# ProCLIP: Contrastive Multimodal Prediction of Spatial Proteomics from H&E and Spatial Transcriptomics

Zhaoqi Song 1, Lecheng Li 1, Yuxiang Lin 4, Zixiang Wang 1, Huimin Zhang 2 3#, Mengsha Tong1 5#, Chaoyong Yang 2 4#

1 State Key Laboratory of Cellular Stress Biology, School of Life Sciences, Faculty of Medicine and Life Sciences, Xiamen University, Xiamen, 361102, China

2 The MOE Key Laboratory of Spectrochemical Analysis & Instrumentation, State Key Laboratory of Physical Chemical of Solid Surfaces, Department of Chemical Biology, College of Chemistry and Chemical Engineering, Xiamen University, Xiamen, China

3 State Key Laboratory of Vaccines for Infectious Diseases, Fujian Provincial Key Laboratory of Innovative Drug Target Research, School of Pharmaceutical Sciences, School of Life Sciences, Faculty of Medicine and Life Sciences, Xiamen University, Xiamen 361005, China

4 Institute of Molecular Medicine, Shanghai Key Laboratory for Nucleic Acid Chemistry and Nanomedicine, Department of Urology, Renji Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai, China

5 National Institute for Data Science in Health and Medicine, Xiamen University, Xiamen, 361102, Fujian, China

## Abstract

Spatial proteomics resolves functional cell states and tissue niches, but high-plex imaging remains costly and limited to predefined antibody panels. Histopathology and spatial transcriptomics provide complementary morphological and molecular information, but neither directly reveals unmeasured proteins. We developed ProCLIP, a protein-conditioned framework for virtual spatial proteomics. A dual-encoder architecture integrates H&E morphology, spatial transcriptomic features and coordinates while representing each queried protein through its identity and biological annotations. A CLIP-style objective aligns tissue and protein representations, and feature-wise linear modulation conditions a shared regressor on proteins excluded from abundance supervision. Across six spatial-omics datasets, ProCLIP achieved the highest mean protein-wise Pearson correlation in five comparisons and closely matched the leading method in the sixth. In a held-out HCC sample, ProCLIP-derived profiles improved recovery of CODEX-defined cancer-associated fibroblast spatial subtypes compared with raw spatial transcriptomic features. These results support protein-conditioned prediction as a strategy for generating biologically informative estimates when direct multiplexed measurements are unavailable.

## Introduction

Tumours are spatially organized ecosystems in which malignant, stromal and immune cells interact through tissue architecture, extracellular matrix and vascular niches [1], [8], [12]. Spatial transcriptomics measures broad RNA programmes in tissue context [19], whereas multiplexed imaging measures functional proteins directly in situ [4], [5], [10]. RNA captures a wide molecular repertoire, while proteins more directly report cellular phenotype and signalling state. However, high-plex protein imaging remains costly, throughput-limited and constrained by antibody panels selected before all relevant biological questions are known.

Virtual spatial proteomics could extend routinely acquired H&E and spatial transcriptomic data by computationally estimating protein maps. This task is difficult because RNA abundance is an incomplete proxy for protein abundance. Post-transcriptional regulation, protein turnover and local microenvironmental effects can decouple the two measurements. Histology provides orthogonal evidence about tissue organization, cell morphology and stromal context, but cannot identify every protein from morphology alone. A useful model must therefore integrate complementary tissue measurements and support queries for proteins excluded from abundance supervision.

Existing studies establish important components of this problem but do not resolve this queryability gap. Histology-to-omics methods show that H&E images contain predictive molecular information [6], [14], and HEX predicts a predefined protein panel from pathology features [9]. SpatialEx and SpatialEx+ use histology to align spatial-omics sections [11], while RNA–protein models learn paired representations from assays such as CITE-seq [3]. These approaches either bind prediction to a fixed output panel or address a different cross-modality alignment problem. They therefore do not directly support protein-held-out prediction from multimodal tissue context.

We address this limitation by treating protein identity as an explicit conditioning variable rather than a fixed output index. ProCLIP couples a tissue encoder for H&E morphology, spatial transcriptomics, coordinates, cell identity and neighbourhood composition with a protein encoder for marker identity and biological descriptors. A contrastive objective aligns tissue and protein representations, and feature-wise linear modulation conditions a shared regressor on the queried protein [15]. This design separates a protein query from its measured abundance, allowing one model to predict supervised and protein-held-out markers.

This study makes four contributions:

- We formulate virtual spatial proteomics as protein-conditioned regression and develop a dual-encoder architecture for supervised and protein-held-out markers.
- We evaluate the formulation on six tumour spatial-omics datasets, including spot-level protein-held-out prediction and a patient-held-out single-cell cohort [17], [18], [21], [22].
- We compare ProCLIP with dataset-appropriate histology, foundation-model, graph and RNA-to-protein baselines.
- We test whether virtual protein profiles preserve biological organization by evaluating recovery of spatial CAF subtypes beyond marker-wise correlation.

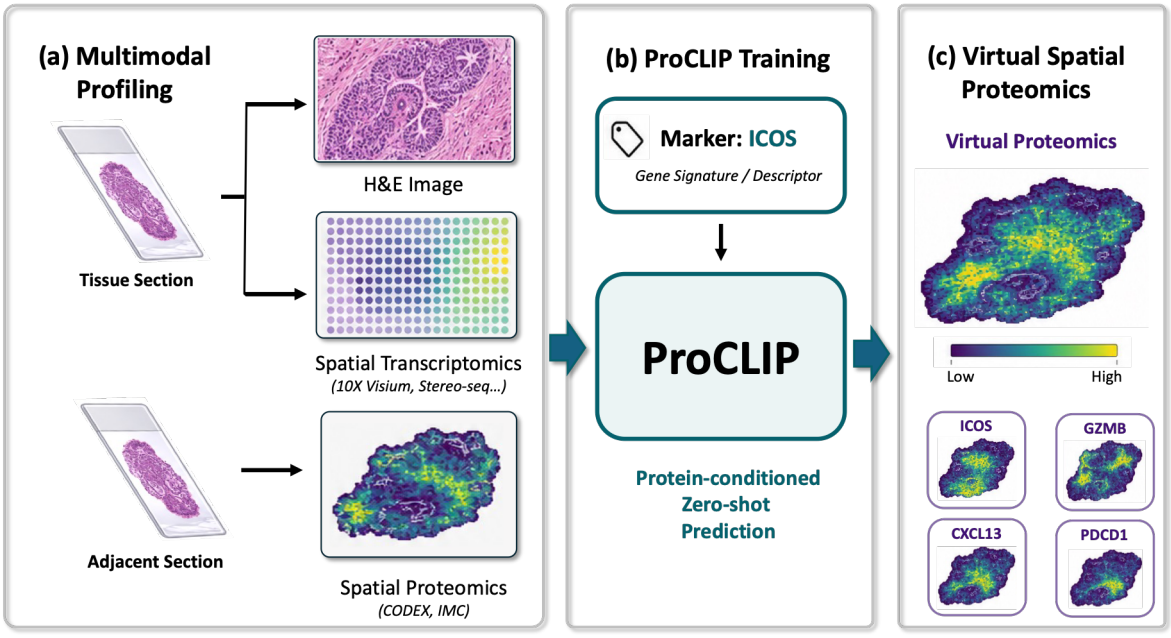



**Figure 1 | Overview of ProCLIP.** ProCLIP generates queryable virtual protein maps from multimodal tissue context. Tissue-side morphology, transcriptomics, coordinates, cell identity and neighbourhood features are paired with protein-side identities and biological descriptors.

# Related Work

### Spatial molecular profiling

Multiplexed imaging technologies quantify tissue proteins through metal-tagged antibodies, DNA-barcoded cyclic imaging or repeated immunofluorescence [4], [5], [8], [10]. These measurements reveal cell phenotypes and neighbourhoods, but coverage is constrained by antibody panels, cost and throughput. Spatial transcriptomics provides broader molecular coverage [19], while histology-to-gene models infer transcriptomic programmes from H&E morphology [6], [14]. Neither approach directly supports queries for an unmeasured protein from multimodal tissue context.

### Prediction of spatial protein profiles from H&E images

HEX predicts a fixed panel of protein biomarkers from H&E images using pathology foundation-model features [9]. SpatialEx uses histology to anchor high-parameter spatial-omics integration, and SpatialEx+ extends this framework across adjacent sections [11]. ProCLIP addresses a related but distinct setting by combining morphology, spatial transcriptomics and neighbourhood context while conditioning prediction on the queried protein. We use protocol-specific comparisons because baseline inputs and target definitions differ across datasets.

### RNA-to-protein prediction and conditional representation learning

Joint single-cell models such as totalVI learn RNA–protein representations from paired CITE-seq measurements [3]. Their output proteins are generally observed during training, and these models do not use tissue morphology or spatial neighbourhoods. ProCLIP instead combines a CLIP-style dual-encoder formulation [16] with FiLM conditioning [15]. Protein metadata define the query, while the shared tissue representation provides spatial evidence for prediction.

# Results

## ProCLIP learns protein-conditioned representations of multimodal tissue context

ProCLIP is a protein-conditioned framework for generating queryable virtual spatial protein profiles from multimodal tissue measurements. We trained the model on tumour datasets containing spatial transcriptomic and proteomic profiles, together with histopathology where available. The dual-encoder architecture processes tissue context and protein attributes separately, then combines them through contrastive alignment and conditional regression (Figs. 1 and 2; Methods).

The tissue encoder integrates three complementary sources of information. H&E patches are processed with the pretrained MUSK pathology model to capture cellular morphology and local tissue architecture. Spatial transcriptomic features represent molecular state, while coordinates and neighbourhood summaries represent local spatial context. A multimodal encoder fuses these inputs into one representation for each cell or spot.

In parallel, the protein encoder constructs a representation for each queried protein. Inputs include the harmonized protein identity, associated gene information, Gene Ontology annotations, functional descriptors and optional protein-associated imaging features. The resulting representation encodes the molecular identity and annotated function of the protein.

ProCLIP projects the tissue and protein embeddings into a shared latent space and optimizes a CLIP-style contrastive objective. The objective assigns greater compatibility to spatial units with stronger measured expression of a supervised protein. This alignment organizes multimodal tissue states in relation to protein identities.

The protein representation then conditions abundance prediction through feature-wise linear modulation. A FiLM generator converts the protein embedding into protein-specific scale and shift parameters that modulate the tissue embedding before a shared readout. The model therefore predicts abundance for a cell–protein pair rather than producing all proteins through fixed output nodes.

This formulation supports prediction for supervised proteins and proteins excluded from abundance supervision. We assessed whether ProCLIP recovered measured protein distributions and preserved biologically relevant tissue organization. The latter analysis focused on the spatial subtype structure of cancer-associated fibroblasts (CAFs).

## ProCLIP predicts proteins excluded from abundance supervision

Spatially resolved proteins define cellular phenotypes, functional states and tissue niches, but direct measurement remains costly and technically demanding. CODEX, imaging mass cytometry and multiplexed immunofluorescence quantify proteins in situ, but their coverage is limited by antibody availability, panel compatibility and throughput. We therefore tested whether ProCLIP could infer proteins that were excluded from supervised prediction during model training.

For each dataset, we divided the measured protein panel into supervised proteins and held-out query proteins. Only supervised protein abundances contributed to the regression objective. At inference, each held-out protein was specified by its identity, associated gene information, Gene Ontology annotations and functional descriptors.

Here, zero-shot denotes the absence of abundance supervision for the queried protein. The protein name and externally defined annotations remain available to the encoder, but the protein's measured spatial abundance does not contribute to model fitting. The task therefore assesses generalization to a protein-held-out target, not to a protein lacking all biological information.

We evaluated this task across six tumour spatial-omics datasets spanning hepatocellular carcinoma (HCC), colon adenocarcinoma (COAD), ovarian cancer (OV), glioma and colorectal cancer liver metastasis. Five datasets assessed within-dataset prediction of held-out proteins. In HCC-Wu, patient 84 and its held-out proteins were reserved for testing, providing a stricter patient- and protein-held-out evaluation. The primary metric was protein-wise Pearson correlation coefficient (PCC), supplemented by Spearman correlation, root mean squared error (RMSE) and spatial maps.

ProCLIP achieved the highest mean protein-wise PCC in five of the six dataset-specific comparisons against MUSK, SpaTranslator, HEX and STProtein. Mean PCC values were 0.916 for glioma, 0.870 for COAD, 0.836 for OV, 0.808 for CRC-Mo and 0.730 for HCC-Wu. In HCC, ProCLIP achieved a PCC of 0.929, compared with 0.935 for HEX. These results show that protein-conditioned multimodal prediction recovered held-out protein variation across the evaluated datasets.

We next examined representative held-out protein maps to determine whether predictions preserved spatial organization. Marker-level PCC values were 0.95 for ICOS in glioma, 0.95 for MPO in HCC and 0.88 for FOXP3 in COAD. The corresponding values were 0.93 for FOXP3 in OV, 0.92 for E-cadherin in CRC-Mo and 0.61 for CXCL13 in HCC-Wu. Predicted maps retained broad gradients and contiguous expression domains, although fine-scale discrepancies remained in heterogeneous regions.

This capability follows from the protein-conditioned formulation. Conventional fixed-output models associate each marker with a supervised output node and cannot directly query an untrained node. ProCLIP instead uses the queried protein embedding to generate FiLM parameters for a shared tissue representation. The same regression function can therefore be applied to supervised and protein-held-out targets.

Together, these results support computational estimation of proteins excluded from abundance supervision. The predictions remain model-based estimates and should not replace direct measurements,

especially when a protein has weak transcriptomic or morphological correlates. Additional proteins lacking matched measurements require independent experimental validation before their predicted maps can be interpreted as panel expansion.

## ProCLIP profiles improve recovery of spatial CAF subtypes

Cancer-associated fibroblast states reflect both intrinsic transcriptional programmes and spatial relationships with malignant, immune and stromal compartments. Distinct CAF populations occupy characteristic niches that can be defined by neighbouring cell composition. Transcriptomic measurements alone may incompletely resolve these states when their distinction depends on protein phenotypes or local organization. We therefore tested whether ProCLIP-derived profiles preserved enough spatial information to recover CODEX-defined CAF subtypes.

We analysed 5,844 CAFs from held-out HCC-Wu patient 84, for which spatial transcriptomic, histological and CODEX-derived reference information was available. Following Liu et al., we assigned each CAF to one of four subtypes according to neighbouring cell composition. The subtypes were tumour-adjacent, myeloid-associated, lymphoid-associated and stromal CAFs (Methods). These assignments served as reference labels for comparing molecular representations.

We compared measured CODEX protein profiles, ProCLIP-predicted virtual protein profiles and raw spatial transcriptomic features using the same classification procedure. Balanced accuracy was the primary metric because subtype frequencies differed.

Classifiers using measured CODEX proteins, ProCLIP profiles and raw spatial transcriptomic features achieved balanced accuracies of 0.804, 0.713 and 0.640, respectively (Fig. Xa). ProCLIP therefore improved balanced accuracy by 7.3 percentage points relative to raw transcriptomic features, but remained below directly measured proteins. This intermediate performance is consistent with virtual proteins recovering only part of the structure captured by direct measurements.

The improvement varied across CAF states. Compared with raw spatial transcriptomic features, ProCLIP increased recall from 0.75 to 0.81 for tumour-adjacent CAFs and from 0.72 to 0.77 for stromal CAFs. Recall for lymphoid-associated CAFs increased from 0.58 to 0.77. By contrast, recall for myeloid-associated CAFs was similar for transcriptomic and ProCLIP features (0.52 and 0.51, respectively; Fig. Xb).

These results indicate that ProCLIP profiles better captured the reference CAF labels than raw transcriptomic features. The largest gain occurred for lymphoid-associated CAFs. However, because the labels were defined from CODEX neighbourhood composition, this result requires controls excluding overlapping neighbourhood or cell-type information from the ProCLIP inputs.

## Ablation analysis of ProCLIP

We evaluated five ablated variants on the OV protein-held-out task. The A0 model achieved a mean protein-wise PCC of $0.836 \pm 0.074$ (Table 2).

*FiLM.* Removing FiLM changed PCC to $0.727 \pm 0.004$ ($\Delta\text{PCC} = -0.109$).

*Spatial context.* Removing spatial-context features changed PCC to $0.749 \pm 0.017$ ($\Delta\text{PCC} = -0.087$).

*CLIP-style alignment.* Removing the contrastive objective changed PCC to $0.798 \pm 0.019$ ($\Delta\text{PCC} = -0.038$).

*H&E-derived MUSK features.* Removing morphology features changed PCC to $0.797 \pm 0.026$ ($\Delta\text{PCC} = -0.039$).

*Residual prior.* Removing the residual-prior component changed PCC to $0.804 \pm 0.010$ ($\Delta\text{PCC} = -0.032$).

**Table 2 | Ablation of ProCLIP components on OV. A0 reports mean** $\pm$ sample standard deviation across four held-out proteins from one run, whereas A1–A5 report mean $\pm$ sample standard deviation across three seeds. $\Delta PCC$ denotes the change from A0.

| Variant | Seeds | Test PCC | ΔPCC |
|---|---|---|---|
| A0 Full | 1 | $\mathbf{0.836 \pm 0.074}$ | 0 |
| A1 w/o CLIP | 3 | $0.798 \pm 0.019$ | $-0.038$ |
| A2 w/o FiLM | 3 | $0.727 \pm 0.004$ | $-0.109$ |
| A3 w/o H&E/MUSK | 3 | $0.797 \pm 0.026$ | $-0.039$ |
| A4 w/o spatial context | 3 | $0.749 \pm 0.017$ | $-0.087$ |
| A5 w/o residual prior | 3 | $0.804 \pm 0.010$ | $-0.032$ |

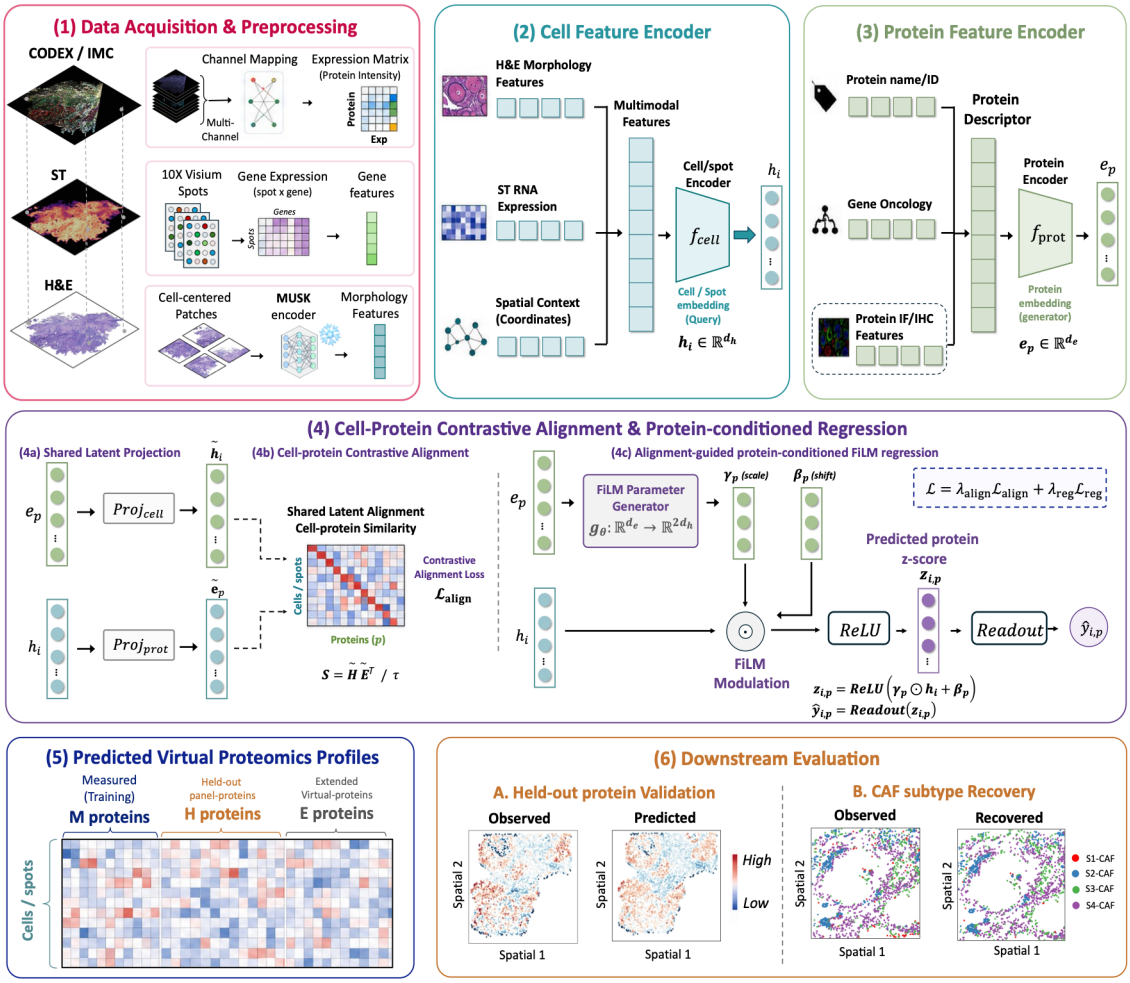



**Figure 2 | ProCLIP framework.** A tissue encoder and a protein encoder generate paired representations. Contrastive alignment organizes the shared latent space, while protein-conditioned FiLM regression predicts abundance for supervised and held-out protein queries.

# Methods

## Spatial multi-omics data collection

We collected six public human tumour datasets containing spatial transcriptomic and multiplexed protein measurements, together with matched histology where available. The datasets covered Stereo-seq, 10x Genomics Visium and Visium HD transcriptomics, and CODEX, MIBI or multiplexed immunofluorescence proteomics. Because spatial resolution and tissue-section configuration differed among cohorts, each dataset was processed separately. Processing comprised data selection, modality-specific standardization, protein nomenclature harmonization, quality control, cross-modal registration and integrated feature construction.

*Data acquisition and selection.* We included tumour datasets that met four criteria. First, spatial transcriptomic and multiplexed protein measurements originated from the same or adjacent tissue sections. Second, spatial coordinates or registration information enabled cross-modal correspondence. Third, protein channels had identifiable marker annotations. Fourth, an H&E or other histological reference image was available when morphology features were required. We excluded datasets with

inadequate spatial correspondence, uninterpretable protein channels or insufficient spatial units after quality control.

We evaluated ProCLIP on six public tumour spatial-omics datasets spanning HCC, COAD, OV, glioma and colorectal cancer liver metastasis.

*SPATCH datasets.* We downloaded paired HCC, COAD and OV datasets from SPATCH (https://spatch.pku-genomics.org/) [17], [18]. Each cohort combined high-resolution Stereo-seq v1.3 transcriptomics with adjacent 16-plex CODEX imaging. HCC contained 390,774 spatially resolved cells and 30,493 profiled genes. OV contained 319,366 cells and 32,008 genes, while COAD contained 361,260 cells and 31,718 genes.

*Glioma dataset.* The current source text describes the metastatic colorectal cancer dataset reported by Mo et al. [21]. It contains one colorectal cancer liver metastasis sample with 4,719 tissue-covered Visium spots, 36,945 genes, matched H&E and a 47-marker MIBI panel.

*CRC-Mo dataset.* The current source text describes the IDH-mutant astrocytoma dataset reported by Tang et al. [13]. We used the initial tumour sample from patient P174511, profiled with 10x Genomics Visium HD at 16 × 16 μm resolution. The sample contains 87,988 bins, 18,085 genes, matched H&E and a 48-channel CODEX image.

*HCC-Wu dataset.* This cohort contains sequential H&E, 10x Visium transcriptomic and 51-channel immunofluorescence measurements from 14 HCC patients sampled around immunotherapy [22]. After preprocessing, we retained 50 protein markers and cell-level annotations. Patient 84 was reserved for testing and patient 86 for validation in the patient- and protein-held-out experiment.

## Multimodal data preprocessing

*Spatial transcriptomic preprocessing.* We processed count matrices separately to accommodate platform, resolution and molecular-coverage differences. We removed spatial units lacking valid coordinates or adequate transcriptomic measurements and resolved duplicated gene symbols. Counts were library-size normalized and log transformed. Depending on the dataset, representations comprised highly variable genes, reduced-dimensional features or predefined gene-program scores. We removed features with zero variance, excessive missingness or inadequate coverage.

*Histology preprocessing.* We mapped H&E images to the corresponding transcriptomic coordinates using platform-provided scale factors or spatial transformations. A patch centred on each cell, bin or spot was processed with the pretrained MUSK pathology model. The resulting embeddings represented local morphology, tissue architecture and stromal organization. We excluded patches outside valid tissue or image boundaries, removed non-finite features and standardized the remaining embeddings.

*Spatial proteomic preprocessing.* We processed CODEX, MIBI, IMC and multiplexed immunofluorescence measurements according to each dataset. Channel labels were mapped to canonical protein names and corresponding gene symbols. We removed platform-specific suffixes, acquisition labels and non-biological channels. Only markers with valid identities and complete spatial measurements were retained. Transformed and standardized protein abundances served as regression targets, not tissue-encoder inputs.

## Cross-modal alignment and model inputs

*Cross-modal spatial alignment.* H&E served as the common spatial reference for histology, transcriptomics and proteomics. Transcriptomic units were linked to H&E using platform coordinates and image scale factors. For proteomics, the nuclear DAPI channel was registered to H&E, and the resulting transformation was applied to protein images and coordinates. We then matched

transcriptomic units to proteomic locations and extracted an H&E-centred patch for each retained unit. All modality matrices were reordered to a common spatial index. For adjacent or sequential sections, matched observations represent registered spatial correspondences rather than identical physical cells.

*Construction of tissue-level inputs.* For each spatial unit i, we combined transcriptomic features, a MUSK-derived H&E embedding and standardized coordinates with local context. We constructed a $k$-nearest-neighbour graph and summarized the molecular state or cellular composition of neighbouring units. Reliable cell-type or compartment annotations were optionally added. These components formed the tissue input:

$$x_i = \left[x_i^{\mathrm{RNA}}, x_i^{\mathrm{H\&E}}, x_i^{\mathrm{coord}}, x_i^{\mathrm{neigh}}, x_i^{\mathrm{celltype}}\right],$$

The tissue encoder transformed this input into representation $h_i$.

*Construction of protein-level inputs and targets.* Because panels differed among datasets, each protein was represented as an independent query. For protein p, the descriptor comprised its harmonized identity, gene information, signatures, Gene Ontology annotations and functional descriptors. Optional protein-associated imaging features were added where available. These attributes formed the protein input

$$a_p = \left[a_p^{\mathrm{ID}}, a_p^{\mathrm{gene}}, a_p^{\mathrm{GO}}, a_p^{\mathrm{function}}, a_p^{\mathrm{image}}\right],$$

The protein encoder transformed this input into representation $e_p$. Standardized protein abundance $y_{i,p}$ was stored separately as the regression target. Held-out protein measurements were excluded from model optimization and used only for evaluation. Target abundance was absent from both the tissue and protein descriptors.

# ProCLIP model architecture

ProCLIP uses a dual-encoder architecture for protein-conditioned prediction from multimodal tissue information (Fig. 2). The tissue encoder integrates H&E morphology, spatial transcriptomic features and spatial context for each cell or spot. The protein encoder represents each query using its identity, biological annotations and optional protein-associated imaging features. A contrastive objective aligns the two representations, after which the protein embedding conditions a shared FiLM regressor.

***Tissue encoder.*** For spatial unit $i$, the input comprised three feature groups:

$$x_i = \left[x_i^{\mathrm{H\&E}}, x_i^{\mathrm{RNA}}, x_i^{\mathrm{spatial}}\right],\ i = 1, \ldots, N, \qquad (1)$$

Here, $x_i^{\mathrm{H\&E}}$ denotes the MUSK embedding of the centred H&E patch, $x_i^{\mathrm{RNA}}$ denotes processed transcriptomic features and $x_i^{\mathrm{spatial}}$ denotes standardized coordinates.

The three feature groups are concatenated into a multimodal tissue descriptor:

$$m_i = \mathrm{Concat}\left(x_i^{\mathrm{H\&E}}, x_i^{\mathrm{RNA}}, x_i^{\mathrm{spatial}}\right). \qquad (2)$$

A tissue encoder $f_{\mathrm{cell}}$ transformed the multimodal descriptor into a compact representation:

$$h_i = f_{\mathrm{cell}}(m_i),\ h_i \in \mathbb{R}^{d_h}, \qquad (3)$$

Here, $d_h$ denotes the tissue-embedding dimension. The resulting $h_i$ summarizes the morphological, transcriptional and spatial context of unit $i$.

***Protein encoder.*** For each queried protein $p$, ProCLIP constructed a descriptor from three sources:

$$a_p = \left[a_p^{\mathrm{ID}}, a_p^{\mathrm{GO}}, a_p^{\mathrm{image}}\right],\ p = 1, \ldots, P, \qquad (4)$$

Here, $a_p^{\mathrm{ID}}$ encodes the harmonized protein identity, $a_p^{\mathrm{GO}}$ contains Gene Ontology annotations and $a_p^{\mathrm{image}}$ denotes optional protein-associated imaging features.

These features are concatenated into a protein descriptor and processed by the protein encoder:

$$e_p = f_{\mathrm{prot}}(a_p),\ e_p \in \mathbb{R}^{d_e}, \qquad (5)$$

Here, $d_e$ is the protein-embedding dimension. Protein abundance is excluded from $a_p$ and retained only as a regression target. A protein can therefore be queried without supplying its measured abundance to the encoder.

***Shared latent projection.*** Separate projection heads map the tissue and protein embeddings into a common latent space:

$$v_i = \mathrm{Proj}_{\mathrm{cell}}(h_i),\ u_p = \mathrm{Proj}_{\mathrm{prot}}(e_p). \qquad (6)$$

The projected representations are L2 normalized:

$$\tilde{h}_i = \frac{v_i}{\| v_i \|_2},\ \tilde{e}_p = \frac{u_p}{\| u_p \|_2}, \qquad (7)$$

where

$$\tilde{h}_i, \tilde{e}_p \in \mathbb{R}^{d_c}.$$

Here, $\mathrm{d_c}$ denotes the dimension of the shared contrastive space.

***Cell–protein contrastive alignment.*** Compatibility between spatial unit i and protein p was calculated using temperature-scaled cosine similarity:

$$S_{i,p} = \frac{\tilde{h}_i^T \tilde{e}_p}{\tau}, \qquad (8)$$

Here, $\tau > 0$ is the contrastive temperature. For a batch of $B$ spatial units and $P_{\mathrm{train}}$ supervised proteins, the similarity matrix is

$$S = \frac{\tilde{H}\tilde{E}^T}{\tau},\ S \in \mathbb{R}^{B \times P_{\mathrm{train}}}. \qquad (9)$$

Cell–protein relationships are many-to-many rather than one-to-one. We therefore derived normalized association weights $q_{i,p}$ from measured supervised-protein abundances:

$$q_{i,p} = \frac{\exp\left(y_{i,p}/\tau_y\right)}{\sum_{p' \in \mathcal{P}_{\text{train}}} \exp\left(y_{i,p'}/\tau_y\right)}, \qquad (10)$$

Here, $y_{i,p}$ is the standardized abundance of protein $p$ at spatial unit $i$, and $\tau_y$ controls the concentration of the target distribution.

The predicted cell-to-protein distribution is

$$\hat{q}_{i,p} = \frac{\exp\left(S_{i,p}\right)}{\sum_{p' \in \mathcal{P}_{\text{train}}} \exp\left(S_{i,p'}\right)}. \qquad (11)$$

The corresponding alignment loss is

$$\mathcal{L}_{\text{cell}\to\text{prot}} = -\frac{1}{B} \sum_{i=1}^{B} \sum_{p \in \mathcal{P}_{\text{train}}} q_{i,p} \log \hat{q}_{i,p}\,. \qquad (12)$$

For the reverse protein-to-cell direction, the normalized target distribution is

$$r_{p,i} = \frac{\exp\left(y_{i,p}/\tau_y\right)}{\sum_{j=1}^{B} \exp\left(y_{j,p}/\tau_y\right)}, \qquad (13)$$

and the model-derived distribution is

$$\hat{r}_{p,i} = \frac{\exp\left(S_{i,p}\right)}{\sum_{j=1}^{B} \exp\left(S_{j,p}\right)}. \qquad (14)$$

The protein-to-cell alignment loss is

$$\mathcal{L}_{\text{prot}\to\text{cell}} = -\frac{1}{|\,\mathcal{P}_{\text{train}}\,|} \sum_{p \in \mathcal{P}_{\text{train}}} \sum_{i=1}^{B} r_{p,i} \log \hat{r}_{p,i}\,. \qquad (15)$$

The bidirectional contrastive objective is

$$\mathcal{L}_{\text{align}} = \frac{1}{2}\left(\mathcal{L}_{\text{cell}\to\text{prot}} + \mathcal{L}_{\text{prot}\to\text{cell}}\right). \qquad (16)$$

This objective assigns greater compatibility to spatial units with stronger measured expression of a supervised protein.

***Protein-conditioned FiLM regression.*** Contrastive alignment used the shared projections, whereas abundance prediction used tissue embedding $h_i$ and protein embedding $e_p$.

A FiLM generator $g_\theta$ transformed the protein embedding into protein-specific scale and shift vectors:

$$[\gamma_p; \beta_p] = g_\theta(e_p), \qquad (17)$$

where

$$g_\theta: \mathbb{R}^{d_e} \to \mathbb{R}^{2d_h},\ \gamma_p, \beta_p \in \mathbb{R}^{d_h}. \qquad (18)$$

The protein-specific parameters modulate the tissue embedding through a feature-wise affine transformation:

$$u_{i,p} = \gamma_p \odot h_i + \beta_p, \qquad (19)$$

Here, $\odot$ denotes element-wise multiplication. A nonlinear activation was then applied:

$$z_{i,p} = \mathrm{ReLU}(\gamma_p \odot h_i + \beta_p). \qquad (20)$$

A shared readout predicted the standardized abundance of protein $p$ at spatial unit $i$:

$$\hat{y}_{i,p} = \mathrm{Readout}(z_{i,p}). \qquad (21)$$

For a linear readout,

$$\hat{y}_{i,p} = w_{\mathrm{out}}^T z_{i,p} + b_{\mathrm{out}}, \qquad (22)$$

Parameters $w_{\mathrm{out}}$ and $b_{\mathrm{out}}$ were shared across proteins. Protein specificity arose from FiLM parameters $\gamma_p$ and $\beta_p$, not separate output nodes.

For a held-out protein $p^*$, prediction followed the same pathway:

$$a_{p^*} \to e_{p^*} \to (\gamma_{p^*}, \beta_{p^*}) \to \hat{y}_{i,p^*}. \qquad (23)$$

Explicitly,

$$\hat{y}_{i,p^*} = \mathrm{Readout}[\mathrm{ReLU}(\gamma_{p^*} \odot h_i + \beta_{p^*})]. \qquad (24)$$

Thus, no protein-specific output head was trained for $p^*$.

Protein-abundance regression objective. Let

$$\Omega_{\mathrm{train}} = \{(i, p): p \in \mathcal{P}_{\mathrm{train}}\} \qquad (25)$$

denote the set of supervised cell–protein pairs. The regression objective is evaluated only for proteins in the training panel.

Using a Huber loss, the regression term is

$$\mathcal{L}_{\text{reg}} = \frac{1}{\mid \Omega_{\text{train}} \mid} \sum_{(i,p)\in\Omega_{\text{train}}} \rho_{\delta}\left(\hat{y}_{i,p} - y_{i,p}\right), \qquad (26)$$

where

$$\rho_{\delta}(\varepsilon) = \begin{cases} \frac{1}{2}\varepsilon^{2}, & \mid \varepsilon \mid \leq \delta, \\ \delta\left(\mid \varepsilon \mid -\frac{1}{2}\delta\right), & \mid \varepsilon \mid > \delta. \end{cases} \qquad (27)$$

$$\mathcal{L}_{\text{reg}} = \frac{1}{\mid \Omega_{\text{train}} \mid} \sum_{(i,p)\in\Omega_{\text{train}}} \left(\hat{y}_{i,p} - y_{i,p}\right)^{2}. \qquad (28)$$

Joint learning objective. The contrastive alignment and protein-conditioned regression modules are optimized jointly:

$$\mathcal{L} = \lambda_{\text{align}}\mathcal{L}_{\text{align}} + \lambda_{\text{reg}}\mathcal{L}_{\text{reg}}, \qquad (29)$$

Here, $\lambda_{\text{align}}$ and $\lambda_{\text{reg}}$ control the contributions of contrastive alignment and abundance regression.

The contrastive term aligns tissue states with protein descriptors, while the regression term predicts continuous abundance. Both objectives are optimized end to end.

Virtual protein profile generation. After training, ProCLIP produces a cell- or spot-by-protein matrix:

$$\hat{Y} = \left[\hat{y}_{i,p}\right] \in \mathbb{R}^{N\times P_{\text{query}}}. \qquad (30)$$

The queried protein set can contain three categories:

$$\mathcal{P}_{\text{query}} = \mathcal{P}_{\text{measured}} \cup \mathcal{P}_{\text{held-out}} \cup \mathcal{P}_{\text{extended}}. \qquad (31)$$

Here, $\mathcal{P}_{\text{measured}}$ denotes supervised proteins, $\mathcal{P}_{\text{held-out}}$ denotes measured proteins excluded from abundance supervision and $\mathcal{P}_{\text{extended}}$ denotes descriptor-only proteins without matched measurements.

We quantitatively validated $\mathcal{P}_{\text{held-out}}$ because measured abundance maps were available. Predictions for $\mathcal{P}_{\text{extended}}$ represent unvalidated computational panel expansion and require independent experimental confirmation.

# CAF spatial subtype recovery

We assessed whether ProCLIP-derived profiles preserved biologically meaningful organization beyond individual protein maps. The analysis used held-out HCC-Wu patient 84, comprising 5,844 CAFs with spatial coordinates, CODEX-derived cell annotations and ProCLIP-predicted profiles.

***Reference CAF subtype reconstruction.*** We defined reference labels from the cellular composition of each fibroblast's local CODEX neighbourhood, following Liu et al. For every CAF, we identified the 15 nearest cells and calculated the proportions of malignant or epithelial, stromal, myeloid, and

lymphoid or plasma cells. Each CAF was assigned to the subtype corresponding to its dominant local cellular context.

The labels describe the spatial niche occupied by each CAF rather than its intrinsic transcriptomic state. We defined four CAF subtypes:

*Tumour-adjacent CAFs (s1).* These CAFs occupied cancer-rich neighbourhoods with many malignant or epithelial cells. Representative niche-associated markers included ALB, CDH1 and SLPI. These markers supported biological interpretation but did not directly define the reference labels.

*Stromal/ECM-rich CAFs (s2).* These CAFs occupied stroma-dominated neighbourhoods with extracellular-matrix and matrix-remodelling signals. Representative markers included COL6A1, DCN, BGN and MMP2.

*Myeloid-inflammatory CAFs (s3).* These CAFs occupied neighbourhoods enriched for monocyte- and macrophage-associated cells. Representative markers included CD68, LYZ, CD14, S100A8 and S100A9. This was the most difficult subtype to recover from ProCLIP profiles.

*Lymphoid/TLS-like CAFs (s4).* These CAFs occupied neighbourhoods enriched for T cells, B cells and plasma cells. Representative markers included CD3D, CD3E, CD79A, JCHAIN and CXCL13, consistent with organized lymphoid niches.

***CAF subtype classification.*** Each CAF was represented by a 560-dimensional ProCLIP profile. A class-balanced multinomial logistic-regression classifier was evaluated using fivefold cross-validation. Predictions from the held-out folds were concatenated to obtain one out-of-fold subtype prediction per CAF.

CODEX neighbourhood labels served as the classification reference. Measured CODEX protein intensities and true neighbouring-cell identities were not supplied to the classifier.

We evaluated performance using balanced accuracy, adjusted Rand index and subtype-specific recall. For four CAF subtypes, balanced accuracy was defined as

$$\text{Balanced Accuracy} = \frac{1}{4}\sum_{k=1}^{4}\frac{\text{TP}_k}{\text{TP}_k + \text{FN}_k},$$

Here, $\text{TP}_k$ and $\text{FN}_k$ denote correctly recovered and missed CAFs for subtype $k$. A row-normalized confusion matrix summarized assignments for each reference subtype.

## Baseline methods

We compared ProCLIP with four methods spanning histology-based and transcriptome-to-proteome prediction. HEX predicts a predefined protein panel from H&E using pathology foundation-model features. The MUSK baseline uses frozen morphology embeddings with a supervised regression head. SpaTranslator performs supervised translation from transcriptomic expression to protein abundance. STProtein combines an RNA-based GraphSAGE encoder with a supervised multi-output prediction head.

## Evaluation metrics

We evaluated protein prediction using Pearson correlation coefficient (PCC), Spearman rank correlation coefficient (SCC) and root mean squared error (RMSE). Each metric was calculated separately for every

evaluated protein across test spatial units. Protein-wise PCC was the primary metric because it measures recovery of spatial variation independently of absolute signal scale.

For protein $p$, PCC between measured abundance $y_{i,p}$ and predicted abundance $\hat{y}_{i,p}$ across $N$ test units was calculated as

$$\mathrm{PCC}_p = \frac{\sum_{i=1}^{N}\left(y_{i,p} - \bar{y}_p\right)\left(\hat{y}_{i,p} - \bar{\hat{y}}_p\right)}{\sqrt{\sum_{i=1}^{N}\left(y_{i,p} - \bar{y}_p\right)^2}\sqrt{\sum_{i=1}^{N}\left(\hat{y}_{i,p} - \bar{\hat{y}}_p\right)^2}}, \qquad (1)$$

Here, $\bar{y}_p$ and $\bar{\hat{y}}_p$ denote the mean measured and predicted abundance of protein $p$, respectively.

SCC was calculated by applying the Pearson correlation formula to rank-transformed measured and predicted values:

$$\mathrm{SCC}_p = \mathrm{PCC}\left[\mathrm{rank}\left(y_{:,p}\right)\mathrm{rank}\left(\hat{y}_{:,p}\right)\right]. \qquad (2)$$

SCC measures whether the model preserves spatial-unit rankings when the measured and predicted abundances are not linearly related.

RMSE was calculated as

$$\mathrm{RMSE}_p = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\hat{y}_{i,p} - y_{i,p}\right)^2}. \qquad (3)$$

RMSE was reported only when predictions and measurements used the same target scale. For ProCLIP, both were evaluated in the standardized abundance space used for training. No post hoc calibration used measurements from the test protein.

For held-out protein set $\mathcal{P}_{\mathrm{test}}$, mean protein-wise PCC was calculated as

$$\bar{\mathrm{PCC}} = \frac{1}{|\,\mathcal{P}_{\mathrm{test}}\,|}\sum_{p\in\mathcal{P}_{\mathrm{test}}}\mathrm{PCC}_p\,, \qquad (4)$$

Mean SCC and RMSE were calculated analogously. Unless stated otherwise, variation across proteins was summarized by the sample standard deviation:

$$s_{\mathrm{PCC}} = \sqrt{\frac{1}{|\,\mathcal{P}_{\mathrm{test}}\,| - 1}\sum_{p\in\mathcal{P}_{\mathrm{test}}}\left(\mathrm{PCC}_p - \bar{\mathrm{PCC}}\right)^2}. \qquad (5)$$

## Ablation study design

We designed an ablation study on OV to assess major ProCLIP components under protein-held-out evaluation. The held-out proteins were FOXP3, HLA-DR, IDO1 and MPO. Mean protein-wise PCC across these markers was the primary metric.

The complete model was denoted A0. Five ablated variants were evaluated:

- A1, without CLIP, removed the cell–protein contrastive objective while retaining abundance regression.
- A2, without FiLM, removed protein-conditioned scale-and-shift modulation.
- A3, without H&E/MUSK, removed the pretrained morphology embeddings from the tissue input.
- A4, without spatial context, removed coordinate-derived and neighbourhood features.
- A5, without the residual prior, removed the dataset-specific residual-prior component used for OV.

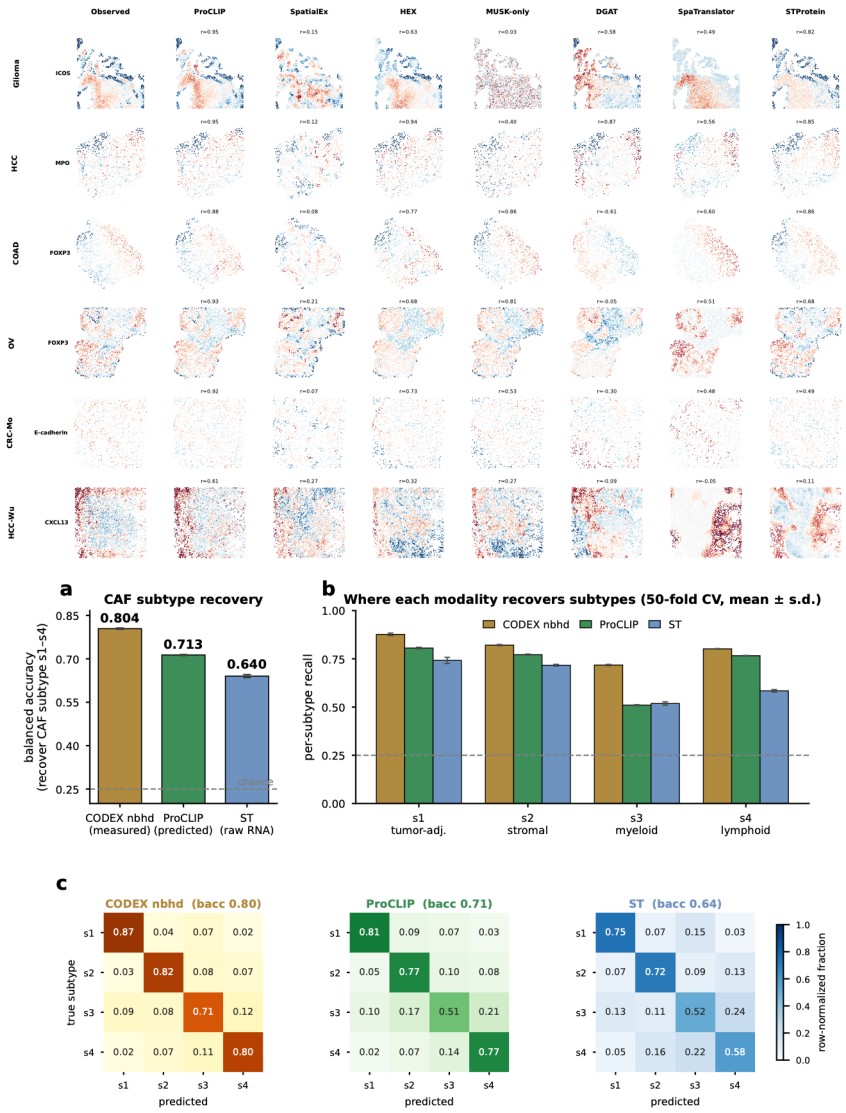
All other components were held constant within the redesigned ablation branch. Each ablated variant was trained with three random seeds. Test PCC was summarized as the mean and sample standard deviation across seeds.

$$\bar{\mathrm{PCC}}_{\mathrm{variant}} = \frac{1}{R}\sum_{r=1}^{R} \mathrm{PCC}_r\,,\ R = 3, \qquad (7)$$

$$s_{\mathrm{variant}} = \sqrt{\frac{1}{R-1}\sum_{r=1}^{R} (\mathrm{PCC}_r - \bar{\mathrm{PCC}}_{\mathrm{variant}})^2}\,. \qquad (8)$$

The change relative to A0 was calculated as

$$\Delta\mathrm{PCC}_v = \bar{\mathrm{PCC}}_v - \bar{\mathrm{PCC}}_{\mathrm{A0}}\,. \qquad (9)$$




# References


Chen, Jun et al. 2024. “Spatial Landscapes of Cancers: Insights and Opportunities.” *Nature Reviews Clinical Oncology*.

Ganin, Yaroslav, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. 2016. “Domain-Adversarial Training of Neural Networks.” *Journal of Machine Learning Research* 17 (59): 1–35.

Gayoso, Adam, Zoe Steier, Romain Lopez, Jeffrey Regier, Kristopher L. Nazor, Aaron Streets, and Nir Yosef. 2021. “Joint Probabilistic Modeling of Single-Cell Multi-Omic Data with totalVI.” *Nature Methods* 18: 272–82. https://doi.org/10.1038/s41592-020-01050-x.

Giesen, Charlotte, Hao A. O. Wang, Denis Schapiro, Nevena Zivanovic, Andrea Jacobs, Bodo Hattendorf, Peter J. Schüffler, et al. 2014. “Highly Multiplexed Imaging of Tumor Tissues with Subcellular Resolution by Mass Cytometry.” *Nature Methods* 11: 417–22. https://doi.org/10.1038/nmeth.2869.

Goltsev, Yury, Nikolay Samusik, Julia Kennedy-Darling, Salil Bhate, Mark Hale, Gustavo Vazquez, Sarah Black, and Garry P. Nolan. 2018. “Deep Profiling of Mouse Splenic Architecture with CODEX Multiplexed Imaging.” *Cell* 174 (4): 968–981.e15. https://doi.org/10.1016/j.cell.2018.07.010.

He, Bryan, Ludvig Bergenstråhle, Linnea Stenbeck, Abubakar Abid, Alma Andersson, Åke Borg, Jonas Maaskola, Joakim Lundeberg, and James Zou. 2020. “Integrating Spatial Gene Expression and Breast Tumour Morphology via Deep Learning.” *Nature Biomedical Engineering* 4: 827–34. https://doi.org/10.1038/s41551-020-0578-x.

Jiang, Zhaorui, Yingfang Yuan, Lei Hu, and Wei Pang. 2026. “STProtein: Predicting Spatial Protein Expression from Multi-Omics Data.” *arXiv Preprint arXiv:2602.05811*. https://arxiv.org/abs/2602.05811.

Keren, Leeat, Marc Bosse, Diana Marquez, Roshan Angoshtari, Samir Jain, Sushama Varma, Soo-Ryum Yang, et al. 2018. “A Structured Tumor-Immune Microenvironment in Triple Negative Breast Cancer Revealed by Multiplexed Ion Beam Imaging.” *Cell* 174 (6): 1373–1387.e19. https://doi.org/10.1016/j.cell.2018.08.039.

Li, Z., Y. Li, J. Xiang, et al. 2026. “AI-Enabled Virtual Spatial Proteomics from Histopathology for Interpretable Biomarker Discovery in Lung Cancer.” *Nature Medicine*. https://doi.org/10.1038/s41591-025-04060-4.

Lin, Jia-Ren, Mohammad Fallahi-Sichani, and Peter K. Sorger. 2018. “Highly Multiplexed Immunofluorescence Imaging of Human Tissues and Tumors Using t-CyCIF and Conventional Optical Microscopes.” *eLife* 7: e31657. https://doi.org/10.7554/eLife.31657.

Liu, Yonghao, Chuyao Wang, Zhikang Wang, et al. 2026. “High-Parameter Spatial Multi-Omics Through Histology-Anchored Integration.” *Nature Methods*. https://doi.org/10.1038/s41592-025-02926-6.

Liu, Yunhe, Ansam Sinjab, Jimin Min, et al. 2025. “Conserved Spatial Subtypes and Cellular Neighborhoods of Cancer-Associated Fibroblasts Revealed by Single-Cell Spatial Multi-Omics.” *Cancer Cell* 43: 905–24. https://doi.org/10.1016/j.ccell.2025.03.004.

Mo, Chia-Kuei, Jingxian Liu, Siqi Chen, et al. 2024. “Tumour Evolution and Microenvironment Interactions in 2D and 3D Space.” *Nature* 634: 1178–86. https://doi.org/10.1038/s41586-024-08087-4.

Pang, Min et al. 2025. “Spatial Gene Expression at Single-Cell Resolution from Histology Images.” *Nature Methods*.

Perez, Ethan, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron Courville. 2018. “FiLM: Visual Reasoning with a General Conditioning Layer.” In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Radford, Alec, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, et al. 2021. “Learning Transferable Visual Models from Natural Language Supervision.” In *Proceedings of the 38th International Conference on Machine Learning*, 8748–63.

Ren, Pengfei, Rui Zhang, Yunfeng Wang, et al. 2025. “Systematic Benchmarking of High-Throughput Subcellular Spatial Transcriptomics Platforms Across Human Tumors.” *Nature Communications* 16: 9232. https://doi.org/10.1038/s41467-025-64292-3.

SPATCH Database. 2025. “SPATCH Database.” http://spatch.pku-genomics.org/.

Ståhl, Patrik L., Fredrik Salmén, Sanja Vickovic, Anna Lundmark, José Fernández Navarro, Jens Magnusson, Stefania Giacomello, et al. 2016. “Visualization and Analysis of Gene Expression in Tissue Sections by Spatial Transcriptomics.” *Science* 353 (6294): 78–82. https://doi.org/10.1126/science.aaf2403.

Sun, Baochen, and Kate Saenko. 2016. “Deep CORAL: Correlation Alignment for Deep Domain Adaptation.” In *Computer Vision – ECCV 2016 Workshops*, 443–50. https://doi.org/10.1007/978-3-319-49409-8_35.

Tang, Jihong, Wenhua Fan, Yuyan Ruan, et al. 2025. “Protein-Based Classification Reveals an Immune-Hot Subtype in IDH Mutant Astrocytoma with Worse Prognosis.” *Cancer Cell* 43 (11): 2136–2155.e14. https://doi.org/10.1016/j.ccell.2025.08.006.

Wu, Zhenqin, Joseph Boen, Sonali Jindal, et al. 2025. “Spatial Multi-Omics and Deep Learning Reveal Fingerprints of Immunotherapy Response and Resistance in Hepatocellular Carcinoma.” https://doi.org/10.1101/2025.06.11.656869.