriel K. Griffin, Åke Borg, Mostafa Ronaghi, Jonas Frisén, Joakim Lundeberg, Aviv Regev, and Patrik L. Ståhl. High-definition spatial transcriptomics for in situ tissue profiling. *Nature Methods*, 16(10):987–990, 2019.

[10] Robert R. Stickels, Evan Murray, Pawan Kumar, Jilong Li, Jamie L. Marshall, Daniela J. Di Bella, Paola Arlotta, Evan Z. Macosko, and Fei Chen. Highly sensitive spatial transcriptomics at near-cellular resolution with Slide-seqV2. *Nature Biotechnology*, 39(3):313–319, 2021.

[11] Yang Liu, Mingyu Yang, Yanxiang Deng, Graham Su, Archibald Enninful, Cindy C. Guo, Toma Tebaldi, Di Zhang, Dongjoo Kim, Zhiliang Bai, Eileen Norris, Alisia Pan, Jiatong Li, Yang Xiao, Stephanie Halene, and Rong Fan. High-Spatial-Resolution Multi-Omics Sequencing via Deterministic Barcoding in Tissue. *Cell*, 183(6):1665–1681.e18, 2020.

[12] 10x Genomics. Introducing the Visium Spatial Gene Expression Solution. https://www.10xgenomics.com/videos/n2lbrg209s, 2019. Accessed: 2026-10-21.

[13] Amanda Janesick, Robert Shelansky, Andrew D. Gottscho, Florian Wagner, Stephen R. Williams, Morgane Rouault, Ghezal Beliakoff, Carolyn A. Morrison, Michelli F. Oliveira, Jordan T. Sicherman, Andrew Kohlway, Jawad Abousoud, Tingsheng Yu Drennon, Seayar H. Mohabbat, 10x Development Teams, and Sarah E. B. Taylor. High resolution mapping of the tumor microenvironment using integrated single-cell, spatial and in situ analysis. *Nature Communications*, 14(1):8353, 2023.

[14] Michelli Faria de Oliveira, Juan Pablo Romero, Meii Chung, Stephen R. Williams, Andrew D. Gottscho, Anushka Gupta, Susan E. Pilipauskas, Seayar Mohabbat, Nandhini Raman, David J. Sukovich, David M. Patterson, Visium HD Development Team, and Sarah E. B. Taylor. High-definition spatial transcriptomic profiling of immune cell populations in colorectal cancer. *Nature Genetics*, 57(6):1512–1523, 2025.

[15] Ao Chen, Sha Liao, Mengnan Cheng, Kailong Ma, Liang Wu, Yiwei Lai, Xiaojie Qiu, Jin Yang, Jiangshan Xu, Shijie Hao, Xin Wang, Huifang Lu, Xi Chen, Xing Liu, Xin Huang, Zhao Li, Yan Hong, Yujia Jiang, Jian Peng, Shuai Liu, Mengzhe Shen, Chuanyu Liu, Quanshui Li, Yue Yuan, Xiaoyu Wei, Huiwen Zheng, Weimin Feng, Zhifeng Wang, Yang Liu, Zhaohui Wang, Yunzhi Yang, Haitao Xiang, Lei Han, Baoming Qin, Pengcheng Guo, Guangyao Lai, Pura Muñoz-Cánoves, Patrick H. Maxwell, Jean Paul Thiery, Qing-Feng Wu, Fuxiang Zhao, Bichao Chen, Mei Li, Xi Dai, Shuai Wang, Haoyan Kuang, Junhou Hui, Liqun

Wang, Ji-Feng Fei, Ou Wang, Xiaofeng Wei, Haorong Lu, Bo Wang, Shiping Liu, Ying Gu, Ming Ni, Wenwei Zhang, Feng Mu, Ye Yin, Huanming Yang, Michael Lisby, Richard J. Cornall, Jan Mulder, Mathias Uhlén, Miguel A. Esteban, Yuxiang Li, Longqi Liu, Xun Xu, and Jian Wang. Spatiotemporal transcriptomic atlas of mouse organogenesis using DNA nanoball-patterned arrays. *Cell*, 185(10):1777–1792.e21, 2022.

[16] Ruben Dries, Qian Zhu, Rui Dong, Chee-Huat Linus Eng, Huipeng Li, Kan Liu, Yuntian Fu, Tianxiao Zhao, Arpan Sarkar, Feng Bao, Rani E. George, Nico Pierson, Long Cai, and Guo-Cheng Yuan. Giotto: a toolbox for integrative analysis and visualization of spatial expression data. *Genome Biology*, 22(1):78, 2021.

[17] Giovanni Palla, Hannah Spitzer, Michal Klein, David Fischer, Anna Christina Schaar, Louis Benedikt Kuemmerle, Sergei Rybakov, Ignacio L. Ibarra, Olle Holmberg, Isaac Virshup, Mohammad Lotfollahi, Sabrina Richter, and Fabian J. Theis. Squidpy: a scalable framework for spatial omics analysis. *Nature Methods*, 19(2):171–178, 2022.

[18] Jian Hu, Xiangjie Li, Kyle Coleman, Amelia Schroeder, Nan Ma, David J. Irwin, Edward B. Lee, Russell T. Shinohara, and Mingyao Li. SpaGCN: Integrating gene expression, spatial location and histology to identify spatial domains and spatially variable genes by graph convolutional network. *Nature Methods*, 18(11):1342–1351, 2021.

[19] Edward Zhao, Matthew R. Stone, Xing Ren, Jamie Guenthoer, Kimberly S. Smythe, Thomas Pulliam, Stephen R. Williams, Cedric R. Uytingco, Sarah E. B. Taylor, Paul Nghiem, Jason H. Bielas, and Raphael Gottardo. Spatial transcriptomics at subspot resolution with BayesSpace. *Nature Biotechnology*, 39(11):1375–1384, 2021.

[20] Tommaso Biancalani, Gabriele Scalia, Lorenzo Buffoni, Raghav Avasthi, Ziqing Lu, Aman Sanger, Neriman Tokcan, Charles R. Vanderburg, Åsa Segerstolpe, Meng Zhang, Inbal Avraham-Davidi, Sanja Vickovic, Mor Nitzan, Sai Ma, Ayshwarya Subramanian, Michal Lipinski, Jason Buenrostro, Nik Bear Brown, Duccio Fanelli, Xiaowei Zhuang, Evan Z. Macosko, and Aviv Regev. Deep learning and alignment of spatially resolved single-cell transcriptomes with Tangram. *Nature Methods*, 18(11):1352–1362, 2021.

[21] Vitalii Kleshchevnikov, Artem Shmatko, Emma Dann, Alexander Aivazidis, Hamish W. King, Tong Li, Rasa Elmentaite, Artem Lomakin, Veronika Kedlian, Adam Gayoso, Mika Sarkin Jain, Jun Sung Park, Lauma Ramona, Elizabeth Tuck, Anna Arutyunyan, Roser Vento-Tormo, Moritz Gerstung, Louisa James, Oliver Stegle, and Omer Ali Bayraktar. Cell2location maps fine-grained cell types in spatial transcriptomics. *Nature Biotechnology*, 40(5):661–671, 2022.

[22] Ludvig Bergenstråhle, Bryan He, Joseph Bergenstråhle, Xesús Abalo, Reza Mirzazadeh, Kim Thrane, Andrew L. Ji, Alma Andersson, Ludvig Larsson, Nathalie Stakenborg, Guy Boeckxstaens, Paul Khavari, James Zou, Joakim Lundeberg, and Jonas Maaskola. Super-resolved spatial transcriptomics by deep data fusion. *Nature Biotechnology*, 40(4):476–479, 2022.

[23] Reuben Moncada, Dalia Barkley, Florian Wagner, Marta Chiodin, Joseph C. Devlin, Maayan Baron, Cristina H. Hajdu, Diane M. Simeone, and Itai Yanai. Integrating microarray-based spatial transcriptomics and single-cell RNA-seq reveals tissue architecture in pancreatic ductal adenocarcinomas. *Nature Biotechnology*, 38(3):333–342, 2020.

[24] Alma Andersson, Joseph Bergenstråhle, Michaela Asp, Ludvig Bergenstråhle, Aleksandra Jurek, José Fernández Navarro, and Joakim Lundeberg. Single-cell and spatial transcriptomics enables probabilistic inference of cell type topography. *Communications Biology*, 3(1):565, 2020.

[25] Rahul Satija, Jeffrey A Farrell, David Gennert, Alexander F Schier, and Aviv Regev. Spatial reconstruction of single-cell gene expression data. *Nature Biotechnology*, 33(5):495–502, 2015.

[26] Ziheng Duan, Dylan Riffle, Ren Li, Junhao Liu, Martin Renqiang Min, and Jing Zhang. Impeller: a path-based heterogeneous graph learning method for spatial transcriptomic data imputation. *Bioinformatics*, 40(6):btae339, 2024.

[27] Ziheng Duan, Siwei Xu, Cheyu Lee, Dylan Riffle, and Jing Zhang. iMIRACLE: An Iterative Multi-View Graph Neural Network to Model Intercellular Gene Regulation From Spatial Transcriptomic Data. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 538–548. ACM, 2024.

[28] Ziheng Duan, Xi Li, Zhuoyang Zhang, James Song, and Jing Zhang. DISCO: A Diffusion Model For Spatial Transcriptomics Data Completion. In *2025 IEEE International Conference on Image Processing (ICIP)*, pages 19–24. IEEE, 2025.

[29] Ziheng Duan, Xi Li, Zhiqing Xiao, Rex Ying, and Jing Zhang. MUSE: A Multi-slice Joint Analysis Method for Spatial Transcriptomics Experiments. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, pages 625–634. ACM, 2025.

[30] Tim Stuart, Andrew Butler, Paul Hoffman, Christoph Hafemeister, Efthymia Papalexi, William M. Mauck, III, Yuhan Hao, Marlon Stoeckius, Peter Smibert, and Rahul Satija. Comprehensive Integration of Single-Cell Data. *Cell*, 177(7):1888–1902.e21, 2019.

[31] Yuhan Hao, Stephanie Hao, Erica Andersen-Nissen, William M. Mauck, III, Shiwei Zheng, Andrew Butler, Maddie J. Lee, Aaron J. Wilk, Charlotte Darby, Michael Zager, Paul Hoffman, Marlon Stoeckius, Efthymia Papalexi, Eleni P. Mimitou, Jaison Jain, Avi Srivastava, Tim Stuart, Lamar M. Fleming, Bertrand Yeung, Angela J. Rogers, Juliana M. McElrath, Catherine A. Blish, Raphael Gottardo, Peter Smibert, and Rahul Satija. Integrated analysis of multimodal single-cell data. *Cell*, 184(13):3573–3587.e29, 2021.

[32] F. Alexander Wolf, Philipp Angerer, and Fabian J. Theis. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biology*, 19(1):15, 2018.

[33] Aviv Regev, Sarah A Teichmann, Eric S Lander, Ido Amit, Christophe Benoist, Ewan Birney, Bernd Bodenmiller, Peter Campbell, Piero Carninci, Menna Clatworthy, Hans Clevers, Bart Deplancke, Ian Dunham, James Eberwine, Roland Eils, Wolfgang Enard, Andrew Farmer, Lars Fugger, Berthold Göttgens, Nir Hacohen, Muzlifah Haniffa, Martin Hemberg, Seung Kim, Paul Klenerman, Arnold Kriegstein, Ed Lein, Sten Linnarsson, Emma Lundberg, Joakim Lundeberg, Partha Majumder, John C Marioni, Miriam Merad, Musa Mhlanga, Martijn Nawijn, Mihai Netea, Garry Nolan, Dana Pe'er, Anthony Phillipakis, Chris P Ponting, Stephen Quake, Wolf Reik, Orit Rozenblatt-Rosen, Joshua Sanes, Rahul Satija, Ton N Schumacher, Alex Shalek, Ehud Shapiro, Padmanee Sharma, Jay W Shin, Oliver Stegle, Michael Stratton, Michael J T Stubbington, Fabian J Theis, Matthias Uhlen, Alexander van Oudenaarden, Allon Wagner, Fiona Watt, Jonathan Weissman, Barbara Wold, Ramnik Xavier, Nir Yosef, and Human Cell Atlas Meeting Participants. The Human Cell Atlas. *eLife*, 6:e27041, 2017.

[34] The Tabula Sapiens Consortium, Robert C. Jones, Jim Karkanias, Mark A. Krasnow, Angela Oliveira Pisco, Stephen R. Quake, Julia Salzman, Nir Yosef, Bryan Bulthaup, Phillip Brown, William Harper, Marisa Hemenez, Ravikumar Ponnusamy, Ahmad Salehi, Bhavani A. Sanagavarapu, Eileen Spallino, Ksenia A. Aaron, Waldo Concepcion, James M. Gardner, Burnett Kelly, Nikole Neidlinger, Zifa Wang, Sheela Crasta, Saroja Kolluru, Maurizio Morri, Serena Y. Tan, Kyle J. Travaglini, Chenling Xu, Marcela Alcántara-Hernández, Nicole Almanzar, Jane Antony, Benjamin Beyersdorf, Deviana Burhan, Kruti Calcuttawala, Matthew M. Carter, Charles K. F. Chan, Charles A. Chang, Stephen Chang, Alex Colville, Rebecca N. Culver, Ivana Cvijović, Gaetano D'Amato, Camille Ezran, Francisco X. Galdos, Astrid Gillich, William R. Goodyer, Yan Hang, Alyssa Hayashi, Sahar Houshdaran, Xianxi Huang, Juan C. Irwin, SoRi Jang, Julia Vallve Juanico, Aaron M. Kershner, Soochi Kim, Bernhard Kiss, William Kong, Maya E. Kumar, Angera H. Kuo, Rebecca Leylek, Baoxiang Li, Gabriel B. Loeb, Wan-Jin Lu, Sruthi Mantri, Maxim Markovic, Patrick L. McAlpine, Antoine de Morree, Karim Mrouj, Shravani Mukherjee, Tyler Muser, Patrick Neuhöfer, Thi D. Nguyen, Kimberly Perez, Ragini Phansalkar, Nazan Puluca, Zhen Qi, Poorvi Rao, Hayley Raquer-McKay, Nicholas Schaum, Bronwyn Scott, Bobak Seddighzadeh, Joe Segal, Sushmita Sen, Shaheen Sikandar, Sean P. Spencer, Lea C. Steffes, Varun R. Subramaniam, Aditi Swarup, Michael Swift, Will Van Treuren, Emily Trimm, Stefan Veizades, Sivakamasundari Vijayakumar, Kim Chi Vo, Sevahn K. Vorperian, Wanxin Wang, Hannah N. W. Weinstein, Juliane Winkler, Timothy T. H. Wu, Jamie Xie, Andrea R. Yung, Yue Zhang, Angela M. Detweiler, Honey Mekonen, Norma F. Neff, Rene V. Sit, Michelle Tan, Jia Yan, Gregory R. Bean, Vivek Charu, Erna Forgó, Brock A. Martin, Michael G. Ozawa, Oscar Silva, Angus Toland, Venkata N. P. Vemuri, Shaked Afik, Kyle Awayan, Olga Borisovna Botvinnik, Ashley Byrne, Michelle Chen, Roozbeh Dehghannasiri, Adam Gayoso, Alejandro A. Granados, Qiqing Li, Gita Mahmoudabadi, Aaron McGeever, Julia Eve Olivieri, Madeline Park, Neha Ravikumar, Geoff Stanley, Weilun Tan, Alexander J. Tarashansky, Rohan Vanheusden, Peter Wang, Sheng Wang, Galen Xing, Les Dethlefsen, Po-Yi Ho, Shixuan Liu, Jonathan S. Maltzman, Ross J. Metzger, Koki Sasagawa, Rahul Sinha, Hanbing Song, Bruce Wang, Steven E. Artandi, Philip A. Beachy, Michael F. Clarke, Linda C. Giudice, Franklin W. Huang, Kerwyn Casey Huang, Juliana Idoyaga, Seung K. Kim, Christin S. Kuo, Patricia Nguyen, Thomas A. Rando, Kristy Red-Horse, Jeremy Reiter, David A. Relman, Justin L. Sonnenburg, Albert Wu, Sean M. Wu, and Tony Wyss-Coray. The Tabula Sapiens: A multiple-organ, single-cell transcriptomic atlas of humans. *Science*, 376(6594):eabl4896, 2022.

[35] Orit Rozenblatt-Rosen, Aviv Regev, Philipp Oberdoerffer, Tal Nawy, Anna Hupalowska, Jennifer E. Rood, Orr Ashenberg, Ethan Cerami, Robert J. Coffey, Emek Demir, Li Ding, Edward D. Esplin, James M. Ford, Jeremy Goecks, Sharmistha Ghosh, Joe W. Gray, Justin Guinney, Sean E. Hanlon, Shannon K. Hughes, E. Shelley Hwang, Christine A. Iacobuzio-Donahue, Judit Jané-Valbuena, Bruce E. Johnson, Ken S. Lau, Tracy Lively, Sarah A. Mazzilli, Dana Pe'er, Sandro Santagata, Alex K. Shalek, Denis Schapiro, Michael P. Snyder, Peter K. Sorger, Avrum E. Spira, Sudhir Srivastava, Kai Tan, Robert B. West, Elizabeth H. Williams, and Human Tumor Atlas Network. The Human Tumor Atlas Network: Charting tumor transitions across space and time at single-cell resolution. *Cell*, 181(2):236–249, 2020.

[36] Benoît Schmauch, Alberto Romagnoni, Elodie Pronier, Charlie Saillard, Pascale Maillé, Julien Calderaro, Aurélie Kamoun, Meriem Sefta, Sylvain Toldo, Mikhail Zaslavskiy, Thomas Clozel, Matahi Moarii, Pierre Courtiol, and Gilles Wainrib. A deep learning model to predict RNA-Seq expression of tumours from whole slide images. *Nature Communications*, 11(1):3877, 2020.

[37] 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.

[38] 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. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 70626–70637. Curran Associates, Inc., 2023.

[39] Daiwei Zhang, Amelia Schroeder, Hanying Yan, Haochen Yang, Jian Hu, Michelle Y. Y. Lee, Kyung S. Cho, Katalin Susztak, George X. Xu, Michael D. Feldman, Edward B. Lee, Emma E. Furth, Linghua Wang, and Mingyao Li. Inferring super-resolution tissue architecture by integrating spatial transcriptomics with histology. *Nature Biotechnology*, 42(9):1372–1377, 2024.

[40] 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.

[41] Gabriele Campanella, Matthew G. Hanna, Luke Geneslaw, Allen Miraflor, Vitor Werneck Krauss Silva, Klaus J. Busam, Edi Brogi, Victor E. Reuter, David S. Klimstra, and Thomas J. Fuchs. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nature Medicine*, 25(8):1301–1309, 2019.

[42] Ming Y. Lu, Drew F. K. Williamson, Tiffany Y. Chen, Richard J. Chen, Matteo Barbieri, and Faisal Mahmood. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nature Biomedical Engineering*, 5(6):555–570, 2021.

[43] Zhuchen Shao, Hao Bian, Yang Chen, Yifeng Wang, Jian Zhang, Xiangyang Ji, and Yongbing Zhang. TransMIL: Transformer based Correlated Multiple Instance Learning for Whole Slide Image Classification. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 2136–2147. Curran Associates, Inc., 2021.

[44] Maximilian Ilse, Jakub Tomczak, and Max Welling. Attention-based Deep Multiple Instance Learning. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2127–2136. PMLR, 2018.

[45] Ozan Ciga, Tony Xu, and Anne Louise Martel. Self supervised contrastive learning for digital histopathology. *Machine Learning with Applications*, 7:100198, 2022.

[46] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, volume 9351 of *Lecture Notes in Computer Science*, pages 234–241. Springer, 2015.

[47] Simon Graham, Quoc Dang Vu, Shan E Ahmed Raza, Ayesha Azam, Yee Wah Tsang, Jin Tae Kwak, and Nasir Rajpoot. HoVer-Net: Simultaneous segmentation and classification of nuclei in multi-tissue histology images. *Medical Image Analysis*, 58:101563, 2019.

[48] Fabian Hörst, Moritz Rempe, Lukas Heine, Constantin Seibold, Julius Keyl, Giulia Baldini, Selma Ugurel, Jens Siveke, Barbara Grünwald, Jan Egger, and Jens Kleesiek. CellViT: Vision Transformers for precise cell segmentation and classification. *Medical Image Analysis*, 94:103143, 2024.

[49] Babak Ehteshami Bejnordi, Mitko Veta, Paul Johannes van Diest, Bram van Ginneken, Nico Karssemeijer, Geert Litjens, Jeroen A. W. M. van der Laak, and the CAMELYON16 Consortium. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *JAMA*, 318(22):2199–2210, 2017.

[50] Dayong Wang, Aditya Khosla, Rishab Gargeya, Humayun Irshad, and Andrew H. Beck. Deep learning for identifying metastatic breast cancer, 2016.

[51] Daisuke Komura and Shumpei Ishikawa. Machine Learning Methods for Histopathological Image Analysis. *Computational and Structural Biotechnology Journal*, 16:34–42, 2018.

[52] Jakob Nikolas Kather, Alexander T. Pearson, Niels Halama, Dirk Jäger, Jeremias Krause, Sven H. Loosen, Alexander Marx, Peter Boor, Frank Tacke, Ulf Peter Neumann, Heike I. Grabsch, Takaki Yoshikawa, Hermann Brenner, Jenny Chang-Claude, Michael Hoffmeister, Christian Trautwein, and Tom Luedde. Deep learning can predict microsatellite instability directly from histology in gastrointestinal cancer. *Nature Medicine*, 25(7):1054–1056, 2019.

[53] Wisdom Ikezogwo, Saygin Seyfioglu, Fatemeh Ghezloo, Dylan Geva, Fatwir Sheikh Mohammed, Pavan Kumar Anand, Ranjay Krishna, and Linda Shapiro. Quilt-1M: One million image-text pairs for histopathology. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 37995–38017. Curran Associates, Inc., 2023.

[54] Xiyue Wang, Sen Yang, Jun Zhang, Minghui Wang, Jing Zhang, Wei Yang, Junzhou Huang, and Xiao Han. Transformer-based unsupervised contrastive learning for histopathological image classification. *Medical Image Analysis*, 81:102559, 2022.

[55] Richard J. Chen, Tong Ding, Ming Y. Lu, Drew F. K. Williamson, Guillaume Jaume, Andrew H. Song, Bowen Chen, Andrew Zhang, Daniel Shao, Muhammad Shaban, Mane Williams, Lukas Oldenburg, Luca L. Weishaupt, Judy J. Wang, Anurag Vaidya, Long Phi Le, Georg Gerber, Sharifa Sahai, Walt Williams, and Faisal Mahmood. Towards a general-purpose foundation model for computational pathology. *Nature Medicine*, 30(3):850–862, 2024.

[56] Mahmood Lab. UNI2-h. https://huggingface.co/MahmoodLab/UNI2-h, 2025. Accessed: 2026-10-21.

[57] Alexandre Filiot, Paul Jacob, Alice Mac Kain, and Charlie Saillard. Phikon-v2, a large and public feature extractor for biomarker prediction. *arXiv preprint arXiv:2409.09173*, 2024.

[58] Hanwen Xu, Naoto Usuyama, Jaspreet Bagga, Sheng Zhang, Rajesh Rao, Tristan Naumann, Cliff Wong, Zelalem Gero, Javier González, Yu Gu, Yanbo Xu, Mu Wei, Wenhui Wang, Shuming Ma, Furu Wei, Jianwei Yang, Chunyuan Li, Jianfeng Gao, Jaylen Rosemon, Tucker Bower, Soohee Lee, Roshanthi Weerasinghe, Bill J. Wright, Ari Robicsek, Brian Piening, Carlo Bifulco, Sheng Wang, and Hoifung Poon. A whole-slide foundation model for digital pathology from real-world data. *Nature*, 630(8015):181–188, 2024.

[59] Eugene Vorontsov, Alican Bozkurt, Adam Casson, George Shaikovski, Michal Zelechowski, Kristen Severson, Eric Zimmermann, James Hall, Neil Tenenholtz, Nicolo Fusi, Ellen Yang, Philippe Mathieu, Alexander van Eck, Donghun Lee, Julian Viret, Eric Robert, Yi Kan Wang, Jeremy D. Kunz, Matthew C. H. Lee, Jan H. Bernhard, Ran A. Godrich, Gerard Oakley, Ewan Millar, Matthew Hanna, Hannah Wen, Juan A. Retamero, William A. Moye, Razik Yousfi, Christopher Kanan, David S. Klimstra, Brandon Rothrock, Siqi Liu, and Thomas J. Fuchs. A foundation model for clinical-grade computational pathology and rare cancers detection. *Nature Medicine*, 30(10):2924–2935, 2024.

[60] Eric Zimmermann, Eugene Vorontsov, Julian Viret, Adam Casson, Michal Zelechowski, George Shaikovski, Neil Tenenholtz, James Hall, David Klimstra, Razik Yousfi, Thomas Fuchs, Nicolo Fusi, Siqi Liu, and Kristen Severson. Virchow2: Scaling self-supervised mixed magnification models in pathology. *arXiv preprint arXiv:2408.00738*, 2024.

[61] Charlie Saillard, Rodolphe Jenatton, Felipe Llinares-López, Zelda Mariet, David Cahané, Eric Durand, and Jean-Philippe Vert. H-optimus-0. https://github.com/bioptimus/releases/tree/main/models/h-optimus/v0, 2024.

[62] Ming Y. Lu, Bowen Chen, Drew F. K. Williamson, Richard J. Chen, Ivy Liang, Tong Ding, Guillaume Jaume, Igor Odintsov, Long Phi Le, Georg Gerber, Anil V. Parwani, Andrew Zhang, and Faisal Mahmood. A visual-language foundation model for computational pathology. *Nature Medicine*, 30(3):863–874, 2024.

[63] Zhi Huang, Federico Bianchi, Mert Yuksekgonul, Thomas J. Montine, and James Zou. A visual–language foundation model for pathology image analysis using medical Twitter. *Nature Medicine*, 29(9):2307–2316, 2023.

[64] Yuxuan Sun, Chenglu Zhu, Sunyi Zheng, Kai Zhang, Lin Sun, Zhongyi Shui, Yunlong Zhang, Honglin Li, and Lin Yang. PathAsst: A generative foundation ai assistant towards artificial general intelligence of pathology. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 5034–5042, 2024.

[65] Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, Cliff Wong, Andrea Tupini, Yu Wang, Matt Mazzola, Swadheen Shukla, Lars Liden, Jianfeng Gao, Angela Crabtree, Brian Piening, Carlo Bifulco, Matthew P. Lungren, Tristan Naumann, Sheng Wang, and Hoifung Poon. BiomedCLIP: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. *arXiv preprint arXiv:2303.00915*, 2023.

[66] Yuxuan Sun, Yunlong Zhang, Yixuan Si, Chenglu Zhu, Kai Zhang, Zhongyi Shui, Jingxiong Li, Xuan Gong, Xinheng Lyu, Tao Lin, and Lin Yang. PathGen-1.6M: 1.6 Million Pathology Image-text Pairs Generation through Multi-agent Collaboration. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu, editors, *International Conference on Learning Representations*, volume 2025, pages 94611–94653, 2025.

[67] Shekoofeh Azizi, Laura Culp, Jan Freyberg, Basil Mustafa, Sebastien Baur, Simon Kornblith, Ting Chen, Nenad Tomasev, Jovana Mitrović, Patricia Strachan, S. Sara Mahdavi, Ellery Wulczyn, Boris Babenko, Megan Walker, Aaron Loh, Po-Hsuan Cameron Chen, Yuan Liu, Pinal Bavishi, Scott Mayer McKinney, Jim Winkens, Abhijit Guha Roy, Zach Beaver, Fiona Ryan, Justin Krogue, Mozziyar Etemadi, Umesh Telang, Yun Liu, Lily Peng, Greg S. Corrado, Dale R. Webster, David Fleet, Geoffrey Hinton, Neil Houlsby, Alan Karthikesalingam, Mohammad Norouzi, and Vivek Natarajan. Robust and data-efficient generalization of self-supervised machine learning for diagnostic imaging. *Nature Biomedical Engineering*, 7(6):756–779, 2023.

[68] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255. IEEE, 2009.

[69] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, volume 8693 of *Lecture Notes in Computer Science*, pages 740–755. Springer, 2014.

[70] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778. IEEE, 2016.

[71] Karen Simonyan and Andrew Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *International Conference on Learning Representations*, 2015.

[72] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going Deeper With Convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–9, June 2015.

[73] Mingxing Tan and Quoc Le. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 6105–6114. PMLR, 09–15 Jun 2019.

[74] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

[75] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Herve Jegou. Training data-efficient image transformers & distillation through attention. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 10347–10357. PMLR, 2021.

[76] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin Transformer: Hierarchical Vision Transformer using shifted windows. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9992–10002. IEEE, 2021.

[77] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A ConvNet for the 2020s. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11966–11976. IEEE, 2022.

[78] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9726–9735. IEEE, 2020.

[79] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A Simple Framework for Contrastive Learning of Visual Representations. In Hal Daumé, III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1597–1607. PMLR, 2020.

[80] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Remi Munos, and Michal Valko. Bootstrap Your Own Latent - A New Approach to Self-Supervised Learning. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 21271–21284. Curran Associates, Inc., 2020.

[81] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stephane Deny. Barlow Twins: Self-Supervised Learning via Redundancy Reduction. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 12310–12320. PMLR, 2021.

[82] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep Clustering for Unsupervised Learning of Visual Features. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 139–156, September 2018.

[83] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised Learning of Visual Features by Contrasting Cluster Assignments. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 9912–9924. Curran Associates, Inc., 2020.

[84] Xinlei Chen and Kaiming He. Exploring Simple Siamese Representation Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15750–15758, 2021.

[85] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging Properties in Self-Supervised Vision Transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9650–9660, 2021.

[86] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning Robust Visual Features without Supervision. *Transactions on Machine Learning Research*, 2024.

[87] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked Autoencoders Are Scalable Vision Learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16000–16009, 2022.

[88] Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. data2vec: A General Framework for Self-supervised Learning in Speech, Vision and Language. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 1298–1312. PMLR, 2022.

[89] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 2021.

[90] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, jun 2019. Association for Computational Linguistics.

[91] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020.

[92] OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael (Rai) Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such,

Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. GPT-4 Technical Report. *arXiv preprint arXiv:2303.08774*, 2023.

[93] Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. On the Opportunities and Risks of Foundation Models. *arXiv preprint arXiv:2108.07258*, 2021.

[94] Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, Etienne David, Ian Stavness, Wei Guo, Berton Earnshaw, Imran Haque, Sara M Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. WILDS: A Benchmark of in-the-Wild Distribution Shifts. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 5637–5664. PMLR, 18–24 Jul 2021.

[95] Jevgenij Gamper, Navid Alemi Koohbanani, Ksenija Benes, Simon Graham, Mostafa Jahanifar, Syed Ali Khurram, Ayesha Azam, Katherine Hewitt, and Nasir Rajpoot. PanNuke Dataset Extension, Insights and Baselines. arXiv preprint arXiv:2003.10778, 2020.

[96] The Cancer Genome Atlas Research Network, John N Weinstein, Eric A Collisson, Gordon B Mills, Kenna R Mills Shaw, Brad A Ozenberger, Kyle Ellrott, Ilya Shmulevich, Chris Sander, and Joshua M Stuart. The Cancer Genome Atlas Pan-Cancer analysis project. *Nature Genetics*, 45(10):1113–1120, 2013.

[97] 10x Genomics. Visium HD Spatial Gene Expression Libraries, Post-Xenium, Human Lung Cancer (FFPE). https://www.10xgenomics.com/datasets/visium-hd-cytassist-gene-expression-human-lung-cancer-post-xenium-expt, 2024. Public dataset, analyzed using Space Ranger 3.0.0. Published September 24, 2024; accessed October 21, 2026.

[98] 10x Genomics. Visium HD Spatial Gene Expression Library, Human Lung Cancer (Fixed Frozen). https://www.10xgenomics.com/datasets/visium-hd-cytassist-gene-expression-human-lung-cancer-fixed-frozen, 2024. Published: 2024-09-17; Space Ranger 3.1.1; accessed: 2026-10-21.

Appendix

A Ethical Considerations, Intended Usage, and License

We developed sMCC40M for research. Its annotations have not been validated for clinical diagnosis. Some components (e.g., segmentation or classification modules) are adapted from public models and should not be treated as gold standards. Users should exercise caution when applying these resources to downstream analyses.

All personally identifiable information—such as patient names, addresses, and social security numbers—has been removed. Users are **prohibited** from attempting to recover private information through reverse engineering or related techniques. When used under these conditions, no evidence of social or ethical risks has been found.

Code and benchmark documentation are provided through https://github.com/aicb-ZhangLabs/sMCC40M. Source-data URLs and provenance are retained in the acquisition inventory; access and reuse follow the applicable upstream terms.

The appendix release distinguishes between *public source studies*, *processed derivatives*, and *newly generated benchmark artifacts*. Public Xenium and Visium HD studies remain governed by their upstream portal and study-specific terms. Cell-aligned patches, harmonized expression matrices, split files, captions, and communication-graph summaries are processed derivatives or generated artifacts released for non-diagnostic research use. Clinical deployment, patient re-identification, and any safety-critical decision making are explicitly prohibited.

The release also inherits the long-tail structure of the underlying studies. Organ coverage, platform coverage, and panel size are highly uneven, so benchmark scores should not be interpreted as uniformly representative across all biological settings. Results should therefore be read together with the cohort and protocol definitions in Section C. Some categories support only within-acquisition evaluation; the combined 23-category mean does not establish cross-platform transfer performance.

B Data Source

In this section, we present two data sources collected from the 10x Genomics portal: **Xenium** and **Visium HD**. These datasets provide spatial transcriptomic information across various tissue types. A complete, deduplicated source-acquisition inventory is presented in Table S2. It fixes the official sample identifiers, assay, species/tissue, preservation, preferred analysis version, and source/download links for the 12 October 2026 catalogue snapshot. The processed freeze contains 124 acquisitions (53 HD and 71 Xenium) and 38,472,192 gene-positive cell/nucleus observations; 11 Xenium catalogue acquisitions are outside this processed freeze.

The frozen sMCC40M release has 23 harmonized reporting categories, including developmental, plant, whole-organism, and xenograft categories. These labels do not imply 23 distinct anatomical organs.

B.1 Xenium Data Source

The processed freeze contains 71 Xenium acquisitions and 24,811,436 gene-positive cellular observations. Native cellular counts, measured gene/feature identities, and cell boundaries are retained. H&E and DAPI identify image-display branches, not different molecular units: both retain cell-level expression. Coordinate transforms map native boundaries into the selected image. Targeted panels differ across acquisitions, so unmeasured genes are not filled with zeros. The current gene-prediction benchmark uses 13 H&E acquisitions in five categories absent from HD18; its panel handling is specified in Section C.1.

B.2 Visium HD Data Source

The processed freeze contains 53 Visium HD acquisitions and 13,660,756 gene-positive cell/nucleus observations. The H&E branch uses VitaminP cell footprints; the IF branch is nucleus-aligned. Filtered native 2 μm counts and supplied image transforms map each complete bin to the capture

grid. A bin contributes at most once, only when one footprint covers strictly more than half its area. Aggregation preserves raw integer counts, mask holes, disconnected components, and unassigned-bin status. All segmentation candidates remain indexed, with molecular support and image/crop eligibility recorded separately. The current H&E benchmark and training-only HVG selection are specified in Section C.2.

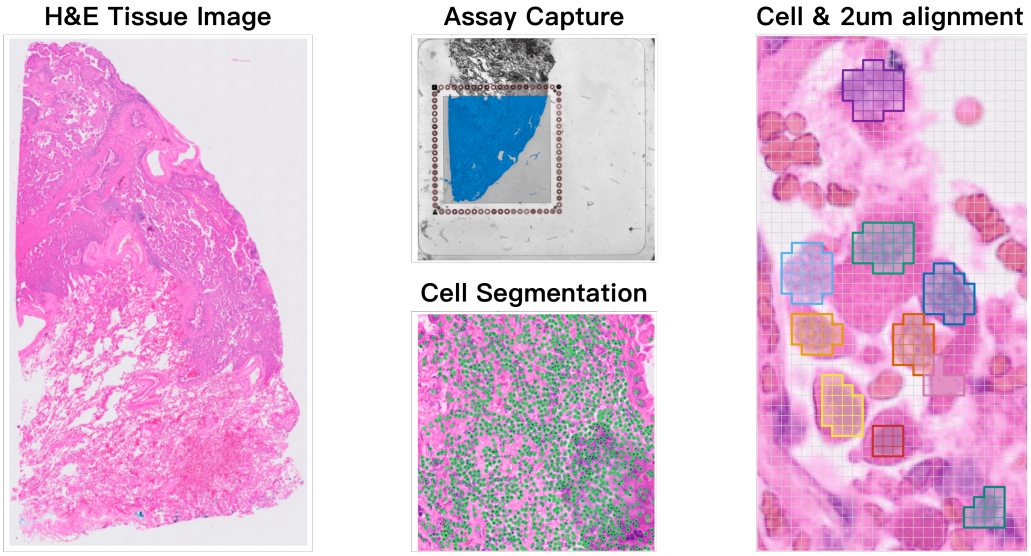


Figure S1: Schematic of the frozen Visium HD bin-to-cell construction. Native 2 μm bins are mapped to segmented cell or nuclear footprints and assigned only when more than half of the complete bin is covered. Expression availability and image/crop eligibility are recorded separately.

B.3 Frozen 10x catalogue and acquisition-level deduplication

The catalogue snapshot covers all 103 official Visium HD and Xenium dataset pages available on 12 October 2026. Expanding these pages into downloadable molecular sample records gives 146 sample-version observations. We consolidate eight analysis reruns, two sequencing-depth variants, and one spatial demo subset, yielding **135 distinct source acquisitions: 53 Visium HD and 82 Xenium**. Shared input-only filesets are not counted as samples. These are assay acquisitions, tissue sections, or ROIs; they are not counts of independent donors or biological studies.

For Visium HD, identity is established by the official slide serial number and capture area. For Xenium, we use the official molecular resource identifier, retaining the evidence for confirmed reruns and derived spatial subsets. Adjacent sections, experimental controls, different panels/assays, and RNA–protein acquisitions remain separate. In particular, the three mouse-brain replicate sections, the four HD lung control/post-Xenium sections, and the twelve breast-biomarker sections are retained individually. An ovarian capture's deep and minimum sequencing outputs are represented once, with deep sequencing selected; the cropped mouse-brain demo is linked to its full coronal section.

Each inventory row has a snapshot-scoped sample ID and one preferred full molecular download. We select the latest official analysis version available in this snapshot, preserve alternate versions and URLs in the provenance ledger, and keep full spatial coverage and deep sequencing for subset/depth variants. The assay codes are HD-WT (probe-based Visium HD), HD-P2 (its Protocol 2.0 workflow), HD-3prime, X-v1, X-Prime5K, and X-RNA-protein. SR and XOA denote Space Ranger and Xenium Onboard Analysis, respectively.

The machine-readable download specification is supplied with the manuscript sources in `data_sources/tenx_2026-10-12/`: `sample_inventory.csv` lists the deduplicated samples, `sample_versions.csv` retains all 146 sample-version observations, and `download_manifest.csv` provides exact asset URLs, provider byte counts/MD5 checksums, shared-file associations, and selected downloads. Observed checksums/download status and sample-level QC/inclusion are recorded separately in `download_tracking.csv` and `sample_qc_tracking.csv`; rebuilding the catalogue does not reset these ledgers. Identical shared assets are

listed once. Full molecular bundles, available histology/registration files, panel definitions, and source summaries are selected; FASTQs and viewer-only exports remain optional.

This freezes the *source catalogue and download specification*, distinct from the processed 28 October 2026 release of 124 acquisitions and 38,472,192 gene-positive observations. The processed sample index records acquisition membership, source URLs, measured feature axes, QC and image eligibility; checksum manifests bind the versioned files. The 11 remaining Xenium catalogue entries do not contribute to the reported processed totals.

B.3.1 Acquisition-level overlap with HEST-1k and STimage-1K4M

We compared our 124 frozen acquisitions with the 1,229 profiles reported in Figure 1a of the HEST-1k paper [1] and the 1,149 records in the corrected STimage-1K4M catalogue [2]. The September 2024 HEST manifest reproduces the paper's platform counts (602 Visium, 552 Spatial Transcriptomics, 10 Visium HD, and 65 Xenium). We screened all 75 Visium HD and Xenium profiles for shared acquisitions. The manifest revisions, checksums, and per-acquisition matches are in `data_sources/source_overlap_2026-11-02/`.

We matched native sample identifiers or Xenium ROI UUIDs where available. For multi-sample source pages, we resolved individual sections and panels using provider labels and image geometry. Where a HEST source URL was missing, we checked the provider title and image content (one record). For STimage, we checked native barcode coordinates and provider reports. Coordinates or tissue selection differed across processing runs for the brain and colon acquisitions (2 of 4 HD profiles). Each acquisition contributes once to Table S1.

Table S1: **Source overlap with the HEST-1k publication cohort and STimage-1K4M.** The four groups are mutually exclusive. The final column counts gene-positive cell or nucleus records in our frozen release for each group.

Catalogue membership	HD	Xenium	Total	Our observations
HEST only	3	36	39	13,317,654
STimage only	1	0	1	483,460
Both	3	0	3	565,951
No catalogue match	46	35	81	24,105,127
Total	**53**	**71**	**124**	**38,472,192**

Most HEST matches are Xenium acquisitions (36 Xenium and 6 HD, totaling 42 acquisitions and 13,883,605 gene-positive records in our release). HD-011 (brain), HD-022 (small intestine), and HD-043 (colon) occur in both catalogues. STimage also includes HD-028 (lung, IF/nuclear branch). Most release acquisitions have no identified match in either catalogue (81 unmatched and 43 matched, Table S1). Shared acquisitions can have different segmentation masks, expression targets, or registration.

Table S2: Deduplicated 10x Genomics source-acquisition inventory, frozen on 12 October 2026: 53 Visium HD and 82 Xenium acquisitions. Each row is an assay acquisition or ROI, not a donor. Analysis reruns, two sequencing-depth variants, and one spatial demo subset are consolidated. Distinct sections, controls, and assays are retained. Links identify the preferred source page and molecular download. The processed freeze includes 124 of these acquisitions; its counts and QC are reported separately.

ID	Source / sample	Species / tissue	Assay / preservation	Analysis / files
HD-001	Mouse Brain (Fixed Frozen)	Mouse Brain	HD-WT Fixed Frozen	SR 3.1.1 page / data
HD-002	Mouse Ovary (Fresh Frozen)	Mouse Ovary	HD-3prime Fresh Frozen	SR 4.1.0 page / data
HD-003	Human Tonsil (Fresh Frozen)	Human Tonsil	HD-WT Fresh Frozen	SR 3.1.1 page / data
HD-004	Human Breast Cancer (Fresh Frozen), Ultima Sequencing	Human Breast	HD-WT Fresh Frozen	SR 4.0.1 page / data

Continued on next page

Table S2 continued

ID	Source / sample	Species / tissue	Assay / preservation	Analysis / files
HD-005	Cross-Platform Comparison: Visium HD Spatial Gene Expression Library, Human Ovarian Cancer (Fresh Frozen)	Human Ovary	HD-WT Fresh Frozen	SR 4.0.1 page / data
HD-006	Human Brain Cancer (FFPE), 6.5 mm Capture Area	Human Brain	HD-P2 FFPE	SR 4.1.0 page / data
HD-007	Human Tonsil (Fresh Frozen)	Human Tonsil	HD-3prime Fresh Frozen	SR 4.1.0 page / data
HD-008	Mouse Brain (Fresh Frozen)	Mouse Brain	HD-WT Fresh Frozen	SR 3.1.1 page / data
HD-009	Human Lymph Node (Fresh Frozen)	Human Lymph Node	HD-3prime Fresh Frozen	SR 4.1.0 page / data
HD-010	Post-Xenium, Human Lung Cancer (FFPE) / Experiment 2, HD only (control)	Human Lung	HD-WT FFPE	SR 3.0.0 page / data
HD-011	Mouse Brain (FFPE)	Mouse Brain	HD-WT FFPE	SR 4.0.1 page / data
HD-012	Mouse Kidney (Fresh Frozen)	Mouse Kidney	HD-3prime Fresh Frozen	SR 4.0.1 page / data
HD-013	Post-Xenium, Human Lung Cancer (FFPE) / Experiment 1, HD only (control)	Human Lung	HD-WT FFPE	SR 3.0.0 page / data
HD-014	Post-Xenium, Human Lung Cancer (FFPE) / Experiment 1, post-Xenium v1	Human Lung	HD-WT FFPE	SR 3.0.0 page / data
HD-015	Human Tonsil (Fresh Frozen), Ultima Sequencing	Human Tonsil	HD-WT Fresh Frozen	SR 3.1.2 page / data
HD-016	Mouse Brain (FFPE), 6.5 mm Capture Area	Mouse Brain	HD-P2 FFPE	SR 4.1.0 page / data
HD-017	Human Breast Cancer (Fixed Frozen)	Human Breast	HD-WT Fixed Frozen	SR 3.1.1 page / data
HD-018	Mouse Embryo (Fresh Frozen)	Mouse Embryo	HD-3prime Fresh Frozen	SR 4.0.1 page / data
HD-019	Rat Liver (FFPE), 6.5 mm Capture Area	Rat Liver	HD-P2 FFPE	SR 4.1.0 page / data
HD-020	Mouse Embryo (FFPE)	Mouse Whole Embryo	HD-WT FFPE	SR 4.0.1 page / data
HD-021	Human Pancreatic Cancer (Fresh Frozen)	Human Pancreas	HD-3prime Fresh Frozen	SR 4.0.1 page / data
HD-022	Mouse Small Intestine (FFPE)	Mouse Intestine	HD-WT FFPE	SR 3.0.0 page / data
HD-023	Rhesus Macaque Liver (Fresh Frozen)	Rhesus macaque Liver	HD-3prime Fresh Frozen	SR 4.1.0 page / data
HD-024	Human Colon Cancer (FFPE), 6.5 mm Capture Area	Human Colon	HD-P2 FFPE	SR 4.1.0 page / data
HD-025	Human Breast Cancer (Fresh Frozen)	Human Breast	HD-WT Fresh Frozen	SR 3.1.1 page / data
HD-026	Human + Mouse Xenograft (Fresh Frozen)	Mouse; Human Flank (subcutaneous)	HD-3prime Fresh Frozen	SR 4.0.1 page / data
HD-027	Human Pancreas (FFPE)	Human Pancreas	HD-WT FFPE	SR 4.0.1 page / data
HD-028	Human Lung Cancer, IF Stained (FFPE)	Human Lung	HD-WT FFPE	SR 3.0.0 page / data
HD-029	Arabidopsis thaliana (Fresh Frozen)	Arabidopsis thaliana not specified	HD-3prime Fresh Frozen	SR 4.0.1 page / data
HD-030	Mouse Brain (Fresh Frozen)	Mouse Brain	HD-3prime Fresh Frozen	SR 4.0.1 page / data
HD-031	Post-Xenium, Human Lung Cancer (FFPE) / Experiment 2, post-Xenium Prime 5K	Human Lung	HD-WT FFPE	SR 3.0.0 page / data
HD-032	Human Prostate Cancer (FFPE)	Human Prostate	HD-WT FFPE	SR 4.0.1 page / data
HD-033	Human Tonsil, IF (Fresh Frozen)	Human Tonsil	HD-WT Fresh Frozen	SR 3.1.1 page / data
HD-034	Human Ovarian Cancer (Fresh Frozen)	Human Ovary	HD-3prime Fresh Frozen	SR 4.0.1 page / data
HD-035	Human Kidney (FFPE)	Human Kidney	HD-WT FFPE	SR 4.0.1 page / data

Continued on next page

Table S2 continued

ID	Source / sample	Species / tissue	Assay / preservation	Analysis / files
HD-036	Zebrafish Head (Fresh Frozen)	Zebrafish Head	HD-3prime Fresh Frozen	SR 4.0.1 page / data
HD-037	Human Lymph Node (FFPE)	Human Lymph Node	HD-WT FFPE	SR 4.0.1 page / data
HD-038	Mouse Lung (Fresh Frozen)	Mouse Lung	HD-WT Fresh Frozen	SR 3.1.1 page / data
HD-039	Soybean Seed (Fresh Frozen)	Soybean not specified	HD-3prime Fresh Frozen	SR 4.0.1 page / data
HD-040	Rhesus Macaque Kidney (Fresh Frozen)	Rhesus macaque Kidney	HD-3prime Fresh Frozen	SR 4.0.1 page / data
HD-041	Mouse Kidney (FFPE)	Mouse Kidney	HD-WT FFPE	SR 4.0.1 page / data
HD-042	Human Lung Cancer (Fixed Frozen)	Human Lung	HD-WT Fixed Frozen	SR 3.1.1 page / data
HD-043	Human Colorectal Cancer (FFPE)	Human Colon	HD-WT FFPE	SR 4.0.1 page / data
HD-044	Human Breast Cancer, IF (FFPE)	Human Breast	HD-WT FFPE	SR 3.1.2 page / data
HD-045	Cross-Platform Comparison: Visium HD 3' Gene Expression Library, Human Ovarian Cancer (Fresh Frozen)	Human Ovary	HD-3prime Fresh Frozen	SR 4.0.1 page / data
HD-046	Human Heart (FFPE), 6.5 mm Capture Area	Human Heart	HD-P2 FFPE	SR 4.1.0 page / data
HD-047	Mouse Embryo (FFPE), 11 mm Capture Area	Mouse Embryo	HD-P2 FFPE	SR 4.1.0 page / data
HD-048	Human Breast Cancer Tissue Microarray (FFPE), 11 mm Capture Area	Human Breast	HD-P2 FFPE	SR 4.1.0 page / data
HD-049	Human Colon Cancer (FF), 11 mm Capture Area	Human Colon	HD-P2 Fresh Frozen	SR 4.1.0 page / data
HD-050	Human Breast Cancer (Fixed Frozen), 11 mm Capture Area	Human Breast	HD-P2 Fixed Frozen	SR 4.1.0 page / data
HD-051	Human Pancreas (FFPE), 11 mm Capture Area	Human Pancreas	HD-P2 FFPE	SR 4.1.0 page / data
HD-052	Human Lung Cancer, Breast Cancer and Colon Cancer Tissue Array (FFPE), 11 mm Capture Area	Human Lung; Breast; Colon	HD-P2 FFPE	SR 4.1.0 page / data
HD-053	Human Ovarian Cancer (FF) with Immunofluorescent staining, 11 mm Capture Area	Human Ovary	HD-P2 Fresh Frozen	SR 4.1.0 page / data
XE-001	Xenium v1 Human Breast FFPE with Biomarkers & Housekeeping Genes Custom Panel / Section 1, bottom	Human Breast	X-v1 FFPE	XOA 4.0.0 page / data
XE-002	Xenium v1 Human Breast FFPE with Biomarkers & Housekeeping Genes Custom Panel / Section 1, middle	Human Breast	X-v1 FFPE	XOA 4.0.0 page / data
XE-003	Xenium v1 Human Breast FFPE with Biomarkers & Housekeeping Genes Custom Panel / Section 1, top	Human Breast	X-v1 FFPE	XOA 4.0.0 page / data
XE-004	Xenium v1 Human Breast FFPE with Biomarkers & Housekeeping Genes Custom Panel / Section 2, bottom	Human Breast	X-v1 FFPE	XOA 4.0.0 page / data
XE-005	Xenium v1 Human Breast FFPE with Biomarkers & Housekeeping Genes Custom Panel / Section 2, middle	Human Breast	X-v1 FFPE	XOA 4.0.0 page / data
XE-006	Xenium v1 Human Breast FFPE with Biomarkers & Housekeeping Genes Custom Panel / Section 2, top	Human Breast	X-v1 FFPE	XOA 4.0.0 page / data
XE-007	Xenium v1 Human Breast FFPE with Biomarkers & Housekeeping Genes Custom Panel / Section 3, bottom	Human Breast	X-v1 FFPE	XOA 4.0.0 page / data
XE-008	Xenium v1 Human Breast FFPE with Biomarkers & Housekeeping Genes Custom Panel / Section 3, middle	Human Breast	X-v1 FFPE	XOA 4.0.0 page / data

Continued on next page

Table S2 continued

ID	Source / sample	Species / tissue	Assay / preservation	Analysis / files
XE-009	Xenium v1 Human Breast FFPE with Biomarkers & Housekeeping Genes Custom Panel / Section 3, top	Human Breast	X-v1 FFPE	XOA 4.0.0 page / data
XE-010	Xenium v1 Human Breast FFPE with Biomarkers & Housekeeping Genes Custom Panel / Section 4, bottom	Human Breast	X-v1 FFPE	XOA 4.0.0 page / data
XE-011	Xenium v1 Human Breast FFPE with Biomarkers & Housekeeping Genes Custom Panel / Section 4, middle	Human Breast	X-v1 FFPE	XOA 4.0.0 page / data
XE-012	Xenium v1 Human Breast FFPE with Biomarkers & Housekeeping Genes Custom Panel / Section 4, top	Human Breast	X-v1 FFPE	XOA 4.0.0 page / data
XE-013	Xenium Human Lung Preview Data / Lung Cancer	Human Lung	X-v1 FFPE	XOA 1.3.0 page / data
XE-014	Xenium Human Lung Preview Data / Non-diseased Lung	Human Lung	X-v1 FFPE	XOA 1.3.0 page / data
XE-015	FFPE Human Breast Cancer with 5K Human Pan Tissue and Pathways Panel plus 100 Custom Genes / Breast cancer	Human Breast	X-Prime5K FFPE	XOA 3.0.0 page / data
XE-016	FFPE Human Cervical Cancer with 5K Human Pan Tissue and Pathways Panel plus 100 Custom Genes / Cervical cancer	Human Cervix	X-Prime5K FFPE	XOA 3.0.0 page / data
XE-017	Post-Xenium Technical Note: Xenium v1 and Xenium Prime 5K for FFPE Human Lung Cancer / Experiment 2, Xenium Prime 5K	Human Lung	X-Prime5K FFPE	XOA 3.0.0 page / data
XE-018	Preview Data: FFPE Human Lymph Node with 5K Pan Tissue and Pathways Panel / Lymph Node	Human Lymph Node	X-Prime5K FFPE	XOA 3.0.0 page / data
XE-019	Cross-Platform Comparison: FF Human Ovarian Cancer with Xenium v1 and Xenium Prime 5K / Xenium Prime 5K sample	Human Ovary	X-Prime5K Fresh Frozen	XOA 4.0.0 page / data
XE-020	Fresh Frozen Human Ovarian Adenocarcinoma with 5K Human Pan Tissue and Pathways Panel / Ovarian adenocarcinoma	Human Ovary	X-Prime5K Fresh Frozen	XOA 3.0.0 page / data
XE-021	Preview Data: FFPE Human Prostate Adenocarcinoma with 5K Human Pan Tissue and Pathways Panel / Prostate adenocarcinoma	Human Prostate	X-Prime5K FFPE	XOA 3.0.0 page / data
XE-022	Preview Data: FFPE Human Skin Primary Dermal Melanoma with 5K Human Pan Tissue and Pathways Panel / Skin melanoma	Human Skin	X-Prime5K FFPE	XOA 3.0.0 page / data
XE-023	Fresh Frozen Mouse Brain Hemisphere with 5K Mouse Pan Tissue and Pathways Panel / Mouse brain	Mouse Brain	X-Prime5K Fresh Frozen	XOA 3.0.0 page / data
XE-024	Whole Mouse Pup (FFPE) with 5K Mouse Pan Tissue and Pathways Panel and Snap25a/Snap25b Add-on / Whole Mouse Pup	Mouse Whole Organism	X-Prime5K FFPE	XOA 3.0.0 page / data
XE-025	FFPE Human Ovarian Cancer with 5K Human Pan Tissue and Pathways Panel plus 100 Custom Genes / Ovarian cancer	Human Ovary	X-Prime5K FFPE	XOA 3.0.0 page / data
XE-026	Xenium Human Brain Preview Data / Alzheimer's	Human Brain	X-v1 FFPE	XOA 1.3.0 page / data
XE-027	Xenium Human Brain Preview Data / Glioblastoma	Human Brain	X-v1 FFPE	XOA 1.3.0 page / data
XE-028	Xenium Human Brain Preview Data / Healthy Brain	Human Brain	X-v1 FFPE	XOA 1.3.0 page / data
XE-029	FFPE Human Breast with Pre-designed Panel / Tissue sample 1	Human Breast	X-v1 FFPE	XOA 1.0.2 page / data
XE-030	FFPE Human Breast using the Entire Sample Area / Replicate 1	Human Breast	X-v1 FFPE	XOA 1.0.2 page / data

Continued on next page

Table S2 continued

ID	Source / sample	Species / tissue	Assay / preservation	Analysis / files
XE-031	FFPE Human Breast using the Entire Sample Area / Replicate 2	Human Breast	X-v1 FFPE	XOA 1.0.2 page / data
XE-032	Xenium FFPE Human Breast with Custom Add-on Panel / Tissue sample 1 (IDC)	Human Breast	X-v1 FFPE	XOA 1.3.0 page / data
XE-033	FFPE Human Breast with Pre-designed Panel / Tissue sample 2	Human Breast	X-v1 FFPE	XOA 1.0.2 page / data
XE-034	FFPE Human Breast with Custom Add-on Panel / Tissue sample 2	Human Breast	X-v1 FFPE	XOA 1.0.2 page / data
XE-035	Xenium In Situ Analysis of Alzheimer's Disease Mouse Model Brain Coronal Sections from One Hemisphere Over a Time Course / TgCRND8 17.9 months	Mouse Brain	X-v1 FFPE	XOA 1.4.0 page / data
XE-036	Xenium In Situ Analysis of Alzheimer's Disease Mouse Model Brain Coronal Sections from One Hemisphere Over a Time Course / TgCRND8 2.5 months	Mouse Brain	X-v1 FFPE	XOA 1.4.0 page / data
XE-037	Xenium In Situ Analysis of Alzheimer's Disease Mouse Model Brain Coronal Sections from One Hemisphere Over a Time Course / TgCRND8 5.7 months	Mouse Brain	X-v1 FFPE	XOA 1.4.0 page / data
XE-038	Xenium In Situ Analysis of Alzheimer's Disease Mouse Model Brain Coronal Sections from One Hemisphere Over a Time Course / Wild type 13.4 months	Mouse Brain	X-v1 FFPE	XOA 1.4.0 page / data
XE-039	Xenium In Situ Analysis of Alzheimer's Disease Mouse Model Brain Coronal Sections from One Hemisphere Over a Time Course / Wild type 2.5 months	Mouse Brain	X-v1 FFPE	XOA 1.4.0 page / data
XE-040	Xenium In Situ Analysis of Alzheimer's Disease Mouse Model Brain Coronal Sections from One Hemisphere Over a Time Course / Wild type 5.7 months	Mouse Brain	X-v1 FFPE	XOA 1.4.0 page / data
XE-041	Fresh Frozen Mouse Brain for Xenium Explorer Demo / Full coronal section	Mouse Brain	X-v1 Fresh Frozen	XOA 1.0.2 page / data
XE-042	Fresh Frozen Mouse Brain Replicates / Replicate 1	Mouse Brain	X-v1 Fresh Frozen	XOA 1.0.2 page / data
XE-043	Fresh Frozen Mouse Brain Replicates / Replicate 2	Mouse Brain	X-v1 Fresh Frozen	XOA 1.0.2 page / data
XE-044	Fresh Frozen Mouse Brain Replicates / Replicate 3	Mouse Brain	X-v1 Fresh Frozen	XOA 1.0.2 page / data
XE-045	FFPE Human Brain Cancer Data with Human Immuno-Oncology Profiling Panel and Custom Add-on / Glioblastoma Multiforme	Human Brain	X-v1 FFPE	XOA 2.0.0 page / data
XE-046	Human FFPE Clear Cell Renal Cell Carcinoma (ccRCC) / Clear Cell Renal Cell Carcinoma	Human Kidney	X-RNA-protein FFPE	XOA 4.0.0 page / data
XE-047	FFPE Human Colorectal Cancer Data with Human Immuno-Oncology Profiling Panel and Custom Add-on / Colorectal cancer	Human Colon	X-v1 FFPE	XOA 2.0.0 page / data
XE-048	FFPE Human Pancreatic Ductal Adenocarcinoma Data with Human Immuno-Oncology Profiling Panel / Ductal adenocarcinoma	Human Pancreas	X-v1 FFPE	XOA 2.0.0 page / data
XE-049	FFPE Human Renal Cell Carcinoma / Renal cell carcinoma	Human Kidney	X-RNA-protein FFPE	XOA 4.0.0 page / data
XE-050	FFPE Human Lung Cancer Data with Human Immuno-Oncology Profiling Panel and Custom Add-on / Lung cancer	Human Lung	X-v1 FFPE	XOA 2.0.0 page / data
XE-051	Post-Xenium Technical Note: Xenium v1 and Xenium Prime 5K for FFPE Human Lung Cancer / Experiment 1, Xenium v1	Human Lung	X-v1 FFPE	XOA 3.0.0 page / data

Continued on next page

Table S2 continued

ID	Source / sample	Species / tissue	Assay / preservation	Analysis / files
XE-052	FFPE Human Ovarian Cancer Data with Human Immuno-Oncology Profiling Panel and Custom Add-on / Ovarian cancer	Human Ovary	X-v1 FFPE	XOA 2.0.0 page / data
XE-053	Cross-Platform Comparison: FF Human Ovarian Cancer with Xenium v1 and Xenium Prime 5K / Xenium v1 sample	Human Ovary	X-v1 Fresh Frozen	XOA 4.0.0 page / data
XE-054	Human Bone and Bone Marrow Data with Custom Add-on Panel / Acute Lymphoid Leukemia Bone Marrow	Human Bone; Bone Marrow	X-v1 FFPE	XOA 1.9.0 page / data
XE-055	Human Bone and Bone Marrow Data with Custom Add-on Panel / Non-diseased Bone Marrow	Human Bone; Bone Marrow	X-v1 FFPE	XOA 1.9.0 page / data
XE-056	Human Bone and Bone Marrow Data with Custom Add-on Panel / Non-diseased Femur Bone	Human Bone; Bone Marrow	X-v1 FFPE	XOA 1.9.0 page / data
XE-057	Human Colon Preview Data (Xenium Human Colon Gene Expression Panel) / Cancer, pre-designed + add-on panel	Human Intestine (colon)	X-v1 FFPE	XOA 1.6.0 page / data
XE-058	Human Colon Preview Data (Xenium Human Colon Gene Expression Panel) / Cancer, pre-designed panel	Human Intestine (colon)	X-v1 FFPE	XOA 1.6.0 page / data
XE-059	Human Colon Preview Data (Xenium Human Colon Gene Expression Panel) / Non-diseased, pre-designed + add-on panel	Human Intestine (colon)	X-v1 FFPE	XOA 1.6.0 page / data
XE-060	Human Colon Preview Data (Xenium Human Colon Gene Expression Panel) / Non-diseased, pre-designed panel	Human Intestine (colon)	X-v1 FFPE	XOA 1.6.0 page / data
XE-061	Human Heart Data with Xenium Human Multi-Tissue and Cancer Panel / Non-diseased	Human Heart	X-v1 FFPE	XOA 1.9.0 page / data
XE-062	Human Kidney Preview Data (Xenium Human Multi-Tissue and Cancer Panel) / Kidney Cancer (PRCC)	Human Kidney	X-v1 FFPE	XOA 1.5.0 page / data
XE-063	Human Kidney Preview Data (Xenium Human Multi-Tissue and Cancer Panel) / Non-diseased Kidney	Human Kidney	X-v1 FFPE	XOA 1.5.0 page / data
XE-064	Human Liver Data with Xenium Human Multi-Tissue and Cancer Panel / Liver cancer	Human Liver	X-v1 FFPE	XOA 1.9.0 page / data
XE-065	Human Liver Data with Xenium Human Multi-Tissue and Cancer Panel / Non-diseased	Human Liver	X-v1 FFPE	XOA 1.9.0 page / data
XE-066	Human Lung Cancer Preview Data (Xenium Human Multi-Tissue and Cancer Panel) / Lung Cancer (Carcinoma)	Human Lung	X-v1 FFPE	XOA 1.5.0 page / data
XE-067	Human Lymph Node Preview Data (Xenium Human Multi-Tissue and Cancer Panel) / Non-diseased Lymph Node	Human Lymph Node	X-v1 FFPE	XOA 1.5.0 page / data
XE-068	Pancreatic Cancer with Xenium Human Multi-Tissue and Cancer Panel / Adenocarcinoma	Human Pancreas	X-v1 FFPE	XOA 1.6.0 page / data
XE-069	Human Pancreas Preview Data (Xenium Human Multi-Tissue and Cancer Panel) / Non-diseased Pancreas	Human Pancreas	X-v1 FFPE	XOA 1.5.0 page / data
XE-070	Human Skin Preview Data (Xenium Human Skin Gene Expression Panel) / Melanoma, pre-designed panel	Human Skin	X-v1 FFPE	XOA 1.6.0 page / data
XE-071	Human Skin Data with Xenium Human Multi-Tissue and Cancer Panel / Sample 1	Human Skin	X-v1 FFPE	XOA 1.9.0 page / data
XE-072	Human Skin Data with Xenium Human Multi-Tissue and Cancer Panel / Sample 2	Human Skin	X-v1 FFPE	XOA 1.9.0 page / data
XE-073	Human Tonsil Data with Xenium Human Multi-Tissue and Cancer Panel / Follicular lymphoid hyperplasia	Human Tonsil	X-v1 FFPE	XOA 1.9.0 page / data

Continued on next page

Table S2 continued

ID	Source / sample	Species / tissue	Assay / preservation	Analysis / files
XE-074	Human Tonsil Data with Xenium Human Multi-Tissue and Cancer Panel / Reactive follicular hyperplasia	Human Tonsil	X-v1 FFPE	XOA 1.9.0 page / data
XE-075	Preview Data: FFPE Human Lung Cancer with Xenium Multimodal Cell Segmentation / Lung Cancer (Adenocarcinoma)	Human Lung	X-v1 FFPE	XOA 2.0.0 page / data
XE-076	FFPE Human Pancreas with Xenium Multimodal Cell Segmentation / Pancreas	Human Pancreas	X-v1 FFPE	XOA 2.0.0 page / data
XE-077	Mouse Bone Data with Custom Add-on Panel / 0.5M EDTA decalcification	Mouse Bone	X-v1 FFPE	XOA 1.9.0 page / data
XE-078	Mouse Bone Data with Custom Add-on Panel / 15% EDTA/0.4% PFA decalcification	Mouse Bone	X-v1 FFPE	XOA 1.9.0 page / data
XE-079	Mouse Bone Data with Custom Add-on Panel / 8% Formic acid decalcification	Mouse Bone	X-v1 FFPE	XOA 1.9.0 page / data
XE-080	Fresh Frozen Mouse Colon with Xenium Multimodal Cell Segmentation / Mouse	Mouse Colon	X-v1 Fresh Frozen	XOA 2.0.0 page / data
XE-081	Whole Mouse Pup Preview Data (Xenium Mouse Tissue Atlassing Panel) / Mouse Pup	Mouse Whole Organism	X-v1 FFPE	XOA 1.6.0 page / data
XE-082	Human Skin Preview Data (Xenium Human Skin Gene Expression Panel with Custom Add-On) / Melanoma, pre-designed panel with add-on	Human Skin	X-v1 FFPE	XOA 1.7.0 page / data

B.4 Release-wide quality control

Snapshot and observation units. The corrected 25 October 2026 snapshot contains **53 Visium HD and 71 Xenium acquisitions**, corresponding to processed sections, capture areas, or ROIs rather than verified independent donors or studies. The 135-entry source catalogue (Table S2) is a separate download inventory. HD H&E targets are cell-aligned; HD IF targets are nucleus-aligned. All Xenium expression observations are native *cells*, including those displayed against DAPI. The rounded sMCC40M name does not replace exact counts.

Table S3: Audited release census. Candidates include objects without a stored expression row. A matrix row has a positive count for at least one typed feature; a gene-positive row has a positive Gene Expression count. Cells and nuclei are separate observation types.

Branch	Acq.	Candidates	Matrix rows	Gene-positive
HD cells, H&E	49	18,765,683	11,619,603	11,619,603
HD nuclei, IF	4	2,635,758	2,041,153	2,041,153
Xenium cells, H&E	56	21,132,812	21,078,920	21,064,472
Xenium cells, DAPI	15	3,750,373	3,746,978	3,746,964
Total	124	46,284,626	38,486,654	38,472,192

Of the 38,486,654 nonzero-feature matrix rows, 36,445,501 are cell observations and 2,041,153 are HD nuclei. There are 38,472,192 gene-positive rows; 14,462 Xenium rows contain only other typed features. Another 57,287 measured-zero Xenium candidates retain metadata with null matrix-row links. HD candidates include segmentation objects outside capture support. Candidates, molecular observations, visible labels, and image-eligible pairs have different denominators.

Audit scope and units. The release-wide audit scanned all H5AD CSR matrices, cell/gene indices and selected-image masks, and all 729,837,032 rows of the 53 HD bin-to-cell tables. Checks cover integer counts, feature types/order, row links, labels, image axes, coordinate transforms, pixel counts, and assignment status. Under the strict majority-area rule (Section C.2), each 2 μm bin is assigned at most once, only when a cell/nucleus covers more than half of the *complete* bin. This audit checks stored assignments; original raw-count conservation and mask–bin intersections are separate construction checks. Native distances are in μm, areas in μm^2, and image coordinates in level-zero pixels using per-axis physical pixel sizes.

Reproducibility. Per-acquisition metrics, feature definitions, and reproducible plotting scripts are provided with the accompanying code and data. The analyses cover the 124 public Visium HD and Xenium acquisitions and report cell and nucleus geometry separately.

B.4.1 Sensitivity of Visium HD expression to polygon boundaries

We analyzed 9,000 CellViT polygons from three Visium HD samples (3,000 per sample). We shifted the contours by $\pm 1\ \mu$m relative to the native $2\ \mu$m bin grid and recomputed gene counts from intersecting bins, assigning contested bins to the nearest centroid. The table reports the fraction of all polygons with at least one bin before perturbation. For supported polygons, we compared bin membership (Jaccard index), gene expression profiles (cosine similarity), and total UMI counts with their unshifted values.

Sample	Bin support	Bin Jaccard	Expression cosine	Median $\lvert\Delta\mathrm{UMI}\rvert$ (%)
Human lung cancer	49.4%	0.727–0.733	0.954–0.957	11.1–11.3
Mouse brain	97.5%	0.806–0.816	0.936–0.939	6.0–6.1
Human pancreas	50.0%	0.706–0.714	0.994	13.0–13.7

Expression profiles remained similar after translation even as bin membership changed. Eroding or dilating the masks had a larger effect on total counts (median absolute UMI change, 32.8–66.6%) and yielded bin Jaccard indices of 0.462–0.720 and expression cosine similarities of 0.871–0.993.

B.4.2 Molecular support and measured panels

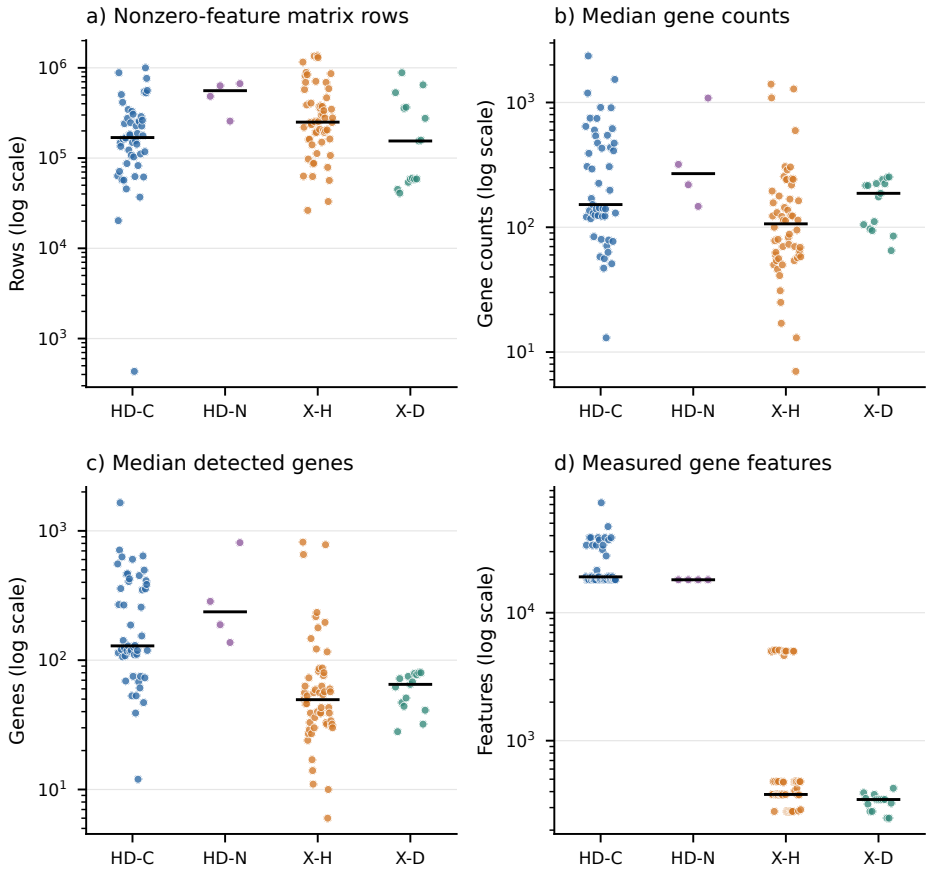


Figure S2: **Molecular support across all 124 acquisitions.** Each dot is one acquisition; black segments are medians across acquisition summaries, not pooled cell medians or confidence intervals. Branches: HD-C, HD cells/H&E ($n = 49$); HD-N, HD nuclei/IF ($n = 4$); X-H, Xenium cells/H&E ($n = 56$); X-D, Xenium cells/DAPI ($n = 15$). All axes are logarithmic. (a) Stored nonzero-feature matrix rows. (b,c) Median Gene Expression counts and detected genes over those rows, including gene-zero rows. (d) Measured Gene Expression features.

HD counts are aggregated UMIs; Xenium counts are gene-assigned transcripts. Controls, unassigned-codeword features, and protein features remain separately typed and are excluded from gene-only summaries. Different measured panels are not zero-padded into a union. Mitochondrial fractions are unavailable when the recognized mitochondrial features are absent from a panel, rather than treated as zero contamination. Species, panel size, and assay chemistry differ, so these distributions describe the release rather than rank assay sensitivity. The full CSVs additionally provide gene-positive counts, upper percentiles, feature-type counts, and detected/measured gene fractions.

B.4.3 Image coverage and delivered-mask geometry

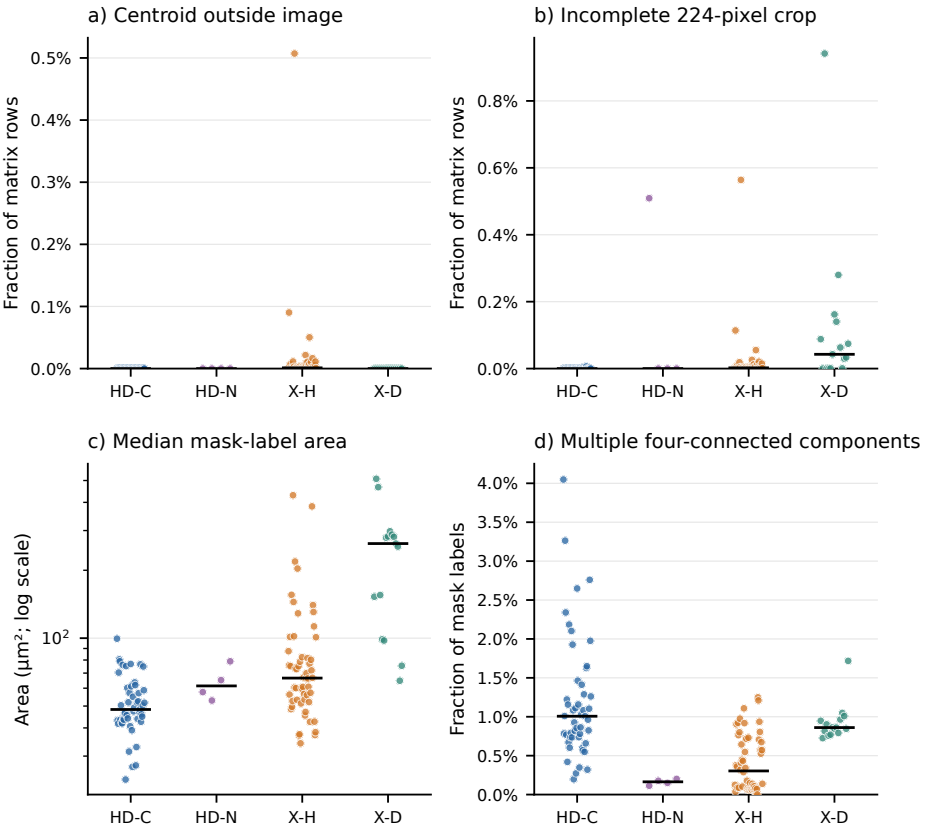


Figure S3: **Spatial support and categorical-label geometry after correction.** Dots, medians, and branch codes follow Figure S2. (a,b) Fractions of all nonzero-feature matrix rows with centroids outside the selected image or incomplete centered 224 × 224-pixel windows. (c) Median area of observed nonzero mask labels, using physical pixel sizes (logarithmic axis). (d) Fraction of labels with multiple four-connected components. These are coverage and raster-geometry diagnostics.

Coverage uses TIFF Y/X dimensions and a continuous centered window wholly contained within the image extent. A fixed pixel window spans different physical sizes across acquisitions; 50- and 100 μm windows and missing-mask counts are also in the full CSV. Boundary containment does not establish tissue content, histological quality, or independently measured registration accuracy.

Mask statistics retain original labels, holes, and disconnected components; four-connected components are merged exactly across processing-tile boundaries. Diagonal contact remains disconnected. The fragmentation fraction is neither an oversegmentation error rate nor a spindle-cell recall estimate. Xenium display masks can change area through resampling or clipping without reassigning native transcripts. Plant and zebrafish segmentation domains remain exploratory.

B.4.4 Separate native Xenium cell and nucleus consistency

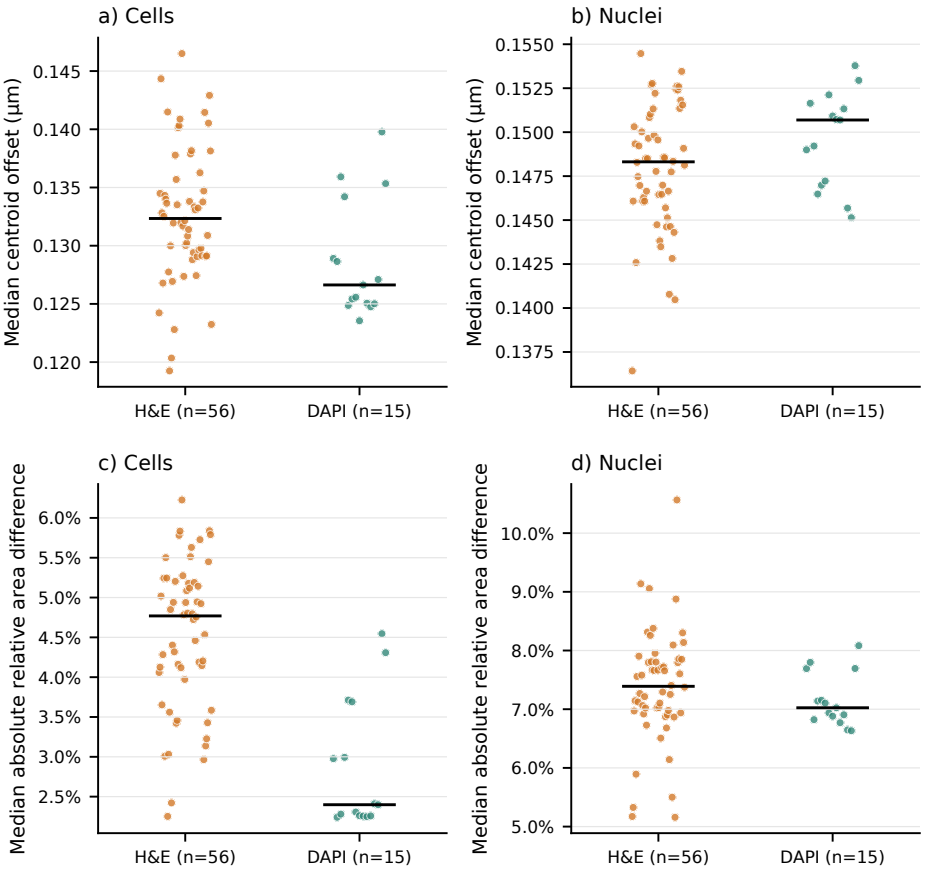
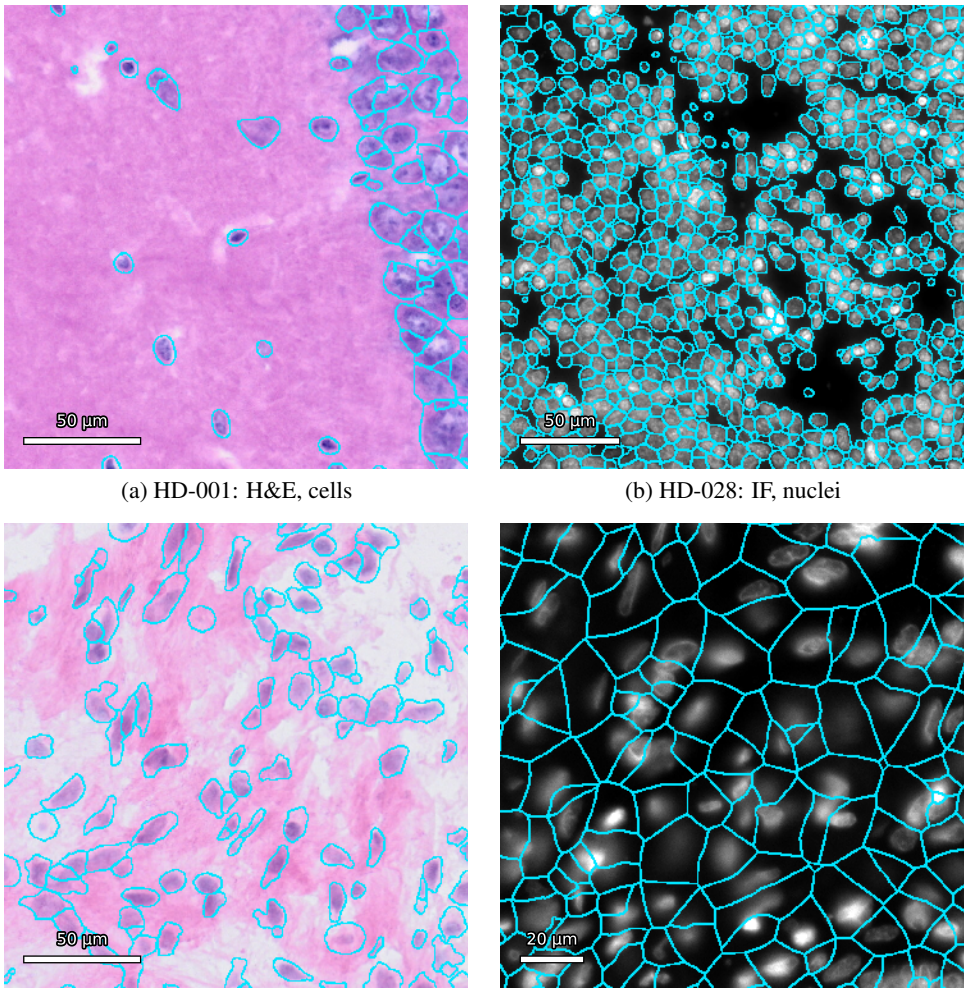


Figure S4: **Native polygon/summary consistency in 71 Xenium acquisitions.** Cells and nuclei are measured separately against their corresponding native summary centroid and area. Each dot is an acquisition summary; black segments are medians across acquisitions. Upper panels show median centroid differences in μm, lower panels median absolute relative area differences. H&E and DAPI identify display branches; all expression assignments remain cell-based.

Geometry	Sampled candidates	Eligible measurements	Median offset range (μm)
Cell	355,000	354,717	0.119–0.147
Nucleus reference	355,000	346,185	0.136–0.154

For each acquisition, 5,000 native candidates are sampled uniformly before expression filtering (seed 20261024). Polygon centroids use area rather than vertex means. Missing/invalid geometry and unsupported multiple-nucleus reference scopes are recorded separately. The nucleus-reference denominator is associated with sampled native cell entries; it is not a census of distinct nuclei. These are internal consistency checks, not independent H&E registration errors or biological boundary accuracy; Dice/PQ is not inferred from them.

B.4.5 Fixed-region image inspection across four branches



(a) HD-001: H&E, cells

(b) HD-028: IF, nuclei

(c) FFHO026: H&E, cells

(d) XE-013: DAPI, cells

Figure S5: **Deterministically selected local overlays.** The first lexicographic acquisition in each branch is shown. Within each, expression- available centroids supporting a complete 768×768-pixel window are sorted by (y, x, label) and the lower median is selected. No appearance or performance ranking is used. Cyan shows delivered label boundaries, expanded by one image pixel in each direction for visibility only; segmentation and QC values are unchanged. Bars: 50 μm in (a–c), 20 μm in (d). IF/DAPI intensities use per-patch percentile clipping.

Only the original overlay panels are cropped for this layout, without resampling the images or changing their boundaries. All eight original paired views (first and last acquisitions per branch), region provenance, and selection records remain in the accompanying sources. These are qualitative inspections, not independently annotated segmentation assessments. DAPI cell boundaries are not nucleus labels; native cell segmentation can include nucleus expansion and is not independent whole-cell ground truth.

B.5 Dataset construction and alignment pipeline

The frozen construction comprises source and metadata auditing, coordinate registration, segmentation or platform-mask ingestion, expression aggregation or native-count retention, and separate image/crop eligibility assessment. Its expression units are cells or nuclei, explicitly typed by branch. Images, labels, coordinates, and counts support crop generation on demand. Benchmark partitions

and optional caption or communication analyses are derived artifacts with their own coverage and protocols.

B.6 Data schema and release package

The versioned package links H5AD raw-count matrices to Parquet cell/nucleus indices, sample context, source evidence, and eligibility flags through stable sample and observation identifiers. Image references and physical scales support reproducible on-demand crops. A separate benchmark package retains exact observation IDs, measured-gene targets, fold memberships, and evaluation settings for the protocols in Section C.

Table S4: Main artifact types in the sMCC40M benchmark release.

Artifact	Format	Granularity	Required	Contents
Images and label masks	TIFF	acquisition	yes	selected images, transforms, and labels
Expression matrices	H5AD	cell/nucleus × feature	yes	integer CSR counts; original feature IDs/types
Metadata and evidence	Parquet / JSON	observation or sample	yes	coordinates, context, source URLs, evidence status
Image eligibility	Parquet	candidate object	yes	RNA support, unit, image pairing, crop completeness
Benchmark package	Python / Parquet / JSON / NPZ	protocol or fold	yes	exact IDs, genes, splits, environments, scoring
Cell-type annotations	H5AD / Parquet	observation	partial	predicted labels and annotation evidence
Crops, captions and communication	protocol-specific	cell or neighborhood	optional	analysis-specific; coverage varies

C Current benchmark protocols and full results

C.1 Xenium-only categories in the 23-category benchmark

This section consolidates frozen CONCH–ridge benchmarks by platform and evaluation regime. Acquisitions are sections/capture areas, not verified independent donors. Gene P/S denote gene-wise Pearson/Spearman, Cell P is Pearson across genes within a cell, Coord/KNN P are gene Pearson for spatial controls, and RMSE is on the log1p-count scale. Detection F1 uses count > 0 as truth and prediction $> \log(1.5)$ as positive. Undefined correlations are `U`; unavailable or superseded results are `NA`. Negative scores are retained. The current evaluation comprises the 18-category HD benchmark, five additional Xenium categories, and HD cross-acquisition validation. Code, fold plans, and numerical results are provided through https://github.com/aicb-ZhangLabs/sMCC40M.

Updated Xenium-only extension. The five categories absent from HD18 contribute 13 current H&E acquisitions and 156,000 evaluated cells. They use the HD-style 20-by-20 spatial sampler, four edge folds, 50 training-only HVGs, and identical predictors/metrics. Only gene-positive rows with centroids inside the selected image are eligible; partially clipped 150-pixel crops retain white padding. Gene identities and raw integer targets are checked against the typed release axes.

Table S5: Updated native-Xenium extension: all 13 eligible H&E acquisitions in the five categories added to HD18. †Xenium data only.

Category	n	Gene P	Coord P	KNN P	Gene S	Cell P	F1	R^2	RMSE
Bone†	4	0.417	0.058	0.089	0.380	0.524	0.480	0.157	0.538
Bone Marrow†	1	0.116	0.009	0.017	0.101	0.249	0.022	-0.012	0.221
Cervix†	1	0.308	0.042	0.041	0.268	0.368	0.231	0.072	0.329
Skin†	5	0.402	0.060	0.102	0.347	0.476	0.464	-0.460	0.515
Whole Organism†	2	0.274	0.038	0.056	0.218	0.477	0.249	-0.517	0.393

HVGs are shared within organ/species using the intersection of measured gene IDs and equal-acquisition mean training variance. Skin is the sole panel exception: its five acquisitions share only 47 measured genes. To retain all acquisitions and exactly 50 targets, Skin HVGs are selected within four exact measured-panel groups. Unmeasured genes are not zero-filled. This metadata-based adaptation

was fixed before evaluation; it is not score-based model selection. No new confidence intervals were computed for these five categories or the combined 23-category macro.

C.2 HD within- and cross-acquisition benchmark

Expanded cohort and within-acquisition evaluation. Of 53 deduplicated HD acquisitions, 49 have H&E/cell segmentations; the four IF/nucleus-only acquisitions (HD-028/033/044/053) are excluded from the H&E encoder experiment. The primary table has 48 acquisitions in 18 reporting categories. HD-052 is a mixed lung/breast/colon TMA without cell-level organ labels and is evaluated separately. The categories include Embryo, Head, Xenograft, Arabidopsis, and Soybean rather than 18 distinct anatomical organs. Whole Embryo maps to Embryo, Intestine to Small Intestine, and flank xenograft to Xenograft. Plant labels do not imply seed-specific anatomy; plant and zebrafish segmentation domains remain exploratory.

The 20-by-20 sampler uses seed 20260723 plus the frozen acquisition index. The primary experiment evaluates 564,433 sampled cells; Arabidopsis contributes all 433 positive-expression cells, and the other acquisitions contribute 12,000 each. Four directional edge tests (high-x, high-y, low-x, low-y) hold out 20% with a 5%-span training buffer. Fifty HVGs maximize equal-acquisition mean training-only log1p variance over organ/species measured-gene intersections, with gene-ID tie breaks. Table S6 reports all six image-prediction metrics and gene Pearson for both spatial controls; their full metrics and uncertainty bounds remain in the CSVs.

Table S6: All 18 HD tissue categories (48 primary H&E acquisitions). Gene P includes approximate marginal 95% conditional intervals where supported; the other columns are point estimates. ‡Ovary has limited block support; Arabidopsis intervals are unavailable.

Category	n	Gene P [95% CI]	Coord P	KNN P	Gene S	Cell P	F1	R^2	RMSE
Brain	6	0.505 [0.388, 0.591]	0.076	0.072	0.492	0.471	0.617	0.238	0.575
Breast	5	0.395 [0.292, 0.511]	0.074	0.057	0.378	0.613	0.503	0.160	0.605
Colon	3	0.424 [0.391, 0.452]	0.110	0.116	0.400	0.526	0.472	0.042	0.505
Embryo	3	0.320 [0.195, 0.391]	0.041	0.060	0.273	0.519	0.418	-0.303	0.496
Head	1	0.088 [0.020, 0.109]	-0.012	-0.001	0.057	0.131	0.023	-1.891	0.138
Heart	1	0.372 [0.352, 0.390]	0.111	0.081	0.312	0.446	0.287	0.097	0.403
Kidney	4	0.368 [0.318, 0.415]	0.061	0.045	0.347	0.545	0.374	0.127	0.461
Liver	2	0.477 [0.462, 0.490]	0.008	0.023	0.466	0.627	0.781	0.226	0.622
Lung	6	0.432 [0.372, 0.478]	0.009	0.058	0.370	0.497	0.479	0.140	0.494
Lymph Node	2	0.151 [0.128, 0.167]	0.015	0.018	0.140	0.334	0.036	-0.001	0.232
Ovary‡	4	0.393 [0.245, 0.535]	0.021	0.036	0.385	0.550	0.517	0.130	0.524
Pancreas	3	0.403 [0.248, 0.495]	0.101	0.092	0.368	0.724	0.528	0.117	0.544
Arabidopsis	1	0.388 NA	-0.033	-0.013	0.376	0.233	0.519	-1.051	0.716
Soybean	1	0.142 [0.129, 0.153]	0.057	0.033	0.128	0.251	0.026	-0.133	0.218
Prostate	1	0.260 [0.240, 0.274]	0.033	0.065	0.236	0.292	0.090	-0.174	0.257
Small Intestine	1	0.509 [0.498, 0.518]	0.047	0.024	0.481	0.537	0.622	0.246	0.526
Tonsil	3	0.282 [0.219, 0.320]	0.023	0.040	0.252	0.510	0.382	0.047	0.530
Xenograft	1	0.655 [0.591, 0.682]	0.171	0.198	0.538	0.196	0.731	0.364	0.402

CONCH–ridge exceeds both spatial controls in gene Pearson in all 18 categories, but the HD18 mean R^2 remains negative. Correlation therefore does not imply calibrated absolute expression, biological segmentation accuracy, or generalization to unseen donors. The interval scopes and full-control summaries follow below.

Cross-sample evaluation within organ and species. Ten organ/species strata support 34 leave-one-acquisition-out fits. Every fit holds out all 12,000 sampled cells of one acquisition. Other acquisitions in the stratum supply training data; scaling, PCA and ridge are refitted, and 50 HVGs are independently selected from training-acquisition log1p-count variance before target expression is loaded. Gene-ID universes use measured panel metadata only. Liver cannot support this same-species test because its two acquisitions are from different species; singleton species strata, IF acquisitions, and mixed HD-052 are excluded.

Table S7: HD whole-acquisition holdouts: all 34 targets in ten eligible organ/species strata. H/M denote human/mouse. Gene-Pearson brackets are fixed-cohort test-tissue spatial intervals; NA denotes unsupported intervals. All scores, including extreme negative R^2, are retained.

Category (species)	n	Gene P [95% CI]	Gene S	Cell P	F1	R^2	RMSE
Brain (M)	5	0.462 [0.452, 0.473]	0.455	0.420	0.642	-0.471	0.701
Breast (H)	5	0.252 NA	0.247	0.411	0.422	-6.440	0.825
Colon (H)	3	0.317 NA	0.313	0.396	0.368	-3.983	0.580
Embryo (M)	3	0.226 [0.217, 0.234]	0.210	0.363	0.406	-4.619	0.600
Kidney (M)	2	0.233 [0.221, 0.245]	0.230	0.267	0.385	-0.784	0.697
Lung (H)	5	0.424 NA	0.362	0.361	0.439	-1.302	0.483
Lymph Node (H)	2	0.064 [0.058, 0.069]	0.064	0.232	0.023	-1.723	0.272
Ovary (H)	3	0.343 [0.333, 0.352]	0.337	0.476	0.553	-4.024	0.775
Pancreas (H)	3	0.234 NA	0.220	0.478	0.427	-1400.866	1.118
Tonsil (H)	3	0.048 [0.037, 0.058]	0.038	0.122	0.105	-0.904	0.759
10-category mean	34	0.260	0.248	0.353	0.377	-142.512	0.681

Cross-sample scoring uses float64 to avoid false variance from float32 centering; constant targets/predictions retain undefined statistics. Table brackets resample only the fixed held-out test tissue, not acquisitions or fitted models. The unusually negative Pancreas R^2 is retained and indicates poor absolute calibration. Scores from the original HD18 table cannot be subtracted as a matched transfer effect because genes, strata and test cells differ.

Table S8: Matched HD comparison over ten strata: both predictors are evaluated on the same externally selected genes and edge-test cells. This test differs from the whole-acquisition holdouts above and from the original HD18 target selection.

Regime	Gene P	Gene S	Cell P	F1	R^2	RMSE
Within acquisition	0.354	0.328	0.563	0.407	0.059	0.480
Other acquisitions	0.248	0.234	0.349	0.370	-26.919	0.676

The matched comparison instead uses each holdout's externally selected genes and the same four edge-test cell sets for within- and cross-acquisition fits. Its gene Pearson changes from 0.354 to 0.248; training-data amounts also differ, so this is not an isolated causal effect of domain shift. Whole-acquisition coordinates have no common anatomical frame, so within-slide coordinate/KNN controls are not reused as cross-slide baselines. A related-source sensitivity holds out all four linked lung control/post-Xenium acquisitions together; lung gene Pearson decreases from 0.424 to 0.228. Source pages do not establish donor independence; all source-page-out metrics and support flags are archived.

Predictors, targets, uncertainty, and the combined coverage summary. HD RNA targets aggregate raw integer UMIs from the official filtered 2 μm bins using registered Vitamin-P footprints. Each bin contributes at most once and only when a cell covers strictly more than half of its *complete* area; holes and disconnected components are retained. These are segmentation-dependent cell-aligned targets, not independent cell-boundary ground truth. Counts are transformed with log1p without

library-size normalization. The frozen CONCH v1 encoder uses 150-by-150 native-pixel RGB crops, white edge padding, its official 448-pixel validation transform, and 512-dimensional features before contrastive projection, without normalization. Physical fields of view vary by acquisition. The head is StandardScaler, randomized PCA(128), and ridge $\alpha = 10$; spatial controls use quadratic coordinates or the mean of eight nearest training cells. No target-based feature fitting or post-prediction clipping is applied.

Table S9: Equal-category HD18 macro averages for every predictor. U denotes intrinsically undefined correlation. No formal HD18 macro interval is reported because it includes the spatially unsupported Arabidopsis category.

Method	Gene P	Gene S	Cell P	F1	R^2	RMSE
CONCH + ridge	0.365	0.333	0.445	0.411	-0.090	0.458
Coordinate + ridge	0.051	0.046	0.346	0.257	-0.706	0.586
Spatial KNN	0.056	0.053	0.329	0.287	-0.768	0.584
Training mean	U	U	0.400	0.224	-0.160	0.543

Table S10: Descriptive coverage summary of the HD18 and updated Xenium-only extension. All 23 categories are equally weighted using unrounded estimates. Shared organs retain their original HD results; the five added categories use only Xenium. No combined confidence interval is claimed.

Coverage	Gene P	Coord P	KNN P	Gene S	Cell P	F1	R^2	RMSE
HD18 (48 acquisitions)	0.365	0.051	0.056	0.333	0.445	0.411	-0.090	0.458
Xenium-only 5 (13)	0.303	0.042	0.061	0.263	0.419	0.289	-0.152	0.399
Combined 23 (61)	0.351	0.049	0.057	0.318	0.439	0.385	-0.103	0.445

The combined 23-category mean uses the original 18 HD rows plus the five updated Xenium-only rows; it is not a pooled-cell or 61-acquisition mean. HD aggregated UMIs, native Xenium transcripts, species-specific genes, and panel-dependent targets differ. The macro describes coverage across these conditions and is not a controlled comparison of assay sensitivity.

Conditional intervals and reproducibility. Within-acquisition HD intervals use 10,000 percentile-bootstrap draws (seed 20261019) with 500 μm spatial blocks anchored at physical origin. Blocks in the union of the four tests receive shared multiplicities across folds and methods; Spearman midranks are recomputed. Acquisitions are resampled within species, preserving equal acquisition weights. Fewer than 20 occupied or effective blocks flags limited support; single-acquisition categories with fewer than 10 suppress their main interval. At least 9,900 finite replicates are required. Ovary is limited and Arabidopsis unsupported; the HD18 and 23-category macros have no formal interval. Cross-sample intervals use fixed-cohort held-out-tissue resampling, with the original predictors/HVGs held fixed, and suppress unsupported bounds. Acquisition SD is descriptive and is not a confidence interval. Neither design captures refitting, gene selection, segmentation, new cell sampling or new-donor uncertainty. Full six-metric bounds, paired contrasts, 250/1,000 μm sensitivity, valid-replicate counts, and execution records accompany the results and data-side provenance described in the repository documentation.

The 49 HD acquisitions produced 196 fold predictions and 784 fold/method summaries; feature extraction used 6.90 GPU-hours including storage waits. The new Xenium extension produced 52 prediction packets and 208 summaries; 1,248 independently recomputed metrics and 36 HVG selections passed checks. All primary HD point estimates remain unchanged. Cohort/axis mappings, unrounded values, all control outputs, checkpoints/software identifiers, seeds and audit hashes accompany the tables; upstream images and model weights are not redistributed here.

D Supplementary visualizations and downstream analyses

D.1 Illustrative spatial-expression maps on TCGA histology

Figure S6 illustrates how predicted expression can be displayed alongside whole-slide histology. These panels are retained as a visualization example from the earlier STBoost workflow, not as

quantitative validation of the current CONCH–ridge benchmark. Paired spatial-expression ground truth is unavailable for this illustration.

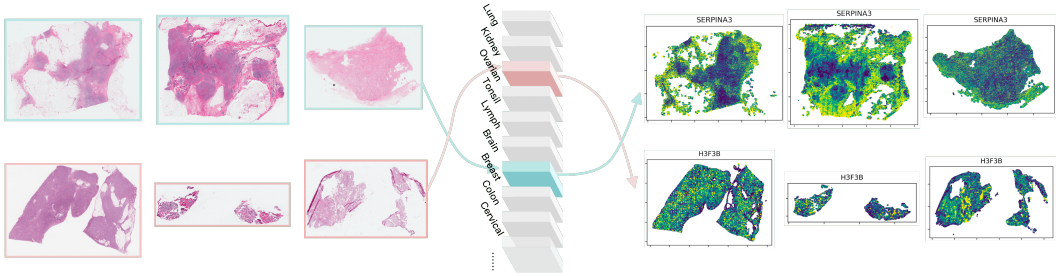


Figure S6: **Illustrative prediction maps on TCGA histology.** Example whole-slide images and model-generated gene-expression maps demonstrate the STBoost inference and visualization workflow. No paired spatial-expression reference is used to assess prediction accuracy in these panels.

D.2 Cell-type downstream analysis

Figure 6 presents the coarse-to-fine expression-prediction analysis on lung acquisition K. Observed and predicted expression are converted to marker scores over the restricted 50-gene target subset, retaining *Club*, *AT2*, and *Immune*. The reference assignments are expression-derived labels, not independent cell-type annotations. Both spatial cluster maps are computed from predicted expression: one uses a global clustering fit and the other a regional fit applied across the observations, with cluster IDs aligned by maximum overlap. Consequently, this panel illustrates clustering stability and coarse marker-label recovery rather than GT-versus-prediction clustering accuracy.

Table S11: Descriptive agreement metrics for the lung downstream example; reference labels are marker-derived and the scope includes the full analysis cohort.

Metric	Value
Spatial observations	111,772
Clusters	10
Genes in downstream target	50
Observed vs. predicted label agreement	0.5935
Balanced label agreement	0.5922
Cluster vs. Pred celltype ARI	0.1362
Cluster vs. Pred celltype NMI	0.2154
Retained marker-derived classes	Club, AT2, Immune

All ten clusters retain the same dominant coarse class. Cluster 3 has the highest agreement (0.934); clusters 1, 5, and 8 have agreement below 0.5. The complete ten-cluster summaries are retained in `data_sources/celltype_2026-10-31/cluster_agreement_summary.csv`.

D.3 Context-stratified ovarian transfer analysis

Figure S7 expands the four-metric summary in Table 3. The analysis uses UNI2 image features with a top-10 retrieval predictor and reference observations from the left two thirds of ovarian acquisition AK (60+). It evaluates 5,000 observations from each target: AD (60+) and AL (< 40). The evaluated gene sets contain 18,137 and 38,606 genes, respectively. This protocol is separate from the current CONCH–ridge benchmark.

The younger target has lower F1 and rank correlations together with a much smaller predicted nonzero fraction. The nonzero-rate gap is the absolute difference between predicted and observed nonzero fractions. Matrix distance, energy distance, and MMD also decrease in these summaries, so their numerical ordering alone does not establish improved prediction. Because target gene sets and other sample properties differ, these results are descriptive examples of context-associated transfer differences; they do not establish a causal effect of age. Each bar is a single target summary, and these runs provide no replicate-based confidence intervals.

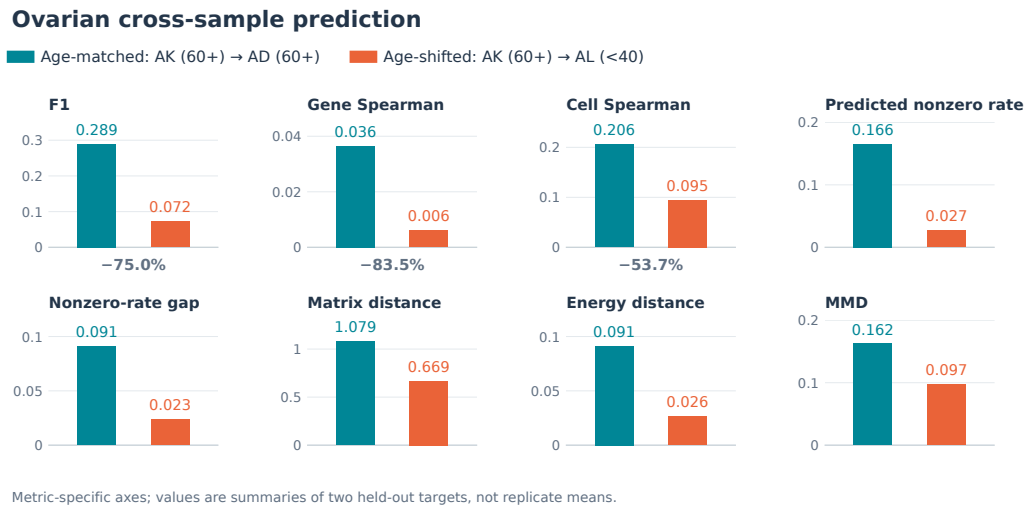
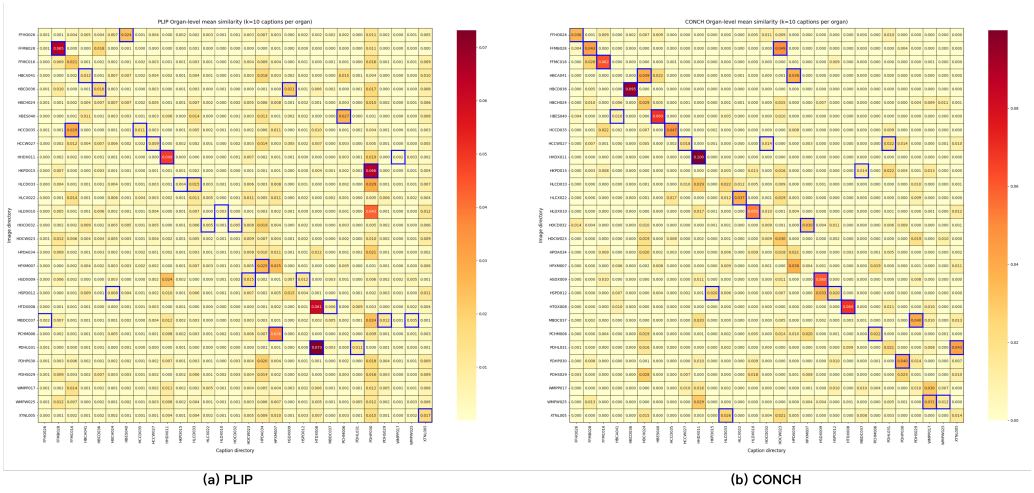


Figure S7: **Complete context-stratified ovarian transfer results.** The same AK source is evaluated on AD (matched older age band) and AL (younger age band). The eight panels show F1, gene- and cell-wise Spearman correlations, the predicted nonzero fraction, its absolute gap from the observed fraction, and three global distribution distances. Each panel uses its own native-scale axis; displayed percentages are relative changes from AD to AL. Values come from the original run summaries, with 5,000 observations per target and different evaluated gene sets. Lower global distances coexist with poorer predictive agreement; the comparison does not isolate age from other sample differences.

D.4 Extended Results for Captioning Evaluation

Metadata-grounded captions describe tissue identity and local image context. We assess image–caption relevance using two complementary encoders: PLIP for sample-level comparisons and CONCH for organ-level comparisons with $k = 10$ captions per organ. Matched image–caption pairs generally score higher than mismatched pairs at both levels (Figure S8), supporting the relevance of structured text for dataset browsing and multimodal analysis.



(a) PLIP (b) CONCH

Figure S8: **Image–caption relevance assessed using PLIP and CONCH.** The upper panel reports sample-level similarity using PLIP; the lower panel reports organ-level similarity using CONCH.

E Datasheet for sMCC40M

We provide a DataSheet for **sMCC40M**, summarizing its objectives, scope, and design choices, along with potential applications and ethical considerations. The datasheet below emphasizes the

benchmark-facing aspects of the release: the units are explicitly typed cell/nucleus observations, benchmark tracks are defined through explicit split files, and the released schema is versioned so that the same processed assets can be used for both modeling and evaluation.

E.1 Motivation for dataset creation

Why was the dataset created?
sMCC40M is a large-scale single-cell dataset featuring paired imaging and gene expression data, with multi-level imaging modalities that capture detailed cellular structures. This comprehensive multimodal resource incorporates elements such as captioning, CCC (cell-cell communication), and attribute annotations, creating a rich interconnected framework. Moreover, sMCC40M serves as a versatile benchmark for foundation models, broadening the scope of existing case studies and facilitating research across multiple tissues and modalities.

- **Extensive Single-Cell Dataset:** Provides a vast collection of paired image and gene expression data, capturing cellular characteristics at multiple layers and scales.
- **Multimodal Content:** Integrates captioning, cell-cell communication (CCC), attribute annotations, and multi-tier imaging, enabling deeper insights into phenotypic and molecular interrelationships.
- **Benchmark Utility:** Offers a foundational platform for evaluating and enhancing current models, expanding potential applications and analytical perspectives across tissues and modalities.

What tasks could the dataset be used for?
Potential tasks include cross-modal retrieval, cell-type classification, morphological phenotype analysis, spatial domain segmentation, and gene expression prediction from histology images. Users are welcome to devise new applications but are strictly prohibited from attempting to re-identify patients.

Has the dataset been used for any tasks already?
The main paper evaluates encoder scaling and within- and cross-acquisition gene-expression prediction, with separate downstream illustrations of coarse cell-type recovery and context-associated transfer differences.

Who funded the creation of the dataset?
sMCC40M was developed with institutional support from multiple sources; details are provided in the Acknowledgments section of the main manuscript.

sMCC40M Bench is intended to provide predefined splits, target definitions, and evaluation scripts for single-cell histology-to-gene prediction under standardized regimes. The benchmark is therefore not only a collection of multimodal data, but also a controlled evaluation package with fixed preprocessing and explicit held-out partitions.

E.2 Dataset composition

What are the instances?
Instances in sMCC40M are cell- or nucleus-aligned observations linked by stable identifiers to:

- Source images, coordinates, and eligibility flags for on-demand multi-scale crops.
- Gene expression profiles and structured metadata.
- Optional textual captions for specified analyses.
- Optional cell-cell communication summaries for specified analyses.

Are relationships between instances made explicit?
Yes. Each cell is uniquely mapped to its corresponding gene expression vector, tissue type, and spatial coordinates. Cell-to-cell communication edges (ligand-receptor pairs) are also provided, capturing local interactions within a 200 μm radius.

What does each instance consist of?

- **Imaging data:** Standard TIFF patches (cell/tissue/WSI) at multiple scales.
- **Expression data:** H5AD matrices storing raw integer counts on the original measured feature axes.
- **Metadata:** Parquet/JSON tables with coordinates, source context, eligibility, and annotation evidence.
- **Captions:** Automated textual descriptions of morphological features and context.

Is there a label/target associated with instances?
Each cell is linked to a variety of annotations (e.g., `cell_type`, `cell_disease_state`). Demographic information is partially available (for certain human samples) and subject to IRB and licensing constraints.

Are external resources required?
The complete processed data and benchmark caches are separate from the public code repository. Source URLs are recorded in the inventory, and source-specific access and redistribution terms apply; pretrained encoder weights must be obtained separately under their own terms.

Recommended data splits or evaluation measures?
The current benchmark uses four directional spatial holdouts and HD leave-one-acquisition-out evaluation. Public JSON fold plans identify the partitions; the complete data-side package retains exact observation memberships and target genes. Details appear in Section C.

What experiments were initially run on this dataset?
The main paper reports encoder scaling, cell-aligned CONCH–ridge prediction, and downstream visualization examples. The benchmark tables retain the current HD18, five-category Xenium extension, and HD cross-acquisition protocols.

Benchmark metadata link observations to acquisitions, platforms, categories, and spatial coordinates. The public code supports replay from saved predictions and refitting from frozen features; both require the corresponding data-side assets.

E.3 Data collection process

How was the data collected?
sMCC40M aggregates single-cell data primarily from 10x Genomics (Xenium & Visium HD) and in-house experiments. Each study provides histopathology images, transcriptomic profiles, and relevant sample metadata.

Who was involved?
All authors of this paper contributed to data curation and validation. Original acquisition protocols (e.g., IRB approvals, tissue preparation) are documented in the source publications.

Over what time frame was the data collected?
The processed release is frozen as 28 October 2026. Acquisition and source-publication dates are recorded where supported by the upstream metadata.

Is the dataset a sample from a larger set?
We aim to include all single-cell data from the selected studies. Instances are unique pairings of morphological patches and expression profiles. Minor omissions may arise if crucial alignment files are missing.

Known errors or sources of noise?
Whole-slide images vary in quality due to staining differences, compression artifacts, and resolution limits. Gene expression data may exhibit batch effects or sequencing noise. Users are encouraged to apply standard normalization/denoising procedures as needed.

Studies were included only when both histology and spatial-molecular coordinates were available and could be aligned at cell level. Samples lacking reliable coordinate transforms, segmentation support, or sufficient image coverage were excluded from benchmark construction. Known related source material is flagged. Acquisition holdouts alone do not establish donor independence, and the current benchmark reports a related-source sensitivity analysis.

E.4 Data preprocessing

What preprocessing was done?

- **Image processing:** Selected TIFF images retain acquisition-specific physical pixel sizes and coordinate transforms.
- **Gene expression standardization:** Raw integer counts retain original gene/feature IDs and types in H5AD; benchmark transformations are specified separately.
- **Cell-level alignment:** Each cell's bounding box and expression profile carefully cross-validated to ensure correct mapping.

Were raw data retained?
The source inventory retains public re-download links. The code repository contains documentation and compact result tables; full images, masks, expression matrices, and benchmark caches are separate data-side assets.

Is preprocessing software available?
Preprocessing and benchmark code are documented in https://github.com/aicb-ZhangLabs/sMCC40M; the repository specifies the supported workflows and required inputs.

The current CONCH–ridge benchmark uses 150-pixel native H&E crops, frozen 512-dimensional features, training-only scaling and PCA, and ridge regression. HD targets aggregate 2 μm bins by strict majority area; Xenium targets retain native cellular counts. Both use log1p raw counts and training-selected HVGs, with no library-size normalization or zero-filling of unmeasured genes.

E.5 Dataset distribution

How is the dataset distributed?
Download the processed acquisitions and benchmark inputs from https://huggingface.co/datasets/zhang-lab/sMCC40M. The dataset card provides a sample index and commands to download either a complete acquisition or its H5AD expression matrix. Code and evaluation instructions are available at https://github.com/aicb-ZhangLabs/sMCC40M.

When will it be released?
The frozen October 28, 2026 release is publicly available. The archived revision identifies the files used here.

Under what license?
All 91 source dataset pages specify CC BY 4.0. The release provides per-acquisition attribution, source links, and descriptions of processing in its `licensing/` directory. Third-party software and pretrained weights retain their own terms.

Who maintains the dataset?
The dataset is actively maintained by the authors. For questions or inquiries, please submit an issue on the GitHub page.

Will the dataset be updated?
Yes. We plan periodic updates as new single-cell samples are curated. Dataset versioning will track changes, documented via GitHub releases.

Repository documentation identifies the inputs and settings for each supported benchmark. The current protocols are specified in Section C.

E.6 Legal and ethical considerations

Ethical review approvals?
All in-house data adhere to institutional IRB protocols. Public data rely on the ethical compliance of the original sources. Detailed references to each study's IRB are included where available.

Could this dataset expose individuals to harm?
No patient identifiers (name, address, etc.) exist in sMCC40M. Users are strictly forbidden from attempting re-identification or reverse engineering of personal data.

Sensitive or offensive content?
The resource concerns histology and molecular measurements. Source-backed sample attributes may include age, sex, disease, and tissue preparation; missing or conflicting metadata are marked explicitly.

Social bias or fairness concerns?
Demographics (e.g., age, gender) are only reported if originally included by the source. Uneven tissue and demographic coverage can bias evaluation; the context-stratified example does not establish performance for a population.

Licensing fees or restrictions?
Source-specific access and reuse conditions apply. Repository availability does not replace upstream permissions or permit unrestricted redistribution.

Future updates or obsolescence?
If the dataset is ever discontinued, we will note this in the GitHub README. We encourage the community to mirror or fork the repository for long-term availability.

Benchmark scores should not be interpreted as uniformly representative across all biological domains. Organ, species, and platform coverage are deliberately broad but highly uneven, and demographic metadata are sparse for many public studies. The benchmark is therefore best understood as a reproducible stress test for multimodal pathology models, not as a clinically calibrated estimate of performance for any individual population.

F Author Statement

The authors of this paper bear all responsibility in case of violation of rights associated with the sMCC40M dataset.

NeurIPS Paper Checklist

1. **Claims.**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: **Yes.**

   Justification: The abstract and Introduction state the paper's scope as a context-aware, cell-aligned spatial-transcriptomics resource and benchmark for image-to-single-cell gene-expression prediction. The claims are supported by the release manifest, dataset construction, quality-control analyses, single-cell alignment protocol, encoder benchmarking, CONCH–ridge within- and cross-acquisition evaluation, and context-stratified ovarian analysis.

2. **Limitations.**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: **Yes.**

   Justification: The paper states that robust histology-to-gene generalization remains limited by biological context, platform shift, patient shift, organ coverage, and cellular-resolution alignment quality. Acquisition holdouts and the context-stratified ovarian example do not establish donor independence or a causal effect of age. The release is presented as a research benchmark, not as a clinical prediction system.

3. **Theory assumptions and proofs.**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete proof?

   Answer: **N/A.**

   Justification: The paper does not make theoretical theorem or proof claims. The mathematical and algorithmic content defines data alignment, cell-centered histology records, benchmark splits, model inputs, training objectives, and evaluation metrics.

4. **Experimental result reproducibility.**

   Question: Does the paper fully disclose the information needed to reproduce the main experimental results to the extent that it affects the main claims and conclusions?

   Answer: **Yes.**

   Justification: The paper and appendix describe the standardized datasets, cell-alignment procedure, train/validation/test split definitions, organ-level benchmark coverage, target gene construction, shared-gene filtering, baseline methods, random seed, metrics, and evaluation protocols. The current benchmark covers 23 reporting categories and 34 HD acquisition holdouts in ten organ/species strata; public fold plans and repository documentation identify the required data-side inputs.

5. **Open access to data and code.**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to reproduce the main experimental results?

   Answer: **Yes.**

   Justification: Section 5 links to the public processed acquisitions, frozen benchmark inputs, and code. The dataset card gives sample download commands, and the benchmark guide documents evaluation and access to external encoder weights.

6. **Experimental setting and details.**

   Question: Does the paper specify all training and test details necessary to understand the results?

   Answer: **Yes.**

   Justification: The paper specifies the image-to-single-cell gene-expression prediction task, within- and cross-acquisition regimes, sample identities, spatial folds and buffers, acquisition holdouts, training-selected gene sets, and reported gene-wise and cell-wise metrics.

7. **Experiment statistical significance.**

   Question: Does the paper report error bars or other appropriate information about statistical significance?

Answer: **No.**

Justification: Section C.2 reports approximate marginal 95% spatial-block bootstrap intervals for the expanded HD CONCH–ridge experiment and all three controls, with explicit resampling units, species stratification, support flags, and conditional scope. Arabidopsis and the 18-category macro do not receive formal intervals because of insufficient spatial support. The answer remains No because uncertainty is not systematically quantified for all main-text experiments and downstream examples; the new HD intervals do not establish uncertainty or significance for those experiments.

8. **Experiments compute resources.**

   Question: For each experiment, does the paper provide sufficient information on compute resources needed to reproduce the experiments?

   Answer: **No.**

   Justification: The paper provides data construction details, benchmark splits, model settings, and reproduction scripts, but it does not fully itemize hardware, memory usage, wall-clock runtime, or storage requirements for every reported experiment. Exact runtime will depend on local GPU, CPU, memory, storage, and I/O configuration.

9. **Code of ethics.**

   Question: Does the research conform to the NeurIPS Code of Ethics?

   Answer: **Yes.**

   Justification: The work uses existing public or externally sourced research datasets and does not involve new participant recruitment, human-subject intervention, clinical deployment, or personally identifiable private information. The model and benchmark are presented for research evaluation of spatial-transcriptomics prediction, not for diagnosis, treatment selection, or clinical decision-making.

10. **Broader impacts.**

    Question: Does the paper discuss potential positive and negative societal impacts of the work?

    Answer: **Yes.**

    Justification: The paper motivates positive impact through scalable single-cell spatial-transcriptomics benchmarking, improved evaluation of histology-to-gene models, and reusable cell-aligned data assets. It also identifies negative or misuse risks: benchmark performance may not imply clinical robustness, cross-patient and cross-platform generalization remains difficult, and predictions should not be used for medical or therapeutic decisions without additional biological, clinical, and regulatory validation.

11. **Safeguards.**

    Question: Does the paper describe safeguards for responsible release of high-risk models or datasets?

    Answer: **N/A.**

    Justification: The paper does not release a high-risk generative model, clinical decision system, human-facing model, or private scraped dataset. The released assets are research code, release manifests, cell-aligned derived artifacts, benchmark splits, and configuration files constructed from public or externally sourced biological datasets. Upstream access and redistribution are governed by the original data-source terms.

12. **Licenses for existing assets.**

    Question: Are creators or original owners of used assets properly credited, and are license and terms of use respected?

    Answer: **Yes.**

    Justification: The paper cites upstream spatial-transcriptomics datasets, 10x Genomics example datasets, pathology and spatial-analysis resources, and external baseline methods. Raw upstream biological samples are not claimed as new assets. The released manifests and benchmark artifacts document provenance and are intended to respect original data licenses and redistribution terms.

13. **New assets.**

Question: Are new assets introduced in the paper well documented and provided alongside the assets?

Answer: **Yes.**

Justification: The paper introduces sMCC40M, a standardized cell-aligned multimodal resource comprising 38.53 million cell/nucleus observations; release manifests; benchmark splits; configuration files; and evaluation scripts. The documentation describes the standardized biological datasets, aligned-cell records, platform-specific files, organ-level benchmark coverage, split construction, target details, and derived benchmark packages.

14. **Crowdsourcing and research with human subjects.**

Question: For crowdsourcing experiments and research with human subjects, does the paper include participant instructions, screenshots, and compensation details?

Answer: **N/A.**

Justification: This work does not involve crowdsourcing, user studies, human annotation labor, participant recruitment, or compensation of human subjects.

15. **Institutional review board approvals or equivalent for research with human subjects.**

Question: Does the paper describe participant risks and IRB approvals, if applicable?

Answer: **N/A.**

Justification: The study analyzes existing public or externally sourced research datasets and does not collect new human-subject data, recruit participants, intervene on participants, or use identifiable private participant information. IRB review is therefore not applicable to the work reported in this paper.

16. **Declaration of LLM usage.**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods?

Answer: **N/A.**

Justification: LLMs are not an important, original, or non-standard component of the core method, dataset construction, benchmark definition, model training, or scientific claims in this research. The work is based on spatial transcriptomics, histology-image encoders, cell-level alignment, benchmark construction, and image-to-gene prediction evaluation.