

Supplementary Information for:

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

Wenlin Li^{1,2,†}, Hengrui Liu^{1,†}, Yan Lu^{1,†}, Xu Liao¹, Xiaobo Sun³, Weiwei
Zhai⁴, Chengqi Lin⁵, and Jin Liu^{1,2,*}

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

²Shenzhen Loop Area Institute, Shenzhen, China

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

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

⁵Key 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

16 Contents

17	S1	Supplementary Methods	3
18	S1.1	HistAgent model architecture	3
19	S1.2	HistAgent training and optimization	3
20	S1.3	Evidence-grounding and hallucination evaluation	4
21	S1.4	Standard ST analyses	4
22	S1.5	Local interpretation of clinical predictions	6
23	S1.6	Survival stratification and multivariable Cox regression	6
24	S1.7	Comparisons of methods	6
25	S1.7.1	Compared methods for spatial molecular recovery	6
26	S1.7.2	Compared methods for clinical prediction	7
27	S1.7.3	Compared method for natural-language atlas retrieval	8
28	S1.8	Evaluation metrics	8
29	S1.8.1	Ranked molecular-recovery metrics	8
30	S1.8.2	Spatial-profile agreement	9
31	S1.8.3	Metrics for standard ST analyses	9
32	S1.8.4	Clinical prediction metrics	11
33	S1.8.5	Atlas-retrieval metrics	12
34	S1.9	Datasets and preprocessing for downstream evaluation	12
35	S1.9.1	Molecular-recovery evaluation data	12
36	S1.9.2	Clinical cohorts and whole-slide preprocessing	12
37	S1.9.3	Atlas and natural-language retrieval benchmark	13
38	S2	Supplementary Figures	14
39	S1	Spatial expression concordance across increasing HVG set sizes	14
40	S2	Slide-level distributions of spot-level gene-ranking recovery	15
41	S3	Gene-level spatial recovery across increasing highly variable gene sets	17
42	S4	Slide-level distributions of downstream spatial transcriptomic agreement	19
43	S5	HistAgent spatial readouts for organ cohorts not shown in the main figure	22
44	S6	Baseline spatial readouts for all organ cohorts in the Figure 3 benchmark	28
45	S7	HistAgent GS subtype scores across annotated STAD differentiation regions	37
46	S8	HistAgent overall survival risk scores across annotated BRCA tissue regions	38
47	S9	Biological query groups used for natural-language retrieval evaluation	39
48	S10	Slide-level natural-language retrieval performance	40
49	S11	Natural-language retrieval performance across evaluation cutoffs	41
50	S12	Evidence-grounding audit of the spatial agentic module	42

Supplementary Methods

S1.1 HistAgent model architecture

HistAgent contains approximately 1.23 billion parameters, of which approximately 94.59 million (7.69%) are trainable. Each spot is represented by a local H&E view centered on the nominal Visium capture region and a contextual H&E view covering four times the linear field of view. Both views were converted to $3 \times 224 \times 224$ tensors before encoding.

A shared GigaPath tile encoder mapped each image view to a 1,536-dimensional feature vector. Parameters of the pretrained backbone remained frozen, except for inserted low-rank adaptation (LoRA) modules. Separate Perceiver-style resamplers transformed the local and contextual visual vectors. Each resampler used eight attention heads and returned 16 latent tokens with a width of 512. The 16 local tokens and 16 contextual tokens were concatenated with one learned species token and one learned organ token, producing a 34×512 contextual memory for each spot.

The autoregressive Transformer decoder comprised six layers, eight attention heads with a hidden width of 512, with dropout of 0.1. Causal self-attention modeled the preceding gene tokens, whereas cross-attention conditioned each prediction on the contextual memory. The decoder generated a ranked molecular readout containing at most 50 non-repeating canonical gene tokens. Its cleaned vocabulary contained 32,019 genes plus 4 special tokens. During inference, greedy decoding masked padding, beginning-of-sequence, unknown, and previously generated gene tokens and stopped at the end-of-sequence token or after 50 genes.

S1.2 HistAgent training and optimization

LoRA modules with rank 16, scaling parameter 32, and dropout of 0.05 were applied to the query-key-value projection, attention output projection, and two feed-forward projections of GigaPath. Training used teacher forcing with a normalized rank-weighted cross-entropy and label smoothing of 0.1, leading to larger loss weights being assigned to earlier positions in the measured ranking.

We trained HistAgent for 30 epochs with AdamW, at 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 optimization steps and then followed cosine decay. The per-GPU batch size was 24. During training, Gaussian noise with a standard deviation of 0.02 was added to the visual features, and species and organ conditions were independently replaced by their unknown embeddings with a probability of 0.15. The complete training run lasted approximately 103 hours on eight NVIDIA H100 GPUs.

The release and held-out-evaluation checkpoints used the same architecture and optimization procedure. The held-out-evaluation checkpoint excluded all 135 molecular-evaluation slides from optimization, validation, and model selection.

S1.3 Evidence-grounding and hallucination evaluation

We evaluated whether responses from the spatial agentic module depended on the supplied evidence rather than on query wording alone. The evaluation contained 22 scenarios selected before response scoring from the natural-language retrieval benchmark. For each scenario, the three highest-ranked measured spatial transcriptomics (ST) evidence cards were supplied to the fine-tuned agentic module, which was instructed to support molecular and spatial statements with structured `evidence_refs`. Responses were generated by greedy decoding.

Each scenario was evaluated under three conditions. The full-evidence condition used the retrieved cards without modification. The evidence-removal condition deleted the fields that directly supported the query, including relevant annotations, genes, inferred cell composition, functional programs, and spatial context. The counterfactual condition replaced those fields with manually constructed, biologically incompatible evidence for the same query type. Claim traceability was the proportion of molecular claims linked to a cited field that supported the claim. A molecular claim was classified as a hallucination when a stated gene, cell type, or functional program could not be traced to any cited evidence-card field. We also assessed whether responses changed appropriately or expressed uncertainty after evidence removal or counterfactual replacement. The evaluation scenarios were excluded from supervised fine-tuning and model selection.

S1.4 Standard ST analyses

We evaluated spatially variable gene (SVG) detection, spatial domain identification, cell-type deconvolution, differential expression, and pathway enrichment with standard ST. For method m and spot i , the top-50 gene window was fixed before performing any task-specific calculations. Within that window, we canonicalized the gene symbols and removed genes in the fixed housekeeping list, genes matching the predefined technical prefixes, duplicate symbols, and genes absent from the common expression axis without backfilling. Let $\mathbf{Y}^{(m)}$ be the continuous expression matrix, comprising reference-guided expression estimates for HistAgent and native continuous predictions for STPath and OmiCLIP. Let $\mathbf{B}^{(m)}$ be the binary support matrix, with an entry of one when a gene was retained from the method’s original top-50 prediction for that spot after filtering and zero otherwise. The common input to all five tasks was

$$\mathbf{X}^{(m)} = \mathbf{Y}^{(m)} \odot \mathbf{B}^{(m)}, \quad (\text{S1})$$

where $\odot$ denotes element-wise multiplication. Thus, continuous expression values were retained only for genes supported by the corresponding top-50 predictions, and all other values were set to zero. Corresponding reference results were derived from measured ST using the same task definitions and evaluation procedures.

For SVG detection, we constructed a common spatial graph from the measured coordinates using six nearest neighbors, then symmetrized its adjacency matrix. Using spatially

variable gene detection, Moran's I [1] was calculated for every gene included in the analysis using this common spatial graph.

For spatial domain identification, measured ST data and each predicted readout were processed using the same joint molecular-spatial clustering procedure. For each readout r , the expression features were standardized, restricted to genes with nonzero variance, and reduced to at most 30 principal components. Let $\mathbf{W}_{\text{mol}}^{(r)}$ be the binary symmetric 12-neighbor graph constructed from these features and $\mathbf{W}_{\text{spatial}}$ the binary symmetric six-neighbor graph constructed from the measured coordinates. The joint affinity was

$$\mathbf{W}_{\text{joint}}^{(r)} = \mathbf{W}_{\text{mol}}^{(r)} + 2\mathbf{W}_{\text{spatial}}, \quad (\text{S2})$$

with diagonal entries set to one. If $\lambda_1 \leq \dots \leq \lambda_n$ are the eigenvalues of its normalized Laplacian, the number of domains was

$$K_r^* = \arg \max_{K \in \{2, \dots, \min(12, n-1)\}} (\lambda_{K+1} - \lambda_K), \quad (\text{S3})$$

with ties resolved in favor of the smaller K . Spectral clustering [2] with K_r^* clusters produced the spatial-domain partition for each measured or predicted readout.

For cell-type deconvolution, let $\mathcal{M}_c$ denote the marker set for cell type c , let C be the number of retained marker sets, and let f_g be the number of those sets containing gene g . For the amplitude-ordered genes $(g_{i,1}, \dots, g_{i,L_i})$ at spot i , the unnormalized marker score was

$$\tilde{a}_{i,c} = \frac{1}{\sqrt{|\mathcal{M}_c|}} \sum_{r: g_{i,r} \in \mathcal{M}_c} \frac{\log[1 + (C+1)/(1 + f_{g_{i,r}})]}{\log_2(r+1)}. \quad (\text{S4})$$

Scores were normalized within each spot,

$$a_{i,c} = \begin{cases} \frac{\tilde{a}_{i,c}}{\sum_{d=1}^C \tilde{a}_{i,d}}, & \sum_{d=1}^C \tilde{a}_{i,d} > 0, \\ 0, & \sum_{d=1}^C \tilde{a}_{i,d} = 0, \end{cases} \quad (\text{S5})$$

to obtain relative marker-composition profiles rather than absolute cell fractions.

For differential expression, the fixed measured ST spatial domains were used for every method. Genes were standardized across spots, and one-versus-rest domain effects were calculated for each gene. For pathway enrichment, we retained Reactome gene sets $\mathcal{P}_p$ containing 5–500 genes. Let $h_{i,p}$ be the number of retained genes at spot i that occurred in pathway p , let L_i be the number of retained genes, and let U be the size of the pathway-gene universe. The pathway score was

$$q_{i,p} = \begin{cases} h_{i,p} \log_2 \left[1 + \frac{h_{i,p}}{L_i |\mathcal{P}_p| / U} \right], & h_{i,p} \geq 2, \\ 0, & h_{i,p} < 2. \end{cases} \quad (\text{S6})$$

Pathways were ranked in descending order of $q_{i,p}$, with the predefined pathway index

150 resolving ties. The output of each task was compared with the corresponding measured
151 ST result using the metrics defined below.

152 **S1.5 Local interpretation of clinical predictions**

153 Local scores were generated for held-out slides using the ABMIL model, trained without the
154 corresponding patient’s outer fold. The trained instance encoder and linear prediction head
155 were applied to each tile representation without attention aggregation. This yielded a local
156 target-class score for classification and a local risk score for survival. Region comparisons
157 used the individual tile scores intersecting the supplied annotations.

158 For stomach adenocarcinoma, local genomically stable subtype scores were compared
159 between poorly differentiated and highly differentiated regions using separate one-sided
160 Mann–Whitney U tests for each displayed slide. For breast invasive carcinoma, the local
161 overall survival risk scores were compared between tumor and stromal regions defined by
162 the Breast Cancer Semantic Segmentation annotations [3]. These tile-level comparisons
163 were descriptive, and patient-level clinical performance was evaluated only from the out-
164 of-fold predictions described above.

165 **S1.6 Survival stratification and multivariable Cox regression**

166 Kaplan–Meier and Cox analyses involved the same patient-level out-of-fold risk scores used
167 to calculate the clinical concordance indices. For clearer visualization, patients with liver
168 hepatocellular carcinoma and lower-grade glioma were divided at the cohort median of
169 the out-of-fold risk score. Kaplan–Meier curves included 95% confidence intervals and
170 numbers at risk, and high- and low-risk groups were compared using a two-sided log-rank
171 test [4, 5]. The cohort-wide median split was descriptive and was not used to calculate the
172 concordance index.

173 For multivariable analysis, the continuous HistAgent risk score was standardized so
174 that one unit represented one standard deviation. Complete-case Cox proportional-hazards
175 models were fitted separately for the two cohorts. The liver hepatocellular carcinoma model
176 included the HistAgent risk score and American Joint Committee on Cancer stage, and the
177 lower-grade glioma model included the HistAgent risk score and isocitrate dehydrogenase
178 status. Hazard ratios, 95% confidence intervals, and two-sided Wald P values were reported
179 for a one-standard-deviation increase in the HistAgent risk score. The proportional-hazards
180 assumption was assessed using scaled Schoenfeld residuals [6].

181 **S1.7 Comparisons of methods**

182 **S1.7.1 Compared methods for spatial molecular recovery**

183 We compared HistAgent with the frozen visual-omics foundation models STPath [7] and
184 OmiCLIP [8]. All three methods received the same H&E slides, were aligned to the same

capture-spot coordinates, and were evaluated over canonical gene symbols. In the slide analysis, HistAgent used its directly generated ranked molecular readout for ranked recovery and the reference-guided expression estimates described above for all analyses requiring continuous values.

Paired measured ST data were used to define the evaluation targets, rather than providing a competing prediction. For each query study, the independent ST reference used for reference-guided expression inference excluded the entire query study. No baseline was fine-tuned on the evaluation slides.

S1.7.2 Compared methods for clinical prediction

For slide j , method m produced a bag of frozen tile representations $\mathbf{Z}_j^{(m)} \in \mathbb{R}^{n_j \times d_m}$, where only the representation dimension d_m and the model-specific encoder differed. We compared HistAgent’s 512-dimensional intermediate visual-omics representation with eight methods:

1. **GigaPath** [9], using the 1,536-dimensional tile representation produced by its released pretrained tile encoder.
2. **UNI v2** [10], using its released 1,536-dimensional pathology tile representation.
3. **MUSK** [11], using its released 2,048-dimensional pathology representation.
4. **CHIEF** [12], using its released 768-dimensional pathology representation.
5. **CONCH** [13], using its released 512-dimensional pathology representation.
6. **OmiCLIP** [8], using its released 768-dimensional H&E-omics representation.
7. **STPath** [7], using its 512-dimensional contextual tile representation from the final spatial-transformer layer before the gene-expression prediction head.
8. **Path2Space** [14], using the organ-specific predicted gene-expression vector generated for each tile. For non-breast endpoints, organ-specific Path2Space models were trained using external organ-matched paired H&E-ST reference slides that excluded all evaluation patients, and the resulting models were frozen before clinical prediction.

We used the official implementations and released weights for GigaPath (<https://github.com/prov-gigapath/prov-gigapath>; <https://huggingface.co/prov-gigapath/prov-gigapath>), UNI v2 (<https://github.com/mahmoodlab/UNI>; <https://huggingface.co/MahmoodLab/UNI2-h>), MUSK (<https://github.com/lilab-stanford/MUSK>; <https://huggingface.co/xiangjx/musk>), CHIEF (<https://github.com/hms-dbmi/CHIEF>), CONCH (<https://github.com/mahmoodlab/CONCH>; <https://huggingface.co/MahmoodLab/CONCH>), OmiCLIP (<https://github.com/GuangyuWangLab2021/Loki>; <https://huggingface.co/WangGuangyuLab/Loki>), STPath (<https://github.com/Graph-and-Geometric-Learning/STPath>; <https://huggingface.co/tlhuang/STPath>) and Path2Space (<https://github.com>).

220 [com/eldadshulman/path2space-companion; https://doi.org/10.5281/zenodo.20174301](https://doi.org/10.5281/zenodo.20174301)).
 221 Model-specific image normalization and input resolution followed the corresponding re-
 222 leased implementation.

223 All encoders were frozen. Within each endpoint, the nine methods used the same
 224 patient intersection, patient folds, random seeds, ABMIL architecture, tile caps, and early-
 225 stopping rule. The same tile coordinates were used whenever supported by the encoder.
 226 External ST reference data required by an organ-specific method excluded all evaluation
 227 patients.

228 S1.7.3 Compared method for natural-language atlas retrieval

229 We compared HistAgent with CellWhisperer, a multimodal model that aligns transcrip-
 230 tomes and text [15], using its official implementation and released model (<https://github.com/epigen/CellWhisperer>).
 231

232 S1.8 Evaluation metrics

233 We used task-specific metrics for ranked molecular recovery, spatial-profile agreement,
 234 clinical prediction, atlas retrieval, and evidence grounding by the spatial agentic module.
 235 Unless stated otherwise, spot- or gene-level values were first averaged within each tissue
 236 slide. Slide-level values in the 135-slide molecular analysis were summarized without spot
 237 weighting within each species–organ cohort. Higher values indicate better performance,
 238 except for unsupported molecular claims, for which lower values indicate better grounding.

239 S1.8.1 Ranked molecular-recovery metrics

240 For spot i , let $\mathcal{M}_i^K$ be the set of the K most highly expressed genes in measured ST and
 241 let $\hat{\pi}_i = (\hat{g}_{i,1}, \dots, \hat{g}_{i,L_i})$ be the predicted ranking. We used $K = 50$. The relevance at
 242 predicted rank r was

$$R_i(r) = \mathbf{1}(\hat{g}_{i,r} \in \mathcal{M}_i^K). \quad (\text{S7})$$

243 HitRate@ K measured the fraction of the measured top- K genes recovered in the predicted
 244 top- K list:

$$\text{HitRate}_i @ K = \frac{\sum_{r=1}^{\min(K, L_i)} R_i(r)}{|\mathcal{M}_i^K|}. \quad (\text{S8})$$

245 Precision at predicted rank r and average precision at K were

$$P_i(r) = \frac{1}{r} \sum_{t=1}^r R_i(t), \quad \text{AP}_i @ K = \frac{1}{|\mathcal{M}_i^K|} \sum_{r=1}^{\min(K, L_i)} P_i(r) R_i(r). \quad (\text{S9})$$

246 For slide s containing n_s evaluated spots, the mean average precision was

$$\text{mAP}_s @ K = \frac{1}{n_s} \sum_{i=1}^{n_s} \text{AP}_i @ K. \quad (\text{S10})$$

HitRate was aggregated over spots in the same way:

$$\text{HitRate}_s @K = \frac{1}{n_s} \sum_{i=1}^{n_s} \text{HitRate}_i @K. \quad (\text{S11})$$

For either slide-level metric $M_s^{(m)}$, the overall score for method m was the unweighted mean over the $S = 135$ slides,

$$\overline{M}^{(m)} = \frac{1}{S} \sum_{s=1}^S M_s^{(m)}. \quad (\text{S12})$$

Species–organ summaries used the same unweighted mean restricted to the slides in that cohort.

S1.8.2 Spatial-profile agreement

For gene g in slide s , let $\mathbf{x}_{s,g}$ and $\widehat{\mathbf{x}}_{s,g}$ denote the measured and predicted profiles over the same n_s spots. The Pearson correlation coefficient (PCC) was

$$\text{PCC}_{s,g} = \frac{\sum_{i=1}^{n_s} (x_{s,i,g} - \bar{x}_{s,g})(\widehat{x}_{s,i,g} - \widehat{\bar{x}}_{s,g})}{\sqrt{\sum_{i=1}^{n_s} (x_{s,i,g} - \bar{x}_{s,g})^2} \sqrt{\sum_{i=1}^{n_s} (\widehat{x}_{s,i,g} - \widehat{\bar{x}}_{s,g})^2}}. \quad (\text{S13})$$

PCCs were averaged across the genes included in the analysis for each slide. A gene with an undefined PCC because the predicted profile was constant was assigned a PCC of zero rather than excluded. The overall PCC was the unweighted mean of the 135 slide-level means, and species–organ summaries were unweighted means over slides within each cohort.

For published spatial biological findings, the area under the precision–recall curve (AUPRC) compared continuous signature scores with the corresponding binary expert annotation. To preserve tied model scores, let $\tau_1 > \dots > \tau_J$ be their distinct thresholds and let P_j and R_j denote precision and recall after all spots with scores of at least τ_j were included. With $R_0 = 0$, non-interpolated AUPRC was

$$\text{AUPRC} = \sum_{j=1}^J (R_j - R_{j-1}) P_j. \quad (\text{S14})$$

For each slide and biological finding, the AUPRC was calculated from the corresponding spot-level scores and binary annotation.

S1.8.3 Metrics for standard ST analyses

For SVG analysis, let $\mathbf{W} = (w_{ij})$ be the spatial adjacency matrix, $W_+ = \sum_i \sum_j w_{ij}$ and $\bar{x}_g$ the mean value for gene g . Moran’s I was

$$I_g(\mathbf{x}) = \frac{n_s}{W_+} \frac{\sum_i \sum_j w_{ij} (x_{i,g} - \bar{x}_g)(x_{j,g} - \bar{x}_g)}{\sum_i (x_{i,g} - \bar{x}_g)^2}. \quad (\text{S15})$$

270 The reported metric was the Spearman correlation coefficient between the predicted and
 271 measured gene-wise Moran's I vectors:

$$S_s^{\text{SVG}} = \rho_s(\{I_g(\hat{\mathbf{x}})\}_{g \in \mathcal{G}_s}, \{I_g(\mathbf{x})\}_{g \in \mathcal{G}_s}), \quad (\text{S16})$$

272 where gene set $\mathcal{G}_s$ was defined only from the measured ST. Spearman correlation ρ_s was
 273 calculated as the PCC between the midrank-transformed values. Candidate Moran's I
 274 values that were undefined because of a constant predicted profile were imputed zero
 275 rather than removed from the evaluation.

276 Spatial-domain agreement was evaluated using normalized mutual information (NMI)
 277 and adjusted Rand index (ARI). For predicted partition U and fixed measured ST partition
 278 V ,

$$\text{NMI}(U, V) = \frac{2I(U; V)}{H(U) + H(V)}, \quad (\text{S17})$$

279 where $I(U; V)$ is mutual information and $H(\cdot)$ is entropy. Let n_{uv} be the number of spots
 280 assigned to predicted domain u and measured domain v , with $a_u = \sum_v n_{uv}$, $b_v = \sum_u n_{uv}$
 281 and $n = \sum_{u,v} n_{uv}$. The ARI was

$$\text{ARI} = \frac{\sum_{u,v} \binom{n_{uv}}{2} - \frac{(\sum_u \binom{a_u}{2}) (\sum_v \binom{b_v}{2})}{\binom{n}{2}}}{\frac{1}{2} \left[\sum_u \binom{a_u}{2} + \sum_v \binom{b_v}{2} \right] - \frac{(\sum_u \binom{a_u}{2}) (\sum_v \binom{b_v}{2})}{\binom{n}{2}}}. \quad (\text{S18})$$

282 For cell-type deconvolution, let $\mathbf{a}_{s,i}$ and $\hat{\mathbf{a}}_{s,i}$ be the measured and predicted relative
 283 cell-type composition profiles over the same retained cell types at spot i . The slide score
 284 was

$$S_s^{\text{cell}} = \frac{1}{|\mathcal{I}_s|} \sum_{i \in \mathcal{I}_s} \rho_s(\hat{\mathbf{a}}_{s,i}, \mathbf{a}_{s,i}), \quad (\text{S19})$$

285 where $\mathcal{I}_s$ contained spots with at least three active marker signatures in measured ST and
 286 a nonconstant measured profile. An undefined predicted correlation was assigned a value
 287 of zero rather than removed.

288 For differential expression, let $\mathcal{I}_{s,d}$ be the spots in a fixed measured ST domain d and
 289 $\mathcal{I}_{s,-d}$ the remaining spots. After gene-wise standardization across spots, the one-versus-rest
 290 effect was

$$\delta_{s,d,g} = \frac{1}{|\mathcal{I}_{s,d}|} \sum_{i \in \mathcal{I}_{s,d}} z_{s,i,g} - \frac{1}{|\mathcal{I}_{s,-d}|} \sum_{i \in \mathcal{I}_{s,-d}} z_{s,i,g}. \quad (\text{S20})$$

291 Agreement was the mean domain-wise Spearman correlation between predicted and mea-
 292 sured gene-effect vectors:

$$S_s^{\text{DE}} = \frac{1}{|\mathcal{D}_s|} \sum_{d \in \mathcal{D}_s} \rho_s(\{\hat{\delta}_{s,d,g}\}_{g \in \mathcal{G}_s}, \{\delta_{s,d,g}\}_{g \in \mathcal{G}_s}). \quad (\text{S21})$$

293 Only domains containing at least five spots and leaving at least five spots in the one-

versus-rest comparison were included. An undefined predicted domain-wise correlation was imputed zero rather than removed.

For pathway analysis, let $K' = \min(K, N)$, where N is the number of candidate pathways. The discounted cumulative gain (DCG) and normalized discounted cumulative gain (nDCG) at K [16] for a predicted pathway ranking were

$$\text{DCG @}K = \sum_{r=1}^{K'} \frac{\gamma_{\pi_r}}{\log_2(r+1)}, \quad \text{nDCG @}K = \frac{\text{DCG @}K}{\text{IDCG @}K}, \quad (\text{S22})$$

where γ_{π_r} was the nonnegative pathway score from measured ST and IDCG @ K was obtained from the first K' positions after sorting the measured scores in descending order. Linear gain was used because pathway scores were continuous and unbounded. Spot-level nDCG@10 values were averaged within each slide.

The six reported outcomes from the five standard ST analyses were Moran's- I concordance, NMI, ARI, cell-type composition agreement, differential-expression agreement, and pathway nDCG. For outcome a , the overall value was

$$\overline{M}_a^{(m)} = \frac{1}{135} \sum_{s=1}^{135} M_{s,a}^{(m)}. \quad (\text{S23})$$

Species-organ values were calculated using the same unweighted mean over the slides in each cohort.

S1.8.4 Clinical prediction metrics

Binary classification was the area under the receiver operating characteristic curve (AUROC). For scores s^+ and s^- from randomly selected positive and negative patients,

$$\text{AUROC} = \Pr(s^+ > s^-) + \frac{1}{2} \Pr(s^+ = s^-). \quad (\text{S24})$$

For a multiclass endpoint with C classes, one-versus-rest AUROCs were macro-averaged:

$$\text{macro-AUROC} = \frac{1}{C} \sum_{c=1}^C \text{AUROC}_c. \quad (\text{S25})$$

To summarize performance across the $E = 20$ endpoints, method m received an endpoint-specific rank

$$r_{e,m} = 1 + \sum_{h \neq m} \mathbf{1}(\hat{\theta}_{e,h} > \hat{\theta}_{e,m}) + \frac{1}{2} \sum_{h \neq m} \mathbf{1}(\hat{\theta}_{e,h} = \hat{\theta}_{e,m}), \quad (\text{S26})$$

where $\hat{\theta}_{e,m}$ was AUROC, macro-AUROC, or the concordance index, as appropriate. Tied

315 methods therefore received their average rank. The mean rank was

$$\bar{r}_m = \frac{1}{E} \sum_{e=1}^E r_{e,m}. \quad (\text{S27})$$

316 For survival prediction, the concordance index was used. With observed time T_i , event
317 indicator δ_i , risk η_i and comparable-pair set $\mathcal{A} = \{(i, j) : T_i < T_j, \delta_i = 1\}$,

$$\text{C-index} = \frac{1}{|\mathcal{A}|} \sum_{(i,j) \in \mathcal{A}} \left[\mathbf{1}(\eta_i > \eta_j) + \frac{1}{2} \mathbf{1}(\eta_i = \eta_j) \right]. \quad (\text{S28})$$

318 **S1.8.5 Atlas-retrieval metrics**

319 For a ranked retrieval result, let $K' = \min(K, N)$ for a candidate set of size N . With
320 binary expert relevance y_j and item π_r at retrieval rank r , precision at K was

$$\text{Precision @}K = \frac{1}{K'} \sum_{r=1}^{K'} y_{\pi_r}. \quad (\text{S29})$$

321 AUPRC used the non-interpolated definition above. For retrieval nDCG, binary expert
322 relevance was used as the gain in the DCG definition, with the same K' convention and
323 $K = 50$. The sensitivity analyses involved nDCG recalculated at 10, 50, and 100 retrieved
324 spots and at cutoffs corresponding to the top 1% and 5% of candidate spots within each
325 slide.

326 **S1.9 Datasets and preprocessing for downstream evaluation**

327 **S1.9.1 Molecular-recovery evaluation data**

328 The molecular analysis comprised 135 ST slides, including 79 human and 56 mouse slides
329 from 15 species–organ cohorts. Within each slide, all three predictions and measured ST
330 were restricted to their ordered four-way spot intersection. The slide identifiers and cohort
331 assignments are listed in Supplementary Table 2.

332 For SVG analysis, genes were required to be nonconstant in measured ST and detected
333 in at least the largest of ten spots or 1% of the measured spots in that slide. For cell-type
334 deconvolution, CellMarker sets containing 5–300 genes were retained. At least three active
335 marker signatures in measured ST data were required for a spot to enter the comparison.
336 For pathway analysis, Reactome sets containing 5–500 genes were retained, and a spot
337 required a positive ideal DCG to be evaluated.

338 **S1.9.2 Clinical cohorts and whole-slide preprocessing**

339 The clinical benchmark contained 20 endpoints spanning molecular subtype, histologic
340 subtype, tumor grade, and prognostic risk across The Cancer Genome Atlas, the Clinical
341 Proteomic Tumor Analysis Consortium, and associated image repositories [17–19]. The

342 endpoint definitions, cohorts, patient counts, and data portals are provided in Supple-
343 mentary Table 4. All slides from each patient were assigned to the same cross-validation
344 fold.

345 Whole-slide images were processed at 0.25 μm per pixel using a non-overlapping grid
346 of 224×224 -pixel tiles over tissue regions identified from a low-resolution thumbnail. For
347 HistAgent, each coordinate defined a local tile and a centered contextual view covering
348 four times the linear field of view. A slide was represented by the resulting bag of frozen
349 tile representations.

350 **S1.9.3 Atlas and natural-language retrieval benchmark**

351 Natural-language retrieval was evaluated on 40 unique ST slides using five biological query
352 groups, targeting tumor epithelium, normal epithelium, tumor pathology, cortical layers,
353 and TLS niches. Expert annotations defined the relevant spots for each query. Query
354 wording, annotation targets, data sources, and slide counts are provided in Supplementary
355 Table 5.

Supplementary Figures

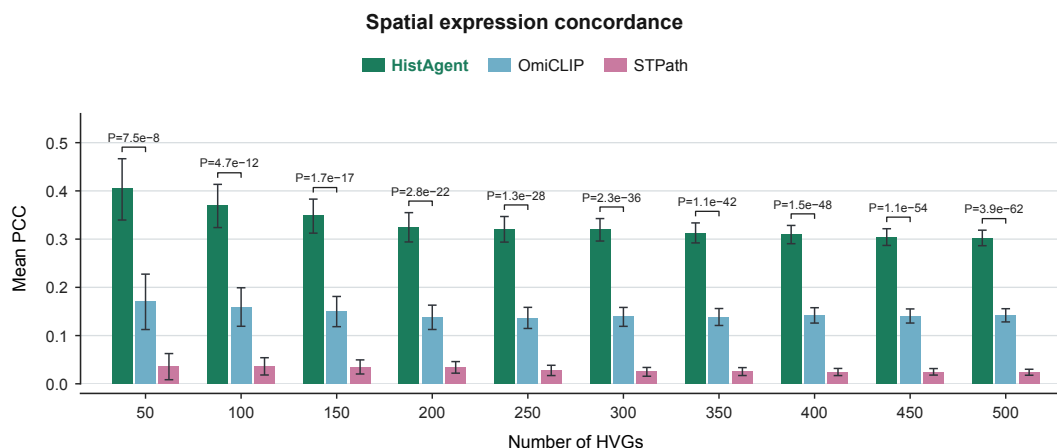


Figure S1: **Spatial expression concordance across increasing highly variable gene set sizes.** Mean gene-wise Pearson correlation coefficients (PCCs) between predicted and measured spatial expression compared across nested highly variable gene (HVG) sets in a representative renal cell carcinoma tertiary lymphoid structure slide. Error bars show 95% bootstrap confidence intervals. P values are from paired two-sided Wilcoxon signed-rank tests with Bonferroni correction.



Figure S2: **Slide-level distributions of spot-level gene-ranking recovery.** HitRate@50 indicates coverage of the measured top-50 expressed genes in each predicted ranking, and mean average precision at 50 (mAP@50) also considers their ranked positions. a, b, Human and mouse HitRate@50 distributions. Scores are shown for held-out human and mouse organ cohorts. Each point denotes one slide; violin outlines show the score distributions, thick lines show interquartile ranges, and white points show medians.

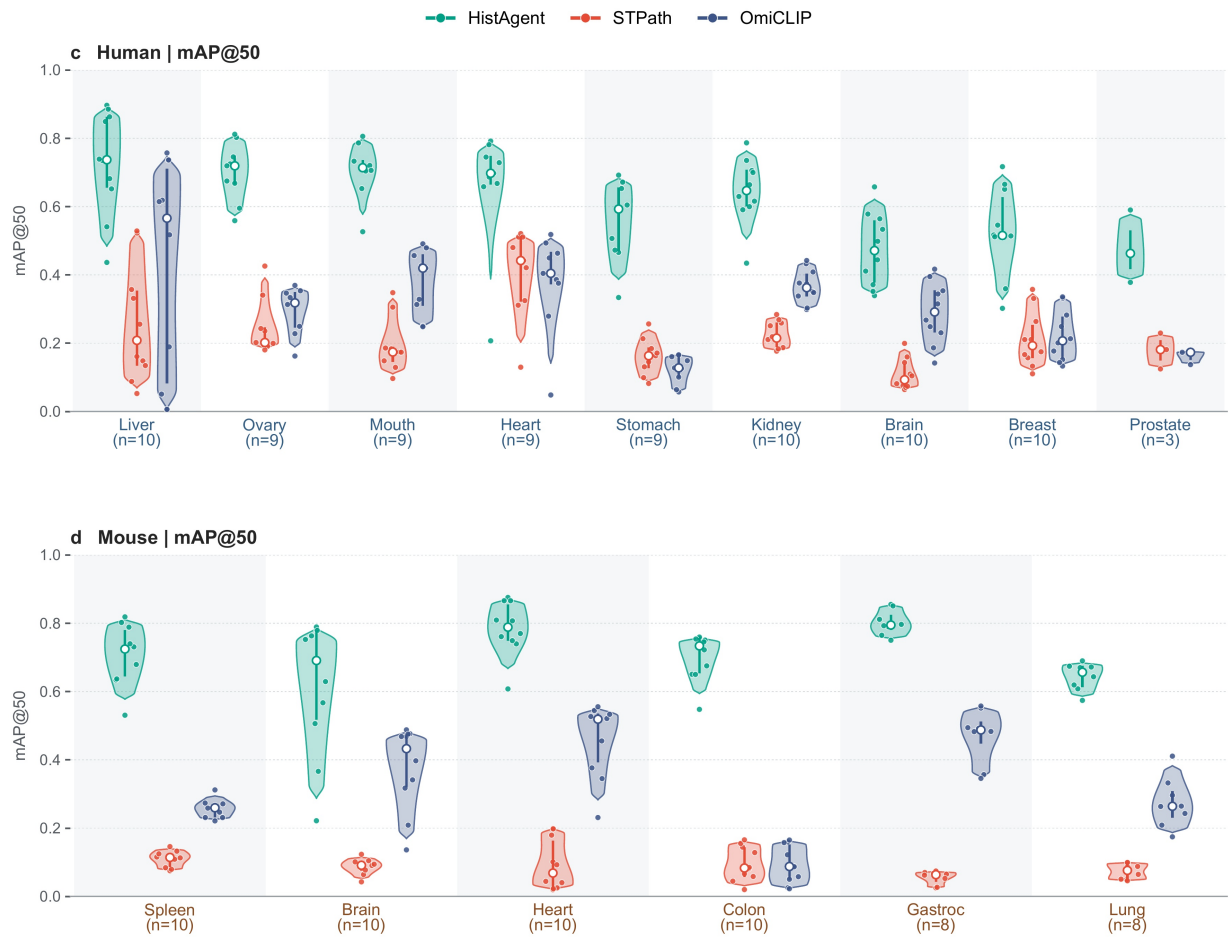


Figure S2: Slide-level distributions of spot-level gene-ranking recovery (continued). c,d, Human and mouse mAP@50 distributions.

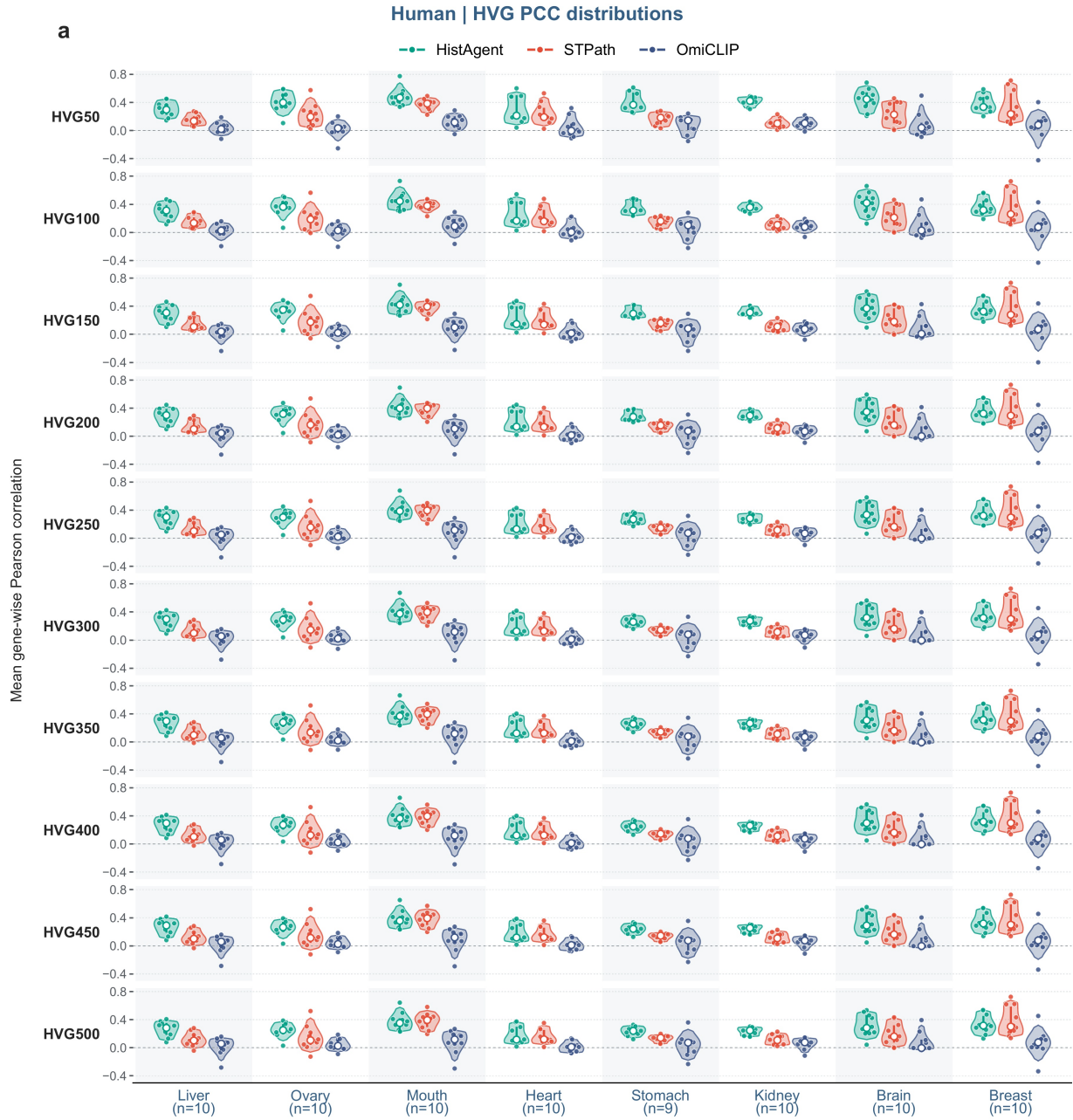


Figure S3: **Gene-level spatial recovery across increasing highly variable gene set sizes.** Mean gene-wise Pearson correlation coefficients (PCCs) between predicted and measured spatial expression are shown across nested highly variable gene (HVG) sets for held-out human (a) and mouse organ cohorts. Each point denotes one slide; violin outlines show the score distributions, thick lines show interquartile ranges, and white points show medians.

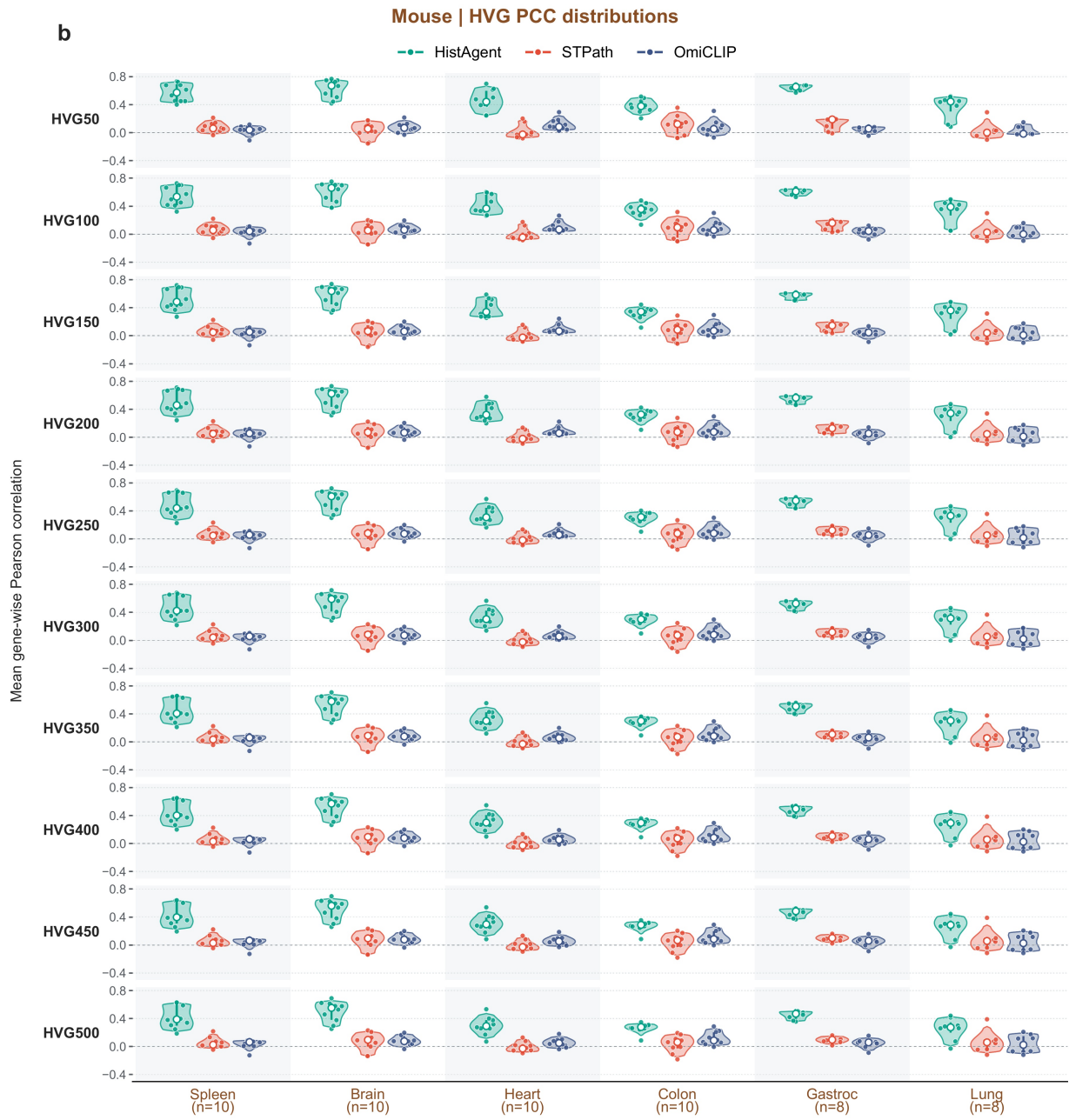


Figure S3: Gene-level spatial recovery across increasing highly variable gene set sizes (continued). **b**, Mouse organ cohorts.



Figure S4: Slide-level distributions of downstream spatial transcriptomic agreement. Agreement between each predicted readout and measured spatial transcriptomics (ST) for five analyses across human and mouse cohorts. Spatially variable gene (SVG) agreement is according to the Spearman correlation of gene-wise Moran's I ; spatial-domain agreement is according to the adjusted Rand index (ARI) and normalized mutual information (NMI); cell-type composition and differential-expression agreement is according to Spearman correlations; and pathway agreement is according to the normalized discounted cumulative gain (nDCG). Each point denotes one slide, and higher values indicate closer agreement with measured ST data. Violin outlines show the score distributions, thick lines show interquartile ranges, and white points show medians.

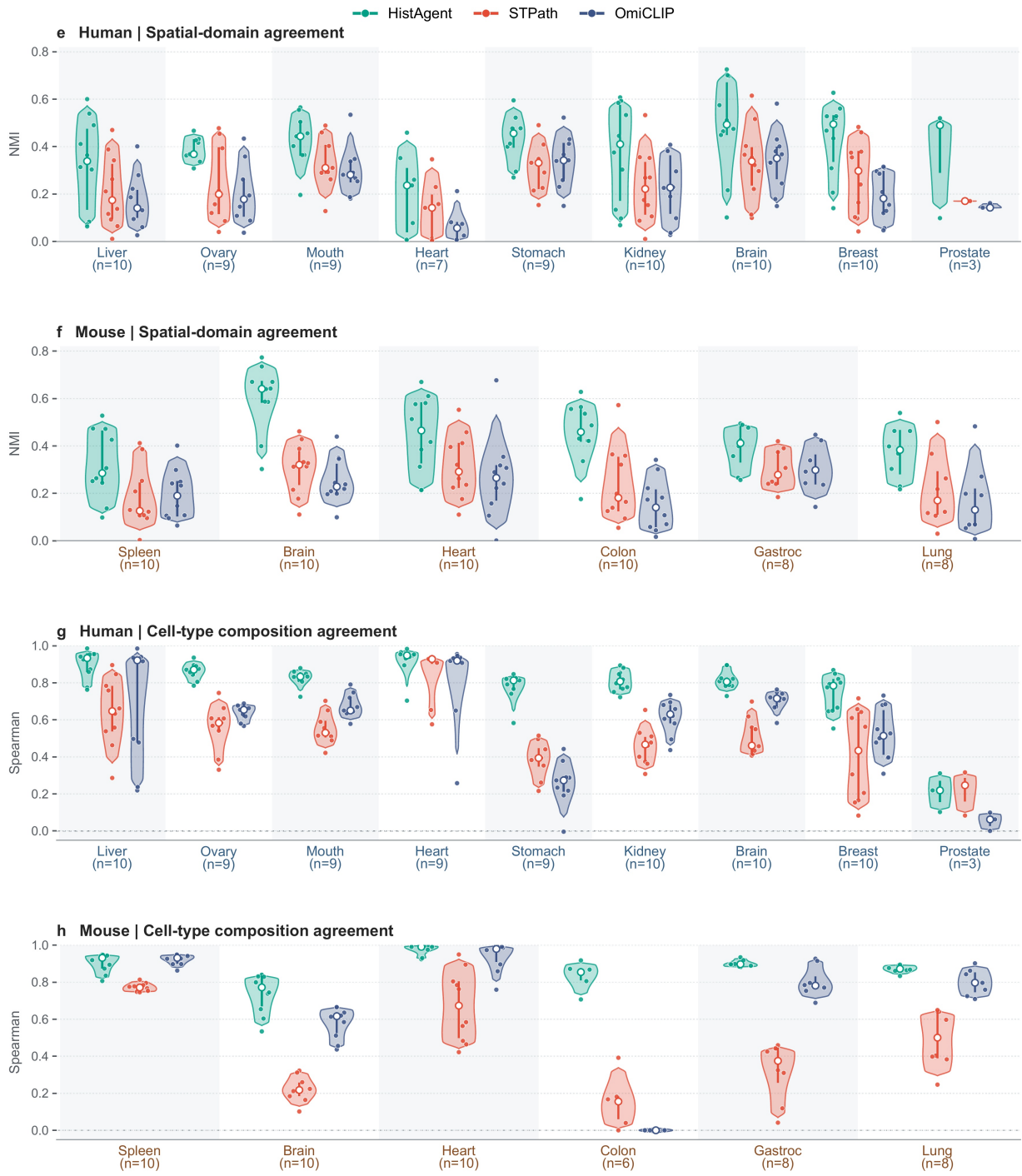


Figure S4: Slide-level distributions of downstream spatial transcriptomic agreement (continued). e–h, Spatial-domain NMI and cell-type composition agreement in human and mouse cohorts.



Figure S4: Slide-level distributions of downstream spatial transcriptomic agreement (continued). i–l, Differential-expression and pathway-ranking agreement in human and mouse cohorts.

a Human I | Spatial organization

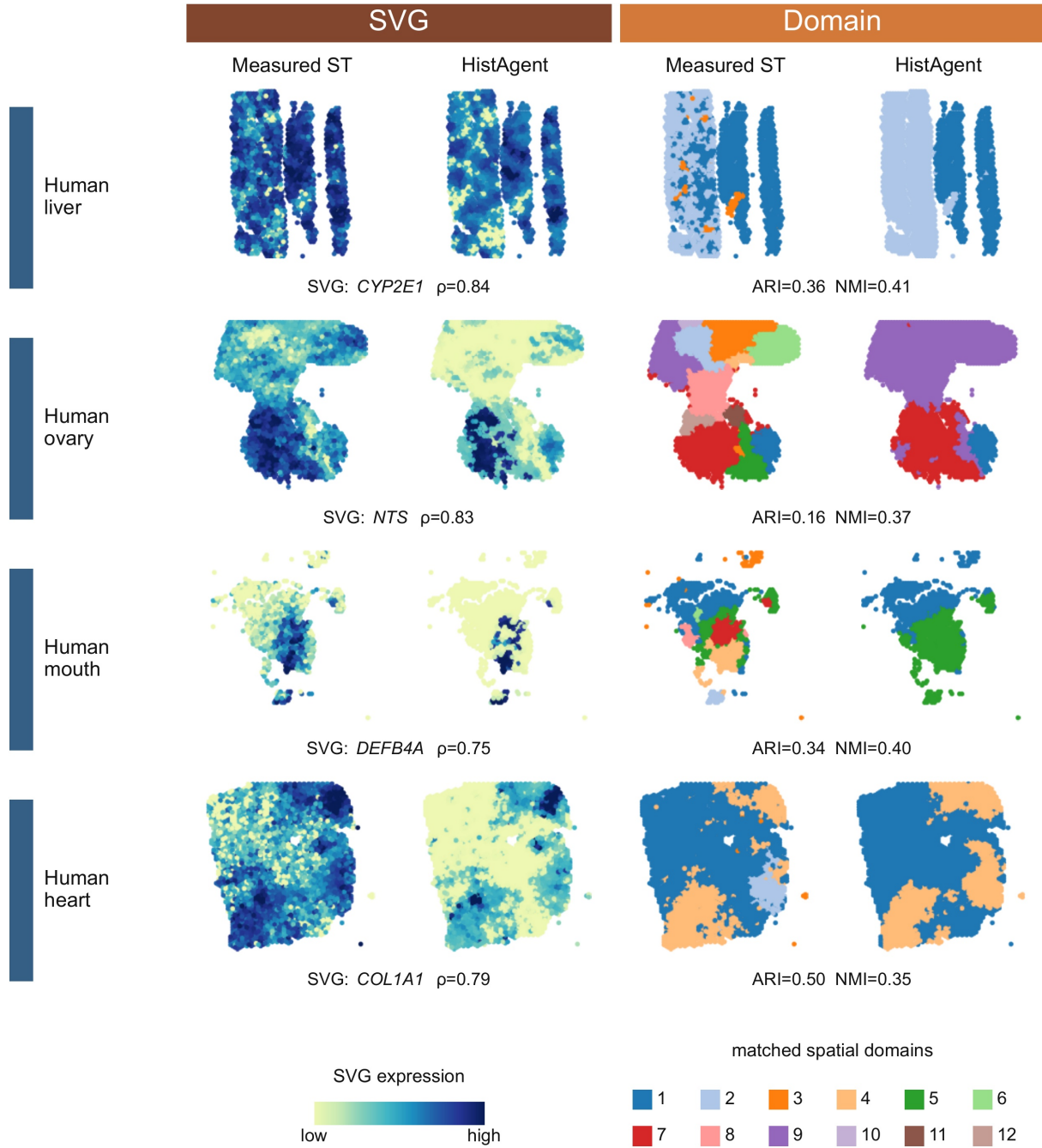


Figure S5: **HistAgent** spatial readouts for organ cohorts not shown in Fig. 3d. One held-out slide is shown for each remaining organ cohort. Paired measured ST and HistAgent visualizations of spatially variable gene (SVG) expression, spatial domains, cell-type composition, differential-expression marker expression and pathway scores. a, Spatial organization in human liver, ovary, mouth, and heart. Slides near the median HistAgent performance were selected without reference to baseline performance, and displayed features were selected for high spatial heterogeneity with measured ST. Values below the continuous maps are spatial Spearman correlation coefficients (ρ); domains were evaluated via ARI and NMI. Domain colors are matched within each pair, and continuous color scales are feature-specific.

b Human I | Cellular and functional state

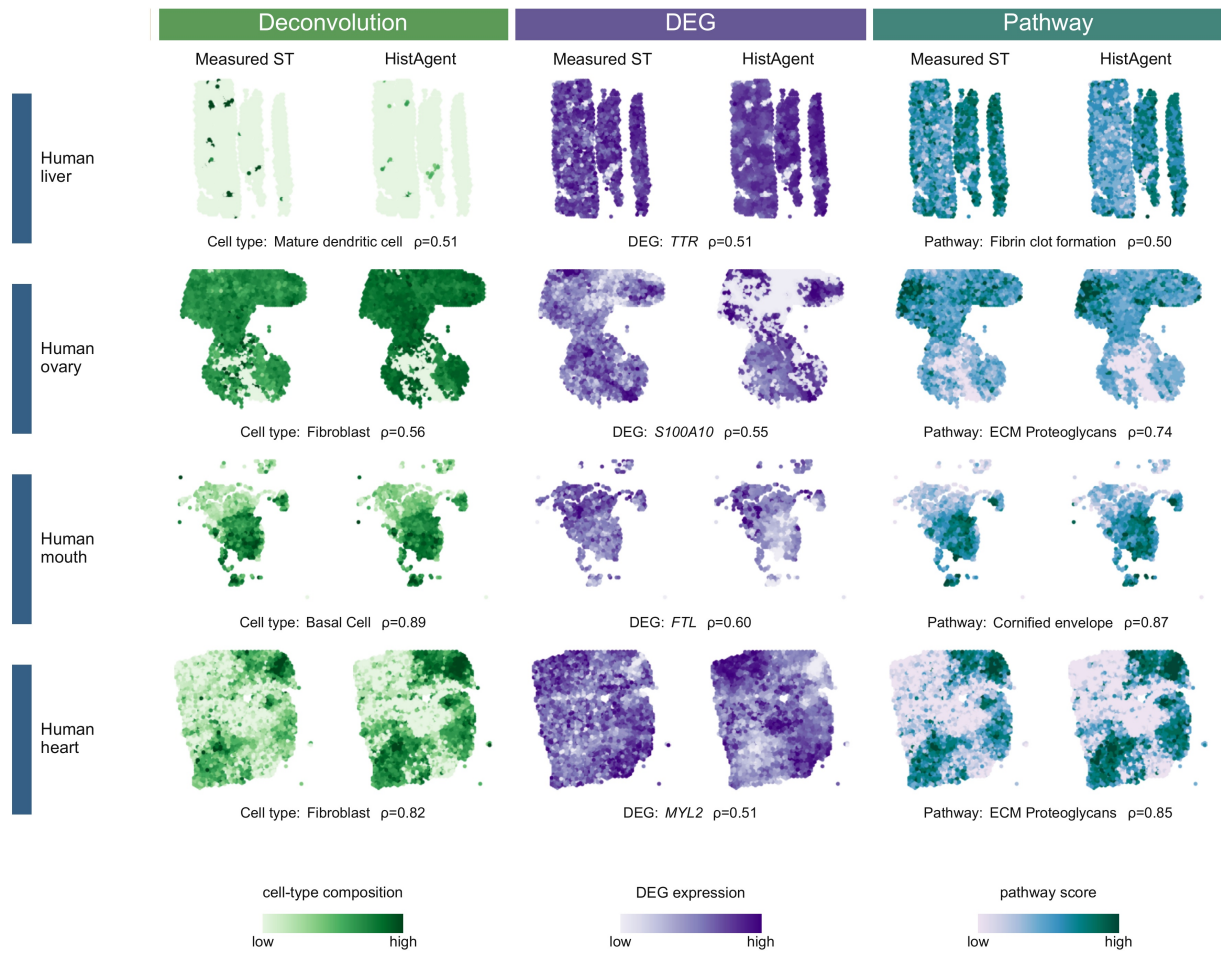


Figure S5: **HistAgent** spatial readouts for organ cohorts not shown in Fig. 3d (continued). b, Cellular and functional readouts for human liver, ovary, mouth, and heart.

C Human II | Spatial organization

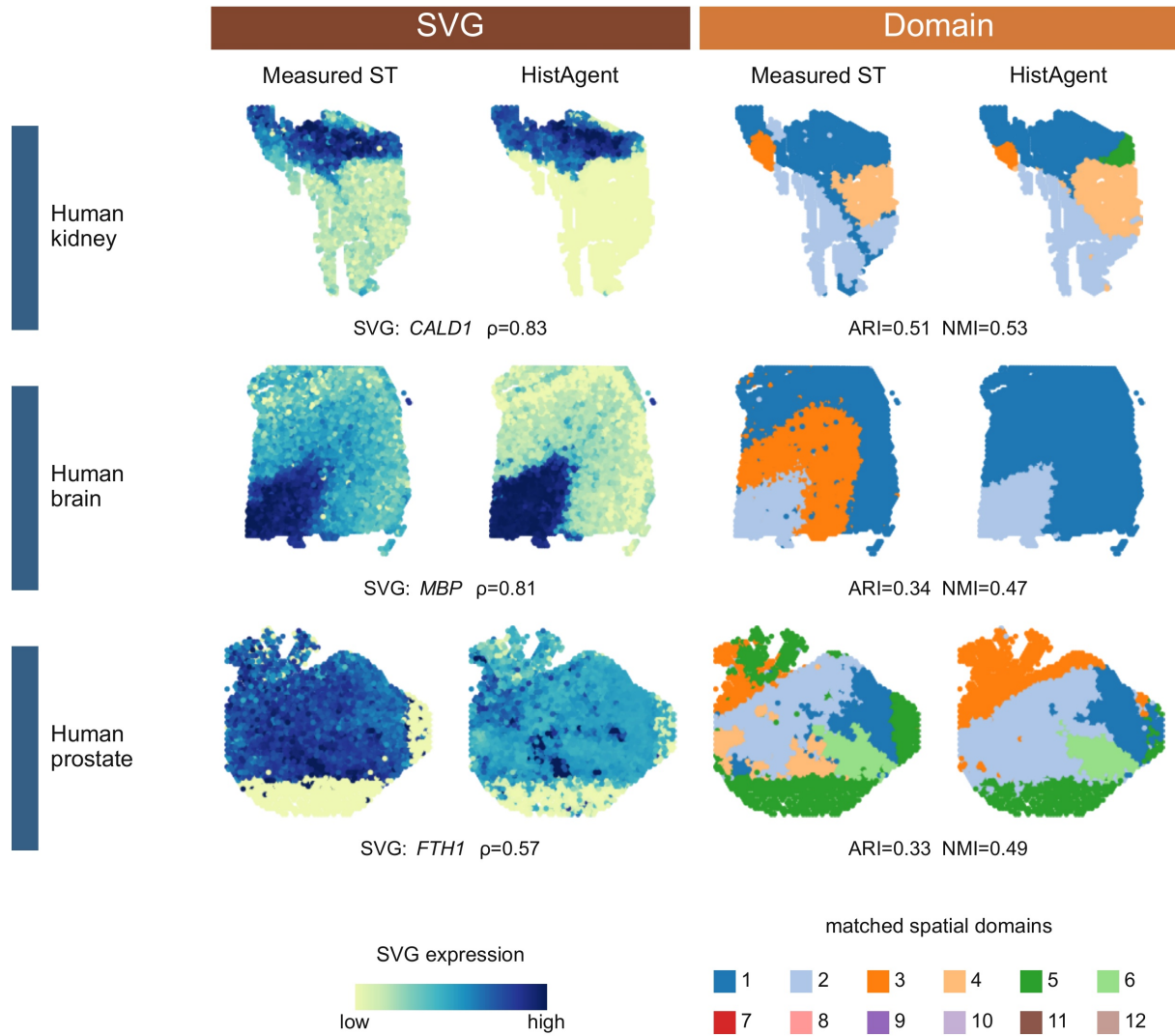


Figure S5: **HistAgent** spatial readouts for organ cohorts not shown in Fig. 3d (continued). c, Spatial organization in kidney, brain and prostate.

d Human II | Cellular and functional state

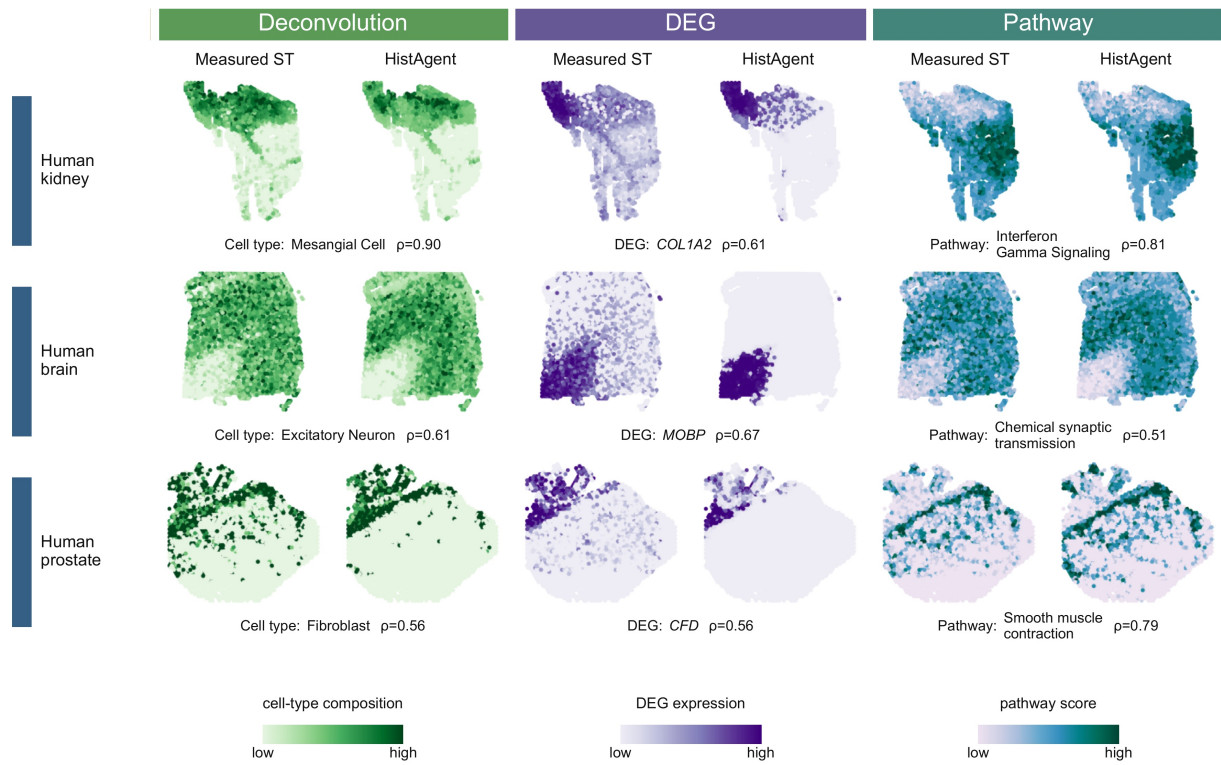


Figure S5: **HistAgent** spatial readouts for organ cohorts not shown in Fig. 3d (continued). d, Cellular and functional readouts in human kidney, brain and prostate.

e Mouse | Spatial organization

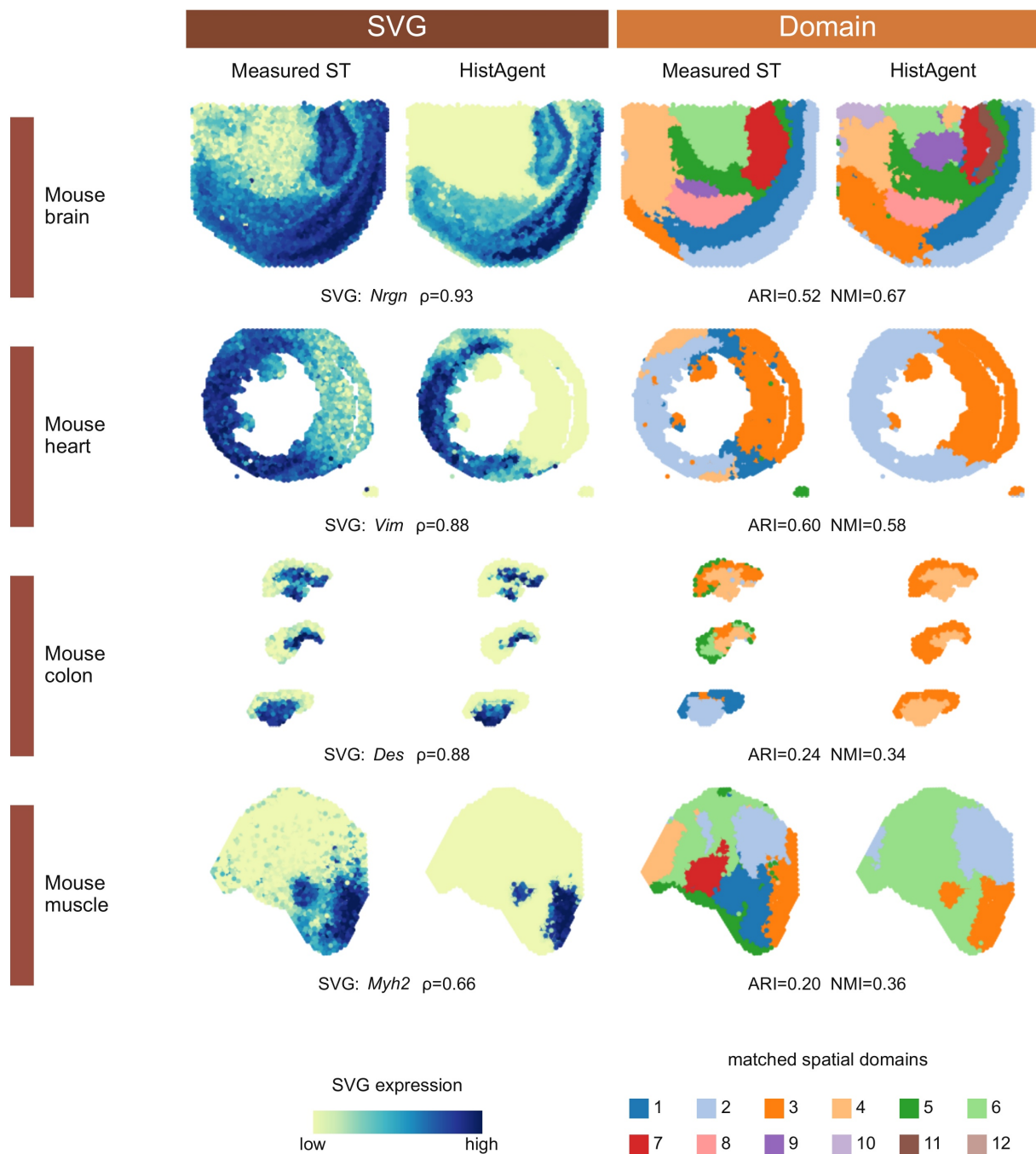


Figure S5: **HistAgent** spatial readouts for organ cohorts not shown in Fig. 3d (continued). **e**, Spatial organization in mouse brain, heart, colon and muscle.

f Mouse | Cellular and functional state

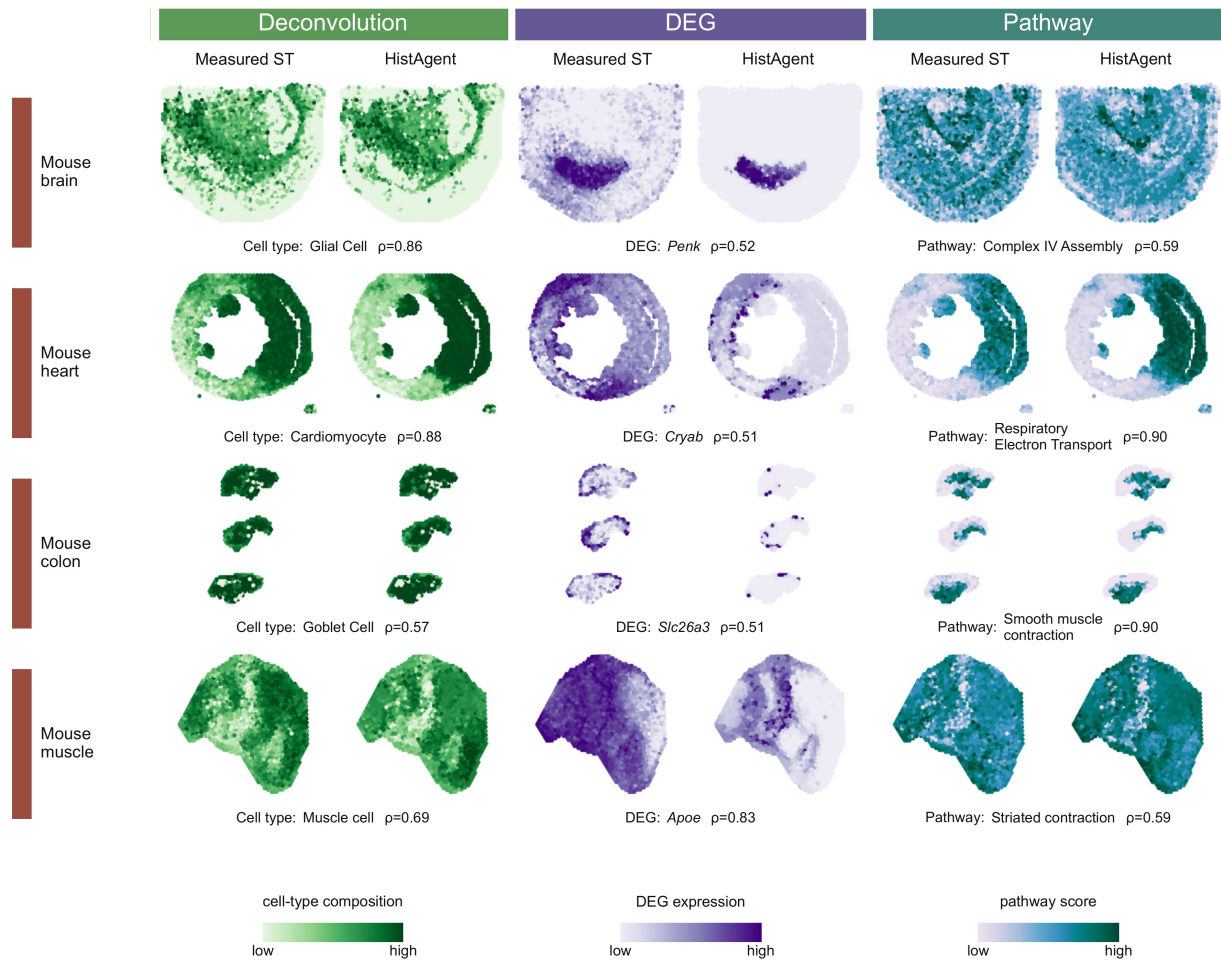


Figure S5: **HistAgent** spatial readouts for organ cohorts not shown in Fig. 3d (continued). **f**, Cellular and functional readouts in mouse brain, heart, colon and muscle.

a Human I | Spatial organization

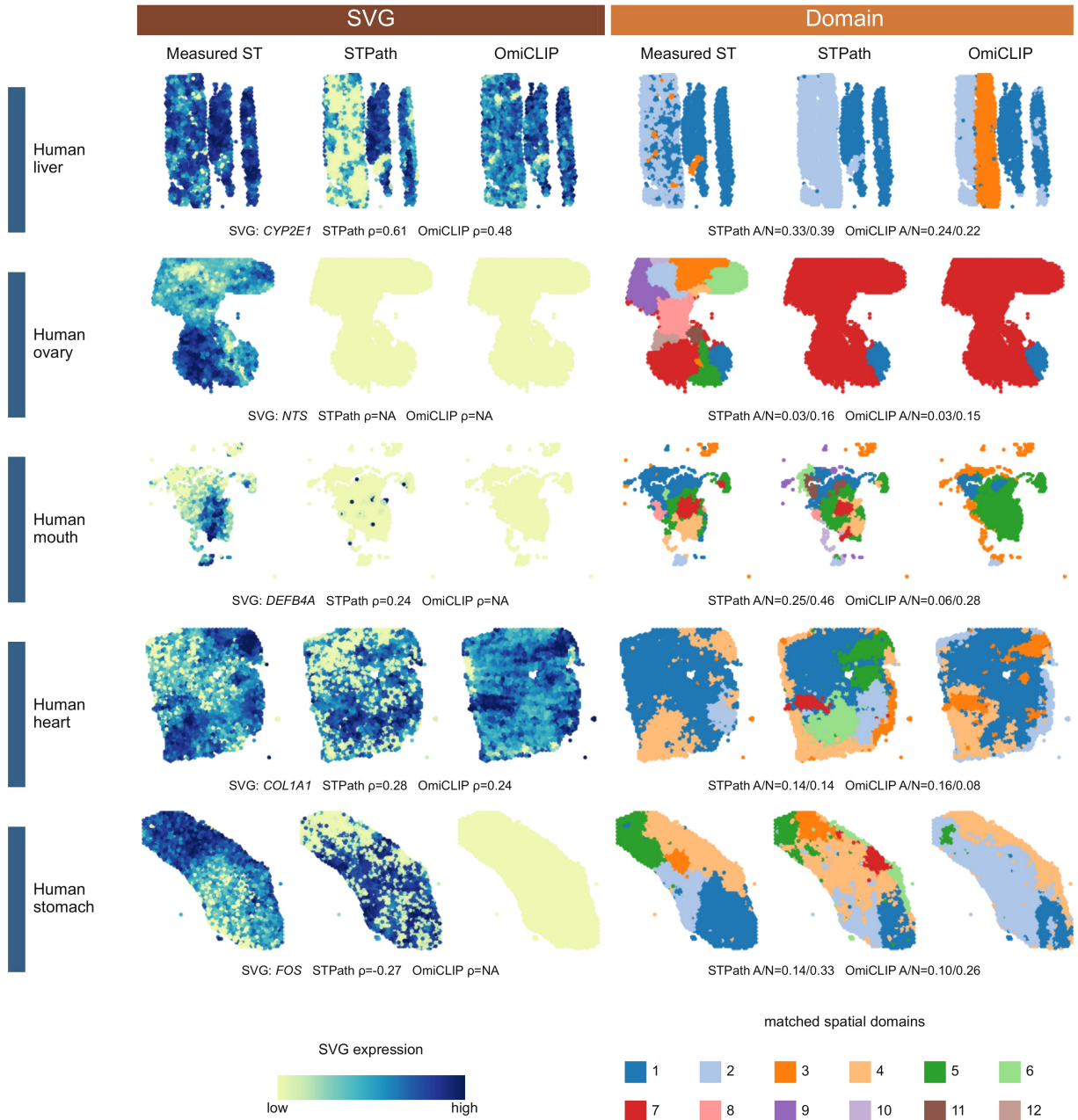


Figure S6: **STPath** and **OmiCLIP** spatial readouts for the **Figure 3** benchmark. Measured ST, STPath, and OmiCLIP visualizations for the same slides and features used in Fig. 3d and Supplementary Fig. S5. The comparisons cover spatially variable gene (SVG) expression, spatial domains, cell-type composition, differential-expression marker expression and pathway scores. Values below the continuous maps are spatial Spearman correlations; domains are evaluated by adjusted Rand index (ARI) and normalized mutual information (NMI). Domain colors are matched within each comparison, and continuous color scales are feature-specific.

b Human I | Cellular and molecular state

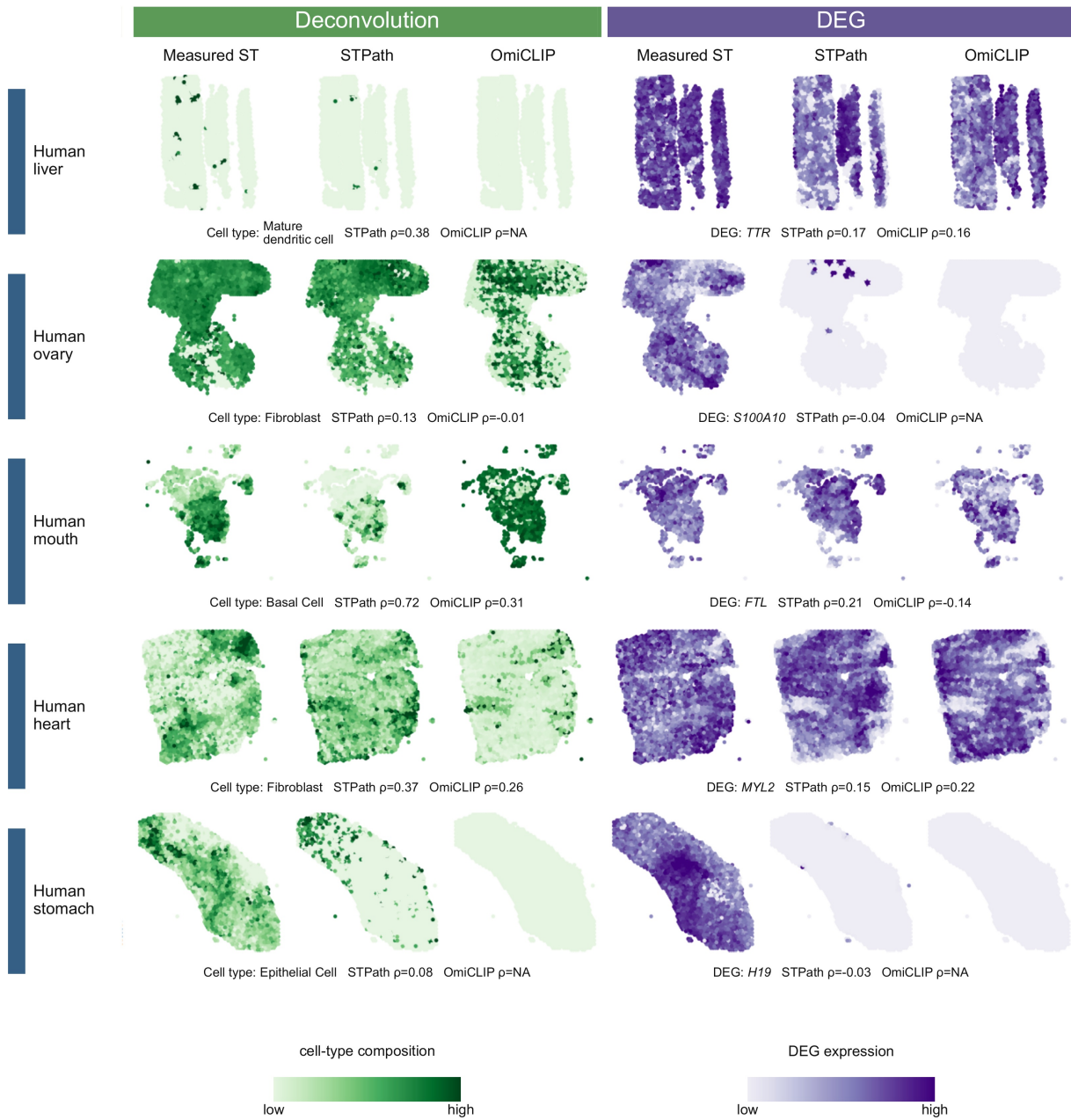


Figure S6: STPath and OmiCLIP spatial readouts for all organ cohorts in the Figure 3 benchmark (continued). b, Cellular and molecular state of human liver, ovary, mouth, heart and stomach.

C Human I | Pathway activity



Figure S6: STPath and OmiCLIP spatial readouts for all organ cohorts in the **Figure 3 benchmark (continued)**. c, Pathway activity in human liver, ovary, mouth, heart and stomach.

d Human II | Spatial organization

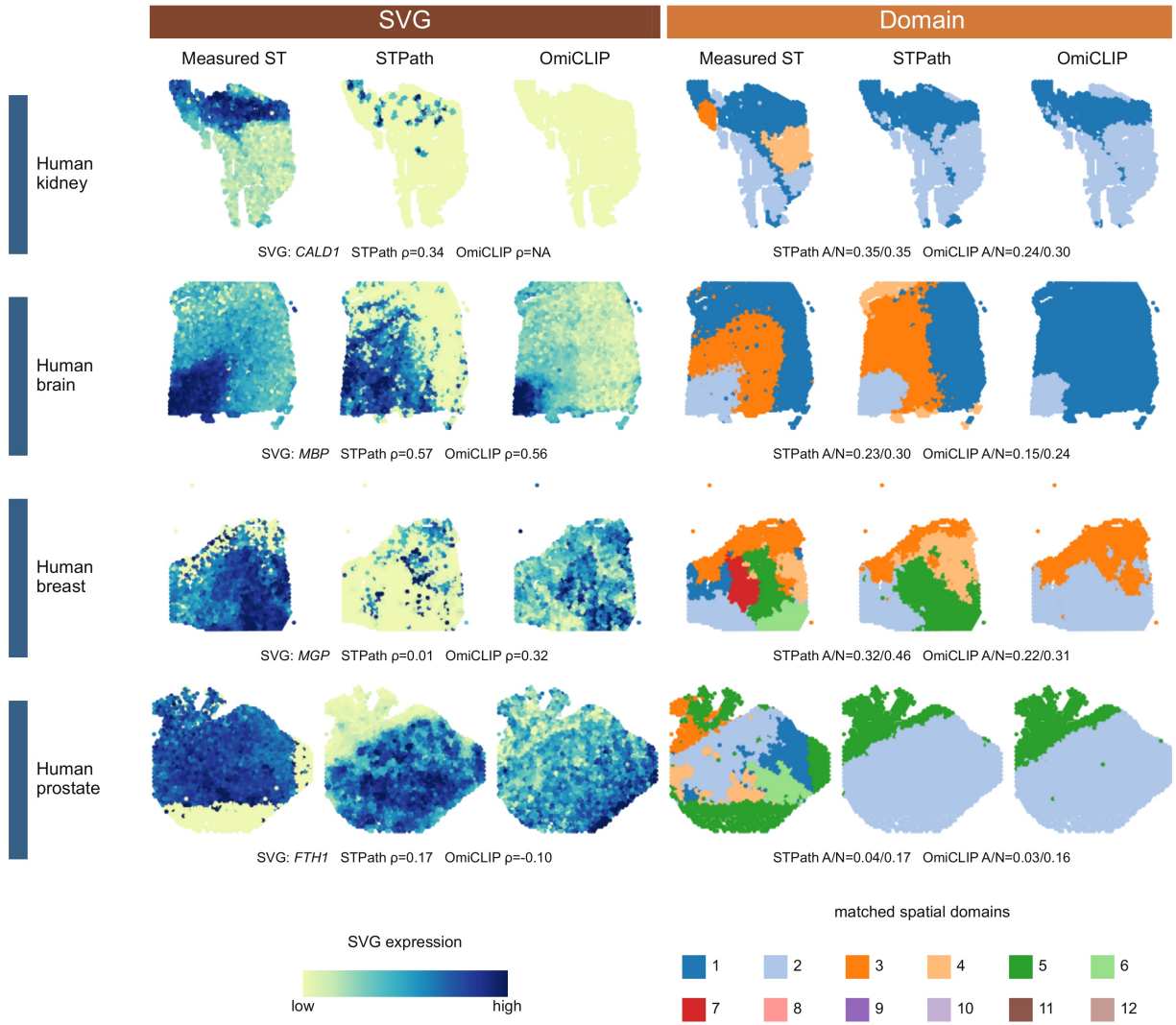


Figure S6: STPath and OmiCLIP spatial readouts for all organ cohorts in the Figure 3 benchmark (continued). d, Spatial organization in human kidney, brain, breast and prostate.

e Human II | Cellular and molecular state

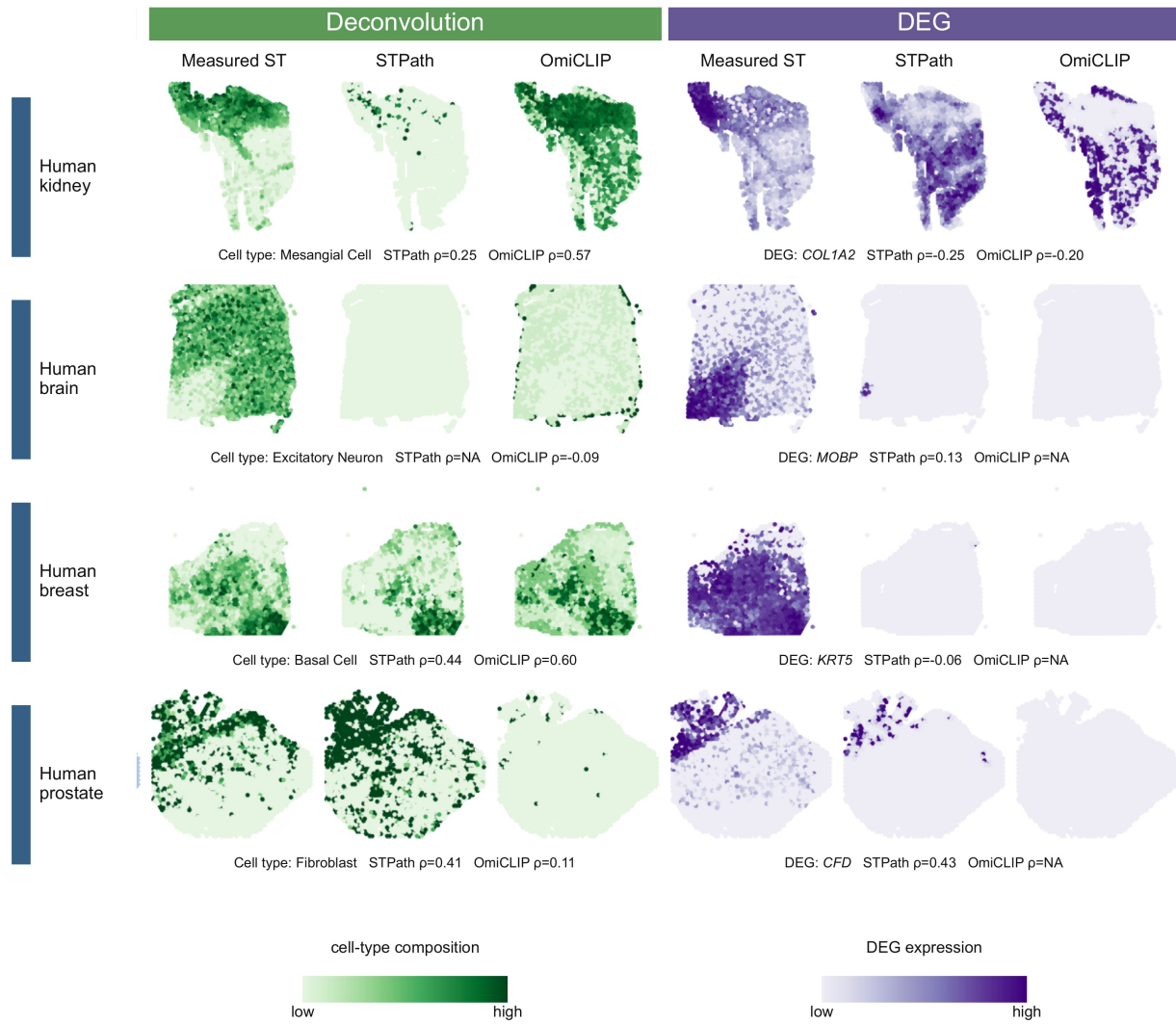


Figure S6: STPath and OmiCLIP spatial readouts for all organ cohorts in the Figure 3 benchmark (continued). e, Cellular and molecular state in human kidney, brain, breast and prostate.

f Human II | Pathway activity

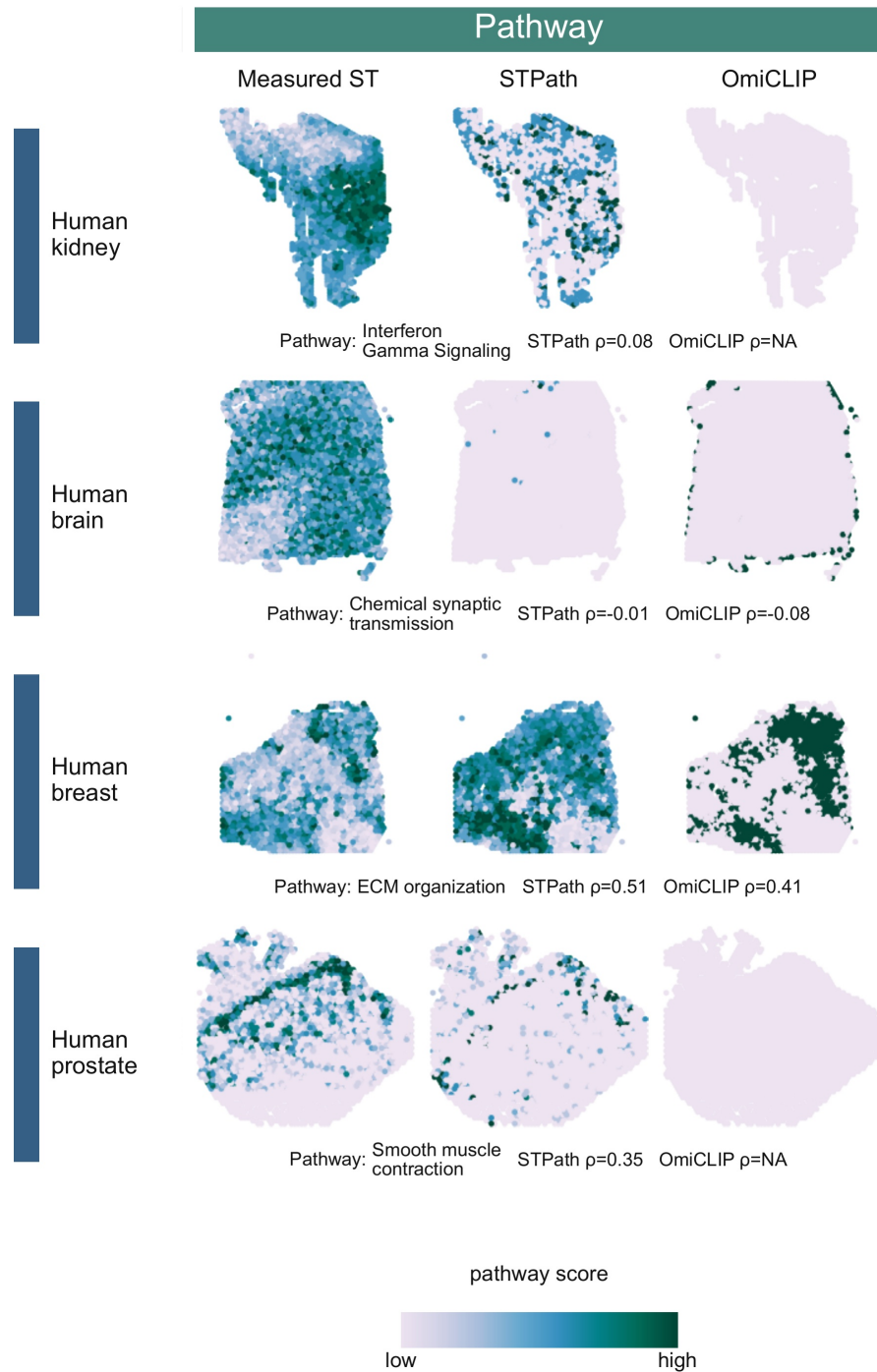


Figure S6: STPath and OmiCLIP spatial readouts for all organ cohorts in the **Figure 3 benchmark (continued)**. **f**, Pathway activity in human kidney, brain, breast and prostate.

g Mouse | Spatial organization

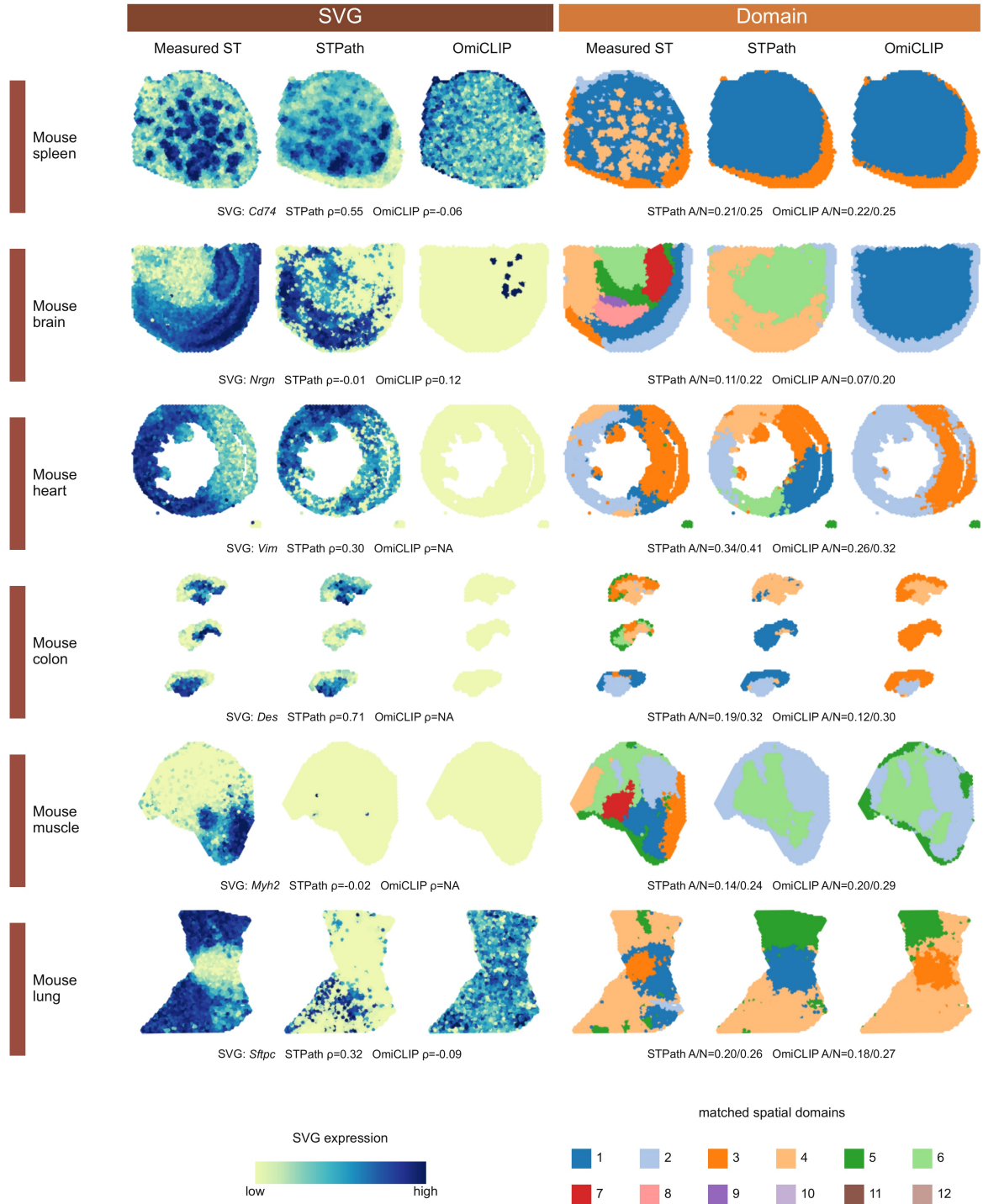


Figure S6: STPath and OmiCLIP spatial readouts for all organ cohorts in the Figure 3 benchmark (continued). g, Spatial organization in mouse spleen, brain, heart, colon, muscle and lung.

h Mouse | Cellular and molecular state

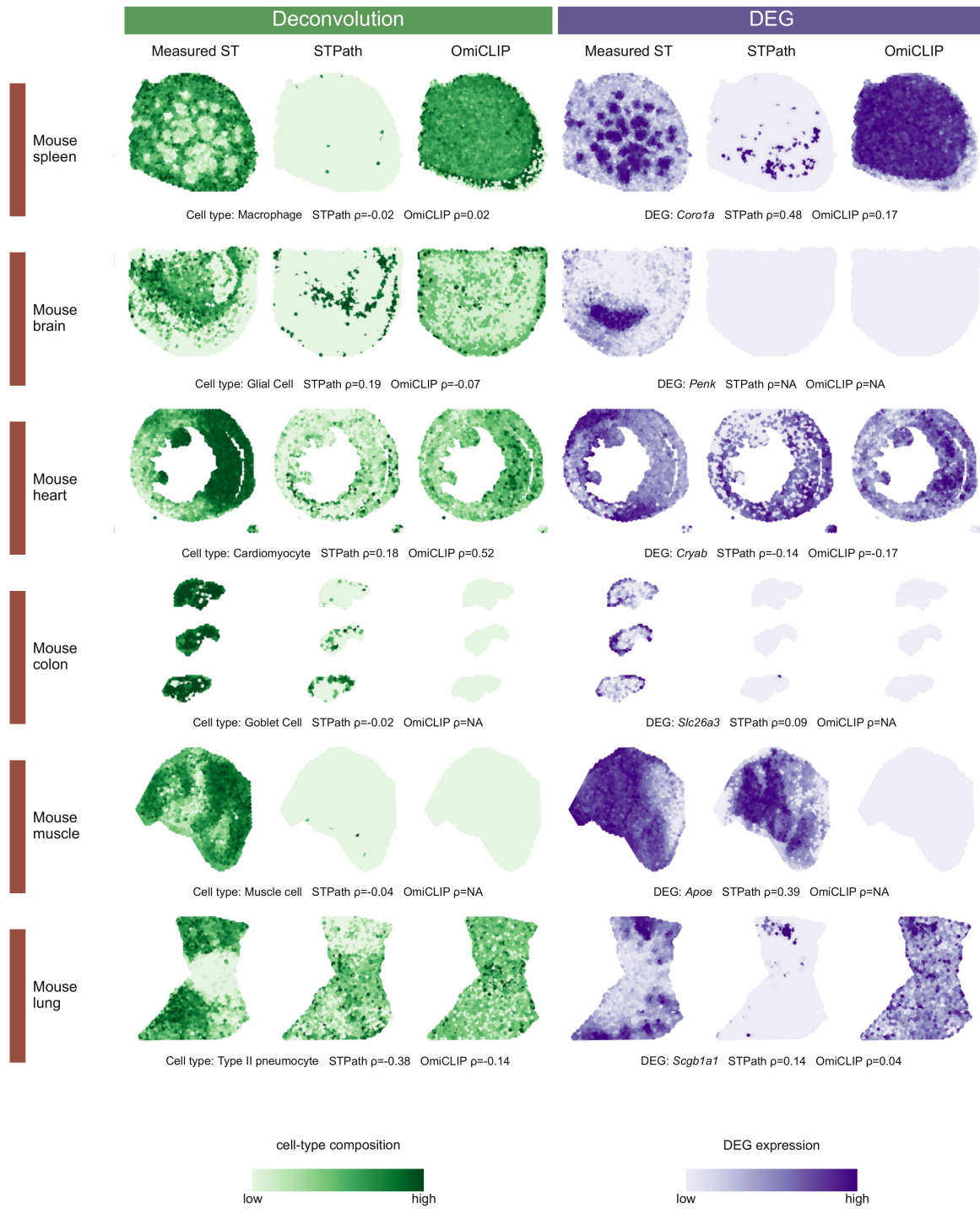


Figure S6: STPath and OmiCLIP spatial readouts for all organ cohorts in the Figure 3 benchmark (continued). h, Cellular and molecular state in mouse spleen, brain, heart, colon, muscle and lung.

i Mouse | Pathway activity

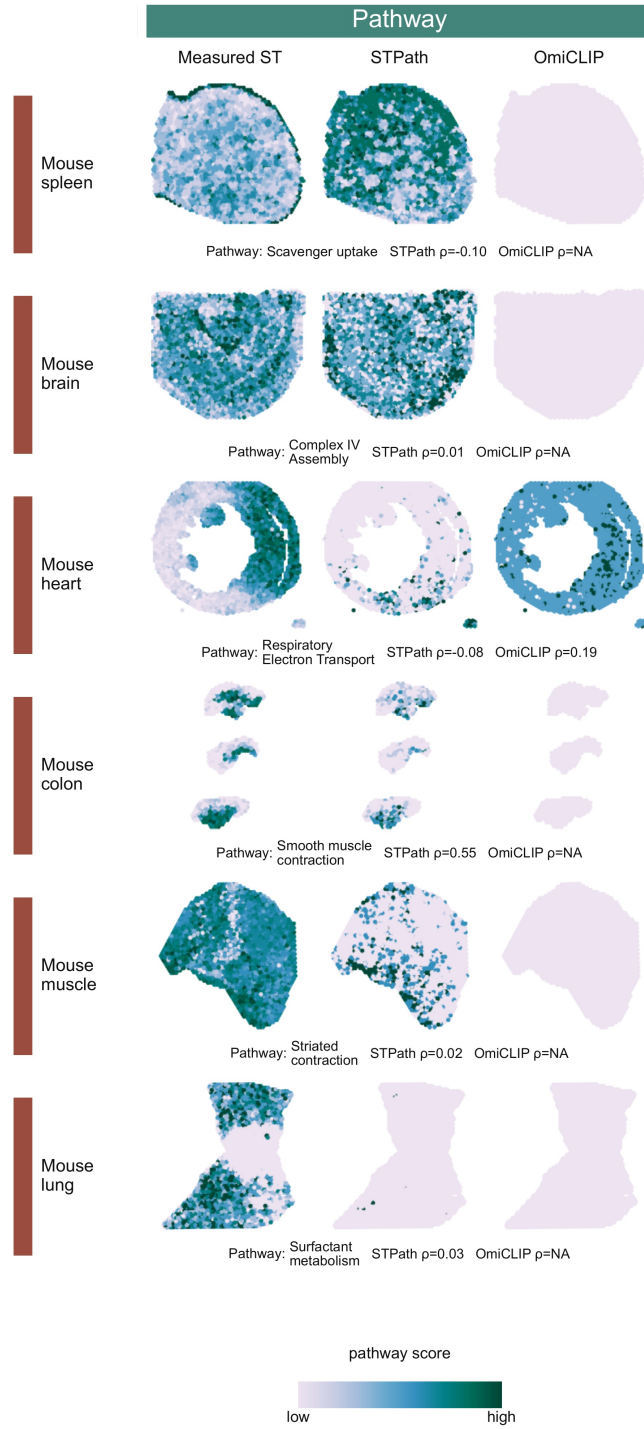


Figure S6: STPath and OmiCLIP spatial readouts for all organ cohorts in the **Figure 3 benchmark (continued)**. i, Pathway activity in mouse spleen, brain, heart, colon, muscle and lung.

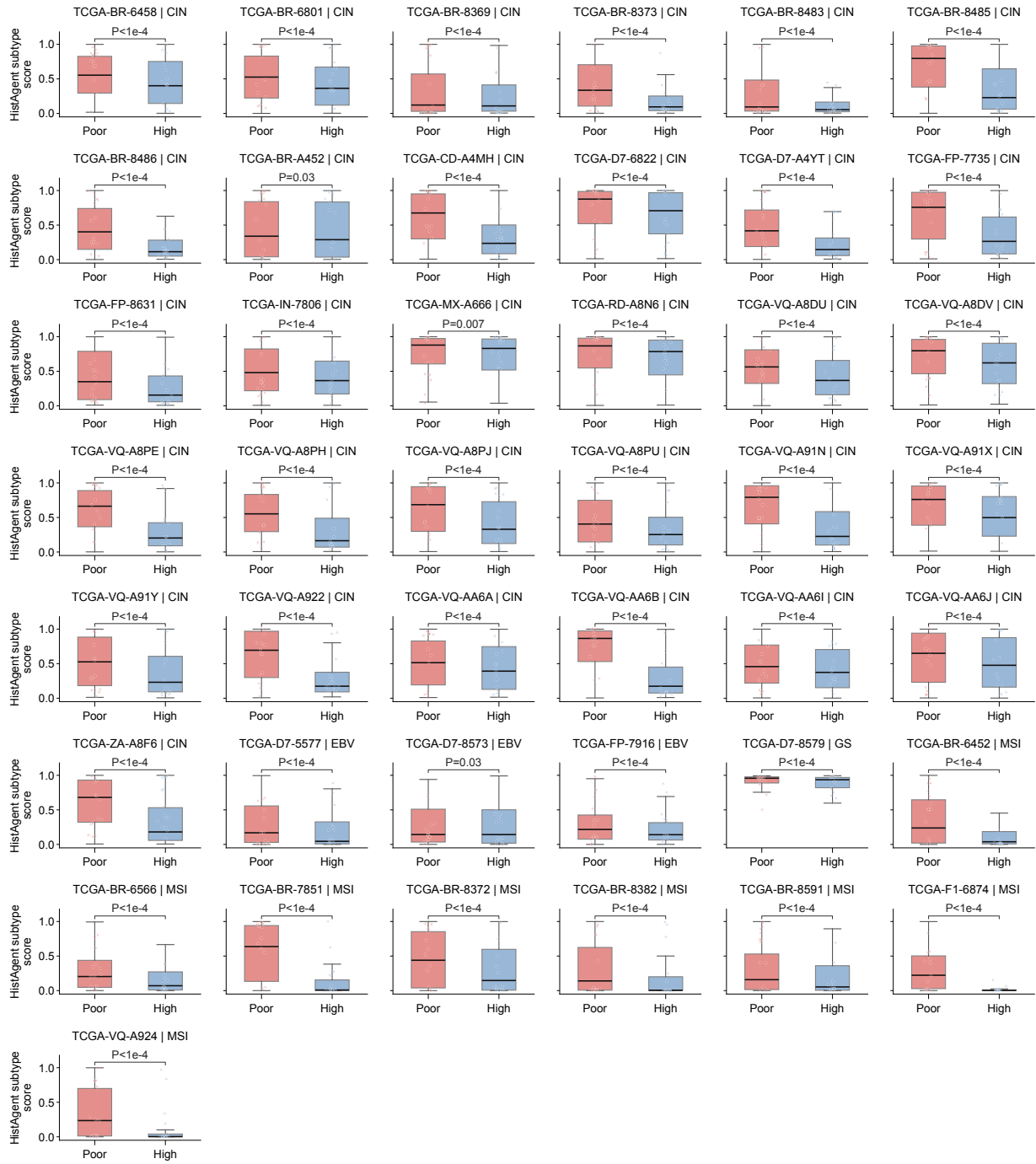


Figure S7: HistAgent GS subtype scores across annotated STAD differentiation regions. Tile-level genomically stable (GS) subtype scores compared between manually annotated poorly and highly differentiated regions in The Cancer Genome Atlas stomach adenocarcinoma (TCGA-STAD) slides. Scores were generated by out-of-fold models and mapped to their tissue coordinates. 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 are from separate one-sided Mann–Whitney U tests for each slide

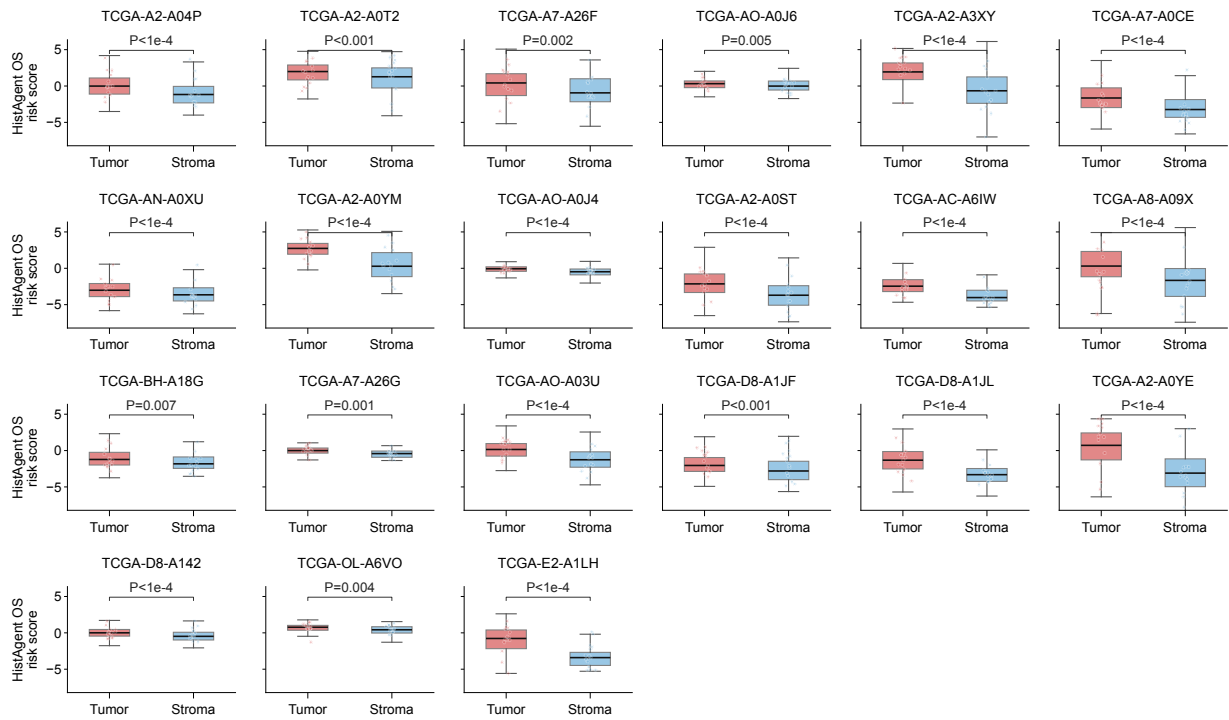


Figure S8: **HistAgent overall survival risk scores across annotated BRCA tissue regions.** Tile-level overall survival (OS) risk scores compared between tumor and stromal regions in The Cancer Genome Atlas breast invasive carcinoma (TCGA-BRCA) slides with Breast Cancer Semantic Segmentation (BCSS) annotations. Scores were generated by out-of-fold models and mapped to their tissue coordinates. 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 are from separate one-sided Mann–Whitney U tests for each slide.

Tumor epithelium 10 sections; 30 query-section units	Normal epithelium 3 sections; 9 query-section units
Query Retrieve breast cancer spatial spots annotated as malignant epithelial tumor regions or in situ carcinoma. Relevance Pathologist-annotated malignant epithelial tumor regions Source GSE213688 breast cancer	Query Retrieve breast tissue spots annotated as normal epithelium or benign epithelial glands. Relevance Pathologist-annotated normal epithelial regions Source GSE213688 breast cancer
Tumor pathology 7 sections; 21 query-section units	Cortical layers 36 sections; 72 query-section units
Query Retrieve malignant breast or prostate tumor regions, including invasive carcinoma and Gleason-grade cancer annotations. Relevance Tumor/invasive breast regions and malignant Gleason-grade prostate regions Source Human Breast 10X; Human Prostate Erickson	Query Retrieve human DLPFC spatial spots annotated as cortical layer 3, 4 or 5. Relevance Published cortical layer III, IV and V labels Source Human DLPFC Maynard
TLS niche 8 sections; 32 query-section units	Inclusion rule
Query Retrieve renal cell carcinoma spatial spots annotated as tertiary lymphoid structure niches. Relevance Published tertiary lymphoid structure (TLS) spots Source GSE175540 renal cell carcinoma	Each query group maps to one Figure 5f category. Query-section units with positive-label fraction $\geq 5\%$ were included. The source table stores query variants and source labels.

Figure S9: **Biological query groups used for natural-language retrieval evaluation.** The five query groups define the retrieval intent, expert-annotated relevant spots, source datasets, query counts, and evaluated slides used to compare HistAgent with Cell-Whisperer. Each evaluation unit comprises one natural-language query applied to one ST slide.

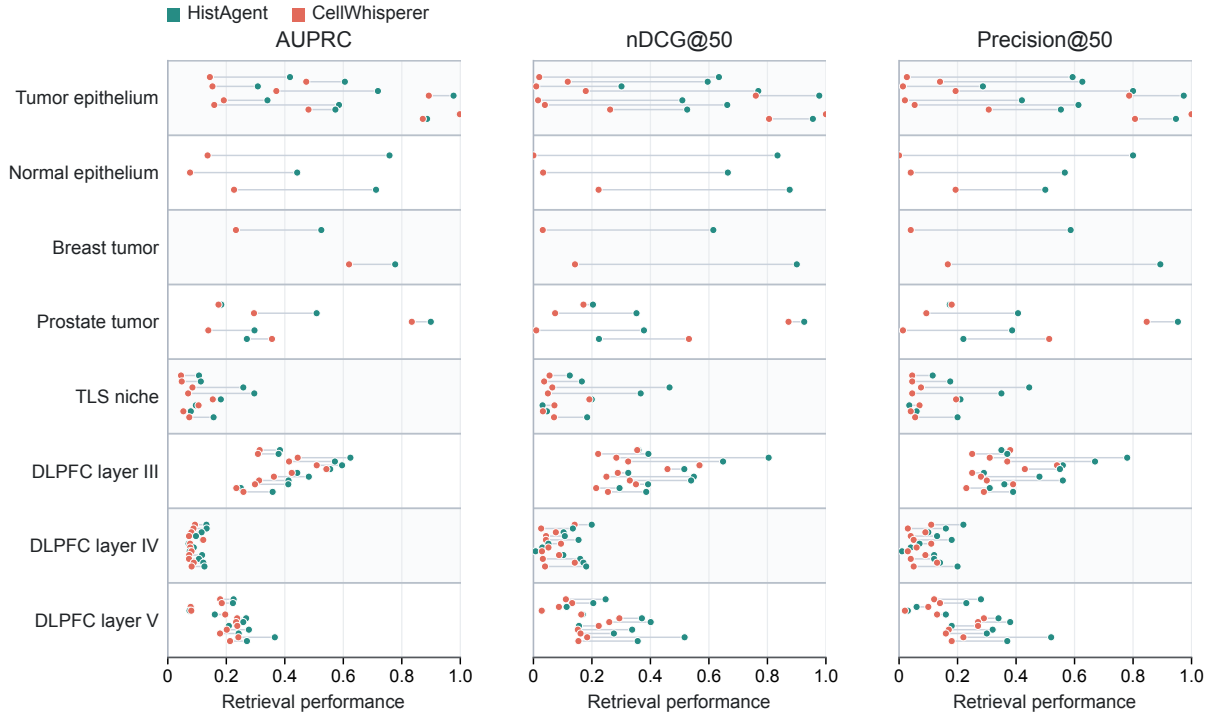


Figure S10: **Slide-level natural-language retrieval performance.** HistAgent and CellWhisperer compared for matched query–slide units using the area under the precision–recall curve (AUPRC), normalized discounted cumulative gain at 50 (nDCG@50), and Precision@50. nDCG@50 gives greater weight to relevant spots ranked near the top, whereas Precision@50 is the fraction of the top 50 spots annotated as relevant. Each point denotes one query–slide unit; linked points denote paired evaluations of the two methods, and higher values indicate better retrieval. DLPFC, dorsolateral prefrontal cortex.

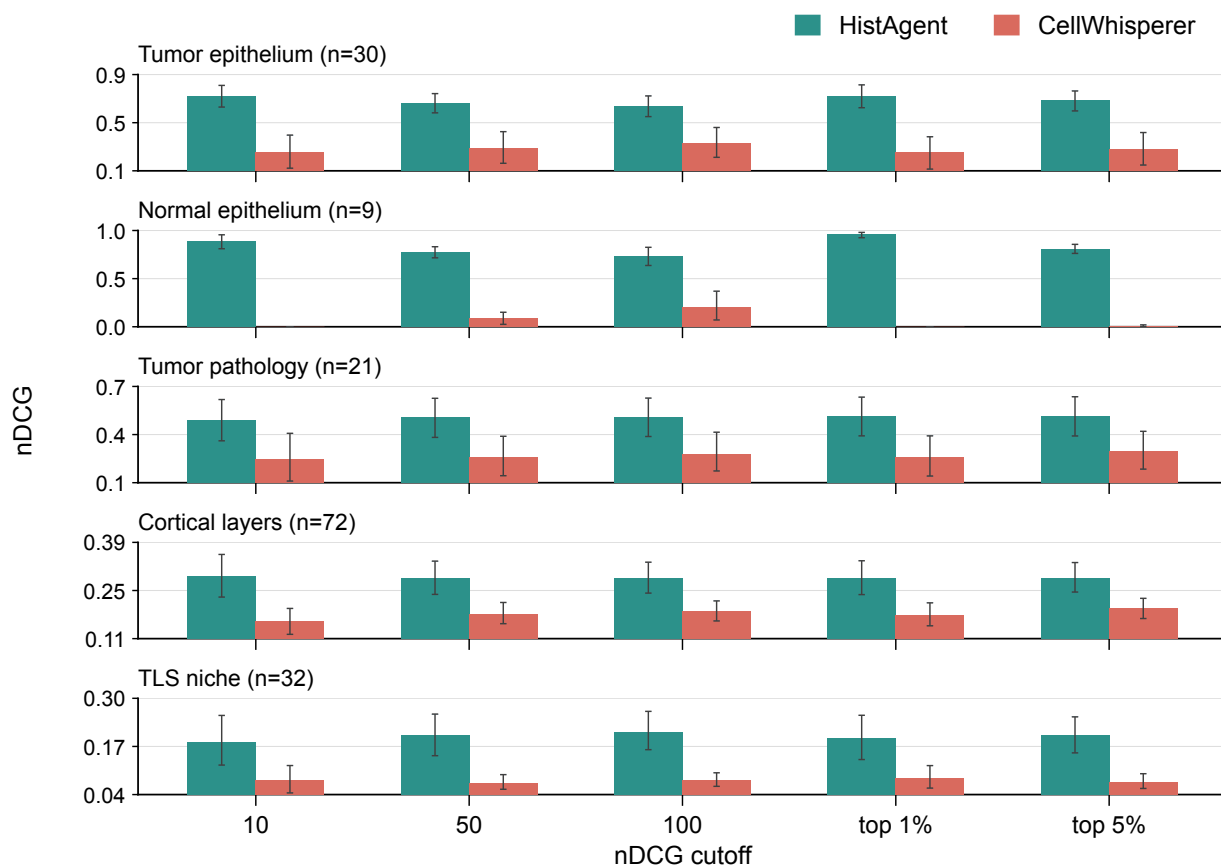


Figure S11: Natural-language retrieval performance across evaluation cutoffs. Normalized discounted cumulative gain (nDCG) is compared between HistAgent and CellWhisperer at fixed cutoffs of 10, 50, and 100 retrieved spots and at slide-normalized cutoffs comprising the top 1% and 5% of spots. Bars show means across matched query–slide units within each biological query group, and error bars denote 95% bootstrap confidence intervals. Higher values indicate that expert-annotated relevant spots occur nearer the top of the retrieved list.

Biological query group	Scenarios	Claim traceability (full evidence)	Unsupported gene mentions	Limited after evidence removal	Shifted after counterfactual evidence
Tumor epithelium	3	100%	0	3/3	3/3
Normal epithelium	3	100%	0	3/3	3/3
Tumor pathology	6	100%	0	6/6	6/6
Cortical layers	6	100%	0	6/6	6/6
TLS niches	4	100%	0	4/4	4/4
Overall	22	100%	0	22/22	22/22

Figure S12: **Evidence-grounding audit of the spatial agentic module.** The audit included 22 query scenarios across five biological query groups. Responses were evaluated using the original retrieved evidence cards, after removing the query-defining evidence and after replacing it with biologically incompatible evidence. Claim traceability measures whether supported claims cite the corresponding evidence-card fields, whereas the evidence-removal and counterfactual conditions assess whether the module limits or revises its interpretation when the supporting evidence changes.

Supplementary References

- [1] P.A.P. Moran, Notes on continuous stochastic phenomena. *Biometrika* **37**(1/2), 17–23 (1950)
- [2] A.Y. Ng, M.I. Jordan, Y. Weiss, *On Spectral Clustering: Analysis and an Algorithm*, in *Advances in Neural Information Processing Systems*, vol. 14 (2001), pp. 849–856
- [3] M. Amgad, H. Elfandy, H. Hussein, et al., Structured crowdsourcing enables convolutional segmentation of histology images. *Bioinformatics* **35**(18), 3461–3467 (2019)
- [4] E.L. Kaplan, P. Meier, Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association* **53**(282), 457–481 (1958)
- [5] N. Mantel, Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemotherapy Reports* **50**(3), 163–170 (1966)
- [6] P.M. Grambsch, T.M. Therneau, Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika* **81**(3), 515–526 (1994)
- [7] 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)
- [8] 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)
- [9] 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)
- [10] 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)
- [11] 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)

- [12] 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)
- [13] 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)
- [14] 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)
- [15] 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)
- [16] K. Järvelin, J. Kekäläinen, Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems* **20**(4), 422–446 (2002)
- [17] A.P. Heath, V. Ferretti, S. Agrawal, et al., The NCI genomic data commons. *Nature Genetics* **53**(3), 257–262 (2021)
- [18] K. Clark, B. Vendt, K. Smith, J. Freymann, J. Kirby, P. Koppel, S. Moore, S. Phillips, D. Maffitt, M. Pringle, L. Tarbox, F. Prior, The cancer imaging archive (TCIA): Maintaining and operating a public information repository. *Journal of Digital Imaging* **26**(6), 1045–1057 (2013)
- [19] J. Liu, T. Lichtenberg, K.A. Hoadley, L.M. Poisson, A.J. Lazar, A.D. Cherniack, et al., An integrated TCGA pan-cancer clinical data resource to drive high-quality survival outcome analytics. *Cell* **173**(2), 400–416.e11 (2018)