Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

HistAgent: an evidence-grounded agent for spatial molecular reasoning from routine histology

Jin Liu

liujinlab@cuhk.edu.cn

The Chinese University of Hong Kong, Shenzhen https://orcid.org/0000-0002-5707-2078

Wenlin Li

The Chinese University of Hong Kong, Shenzhen

Hengrui Liu

The Chinese University of Hong Kong, Shenzhen

Yan Lu

The Chinese University of Hong Kong, Shenzhen

Xu Liao

The Chinese University of Hong Kong Shenzhen

Xiaobo Sun

Emory University

Weiwei Zhai

Key Laboratory of Zoological Systematics and Evolution, Institute of Zoology, Chinese Academy of Sciences https://orcid.org/0000-0001-7938-0226

Chengqi Lin

Southeast University https://orcid.org/0000-0002-2121-584X

Article

Keywords:

Posted Date: August 18th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10677502/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: There is **NO** Competing Interest.

HistAgent: an evidence-grounded agent for spatial molecular reasoning from routine histology

Wenlin Li[1,2,†], Hengrui Liu[1,†], Yan Lu[1,†], Xu Liao[1], Xiaobo Sun[3], Weiwei Zhai[4], Chengqi Lin[5], and Jin Liu[1,2,*]

[1]School of Data Science, The Chinese University of Hong Kong, Shenzhen, Shenzhen, China

[2]Shenzhen Loop Area Institute, Shenzhen, China

[3]Department of Human Genetics, Emory University School of Medicine, Atlanta, GA, United States

[4]Institute of Zoology, Chinese Academy of Sciences, Beijing, China

[5]Key Laboratory of Developmental Genes and Human Disease, School of Life Science and Technology, Southeast University, Nanjing, China

[†]These authors contributed equally

[*]Corresponding author: liujinlab@cuhk.edu.cn

Abstract

Routine hematoxylin and eosin (H&E) staining is a widely used histology test in clinical practice and biomedical research, but it does not provide direct spatial molecular information. Spatial transcriptomics (ST) provides such data while preserving the tissue context, yet its high cost, technical complexity, and limited accessibility restrict its broad application. Methods for inferring spatial molecular information from H&E images could bridge this gap; however, current visual-omics algorithms often exhibit limited cross-dataset generalizability and do not sufficiently facilitate traceable biological interpretation. Here, we present HistAgent, a unified framework that couples a visual-omics foundation model with a spatial artificial intelligence (AI) module. The foundation model, trained on 2.23 million paired H&E–ST spots, generates ranked molecular readouts from local and contextual H&E images. The spatial AI agentic module structures these readouts as evidence cards comprising data on the inferred cell composition, functional programs, and spatial context, thereby enabling question-driven multi-turn analysis. In tests, HistAgent recovered and interpreted literature-defined spatial findings, spanning immune, stromal, tumor, and anatomical states across tissues at lower cost and in less time than conventional ST

workflows, and its molecular readouts accurately reproduced five standard ST analyses. Its intermediate visual-omics representations supported whole-slide predictions of molecular and histologic subtypes, tumor grade, and prognostic risk. Providing the module access to an evidence bank constructed from experimentally measured ST profiles enabled further natural-language and H&E–image retrieval as well as an analysis of molecular states. Overall, HistAgent enables generalizable spatial molecular analyses and traceable biological interpretation from routine histology.

Introduction

Hematoxylin and eosin (H&E)-stained tissue slides are routinely available in clinical and research settings [1, 2]; these provide morphological information but do not directly reveal the spatial gene-expression configurations that underlie local tissue states. In contrast, spatial transcriptomics (ST) can profile gene expression while preserving the tissue context [3], enabling the analysis of tissue architecture, cellular niches, and disease microenvironments [4–6]. Currently, its high cost, technical complexity, and limited accessibility hinder its routine use [7, 8]. Nevertheless, paired H&E–ST data archives continue to grow, prompting routine histology analysis to move beyond simple descriptive imaging toward platforms with two complementary capabilities: inferring omics readouts and generating evidence-based biological interpretations to address direct, region-specific queries.

The reliable inference of omics readouts solely from routine histology is a critical prerequisite for deriving downstream biological evidence. Pioneering visual-omics methods have demonstrated the feasibility of predicting transcriptomes from histology images [9–14]. However, the methods' transferability across datasets and batches remains limited, in part because they rely on predefined, dataset-specific gene sets [15]. With the rapid expansion of paired histology–ST data repositories, visual-omics foundation models are beginning to benefit from large-scale pretraining, as exemplified by STPath [16] and OmiCLIP [17]. However, the continuous gene-expression predictions produced by STPath can be sensitive to expression scale, normalization procedures, and batch effects [15, 18, 19], and the CLIP-based contrastive image-omics framework [20] of OmiCLIP limits its direct omics generation. OmiCLIP also neglects the spatial microenvironmental context, which may restrict its utility in context-dependent downstream applications. These limitations highlight the unmet need for image-conditioned, rank-based molecular readouts that include the local morphology within the surrounding tissue context while remaining robust across expression scales and datasets.

Reliable molecular readouts provide the initial evidence, but answering region-specific questions requires users to select relevant signals; integrate genes, cell states, and pathways with the spatial context; and trace each interpretation back to its supporting evidence. For example, elevated *CXCL13* expression alone is not sufficient to indicate a tertiary lymphoid structure (TLS)-like niche; rather, this interpretation requires concordant evidence of immune-cell composition, immunoglobulin programs, and spatial organization [21–25]. The processes of evidence selection and integration currently depend on expert judgment,

making it difficult to scale across regions and questions. Recent large language model (LLM)-based systems, including PathChat, GeneAgent, and CellVoyager [26–28], show that agentic frameworks can select task-relevant information, integrate heterogeneous evidence, and support iterative analysis. These systems can address the questions answerable by pathology images, gene-set interpretation, and single-cell analyses. However, they were not designed to reason based on histology-derived molecular readouts or to link region-level interpretations to the supporting molecular and spatial evidence.

Addressing these gaps requires a framework that both generates robust molecular readouts from routine H&E images and organizes question-relevant molecular and spatial evidence into traceable interpretations. Here, we present HistAgent, a unified framework that couples image-conditioned molecular inference with evidence-grounded, question-driven analyses of local tissue states. HistAgent comprises two modules: a visual-omics foundation model that generates ranked molecular readouts from local and contextual H&E views and a spatial artificial intelligence agent that organizes these readouts into structured evidence for question-specific interpretation. The visual-omics model formulates each histology-to-molecular inference via the conditional generation of ranked gene readouts linked to the local and contextual morphology. Trained on 2.23 million spatially matched H&E–ST region pairs, the model autoregressively ranks the top 50 genes from a unified vocabulary of 32,019 standardized gene symbols. This rank-based representation is not restricted to a dataset-specific target panel and reduces the model's direct dependence on expression scale and normalization. The agentic module first organizes each ranked readout into an evidence card containing the top-ranked genes, inferred cell composition, functional programs, and spatial context. Fine-tuned on teacher-generated multi-turn traces, the agent uses these cards to offer question-specific interpretations with links to supporting molecular and spatial evidence. It supports the iterative analysis of researcher-defined regions and links each interpretation to the specific evidence fields on which it is based. For quantitative spatial transcriptomics (ST) analyses, a separate reference-guided expression inference procedure is applied to map the ranked molecular readouts to reference-guided expression estimates using an independent ST reference. Overall, HistAgent takes routine histology beyond static molecular readout predictions by enabling an interactive, query-driven, and evidence-grounded process for spatial molecular analysis.

We applied HistAgent to four complementary settings — spatial biological discovery, molecular recovery, clinical prediction, and ST atlas retrieval. Using H&E images alone, HistAgent recovered the spatial biological findings reported in published studies and provided supporting molecular evidence for TLS-like immune niches, stromal structures, tumor epithelial states, and anatomical regions [25, 29–34]. Across 135 held-out ST slides, HistAgent recovered spot-level gene rankings from H&E images. Using an independent ST reference, these rankings were mapped to spatial molecular profiles for five standard ST analyses [35, 36]. The intermediate visual-omics representations further supported the prediction of histological and molecular subtypes, tumor grade, and prognostic risk across organs and cohorts. At the atlas scale, an evidence bank constructed from experimentally

measured ST profiles was shown to support HistAgent in natural-language and image-based queries. We quantitatively evaluated the model's natural-language retrieval ability and used image-based queries to demonstrate its evidence-grounded atlas exploration skills. Across these settings, HistAgent provides a unified route from routine histology to ranked molecular inference, evidence-linked regional analysis, and tissue exploration at the spot, slide, and atlas scales.

Results

HistAgent enables evidence-grounded spatial molecular reasoning from routine histology

HistAgent is a unified framework that relates molecular inference based on routine H&E images to evidence-grounded reasoning of local tissue states. HistAgent comprises a visual-omics foundation model and a spatial agentic module, which together form a coupled workflow (Fig. 1a). The foundation model first generates an image-conditioned, rank-based molecular readout for each local tissue region. The agentic module then organizes each readout and its rank-derived analyses into a structured evidence card. It uses this card to answer region-specific questions while linking each conclusion to supporting molecular and spatial evidence.

The visual-omics foundation model forms spatial molecular predictions via the conditional generation of ranked molecular readouts from multi-scale H&E morphology. The spot-centered H&E view is analyzed within the surrounding tissue context to capture the morphology within each 55-µm spot, as well as its immediate microenvironment. Each ST profile is represented as a top 50 gene ranking, rather than as continuous expression values, thereby reducing the model's sensitivity to expression scale, normalization procedures, and batch effects. As it is conditioned on various image scales and previously generated genes, the model can autoregressively predict the top 50 genes from a unified vocabulary of 32,019 standardized gene symbols. This design explicitly models dependencies within the ranking without restricting predictions to a dataset-specific gene panel. We trained the foundation model on a paired H&E–ST corpus assembled from human and mouse 10x Visium datasets, comprising 2.23 million paired spots from 936 tissue slides across 32 tissue categories (Fig. 1b, left). When continuous values are required for quantitative ST analyses, a separate reference-guided inference procedure uses an independent species-matched ST reference to map each ranked readout to expression estimates for 15,482 human genes or 16,811 mouse genes (see Methods). Together, this architecture and training corpus enable the direct generation of context-aware, rank-based molecular readouts across diverse tissue settings.

Building on these ranked readouts, the spatial agentic module constructs structured molecular and spatial evidence and uses this evidence to answer questions about local tissue states. First, the module organizes the information on top-ranked genes, inferred

cell composition, functional programs, and spatial context into spot-level evidence cards [37, 38]. It then uses an adapted LLM to select and integrate question-relevant evidence from these cards and generate answers across multiple conversational turns, with each conclusion linked to its supporting molecular and spatial basis. To develop this capability, we fine-tuned the LLM on 50,000 multi-turn evidence-reasoning traces generated by a teacher LLM (Fig. 1b, right; see Methods). This design supports the iterative, multi-turn analysis of local tissue states while preserving a traceable link between each interpretation and its molecular and spatial evidence.

We evaluated the complementary capabilities of HistAgent across spot, whole-slide, and atlas scales (Fig. 1c,d). At the spot level, we assessed its recovery of literature-defined biological states, ranked molecular readouts, and downstream spatial analyses. At the whole-slide level, we tested whether the learned foundation-model representations support the prediction of histologic and molecular subtypes, tumor grade, and prognostic risk. At the atlas scale, we quantitatively evaluated its natural-language retrieval from a measured ST evidence bank and used image-based queries to demonstrate its atlas exploration. Across quantitative spot-level and whole-slide benchmarks, HistAgent achieved the best overall mean rank of all evaluated methods (Fig. 1d). Overall, HistAgent accurately connects routine histology data to ranked molecular inference, evidence-grounded regional reasoning, and cross-scale tissue exploration.

HistAgent recovers and interprets spatial biological findings from routine histology

HistAgent recovered and interpreted a TLS-like immune niche from the H&E image of a renal cell carcinoma slide. We used the paired 10x Visium profile as the evaluation reference and focused on TLS localization and interpretation [25] to test the complete workflow — from molecular-readout generation to evidence-linked regional interpretation. During molecular-readout acquisition, HistAgent demonstrated a reduction in both time and cost relative to conventional ST. This it achieved by generating a top 50 gene ranking for each 55-µm spot directly from a digitized H&E slide within minutes, whereas the conventional workflow required ST profiling. The HistAgent workflow, including slide scanning, took less than one day and cost approximately US$10–30 per slide, compared with several weeks and US$5,000–7,000 for conventional ST [39] (Fig. 2a).

We first quantified the concordance between the spatial gene-expression patterns inferred by HistAgent and those measured by ST profiling. To obtain continuous spatial expression estimates from each top 50 ranked gene, we applied reference-guided expression inference using an independent ST reference. We examined eight representative genes reflecting the TLS-associated immune and stromal environment: the immunoglobulin-related genes *IGHG2*, *IGHG4*, and *IGLV3-1*; the complement gene *C7*; and the stromal and extracellular matrix (ECM) genes *COL1A1*, *DCN*, *FN1*, and *TGFBI* [25]. Across all eight genes, the HistAgent-derived expression estimates showed high spatial concordance with

the measured ST, with a mean Pearson correlation coefficient (PCC) of 0.58 (Fig. 2b). This concordance persisted when the evaluation was expanded to encompass progressively broader sets of 50 – 500 highly variable genes defined from measured ST. HistAgent consistently achieved higher mean PCCs than the visual-omics foundation models OmiCLIP and STPath (Supplementary Fig. S1).

Next, we examined whether HistAgent-derived molecular readouts could accurately locate a TLS. Using 25 literature-supported marker genes [25] (Supplementary Table 1), we calculated the TLS signature scores from both HistAgent-derived and measured ST profiles. Both scores were strongly enriched in expert-annotated TLS spots relative to non-TLS regions (two-sided Mann-Whitney U test, $P < 10^{-100}$ for each readout). HistAgent achieved a comparable TLS-localization performance to that of measured ST, with areas under the precision-recall curve (AUPRCs) of 0.586 and 0.592, respectively (Fig. 2c). Thus, HistAgent-derived molecular readouts retained sufficient information to localize a TLS and provided a basis for subsequent evidence-grounded interpretation.

Building on this localization, we used the spatial agentic module to interpret the molecular and spatial context of the identified TLS region. Conventional workflows require researchers to manually integrate separate analytical outputs with biological interpretations [25, 40]. By contrast, HistAgent organizes the data on the top-ranked genes, inferred cell composition, biological programs, and spatial context into spot-level evidence cards. The fine-tuned LLM then uses these cards as structured context to select and integrate the evidence relevant to region-specific questions. To illustrate this capability, we examined the spot with the highest HistAgent-derived TLS signature score within the annotated TLS region (Fig. 2d). In response to the query, "Why TLS-like?", HistAgent used the expression of immunoglobulin and antigen-presentation genes along with evidence of B cells, plasma cells, and cells with antigen-presenting features to interpret the region as a TLS-like immune niche. The agentic module also supports multi-turn dialogue, allowing researchers to progressively examine the same biological finding from its molecular basis to its spatial context. For the follow-up query, "What context supports it?", HistAgent integrated programs related to humoral immunity, complement activity, and ECM remodeling with evidence from the local stromal and immune neighborhood. This evidence was used to explain how the surrounding tissue supported the TLS-like interpretation. Both answers were returned within minutes and were broadly consistent with the TLS-associated immune and stromal features reported in the original study [25].

Finally, to assess the model's generalizability beyond renal cell carcinoma and TLS, we constructed a benchmark for recovering spatial biological findings across studies (Fig. 2e). The benchmark included 22 slides from five published studies [25, 30–33] and one public 10x Genomics resource [34], which covered 12 biological findings spanning immune states, stromal states, tumor and epithelial states, and tissue anatomy (see Supplementary Table 1 for details). For each finding, we applied the same literature-defined gene signature to HistAgent, OmiCLIP, STPath, and measured ST, and evaluated the resulting localization data against the corresponding manual annotation using AUPRC. Across the 12 findings,

HistAgent achieved the highest mean AUPRC (0.70), followed by the measured ST (0.65), OmiCLIP (0.41), and STPath (0.35). By deriving molecular readouts from the H&E morphology, HistAgent achieved higher AUPRCs than the measured ST for findings with morphology-informed annotations such as dorsolateral prefrontal cortex white matter and the prostate stromal compartment [30, 32]. Together, the benchmark and TLS cases showed that HistAgent can recover spatial biological findings from routine H&E images across diverse tissues and organize the regional evidence to support its interpretation, therefore complementing measured ST and expert analysis.

HistAgent recovers spatial molecular readouts and recapitulates spatial transcriptomic analyses from routine histology

We next evaluated whether HistAgent's rank-based molecular readouts could recover spatial molecular structure across 135 held-out ST slides and support standard downstream analyses (Fig. 3a). The benchmark comprised 79 human and 56 mouse slides across 15 organ cohorts (Supplementary Table 2). Using the measured ST profiles as the reference, we compared the molecular readouts generated by HistAgent, STPath, and OmiCLIP after analysis of the same slides under a unified evaluation framework. The benchmark first assessed the direct recovery of molecular information, as measured by ST, and then evaluated whether the predicted readouts would reproduce the tissue organization and biological states revealed by standard ST analyses.

We examined the models' molecular recovery performance at both the spot and gene levels. These complementary evaluations respectively captured the local molecular composition and gene-specific spatial variation, which revealed the predicted readouts' preservation of both the local tissue states and tissue-wide spatial organization. At the spot level, HistAgent directly generated a top 50 gene ranking for each spot. The corresponding STPath and OmiCLIP rankings were obtained by ordering their continuous expression predictions. For each spot, we used the 50 most highly expressed genes in the measured ST profile as the reference. HitRate@50 was applied to measure the coverage of these reference genes by the predicted top 50 list, and we used the mean average precision at 50 (mAP@50) to additionally account for their positions within the ranking. Both metrics showed that HistAgent recovered the local molecular composition more accurately than either baseline method (Fig. 3b; Supplementary Fig. S2). Across the 135 held-out slides, HistAgent achieved the highest mean HitRate@50 (0.699) and mAP@50 (0.655), substantially outperforming both baselines. This stronger performance can be attributed to its training on more data from more human and mouse organs than either baseline model. The rank-based objective was also designed to reduce HistAgent's direct dependence on expression scale and batch effects [17, 41], unlike STPath, which directly predicts continuous expression values.

At the gene level, we used the previously described reference-guided expression inference procedure to derive spatial expression estimates from HistAgent's top 50 rankings. PCCs

across spots showed the level of agreement between the predicted and measured spatial expression for each gene. For the top 50 highly variable genes, as defined using measured ST, the mean PCC was 0.440 for HistAgent, 0.157 for STPath, and 0.064 for OmiCLIP. HistAgent retained the highest mean PCC when the evaluation was expanded from 50 to 500 highly variable genes (Supplementary Fig. S3). HistAgent's higher spatial concordance was consistent with its multi-scale H&E input, which integrates local morphology with the surrounding tissue context. By comparison, OmiCLIP processes each spot independently and does not explicitly model the relationships among neighboring spots, which may limit its recovery of gene-specific spatial variation.

We next evaluated whether the recovered molecular readouts would agree with five standard ST analyses: spatially variable gene detection, spatial domain identification, cell-type deconvolution, differential expression analysis, and pathway enrichment analysis. We applied the same workflows to each slide readout from HistAgent, STPath, and OmiCLIP, as well as measured ST. The agreement between the predicted and measured results was quantified using six task-specific metrics, with higher values indicating greater concordance (Supplementary Table 3). Across this benchmark, HistAgent achieved the highest mean value for all six metrics and was superior in 88 of the 90 organ-level comparisons (Fig. 3c; the corresponding score distributions are shown in Supplementary Fig. S4). The relative gains over the best-performing baseline ranged from 29% to 235%, with the largest improvements observed for differential expression analysis and spatially variable gene detection. These results show that HistAgent's molecular readouts retained biologically meaningful information about tissue organization, cell-type composition, and functional programs.

To visualize these aggregated results within tissue space, we selected four slides, comprising human stomach [42] and breast tissues [33] and mouse lung [43] and spleen tissues [44], with a median overall performance across five downstream analyses (Fig. 3d). For each slide, we displayed the molecular features exhibiting the greatest task-specific spatial variation in the measured ST findings. HistAgent reproduced the spatial expression patterns of *FOS* and *H19* in human stomach, *MGP* and *KRT5* in human breast, *Sftpc* and *Scgb1a1* in mouse lung, and *Cd74* and *Coro1a* in mouse spleen tissues. It also recovered the corresponding distributions of epithelial cells, basal cells, type II pneumocytes, and macrophages, together with pathway-score patterns for the complement cascade, ECM organization, surfactant metabolism, and scavenger receptor uptake. Visualizations of the remaining organ cohorts, together with matched STPath and OmiCLIP outputs, further demonstrated the stronger spatial agreement between the HistAgent-derived readouts and measured ST data (Supplementary Fig. S5 and S6).

HistAgent supports clinical prediction across cohorts and organs

To assess whether the morphological and molecular information obtained by HistAgent would support clinical predictions, we used the model's intermediate visual-omics representations to predict the clinical endpoints of patients from whole-slide images (Fig. 4a).

HistAgent produced a 512-dimensional representation of each H&E tile by encoding both the local morphology and surrounding tissue context before generating a ranked molecular readout. Across all tasks, HistAgent was kept frozen, and slide-level representations were obtained by aggregating the tile-level representations using attention-based multiple-instance learning (ABMIL) [45, 46]. The benchmark comprised 20 clinical endpoints spanning a range of histologic and molecular subtypes, tumor grades, and prognostic risk across various cohorts and organs (Supplementary Table 4). Classification performance was assessed using the area under the receiver operating characteristic curve (AUROC), macro-averaged for multiclass tasks, whereas prognostic performance was assessed using the concordance index.

Using the same patient sets, patient-level folds, and ABMIL framework, we compared HistAgent with the pathology foundation models GigaPath [47], UNI v2 [48], MUSK [49], CHIEF [50], and CONCH [51]; the visual-omics foundation models OmiCLIP and STPath; and the organ-specific visual-omics method Path2Space [11]. Predictions were averaged across seeds and slides within each patient to obtain patient-level out-of-fold probabilities or survival-risk scores. Across the 20 endpoint-level estimates, HistAgent achieved the best mean rank of 1.9 and ranked first for 15 endpoints, including 8 of the 11 molecular endpoints, 1 of the 2 histologic-subtype endpoints, both tumor-grade endpoints, and 4 of the 5 prognostic-risk endpoints (Fig. 4b). HistAgent showed significant gains in one-sided paired tests ($P < 0.05$) at the patient level for 12 of the 15 comparisons against the strongest baseline. These included six molecular endpoints, one histologic-subtype endpoint, both tumor-grade endpoints, and three prognostic-risk endpoints. The gains spanned all four endpoint categories, including molecular subtyping of stomach adenocarcinoma (STAD) [52] (macro-AUROC, 0.900 versus 0.860; $P = 0.001$), distinguishing invasive ductal carcinoma from invasive lobular carcinoma in breast invasive carcinoma (BRCA) [53] (AUROC, 0.970 versus 0.950; $P < 0.001$), tumor grading in the Clinical Proteomic Tumor Analysis Consortium head and neck squamous cell carcinoma cohort [54] (macro-AUROC, 0.803 versus 0.768; $P = 0.026$), and prognostic-risk prediction in BRCA (concordance index, 0.716 versus 0.658; $P = 0.002$). All endpoint-level estimates are shown in Fig. 4b.

To interpret HistAgent's slide-level predictions, we applied the trained ABMIL models to individual tile representations and compared the resulting subtype scores and survival-risk scores across manually annotated regions in STAD and BRCA [55, 56] (see Methods; Fig. 4c). In both STAD slides, the predicted scores of the genomically stable subtype were higher in poorly differentiated regions than in highly differentiated regions (separate one-sided Mann–Whitney U tests, $P < 0.001$ for both). This difference is consistent with the enrichment of diffuse-type histology in genomically stable gastric cancers [52]. In the BRCA slide, the tile-level overall survival-risk scores were higher in the annotated tumor region than in the stroma (one-sided Mann–Whitney U test, $P < 0.001$), indicating the prognostic signal was localized to the tumor compartment. Additional STAD and BRCA cases are shown in Supplementary Figs. S7 and S8.

Finally, we investigated whether patient-level out-of-fold risk scores from HistAgent

could distinguish among patients with different overall survival outcomes. We were particularly interested in its prognostic value beyond the established clinical factors in liver hepatocellular carcinoma (LIHC) and lower-grade glioma (LGG) [57, 58] (Fig. 4d). Kaplan–Meier analysis showed that patients with risk scores above the cohort median had shorter overall survival times than those with scores below the median in both LIHC ($n = 356$) and LGG ($n = 486$) (two-sided log-rank $P < 0.001$ for both). To determine whether these scores provided prognostic information beyond the established clinical predictors, we fitted a multivariate Cox regression adjusted for American Joint Committee on Cancer stage in LIHC and isocitrate dehydrogenase status in LGG. After adjustment, the hazard ratio per one-standard-deviation increase in the continuous HistAgent-derived risk score was 2.18 in LIHC (95% confidence interval, 1.80–2.65; $n = 336$) and 1.33 in LGG (95% confidence interval, 1.09–1.62; $n = 480$). Together, these findings show that HistAgent's intermediate visual-omics representations achieve more than just molecular readout and can support clinically relevant prediction across cohorts and organs.

HistAgent enables atlas-scale retrieval of measured ST evidence

To extend HistAgent's evidence-grounded reasoning beyond individual slides, we developed an ST atlas-retrieval framework that maps measured atlas spots and natural-language or H&E–image queries onto a shared retrieval-embedding space (Fig. 5a). During offline atlas construction, the evidence cards derived from 2.23 million measured ST spots were embedded as 4,096-dimensional vectors using Qwen3-Embedding-8B [59] to form a measured ST evidence bank. During online query processing, natural-language queries were embedded directly using Qwen3-Embedding-8B. H&E image queries were first processed by HistAgent to generate predicted evidence cards, which were then embedded using the same model. Both query types were subsequently matched by cosine similarity to reference atlas spots with similar molecular states, thereby providing measured ST evidence for follow-up analysis.

To make the reference evidence bank accessible for interactive exploration, we developed the HistAgent Atlas Explorer (`https://histagent.bio/atlas-explorer/`). The web interface combines a natural-language search engine with spatial visualizations of retrieved atlas spots and inspections of their evidence cards (Fig. 5b). For example, a query for tumor-adjacent TLS-like regions highlighted the retrieved spots and displayed their ranked evidence cards for retrieval-grounded analysis. To visualize the reference evidence bank of 2.23 million measured ST spots, we projected the evidence-card embeddings in two dimensions using t-distributed stochastic neighbor embedding (t-SNE) [60] (Fig. 5c). The projection was broadly organized by tissue context, with the finest local structure reflecting distinct molecular states.

We next examined whether the natural-language and H&E–image queries would retrieve biologically relevant atlas spots and support follow-up analyses (Fig. 5d,e). For a natural-language query seeking heart spots with active muscle pathways, the highest-scoring spots

were localized to myocardial regions expressing cardiomyocyte markers *MYH6*, *TNNT2*, *ACTC1*, and *MYL2*, together with muscle contraction and oxidative phosphorylation programs [61]. The follow-up questions focused on the top-ranked region, which HistAgent interpreted as a metabolically active myocardial microenvironment dominated by cardiomyocyte programs. From an H&E image of a breast tumor region, HistAgent retrieved measured ST atlas regions with the epithelial markers *EPCAM*, *KRT8*, and *KRT18* and proliferation markers *MKI67* and *TOP2A* [62], providing molecular context for subsequent analyses of proliferative epithelial and tumor-stromal programs.

Finally, we evaluated HistAgent's natural-language retrieval performance across 40 ST slides using five groups of biological queries targeting tumor epithelium, normal epithelium, tumor pathology, cortical layers, and TLS niches (Fig. 5f and Supplementary Table 5). Expert annotations defined the relevant spots for each query [25, 32, 33]. Using identical query strings and an atlas-wide candidate bank, we compared HistAgent with CellWhisperer [63], a multimodal model that aligns single-cell transcriptomes with text (Supplementary Fig. S9). HistAgent retrieved spots using embeddings of evidence cards containing the molecular and spatial context, whereas CellWhisperer used transcriptome embeddings alone. Across the five query groups, HistAgent achieved a higher AUPRC than CellWhisperer for tumor epithelium (0.641 versus 0.473), normal epithelium (0.637 versus 0.146), tumor pathology (0.494 versus 0.378), cortical layers (0.262 versus 0.213), and TLS niches (0.161 versus 0.079). It also achieved a higher normalized discounted cumulative gain at 50 (nDCG@50) and Precision@50 in every group, outperforming CellWhisperer in all 15 group-by-metric comparisons (one-sided paired tests, $P \leq 0.002$; slide-level results are shown in Supplementary Fig. S10). HistAgent's nDCG advantage persisted across all evaluated cutoffs (Supplementary Fig. S11). Moreover, an evidence-grounding audit performed across 22 query scenarios found that all supported claims were traceable to cited evidence-card fields, with no unsupported gene mentions (Supplementary Fig. S12). Collectively, these results demonstrate HistAgent's ability to extend beyond regional interpretation to include atlas-scale retrieval of measured ST evidence.

Discussion

In this study, we developed HistAgent — a unified framework that makes molecular inferences based on routine H&E images and performs evidence-grounded analyses of local tissue states. HistAgent is able to recover the spatial biological findings within tissues and molecular information obtained via ST and supports standard ST analyses. Its intermediate visual-omics representations further support whole-slide clinical predictions, and natural-language and H&E–image queries can be used to retrieve data from a measured ST evidence bank. HistAgent therefore transforms routine histology from conventional image-based prediction into spatially resolved molecular analysis grounded in molecular and spatial evidence.

At the spot level, HistAgent recovers the local molecular composition more accurately

than the benchmarks STPath and OmiCLIP. Its top 50 gene ranking depends less directly on expression scale and continuous expression values, reducing its sensitivity to normalization and batch effects across ST datasets [15, 17]. HistAgent also learns to generate a list of top 50 genes directly from H&E images. By contrast, STPath predicts continuous gene expression through masked gene-expression modeling [16], while OmiCLIP uses contrastive learning to align H&E images with the top 50 gene rankings and obtains expression predictions through reference retrieval [17]. HistAgent further conditions each gene on those generated previously, allowing dependencies within the ranking to inform subsequent predictions. Its training corpus also includes both human and mouse tissues, whereas the reported pretraining data for OmiCLIP are restricted to human samples. These differences are consistent with HistAgent's stronger recovery of the top-ranked genes across both human and mouse organ cohorts.

At the gene level, HistAgent more accurately reproduces the variation in spatial expression of individual genes. The spot-centered H&E view captures the morphology within each ST capture spot, whereas the contextual view provides information about the surrounding tissue context. This multi-scale input allows each prediction to reflect both the local morphology and the surrounding epithelial, stromal, or immune environment. OmiCLIP instead encodes each tissue patch independently during image–transcriptome alignment [17]. HistAgent's higher PCCs indicate that its ranked molecular readouts preserve not only the dominant gene expression at each spot but also its spatial variation across the tissue. The improved recovery of standard ST analyses further shows that information on tissue organization, cell-type composition, and functional programs is retained.

The spatial agentic module uses these ranked molecular readouts to answer region-specific questions. PathChat answers questions using pathology images, GeneAgent interprets gene sets using domain databases, and CellWhisperer supports natural-language explorations of measured single-cell transcriptomes [26, 27, 63]. In contrast, HistAgent organizes the top ranked genes, inferred cell composition, functional programs, and spatial context into a spot-level evidence card. Fine-tuning on multi-turn evidence-reasoning traces enables the LLM to select question-relevant information and link each conclusion to its molecular and spatial evidence. For instance, in the TLS case, HistAgent combined immunoglobulin and antigen-presentation genes with B-cell, plasma-cell, immune-program, and stromal-context evidence associated with TLS organization [25]. The evidence cards therefore support multi-turn analysis while preserving the evidence behind each interpretation.

The whole-slide results show that HistAgent's intermediate visual-omics representations also retain information relevant to the formation of clinical predictions. Pathology foundation models learn representations through pretraining with histology images only or combined image and text [47–51, 64]. HistAgent is instead trained on spatially paired H&E and ST data, encouraging its representations to retain H&E features associated with local gene expression, cell composition, and functional programs. Its multi-scale input additionally captures the surrounding tissue context before slide-level aggregation occurs.

These properties help to explain its stronger performance during molecular subtype, tumor grade, and prognostic-risk predictions, which depend on both the local morphology and the organization of tumor, stromal, and immune states.

The measured ST evidence bank extends the scope of this analysis beyond individual tissue slides. CellWhisperer aligns transcriptomes with textual descriptions but does not incorporate the spatial tissue context or support H&E image queries [63]. HistAgent retrieves evidence cards containing the top-ranked genes, inferred cell composition, functional programs, and spatial context. Natural-language queries can therefore be matched against multiple forms of biological evidence. For an H&E image query, HistAgent first generates a predicted evidence card and then retrieves atlas spots with related measured ST evidence. This design is consistent with its higher retrieval performance across all evaluated biological query groups. HistAgent enables the tissue states identified from H&E images to be examined against measured molecular profiles from other slides.

Several directions remain for extending the role of HistAgent. The visual-omics foundation model was trained primarily on human and mouse 10x Visium data, and broader coverage of ST platforms, spatial resolutions, tissues, and diseases will be beneficial. Direct generation is limited to the top 50 genes and may omit lower-ranked signals associated with rare cell populations or subtle biological processes. Reference-guided expression inference expands these rankings for standard ST analyses, but the resulting estimates depend on the independent ST reference and are not directly measured transcript counts. Agent responses likewise depend on predicted molecular readouts and rank-derived analyses, whereas atlas retrieval depends on the coverage of the measured ST evidence bank. Whole-slide clinical prediction will also require prospective evaluations in external clinical settings. Expanding the paired H&E–ST training corpus and evidence bank, and providing more expert and experimental validation, could broaden the questions that HistAgent can address, while complementing measured ST analyses and expert investigation.

Overall, HistAgent provides a unified framework for spatial molecular analysis from routine H&E images, spanning ranked molecular readout generation, traceable biological interpretation, whole-slide clinical prediction, and atlas-scale retrieval. Its substantially lower cost and shorter turnaround time relative to conventional ST make these capabilities more accessible across biological and clinical research.

Methods

Development of the visual-omics foundation model for HistAgent

Paired H&E–ST data curation and image preprocessing. We constructed and curated the HistAgent development corpus by integrating paired hematoxylin and eosin (H&E) images and spatial transcriptomics (ST) profiles obtained primarily from the Gene Expression Omnibus and 10x Genomics databases. The processed corpus and supporting resources are available at https://huggingface.co/wli13/HistAgent. The corpus comprised data on human and mouse tissue slides profiled using the 10x Genomics Visium

platform. Each spot-level record linked one tissue-covered capture spot to its measured ST profile, a local H&E image centered on the spot, and a contextual H&E image capturing the surrounding tissue. Species labels were restricted to human or mouse. We harmonized synonymous source labels to 32 canonical organ categories and assigned mixed or unspecified tissues to an *Unknown* category.

The final development corpus contained 2,226,206 paired spots from 936 tissue slides, encompassing 1,301,160 human spots and 925,046 mouse spots. We partitioned the data at the slide level. The model-development partition contained 1,889,336 spots from 801 slides, and the held-out test set contained 331,560 spots from 135 slides, with no slide shared between the two partitions.

For spot i, we denote the local and contextual image inputs by $\mathbf{I}_i^{\text{local}}$ and $\mathbf{I}_i^{\text{context}}$, respectively. The local view was centered on the nominal 55-µm Visium capture region. The contextual view was centered to the same coordinate and covered four times the width and height of the local view. Local images were resized to 224×224 pixels. Contextual images were stored at 256×256 pixels and center-cropped to 224×224 pixels. After RGB conversion and ImageNet normalization, both inputs had shape $3 \times 224 \times 224$.

Ranked molecular targets and gene-token vocabulary. Let $x_{i,g}$ denote the measured count for gene g at spot i. Gene symbols were converted to uppercase, species prefixes were removed, and counts for duplicate symbols created by harmonization were summed. We removed genes that appeared in a fixed 2,177-gene housekeeping exclusion list [65] and genes matching the predefined technical prefixes `MT-`, `MTRNR`, `RPL`, `RPS`, `ENSMUSG`, `ENSGO`, and `DEPRECATED_`. Retained counts were normalized and transformed as

$$\widetilde{x}_{i,g} = \log\left(1 + 10^4 \frac{x_{i,g}}{\sum_h x_{i,h}}\right). \tag{1}$$

Genes with nonzero expression were ordered from highest to lowest $\widetilde{x}_{i,g}$. The first 50 genes defined the measured ranked molecular target

$$\boldsymbol{\pi}_i = (g_{i,1}, \ldots, g_{i,L_i}), \qquad 1 \le L_i \le 50, \tag{2}$$

where $g_{i,r}$ is the gene at rank r. If fewer than 50 retained genes were detected, the shorter ranking was preserved without inserting artificial genes.

After harmonization and gene exclusion, the unified gene-token vocabulary contained 32,019 canonical genes. Four special tokens were added: padding (`[PAD]`), beginning-of-sequence (`[BOS]`), end-of-sequence (`[EOS]`) and unknown (`[UNK]`), giving a decoder vocabulary size of $V = 32,023$. We indexed the shifted decoder sequence as

$$y_{i,0} = \texttt{[BOS]}, \qquad y_{i,r} = g_{i,r} \;\; (1 \le r \le L_i), \qquad y_{i,L_i+1} = \texttt{[EOS]}. \tag{3}$$

The serialized sequence $(y_{i,0}, \ldots, y_{i,L_i+1})$ was padded to a maximum length of 52 tokens. The same unified vocabulary was used across species, organs, and datasets.

Multi-scale visual encoding and contextual memory. The visual-omics foundation model transforms the two H&E inputs and their metadata into a fixed-width contextual memory. A shared GigaPath tile encoder E_ϕ [47] maps each image to a 1,536-dimensional visual vector:

$$\mathbf{v}_i^{\text{local}} = E_\phi(\mathbf{I}_i^{\text{local}}), \qquad \mathbf{v}_i^{\text{context}} = E_\phi(\mathbf{I}_i^{\text{context}}), \qquad \mathbf{v}_i^{(\cdot)} \in \mathbb{R}^{1536}. \tag{4}$$

The two vectors are processed by separate Perceiver-style resamplers, P_{local} and P_{context} [66], which produce 16 latent tokens per view:

$$\mathbf{L}_i^{\text{local}} = P_{\text{local}}(\mathbf{v}_i^{\text{local}}), \qquad \mathbf{L}_i^{\text{context}} = P_{\text{context}}(\mathbf{v}_i^{\text{context}}), \qquad \mathbf{L}_i^{(\cdot)} \in \mathbb{R}^{16\times 512}. \tag{5}$$

Each resampler used eight attention heads. Species s_i and organ o_i were represented by learned 512-dimensional embeddings $\mathbf{e}_{s_i}$ and $\mathbf{e}_{o_i}$. Concatenating the metadata and visual tokens formed the decoder memory

$$\mathbf{M}_i = \left[\mathbf{e}_{s_i}; \mathbf{e}_{o_i}; \mathbf{L}_i^{\text{local}}; \mathbf{L}_i^{\text{context}}\right] \in \mathbb{R}^{34\times 512}. \tag{6}$$

Thus, the visual encoder has the explicit mapping

$$(\mathbf{I}_i^{\text{local}}, \mathbf{I}_i^{\text{context}}, s_i, o_i) \longmapsto \mathbf{M}_i \in \mathbb{R}^{34\times 512}. \tag{7}$$

We adapted GigaPath using low-rank adaptation (LoRA) modules [67]. For an adapted weight $\mathbf{W}$, LoRA parameterized the update as

$$\mathbf{W}' = \mathbf{W} + \frac{\alpha_{\text{L}}}{r_{\text{L}}}\mathbf{BA}, \tag{8}$$

where $\mathbf{A} \in \mathbb{R}^{r_{\text{L}}\times d_{\text{in}}}$, $\mathbf{B} \in \mathbb{R}^{d_{\text{out}}\times r_{\text{L}}}$, $r_{\text{L}} = 16$, $\alpha_{\text{L}} = 32$ and the LoRA dropout probability was 0.05. LoRA modules were applied to the attention query–key–value and output projections and to both feed-forward layers. GigaPath parameters outside these modules remained frozen.

Autoregressive generation of ranked molecular readouts. An autoregressive Transformer decoder D_θ [68] maps the contextual memory to a ranked molecular readout. The decoder comprised six layers, eight attention heads, a hidden dimension of 512 and dropout of 0.1. Gene tokens used learned 512-dimensional embeddings and learned positional embeddings. Causal self-attention encoded the preceding gene tokens, and cross-attention connected the partial output sequence to $\mathbf{M}_i$. For output position r, the decoder produced

$$\mathbf{p}_{i,r} = \text{softmax}\left[D_\theta(\mathbf{y}_{i,<r}, \mathbf{M}_i)\right] \in [0,1]^V. \tag{9}$$

Here, $\mathbf{y}_{i,<r} = (y_{i,0}, \ldots, y_{i,r-1})$ contains the beginning-of-sequence token and all preceding targets. The conditional probability of a measured target is therefore

$$p_\theta(\mathbf{y}_i \mid \mathbf{I}_i^{\text{local}}, \mathbf{I}_i^{\text{context}}, s_i, o_i) = \prod_{r=1}^{L_i+1} p_\theta(y_{i,r} \mid \mathbf{y}_{i,<r}, \mathbf{M}_i), \tag{10}$$

where the final predicted token is `[EOS]`. Excluding special tokens, the model output is the ordered gene sequence

$$\widehat{\boldsymbol{\pi}}_i = (\widehat{g}_{i,1}, \ldots, \widehat{g}_{i,\widehat{L}_i}), \qquad \widehat{L}_i \le 50. \tag{11}$$

The decoder contained a final linear layer over the 32,023-token vocabulary.

Rank-weighted objective and model optimization. We optimized the visual-omics foundation model by teacher forcing. Let $m_{i,r} = 1$ for a non-padding target at shifted decoder position r and $m_{i,r} = 0$ otherwise. For the fixed shifted length $T = 51$, the rank weight was

$$w_r = \frac{T - r + 1}{T}. \tag{12}$$

The training objective was the normalized rank-weighted cross-entropy

$$\mathcal{L}_{\text{rank}} = \frac{\sum_i \sum_{r=1}^{T} m_{i,r} w_r \ \text{CE}_\epsilon(\mathbf{p}_{i,r}, y_{i,r})}{\sum_i \sum_{r=1}^{T} m_{i,r} w_r}, \tag{13}$$

where CE_ϵ denotes cross-entropy with label-smoothing parameter $\epsilon = 0.1$. This weighting means earlier genes in the measured ranking have a larger contribution to the loss.

During training, Gaussian noise with a standard deviation of 0.02 was added to the visual features. Species and organ conditions were independently replaced by their unknown embedding with a probability of 0.15. We trained the model for 30 epochs with a per-device batch size of 24. Optimization used AdamW [69] with an initial learning rate of 1×10^{-4}, a weight decay of 0.05, and gradient clipping at a maximum norm of 1.0. The learning rate increased linearly during the first 5% of the optimization steps and then followed cosine decay.

At inference, decoding began with `[BOS]` and proceeded greedily. At each step, `[PAD]`, `[BOS]`, `[UNK]` and every previously generated gene token were masked. Decoding stopped when `[EOS]` was emitted or after 50 genes. Because the vocabulary was constructed from canonical genes after housekeeping-gene and technical-prefix exclusion, the generated rankings required no additional gene filtering before downstream analysis.

Release and held-out-evaluation checkpoints. We developed a release checkpoint using the 2.23-million-spot corpus. For the molecular-recovery evaluation, we trained a separate held-out-evaluation checkpoint after removing all 135 evaluation slides from the development corpus. No spots from these slides contributed to optimization, validation,

or model selection. Analyses of the 135 held-out slides employed the held-out-evaluation checkpoint. Both checkpoints used the same architecture and optimization procedure.

Construction of the spatial agentic module

Evidence-card representation. The spatial agentic module organizes each ranked molecular readout and its spatial metadata into a structured evidence card. For spot i, let $\mathbf{c}_i$ denote its spatial coordinate, $\mathcal{N}_i$ its spatial neighborhood, $\mathcal{D}_{\text{marker}}$ the cell-marker resource, and $\mathcal{D}_{\text{program}}$ the functional gene-set resource. Evidence-card construction is represented as

$$\mathbf{C}_i = \mathcal{C}\left(\boldsymbol{\pi}_i, \mathbf{c}_i, s_i, o_i, \mathcal{N}_i; \mathcal{D}_{\text{marker}}, \mathcal{D}_{\text{program}}\right), \tag{14}$$

where $\boldsymbol{\pi}_i$ denotes the gene ranking derived from a measured ST spot and is replaced by $\widehat{\boldsymbol{\pi}}_i$ for an H&E-derived spot. Each evidence card contains data on the top-ranked genes, inferred cell composition, functional programs and spatial context. The inferred cell composition was represented by marker-based cell-state scores derived from human and mouse marker sets in CellMarker 2.0 [70]. Functional programs were derived from species-matched Reactome pathways and Gene Ontology biological-process gene sets [37, 38]. The spatial context represented a summary of the six nearest spots within the same slide, including their distances, inferred cell composition, and local heterogeneity. Each card retained its spot and slide identifiers, species, canonical organ category, coordinates, and provenance, as measured ST or H&E-derived evidence.

Conversational supervision. We fine-tuned Qwen3-8B [71] on 50,000 five-turn evidence-reasoning traces generated by GPT-5.5 with xHigh reasoning effort. Each trace paired an evidence card derived from measured ST with questions addressing the spot identity, supporting molecular evidence, functional programs, local microenvironment, and gene–program relationships (Supplementary Table 6). Assistant responses contained natural-language answers and structured `evidence_refs` that linked molecular and spatial claims to specific evidence-card fields.

Training records followed fixed dialogue and evidence-card schemas. Automated checks rejected malformed records, invalid evidence references, genes absent from the associated card, inconsistent gene–program relationships, unsupported spatial statements, and answers that omitted uncertainty when evidence was limited. A separate Qwen-Plus assessed the evidence alignment and internal consistency before a record was retained for supervised fine-tuning. Traces were partitioned at the slide level so that cards from the same slide did not occur in both the optimization and validation data. The complete corpus, prompt templates, schemas and quality-control metadata are available at https://huggingface.co/wli13/HistAgent.

Let $\mathbf{u} = (u_1, \ldots, u_T)$ denote the tokenized dialogue and let $a_t = 1$ for an assistant-

response token and $a_t = 0$ otherwise. The supervised fine-tuning objective was

$$\mathcal{L}_{\mathrm{SFT}} = -\frac{1}{\sum_{t=1}^{T} a_t} \sum_{t=1}^{T} a_t \log p_\psi(u_t \mid u_{<t}, \mathbf{C}_i)\,. \tag{15}$$

Only assistant-response tokens contributed to the loss. Dialogues were formatted with the Qwen3 chat template and truncated to 8,192 tokens. LoRA modules with a rank of 16, scaling parameter of 32, and dropout 0.05 were applied to the attention projections and feed-forward projections. Training used bfloat16 precision, gradient checkpointing, AdamW with a learning rate 1×10^{-4} and weight decay 0.01, an effective batch size of 64, gradient clipping at 1.0, a 3% warm-up, and cosine learning-rate decay. We trained Qwen3 for three epochs and selected the checkpoint with the lowest validation loss.

Multi-turn inference. At inference, the fine-tuned Qwen3-8B model received the user question, conversation history, and relevant evidence cards as structured context. The response policy required molecular and spatial statements to cite the supporting evidence-card fields and required uncertainty when the available evidence was sparse or ambiguous. Conversation history and evidence context were retained across turns, enabling successive questions to be answered about the same region without reconstructing the evidence.

We evaluated the model's evidence grounding in held-out query scenarios under full-evidence, evidence-removal, and counterfactual conditions. Detailed scenario construction, response generation, and scoring definitions are provided in Supplementary Methods.

Reference-guided expression inference for the 135-slide analysis

Mapping ranked molecular readouts to reference-guided expression estimates. For molecular recovery and standard ST analyses requiring a spot-by-gene matrix, we mapped each ranked molecular readout to reference-guided expression estimates. This post hoc procedure did not modify the directly generated ranked molecular readout.

For a rank-feature gene set $\mathcal{G}_R$, the sparse inverse-rank representation of spot i was

$$r_{i,g} = \begin{cases} 1/k, & \text{if gene } g \text{ occurs at rank } k \text{ in } \boldsymbol{\pi}_i, \\ 0, & \text{otherwise,} \end{cases} \qquad g \in \mathcal{G}_R. \tag{16}$$

The 1,500 genes occurring most frequently in the reference rankings defined $\mathcal{G}_R$. For species s, let $\mathbf{R} \in \mathbb{R}^{n \times 1500}$ collect the inverse-rank vectors of n reference spots. We defined a fixed species-specific output axis $\mathcal{G}_X^{(s)}$ from canonical genes present in at least 95% of the selected reference slides for that species after applying the same housekeeping-gene and technical-prefix exclusions used to construct the ranked molecular targets. The resulting axes contained $G_X^{(\mathrm{human})} = 15{,}482$ human genes and $G_X^{(\mathrm{mouse})} = 16{,}811$ mouse genes.

Reference selection and ridge mapping. For the 135-slide benchmark, each query study was removed in full before constructing the independent ST reference. Reference slides were selected from the same species and organ category. At least five same-organ slides were required; otherwise, other organs of the same species were used. Up to 80 slides and 20,000 reference spots were sampled deterministically for each species–organ group. Neither the query slides nor their source studies were used to fit the mapping.

Reference expression targets were normalized to 10,000 counts per spot and log-transformed before fitting. For gene $g \in \mathcal{G}_X^{(s)}$, let $\mathcal{I}_g$ index the reference spots from slides in which g was measured. Structural absence from a reference slide was treated as missing and was not encoded as zero. We fitted the gene-specific ridge coefficient without an intercept as

$$\widehat{\mathbf{w}}_g = \arg\min_{\mathbf{w}_g} \left\| \mathbf{R}_{\mathcal{I}_g} \mathbf{w}_g - \mathbf{x}_{\mathcal{I}_g, g} \right\|_2^2 + \lambda \left\| \mathbf{w}_g \right\|_2^2, \qquad \lambda = 10. \tag{17}$$

Genes sharing the same availability mask could be fitted jointly as a multi-output regression. A gene not measured in any slide of the selected reference subset was omitted from the query-specific output rather than imputed as zero. For query rank matrix $\mathbf{R}_q$, the continuous output for each retained gene g was

$$\widehat{\mathbf{x}}_{q,g} = \left[\mathbf{R}_q \widehat{\mathbf{w}}_g \right]_+, \tag{18}$$

where $[\cdot]_+$ clips negative values to zero. Because the reference-guided expression inference was applied after the ranked molecular readout generation, changing the output axis did not alter the directly generated ranked molecular readout or spot mapping.

Downstream analyses

Recovery of published spatial biological findings. We evaluated 12 published spatial biological findings of 22 ST slides from five studies and one public 10x Genomics resource [25, 30–34]. The evaluation comprised 29 slide–finding pairs spanning immune, stromal, fibrotic, tumor, epithelial, and anatomical states. Each finding was defined by a literature-supported marker set and an expert annotation from the source study or associated data. Slide identifiers, marker sets, and annotation sources are provided in Supplementary Table 1.

HistAgent, STPath, OmiCLIP, and paired measured ST were aligned to the same spots and effective marker genes for each slide–finding pair. HistAgent-ranked molecular readouts were converted into reference-guided expression estimates before signature calculation. For a biological finding defined by marker set $\mathcal{G}_f$, let $x_{i,g}^{(m)}$ denote the value for gene g at spot i from method m. Within each slide and method, every available marker gene was standardized across spots:

$$z_{i,g}^{(m)} = \frac{x_{i,g}^{(m)} - \mu_g^{(m)}}{\sigma_g^{(m)} + 10^{-8}}. \tag{19}$$

The spatial signature score was the mean standardized marker value

$$S_{i,f}^{(m)} = \frac{1}{|\mathcal{G}_f^*|} \sum_{g \in \mathcal{G}_f^*} z_{i,g}^{(m)}, \tag{20}$$

where $\mathcal{G}_f^*$ is the common effective marker set available to all compared methods in that slide. The resulting score vector $\mathbf{S}_f^{(m)}$ was compared with the expert annotation using AUPRC. AUPRC was calculated for each slide–finding pair and then averaged without slide weighting within each finding and across the 12 findings.

For the renal cell carcinoma tertiary lymphoid structure (TLS)-like niche, we used a 25-gene literature-supported TLS signature and the corresponding expert annotation [25]. Signature scores in annotated TLS and non-TLS spots were compared with a two-sided Mann–Whitney U test, and localization was quantified by AUPRC. The illustrative multi-turn analysis used the spot with the highest HistAgent-derived TLS signature score within the annotated TLS region.

Molecular recovery. We evaluated molecular recovery at two complementary levels across the 135 held-out slides. At the spot level, each method produced a ranked molecular readout $\widehat{\boldsymbol{\pi}}_i^{(m)}$, which was compared with the 50 most highly expressed genes in the measured ST data for the same spot. At the gene level, each method produced a continuous vector $\widehat{\mathbf{x}}_g^{(m)}$ over the common spots in a slide, and was compared with the corresponding measured vector $\mathbf{x}_g$. HistAgent supplied its directly generated, ranked molecular readout for the spot-level analysis and reference-guided expression estimates for the gene-level analysis. Highly variable genes were selected only from library-size-normalized and log-transformed measured ST data. We evaluated progressively broader sets containing 50 to 500 highly variable genes in increments of 50.

All methods were aligned to the same ordered intersection of spot barcodes and canonical gene symbols in each slide. The compared model weights remained frozen. Paired measured ST data were used only to define the targets, evaluated genes, and measured profiles; they were not supplied as input to any image-based method.

Standard ST analyses. We evaluated spatially variable gene detection, spatial domain identification, cell-type deconvolution, differential expression, and pathway enrichment. For all image-based methods, each method used the same spot and gene intersections for each task, and continuous estimates were restricted to genes supported by the method's original top-50 prediction before task-specific analysis. Predicted outputs were compared with corresponding results derived from measured ST. Detailed workflows and metric definitions are provided in Supplementary Methods.

Clinical prediction from frozen tile representations. For clinical slide j, let $\mathbf{M}_{j,t} \in \mathbb{R}^{34\times512}$ denote the contextual memory produced by HistAgent's visual-omics foundation model for tile t. The tile-level intermediate visual-omics representation was the mean of

its memory tokens:

$$\mathbf{z}_{j,t}^{\mathrm{HA}} = \frac{1}{34}\sum_{r=1}^{34}\mathbf{M}_{j,t,r} \in \mathbb{R}^{512}. \tag{21}$$

HistAgent's visual-omics foundation model was frozen before representation extraction. A slide containing n_j tiles was represented as a bag of tile representations

$$\mathbf{Z}_j^{(m)} = \left[\mathbf{z}_{j,1}^{(m)}, \ldots, \mathbf{z}_{j,n_j}^{(m)}\right]^\top \in \mathbb{R}^{n_j \times d_m}, \tag{22}$$

where m indexes HistAgent or a comparison encoder, and d_m is its representation dimension.

Every representation was evaluated with the same attention-based multiple-instance learning (ABMIL) model [45, 46]. The instance encoder $f_{\mathrm{inst}}^{(m)}$ projected each input representation onto 512 dimensions, which was followed by layer normalization, a rectified linear activation, and dropout of 0.25, and then onto a 256-dimensional encoded tile representation $\mathbf{h}_{j,t}$. Gated attention assigned the normalized weight

$$a_{j,t} = \frac{\exp\left\{\mathbf{w}^\top\left[\tanh(\mathbf{V}\mathbf{h}_{j,t}) \odot \mathrm{sigmoid}(\mathbf{U}\mathbf{h}_{j,t})\right]\right\}}{\sum_{u=1}^{n_j}\exp\left\{\mathbf{w}^\top\left[\tanh(\mathbf{V}\mathbf{h}_{j,u}) \odot \mathrm{sigmoid}(\mathbf{U}\mathbf{h}_{j,u})\right]\right\}}. \tag{23}$$

The slide representation and prediction were

$$\mathbf{h}_j^{\mathrm{slide}} = \sum_{t=1}^{n_j} a_{j,t}\mathbf{h}_{j,t}, \qquad \boldsymbol{\eta}_j = f_{\mathrm{head}}(\mathbf{h}_j^{\mathrm{slide}}). \tag{24}$$

The prediction head f_{head} was linear. Its output $\boldsymbol{\eta}_j$ was one logit for binary classification, one logit per class for multiclass classification, or one continuous risk score for survival analysis.

For binary label $y_i \in \{0, 1\}$, the predicted probability was $p_i = \mathrm{sigmoid}(\eta_i)$ and the loss was

$$\mathcal{L}_{\mathrm{BCE}} = -\frac{1}{N}\sum_i \left[y_i \log p_i + (1 - y_i)\log(1 - p_i)\right]. \tag{25}$$

For multiclass logits $\boldsymbol{\eta}_i$, the class probabilities were $p_{i,c} = [\mathrm{softmax}(\boldsymbol{\eta}_i)]_c$. With one-hot label $y_{i,c}$, the loss was

$$\mathcal{L}_{\mathrm{CE}} = -\frac{1}{N}\sum_i\sum_c y_{i,c}\log p_{i,c}. \tag{26}$$

For survival endpoints, let T_j be the observed time, δ_j the event indicator, and η_j the model risk associated with training slide j. The negative Cox partial-likelihood loss [72] within a training batch was

$$\mathcal{L}_{\mathrm{Cox}} = -\frac{1}{\sum_j \delta_j}\sum_{j:\delta_j=1}\left[\eta_j - \log\left(\sum_{k:T_k \geq T_j} e^{\eta_k}\right)\right]. \tag{27}$$

Patient-level cross-validation and out-of-fold prediction. We used patient-level outer five-fold cross-validation for every endpoint and method. All slides from each patient were assigned to the same outer fold. For each outer fold, the held-out patients were excluded from model selection, and an inner validation subset drawn only from the remaining four folds was used to select the training epoch. Each outer split was repeated with five random seeds: 42, 142, 242, 342, and 442.

ABMIL models were optimized with Adam [73] at a learning rate of 2×10^{-4} and weight decay of 1×10^{-5} using cosine learning-rate decay. Training continued for at most 30 epochs and stopped after 6 epochs when there was no improvement in the inner-validation metric. During training, at most 4,096 tiles were sampled from each slide. Validation and inference used a maximum of 8,192 tiles per slide.

The selected model for each seed and outer fold was applied once to the held-out patients. Let $\mathcal{S}_i$ denote the slides from patient i and let $R = 5$ denote the number of seeds. The patient-level out-of-fold prediction was

$$\overline{\boldsymbol{p}}_i = \frac{1}{R|\mathcal{S}_i|}\sum_{r=1}^{R}\sum_{j\in\mathcal{S}_i}\boldsymbol{p}_{i,j,r}. \tag{28}$$

Multiclass probabilities were renormalized so sum was one. For survival endpoints, seed-specific risk scores were standardized using the corresponding outer-training patients before being averaged across seeds and slides. This procedure yielded one patient-level out-of-fold probability vector or risk score for each method and endpoint. Point estimates were calculated after pooling the patient-level out-of-fold predictions from all five outer folds rather than by averaging the fold-specific metrics.

For each method and endpoint, 95% confidence intervals were obtained from 10,000 patient-level percentile bootstrap samples. Binary and multiclass endpoints were resampled within class, preserving the observed class counts, whereas survival endpoints were resampled without stratification. HistAgent was compared with the strongest of the eight baselines using paired patient-level bootstrap samples. Each replicate used the same sampled patients for all nine methods and reselected the strongest baseline within that replicate. One-sided P values were used to test the alternative hypothesis that HistAgent outperforms the strongest baseline [74].

Clinical interpretation and survival analyses. We further characterized the clinical predictions through local interpretation and survival analyses. Local tile scores were generated using outer-fold models that excluded the corresponding patient and were used to compare annotated regions. Patient-level clinical performance was evaluated using out-of-fold predictions. Kaplan–Meier and multivariable Cox analyses used the same patient-level out-of-fold risk scores as the clinical benchmark. Detailed procedures are provided in the Supplementary Methods.

Measured ST evidence bank and atlas retrieval. We constructed the measured ST evidence bank from all 2.23 million human and mouse ST spots in the development corpus. Each spot was represented by an evidence card containing the species, canonical organ category, top-ranked genes, inferred cell composition, functional programs and spatial context. Let $\mathcal{B} = \{\mathbf{C}_j\}_{j=1}^{N_B}$ denote these evidence cards. The fields were serialized in a fixed order and embedded with Qwen3-Embedding-8B E_{emb} [59]. The final non-padding hidden state was normalized to give a 4,096-dimensional vector:

$$\mathbf{e}_j = \frac{E_{\text{emb}}(\mathbf{C}_j)}{\|E_{\text{emb}}(\mathbf{C}_j)\|_2} \in \mathbb{R}^{4096}. \tag{29}$$

During online query processing, a natural-language query q was embedded directly with Qwen3-Embedding-8B. Each H&E image query comprised a local tile from a tiled slide and its centered surrounding-tissue view. HistAgent's visual-omics foundation model generated a ranked molecular readout from these images, and the spatial agentic module constructed a predicted evidence card using the same schema as used for the cards derived from measured ST. The predicted card was then embedded with Qwen3-Embedding-8B. For query embedding $\mathbf{e}_q$, the retrieval score was the cosine similarity

$$s(q, j) = \mathbf{e}_q^\top \mathbf{e}_j, \tag{30}$$

because both vectors were L_2 normalized. Both query types were ranked against all 2.23 million measured ST spots in the full 4,096-dimensional shared retrieval-embedding space. The returned entries were the measured ST spots and their associated evidence cards.

Natural-language retrieval benchmark. Natural-language retrieval was evaluated across 40 ST slides and five biological query groups targeting tumor epithelium, normal epithelium, tumor pathology, cortical layers, and TLS niches. Expert annotations defining the relevant spots for each query were available. Because each slide could contribute more than one query or target, the evaluation comprised 164 query–slide units. The query wording, annotation targets, data sources, and slide counts are provided in Supplementary Table 5.

HistAgent and CellWhisperer received identical natural-language query strings and ranked the same atlas-wide candidate bank comprising 2.23 million measured ST spots. HistAgent ranked spots by similarity between the Qwen3-Embedding-8B query vector and the evidence-card embeddings derived from measured ST. CellWhisperer processed the measured spot transcriptomes and query strings with its released transcriptome and text encoders and ranked the same spots using its similarity score [63]. Expert relevance annotations of relevance were used only in the evaluations.

Atlas-wide rankings were evaluated for each query–slide unit and the expert annotations of relevance for the corresponding slide. Metrics were then averaged within each biological query group. Query–slide comparisons between HistAgent and CellWhisperer used query–slide units as matched observations.

Comparisons of methods

We compared HistAgent with STPath and OmiCLIP in spatial molecular recovery; GigaPath, UNI v2, MUSK, CHIEF, CONCH, OmiCLIP, STPath, and Path2Space in clinical prediction; and CellWhisperer for natural-language atlas retrieval. All pretrained encoders remained frozen, and task-specific ABMIL models were trained only within the corresponding outer training folds. Within each task, methods were evaluated on matched inputs and evaluation units, while clinical representations were assessed using the same patient folds and ABMIL framework. Model-specific representations, implementation settings, and software resources are detailed in the Supplementary Methods.

Evaluation metrics

Ranked molecular recovery was evaluated using HitRate@50 and the mean average precision at 50 (mAP@50). For spatial molecular analyses, we used the Pearson correlation coefficient, AUPRC, and task-specific agreement metrics, including Moran's-I concordance, normalized mutual information, adjusted Rand index, cell-type composition agreement, differential-expression agreement, and pathway nDCG. For clinical prediction, we used AUROC, macro-AUROC, and the concordance index, whereas for atlas retrieval we used AUPRC, nDCG@50, and Precision@50. Unless stated otherwise, spot- or gene-level values were first aggregated within each slide, clinical outcomes were evaluated at the patient level, and retrieval comparisons used query–slide units as matched observations. Complete metric definitions, aggregation rules, and the handling of undefined values are described in Supplementary Methods.

Datasets and preprocessing for downstream evaluation

Molecular recovery was evaluated with 135 held-out ST slides comprising 79 human and 56 mouse slides from 15 species–organ cohorts. The clinical benchmark contained 20 endpoints across The Cancer Genome Atlas, the Clinical Proteomic Tumor Analysis Consortium, and associated image repositories. Natural-language retrieval was evaluated on 40 ST slides using five biological query groups and 164 query–slide units. The slide identifiers, endpoint definitions, whole-slide image preprocessing, and task-specific quality-control procedures are provided in the Supplementary Methods and Supplementary Tables 2–5.

Data and code availability

All processed data, model weights, and supporting resources generated in this study are available from the HistAgent repository on Hugging Face (https://huggingface.co/wli13/HistAgent). Source code is available from GitHub (https://github.com/zipging/HistAgent), and the interactive HistAgent interface is available at https://histagent.bio/histagent. Public source datasets remain available from the repositories and

accessions cited for each cohort.

References

[1] J. van der Laak, G. Litjens, F. Ciompi, Deep learning in histopathology: the path to the clinic. Nature Medicine **27**(5), 775–784 (2021)

[2] K. Bera, K.A. Schalper, D.L. Rimm, V. Velcheti, A. Madabhushi, Artificial intelligence in digital pathology—new tools for diagnosis and precision oncology. Nature Reviews Clinical Oncology **16**(11), 703–715 (2019)

[3] P.L. Ståhl, F. Salmén, S. Vickovic, A. Lundmark, J.F. Navarro, J. Magnusson, S. Giacomello, M. Asp, J.O. Westholm, M. Huss, A. Mollbrink, S. Linnarsson, S. Codeluppi, Å. Borg, F. Pontén, P.I. Costea, P. Sahlén, J. Mulder, O. Bergmann, J. Lundeberg, J. Frisén, Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. Science **353**(6294), 78–82 (2016)

[4] A. Rao, D. Barkley, G.S. França, I. Yanai, Exploring tissue architecture using spatial transcriptomics. Nature **596**(7871), 211–220 (2021)

[5] S.K. Longo, M.G. Guo, A.L. Ji, P.A. Khavari, Integrating single-cell and spatial transcriptomics to elucidate intercellular tissue dynamics. Nature Reviews Genetics **22**(10), 627–644 (2021)

[6] L. Larsson, J. Frisén, J. Lundeberg, Spatially resolved transcriptomics adds a new dimension to genomics. Nature Methods **18**(1), 15–18 (2021)

[7] V. Marx, Method of the year: spatially resolved transcriptomics. Nature Methods **18**(1), 9–14 (2021)

[8] L. Moses, L. Pachter, Museum of spatial transcriptomics. Nature Methods **19**(5), 534–546 (2022)

[9] B. He, L. Bergenstråhle, L. Stenbeck, A. Abid, A. Andersson, Å. Borg, J. Maaskola, J. Lundeberg, J. Zou, Integrating spatial gene expression and breast tumour morphology via deep learning. Nature Biomedical Engineering **4**(8), 827–834 (2020)

[10] Y. Zeng, Z. Wei, W. Yu, R. Yin, Y. Yuan, B. Li, Z. Tang, Y. Lu, Y. Yang, Spatial transcriptomics prediction from histology jointly through Transformer and graph neural networks. Briefings in Bioinformatics **23**(5), bbac297 (2022)

[11] E.D. Shulman, E.M. Campagnolo, R. Lodha, Y. Chung, A. Stemmer, T. Cantore, B. Ru, T.G. Chang, S. Biswas, S.R. Dhruba, S. Patiyal, S. Patkar, A. Wang, R.K. Barman, C. Wang, R. Paul, S.C. Kalisetty, T. Hu, M.P. Nasrallah, E. Patrick, J. Yang, Y. Yuan, K. Sargsyan, A. Plotkin, P.S. Rajagopal, S.J. Sammut, S. Lipkowitz, P. Jiang, C. Caldas, S.R.V. Knott, K. Aldape, J.S. Lee, D.T. Hoang, E. Ruppin, AI-predicted spatial transcriptomics unlocks breast cancer biomarkers from pathology. Cell **189**(14), 4225–4240.e25 (2026)

[12] R. Xie, K. Pang, S. Chung, C. Perciani, S. MacParland, B. Wang, G. Bader, *Spatially Resolved Gene Expression Prediction from Histology Images via Bi-modal Contrastive Learning*, in *Advances in Neural Information Processing Systems 36* (2023), pp. 70626–70637

[13] Y. Jia, J. Liu, L. Chen, T. Zhao, Y. Wang, THItoGene: a deep learning method for predicting spatial transcriptomics from histological images. Briefings in Bioinformatics **25**(1), bbad464 (2024)

[14] W. Min, Z. Shi, J. Zhang, J. Wan, C. Wang, Multimodal contrastive learning for spatial gene expression prediction using histology images. Briefings in Bioinformatics **25**(6), bbae551 (2024)

[15] C. Wang, A.S. Chan, X. Fu, S. Ghazanfar, J. Kim, E. Patrick, J.Y.H. Yang, Benchmarking the translational potential of spatial gene expression prediction from histology. Nature Communications **16**(1), 1544 (2025)

[16] T. Huang, T. Liu, M. Babadi, R. Ying, W. Jin, STPath: a generative foundation model for integrating spatial transcriptomics and whole-slide images. npj Digital Medicine **8**(1), 659 (2025)

[17] W. Chen, P. Zhang, T.N. Tran, Y. Xiao, S. Li, V.V. Shah, H. Cheng, K.W. Brannan, K. Youker, L. Lai, L. Fang, Y. Yang, N.T. Le, J.i. Abe, S.H. Chen, Q. Ma, K. Chen, Q. Song, J.P. Cooke, G. Wang, A visual–omics foundation model to bridge histopathology with spatial transcriptomics. Nature Methods **22**(7), 1568–1582 (2025)

[18] M.D. Luecken, F.J. Theis, Current best practices in single-cell RNA-seq analysis: a tutorial. Molecular Systems Biology **15**(6), e8746 (2019)

[19] H.T.N. Tran, K.S. Ang, M. Chevrier, X. Zhang, N.Y.S. Lee, M. Goh, J. Chen, A benchmark of batch-effect correction methods for single-cell RNA sequencing data. Genome Biology **21**(1), 12 (2020)

[20] A. Radford, J.W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, I. Sutskever, *Learning transferable visual models from natural language supervision*, in *Proceedings of the 38th International Conference on Machine Learning*, *Proceedings of Machine Learning Research*, vol. 139 (2021), pp. 8748–8763

[21] C. Sautès-Fridman, F. Petitprez, J. Calderaro, W.H. Fridman, Tertiary lymphoid structures in the era of cancer immunotherapy. Nature Reviews Cancer **19**(6), 307–325 (2019)

[22] R. Cabrita, M. Lauss, A. Sanna, et al., Tertiary lymphoid structures improve immunotherapy and survival in melanoma. Nature **577**(7791), 561–565 (2020)

[23] B.A. Helmink, S.M. Reddy, J. Gao, et al., B cells and tertiary lymphoid structures promote immunotherapy response. Nature **577**(7791), 549–555 (2020)

[24] T.N. Schumacher, D.S. Thommen, Tertiary lymphoid structures in cancer. Science **375**(6576), eabf9419 (2022)

[25] M. Meylan, F. Petitprez, E. Becht, A. Bougoüin, G. Pupier, A. Calvez, I. Giglioli, V. Verkarre, G. Lacroix, J. Verneau, C.M. Sun, P. Laurent-Puig, Y.A. Vano, R. Elaïdi, A. Méjean, R. Sanchez-Salas, E. Barret, X. Cathelineau, S. Oudard, C.A. Reynaud, A. de Reyniès, C. Sautès-Fridman, W.H. Fridman, Tertiary lymphoid structures generate and propagate anti-tumor antibody-producing plasma cells in renal cell cancer. Immunity **55**(3), 527–541.e5 (2022)

[26] M.Y. Lu, B. Chen, D.F.K. Williamson, R.J. Chen, M. Zhao, A.K. Chow, K. Ikemura, A. Kim, D. Pouli, A. Patel, A. Soliman, C. Chen, T. Ding, J.J. Wang, G. Gerber, I. Liang, L.P. Le, A.V. Parwani, L.L. Weishaupt, F. Mahmood, A multimodal generative AI copilot for human pathology. Nature **634**(8033), 466–473 (2024)

[27] Z. Wang, Q. Jin, C.H. Wei, S. Tian, P.T. Lai, Q. Zhu, C.P. Day, C. Ross, R. Leaman, Z. Lu, GeneAgent: self-verification language agent for gene-set analysis using domain databases. Nature Methods **22**(8), 1677–1685 (2025)

[28] S. Alber, B. Chen, E. Sun, A. Isakova, A.J. Wilk, J. Zou, CellVoyager: AI compbio agent generates new insights by autonomously analyzing biological data. Nature Methods **23**(4), 749–759 (2026)

[29] M.J. Stec, Q. Su, C. Adler, L. Zhang, D.R. Golann, N.P. Khan, L. Panagis, S.A. Villalta, M. Ni, Y. Wei, J.R. Walls, A.J. Murphy, G.D. Yancopoulos, G.S. Atwal, S. Kleiner, G. Halasz, M.W. Sleeman, A cellular and molecular spatial atlas of dystrophic muscle. Proceedings of the National Academy of Sciences **120**(29), e2221249120 (2023)

[30] A. Erickson, M. He, E. Berglund, M. Marklund, R. Mirzazadeh, N. Schultz, L. Kvastad, A. Andersson, L. Bergenstråhle, J. Bergenstråhle, L. Larsson, L. Alonso Galicia, A. Shamikh, E. Basmaci, T. Díaz De Ståhl, T. Rajakumar, D. Doultsinos, K. Thrane, A.L. Ji, P.A. Khavari, F. Tarish, A. Tanoglidi, J. Maaskola, R. Colling, T. Mirtti, F.C. Hamdy, D.J. Woodcock, T. Helleday, I.G. Mills, A.D. Lamb, J. Lundeberg, Spatially resolved clonal copy number alterations in benign and malignant tissue. Nature **608**(7922), 360–367 (2022)

[31] M. Dhainaut, S.A. Rose, G. Akturk, A. Wroblewska, S.R. Nielsen, E.S. Park, M. Buckup, V. Roudko, L. Pia, R. Sweeney, J. Le Berichel, C.M. Wilk, A. Bektesevic, B.H. Lee, N. Bhardwaj, A.H. Rahman, A. Baccarini, S. Gnjatic, D. Pe'er, M. Merad, B.D. Brown, Spatial CRISPR genomics identifies regulators of the tumor microenvironment. Cell **185**(7), 1223–1239.e20 (2022)

[32] K.R. Maynard, L. Collado-Torres, L.M. Weber, C. Uytingco, B.K. Barry, S.R. Williams, J.L. Catallini, M.N. Tran, Z. Besich, M. Tippani, J. Chew, Y. Yin, J.E. Kleinman, T.M. Hyde, N. Rao, S.C. Hicks, K. Martinowich, A.E. Jaffe, Transcriptome-scale spatial gene expression in the human dorsolateral prefrontal cortex. Nature Neuroscience **24**(3), 425–436 (2021)

[33] A. Coutant, V. Cockenpot, L. Muller, C. Degletagne, R. Pommier, L. Tonon, M. Ardin, M.C. Michallet, C. Caux, M. Laurent, A.P. Morel, P. Saintigny, A. Puisieux, M. Ouzounova, P. Martinez, Spatial transcriptomics reveal pitfalls and opportunities for the detection of rare high-plasticity breast cancer subtypes. Laboratory Investigation **103**(12), 100258 (2023)

[34] 10x Genomics. Mouse brain serial section 1 (sagittal-anterior). 10x Genomics Visium Spatial Gene Expression dataset (2020). Accessed 6 June 2026

[35] J. Hu, X. Li, K. Coleman, A. Schroeder, N. Ma, D.J. Irwin, E.B. Lee, R.T. Shinohara, M. 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)

[36] T. Biancalani, G. Scalia, L. Buffoni, R. Avasthi, Z. Lu, A. Sanger, N. Tokcan, C.R. Vanderburg, Å. Segerstolpe, M. Zhang, I. Avraham-Davidi, S. Vickovic, M. Nitzan, S. Ma, A. Subramanian, M. Lipinski, J. Buenrostro, N.B. Brown, D. Fanelli, X. Zhuang, E.Z. Macosko, A. Regev, Deep learning and alignment of spatially resolved single-cell transcriptomes with Tangram. Nature Methods **18**(11), 1352–1362 (2021)

[37] B. Jassal, L. Matthews, G. Viteri, C. Gong, P. Lorente, A. Fabregat, K. Sidiropoulos, J. Cook, M. Gillespie, R. Haw, F. Loney, B. May, M. Milacic, K. Rothfels, C. Sevilla, V. Shamovsky, S. Shorser, T. Varusai, J. Weiser, G. Wu, L. Stein, H. Hermjakob, P. D'Eustachio, The reactome pathway knowledgebase. Nucleic Acids Research **48**(D1), D498–D503 (2020)

[38] S.A. Aleksander, J. Balhoff, S. Carbon, J.M. Cherry, H.J. Drabkin, D. Ebert, M. Feuermann, P. Gaudet, N.L. Harris, D.P. Hill, et al., The Gene Ontology knowledgebase in 2023. Genetics **224**(1), iyad031 (2023)

[39] M. Yuan, K. Jin, H. Yan, A. Schroeder, C. Luo, S. Yao, B. Dumoulin, J. Levinsohn, T. Luo, J.R. Clemenceau, I. Jang, M. Kim, Y. Liu, M. Deng, E.E. Furth, P. Wilson, A. Nayak, I. Lubo, L.M. Solis Soto, L. Wang, J.H. Park, K. Susztak, T.H. Hwang, M. Li, Smart spatial omics (S2-omics) optimizes region of interest selection to capture molecular heterogeneity in diverse tissues. Nature Cell Biology **27**(12), 2225–2238 (2025)

[40] Y. Liu, S.Y. Ye, S. He, D.M. Chi, X.Z. Wang, Y.F. Wen, D. Ma, R.C. Nie, P. Xiang, Y. Zhou, Z.H. Ruan, R.J. Peng, C.L. Luo, P.P. Wei, G.W. Lin, J. Zheng, Q. Cui, M.Y. Cai, J.P. Yun, J. Dong, H.Q. Mai, X. Xia, J.X. Bei, Single-cell and spatial transcriptome analyses reveal tertiary lymphoid structures linked to tumour progression and immunotherapy response in nasopharyngeal carcinoma. Nature Communications **15**(1), 7713 (2024)

[41] C.V. Theodoris, L. Xiao, A. Chopra, M.D. Chaffin, Z.R. Al Sayed, M.C. Hill, H. Mantineo, E.M. Brydon, Z. Zeng, X.S. Liu, P.T. Ellinor, Transfer learning enables predictions in network biology. Nature **618**(7965), 616–624 (2023)

[42] S.H. Lee, D. Lee, J. Choi, H.J. Oh, I.H. Ham, D. Ryu, S.Y. Lee, D.J. Han, S. Kim, Y. Moon, I.H. Song, K.Y. Song, H. Lee, S. Lee, H. Hur, T.M. Kim, Spatial dissection of tumour microenvironments in gastric cancers reveals the immunosuppressive crosstalk between CCL2+ fibroblasts and STAT3-activated macrophages. Gut **74**(5), 714–727 (2025)

[43] M.Y. Kasmani, P. Topchyan, A.K. Brown, R.J. Brown, X. Wu, Y. Chen, A. Khatun, D. Alson, Y. Wu, R. Burns, C.W. Lin, M.R. Kudek, J. Sun, W. Cui, A spatial sequencing atlas of age-induced changes in the lung during influenza infection. Nature Communications **14**(1), 6597 (2023)

[44] M.R.M. Du, C. Wang, C.W. Law, D. Amann-Zalcenstein, C.J.A. Anttila, L. Ling, P.F. Hickey, C.J. Sargeant, Y. Chen, L.J. Ioannidis, P. Rajasekhar, R.K.H. Yip, K.L. Rogers, D.S. Hansen, R. Bowden, M.E. Ritchie, Benchmarking spatial transcriptomics technologies with the multi-sample SpatialBenchVisium dataset. Genome Biology **26**(1), 77 (2025)

[45] M. Ilse, J. Tomczak, M. Welling, *Attention-based deep multiple instance learning*, in *Proceedings of the 35th International Conference on Machine Learning*, *Proceedings of Machine Learning Research*, vol. 80 (2018), pp. 2127–2136

[46] M.Y. Lu, D.F.K. Williamson, T.Y. Chen, R.J. Chen, M. Barbieri, F. Mahmood, Data-efficient and weakly supervised computational pathology on whole-slide images. Nature Biomedical Engineering **5**(6), 555–570 (2021)

[47] H. Xu, N. Usuyama, J. Bagga, S. Zhang, R. Rao, T. Naumann, C. Wong, Z. Gero, J. González, Y. Gu, Y. Xu, M. Wei, W. Wang, S. Ma, F. Wei, J. Yang, C. Li, J. Gao, J. Rosemon, T. Bower, S. Lee, R. Weerasinghe, B.J. Wright, A. Robicsek, B. Piening, C. Bifulco, S. Wang, H. Poon, A whole-slide foundation model for digital pathology from real-world data. Nature **630**(8015), 181–188 (2024)

[48] R.J. Chen, T. Ding, M.Y. Lu, D.F.K. Williamson, G. Jaume, A.H. Song, B. Chen, A. Zhang, D. Shao, M. Shaban, M. Williams, L. Oldenburg, L.L. Weishaupt, J.J. Wang, A. Vaidya, L.P. Le, G. Gerber, S. Sahai, W. Williams, F. Mahmood, Towards

a general-purpose foundation model for computational pathology. Nature Medicine **30**(3), 850–862 (2024)

[49] J. Xiang, X. Wang, X. Zhang, Y. Xi, F. Eweje, Y. Chen, Y. Li, C. Bergstrom, M. Gopaulchan, T. Kim, K.H. Yu, S. Willens, F.M. Olguin, J.J. Nirschl, J. Neal, M. Diehn, S. Yang, R. Li, A vision–language foundation model for precision oncology. Nature **638**(8051), 769–778 (2025)

[50] X. Wang, J. Zhao, E. Marostica, W. Yuan, J. Jin, J. Zhang, R. Li, H. Tang, K. Wang, Y. Li, F. Wang, Y. Peng, J. Zhu, J. Zhang, C.R. Jackson, J. Zhang, D. Dillon, N.U. Lin, L. Sholl, T. Denize, D. Meredith, K.L. Ligon, S. Signoretti, S. Ogino, J.A. Golden, M.P. Nasrallah, X. Han, S. Yang, K.H. Yu, A pathology foundation model for cancer diagnosis and prognosis prediction. Nature **634**(8035), 970–978 (2024)

[51] M.Y. Lu, B. Chen, D.F.K. Williamson, R.J. Chen, I. Liang, T. Ding, G. Jaume, I. Odintsov, L.P. Le, G. Gerber, A.V. Parwani, A. Zhang, F. Mahmood, A visual-language foundation model for computational pathology. Nature Medicine **30**(3), 863–874 (2024)

[52] The Cancer Genome Atlas Research Network, Comprehensive molecular characterization of gastric adenocarcinoma. Nature **513**(7517), 202–209 (2014)

[53] The Cancer Genome Atlas Network, Comprehensive molecular portraits of human breast tumours. Nature **490**(7418), 61–70 (2012)

[54] National Cancer Institute Clinical Proteomic Tumor Analysis Consortium (CPTAC). The clinical proteomic tumor analysis consortium head and neck squamous cell carcinoma collection (CPTAC-HNSCC). The Cancer Imaging Archive (2018)

[55] Z. Gao, A. Mao, Y. Dong, H. Clayton, J. Wu, J. Liu, C. Wang, K. He, T. Gong, C. Li, M. Crispin-Ortuzar, SMMILe enables accurate spatial quantification in digital pathology using multiple-instance learning. Nature Cancer **6**(12), 2025–2041 (2025)

[56] M. Amgad, H. Elfandy, H. Hussein, et al., Structured crowdsourcing enables convolutional segmentation of histology images. Bioinformatics **35**(18), 3461–3467 (2019)

[57] A. Ally, M. Balasundaram, R. Carlsen, E. Chuah, A. Clarke, N. Dhalla, R.A. Holt, S.J.M. Jones, D. Lee, Y. Ma, et al., Comprehensive and integrative genomic characterization of hepatocellular carcinoma. Cell **169**(7), 1327–1341.e23 (2017)

[58] D.J. Brat, R.G.W. Verhaak, K.D. Aldape, W.K.A. Yung, S.R. Salama, L.A.D. Cooper, E. Rheinbay, C.R. Miller, M. Vitucci, O. Morozova, et al., Comprehensive, integrative genomic analysis of diffuse lower-grade gliomas. New England Journal of Medicine **372**(26), 2481–2498 (2015)

[59] Y. Zhang, M. Li, D. Long, X. Zhang, H. Lin, B. Yang, P. Xie, A. Yang, D. Liu, J. Lin, F. Huang, J. Zhou, Qwen3 embedding: advancing text embedding and reranking through foundation models. arXiv preprint arXiv:2506.05176 (2025)

[60] L. van der Maaten, G. Hinton, Visualizing data using t-SNE. Journal of Machine Learning Research **9**, 2579–2605 (2008)

[61] M. Litviňuková, C. Talavera-López, H. Maatz, et al., Cells of the adult human heart. Nature **588**(7838), 466–472 (2020)

[62] S.Z. Wu, G. Al-Eryani, D.L. Roden, et al., A single-cell and spatially resolved atlas of human breast cancers. Nature Genetics **53**(9), 1334–1347 (2021)

[63] M. Schaefer, P. Peneder, D. Malzl, S.D. Lombardo, M. Peycheva, J. Burton, A. Hakobyan, V. Sharma, T. Krausgruber, C. Sin, J. Menche, E.M. Tomazou, C. Bock, Multimodal learning enables chat-based exploration of single-cell data. Nature Biotechnology (2025)

[64] T. Ding, S.J. Wagner, A.H. Song, R.J. Chen, M.Y. Lu, A. Zhang, A.J. Vaidya, G. Jaume, M. Shaban, A. Kim, D.F.K. Williamson, H. Robertson, B. Chen, C. Almagro-Pérez, P. Doucet, S. Sahai, C. Chen, C.S. Chen, D. Komura, A. Kawabe, M. Ochi, S. Sato, T. Yokose, Y. Miyagi, S. Ishikawa, G. Gerber, T. Peng, L.P. Le, F. Mahmood, A multimodal whole-slide foundation model for pathology. Nature Medicine **31**(11), 3749–3761 (2025)

[65] B.W. Hounkpe, F. Chenou, F. de Lima, E.V. De Paula, Hrt atlas v1.0 database: redefining human and mouse housekeeping genes and candidate reference transcripts by mining massive rna-seq datasets. Nucleic Acids Research **49**(D1), D947–D955 (2021)

[66] A. Jaegle, F. Gimeno, A. Brock, O. Vinyals, A. Zisserman, J. Carreira, *Perceiver: General Perception with Iterative Attention*, in *Proceedings of the 38th International Conference on Machine Learning*, vol. 139 (2021), pp. 4651–4664

[67] E.J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, *LoRA: Low-Rank Adaptation of Large Language Models*, in *International Conference on Learning Representations* (2022)

[68] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, I. Polosukhin, *Attention Is All You Need*, in *Advances in Neural Information Processing Systems*, vol. 30 (2017)

[69] I. Loshchilov, F. Hutter, *Decoupled Weight Decay Regularization*, in *International Conference on Learning Representations* (2019)

[70] C. Hu, T. Li, Y. Xu, X. Zhang, F. Li, J. Bai, J. Chen, W. Jiang, K. Yang, Q. Ou, X. Li, P. Wang, Y. Zhang, CellMarker 2.0: An updated database of manually curated cell markers in human/mouse and web tools based on scrna-seq data. Nucleic Acids Research **51**(D1), D870–D876 (2023)

[71] A. Yang, A. Li, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Gao, C. Huang, C. Lv, et al., Qwen3 technical report. arXiv preprint arXiv:2505.09388 (2025)

[72] D.R. Cox, Regression models and life-tables. Journal of the Royal Statistical Society: Series B (Methodological) **34**(2), 187–220 (1972)

[73] D.P. Kingma, J. Ba, *Adam: A Method for Stochastic Optimization*, in *International Conference on Learning Representations* (2015)

[74] B. Efron, Bootstrap methods: Another look at the jackknife. The Annals of Statistics **7**(1), 1–26 (1979)

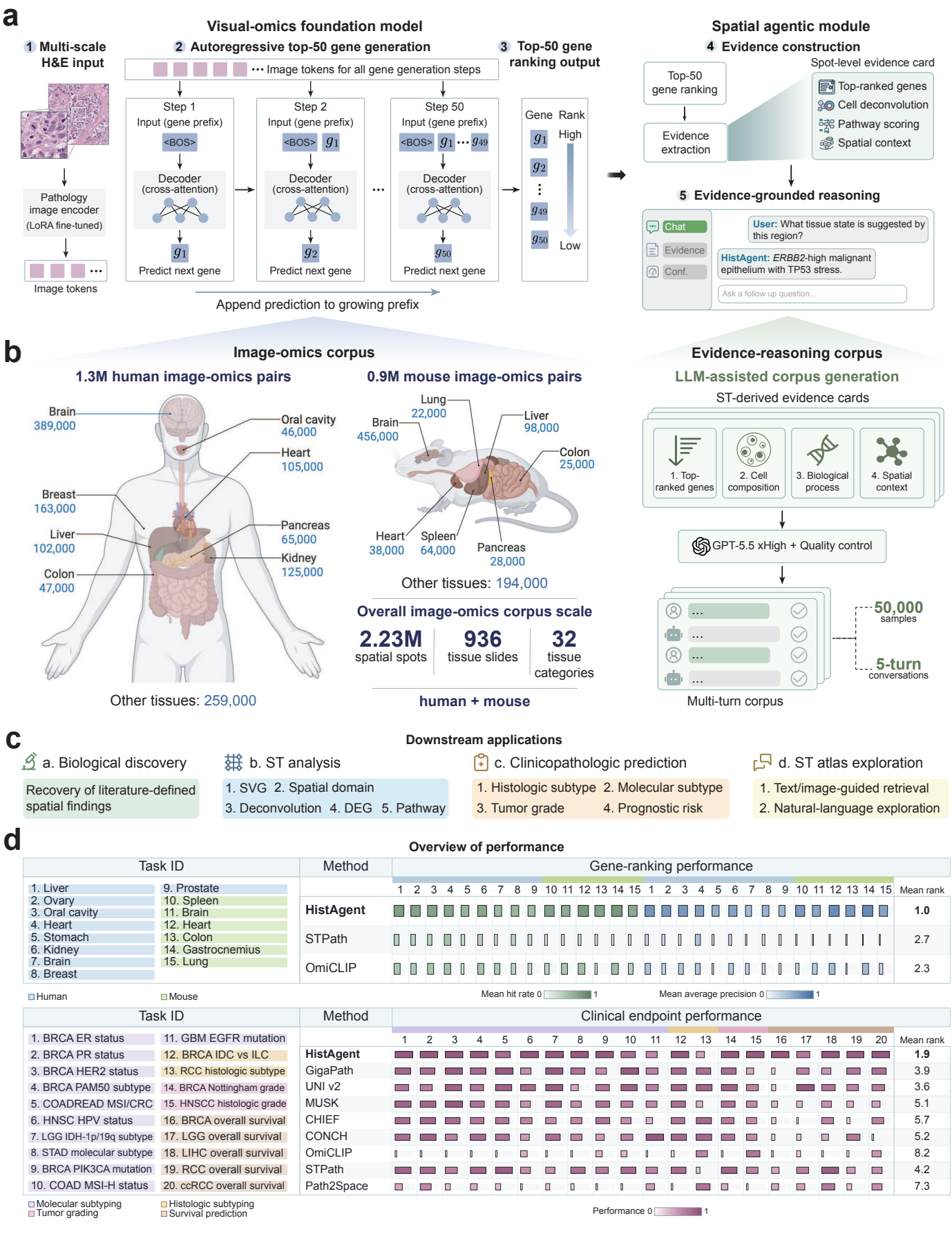
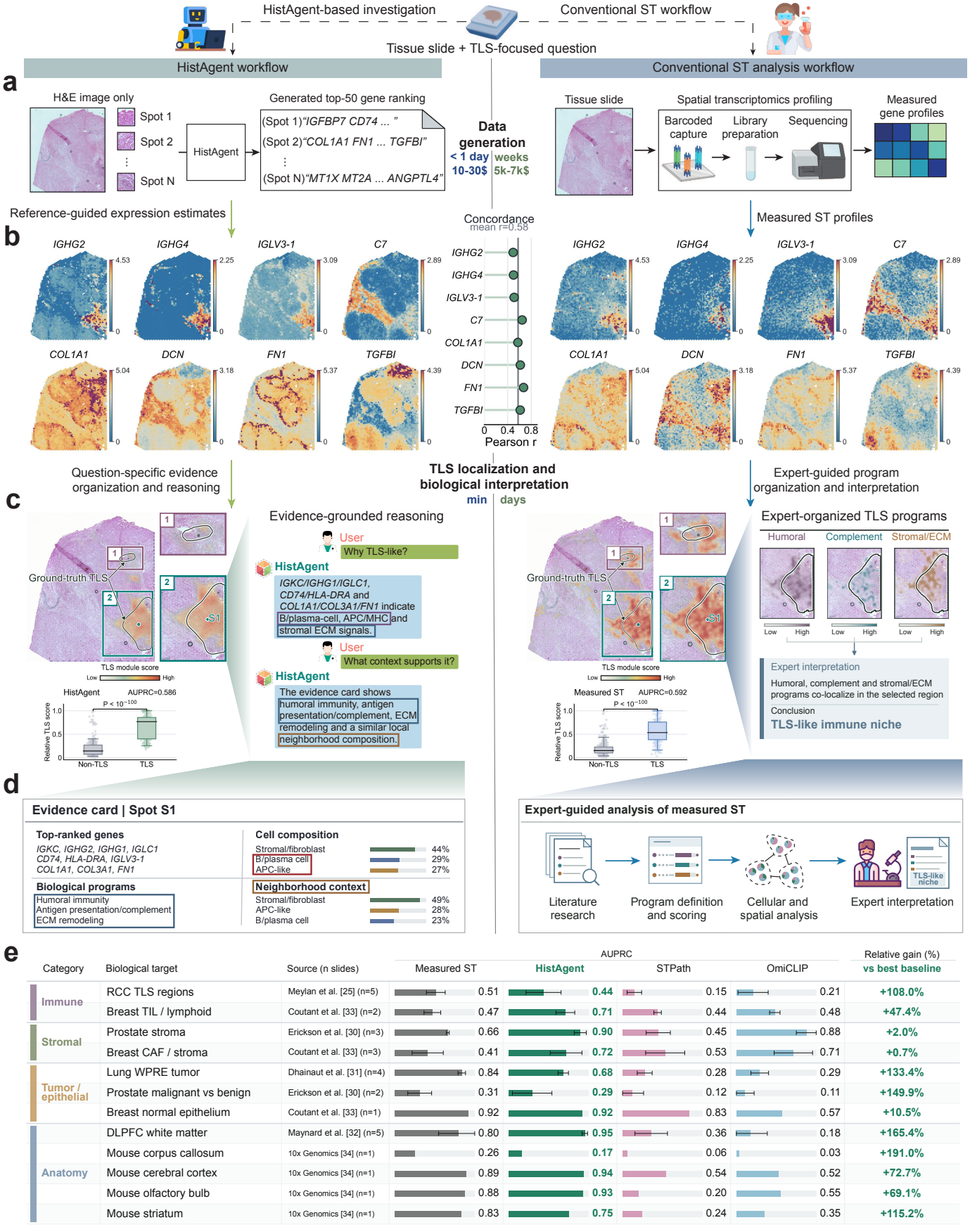


Figure 1— HistAgent integrates image-conditioned molecular inference with evidence-grounded spatial analysis. a, Overview of HistAgent. First, a pathology foundation model adapted using low-rank adaptation (LoRA) encodes local and contextual hematoxylin and eosin (H&E) views. The visual-omics foundation model then autoregressively generates a top-50 gene ranking for each tissue location. The spatial agentic module organizes the ranked genes, inferred cell composition, functional programs and spatial context into a spot-level evidence card, which an adapted large language model (LLM) uses to answer biological

questions across multiple conversational turns. **b**, Construction of the training corpora. The foundation model was trained on 2.23 million paired H&E–spatial transcriptomics (ST) spots from 936 slides across 32 human and mouse tissue categories. The LLM was fine-tuned on 50,000 five-turn evidence–reasoning traces generated by a teacher LLM, followed by quality control. Human and mouse anatomical illustrations were created with BioRender.com. **c**, HistAgent applications spanning biological discovery, downstream ST analyses, whole-slide clinicopathologic prediction, and ST atlas retrieval and exploration. **d**, Benchmark overview for spot-level gene-ranking recovery and slide-level clinical prediction. Gene-ranking performance was measured via the hit rate at 50 and mean average precision at 50 (mAP@50). Clinical performance was measured via the area under the receiver operating characteristic curve (AUROC) for binary endpoints, macro-averaged AUROC for multiclass endpoints, and the concordance index for survival endpoints.



Category	Biological target	Source (n slides)	AUPRC Measured ST	AUPRC HistAgent	AUPRC STPath	AUPRC OmiCLIP	Relative gain (%) vs best baseline
Immune	RCC TLS regions	Meylan et al. [25] (n=5)	0.51	0.44	0.15	0.21	+108.0%
	Breast TIL / lymphoid	Coutant et al. [33] (n=2)	0.47	0.71	0.44	0.48	+47.4%
Stromal	Prostate stroma	Erickson et al. [30] (n=3)	0.66	0.90	0.45	0.88	+2.0%
	Breast CAF / stroma	Coutant et al. [33] (n=3)	0.41	0.72	0.53	0.71	+0.7%
Tumor / epithelial	Lung WPRE tumor	Dhainaut et al. [31] (n=4)	0.84	0.68	0.28	0.29	+133.4%
	Prostate malignant vs benign	Erickson et al. [30] (n=2)	0.31	0.29	0.12	0.11	+149.9%
	Breast normal epithelium	Coutant et al. [33] (n=1)	0.92	0.92	0.83	0.57	+10.5%
Anatomy	DLPFC white matter	Maynard et al. [32] (n=5)	0.80	0.95	0.36	0.18	+165.4%
	Mouse corpus callosum	10x Genomics [34] (n=1)	0.26	0.17	0.06	0.03	+191.0%
	Mouse cerebral cortex	10x Genomics [34] (n=1)	0.89	0.94	0.54	0.52	+72.7%
	Mouse olfactory bulb	10x Genomics [34] (n=1)	0.88	0.93	0.20	0.55	+69.1%
	Mouse striatum	10x Genomics [34] (n=1)	0.83	0.75	0.24	0.35	+115.2%

Figure 2— HistAgent recovers and interprets spatial biological findings from routine histology. **a**, HistAgent and conventional spatial transcriptomics (ST) workflows used for investigating tertiary lymphoid structures (TLS) in a renal cell carcinoma slide. HistAgent generates ranked molecular readouts directly from hematoxylin and eosin (H&E) images. The displayed turnaround and cost estimates are less than one day and US$10–30 per slide for HistAgent, including slide scanning, compared with several weeks and US$5,000–7,000 for conventional ST profiling [39]. **b**, Reference-guided expression estimates derived from HistAgent's ranked molecular readouts and the corresponding measured ST expression patterns for eight

representative genes associated with the TLS immune and stromal microenvironment across 3,206 spots. Pearson correlation coefficients (PCCs) were calculated across spots for each gene, and the displayed mean summarizes the eight gene-specific PCCs. **c**, TLS localization using a literature-supported 25-gene signature with reference-guided HistAgent expression estimates and measured ST profiles. Performance was evaluated against an expert annotation comprising 255 TLS and 2,951 non-TLS spots using the area under the precision–recall curve (AUPRC). The accompanying dialogue illustrates a multi-turn, evidence-grounded interpretation of the recovered TLS-like region. TLS signature-score maps were scaled independently for visualization. Boxplots show signature-score distributions. P values were calculated using two-sided Mann–Whitney U tests. **d**, Evidence supporting the interpretations shown in **c**. HistAgent organizes top-ranked genes, inferred cell composition, functional programs and spatial context into an evidence card, whereas the conventional workflow requires separate analyses followed by expert synthesis. The displayed cell-composition values are normalized marker-based cell-state scores and should not be interpreted as absolute cell fractions. **e**, Cross-study benchmark of 12 literature-defined biological findings evaluated across 29 slide–finding pairs from 22 unique slides, spanning five published studies and one public 10x Genomics resource. The findings represent immune, stromal, tumor and epithelial, and anatomical states. The same signature-scoring procedure was applied to reference-guided HistAgent expression estimates, STPath and OmiCLIP predictions, and measured ST profiles, and AUPRC quantified agreement with the corresponding expert annotation. Bars show means across slides, error bars denote standard deviations, and the number of slides (n) contributing to each finding are in parentheses. Relative gain denotes the percentage improvement over the stronger H&E-based baseline, STPath or OmiCLIP, for each finding.

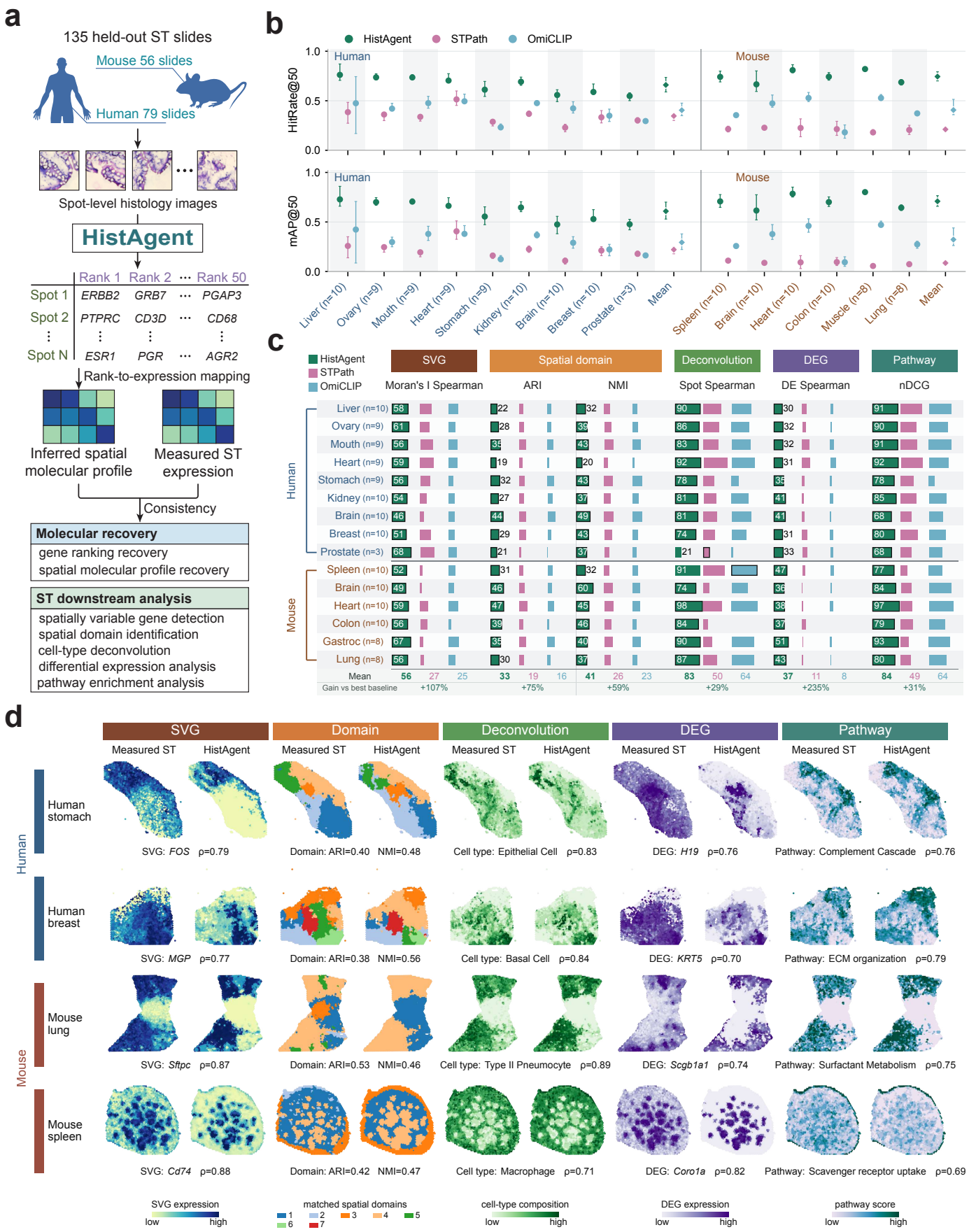


Figure 3— HistAgent recovers spatial molecular readouts from histology for standard spatial transcriptomic analyses. **a**, Benchmark of 135 held-out spatial transcriptomics (ST) slides, comprising 79 human and 56 mouse slides. Readouts predicted from matched hematoxylin and eosin (H&E) images were compared with paired measured ST at the spot and gene levels and through five downstream analyses. Reference-guided inference provided continuous expression estimates when required. **b**, Spot-level recovery by HistAgent, STPath, and OmiCLIP across 15 organ cohorts. HitRate@50 was applied to assess coverage of the measured top-50 expressed genes, and mean average precision at 50 (mAP@50) was used to consider their

positions in the predicted ranking. Dots and vertical bars show organ-level means and interquartile ranges across slides; diamonds show species-level means. Cohort sizes are given below the axis. **c**, Agreement for spatially variable gene (SVG) detection, spatial-domain identification, cell-type deconvolution, differentially expressed gene (DEG) analysis, and pathway enrichment. SVG agreement used the Spearman correlation of gene-wise Moran's I; domain agreement used the adjusted Rand index (ARI) and normalized mutual information (NMI); deconvolution and DEG agreement used Spearman correlation; and pathway agreement used normalized discounted cumulative gain (nDCG). Higher values indicate closer agreement with measured ST. Outlines mark the highest organ-level mean, numbers inside HistAgent bars are scores multiplied by 100, and the bottom rows report overall means and relative gains over the strongest baseline. **d**, Paired measured ST and HistAgent visualizations for four slides with median overall performance. Genes, cell types, and pathways were selected using task-specific measurements of spatial variation in measured ST data. Values below the maps report spatial Spearman correlations (ρ) between measured ST and HistAgent values across aligned spots, or ARI and NMI for domains. Domain colors are matched within each pair, and the continuous color scales are feature-specific.

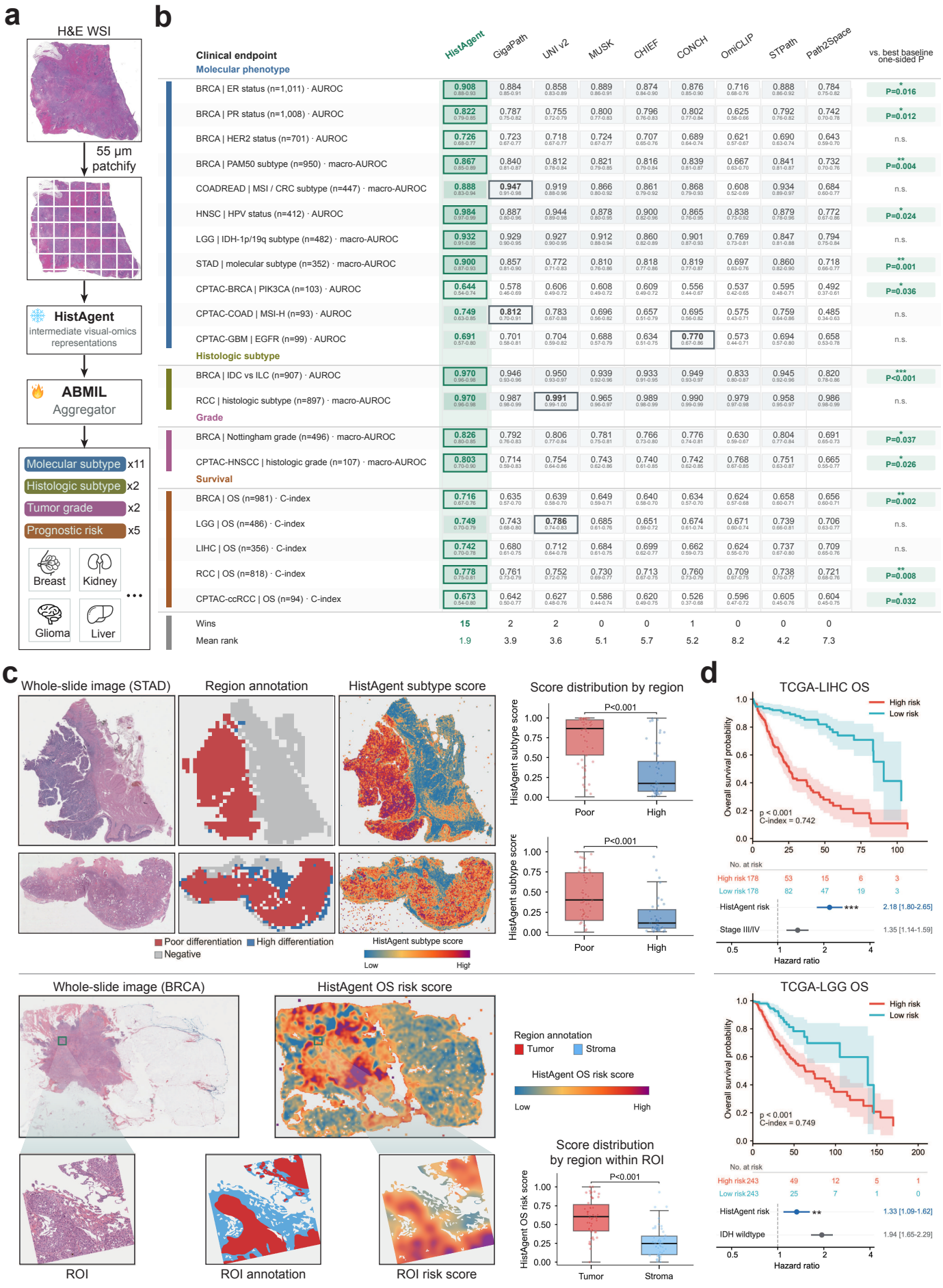


Figure 4— HistAgent supports clinical prediction across cohorts and organs. a, Clinical benchmark design. Hematoxylin and eosin (H&E) whole-slide images (WSI) were divided into 55-µm tiles and encoded by HistAgent. Tile representations were aggregated by an attention-based multiple-instance learning (ABMIL) model to predict molecular subtype, histologic subtype, tumor grade, and prognostic risk.

b, Performance across 20 clinical endpoints. HistAgent was compared with eight baselines using the same patient-level splits, ABMIL model, and outer five-fold evaluation. The n shown for each endpoint denotes the number of patients included in the evaluation. Binary classification was evaluated via the area under the receiver operating characteristic curve (AUROC), and multiclass classification by macro-averaged AUROC (macro-AUROC) and prognostic risk was the concordance index (C-index). The C-index measures agreement between predicted risk ordering and observed survival outcomes, with higher values indicating better prognostic discrimination. Point estimates are shown in large type, with 95% patient-level bootstrap confidence intervals shown beneath in smaller type; outlines mark the highest point estimate for each endpoint. The rightmost column reports one-sided paired bootstrap P values for comparisons against the strongest baseline. Higher estimated win count and lower mean rank indicates stronger overall performance. **c**, Spatial visualization of representative slide-level predictions. Out-of-fold ABMIL models were applied to individual tiles, and the resulting scores were mapped back to their tissue coordinates. Genomically stable subtype scores are shown for stomach adenocarcinoma (STAD) slides with poor- and high-differentiation annotations, and overall survival (OS) risk scores are shown for a breast invasive carcinoma (BRCA) slide with tumor and stromal annotations. Boxplots show medians and interquartile ranges; whiskers extend to 1.5 times the interquartile range, and points denote up to 48 sampled tiles. P values were calculated using one-sided Mann–Whitney U tests in the prespecified direction. ROI, region of interest. **d**, Prognostic-risk analysis for liver hepatocellular carcinoma (LIHC) and lower-grade glioma (LGG) using patient-level out-of-fold risk scores. Kaplan–Meier curves compare groups separated at the cohort median; shaded bands denote 95% confidence intervals, tables give the numbers at risk, and P values are from two-sided log-rank tests. Displayed C-indices were calculated from the continuous risk scores rather than the median split. Multivariable Cox models were used to evaluate the standardized HistAgent risk score after adjustment for American Joint Committee on Cancer (AJCC) stage in LIHC or isocitrate dehydrogenase (IDH) status in LGG. Points and horizontal bars denote hazard ratios and 95% confidence intervals, respectively, and significance was assessed using two-sided Wald tests. For **b,d**, $*P < 0.05$, $**P < 0.01$ and $***P < 0.001$; n.s., not significant.

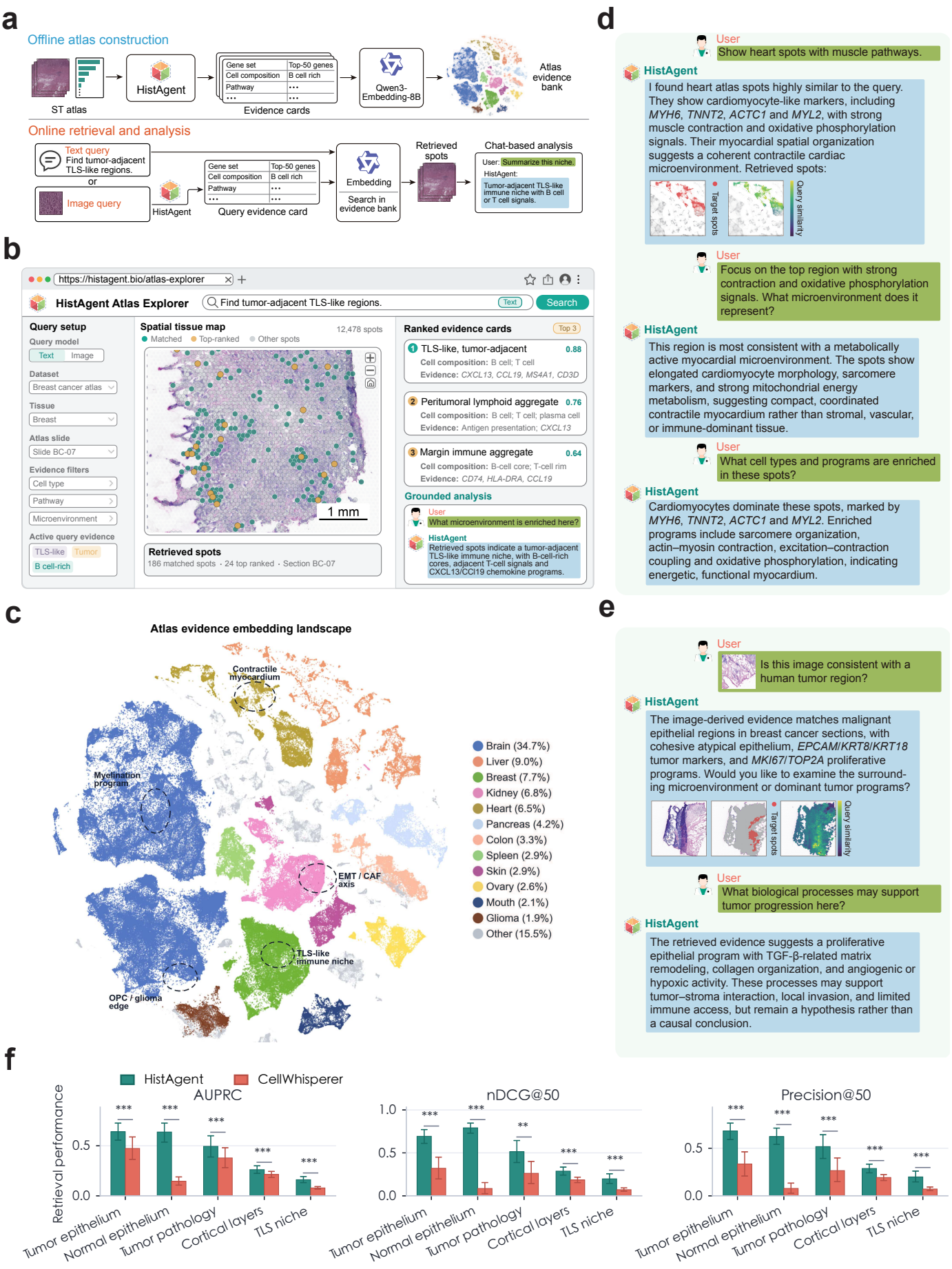


Figure 5— HistAgent supports spatial transcriptomics atlas retrieval using natural-language and H&E image queries. **a**, Framework for spatial transcriptomics (ST) atlas retrieval. During offline atlas construction, evidence cards derived from 2.23 million measured ST spots were embedded using Qwen3-Embedding-8B to form the measured ST evidence bank [59]. During online query processing, natural-language queries were embedded directly, whereas hematoxylin and eosin (H&E) image queries were processed

by HistAgent to generate predicted evidence cards. Both query types were matched to measured atlas spots by cosine similarity. **b**, HistAgent Atlas Explorer. A natural-language query for tumor-adjacent tertiary lymphoid structure (TLS)-like regions retrieves spatial locations and ranked evidence cards from breast cancer atlas slide BC-07 for follow-up analysis. Numbers on the evidence cards denote cosine similarity. **c**, Two-dimensional t-distributed stochastic neighbor embedding (t-SNE) visualization of the evidence-card embeddings [60]. Colors denote tissue context, and dashed contours mark selected biological neighborhoods. The projection was used only for visualization, whereas the full 4,096-dimensional embeddings were used for retrieval. **d**, Natural-language query example. A query for cardiac regions with active muscle programs retrieves cardiomyocyte-like spots and supports follow-up analysis of their molecular programs and spatial context. **e**, H&E image-query example. HistAgent converts a breast tumor region into a predicted evidence card and retrieves measured ST spots with related epithelial tumor states. **f**, Natural-language retrieval benchmark comparing HistAgent with CellWhisperer [63] across five biological query groups and 40 ST slides. Both methods used identical query strings and candidate spots, with expert annotations defining the relevant spots. Performance was evaluated using the area under the precision–recall curve (AUPRC), normalized discounted cumulative gain at 50 (nDCG@50), and Precision@50. Bars show means across matched query–slide units, and error bars denote 95% confidence intervals. P values were calculated using one-sided paired tests. $*P < 0.05$, $**P < 0.01$ and $***P < 0.001$.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- HistAgentsupplementarytables.xlsx
- HistAgentsupplementaryinformation.pdf