

A Deep Learning Framework for Protein Expression Prediction from Histology Images

By

Ian Liu

Thesis

Submitted in partial fulfillment of the requirements for the Degree of
Master of Science in the Department of Biostatistics at Brown University

PROVIDENCE, RHODE ISLAND

May 2026

© Copyright 2026(Ian Liu)

Acknowledgments

I would like to thank Professor Ying Ma for guiding me throughout this project and for supporting my academic journey at Brown. Thank you for introducing me to computational biology and spatial omics, for giving me the freedom to explore my own research interests, and for meeting my uneven progress with patience, generosity, and high expectations. I still remember speaking with you as a first-year MS student after I was the only person in class who had not turned in an overdue homework assignment. I worried that I had made a poor first impression, but over time I came to appreciate the rigor, breadth, and curiosity with which you approach research. I find it unbelievable how you can understand concepts and methods from so many different diverse fields and bridge them together. Working with you, I have learned from the way you connect ideas across fields, ask questions from first principles, and hold ambitious standards for scientific work. Your mentorship made this thesis possible and helped me grow as a researcher. I am honored to have known and worked with you.

Thank you to my thesis reader, Professor Fenghai Duan, for your time and careful attention to this work. I am very grateful for your willingness to read this thesis, for your valuable perspective, and for your broader contributions to the field of biostatistics.

I am grateful to Professor Stavroula A. Chrysanthopoulou for serving as my academic advisor and for providing thoughtful feedback and support throughout my time in the Department of Biostatistics. Your guidance helped shape my academic path at Brown and gave me a steadier foundation as a student and researcher.

I thank Brown University and the Department of Biostatistics for giving me the opportunity to study, work, and grow in a rigorous and supportive environment. I am also grateful to my classmates, instructors, and colleagues whose questions and conversations helped make my time at Brown intellectually rewarding.

Most importantly, I would like to thank my parents for their unwavering support and trust. None of this would have been possible without you. Thank you.

I am deeply thankful to the friends who made my life at Brown fuller and more joyful: Yuan-Jhen Lee, Man-Fang Liang, Chen-En Ma, Yu-Liang Tsai, Ethan Chen, Wenjing Li, Julia (Wanjia) Fu, and many others I cannot fully list here. I also thank the members of Moli Dance Company, the Taiwanese Graduate Student Association, and the classmates I met across my courses for the community, friendship, and memories they gave me.

Finally, I would like to thank my alma mater Fei Tian College Northern Campus for giving me a foundation in academics when I was pivoting from dance to science, and for providing a place where I could explore my interests and develop the skills that eventually brought me to this work.

Contents

Acknowledgments	iv
1 BACKGROUND & MOTIVATION	1
1.1 Motivation & Problem Statement	1
1.1.1 Bridging the Modality Gap	1
1.1.2 Diagonal Integration	2
1.1.3 Scalability and Resolution Limits of Spatial Proteomics	3
1.1.4 Why Protein Expression Prediction Matters	3
1.1.5 Thesis Objectives and Research Questions	4
1.1.6 Main Contributions	4
1.2 Background & Related Work	5
1.2.1 Histopathology as a Computational Modality	5
1.2.2 Spatial Transcriptomics and Spatial Proteomics	6
1.2.3 Algorithmic Evolution of Omics Inference from Histology	7
1.2.4 Multimodal Representation Learning in Computational Pathology	8
1.2.5 Pathology Foundation Models	9
1.2.6 DeepPT and Related Direct-Prediction Baselines	9
1.2.7 Embedding-Based Retrieval/Alignment Methods	10
1.2.8 Evaluation of Virtual Omics	10
1.2.9 Gaps in Existing Methods	11

2	Methods	12
2.1	Overview	12
2.2	Datasets and Study Design	14
2.3	Preprocessing	15
2.4	Two-Stage Latent Alignment Model	16
2.4.1	Rationale	16
2.4.2	Stage 1: Modality-Specific Representation Learning	17
2.4.3	Stage 2: Cross-Modal Alignment	20
2.4.4	Augmentation and Regularization	21
2.5	DeepPT Baseline	22
2.6	Embedding-Derived Shallow Baselines	24
2.7	Ablation Studies	25
2.8	Evaluation Metrics and Statistical Comparisons	26
3	Results	30
3.1	Whole-Section Transfer Benchmark	30
3.2	Overall Predictive Performance	31
3.3	Generalization Was Direction-Dependent	33
3.4	Ablations and Representation Probes	42
3.5	Marker-Specific Predictability and Case Studies	43
3.6	Spatial and Multivariate Structure	47
3.7	Summary	53
4	Discussion	55
4.1	Findings	55
4.2	Interpretation of Model Behavior	56
4.3	Biological Interpretation	57
4.4	Implications for Virtual Spatial Proteomics	58
4.5	Limitations	59

4.6	Future Work	60
4.7	Conclusion	60
A	Supplementary Tables	62
A.1	Detailed Model Evaluation Tables	75

List of Figures

2.1	Stage 1 representation learning. Protein profiles are encoded into a variational latent space, sampled with the reparameterization trick, and decoded back to the original marker panel under reconstruction and MMD regularization. In parallel, augmented H&E tiles are reconstructed through an image encoder-decoder, giving the model a compact morphology representation before cross-modal alignment.	19
2.2	Stage 2 cross-modal alignment. The pretrained protein decoder is reused, while an MLP maps image embeddings into the protein latent space. The image encoder may be fine-tuned during this stage, but the prediction objective is protein reconstruction from image-derived latent coordinates rather than image reconstruction.	23
2.3	Inference workflow for a held-out tissue section. Each H&E tile is encoded, mapped through the trained alignment network into the protein latent space, and decoded into a predicted P -marker protein profile. The protein encoder and image decoder are training-time components and are not needed to generate validation-section predictions.	23
3.1	Validation performance across the two transfer directions. The proposed model improved marker-level association relative to DeepPT, whereas shallow probes of the learned image representation remained competitive and sometimes superior for pointwise error.	32

3.2	Forward-validation marker-level association metrics. Each point is a protein, summarized across the overall, training, and validation partitions for Pearson correlation, Spearman correlation, concordance correlation coefficient, and partial correlation. The proposed model improved held-out marker association relative to DeepPT, although linear and tree-based probes remained competitive for several association summaries.	34
3.3	Forward-validation marker-level error and spatial-autocorrelation metrics. RMSE and MAE quantify pointwise error for each protein map, while the Moran's-I panels compare spatial autocorrelation in the measured, predicted, and absolute-difference summaries. These diagnostics show that lower error and better spatial association were not always achieved by the same model.	35
3.4	Forward-validation marker-level structural and distributional metrics. SSIM and mutual information measure spatial and nonlinear agreement between measured and predicted marker maps, whereas EMD, MMD, and Jensen-Shannon divergence summarize distributional mismatch. The proposed model was strongest against DeepPT on spatial structure, but the distributional metrics emphasize that calibration and dynamic-range recovery remained incomplete.	36
3.5	Forward-validation spot-level association metrics. Each point is a spot-level protein profile. Correlations were high for many models, indicating that spot-wise association is less discriminating than marker-wise spatial recovery in this benchmark.	37
3.6	Forward-validation spot-level pointwise error. RMSE and MAE summarize the magnitude of errors across the full antibody profile at each spot. These metrics favor models that are well calibrated in abundance, which is related to but distinct from recovering marker-specific spatial patterning.	38

3.7	Forward-validation spot-level structural metrics. Spot-level SSIM and mutual information summarize agreement across the marker vector at each spatial location, reinforcing that high spot-profile agreement does not by itself guarantee accurate spatial maps for individual proteins.	38
3.8	Transfer-direction effects for the principal model comparisons. The proposed model was strongest in the forward direction, whereas the reverse section reduced marker-level association and increased pointwise error.	39
3.9	Forward-direction marker scatter plots for the proposed model. Blue points show the training section and orange points show the held-out validation section. Many markers retain a positive validation relationship with measured abundance, but the prediction range is visibly compressed for several targets.	40
3.10	Reverse-direction marker scatter plots for the proposed model. The validation points show weaker calibration and stronger dynamic-range compression than in the forward direction, explaining why pooled marker-level gains did not carry over symmetrically across sections.	41
3.11	Marker-level heterogeneity in the proposed model. Pearson correlations are averaged across forward and reverse validation sections. Proteins associated with visible tissue organization tended to be more predictable, while several controls and markers with weaker morphologic correlates were near zero or negative.	44
3.12	Representative forward validation maps for CD3E and KRT5. Each row compares measured and predicted expression over the held-out section using a shared color scale for that marker. The model captures coarse spatial structure but smooths local variation and under-recovers some high-expression regions.	46

3.13	KRT5 case study for the proposed model. KRT5 illustrates recovery of a basal epithelial/cytokeratin structural marker. The predictions retain broad epithelial spatial organization, especially in the forward direction, but remain smoother and less extreme than the measured maps.	48
3.14	CD19 case study for the proposed model. CD19 shows partial recovery of broad B-cell-associated spatial regions, but predicted maps are dynamically compressed and miss sharper local enrichment and depletion.	49
3.15	PDCD1 case study for the proposed model. PDCD1 recovery is interpreted as morphology-informed prediction of checkpoint-associated immune context rather than direct visualization of PD-1 expression from H&E.	50
3.16	MS4A1 case study for the proposed model. MS4A1 serves as a direction-dependent boundary case: the forward section retains moderate marker-specific structure, whereas the reverse direction shows strong collapse and loss of calibrated marker recovery.	51
3.17	Forward-validation spatial clustering from measured and predicted protein profiles. The proposed model and DeepPT both recover organized tissue partitions, while ablated models show that high cluster separability alone does not guarantee accurate marker recovery.	52
3.18	Reverse-validation spatial clustering from measured and predicted protein profiles. The reverse section again shows organized partitions, but the relationship between cluster compactness and marker-level prediction quality is model dependent.	52
3.19	Forward-validation protein-protein correlation preservation. The proposed model produced a lower correlation matrix distance than DeepPT, indicating better recovery of cross-marker covariance structure in the easier transfer direction.	54

3.20 Reverse-validation protein-protein correlation preservation. DeepPT had the lower correlation matrix distance in this direction, highlighting that multivariate antibody-panel structure was not preserved uniformly by the proposed model across section transfers.	54
---	----

List of Tables

2.1	Configuration summary of the main proposed model and its corresponding ablation studies.	26
3.1	Pooled validation performance across forward and reverse whole-section transfer. Marker-level metrics are averaged across 66 marker-direction pairs; spot-level metrics are averaged across 9,102 held-out spot profiles. Higher values are better for Pearson, Spearman, and SSIM; lower values are better for RMSE and MAE.	32
3.2	Direction-specific validation performance for the proposed model and DeepPT. The forward direction trains on the first section and validates on the second; the reverse direction swaps these roles.	39
3.3	Top pooled validation proteins for the proposed model, ordered by Pearson correlation. Values are averaged across the two transfer directions.	43
A.1	Complete taxonomy of computational methods for spatial omics and histology analysis. Rows are grouped by refined technical family, and architecture/strategy details are merged to reduce table width.	62
A.2	Complete benchmark inventory for histology-to-spatial-omics prediction. MAE and spot-level PCC are omitted to emphasize reported error and gene-level correlation summaries.	68

A.3	Model-level performance summary by transfer direction and split. The table is a compact version of <code>model_summary_by_direction.csv</code> ; secondary standard-deviation columns are omitted to keep the appendix readable. .	75
A.4	Correlation matrix distance by transfer direction and split. Lower CMD indicates better preservation of the measured cross-marker correlation structure.	76
A.5	Validation Moran’s I values for the proposed model. True and predicted marker maps are compared in each transfer direction; absolute error is computed as the absolute difference between those two Moran’s I values.	77
A.6	Held-out per-protein validation metrics for all compared models and both transfer directions. This table is derived from the validation rows of <code>per_protein_metrics_long.csv</code> ; train and pooled rows are omitted because the main text uses held-out transfer as the primary benchmark. . .	78
A.7	Proposed-model forward per-protein metrics	89
A.8	Proposed-model reverse per-protein metrics	92
A.9	Complete evaluation-metric taxonomy for virtual spatial omics. The applicability column is removed, and measurement focus is merged with the metric description.	94

CHAPTER 1

BACKGROUND & MOTIVATION

1.1 Motivation & Problem Statement

1.1.1 Bridging the Modality Gap

Histopathology using hematoxylin and eosin (H&E) staining (Fischer et al., 2008) has been a central modality for clinical oncology and tissue diagnosis for over a century (Golub et al., 1999; Curtis et al., 2012; Doroshow and Doroshow, 2019). H&E images provide high-resolution views of tissue morphology and cellular architecture, but they do not directly measure molecular states or high-dimensional phenotypes. In contrast, multiplexed protein imaging technologies, such as Imaging Mass Cytometry (IMC) (Giesen et al., 2014) and Co-Detection by Indexing (CODEX) (Goltsev et al., 2018), can quantify dozens of protein biomarkers at single-cell or near-single-cell resolution (Tan et al., 2020). These spatial proteomics platforms help characterize the tumor microenvironment (TME) (Hanahan, 2022) by mapping immune-tumor interactions and cellular spatial organization.

However, a fundamental “modality gap” exists between the ubiquitous, low-cost morphological data of H&E and the expensive, high-dimensional molecular data of spatial proteomics. Computational models may help bridge this gap by estimating spatial protein

patterns that are partially encoded in morphology. In the near term, such models are best framed as tools for hypothesis generation, region prioritization, and assay triage rather than as replacements for direct spatial proteomic measurement. In the long term, developing such models that learn the implicit mapping between tissue architecture and molecular profiles would be crucial for democratizing access to spatial omics without reliance on expensive experimental assays.

1.1.2 Diagonal Integration

While computational prediction of omics from histology is promising, integrating multi-omics datasets presents severe technical hurdles, particularly when establishing ground-truth paired data for model training. Because multiple molecular assays (e.g., transcriptomics and proteomics) or complementary high-parameter panels often cannot be applied simultaneously to the exact same tissue section without destructive interference, measurements are typically acquired on adjacent, serial sections (Ekvall et al., 2024). This introduces the spatial diagonal integration problem, defined by the challenge of reconstructing unified multi-omics states when datasets lack shared, overlapping molecular features and suffer from spatial misalignment at the single-cell level (Xu and McCord, 2022).

As Liu et al. (2026) emphasize, standard integration strategies (vertical, horizontal, or mosaic) have seen more progress than spatial diagonal integration. When applying machine learning to integrate adjacent slides, the morphological shifts and cell-to-cell variations between slices mean that pixel-to-pixel exactness is impossible (Andani et al., 2025). H&E histology can therefore serve as a common morphological anchor for relating non-overlapping omics features, while preserving spatial continuity and contextual dependencies across the tissue volume without requiring identically co-measured data (Liu et al., 2026).

1.1.3 Scalability and Resolution Limits of Spatial Proteomics

Despite its biological value, spatial proteomics faces scalability and resolution tradeoffs. Technologies like IMC and multiplex immunofluorescence (mIF) require specialized instrumentation, long processing times, and expensive reagents, which can restrict their use to smaller cohorts and selected regions of interest (ROIs) (Andani et al., 2025; Gong et al., 2024). Furthermore, these high-resolution platforms have throughput and panel-size limitations, often around 40-50 markers (Schürch et al., 2020). Likewise, high-resolution spatial transcriptomics technologies like MERFISH (Chen et al., 2015) and 10x Xenium (Janesick et al., 2023) use predefined gene panels, limiting fully transcriptome-wide discovery (Liu et al., 2026; Bonev et al., 2024; Piwecka et al., 2023; Seferbekova et al., 2023).

Additionally, achieving cellular or sub-cellular resolution across entire gigapixel whole-slide images (WSIs) remains computationally and experimentally difficult. Computational super-resolution and virtual-staining frameworks could help extend coarse, spot-level omics data or spatially restricted measurements into denser expression maps, but these predictions require careful validation before biological or clinical use (Zhang et al., 2024; Zhang and Li, 2025).

1.1.4 Why Protein Expression Prediction Matters

While significant computational effort has been directed toward predicting spatial transcriptomics from H&E (He et al., 2020; Wang et al., 2025; Pang et al., 2021), predicting spatial proteomics is also important for translational applications. As Li et al. (2026b) notes, proteins are more closely related to actual cellular functions and phenotypes than mRNA transcripts. Furthermore, proteins are generally more robust than RNA in routine FFPE tissue workflows, whereas RNA is more susceptible to degradation and technical dropout (Lai et al., 2016; Chung et al., 2008; Qiu, 2020).

If sufficiently reliable, morphology-based protein prediction could support virtual

staining and virtual spatial proteomics workflows (Li et al., 2026b). In practice, current models should be evaluated for more limited uses first: identifying broad morphology-linked protein patterns, prioritizing regions for follow-up multiplexed assays, and screening candidate biomarkers before committing to expensive experimental panels (Valanarasu et al., 2026).

1.1.5 Thesis Objectives and Research Questions

To address this narrower prediction problem, this thesis develops a two-stage latent alignment framework for spot-level protein prediction from H&E histology. The study is a controlled proof-of-concept using two public human tonsil CytAssist sections and 33 shared antibody targets. The intended contribution is not simply the use of a latent space; it is a protein-decoder-constrained alignment design evaluated under whole-section transfer. The core research questions guiding this work are:

1. Does a protein-latent-decoder-constrained alignment framework improve spot-level protein prediction from histology relative to a DeepPT-style direct-prediction baseline (Hoang et al., 2024) under whole-section transfer?
2. Which architectural and training components contribute most to held-out section performance, especially image-encoder fine-tuning and augmentation?

1.1.6 Main Contributions

This thesis makes four main contributions:

1. It develops a two-stage latent alignment framework for morphology-informed spatial protein prediction from H&E. The method first learns modality-specific image and protein representations, then maps histology-derived embeddings into a protein latent space constrained by a learned decoder from the measured antibody panel.
2. It evaluates this framework in a whole-section transfer setting rather than a random

spot split, comparing against DeepPT, shallow probes, and ablations that isolate image-encoder fine-tuning and augmentation.

3. It shows that performance depends strongly on both marker biology and metric choice. The strongest evidence for the model appears in marker-wise spatial association in the forward transfer direction, while pointwise error, calibration, and multivariate protein-structure preservation remain more mixed.
4. It provides marker-specific evidence for the scope and limits of virtual protein prediction: structural markers such as KRT5 are among the most morphology-linked, immune markers such as CD19 and MS4A1 show graded and direction-sensitive recovery, and checkpoint-associated markers such as PDCD1 require interpretation as contextual immune architecture rather than direct molecular visualization from H&E.

1.2 Background & Related Work

1.2.1 Histopathology as a Computational Modality

The digitization of glass histology slides into Whole Slide Images (WSIs) has made histopathology a quantitative computational modality. Because WSIs often exceed $100,000 \times 100,000$ pixels, they cannot be processed directly by standard deep learning architectures due to severe GPU memory constraints.

Consequently, the field heavily relies on patch extraction and tile-based learning, commonly formulated within a Multiple Instance Learning (MIL) paradigm (Ilse et al., 2018). In this approach, a WSI is decomposed into smaller, non-overlapping patches (e.g., 256×256 pixels), which are individually passed through feature extractors before being aggregated to yield slide-level representations (Jaume et al., 2024).

A central challenge in computational pathology is reconciling local versus contextual

morphology to model tissue heterogeneity (Wang et al., 2025; Fang et al., 2026). While a single high-resolution patch captures localized subvisual features, it inherently lacks the broader macro-architectural context of the tissue, such as the spatial organization of the tumor invasive front or overarching gradients of immune infiltration. Accurate molecular prediction requires models that can contextualize highly localized morphological features within the global structure of the tumor microenvironment.

1.2.2 Spatial Transcriptomics and Spatial Proteomics

Spatial transcriptomics (ST) and spatial proteomics (SP) offer complementary layers of molecular information, but they suffer from distinct technical limitations. Sequencing-based ST technologies such as 10x Visium capture transcriptome-scale RNA measurements by tagging molecules with spatial barcodes (Ståhl et al., 2016). However, ST is hindered by coarse spatial resolution, where a single spot aggregates expression from multiple mixed cells, and suffers from sparsity and technical assay noise (dropouts) (He et al., 2020; Wang et al., 2025; Qiu, 2020; Avşar and Pir, 2023).

In contrast, spatial proteomics explicitly measures the functional effectors of cells. While SP platforms exhibit less technical dropout and provide superior single-cell or subcellular resolution, they are constrained by narrow, pre-defined antibody panels, cost, throughput, and marker count (Giesen et al., 2014; Goltsev et al., 2018; Tan et al., 2020; Andani et al., 2025). Predicting protein expression is therefore distinct from gene-expression prediction: proteins display different spatial topologies, reflect post-transcriptional regulation, and are more directly tied to cellular phenotype. Because protein levels do not correlate perfectly with mRNA abundance, models trained only for ST cannot simply be repurposed for SP without accounting for these functional spatial architectures (Li et al., 2026b; Andani et al., 2025; Yuan et al., 2022).

1.2.3 Algorithmic Evolution of Omics Inference from Histology

The algorithmic approach to inferring molecular omics from histology has evolved rapidly through several distinct phases. The most relevant distinction for this thesis is between models that directly regress molecular targets from image features, models that add spatial or generative structure, and models that align images and omics through a shared or translated representation. The proposed method belongs to the third family, but differs from contrastive retrieval methods by decoding predictions through a learned protein manifold.

Appendix Table A.1 expands this overview into a reproducibility-oriented method inventory. The appendix is intended as reference material for readers who want implementation and citation context, not as a necessary extension of the main argument.

Convolution-Based Approaches

The pioneering wave, exemplified by ST-Net, framed histology-to-omics prediction as a local supervised regression problem (He et al., 2020). ST-Net utilizes a pre-trained Convolutional Neural Network (DenseNet-121) to independently predict the gene expression of a given spatial spot using only its corresponding 224×224 image patch. Although it did not explicitly model spatial continuity across neighboring spots, this formulation established the feasibility of histology-to-omics inference (He et al., 2020; Wang et al., 2025).

Subsequent iterations expanded upon this paradigm by implementing deeper architectures like VGG16 (e.g., DeepSpaCE by Monjo et al. (2022)) or, more commonly, ResNet50. While these CNN models effectively extract local morphological textures, their localized receptive fields limit their ability to model long-range spatial dependencies. However, the benchmarking study by Wang et al. (2025) found that CNN-based approaches outperformed more complex vision transformers, suggesting that the gains transformers have shown in other data-rich domains may be more limited in biomedical settings, where

datasets are typically small or moderate- n .

Attention & Transformer-Based Methods

To incorporate broader context and explicitly capture spatial neighborhood relationships, later methods leveraged transformer attention and graph neural networks (Pang et al., 2021; Zeng et al., 2022). Models such as HisToGene and Hist2ST treat tissue patches as sequences or graph nodes, utilizing self-attention mechanisms and message-passing to allow features from adjacent spots to inform local predictions (Pang et al., 2021; Zeng et al., 2022; Wang et al., 2025). This modeling design mitigated isolated and noisy predictions in favor of smoothed expressions.

Generative and Latent-Variable Approaches

The next shift has been toward generative modeling, in which the objective goes beyond predicting marginal expression values. STFlow uses flow matching to model whole-slide joint distributions of gene expression, thereby accounting for cell-cell interactions (Huang et al., 2025). In parallel, virtual proteomics models such as HistoPlexer, which uses conditional Generative Adversarial Networks (cGANs), explicitly generate multiplex protein maps conditioned on histology (Andani et al., 2025; Li et al., 2026b). These approaches are attractive because they can enforce stronger structural fidelity and more realistic co-localization patterns than independent spot-wise regression, although they may require larger paired datasets than are available in the present proof-of-concept.

1.2.4 Multimodal Representation Learning in Computational Pathology

To bridge the gap between images and molecular profiles, recent work has increasingly embraced shared latent spaces for paired histology and omics data (Xie et al., 2023; Valanarasu et al., 2026). A core technique within this domain is contrastive learning,

which aims to construct a shared latent space by encoding matched H&E image patches and corresponding omics vectors through respective modality encoders, maximizing the similarity between paired image-omics representations while pushing non-matching pairs apart (Xie et al., 2023). Such designs are particularly useful when direct pixel-to-molecule supervision is sparse because the model can exploit relational structure in the paired data rather than relying only on pointwise regression. Furthermore, contrastive learning avoids posterior collapse (a common failure mode in autoencoder-based approaches where the latent space fails to capture meaningful variation) by explicitly enforcing that the latent representations of paired samples are informative and discriminative (Ghose and Poupart, 2023; Oord et al., 2018).

1.2.5 Pathology Foundation Models

Historically, many histology models relied on CNNs pretrained on natural images, which introduced an obvious domain shift. More recent benchmark studies show that pathology-specific foundation models provide stronger and more transferable encoders for downstream spatial omics tasks, especially in low-data regimes (Jaume et al., 2024; Wang et al., 2025). In practice, these encoders serve as good visual backbones that stabilize training and reduce overfitting, but they still need task-specific heads or alignment modules to recover fine-grained molecular signals (Jaume et al., 2024; Fang et al., 2026).

1.2.6 DeepPT and Related Direct-Prediction Baselines

DeepPT represents a prominent direct-prediction baseline for histology-to-omics inference (Hoang et al., 2024). Its pipeline extracts image features with a pretrained encoder, compresses them with an autoencoder, and uses a multilayer perceptron to regress molecular targets directly from the compressed representation. This strategy is attractive because it is simple, scalable, and competitive on several benchmark settings, but it does not explicitly encode spatial continuity or the geometry of the predicted molecular

manifold (Hoang et al., 2024; Wang et al., 2025). Representative benchmark patterns are summarized in Appendix Table A.2.

1.2.7 Embedding-Based Retrieval/Alignment Methods

BLEEP (Bi-modal Embedding for Expression Prediction) takes a different approach by treating prediction as retrieval in a joint image-expression space (Xie et al., 2023). Instead of regressing each target *de novo*, BLEEP projects a query image into the shared latent space and imputes expression from nearby reference profiles. This can better preserve biological heterogeneity and cross-modal structure, but its performance depends strongly on the coverage and diversity of the reference dictionary (Xie et al., 2023; Wang et al., 2025).

1.2.8 Evaluation of Virtual Omics

Validating virtual omics requires multiple complementary metrics because no single score captures numerical accuracy, spatial fidelity, distributional agreement, and downstream biological utility simultaneously. Appendix Table A.9 merges the duplicated uploaded metric spreadsheets into a complete taxonomy, retaining the category hierarchy instead of collapsing distinct metric families into a few broad labels.

In practice, correlation metrics quantify whether broad spatial trends are recovered, error metrics measure absolute deviation, structural metrics assess preservation of local tissue architecture, distributional metrics compare population-level profiles, and biological or clinical metrics test whether predictions remain useful for clustering, outcome modeling, or biomarker discovery (Wang et al., 2025; Li et al., 2026b; Andani et al., 2025; Xia et al., 2025).

1.2.9 Gaps in Existing Methods

Despite rapid advancements, several technical gaps remain and motivate the two-stage latent alignment framework proposed in this thesis. The gap addressed here is deliberately specific: whether a small paired CytAssist protein dataset can support decoder-constrained morphology-to-protein alignment that transfers across whole tissue sections.

1. **Preserving spatial structure remains difficult:** direct regression models such as ST-Net and DeepPT can produce localized predictions without explicitly enforcing smooth spatial continuity or neighborhood consistency (He et al., 2020; Hoang et al., 2024).
2. **Protein structure is still under-modeled:** many approaches treat each target as an independent regression output instead of modeling the shared latent geometry and co-expression structure of multiplex proteins (Andani et al., 2025; Li et al., 2026b).
3. **Retrieval methods trade local fidelity for coverage:** embedding-based methods such as BLEEP can preserve global heterogeneity, but they may smooth marker-specific spatial detail when suitable reference exemplars are sparse (Xie et al., 2023; Wang et al., 2025).
4. **Ablation evidence is still limited:** the literature contains relatively few controlled comparisons isolating the effects of augmentation, latent regularization, or encoder fine-tuning, making it difficult to identify which components drive gains across datasets.

CHAPTER 2

Methods

2.1 Overview

This chapter introduces the proposed supervised framework for predicting spatial protein expression from local H&E morphology. For spot i in tissue section s , the observed data consist of an image tile X_{si} centered at that spot, a two-dimensional spatial coordinate $r_{si} = (x_{si}, y_{si})$, and a P -dimensional protein vector $Y_{si} \in \mathbb{R}^P$. The target of inference is a prediction rule

$$f : X_{si} \mapsto \hat{Y}_{si},$$

which estimates spot-level protein expression from morphology alone. Here, s indexes tissue sections and i indexes spots within a section. For notational simplicity, the section index is suppressed in later discussion when there is only a single training or validation slice. The primary scientific question answered in this thesis is whether a learned morphology-to-protein relationship transfers to an independent section. Accordingly, model performance is evaluated on an entire independent tissue section. Because only two compatible sections were available, the experiment should be interpreted as a controlled proof-of-concept

rather than as evidence of broad deployment across tissue types, laboratories, or protein panels.

This formulation relies on several biological and statistical assumptions. First, a subset of proteins is assumed to be partially identifiable from morphology because H&E contains recoverable visual signatures of immune composition (Saltz et al., 2018), epithelial and stromal organization (Wu et al., 2022), and tonsillar/germinal-center architecture (Nave et al., 2001; Brachtel et al., 1996). This assumption is expected to hold unevenly across markers, as lineage and structure-associated markers should be more predictable than technical controls or markers with weak morphologic correlates. Second, the prediction target is treated as a spot-level aggregate rather than a single-cell measurement, so the model estimates the protein composition of the tissue volume captured by each spot. Third, the method assumes a degree of transportability across sections after restricting to shared markers and applying the same preprocessing. This assumption is deliberately tested by evaluating independent slices.

The proposed method uses a two-stage latent alignment strategy rather than direct end-to-end regression. The central inductive bias is that both histology and protein expression can be represented in lower-dimensional spaces, and that the structure of the protein space should guide the final predictions. Rather than treating each marker as an independent scalar target, the model is designed to generate protein profiles that remain close to biologically plausible antibody-expression patterns observed in the data.

This design also deliberately avoids forcing the image and protein latents to coincide through a CLIP-style contrastive objective or an adversarial alignment loss. In a small- n setting, such objectives can overconstrain the model by requiring it to solve reconstruction and cross-modal alignment equally well, even though those goals may not share the same optimal representation and transformation. Morphology-to-protein translation is therefore treated as a distinct mapping problem rather than as exact latent matching: Stage 1 learns stable modality-specific representations, and Stage 2 builds on those representations

while learning only the cross-modal transformation needed to decode protein profiles from image features. The training and inference workflows are summarized in Figures 2.1, 2.2, and 2.3.

2.2 Datasets and Study Design

Two public 10x Genomics Visium CytAssist Spatial Gene and Protein Expression datasets of FFPE human tonsil were used. The first dataset is the standard human tonsil CytAssist FFPE release, reported by 10x Genomics as a female donor in the 30s, imaged with H&E at 20X, processed with Space Ranger 2.1.0, and containing 4,194 tissue-covered spots (10x Genomics, 2023a). The second dataset is the corresponding tonsil release with add-on antibodies, reported as a male 25-year-old donor, imaged with H&E at 20X, processed with Space Ranger 2.1.0, and containing 4,908 tissue-covered spots (10x Genomics, 2023b).

The primary evaluation used two transfer directions. In the forward direction, the standard-panel section was used for training and the add-on-antibody section was held out for validation and final reporting. In the reverse direction, the add-on-antibody section was used for training and the standard-panel section was held out in the same way. This design is more stringent than a within-slice random spot split because it prevents the model from using slide-specific staining, local tissue continuity, or donor-specific morphology as shortcuts. It therefore estimates out-of-section generalization, which is closer to the intended use case of applying a trained model to a new tissue section.

The terminology in the tables follows the original run outputs: the held-out section is reported as “validation” rather than “test.” There was no separate third tissue section for a final test set. The held-out section was also used for early stopping and for comparing model variants, so the reported validation performance should be read as held-out-section evidence within this two-section proof-of-concept, not as an untouched external-cohort

estimate.

The protein prediction task was restricted to markers shared by both sections. This produced $P = 33$ targets: ACTA2, BCL2, CCR7, CD14, CD163, CD19, CD27, CD274, CD3E, CD4, CD40, CD68, CD8A, CEACAM8, CR2, CXCR5, EPCAM, FCGR3A, HLA-DRA, ITGAM, ITGAX, KRT5, MS4A1, PAX5, PCNA, PDCD1, PECAM1, SDC1, VIM, mouse_IgG1k, mouse_IgG2a, mouse_IgG2bk, and rat_IgG2a. The add-on-only antibodies were excluded because they were not observed in both training directions.

2.3 Preprocessing

The raw 10x outputs were converted into a spot-aligned image and protein dataset. Tissue coordinates from Space Ranger were mapped to image pixels using the scale-factor metadata, and only spots marked as tissue-covered and lying within image bounds were retained. Each retained spot was associated with a square H&E tile centered at the spot coordinate. The square tiles had side length 64 pixels. Boundary tiles were padded by reflection before cropping, which avoids introducing artificial black borders at the tissue edge.

No stain normalization was applied in the experiments reported here. This choice keeps the evaluation closer to the raw public 10x data and avoids adding another preprocessing model whose effect would be difficult to separate from the proposed prediction model. The tradeoff is that the learned image representation must tolerate section-level color and staining variability. ImageNet normalization was applied because that normalization matches the scale expected by the pretrained model.

Protein counts were transformed with a centered log-ratio (CLR) transformation using a pseudocount of one. For raw antibody counts C_{ij} at spot i and marker j ,

$$\tilde{Y}_{ij} = \log(C_{ij} + 1) - \frac{1}{P} \sum_{k=1}^P \log(C_{ik} + 1).$$

The CLR transformation is appropriate for compositional measurements because it represents each marker relative to the geometric mean of the marker panel within the same spot (Aitchison, 1982). This reduces sensitivity to spot-specific total antibody capture while preserving relative marker enrichment.

After the whole-section split was defined, each protein was standardized using the training section only:

$$Y_{ij} = \frac{\tilde{Y}_{ij} - \hat{\mu}_{j,\text{train}}}{\hat{\sigma}_{j,\text{train}}}.$$

The same training means and standard deviations were then applied to the validation section. Fitting the scaling parameters only on the training section prevents information leakage from the validation tissue. Neural-network losses were optimized on this standardized scale, while final predictions were converted back to the CLR scale for reporting.

2.4 Two-Stage Latent Alignment Model

2.4.1 Rationale

The proposed model decomposes protein prediction into two problems: learning modality-specific representations and aligning the image representation to the protein representation. This decomposition is motivated by the small number of training spots relative to the complexity of the image-to-protein mapping. A direct neural regressor must simultaneously learn histology features, protein covariance structure, and cross-modal mapping. The two-stage model instead first learns a morphology representation and a protein representation, then learns the cross-modal alignment between them.

The protein latent space is especially important. Protein markers are biologically correlated: for example, immune lineage markers, epithelial markers, stromal markers, and negative-control antibodies are not exchangeable outcomes. Modeling them jointly

allows the decoder to learn feasible multivariate protein profiles. The final prediction is therefore constrained by a learned protein decoder, which acts as an inductive bias toward biologically plausible marker combinations.

2.4.2 Stage 1: Modality-Specific Representation Learning

The image branch uses an ImageNet-pretrained ResNet-50 encoder (He et al., 2016; Deng et al., 2009). Transfer learning is used because the CytAssist datasets contain only thousands of spot-level examples, whereas a fully trainable image encoder would have far more parameters than can be reliably estimated from these data alone. The encoder maps each tile to a 128-dimensional image latent representation, and a convolutional decoder reconstructs the original tile. During Stage 1 image reconstruction, the ResNet wrapper was configured with the final two residual stages trainable.

Image reconstruction was optimized using structural similarity rather than only pixel-wise error. For images X and $\hat{X}$, the structural similarity index can be written as

$$\text{SSIM}(X, \hat{X}) = \frac{(2\mu_X\mu_{\hat{X}} + c_1)(2\sigma_{X\hat{X}} + c_2)}{(\mu_X^2 + \mu_{\hat{X}}^2 + c_1)(\sigma_X^2 + \sigma_{\hat{X}}^2 + c_2)},$$

where μ denotes local mean intensity, σ^2 denotes local variance, $\sigma_{X\hat{X}}$ denotes local covariance, and c_1, c_2 stabilizes the ratio (Wang et al., 2004). The reconstruction loss is

$$\mathcal{L}_{\text{img}}(X, \hat{X}) = 1 - \text{SSIM}(X, \hat{X}).$$

This choice reflects the assumption that spatial arrangement and local contrast in H&E are more relevant for downstream biological prediction than exact RGB reconstruction.

The protein branch uses a variational autoencoder (Kingma and Welling, 2013). The protein encoder and decoder used a 128-dimensional latent space and 1,024 hidden units.

For each standardized protein vector Y_i , the encoder defines an approximate posterior

$$q_\phi(Z_i | Y_i) = \mathcal{N}(\mu_\phi(Y_i), \text{diag}\{\sigma_\phi^2(Y_i)\}),$$

and samples a latent vector using

$$Z_i = \mu_\phi(Y_i) + \sigma_\phi(Y_i) \odot \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0, I).$$

The decoder maps Z_i back to a reconstructed protein vector $\hat{Y}_i$. Protein reconstruction was optimized by mean squared error,

$$\mathcal{L}_{\text{prot}}(Y_i, \hat{Y}_i) = \frac{1}{P} \sum_{j=1}^P (Y_{ij} - \hat{Y}_{ij})^2.$$

Figure 2.1 shows this first stage as two reconstruction problems. The protein pathway learns a regularized latent representation of the antibody panel, while the image pathway learns a compact morphology representation from augmented local tiles.

The protein latent distribution was regularized toward a standard normal prior using maximum mean discrepancy (MMD), a kernel-based discrepancy between probability distributions (Gretton et al., 2012). This makes the protein model a MMD-regularized VAE rather than a standard KL-regularized VAE. The objective function used is related to VAE variants that match the aggregate posterior to the prior using MMD or other distribution-level penalties, including InfoVAE and Wasserstein autoencoders (Zhao et al., 2019; Tolstikhin et al., 2017). In particular, the scalar weight on the latent regularizer follows the same reconstruction-regularization tradeoff motivation as beta-VAE, although here the weighted penalty is MMD rather than KL divergence (Higgins et al., 2017). For encoded samples $Z, Z' \sim q_\phi$ and prior samples $U, U' \sim p(Z) = \mathcal{N}(0, I)$,

$$\text{MMD}^2(q_\phi, p) = \mathbb{E}\{k(Z, Z')\} + \mathbb{E}\{k(U, U')\} - 2\mathbb{E}\{k(Z, U)\}.$$

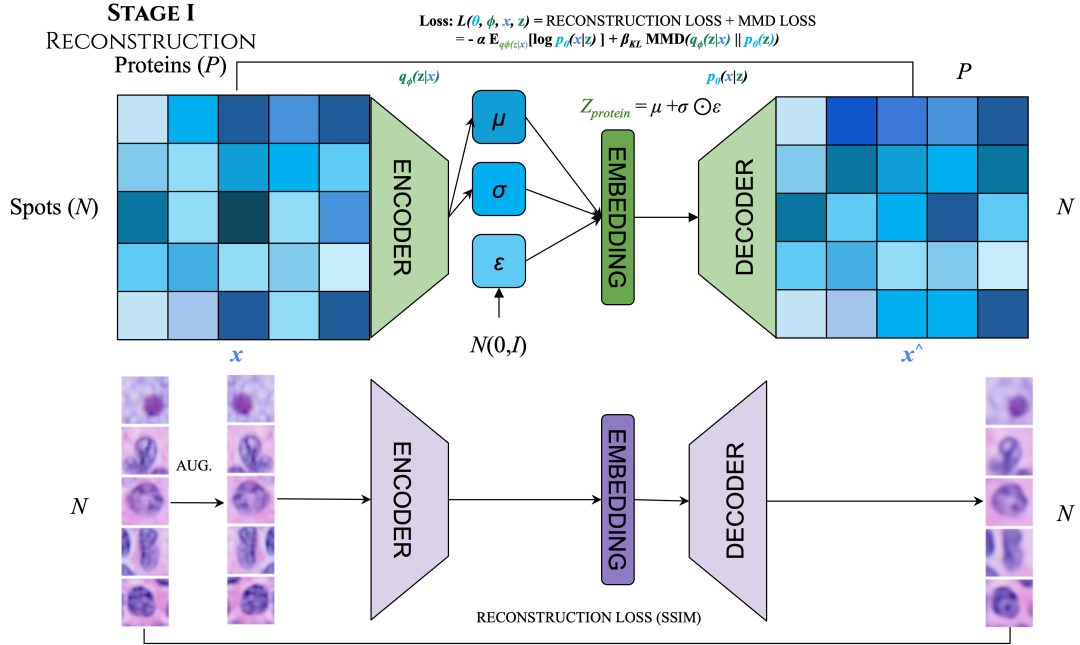


Figure 2.1: Stage 1 representation learning. Protein profiles are encoded into a variational latent space, sampled with the reparameterization trick, and decoded back to the original marker panel under reconstruction and MMD regularization. In parallel, augmented H&E tiles are reconstructed through an image encoder-decoder, giving the model a compact morphology representation before cross-modal alignment.

The experiments used a radial basis function kernel. MMD regularization encourages a smooth, continuous protein latent space. This is useful for cross-modal prediction because the image-derived latent vector produced in Stage 2 may not coincide exactly with an observed protein latent vector. Nevertheless, the decoder should behave smoothly for nearby latent positions.

The Stage 1 objective can be summarized as

$$\min \sum_i [\mathcal{L}_{\text{img}}\{X_i, D_h(E_h(X_i))\} + \mathcal{L}_{\text{prot}}\{Y_i, D_p(Z_i)\} + \beta \text{MMD}^2(q_\phi, p),]$$

where E_h, D_h denote the image encoder and decoder, D_p denotes the protein decoder, and $\beta = 10^{-3}$.

2.4.3 Stage 2: Cross-Modal Alignment

In Stage 2, the model learns a map from image latent space to protein latent space. Let A_θ denote the alignment network. The predicted protein profile is

$$\hat{Y}_i = D_p\{A_\theta(E_h(X_i))\}.$$

The protein decoder learned in Stage 1 is frozen and reused, so the aligner is trained to enter a protein latent space that already encodes multivariate marker structure. The Stage 2 objective is MSE-based:

$$\min_{\theta} \sum_i \frac{1}{P} \sum_{j=1}^P [Y_{ij} - D_p\{A_\theta(E_h(X_i))\}_j]^2.$$

For the main proposed model, the image encoder was unfrozen during this alignment stage to adapt the visual representations for cross-modal protein prediction, while the image decoder was not used for prediction (Figure 2.2). This fine-tuning is intended to adapt the visual representation away from generic image reconstruction and toward

protein-relevant morphology. The aligner was a three-layer MLP that maps the 128-dimensional image latent to the 128-dimensional protein latent, with 1,024 hidden units and 50% dropout. To mitigate the risk of overfitting, fine-tuning is paired with several regularization strategies. The network was trained for up to 500 epochs with a batch size of 64, learning rate 10^{-4} , weight decay 10^{-5} , and early stopping with patience 15 and minimum improvement 10^{-4} . These settings were selected during model development and then applied symmetrically to the forward and reverse transfer directions.

2.4.4 Augmentation and Regularization

The augmentation strategy encodes the assumption that the spot-level protein label should be invariant to simple geometric transformations of the local tile. Rotations (90° , 180° , 270°) and horizontal or vertical flips alter image orientation but should not alter the underlying protein expression measured at the spot. Additional appearance perturbations, including sharpness adjustment and auto-contrast transformations, encourage the image encoder to rely less on staining artifacts and more on stable morphologic structure. During Stage 1 training, the augmentation configuration used exhaustive flips and right-angle rotations plus up to 20 random appearance perturbations per tile. During Stage 2 alignment, the augmented model used the original tile plus right-angle rotations. Augmentation is therefore both a regularizer and an explicit inductive bias about the variation in histology images.

Dropout was used during training (set to 50% in the alignment network), and Monte Carlo dropout was used at inference to estimate predictive uncertainty (Gal and Ghahramani, 2016). Specifically, the model was evaluated repeatedly for 100 inference iterations with a 15% dropout rate active, producing an empirical distribution of predictions for each spot. These stochastic predictions provide an approximate measure of epistemic uncertainty. At inference, only the image encoder, alignment network, protein embedding layer, and protein decoder are required; measured protein expression is not used for the

held-out section (Figure 2.3). Uncertainty-filtered summaries were generated separately from the unfiltered validation predictions. Unless explicitly labeled as filtered, reported performance should be interpreted as the unfiltered validation evaluation.

2.5 DeepPT Baseline

DeepPT was used as the main direct-prediction comparison because it performed best in a recent benchmarking study out of 11 methods (Wang et al., 2025). In this thesis, the DeepPT-style model was adapted from transcriptomic prediction to spot-level protein prediction.

The baseline embodies a different inductive bias from the proposed method. It assumes that a frozen pretrained histology encoder optimized for reconstruction provides sufficiently informative visual features, and that protein expression can be learned by direct regression from a compressed feature representation. Unlike the proposed method in this thesis, it does not explicitly learn a protein latent decoder that constrains predictions to the empirical protein manifold.

For each tile, a frozen ResNet-50 produced a high-dimensional image feature vector. A feature autoencoder compressed this representation into a lower-dimensional bottleneck and reconstructed the original feature vector. A regression multilayer perceptron then predicted the 33-dimensional standardized protein vector from the bottleneck. The feature reconstruction stage regularizes the bottleneck to preserve generic histology information and remove redundancy, while the regression stage learns the protein mapping.

DeepPT was evaluated in both transfer directions. The forward DeepPT run trained on the standard-panel section and evaluated on the add-on-antibody section, while the reverse DeepPT run trained on the add-on-antibody section and evaluated on the standard-panel section. This mirrors the forward and reverse evaluation used for the proposed model and makes the benchmark symmetric across tissue sections.

STAGE II ALIGNMENT

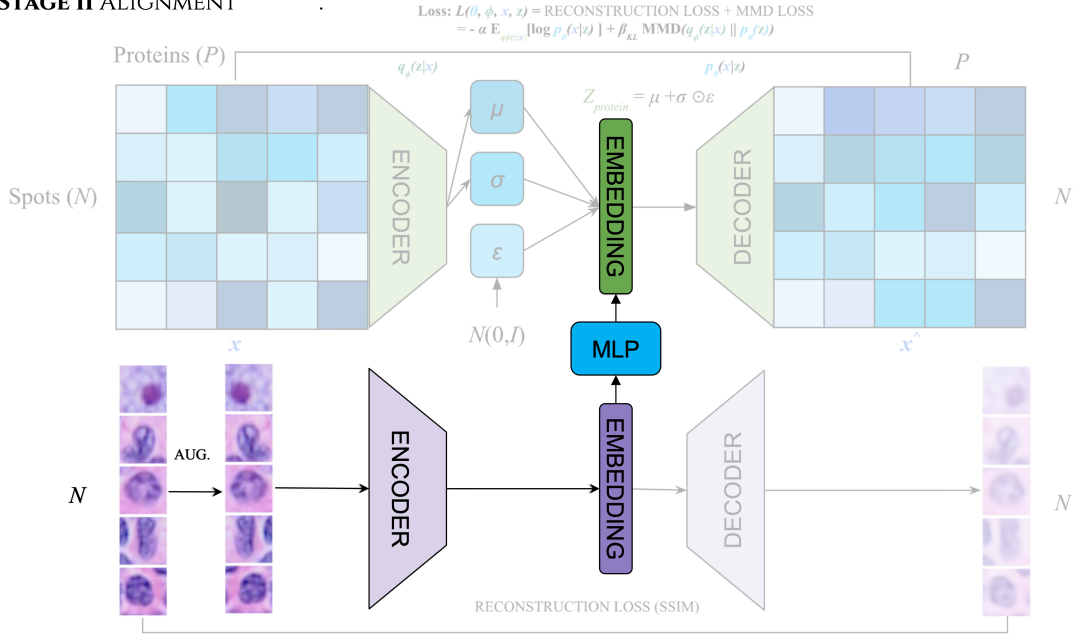


Figure 2.2: Stage 2 cross-modal alignment. The pretrained protein decoder is reused, while an MLP maps image embeddings into the protein latent space. The image encoder may be fine-tuned during this stage, but the prediction objective is protein reconstruction from image-derived latent coordinates rather than image reconstruction.

INFERENCE

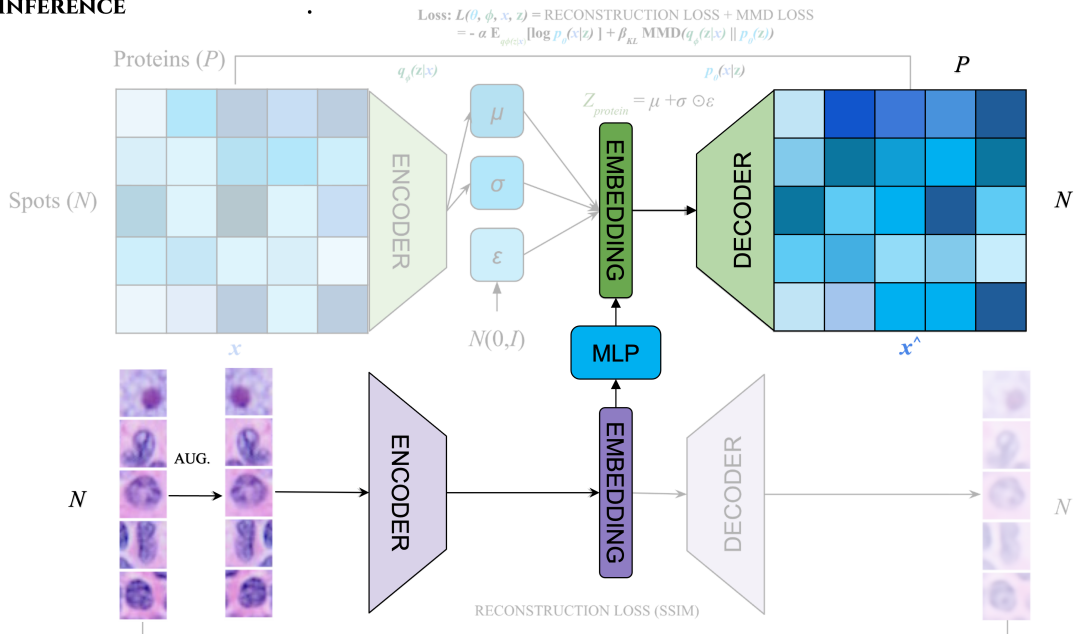


Figure 2.3: Inference workflow for a held-out tissue section. Each H&E tile is encoded, mapped through the trained alignment network into the protein latent space, and decoded into a predicted P -marker protein profile. The protein encoder and image decoder are training-time components and are not needed to generate validation-section predictions.

2.6 Embedding-Derived Shallow Baselines

In addition to DeepPT, two shallow baselines were trained on embeddings derived from the proposed model. Crucially, these embeddings were extracted from the image encoder *after* the completion of Stage 2. This means the image encoder had already engaged in supervised fine-tuning guided by the protein decoder, so the resulting embeddings are adapted specifically for cross-modal alignment rather than being purely reconstruction-optimized. These baselines ask two different questions from the ablations:

1. How much predictive signal is contained in this alignment-adapted image embedding alone?
2. Given this highly adapted image embedding, does the full latent-decoding framework improve protein prediction relative to simpler supervised models applied directly to it?

Let $H_i = E_h(X_i) \in \mathbb{R}^{128}$ denote the fully trained image embedding for spot i . The first embedding-derived baseline is a multi-output ridge regression (Hoerl and Kennard, 1970). Ridge regression estimates an intercept vector a and coefficient matrix B by minimizing

$$\sum_{i \in \mathcal{T}} \|Y_i - a - B^\top H_i\|_2^2 + \lambda \|B\|_F^2,$$

where $\mathcal{T}$ is the training section and $\lambda = 1$. This baseline tests whether the learned image embedding is already approximately linearly predictive of the protein panel. Its low variance and convex objective also make ridge regression a useful check against overstating the benefits of a more flexible neural decoder in a small-sample setting.

The second embedding-derived baseline was CatBoost gradient boosting (Prokhorenkova et al., 2018). A separate boosted-tree regressor was trained for each protein marker:

$$\hat{Y}_{ij} = g_j(H_i), \quad j = 1, \dots, P.$$

CatBoost provides a nonlinear but non-neural probe of the learned image embedding. It can capture threshold effects and interactions among embedding dimensions without using the protein variational decoder. In the experiments, each marker-specific model used squared-error regression with early stopping on the validation section. Thus, the Ridge baseline evaluates linear information in the alignment-adapted image representation, whereas the CatBoost baseline evaluates whether a flexible tabular regressor can exploit the same explicitly supervised representation without the proposed cross-modal latent-decoding structure.

Both embedding-derived baselines used the same training and validation sections as the corresponding proposed-model run. Protein targets were kept on the same standardized scale used for neural-network training during model fitting and were inverse-transformed to the CLR scale for metric calculation.

2.7 Ablation Studies

Ablation studies were designed to isolate two major components of the proposed model: augmentation and image-encoder fine-tuning during alignment. All ablations used the same whole-section transfer design, common protein panel, ResNet-50 image encoder, structural-similarity image reconstruction loss, MMD-regularized protein variational autoencoder, and protein-decoder-based alignment objective.

The first ablation removes augmentation while retaining image-encoder fine-tuning. This tests whether the main model’s performance depends on geometric and appearance-based invariance, holding the capacity to adapt image features fixed. If performance drops, it suggests that regularization through label-preserving transformations is important for out-of-section transfer.

The second ablation keeps augmentation but freezes the image encoder during alignment. This tests whether the pretrained and reconstruction-optimized image representation is

sufficient for cross-modal protein prediction. If performance drops relative to the main model, it supports the hypothesis that protein-relevant morphology requires supervised cross-modal adaptation beyond generic histology features.

The third ablation removes augmentation and freezes the image encoder. This is the most constrained variant and functions as a low-adaptation comparison. It asks how much predictive signal is present in the fixed image representation before either geometric-invariance-enforcing augmentation or supervised image fine-tuning is allowed.

A summary of the model configurations for the ablations compared to the main proposed model is provided in Table 2.1. All models use the same ResNet-50 backbone, structural-similarity reconstruction loss, and MMD-regularized protein VAE.

Model Variant	Augmentation	Img. Enc. Fine-Tuning
Main Proposed Model	Yes (Extensive)	Yes (Unfrozen in Stage 2)
Ablation 1: No Augmentation	No	Yes (Unfrozen in Stage 2)
Ablation 2: Frozen Encoder	Yes (Extensive)	No (Completely Frozen)
Ablation 3: No Aug, Frozen Enc.	No	No (Completely Frozen)

Table 2.1: Configuration summary of the main proposed model and its corresponding ablation studies.

2.8 Evaluation Metrics and Statistical Comparisons

Evaluation was performed at both the marker level and spot level. The primary evidence for spatial protein recovery is marker-wise Pearson correlation, Spearman correlation, and structural similarity (SSIM), because these metrics ask whether each protein’s spatial pattern is recovered across the held-out tissue. RMSE and MAE are interpreted as calibration and magnitude checks. Moran’s I and CMD are structural diagnostics: they test whether spatial autocorrelation and cross-marker covariance are preserved, but they do not by themselves establish accurate local recovery. In all model-comparison tables, the validation section in the dataset-wise split is reported as validation because no separate third test section exists. Results are therefore reported for the training

section, the validation section, and the pooled overall set. Metrics were computed on the inverse-transformed CLR-scale predictions written by each completed run.

Spot-level error measures whether the full multivariate protein profile is accurate for each spot:

$$\text{MSE}_i = \frac{1}{P} \sum_{j=1}^P (Y_{ij} - \hat{Y}_{ij})^2, \quad \text{RMSE}_i = \sqrt{\text{MSE}_i}.$$

Marker-level metrics evaluate whether each protein’s spatial distribution is recovered across the validation tissue. For marker j , Pearson correlation is

$$\rho_j = \frac{\sum_i (Y_{ij} - \bar{Y}_j)(\hat{Y}_{ij} - \bar{\hat{Y}}_j)}{\sqrt{\sum_i (Y_{ij} - \bar{Y}_j)^2} \sqrt{\sum_i (\hat{Y}_{ij} - \bar{\hat{Y}}_j)^2}}.$$

Spearman correlation was also computed to assess rank agreement, which is less sensitive to monotone rescaling. Additional marker-level diagnostics included mean absolute error, Lin’s concordance correlation coefficient, mutual information, one-dimensional structural similarity, earth mover’s distance, maximum mean discrepancy, Jensen-Shannon divergence, and partial correlation after controlling for a proxy of total protein abundance, $\log(1 + \sum_j Y_{ij}^+)$. Here, concordance correlation coefficient penalizes both imperfect association and systematic shifts away from the identity line; mutual information summarizes nonlinear statistical dependence; earth mover’s distance, maximum mean discrepancy, and Jensen-Shannon divergence compare predicted and observed marker distributions rather than spotwise pairing alone. Spot-level diagnostics used the same profile-comparison metrics where they are meaningful, excluding distributional distances and Moran’s I because those require a spatial or sample distribution rather than a single spot profile.

Spatial structure was evaluated with Moran’s I, a measure of spatial autocorrelation (Moran, 1950). For each split of each tissue section, an 8-nearest-neighbor graph was built from the spot pixel coordinates $(x_{\text{px}}, y_{\text{px}})$, excluding each spot’s self-neighbor. The implementation uses directed neighbor pairs from each valid spot to its eight nearest coordinate neighbors. For a marker vector x centered as $x_i - \bar{x}$ over n spots and graph

weights $w_{i\ell} = 1$ for these neighbor pairs,

$$I = \frac{n}{\sum_{i,\ell} w_{i\ell}} \frac{\sum_{i,\ell} w_{i\ell} (x_i - \bar{x})(x_\ell - \bar{x})}{\sum_i (x_i - \bar{x})^2}.$$

Positive Moran’s I indicates that nearby spots tend to have similar marker abundance; values near zero indicate little spatial autocorrelation under the chosen graph; negative values indicate local dissimilarity. The reported column “Moran’s I (true)” is the spatial autocorrelation of the measured marker map, “Moran’s I (predicted)” is the spatial autocorrelation of the model-predicted marker map, and “Moran’s I error” is the absolute difference between the two. Thus, the error is small when a model matches the overall smoothness or clustering strength of the measured map, even if it does not show that the high-expression regions occur in exactly the right locations. For the proposed model, Moran’s I was also computed for image and protein embedding dimensions to assess whether the learned latent spaces retained spatial organization.

To assess multivariate protein structure, protein-protein correlation matrices were computed from true and predicted expression. The correlation matrix distance (CMD), defined as the root mean squared difference between off-diagonal entries of the predicted and true protein correlation matrices, is calculated. This metric penalizes failure to preserve marker co-expression structure, even when individual marker correlations are acceptable.

These metrics capture complementary failure modes. RMSE and MAE penalize absolute error but may favor over-smoothed predictions. Correlation-based metrics emphasize spatial pattern recovery. Distributional distances test whether predicted and observed marker distributions agree. Moran’s I and CMD test spatial and cross-marker structure. The results chapter therefore treats marker-wise association and SSIM as the main evidence for virtual spatial protein recovery, while using pointwise error, distributional distances, Moran’s I, and CMD to diagnose calibration, smoothing, and structural failure modes.

Metrics such as activation false-positive rate and Dice score are appropriate when the target is a binary or pixel-level virtual immunofluorescence map, as in GigaTIME-style virtual mIF translation (Valanarasu et al., 2026). They are less directly suited to the present CytAssist setting because the prediction target is a continuous spot-level antibody abundance vector rather than a registered pixel mask.

For model comparisons, each marker was treated as a paired unit for per-protein tests, and each spot was treated as a paired unit for per-spot tests. Metric differences between the proposed model and each baseline or ablation were tested using the Wilcoxon signed-rank test (Wilcoxon, 1945). Resulting p -values were adjusted with the Benjamini-Hochberg false-discovery-rate procedure (Benjamini and Hochberg, 1995). This comparison strategy reflects the fact that models are evaluated on the same validation markers and spots and controls the expected false-discovery rate across repeated model comparisons.

CHAPTER 3

Results

3.1 Whole-Section Transfer Benchmark

The central experiment asked whether protein expression can be inferred from local H&E morphology in a tissue section that was not used for training. All models were evaluated on the same 33 shared antibody targets in two transfer directions using two public CytAssist human tonsil sections. In the forward direction, 4,194 tissue-covered spots were used for training and 4,908 spots were held out for validation. In the reverse direction, the roles of the two sections were exchanged, producing 4,908 training spots and 4,194 validation spots. Pooled validation summaries therefore reflect 9,102 spot-level predictions and 66 marker-direction measurements. This benchmark is stricter than a random within-section spot split, but it is still a two-section proof-of-concept rather than an external multi-cohort validation.

The analysis included the proposed latent-alignment model, DeepPT, two supervised probes trained on the proposed image representation, and three ablations that isolate augmentation and image-encoder fine-tuning. Throughout this chapter, “validation” refers to transfer into the held-out section rather than to a random split of spots within a section. Because no separate third section was available, these validation results are also the final

reported held-out results.

3.2 Overall Predictive Performance

Following the metric hierarchy defined in Chapter 2, marker-wise Pearson, Spearman, and SSIM are treated as the primary evidence for spatial protein recovery. RMSE, MAE, distributional distances, Moran’s I, and CMD are interpreted as diagnostics of calibration, smoothing, spatial autocorrelation, and cross-marker structure. Under this hierarchy, the proposed model improved the recovery of marker-level spatial patterns relative to DeepPT, although it did not minimize every error metric (Table 3.1 and Figure 3.1). Averaged across both transfer directions, marker Pearson correlation increased from 0.180 for DeepPT to 0.255 for the proposed model, marker Spearman correlation increased from 0.172 to 0.258, and marker SSIM increased from 0.053 to 0.106. These gains are most relevant to virtual spatial proteomics because they measure whether the model reconstructs protein-specific spatial variation across the tissue. Expanded model-level summaries for all train, validation, and pooled splits are provided in Supplementary Table A.3.

Pointwise error told a more conservative story. DeepPT had slightly lower pooled spot RMSE than the proposed model (0.725 versus 0.740), and the CatBoost probe achieved the lowest spot RMSE among all evaluated models (0.687). The ridge probe achieved the highest pooled marker Pearson, Spearman, and SSIM. This result is important because the probes were trained on the proposed image embedding: they are not independent biological baselines, but they show that the learned representation contains substantial protein-predictive information that can be extracted by simpler supervised heads.

The forward-validation diagnostic panels provide a more granular view of this tradeoff (Figures 3.2, 3.3, and 3.4). Marker-level association metrics separated the proposed model from DeepPT most clearly, while the no-augmentation fine-tuned ablation, ridge

Model	Prot. Pearson $\uparrow$	Prot. Spearman $\uparrow$	Prot. SSIM $\uparrow$	Prot. RMSE $\downarrow$	Spot RMSE $\downarrow$	Spot MAE $\downarrow$
Proposed	0.255	0.258	0.106	0.710	0.740	0.588
DeepPT	0.180	0.172	0.053	0.688	0.725	0.569
CatBoost	0.260	0.268	0.083	0.660	0.687	0.537
Ridge	0.280	0.304	0.142	0.747	0.774	0.615
No aug. + tune	0.191	0.238	0.090	0.729	0.740	0.591
2x aug. + frozen enc.	0.121	0.126	0.013	0.704	0.744	0.587
No aug. + frozen enc.	0.100	0.144	0.022	0.705	0.742	0.589

Table 3.1: Pooled validation performance across forward and reverse whole-section transfer. Marker-level metrics are averaged across 66 marker-direction pairs; spot-level metrics are averaged across 9,102 held-out spot profiles. Higher values are better for Pearson, Spearman, and SSIM; lower values are better for RMSE and MAE.

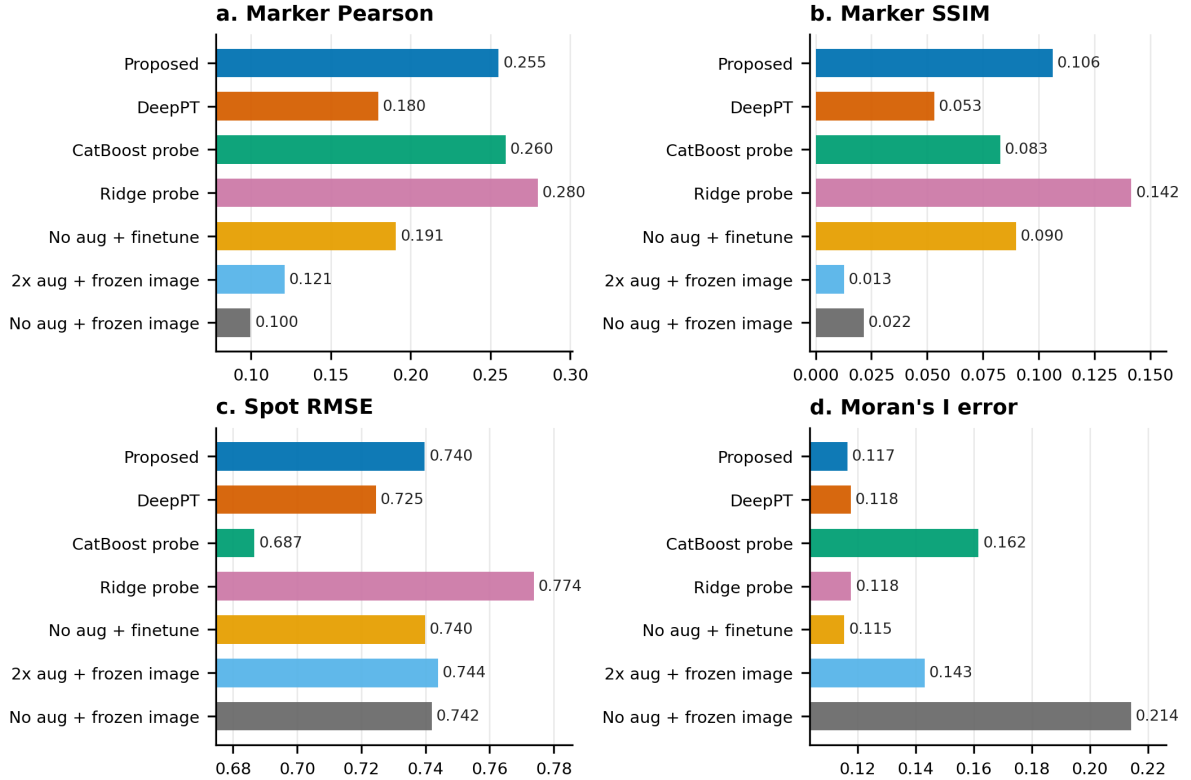


Figure 3.1: Validation performance across the two transfer directions. The proposed model improved marker-level association relative to DeepPT, whereas shallow probes of the learned image representation remained competitive and sometimes superior for pointwise error.

probe, and CatBoost probe remained competitive on some metrics. This pattern suggests that the proposed training strategy improves protein-specific spatial ordering, but the learned image representation also contains signal that simpler prediction heads can exploit. Complete held-out per-protein validation metrics for every compared model are reported in Supplementary Table A.6; the proposed model’s train, validation, and pooled per-protein metrics are reported separately by direction in Supplementary Tables A.7 and A.8.

Spot-level summaries gave a different, less diagnostic view of performance (Figures 3.5, 3.6, and 3.7). Because these metrics compare a 33-marker vector at each spot, shared panel-wide abundance structure can yield high spot Pearson, Spearman, and concordance even when individual marker maps are imperfect. The spot-level panels therefore support using marker-wise spatial metrics as the primary evidence for virtual spatial proteomics, with spot error metrics treated as calibration and magnitude checks.

3.3 Generalization Was Direction-Dependent

The pooled numbers obscure a strong asymmetry between transfer directions (Table 3.2 and Figure 3.8). In the forward validation section, the proposed model achieved marker Pearson 0.344, marker Spearman 0.345, and spot Pearson 0.912. DeepPT reached 0.184, 0.174, and 0.899 on the same metrics. Paired marker-level Wilcoxon tests confirmed that the proposed model improved forward marker Pearson ($q = 6.0 \times 10^{-8}$), Spearman ($q = 7.0 \times 10^{-9}$), and SSIM ($q = 1.4 \times 10^{-3}$) relative to DeepPT.

The reverse direction did not show the same advantage. Marker Pearson was 0.166 for the proposed model and 0.175 for DeepPT, while spot Pearson was 0.854 and 0.880, respectively. The reverse marker-level differences in Pearson, Spearman, and SSIM were not significant after false-discovery-rate correction. In contrast, reverse RMSE and MAE significantly favored DeepPT ($q = 4.1 \times 10^{-3}$ and $q = 7.7 \times 10^{-3}$, respectively). These results show that the proposed representation can improve morphology-to-protein transfer,

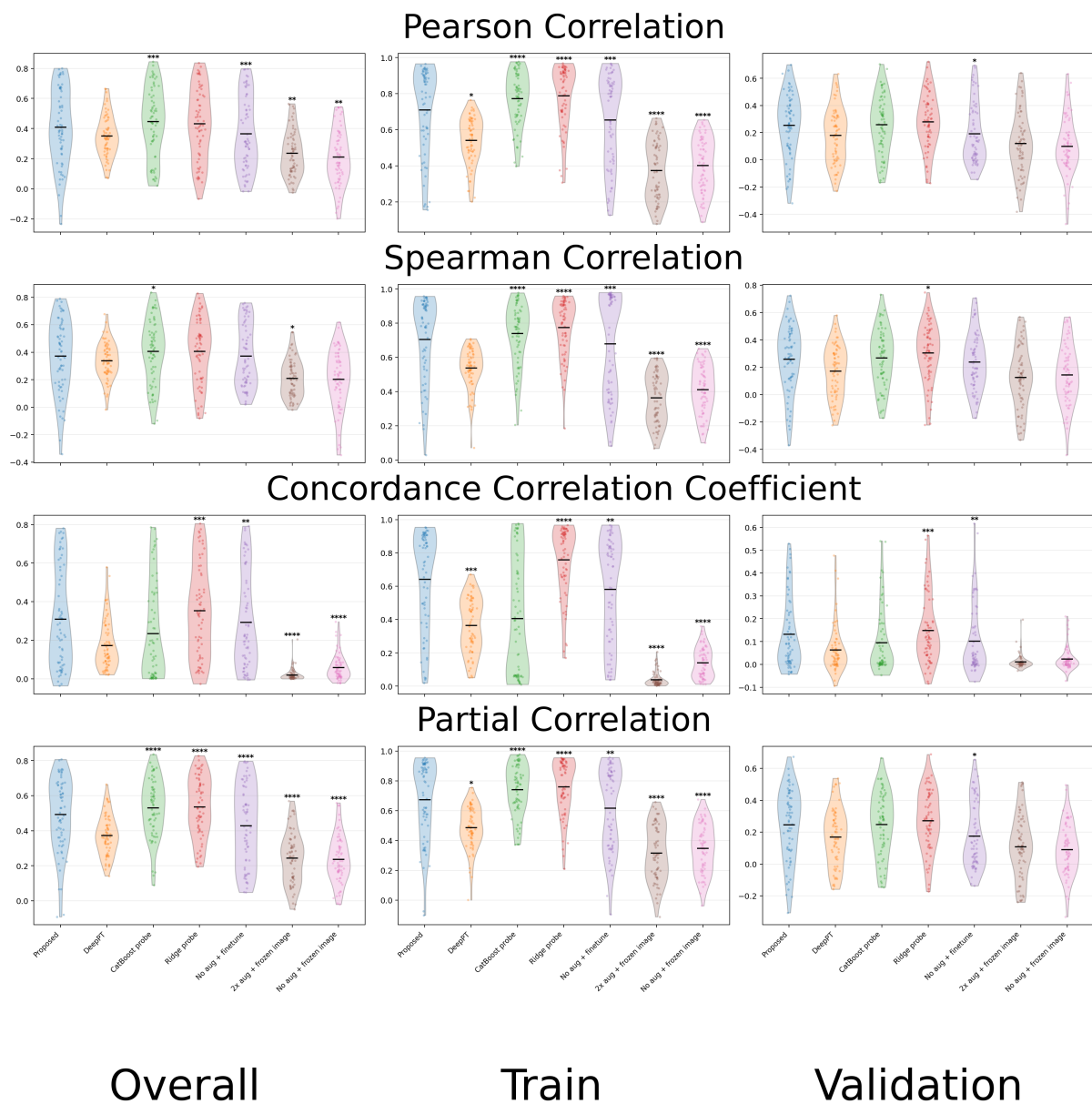
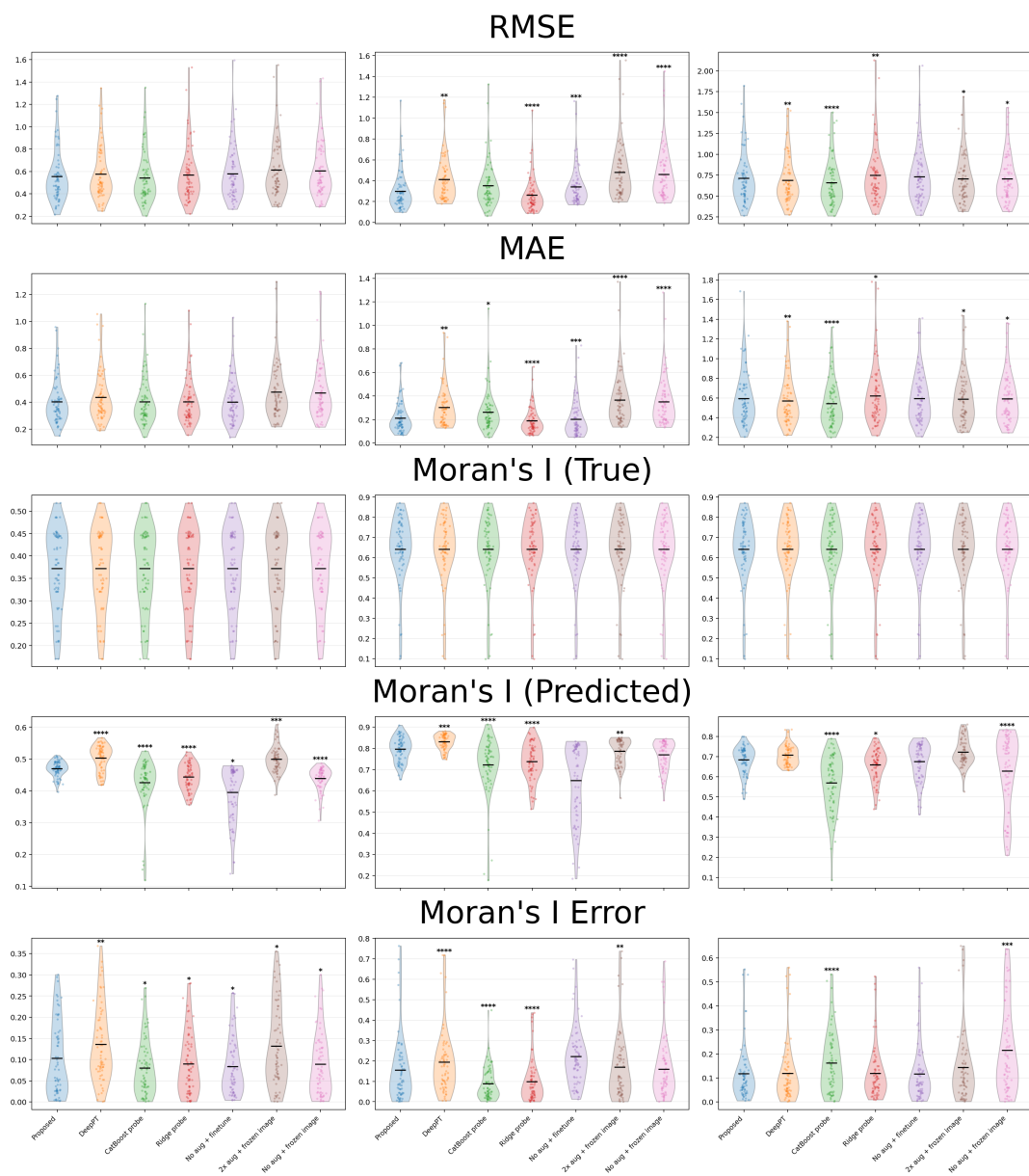


Figure 3.2: Forward-validation marker-level association metrics. Each point is a protein, summarized across the overall, training, and validation partitions for Pearson correlation, Spearman correlation, concordance correlation coefficient, and partial correlation. The proposed model improved held-out marker association relative to DeepPT, although linear and tree-based probes remained competitive for several association summaries.



Overall

Train

Validation

Figure 3.3: Forward-validation marker-level error and spatial-autocorrelation metrics. RMSE and MAE quantify pointwise error for each protein map, while the Moran's-I panels compare spatial autocorrelation in the measured, predicted, and absolute-difference summaries. These diagnostics show that lower error and better spatial association were not always achieved by the same model.

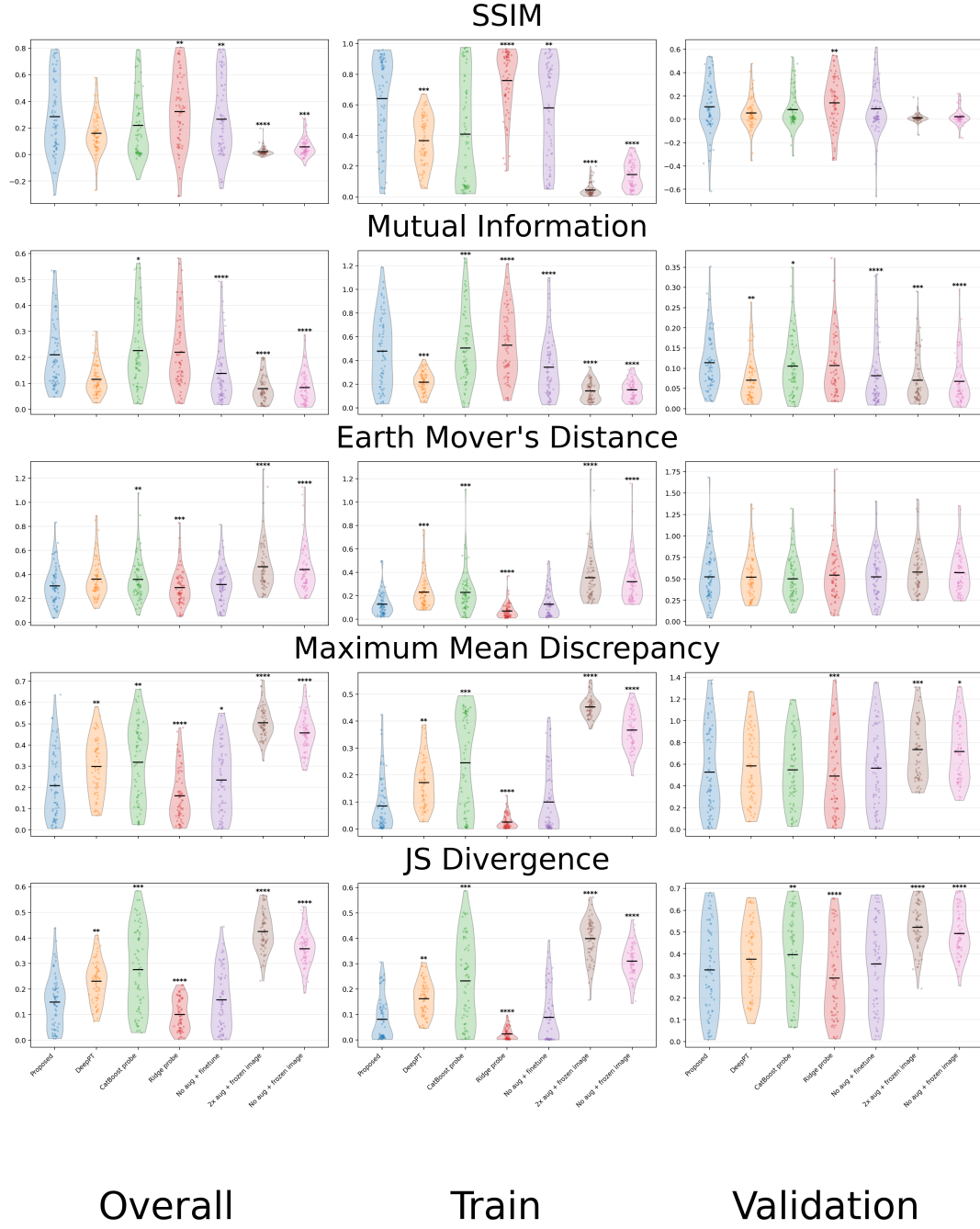


Figure 3.4: Forward-validation marker-level structural and distributional metrics. SSIM and mutual information measure spatial and nonlinear agreement between measured and predicted marker maps, whereas EMD, MMD, and Jensen-Shannon divergence summarize distributional mismatch. The proposed model was strongest against DeepPT on spatial structure, but the distributional metrics emphasize that calibration and dynamic-range recovery remained incomplete.

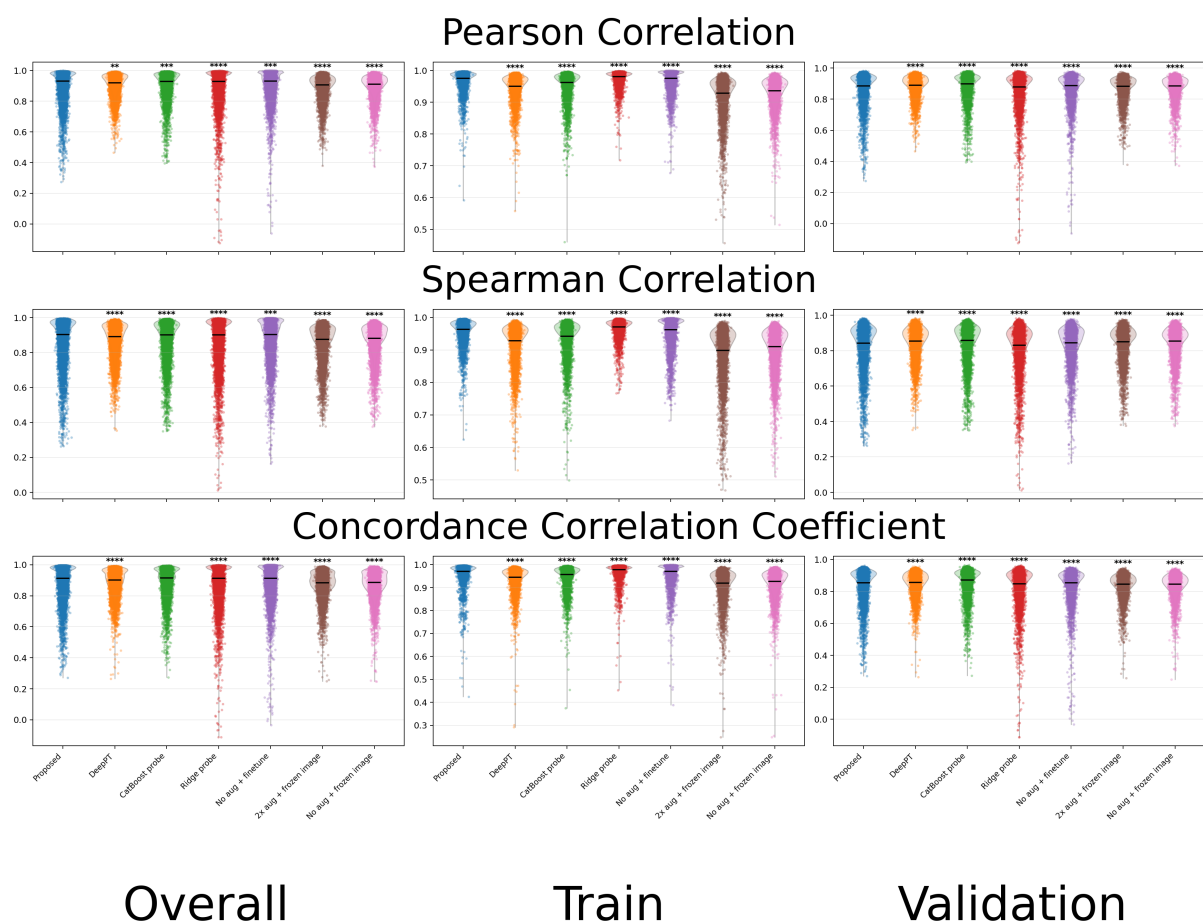


Figure 3.5: Forward-validation spot-level association metrics. Each point is a spot-level protein profile. Correlations were high for many models, indicating that spot-wise association is less discriminating than marker-wise spatial recovery in this benchmark.

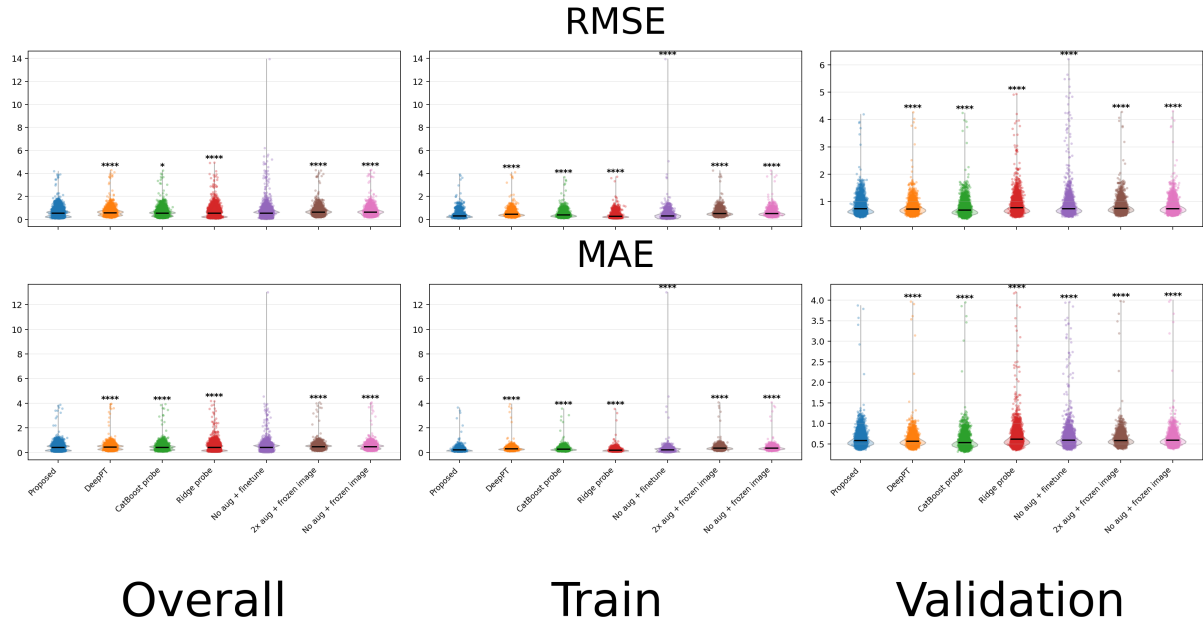


Figure 3.6: Forward-validation spot-level pointwise error. RMSE and MAE summarize the magnitude of errors across the full antibody profile at each spot. These metrics favor models that are well calibrated in abundance, which is related to but distinct from recovering marker-specific spatial patterning.

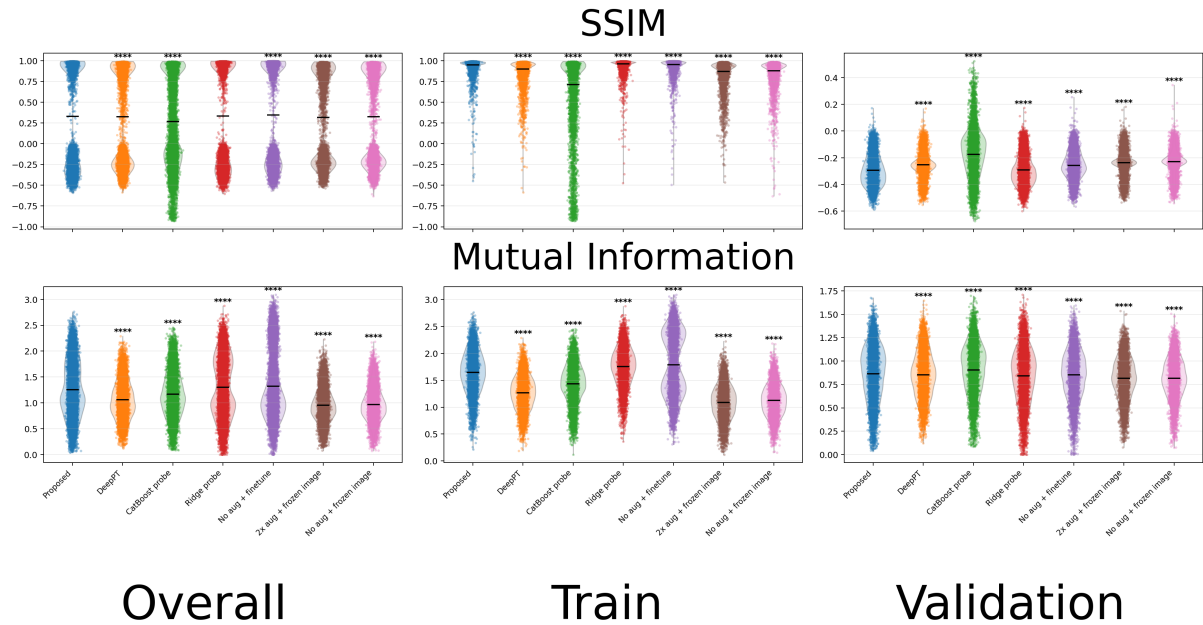


Figure 3.7: Forward-validation spot-level structural metrics. Spot-level SSIM and mutual information summarize agreement across the marker vector at each spatial location, reinforcing that high spot-profile agreement does not by itself guarantee accurate spatial maps for individual proteins.

but that the learned relationship is not yet symmetric across sections.

The marker scatter grids make the asymmetry visible at the protein level (Figures 3.9 and 3.10). In the forward direction, many immune and structural markers retained a monotone relationship between measured and predicted abundance, even though the validation predictions were compressed toward the center of the training range. In the reverse direction, several markers collapsed into narrower predicted bands or section-specific offsets. Thus, the reverse failure mode was not simply increased random noise; it was also loss of calibration and dynamic range under a harder section transfer.

Direction	Model	Prot. Spearman $\uparrow$	Prot. SSIM $\uparrow$	Prot. RMSE $\downarrow$	Spot Pearson $\uparrow$	Spot RMSE $\downarrow$
Forward	Proposed	0.345	0.186	0.664	0.912	0.689
Forward	DeepPT	0.174	0.067	0.680	0.899	0.724
Reverse	Proposed	0.171	0.026	0.756	0.854	0.799
Reverse	DeepPT	0.171	0.040	0.696	0.880	0.725

Table 3.2: Direction-specific validation performance for the proposed model and DeepPT. The forward direction trains on the first section and validates on the second; the reverse direction swaps these roles.

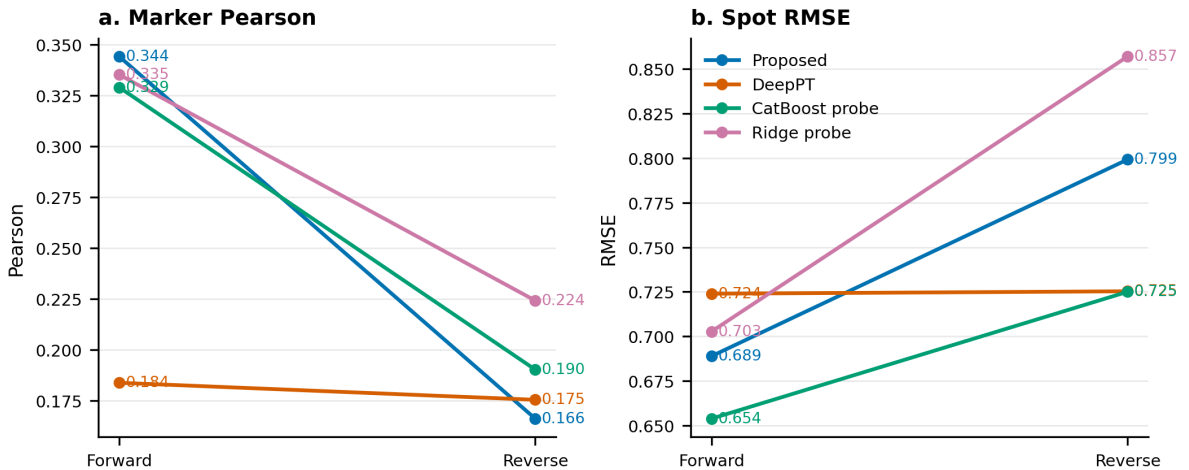


Figure 3.8: Transfer-direction effects for the principal model comparisons. The proposed model was strongest in the forward direction, whereas the reverse section reduced marker-level association and increased pointwise error.

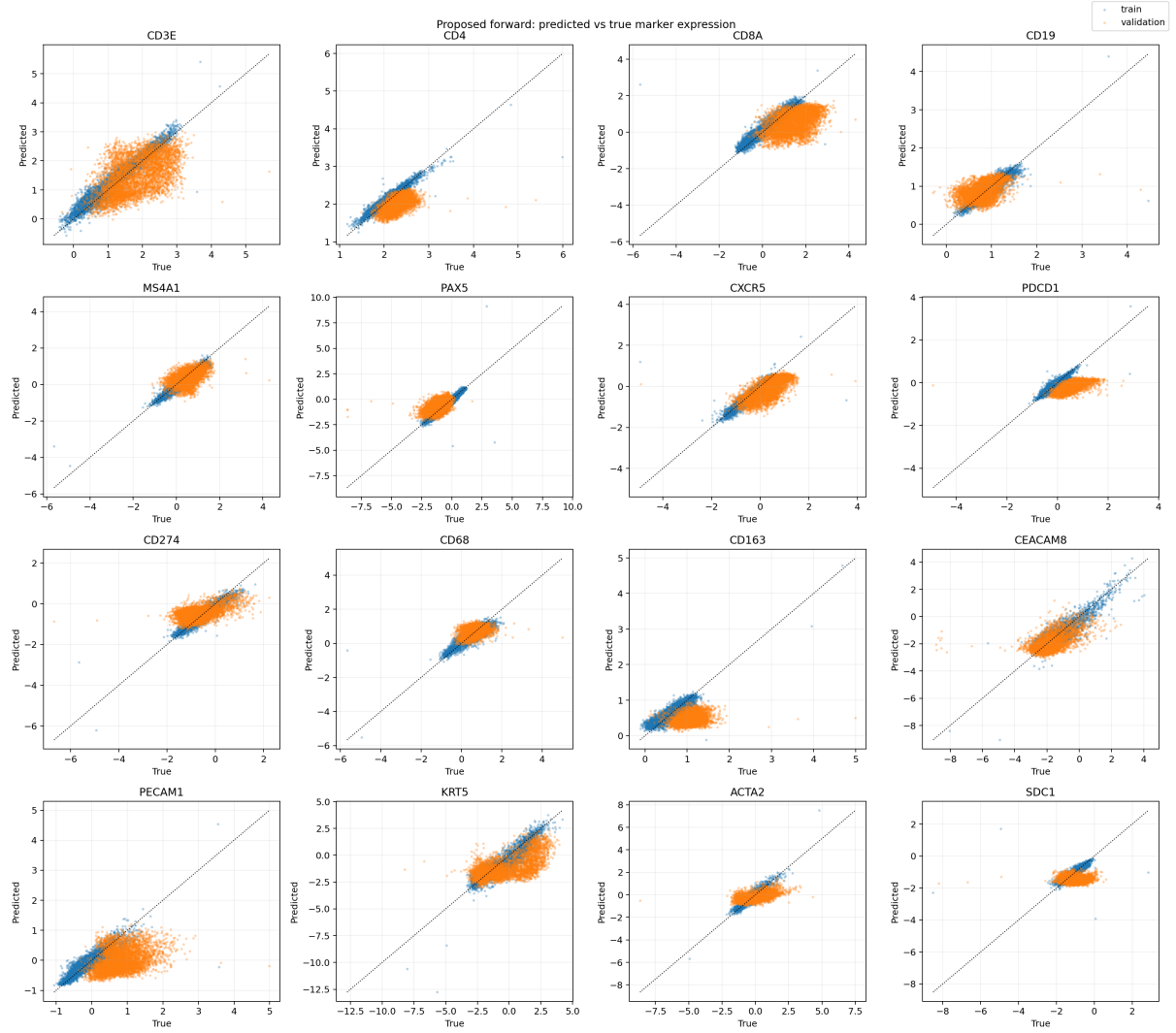


Figure 3.9: Forward-direction marker scatter plots for the proposed model. Blue points show the training section and orange points show the held-out validation section. Many markers retain a positive validation relationship with measured abundance, but the prediction range is visibly compressed for several targets.

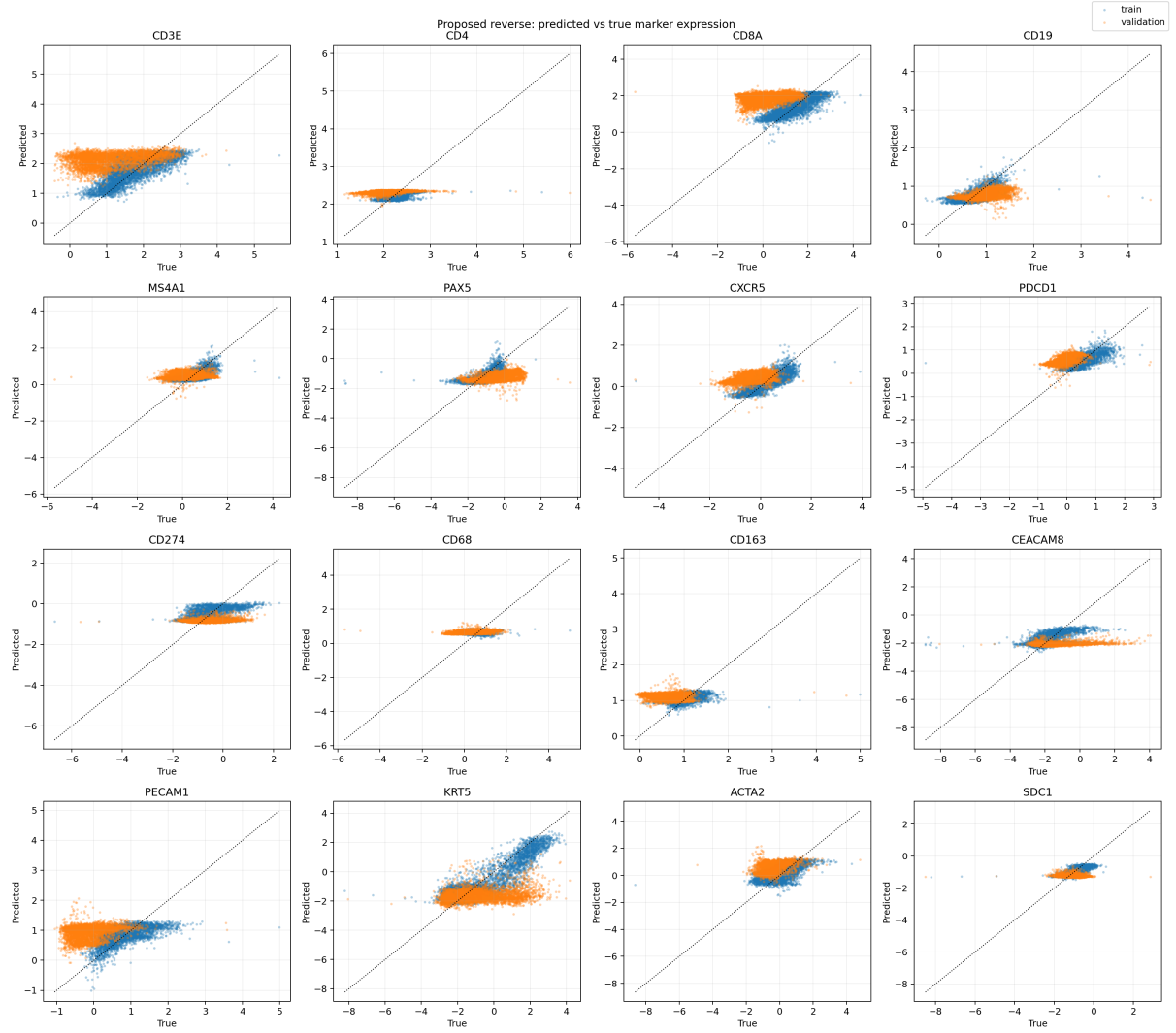


Figure 3.10: Reverse-direction marker scatter plots for the proposed model. The validation points show weaker calibration and stronger dynamic-range compression than in the forward direction, explaining why pooled marker-level gains did not carry over symmetrically across sections.

3.4 Ablations and Representation Probes

The ablations clarify which parts of the proposed model contributed to generalization. Freezing the image encoder during alignment sharply reduced marker-level performance, even when augmentation was retained. The pooled marker Pearson values for the frozen-image variants were 0.121 with augmentation and 0.100 without augmentation, compared with 0.255 for the proposed model. This suggests that protein-relevant morphology was not fully captured by the fixed image representation; supervised adaptation of the image encoder was important.

Removing augmentation while retaining image-encoder fine-tuning produced a mixed result. In the forward direction, the no-augmentation fine-tuned ablation was nearly tied with the proposed model for marker Pearson (0.347 versus 0.344). In the reverse direction, however, it fell to 0.035, well below the proposed model’s 0.166. Augmentation therefore appears less important for the easier forward transfer, but important for robustness when the direction is reversed. This pattern is consistent with augmentation acting as a regularizer against section-specific visual shortcuts.

The ridge and CatBoost probes indicate that the proposed image embedding itself is informative. Ridge achieved the best pooled marker Pearson and SSIM, whereas CatBoost achieved the best pooled spot RMSE. The full latent-alignment model therefore should not be interpreted simply as the best scalar regressor in this experiment. Its stronger claim is that it learns a morphology representation and protein-constrained prediction pathway that improves spatial marker recovery over DeepPT, while leaving room for simpler heads to exploit the same representation.

The expanded metric panels also make the ablation result more specific. Frozen image-encoder variants were weak on marker association and structural map agreement, indicating that the Stage 1 image representation was not sufficient on its own. The no-augmentation fine-tuned model was much closer to the proposed model in the forward

direction, but its reverse-direction collapse indicates that augmentation primarily improved robustness to section shift rather than merely raising the easiest validation score.

3.5 Marker-Specific Predictability and Case Studies

Predictability varied substantially across the antibody panel (Table 3.3 and Figure 3.11). The marker rankings were metric-dependent, but the most predictable proteins were not a random subset of the panel. They included checkpoint markers (PDCD1 and CD274), B-cell and lymphoid markers (CXCR5, PAX5, CR2, and CD19), T-cell markers (CD3E, CD4, and CD8A), epithelial or stromal markers (KRT5 and ACTA2), and the granulocyte marker CEACAM8. These targets plausibly correspond to tissue compartments, immune aggregates, epithelial structures, stromal regions, or inflammatory architecture that leave visible signatures in H&E.

Protein	RMSE ↓	SSIM ↑	Spearman ↑	Pearson ↑
PDCD1	0.740	-0.299	0.653	0.597
CXCR5	0.741	-0.442	0.587	0.551
PAX5	0.908	0.308	0.505	0.466
KRT5	1.529	0.267	0.361	0.435
CR2	0.664	0.251	0.454	0.434
ACTA2	0.981	-0.307	0.451	0.431
CD4	0.364	0.298	0.476	0.431
CD3E	0.973	0.333	0.419	0.431
CD19	0.297	0.347	0.457	0.427
CD8A	1.471	0.177	0.412	0.413
CEACAM8	0.892	0.294	0.300	0.408
CD274	0.527	0.237	0.325	0.396

Table 3.3: Top pooled validation proteins for the proposed model, ordered by Pearson correlation. Values are averaged across the two transfer directions.

The spatial maps show the same pattern qualitatively (Figure 3.12). For CD3E and KRT5, the proposed model recovered broad regions of enrichment rather than independent spot-level noise. At the same time, the predictions were smoother than the measured expression and often compressed the dynamic range. This is an important limitation: the

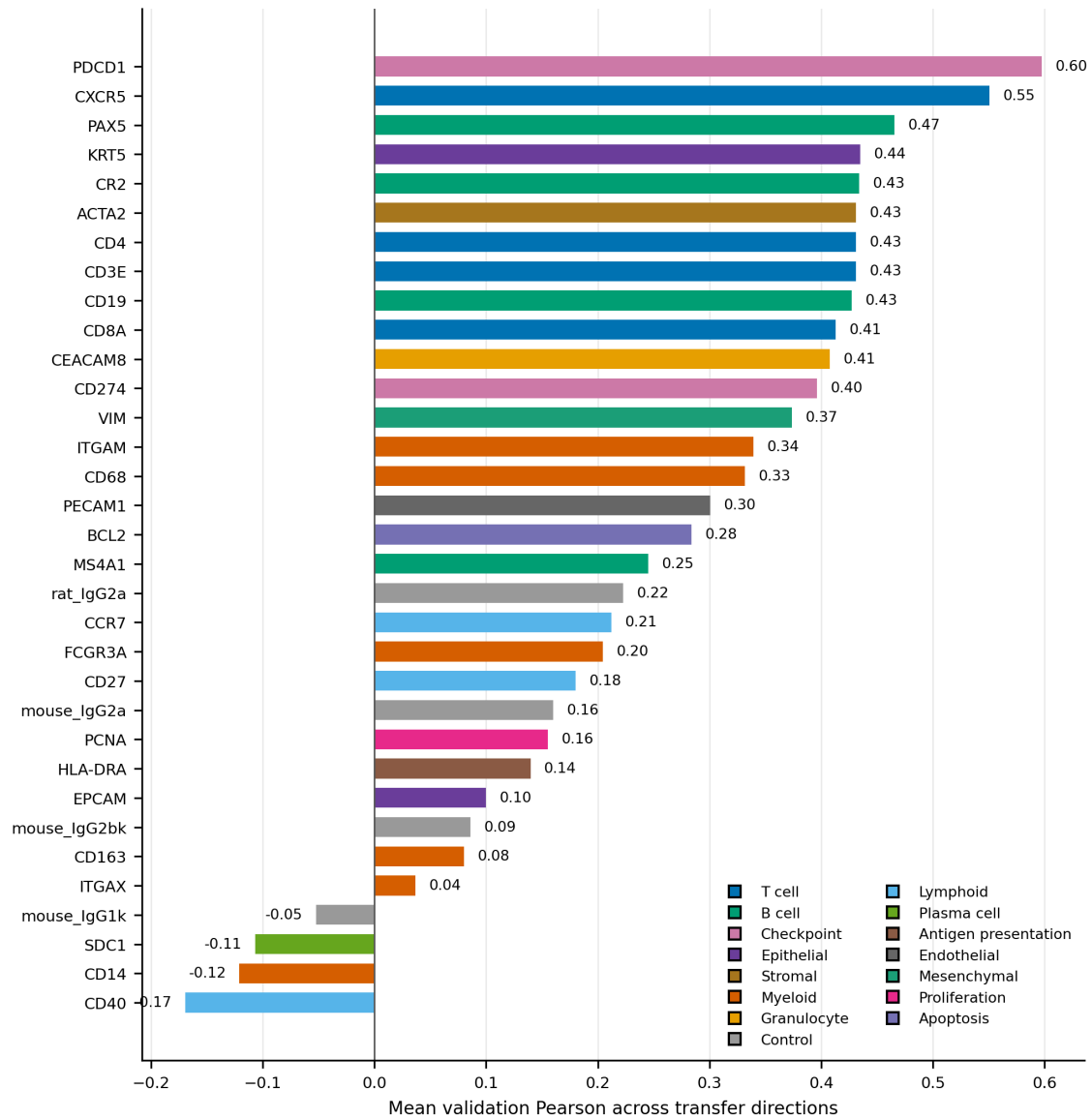


Figure 3.11: Marker-level heterogeneity in the proposed model. Pearson correlations are averaged across forward and reverse validation sections. Proteins associated with visible tissue organization tended to be more predictable, while several controls and markers with weaker morphologic correlates were near zero or negative.

model is better understood as a morphology-informed spatial estimate than as a substitute for measured antibody abundance.

Marker-specific case studies further illustrate why pooled metrics should be interpreted cautiously (Figures 3.13, 3.14, 3.15, and 3.16). KRT5, CD19, PDCD1, and MS4A1 represent different forms of morphology-linked signal: epithelial structure, B-cell organization, checkpoint-associated immune context, and a second B-cell marker with stronger directional instability. Because these diagnostic panels show both training and validation points, the printed panel metrics should be read as case-study summaries; the validation-only aggregates in Table 3.3 remain the primary quantitative benchmark.

KRT5 was the clearest structural success case. The predicted maps recovered broad epithelial enrichment patterns, especially in the forward direction, while still smoothing local extremes and compressing abundance. KRT5 is therefore best interpreted here as a basal epithelial or cytokeratin structural marker, rather than as pan-cytokeratin. This supports the broader conclusion that proteins tied to visible tissue architecture are among the most recoverable from H&E.

Immune markers showed a more graded pattern. CD19 retained moderate spatial association, indicating that broad B-cell-rich regions can sometimes be inferred from morphology. However, predicted CD19 maps were smoother than measured maps and missed sharper local hotspots and low-expression pockets. MS4A1 showed a stronger direction-specific failure. In the forward validation section, MS4A1 was recovered with moderate association and spatial agreement (Pearson 0.437, Spearman 0.444, concordance 0.403, RMSE 0.436, mutual information 0.159, and SSIM 0.410). In the reverse direction, marker-specific structure collapsed (Pearson 0.054, Spearman 0.055, concordance 0.026, RMSE 0.596, mutual information 0.066, and SSIM 0.026). Together, CD19 and MS4A1 show that markers from a similar immune lineage can differ substantially in transfer robustness.

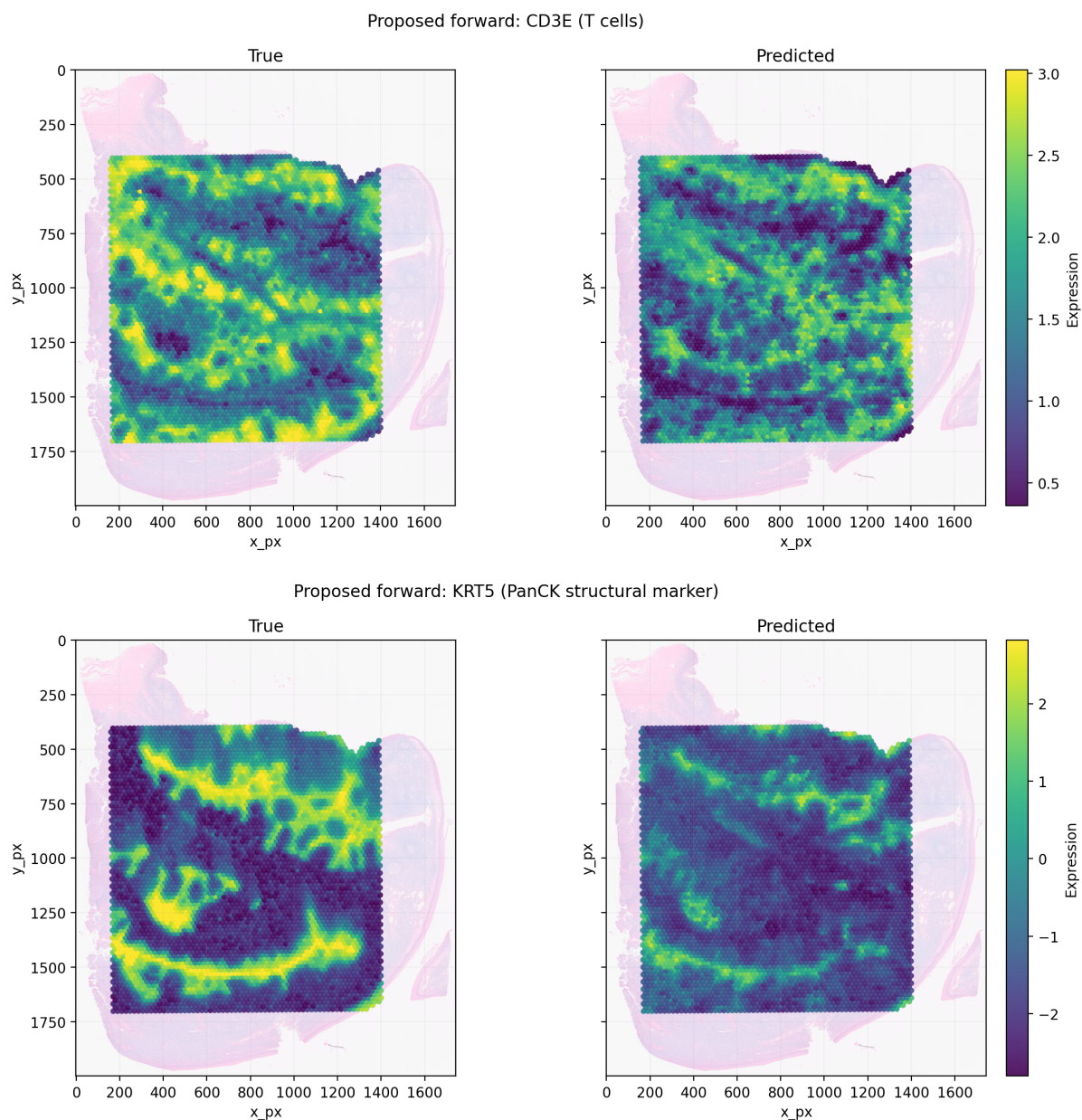


Figure 3.12: Representative forward validation maps for CD3E and KRT5. Each row compares measured and predicted expression over the held-out section using a shared color scale for that marker. The model captures coarse spatial structure but smooths local variation and under-recovers some high-expression regions.

PDCD1 provided a complementary checkpoint-marker example. Its recovery suggests that checkpoint-associated immune architecture can be partially inferred from H&E, but this should not be interpreted as direct visualization of PD-1 expression. More plausibly, the model exploits correlated tissue context such as immune-rich regions, lymphoid organization, and inflammatory architecture. Across all marker examples, predictions were better understood as morphology-informed spatial estimates than as calibrated substitutes for measured antibody abundance: high-expression regions were often attenuated, local variation was smoothed, and calibration varied by marker and transfer direction.

3.6 Spatial and Multivariate Structure

Moran’s-I summaries showed that the proposed model and DeepPT were nearly tied when averaged across both directions: pooled absolute Moran’s-I error was 0.117 for the proposed model and 0.118 for DeepPT. This indicates that both models captured a similar amount of average spatial autocorrelation, even though the proposed model produced stronger marker-wise correlations in the forward direction. The validation Moran’s-I values for measured and predicted maps are listed by marker in Supplementary Table A.5.

Spatial clustering of the predicted protein profiles confirmed that the models learned organized tissue partitions rather than independent spot noise (Figures 3.17 and 3.18). However, silhouette values should not be read as direct accuracy scores. Some ablated models produced very high silhouette values because their predictions formed sharply separated clusters, even when the marker-level maps and scatter plots showed poorer biological recovery. These maps are therefore best interpreted as qualitative checks on tissue organization, complementing the marker-level metrics above.

Protein-protein correlation preservation was more direction-sensitive. In the forward validation section, the correlation matrix distance was 0.382 for the proposed model and 0.581 for DeepPT, indicating better preservation of cross-marker structure by the proposed

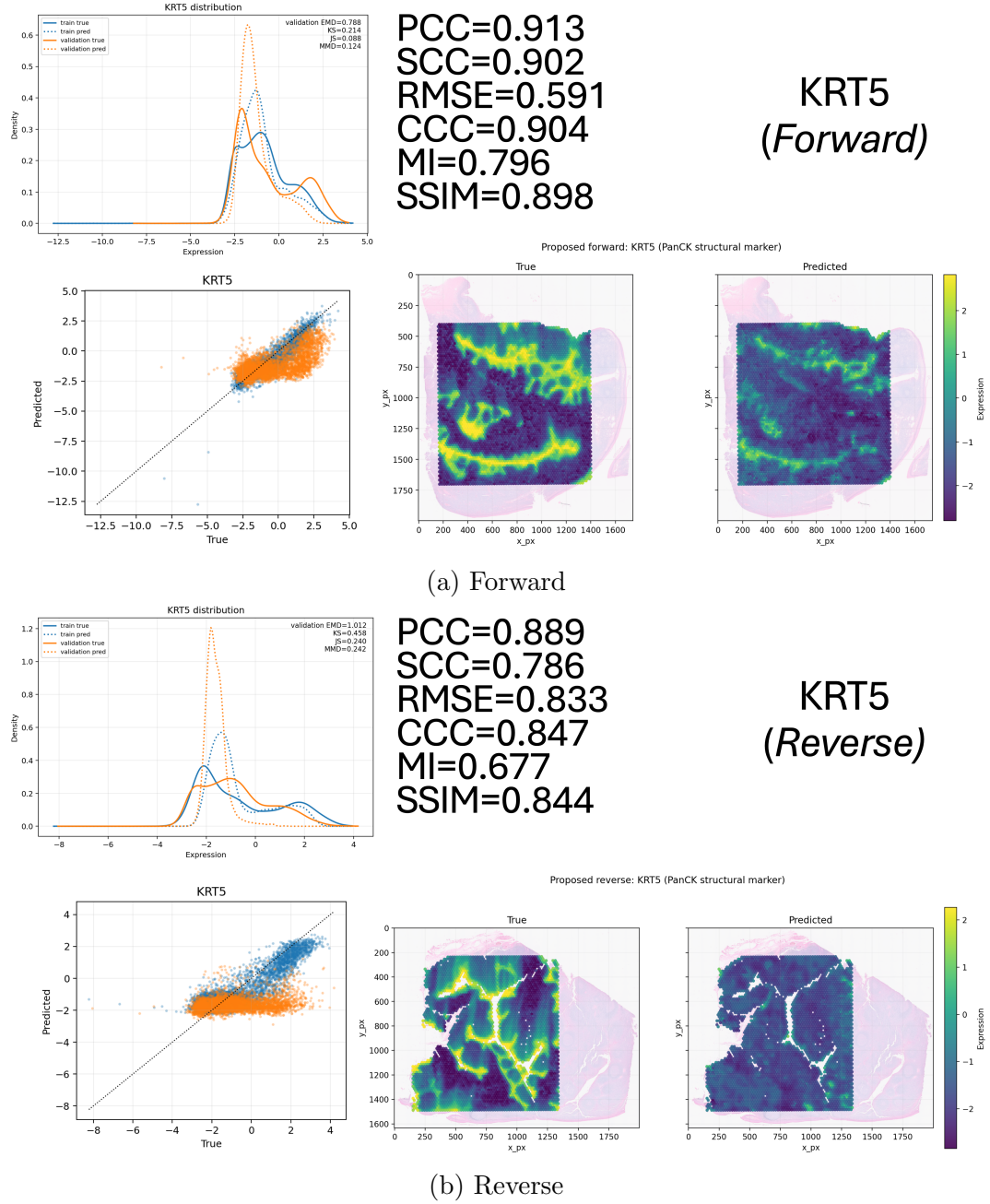


Figure 3.13: KRT5 case study for the proposed model. KRT5 illustrates recovery of a basal epithelial/cytokeratin structural marker. The predictions retain broad epithelial spatial organization, especially in the forward direction, but remain smoother and less extreme than the measured maps.

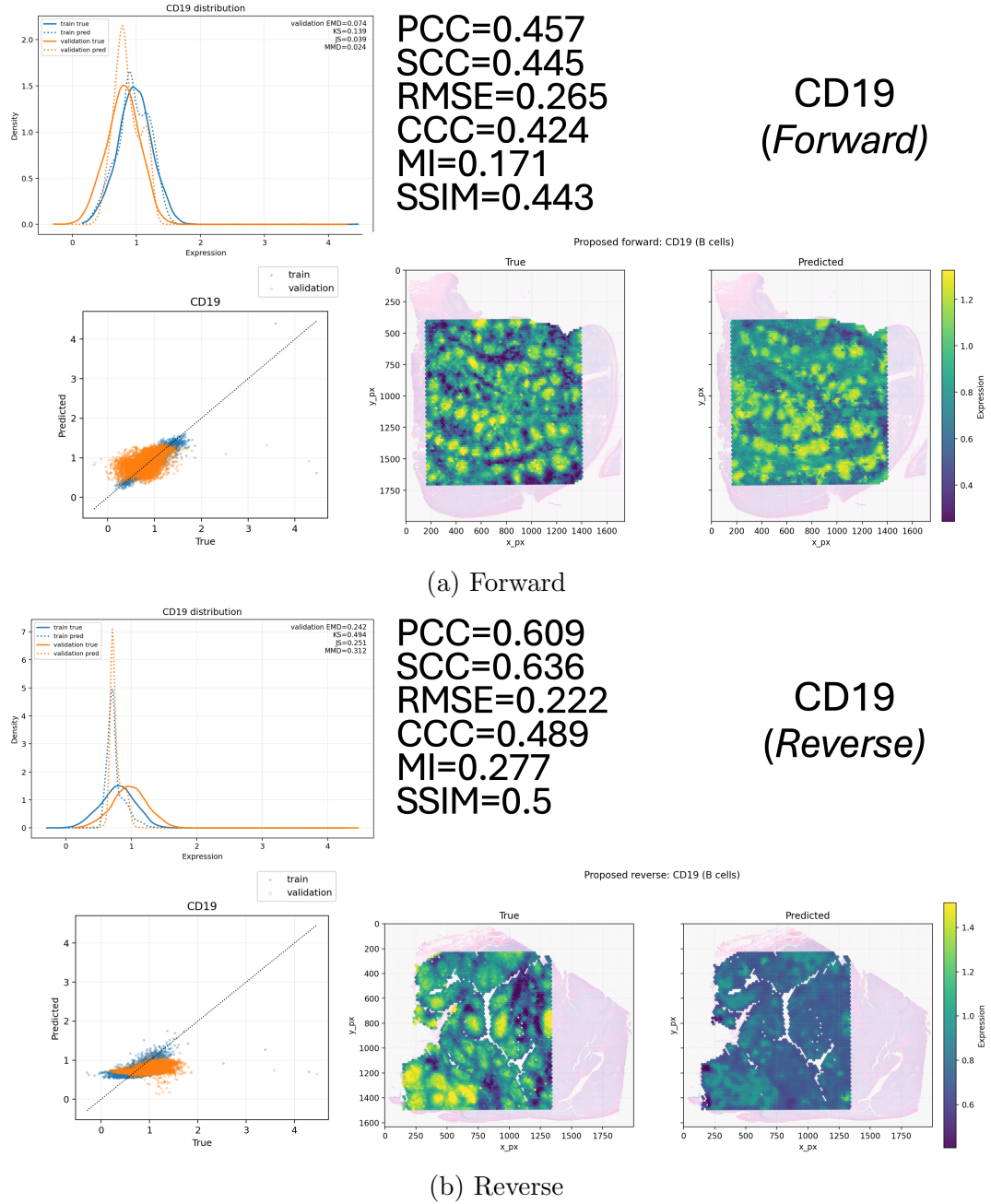


Figure 3.14: CD19 case study for the proposed model. CD19 shows partial recovery of broad B-cell-associated spatial regions, but predicted maps are dynamically compressed and miss sharper local enrichment and depletion.

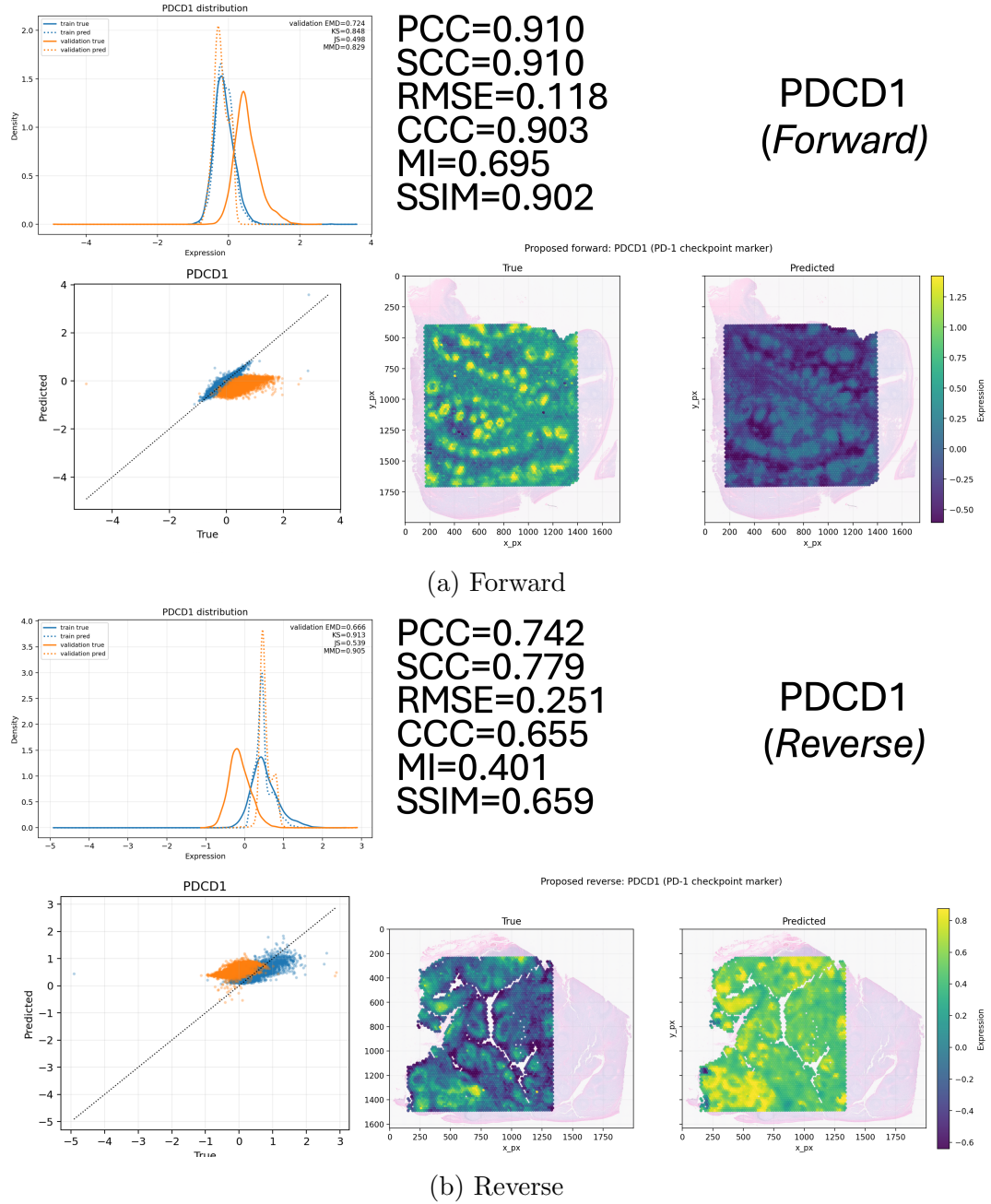


Figure 3.15: PDCD1 case study for the proposed model. PDCD1 recovery is interpreted as morphology-informed prediction of checkpoint-associated immune context rather than direct visualization of PD-1 expression from H&E.

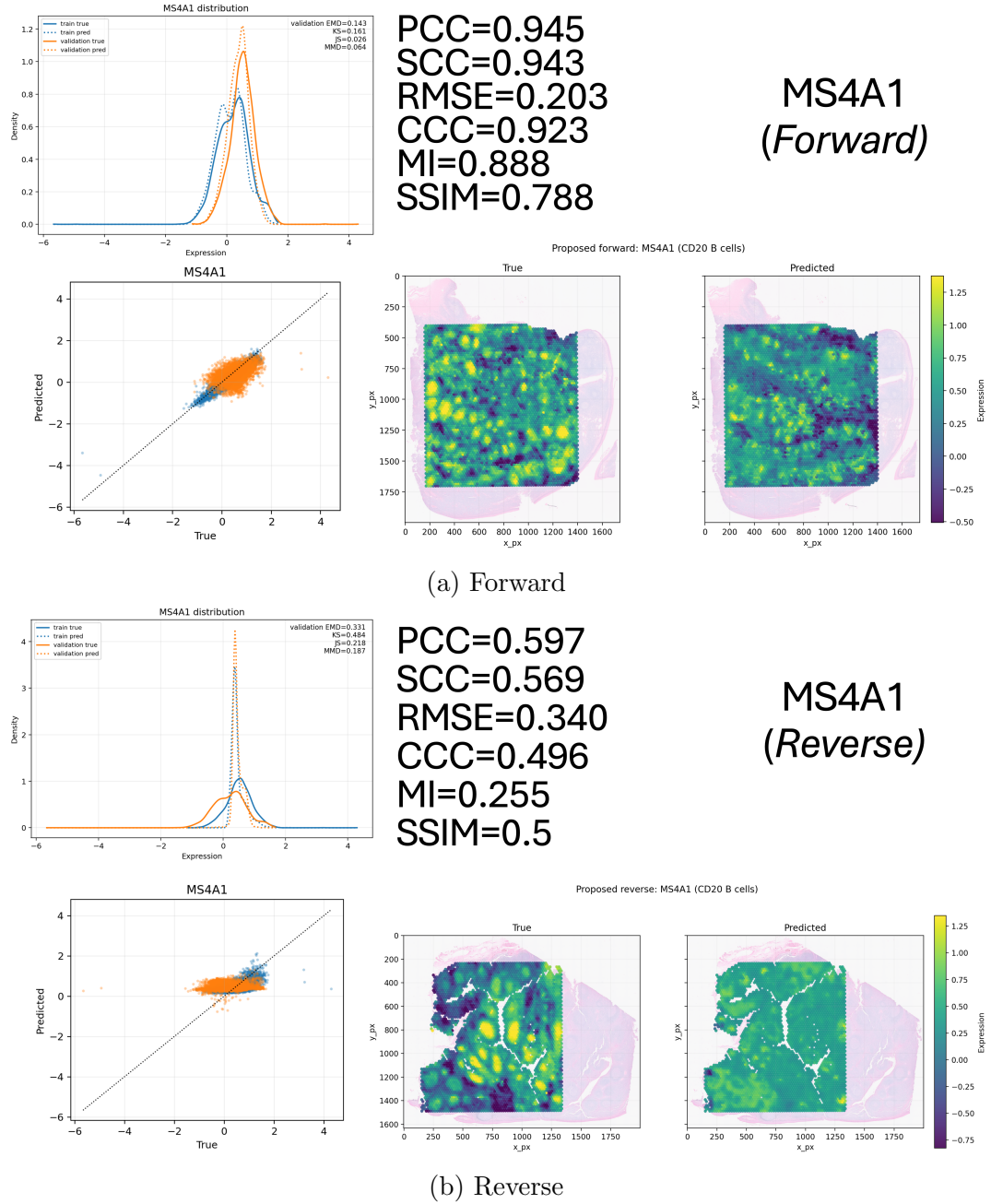


Figure 3.16: MS4A1 case study for the proposed model. MS4A1 serves as a direction-dependent boundary case: the forward section retains moderate marker-specific structure, whereas the reverse direction shows strong collapse and loss of calibrated marker recovery.

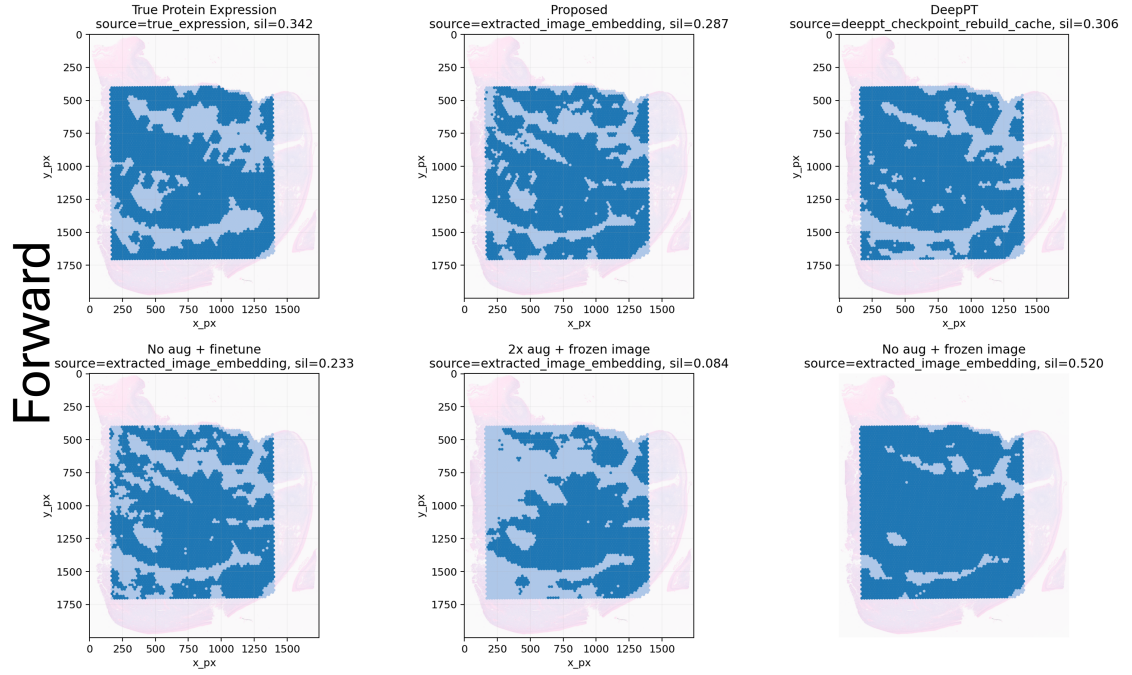


Figure 3.17: Forward-validation spatial clustering from measured and predicted protein profiles. The proposed model and DeepPT both recover organized tissue partitions, while ablated models show that high cluster separability alone does not guarantee accurate marker recovery.

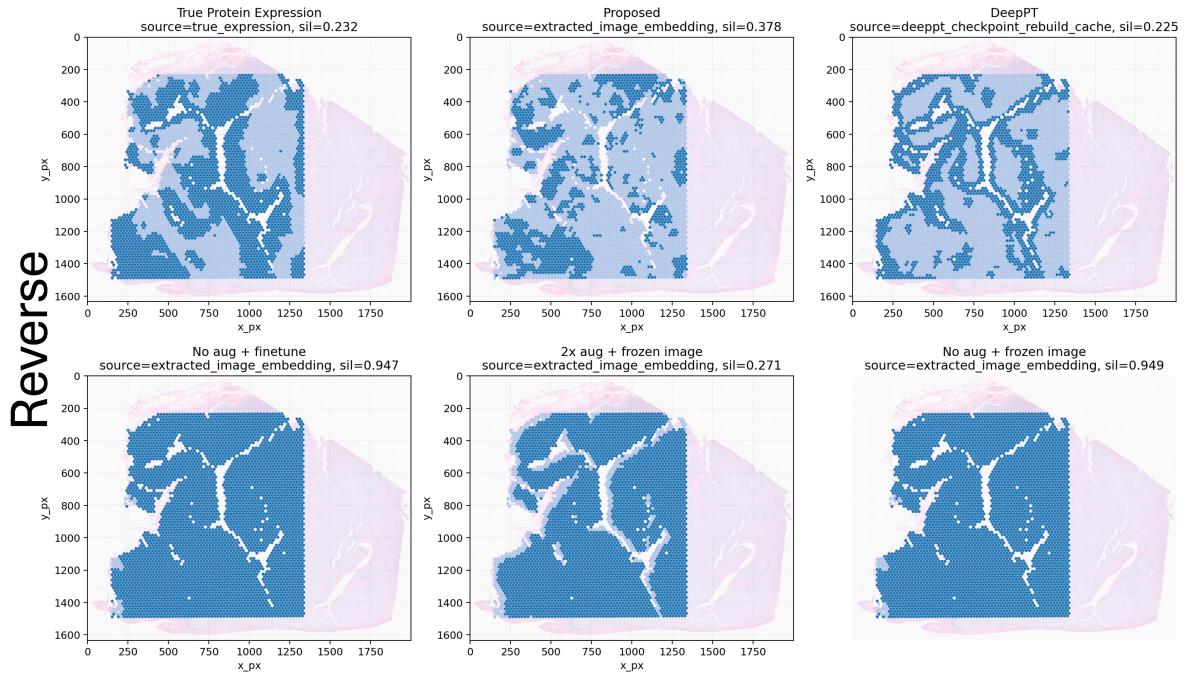


Figure 3.18: Reverse-validation spatial clustering from measured and predicted protein profiles. The reverse section again shows organized partitions, but the relationship between cluster compactness and marker-level prediction quality is model dependent.

model. In the reverse validation section, the ordering changed: the proposed model had distance 0.711, compared with 0.636 for DeepPT. The shallow probes also performed strongly on this diagnostic, particularly in the forward direction. Full CMD values by model, direction, and split are provided in Supplementary Table A.4. Thus, the evidence for the proposed model is strongest for marker-level spatial association, not for uniformly superior multivariate protein-structure recovery.

The correlation-matrix panels visualize this mixed result (Figures 3.19 and 3.20). Forward predictions from the proposed model preserved blocks of correlated immune and structural markers more faithfully than DeepPT, whereas the reverse direction favored DeepPT by this multivariate diagnostic. The proposed model therefore improves one important part of spatial protein recovery, but it does not yet provide direction-invariant preservation of the full antibody-panel covariance structure.

3.7 Summary

These results support the thesis that H&E morphology contains recoverable information about a subset of spatial protein patterns. In this controlled tonsil proof-of-concept, the proposed latent-alignment framework improved marker-level spatial association over DeepPT in the forward whole-section transfer and produced a biologically interpretable marker-specific spectrum. Structural recovery was strongest for KRT5, whereas PDCD1 is best understood as checkpoint-associated immune context rather than direct molecular detection. Spot-level accuracy was often high even when individual marker maps were imperfect, predicted maps were somewhat smoother, linear probing proved to be an effective approach, and multivariate protein-correlation recovery changed with transfer direction. Virtual spatial proteomics from H&E should therefore be treated as a morphology-guided method whose reliability depends on marker biology, tissue section, and evaluation metric.

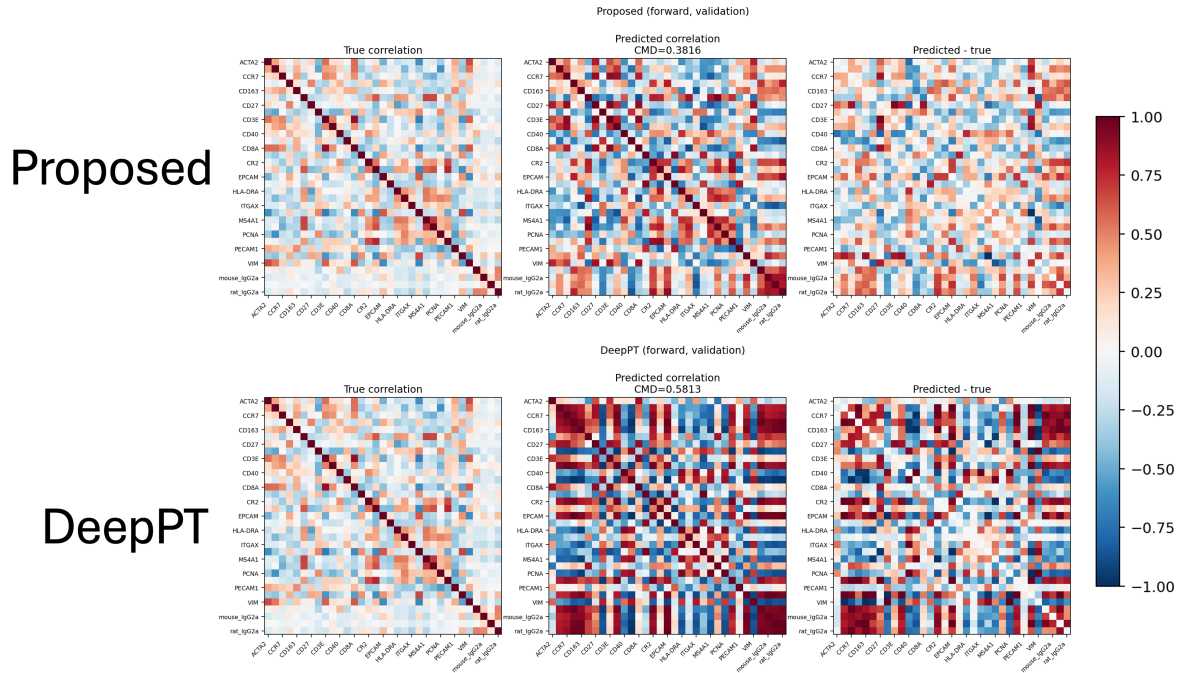


Figure 3.19: Forward-validation protein-protein correlation preservation. The proposed model produced a lower correlation matrix distance than DeepPT, indicating better recovery of cross-marker covariance structure in the easier transfer direction.

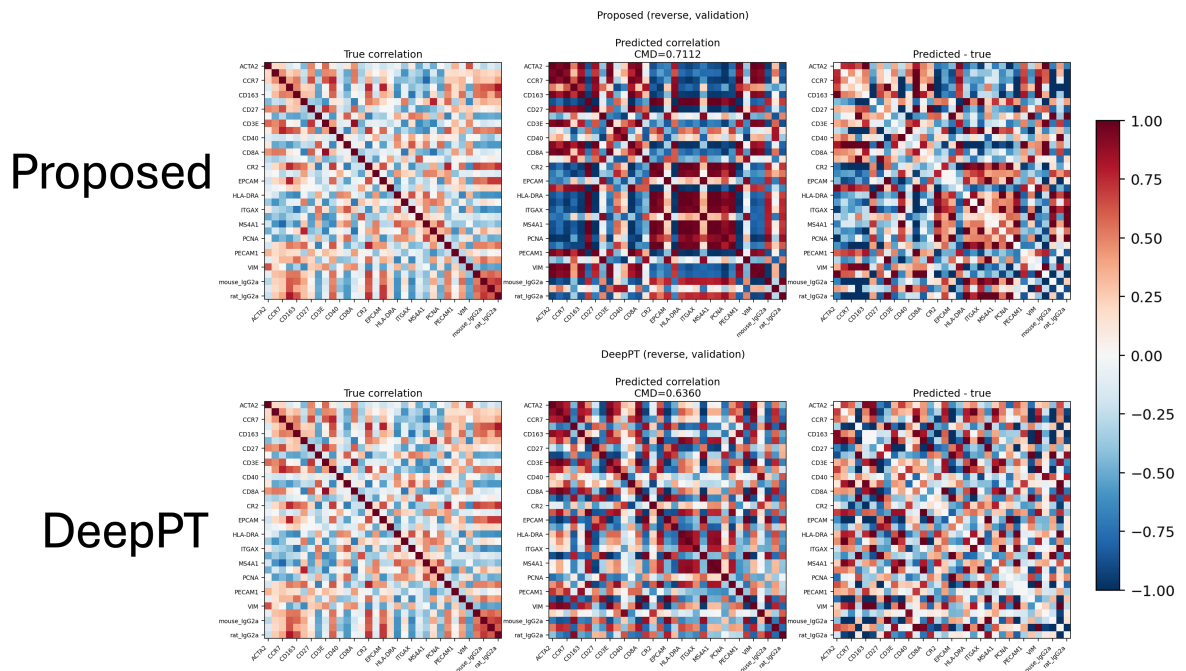


Figure 3.20: Reverse-validation protein-protein correlation preservation. DeepPT had the lower correlation matrix distance in this direction, highlighting that multivariate antibody-panel structure was not preserved uniformly by the proposed model across section transfers.

CHAPTER 4

Discussion

4.1 Findings

This thesis tested whether a two-stage latent alignment framework can predict spatial protein expression from H&E histology. The results partially support the central hypothesis: modeling protein expression through a learned latent space can improve spatially meaningful prediction relative to a direct deep-learning baseline, but the benefit was marker-dependent and direction-dependent. The proposed model outperformed DeepPT most clearly on marker-wise correlation and structural similarity in the forward whole-section transfer, although it did not uniformly minimize spot-level error.

The results also show that performance depends strongly on the metric used. CatBoost achieved the lowest spot-level error in the pooled benchmark, while the proposed model achieved stronger protein-wise spatial association than DeepPT in the easier transfer direction. This distinction matters for virtual spatial proteomics. A model with lower pointwise error may still oversmooth the expression landscape, whereas protein-wise correlation and structural similarity better capture whether predicted maps preserve biologically organized regions.

4.2 Interpretation of Model Behavior

The strongest gains over DeepPT were observed for protein-level Spearman and Pearson correlation. This suggests that the proposed model is most useful for preserving relative spatial variation across tissue, not merely predicting the average expression profile. The latent decoder likely contributes to this behavior by constraining predictions to a protein-expression manifold learned from measured data. Rather than predicting each marker independently, the decoder uses co-expression structure learned from the full panel. This is the main methodological claim of the thesis: protein prediction is routed through a learned protein decoder, and the value of that design is evaluated under whole-section transfer rather than under an easier random spot split.

The ablation results indicate that image encoder fine-tuning is more important than augmentation alone. Frozen-feature linear probes performed poorly even when augmented, suggesting that generic visual features are not sufficient for this task. Histopathology-to-protein prediction requires feature adaptation to morphology that is tissue-specific and marker-specific. Fine-tuning likely allows the image encoder to shift from generic object-recognition features toward tissue architecture, cellular density, staining texture, and compartment organization.

Augmentation produced a smaller and more mixed effect. The augmented fine-tuned model improved MSE and SSIM relative to the no-augmentation fine-tuned ablation, but had slightly lower mean Pearson and Spearman correlation. This suggests that rotation and flip augmentation regularize spatial structure and reduce overfitting, but may also blur orientation-sensitive local cues or interact with the small sample size. The result does not argue against augmentation. Instead, it suggests that augmentation strength and type should be tuned to the spatial biology of the tissue.

The marker case studies clarify the dominant failure mode. Errors were not purely random. Predictions often preserved coarse tissue organization while attenuating local

hotspots and compressing the dynamic range. This behavior is useful for identifying broad morphology-linked regions, but it limits interpretation when small local niches are the biological target.

4.3 Biological Interpretation

The best-predicted proteins were markers with clear tissue-level or compartment-level organization. CXCR5, KRT5, CD3E, PAX5, CEACAM8, ACTA2, PDCD1, and CD274 all showed relatively high Pearson correlations. These proteins are associated with immune, epithelial, stromal, or activation-related structures that can plausibly leave morphological signatures in H&E. For example, tonsil contains organized lymphoid follicles, germinal centers, interfollicular T-cell regions, crypt epithelium, and stromal boundaries that are visible at the tissue-architecture scale (Nave et al., 2001; Brachtel et al., 1996). More generally, computational pathology studies have shown that immune infiltration and epithelial-stromal compartment structure can be recovered from H&E morphology (Saltz et al., 2018; Wu et al., 2022).

The marker-specific examples refine this interpretation. KRT5 was the clearest structural-marker success, consistent with the idea that basal epithelial or cytokeratin-associated architecture is visually recoverable (Chu and Weiss, 2002). CD19, MS4A1, PAX5, and CXCR5 support a related but more heterogeneous B-cell and lymphoid-organization interpretation: these markers are tied to B-lineage biology or follicular localization, but their local protein maps still depend on section-specific immune architecture and assay behavior (Sato et al., 1995; Tedder and Engel, 1994; Nutt et al., 1999; Förster et al., 1996). CD19 and MS4A1 showed that immune-lineage markers are not uniformly easy or uniformly impossible: CD19 retained moderate broad-region recovery, whereas MS4A1 was much more direction-sensitive. PDCD1 was biologically relevant but requires the most caution. Its recovery should be interpreted as checkpoint-associated immune context, such as lymphoid organization or inflammatory neighborhoods, not as

direct detection of PD-1 expression from H&E morphology (Keir et al., 2008).

The weakest markers included control antibodies and markers with less obvious local morphological correlates. This is an important result. Some protein expression is driven by molecular state rather than visible morphology, and some markers may vary at a spatial scale smaller than the H&E tile or larger than the model’s receptive field. Negative or near-zero correlations for several markers indicate that H&E-derived prediction should be interpreted as a morphology-informed estimate, not as a replacement for measured spatial proteomics.

4.4 Implications for Virtual Spatial Proteomics

The practical value of virtual spatial proteomics depends on whether predicted maps can support downstream biological interpretation. The results suggest that H&E can recover a meaningful subset of protein spatial structure, especially for markers tied to visible compartments. This makes the approach potentially useful for hypothesis generation, low-cost screening, and prioritizing regions for follow-up multiplexed assays.

At the same time, the results argue against treating virtual staining as a direct substitute for experimental protein measurements. The current model is best understood as a computational imputation tool whose reliability varies by marker and transfer direction. It can identify likely broad spatial patterns, but uncertainty remains high for proteins whose expression is not strongly encoded in morphology or whose biology is expressed through sparse local hotspots.

The comparison with Ridge and CatBoost also has practical implications. Classical baselines trained on strong embeddings were competitive, and CatBoost achieved the best spot-level error. This suggests that representation quality may be as important as predictor complexity. Future systems should benchmark against strong shallow models rather than only neural baselines, especially when the dataset contains only a few thousand

spots.

4.5 Limitations

This study has several limitations. First, the main experiment used two CytAssist human tonsil protein-expression datasets with the same 33-marker panel. This design is useful for controlled evaluation, but it does not establish generalization across tissue types, cancer indications, staining protocols, or protein panels. Stronger evidence would require external validation on additional organs and disease contexts.

Second, the held-out evaluation used spots from the same overall data source. Even when the split is held out, shared acquisition protocol and tissue type may make the prediction problem easier than true deployment. In addition, because only two compatible sections were available, the held-out section served both as the validation section used during model development and as the final reported evaluation section. A more stringent test would train on one cohort or tissue source, select models without inspecting the final test tissue, and evaluate on an entirely independent dataset collected under different experimental conditions.

Third, the model predicts spot-level protein values rather than single-cell expression. Each spot can aggregate multiple cells, so predicted protein vectors may reflect local mixtures rather than individual cellular phenotypes. This limits direct interpretation of predicted maps as cell-resolved measurements.

Fourth, the current evaluation emphasizes summary metrics and does not yet demonstrate downstream applications in clinical settings. Pearson, Spearman, MSE, and SSIM are useful, but they do not fully capture whether downstream biological conclusions are preserved. The marker case studies show that a model can preserve coarse morphology-linked structure while missing calibrated magnitude, local hotspots, or direction-specific robustness. Future work should test whether predicted maps recover known tissue compartments,

cell-type neighborhoods, immune niches, or clinically meaningful biomarkers.

4.6 Future Work

Future work should expand the model in three directions. First, the model should be trained and evaluated using more data. Scaling up the training sets to include more tissue sections, diverse organ types, varying disease states, and experimental batches would clarify whether the learned visual-protein relationships are broadly robust. Cross-validation across these larger datasets is necessary to assess model generalizability under biological and technical heterogeneity.

Second, the modeling framework should be adapted for higher-resolution spatial generation. Moving beyond spot-level aggregation, applying the latent alignment approach to high-resolution modalities such as CO-Detection by indEXing (CODEX) or multiplexed immunofluorescence (IMF) data could enable single-cell or sub-cellular precision. This structural shift would allow the direct prediction of fine-grained cellular phenotypes and structures from matched H&E microscopic fields.

Third, future studies should evaluate the model in downstream applications that demonstrate clear clinical or biological value. Moving beyond summary error metrics, predicted protein maps should be tested in practical biological workflows, such as tumor microenvironment profiling, cell-neighborhood analysis, or region selection for multiplexed validation. Clinical utility, such as therapy-response prediction, would require independent cohorts and direct comparison against measured protein assays.

4.7 Conclusion

This thesis developed and evaluated a two-stage latent alignment framework for predicting spatial protein expression from H&E histology. The approach separates morphology encoding, protein co-expression modeling, and cross-modal alignment: an image encoder

learns histology-derived features, a protein autoencoder learns the measured protein manifold, and an aligner maps visual latents into the protein latent space.

Across the held-out whole-section transfers, the best two-stage model improved over DeepPT on protein-wise spatial association metrics in the forward direction, while spot-level error and reverse-transfer performance remained more mixed. The strongest predictions occurred for markers with clear tissue-compartment organization, whereas control markers, direction-sensitive immune markers, and less morphology-linked proteins remained difficult. The final takeaway is therefore constrained but useful: H&E can recover some morphology-linked protein structure, but virtual spatial proteomics should currently be treated as marker-specific, section-sensitive, and hypothesis-generating rather than measurement-replacing.

Appendix A

Supplementary Tables

This appendix is intended as reproducibility and reference material rather than continuous narrative reading. The main argument of the thesis is contained in the preceding chapters; the tables below provide expanded method inventories, benchmark summaries, metric definitions, and full model outputs for readers who want to audit the supporting details.

Table A.1: Complete taxonomy of computational methods for spatial omics and histology analysis. Rows are grouped by refined technical family, and architecture/strategy details are merged to reduce table width.

Method	Main task	Inputs	Model design	Target resolution	Reference
<i>Generative flow matching</i>					
HoloTea	3D volumetric spatial transcriptomics imputation from serial histology	H&E images and Spatial Transcriptomics (ST)	Flow-matching framework with ControlNet and Global Set Attention (GSA); Adjacent-section conditioning consistent via morphologically matched context; learned ZINB priors	Spot-level	(Sanian et al., 2025)
STFlow	Whole-slide generative modeling of spatial transcriptomics	H&E images and Spatial Transcriptomics (ST)	Frame Averaging (FA)-based Transformer denoiser; E(2)-invariant spatial attention; ZINB prior; iterative refinement	Spot-level (slide)	(Huang et al., 2025)

continued on next page

Table A.1: Complete taxonomy of computational methods for spatial omics and histology analysis (continued).

Method	Main task	Inputs	Model design	Target resolution	Reference
<i>Diffusion and spatial smoothing</i>					
Stem	Conditional gene expression generation from histology	H&E images and Spatial Transcriptomics (ST)	DiT-based diffusion network; Conditional diffusion guided by UNI/CONCH image embeddings	Spot-level	(Zhu et al., 2025)
sepal	Identification of transcript profiles with spatial patterns	Spatial Transcriptomics (ST) data	Numerical simulation (Fick’s second law); Simulation of transcript diffusion time until convergence	Spot-level	(Andersson and Lundberg, 2021)
Thor	In silico cell spatial transcriptome inference	H&E images and spot-level Spatial Transcriptomics (ST)	Anti-shrinking Markov Diffusion Model; Shared Nearest Neighbors (SNN) graph; feature-preserving Markov diffusion	Single-cell	(Zhang et al., 2025)
HaDM-ST	High-resolution spatial transcriptomics generation	H&E images and Low-Resolution (LR) ST	Transformer semantic encoder / U-Net noise predictor; Semantic Distillation (HSD); Cross-Modal Spatial Alignment (CMSA)	Subcellular (10x enlarge-ment)	(Liu et al., 2025)
HINGE	Histology-to-ST imputation	H&E images and Spatial Transcriptomics (ST)	CellFM (Transformer-based sc-FM) and SoftAdaLN; Layer-wise visual context injection; warm-start curriculum	Spot-level	(Fang et al., 2026)
<i>Spatial-context models: attention, graph, and state-space</i>					
iStar	Super-resolution spatial gene expression prediction and tissue architecture inference	H&E images and Spatial Transcriptomics (ST)	Hierarchical Vision Transformer (HViT); Self-supervised pretraining (DINO/BEiT); superpixel-level feature extraction; weakly supervised level learning	Near-single-cell / Superpixel	(Zhang et al., 2024)
MV_Hybrid	Spatial gene expression prediction from H&E	H&E stained WSIs	Mamba Vision (SSM) and Vision Transformer (ViT); Low-frequency bias; DINOv2 self-supervised pretraining	Patch-level	(Cho et al., 2025)

continued on next page

Table A.1: Complete taxonomy of computational methods for spatial omics and histology analysis (continued).

Method	Main task	Inputs	Model design	Target resolution	Reference
GHIST	Spatially resolved single-cell gene expression prediction	H&E images and Subcellular Spatial Transcriptomics (SST)	UNet 3+ backbone; Multitask learning (cell type, morphology, neighborhood); Multi-head cross attention	Single-cell	(Fu et al., 2025)
HEX	Virtual spatial proteomics and gene expression prediction	H&E images and Spatial Transcriptomics (ST)	Pathology foundation model (MUSK backbone) or Hierarchical ViT; Feature distribution smoothing (FDS); Adaptive loss function (ALF); patches Hierarchical feature extraction	Near-single-cell / 14-pixel	(Li et al., 2026b)
HisToGene	Spatially resolved gene expression prediction and super-resolution	H&E images and Spatial Transcriptomics (ST)	Transformer-based model with Multi-Head Attention; Positional embedding for spot dependency; dense image patch sampling	Sub-spot / Super-resolution	(Pang et al., 2021)
Pixel2Gene	Histology-guided reconstruction and prediction	H&E images and Spatial Transcriptomics (ST)	Vision Transformer (ViT) and MLP; Hierarchical superpixel-based tiling; multi-scale visual cues	Single-cell (8 μm)	(Yao et al., 2026)
TransformerSST	Super-resolved gene expression reconstruction	H&E images and Spatial Transcriptomics (ST)	ViT encoder and adaptive Graph Transformer; Cross-scale internal graph network; co-representation learning	Single-cell	(Zhao et al., 2024)
scTEL	Protein expression prediction and cell type identification	scRNA-seq and CITE-seq	Transformer Encoder and LSTM; Multi-task framework; Attention mechanism; Quantile loss	Single-cell	(Chen et al., 2025b)
SpatialEx	H&E-to-omics prediction (Transcriptomics, Proteomics, Metabolomics)	H&E images and Spatial Omics	UNI foundation model, Hypergraph Neural Networks (HGNNs), and MLP; Multi-neighborhood dependency encoding; cross-omics cycle consistency	Single-cell or Spot-level	(Liu et al., 2026)

continued on next page

Table A.1: Complete taxonomy of computational methods for spatial omics and histology analysis (continued).

Method	Main task	Inputs	Model design	Target resolution	Reference
Hist2ST	Spatial transcriptomics prediction from histology	H&E images and Spatial Transcriptomics (ST)	Convmixer, Transformer, and GraphSAGE; Zero-Inflated Negative Binomial (ZINB) mechanism; self-distillation; global spatial dependencies	Spot-level	(Zeng et al., 2022)
THItGene	Predicting spatial transcriptomics from histology	H&E image patches	Efficient-CapsNet, ViT, and GAT; Multidimensional attention; hierarchical multi-view neighborhood features	Spot-level	(Jia et al., 2024)
Swin Transformer	Spatial Transcriptomics Prediction	H&E image patches	Swin-Tiny; Multi-class classification reformulation; window-based self-attention	Near-cellular / Subcellular	(Li et al., 2026a)
TCGN	Spatial transcriptomics prediction	H&E images	CMT blocks and CensNet (GNN); Graph-node co-embedding; patch-to-patch connection analyzer	Spot-level	(Xiao et al., 2024)
GigaTIME	Virtual multiplex immunofluorescence (mIF) generation	H&E pathology slides	NestedUNet or ViT-giant; Cross-modal translation trained on 40 million paired cells	Pixel-level	(Valanarasu et al., 2026)
SCHAF	H&E to single-cell transcriptomics imputation	H&E images and sc/snRNA-seq	ViT-Large or Adversarial encoder-decoder (UNI/scGPT); Tangram-based mapping; adversarial training; latent space alignment	Single-cell	(Comiter, 2024)
EGN (EGGN)	Spatial gene expression prediction and tissue architecture inference	H&E images and reference Spatial Transcriptomics (ST)	ResNet or ViT backbone and GraphSAGE; Graph Exemplar Bridging (GEB) block; kNN-based exemplar retrieval; StyleGAN-based style code	Fine-grained area (Window)	(Yang et al., 2024)
<i>Direct image-to-omics regression</i>					
DeepPT	Predict mRNA expression and cancer treatment response	H&E images	ResNet-50, Autoencoder, and MLP; Multi-task learning; ensemble learning (bagging); MSE loss optimization	Slide-level (Bulk) to Spot-level	(Hoang et al., 2024)

continued on next page

Table A.1: Complete taxonomy of computational methods for spatial omics and histology analysis (continued).

Method	Main task	Inputs	Model design	Target resolution	Reference
ST-Net	Spatial gene expression prediction	H&E images and Spatial Transcriptomics (ST)	DenseNet-121; Patch-level feature extraction; multivariate regression; transfer learning	Spot-level (100 μm)	(He et al., 2020)
ROSIE	Virtual multiplex immunofluorescence (mIF) staining	H&E images and mIF / CODEX data	ConvNext-Small; Pixel-level multitask regression; masked MSE loss	Pixel-level (3.02 μm)	(Wu et al., 2025)
DeepSpaCE	Super-resolution spatial mapping	H&E images	VGG16; Semi-supervised learning (Noisy Student method)	Super-resolution	(Monjo et al., 2022)
<i>Contrastive and multimodal alignment</i>					
BLEEP	Spatially resolved gene expression prediction and imputation	H&E images and Spatial Transcriptomics (ST)	ResNet50 Image Encoder and MLP Expression Encoder; Bimodal contrastive learning; Query-reference imputation; Joint embedding space construction	Spot-level	(Xie et al., 2023)
OmiCLIP	Spatial transcriptomics prediction	H&E images	Visual-omics foundation model	not reported	(Chen et al., 2025a)
Multi-Embed	Cross-modality inference and integration	H&E images and multi-omics	UNI foundation model, auto-encoder, and MLP; Triplet loss for inter-modality alignment; self-reconstruction	Multilevel (Slide/Spot/Cell)	(Zhang et al., 2026)
CoMTIP	Spatial transcriptomics representation learning	H&E images and gene names	ViT-B/32 and Gene-Text Encoder; Masked Feature Modeling; Pair-aware Adversarial Training (PAAT)	Spot / Cellular	(Qian et al., 2025)
MoBY	H&E to ST imputation and tissue structure discovery	H&E images and Spatial Transcriptomics (ST)	Hierarchical image feature extraction (Swin Transformer / DeiT) or CNN (RetCCL); Self-supervised contrastive learning (MoCo v2 and BYOL); momentum updating	Super-resolution / Spatial	(Xie et al., 2021)
mclSTExp	Spatial gene expression prediction	H&E images and ST data	Transformer and DenseNet-121; Multimodal contrastive learning; weighted distance aggregation	Spot-level	(Min et al., 2024)

continued on next page

Table A.1: Complete taxonomy of computational methods for spatial omics and histology analysis (continued).

Method	Main task	Inputs	Model design	Target resolution	Reference
CHRep	Spatial gene expression prediction from H&E	H&E images and spot coordinates	Image Encoder and Gene Encoder; Symmetric image-expression alignment; spatial topology regularization	Spot-level	(Wang et al., 2026)
HAGE	Gene expression prediction from histology	Histopathology images	UNI foundation model; Gene co-expression priors (Gene2Vec); CyCLIP and CLIP loss	Patch-level	(Dang et al., 2025)
<i>Foundation-model molecular profiling</i>					
Path2Omics	Gene expression and methylation prediction	H&E (FFPE and Fresh Frozen) slides	CTransPath foundation model; Integration of FF and FFPE slide predictions; foundation model feature extraction	Bulk	(Hoang et al., 2025)
<i>Cross-modal image translation and adversarial transport</i>					
HistoPlexer	Virtual proteomics / protein multiplex generation	H&E images and Imaging Mass Cytometry (IMC)	Conditional GAN (cGAN) with U-Net; Gaussian pyramid loss; patch-wise contrastive loss	WSI-level	(Andani et al., 2025)
Unbalanced Monge OT GAN	Unbalanced optimal transport modeling	scRNA-seq or images	Fully-connected feedforward networks; Stochastic transport map; adversarial game; gradient penalties	Continuous measure	(Yang and Uhler, 2018)
Pyramid Pix2pix	H&E to IHC virtual staining	H&E images	ResNet generator / PatchGAN discriminator; Multi-scale Gaussian Pyramid constraints	1024x1024 patches	(Liu et al., 2022)
<i>Coordinate and latent autoencoding</i>					
STAGE	High-density spatial transcriptomics generation	Spatial Transcriptomics and 10x Visium	Spatial location-supervised auto-encoder; Wasserstein distance; L1 loss denoising	High-density / Arbitrary	(Li et al., 2024)
<i>Segmentation-guided single-cell reconstruction</i>					

continued on next page

Table A.1: Complete taxonomy of computational methods for spatial omics and histology analysis (continued).

Method	Main task	Inputs	Model design	Target resolution	Reference
RedeHist	Spatial transcriptomics prediction and imputation	Histology images and scRNA-seq	U-Net and DNN; Histological cell segmentation; deconvolution using scRNA-seq references	Single-cell	(Zhong et al., 2024)
<i>Single-cell probabilistic integration</i>					
totalVI	Integrated analysis and surface protein prediction	scRNA-seq and CITE-seq data	Bayesian probabilistic framework; Variational inference	Single-cell	(Gayoso et al., 2021)
<i>Color and stain-aware modeling</i>					
Neural Stain Learning (NSL)	Spatial transcriptomics gene expression prediction	H&E images	Neural Stain Deconvolution Layer; End-to-end stain deconvolution matrix learning	Spot-level	(Dawood et al., 2021)
<i>Set-based virtual single-cell prediction</i>					
DeepSpot2Cell	Virtual single-cell spatial transcriptomics prediction	H&E images	Permutation-invariant DeepSet network; Modeling spots as bags of cells; frozen foundation model embeddings	Single-cell	(Nonchev et al., 2025)
<i>High-resolution foundation encoders</i>					
HisToSGE	High-resolution spatial gene expression prediction	Histological images	UNI Pathology Image Large Model; Multimodal feature mapping (RGB + Positional + Histological)	Adjustable high resolution	(Shi et al., 2024)

Table A.2: Complete benchmark inventory for histology-to-spatial-omics prediction. MAE and spot-level PCC are omitted to emphasize reported error and gene-level correlation summaries.

Method	Backbone	Target genes/panel	MSE / error	PCC gene	Reference
<i>Dataset: ST-Data</i>					

continued on next page

Table A.2: Complete benchmark inventory for histology-to-spatial-omics prediction (continued).

Method	Backbone	Target genes/panel	MSE / error	PCC gene	Reference
HoloTea	Spatial Transformer with Global Set Attention (GSA)	Top-250 Highly Variable Genes (HVG)	0.364	0.612 $\pm$ 0.132	(Sanian et al., 2025)
STFlow	Within-section k-NN graph attention	Top-250 Highly Variable Genes (HVG)	0.443	0.540 $\pm$ 0.142	(Sanian et al., 2025)
ASIGN 3D	Graph Neural Network (GNN)	Top-250 Highly Variable Genes (HVG)	0.481	0.531 $\pm$ 0.121	(Sanian et al., 2025)
<i>Dataset: HER2</i>					
HoloTea	Spatial Transformer with Global Set Attention (GSA)	Top-250 Highly Variable Genes (HVG)	0.455	0.638 $\pm$ 0.145	(Sanian et al., 2025)
STFlow	Within-section k-NN graph attention	Top-250 Highly Variable Genes (HVG)	0.477	0.620 $\pm$ 0.146	(Sanian et al., 2025)
ASIGN 3D	Graph Neural Network (GNN)	Top-250 Highly Variable Genes (HVG)	0.462	0.607 $\pm$ 0.112	(Sanian et al., 2025)
<i>Dataset: 10xProteomic</i>					
EGGN	GraphSAGE (ResNet-18 Extractor)	250 genes with largest mean	0.0352	0.744	(Yang et al., 2024)
<i>Dataset: cSCC</i>					
HINGE	CellFM (Transformer)	Highly Mean-High Variance Genes (HMHVG)	0.887 $\pm$ 0.081	0.613 $\pm$ 0.008 (PCC-200)	(Fang et al., 2026)
STFlow	Transformer (Flow Matching)	Highly Mean-High Variance Genes (HMHVG)	0.903 $\pm$ 0.096	0.578 $\pm$ 0.012 (PCC-200)	(Fang et al., 2026)
Stem	DiT (Diffusion Transformer)	Highly Mean-High Variance Genes (HMHVG)	1.267 $\pm$ 0.163	0.577 $\pm$ 0.031 (PCC-200)	(Fang et al., 2026)
TRIPLEX	ResNet-based Encoders + Transformer	Highly Mean-High Variance Genes (HMHVG)	0.904 $\pm$ 0.067	0.588 $\pm$ 0.017 (PCC-200)	(Fang et al., 2026)
MERGE	ResNet18 + GNN (GAT)	Highly Mean-High Variance Genes (HMHVG)	1.082 $\pm$ 0.247	0.510 $\pm$ 0.031 (PCC-200)	(Fang et al., 2026)

continued on next page

Table A.2: Complete benchmark inventory for histology-to-spatial-omics prediction (continued).

Method	Backbone	Target genes/panel	MSE / error	PCC gene	Reference
BLEEP	ResNet-50	Highly Mean-High Variance Genes (HMHVG)	1.174 ± 0.109	0.548 ± 0.012 (PCC-200)	(Fang et al., 2026)
mclSTExp	DenseNet-121 and Transformer	Highly Variable Genes (HVG)	0.4302 ± 0.005	0.3235 ± 0.019	(Min et al., 2024)
BLEEP	ResNet/MLP (Contrastive Learning)	Highly Variable Genes (HVG)	0.5163 ± 0.007	0.2449 ± 0.017	(Min et al., 2024)
ST-Net	DenseNet-121	Highly Variable Genes (HVG)	0.6806 ± 0.006	0.0012 ± 0.022	(Min et al., 2024)
HAGE	Hierarchical Alignment Network	Highly Variable Genes (HVG)	0.4248 ± 0.002	0.4607 ± 0.007	(Wang et al., 2026)
CHRep	Cross-modal Transformer/Encoder	Highly Variable Genes (HVG)	0.4183 ± 0.003	0.4682 ± 0.005	(Wang et al., 2026)
TCGN	Transformer w/ Conv and Graph-Node	171 Highly Variable Genes (HVG)	0.080	0.410	(Xiao et al., 2024)
<i>Dataset: Mouse Brain</i>					
HisToSGE	UNI (PILM) + Multi-head Attention	Top 1,000 Highly Variable Genes (HVG)	0.1064	0.91 (e.g. Prkcd)	(Shi et al., 2024)
<i>Dataset: In-house Embryo</i>					
HoloTea	Spatial Transformer with Global Set Attention (GSA)	Top-250 Highly Variable Genes (HVG)	0.199	0.576 ± 0.245	(Sanian et al., 2025)
<i>Dataset: HER2ST</i>					
HINGE	CellFM (Transformer)	Highly Mean-High Variance Genes (HMHVG)	0.926 ± 0.047	0.446 ± 0.010 (PCC-200)	(Fang et al., 2026)
Stem	DiT (Diffusion Transformer)	Highly Mean-High Variance Genes (HMHVG)	0.998 ± 0.160	0.433 ± 0.019 (PCC-200)	(Fang et al., 2026)
STFlow	Transformer (Flow Matching)	Highly Mean-High Variance Genes (HMHVG)	0.929 ± 0.089	0.425 ± 0.024 (PCC-200)	(Fang et al., 2026)
<i>Dataset: HER2+ Breast Cancer</i>					

continued on next page

Table A.2: Complete benchmark inventory for histology-to-spatial-omics prediction (continued).

Method	Backbone	Target genes/panel	MSE / error	PCC gene	Reference
mclSTExp	DenseNet-121 and Transformer	Highly Variable Genes (HVG)	0.3235 - 0.5897	0.2304 - 0.4261	(Min et al., 2024; Dang et al., 2025)
ST-Net	DenseNet-121	Top 1,000 Highly Variable Genes (HVG)	0.0452 - 0.5312	0.0561 - 0.5312	(Yang et al., 2024; Fu et al., 2025; Min et al., 2024; Dang et al., 2025)
HisToGene	Vision Transformer (ViT)	Top 1,000 Highly Variable Genes (HVG)	0.0842 - 0.5202	0.0842 - 0.5202	(Min et al., 2024; Dang et al., 2025; Pang et al., 2021)
Hist2ST	CNN, Transformer, and GNN	Top 1,000 Highly Variable Genes (HVG)	0.1443 - 0.5135	0.1443 - 0.5135	(Min et al., 2024; Dang et al., 2025)
THItGene	Dynamic CNN, Capsule Net, and ViT	Top 1,000 Highly Variable Genes (HVG)	0.1726 - 0.5012	0.1726 - 0.5012	(Min et al., 2024; Dang et al., 2025)
BLEEP	ResNet-50	Top 1,000 Highly Variable Genes (HVG)	0.2449 - 0.6015	0.1873 - 0.3122	(Min et al., 2024; Dang et al., 2025)
HAGE	UNI (Vision Transformer based)	Top 1,000 HVG (Gene2Vec intersect)	0.4830 ± 0.005	0.4458 ± 0.003	(Dang et al., 2025)
NSL	Neural Stain Deconvolution Layer	Top 250 significant expression genes	~ 0.33	0.820	(Dawood et al., 2021)
DenseNet121	DenseNet-121	Top 250 significant expression genes	not reported	0.420	(Dawood et al., 2021)
<i>Dataset: STNet (Breast Cancer)</i>					
EGGN	GraphSAGE (ResNet-18 Extractor)	250 genes with largest mean	0.0394	0.292	(Yang et al., 2024)
EGN	ViT with Exemplar Bridging	Top 250 Mean Expression Genes	0.0410	0.1510	(Yang et al., 2023)
<i>Dataset: STNet</i>					
GraphSAGE	GCN	250 genes with largest mean	0.0379	0.283	(Yang et al., 2024)
<i>Dataset: Alex+10x (Breast Cancer)</i>					
mclSTExp	DenseNet-121 and Transformer	Highly Variable Genes (HVG)	0.2329 ± 0.006	0.1949 ± 0.011	(Min et al., 2024; Wang et al., 2026)
BLEEP	UNI (Pathology Foundation Model)	Highly Variable Genes (HVG)	0.2593 ± 0.013	0.1552 ± 0.009	(Min et al., 2024; Wang et al., 2026)

continued on next page

Table A.2: Complete benchmark inventory for histology-to-spatial-omics prediction (continued).

Method	Backbone	Target genes/panel	MSE / error	PCC gene	Reference
ST-Net	DenseNet-121	Highly Variable Genes (HVG)	0.4721 ± 0.011	0.0452 ± 0.007	(Min et al., 2024; Wang et al., 2026)
<i>Dataset: Alex+10x</i>					
CHRep	Cross-modal Transformer/Encoder	Highly Variable Genes (HVG)	0.2102 ± 0.033	0.4659 ± 0.020	(Wang et al., 2026)
<i>Dataset: HEST-Benchmark (CRC)</i>					
MV-Hybrid	Hybrid Mamba-Vision Transformer	Top 50 Highly Variable Genes	5.011 ± 1.478	0.747 ± 0.082	(Cho et al., 2025)
<i>Dataset: Human Breast Cancer (BC1)</i>					
STAGE	Spatially supervised autoencoder	Top 1,000 Highly Variable Genes (HVG)	0.1036	0.840	(Li et al., 2024)
<i>Dataset: Visium HD Human CRC-P2</i>					
Pixel2Gene	Vision Transformer (ViT)	Top 300 highly expressed genes	0.01 - 0.03 (RMSE)	0.40 - 0.60 (Median)	(Yao et al., 2026)
<i>Dataset: Breast Cancer (Patient-out CV)</i>					
ST-Net	DenseNet-121	Highly Variable Genes (HVG)	0.0961	not reported	(He et al., 2020)
<i>Dataset: Stanford-WSI (NSCLC)</i>					
HEX	MUSK (PFM)	40 protein biomarkers	0.076	0.790	(Li et al., 2026b)
CGAN	Generative Adversarial Network	40 protein biomarkers	0.380	0.541	(Li et al., 2026b)
<i>Dataset: Stanford-TMA</i>					
HEX	MUSK (PFM)	40 protein biomarkers	0.189	0.738	(Li et al., 2026b)
<i>Dataset: Bern Pan-cancer dataset</i>					
HEX	MUSK (PFM)	24 overlapping biomarkers	0.132	0.658	(Li et al., 2026b)
<i>Dataset: HER2+ Breast Tumor</i>					
TCGN	Transformer w/ Conv and Graph-Node	785 Highly Variable Genes (HVG)	0.100	0.460	(Xiao et al., 2024)
<i>Dataset: Lung Adenocarcinoma</i>					

continued on next page

Table A.2: Complete benchmark inventory for histology-to-spatial-omics prediction (continued).

Method	Backbone	Target genes/panel	MSE / error	PCC gene	Reference
GigaTIME	NestedUNet (Encoder-Decoder)	21 protein markers	not reported	0.590	(Valanarasu et al., 2026)
<i>Dataset: Xenium Breast Cancer</i>					
iStar	Hierarchical Vision Transformer	Top 1,000 Highly Variable Genes (HVG)	~0.1 (RMSE)	not reported	(Zhang et al., 2024)
<i>Dataset: Prostate Cancer (SAH)</i>					
ProGENIE	UNI (PFM) & Multi-head Attention	Top 1,000 predictable genes	0.149 (RMSE: 0.386)	0.600	(Han et al., 2026)
<i>Dataset: Lung Cancer (Xenium)</i>					
DeepSpot2CellDeepSet + Phikon-v2 (PFM)		Top 100 most predictive genes	not reported	0.450	(Nonchev et al., 2025)
<i>Dataset: Mouse Pup / ER+ MBC</i>					
Paired SCHAF	Vision Transformer (UNI)	Xenium Panel	not reported	0.402 - 0.445	(Comiter, 2024)
<i>Dataset: Human Breast Cancer</i>					
Thor	Anti-shrinking Markov Diffusion	Xenium Common Panel (306 genes)	0.07 (Normalized)	0.92 (CW-SSIM)	(Zhang et al., 2025)
<i>Dataset: DLPFC</i>					
STAGE	Supervised Auto-encoder	3,000 Highly Variable Genes (HVG)	not reported	0.370	(Li et al., 2024)
<i>Dataset: TCGA (23 cancer types)</i>					
Path2Omics- I	CTransPath foundation model	Highly variable genes (~18,000)	not reported	0.36 - 0.45	(Hoang et al., 2025)
<i>Dataset: HEST-Benchmark (Average)</i>					
H-Optimus-0	ViT-giant (DINOv2)	Top 50 Highly Variable Genes (HVG)	not reported	0.4146 $\pm$ 0.04	(Jaume et al., 2024)
UNIV1.5	ViT-giant (DINOv2)	Top 50 Highly Variable Genes (HVG)	not reported	0.4090 $\pm$ 0.04	(Jaume et al., 2024)
Virchow 2	ViT-Huge (DINOv2)	Top 50 Highly Variable Genes (HVG)	not reported	0.3984 $\pm$ 0.03	(Jaume et al., 2024)
<i>Dataset: Multiple (TCGA, Xenium, HER2ST)</i>					

continued on next page

Table A.2: Complete benchmark inventory for histology-to-spatial-omics prediction (continued).

Method	Backbone	Target genes/panel	MSE / error	PCC gene	Reference
DeepPT	ResNet50	Various (HVG, Tranches, Correlation)	(CMD) 0.18 - 0.35	0.07 - 0.75	(Fu et al., 2025; Liu et al., 2026; Hoang et al., 2024)
<i>Dataset: 10x Xenium Human Breast Cancer</i>					
SpatialEx+	UNI foundation model	Non-overlapping panels	0.12 - 0.22 (CMD)	0.14 - 0.82	(Liu et al., 2026)
<i>Dataset: Metastatic Melanoma (TuPro)</i>					
HistoPlexer	Conditional GAN (U-Net)	Spatiotemporal Protein Markers	0.0068 $\pm$ 0.0004	not reported	(Andani et al., 2025)
<i>Dataset: HER2ST (10x Visium)</i>					
Multi-Embed	UNI (PFM) + MLP Autoencoder	Highly Variable Genes (HVG)	not reported	$\sim$ 0.40	(Zhang et al., 2026)
<i>Dataset: Human Breast Cancer (ST-Data)</i>					
DeepSpaCE	VGG16	ESR1, ERBB2, MKI67	not reported	0.410	(Monjo et al., 2022)

A.1 Detailed Model Evaluation Tables

Table A.3: Model-level performance summary by transfer direction and split. The table is a compact version of `model_summary_by_direction.csv`; secondary standard-deviation columns are omitted to keep the appendix readable.

Model	Dir.	Split	Prot. PCC	Prot. SSIM	Prot. RMSE	Moran err.	Spot PCC	Spot RMSE
Proposed	F	Train	0.901	0.881	0.213	0.080	0.986	0.210
Proposed	F	Val.	0.344	0.186	0.664	0.155	0.912	0.689
Proposed	F	All	0.534	0.418	0.512	0.103	0.946	0.468
Proposed	R	Train	0.519	0.401	0.383	0.229	0.965	0.388
Proposed	R	Val.	0.166	0.026	0.756	0.078	0.854	0.799
Proposed	R	All	0.284	0.150	0.595	0.103	0.914	0.577
DeepPT	F	Train	0.599	0.426	0.408	0.121	0.943	0.427
DeepPT	F	Val.	0.184	0.067	0.680	0.165	0.899	0.724
DeepPT	F	All	0.375	0.189	0.579	0.140	0.920	0.587
DeepPT	R	Train	0.481	0.306	0.412	0.265	0.956	0.429
DeepPT	R	Val.	0.175	0.040	0.696	0.070	0.880	0.725
DeepPT	R	All	0.327	0.130	0.572	0.131	0.921	0.566
CatBoost	F	Train	0.887	0.548	0.280	0.067	0.977	0.278
CatBoost	F	Val.	0.329	0.126	0.621	0.123	0.918	0.654
CatBoost	F	All	0.523	0.300	0.504	0.089	0.945	0.481
CatBoost	R	Train	0.659	0.271	0.424	0.108	0.949	0.462
CatBoost	R	Val.	0.190	0.039	0.698	0.200	0.874	0.725
CatBoost	R	All	0.369	0.137	0.578	0.071	0.914	0.583
Ridge	F	Train	0.921	0.917	0.187	0.060	0.988	0.196
Ridge	F	Val.	0.335	0.188	0.678	0.139	0.908	0.703
Ridge	F	All	0.553	0.426	0.516	0.095	0.945	0.469
Ridge	R	Train	0.656	0.602	0.335	0.138	0.974	0.338
Ridge	R	Val.	0.224	0.095	0.817	0.097	0.844	0.857
Ridge	R	All	0.309	0.218	0.617	0.084	0.914	0.577
No aug + tune	F	Train	0.869	0.846	0.263	0.214	0.994	0.141
No aug + tune	F	Val.	0.347	0.171	0.681	0.136	0.904	0.708
No aug + tune	F	All	0.547	0.430	0.536	0.073	0.945	0.447
No aug + tune	R	Train	0.439	0.313	0.416	0.228	0.958	0.431
No aug + tune	R	Val.	0.035	0.009	0.778	0.095	0.868	0.778
No aug + tune	R	All	0.185	0.102	0.619	0.094	0.916	0.591
2x aug + frozen	F	Train	0.437	0.027	0.492	0.070	0.916	0.517
2x aug + frozen	F	Val.	0.122	0.009	0.707	0.209	0.891	0.756
2x aug + frozen	F	All	0.294	0.013	0.626	0.137	0.902	0.646
2x aug + frozen	R	Train	0.312	0.063	0.469	0.267	0.938	0.509
2x aug + frozen	R	Val.	0.120	0.016	0.702	0.078	0.876	0.730
2x aug + frozen	R	All	0.178	0.030	0.600	0.126	0.909	0.611
No aug + frozen	F	Train	0.437	0.160	0.467	0.079	0.926	0.490
No aug + frozen	F	Val.	0.122	0.023	0.708	0.203	0.893	0.753
No aug + frozen	F	All	0.247	0.061	0.618	0.095	0.908	0.632
No aug + frozen	R	Train	0.366	0.129	0.455	0.240	0.943	0.492
No aug + frozen	R	Val.	0.078	0.020	0.703	0.225	0.875	0.729

continued on next page

Table A.3: Model-level performance summary by transfer direction and split. The table is a compact version of `model_summary_by_direction.csv`; secondary standard-deviation columns are omitted to keep the appendix readable. (continued).

Model	Dir.	Split	Prot. PCC	Prot. SSIM	Prot. RMSE	Moran err.	Spot PCC	Spot RMSE
No aug + frozen	R	All	0.173	0.054	0.594	0.082	0.912	0.601

Table A.4: Correlation matrix distance by transfer direction and split. Lower CMD indicates better preservation of the measured cross-marker correlation structure.

Model	Dir.	Split	CMD
Proposed	F	Train	0.114
Proposed	F	Val.	0.382
Proposed	F	All	0.363
Proposed	R	Train	0.497
Proposed	R	Val.	0.711
Proposed	R	All	0.638
DeepPT	F	Train	0.371
DeepPT	F	Val.	0.581
DeepPT	F	All	0.517
DeepPT	R	Train	0.428
DeepPT	R	Val.	0.636
DeepPT	R	All	0.562
CatBoost	F	Train	0.098
CatBoost	F	Val.	0.326
CatBoost	F	All	0.353
CatBoost	R	Train	0.242
CatBoost	R	Val.	0.457
CatBoost	R	All	0.414
Ridge	F	Train	0.073
Ridge	F	Val.	0.348
Ridge	F	All	0.342
Ridge	R	Train	0.207
Ridge	R	Val.	0.407
Ridge	R	All	0.359
No aug + tune	F	Train	0.300
No aug + tune	F	Val.	0.356
No aug + tune	F	All	0.352

continued on next page

Table A.4: Correlation matrix distance by transfer direction and split. Lower CMD indicates better preservation of the measured cross-marker correlation structure. (continued).

Model	Dir.	Split	CMD
No aug + tune	R	Train	0.690
No aug + tune	R	Val.	0.904
No aug + tune	R	All	0.768
2x aug + frozen	F	Train	0.547
2x aug + frozen	F	Val.	0.698
2x aug + frozen	F	All	0.631
2x aug + frozen	R	Train	0.801
2x aug + frozen	R	Val.	0.985
2x aug + frozen	R	All	0.905
No aug + frozen	F	Train	0.501
No aug + frozen	F	Val.	0.857
No aug + frozen	F	All	0.649
No aug + frozen	R	Train	0.666
No aug + frozen	R	Val.	0.698
No aug + frozen	R	All	0.649

Table A.5: Validation Moran’s I values for the proposed model. True and predicted marker maps are compared in each transfer direction; absolute error is computed as the absolute difference between those two Moran’s I values.

Protein	F true	F pred.	F err.	R true	R pred.	R err.
ACTA2	0.838	0.706	0.132	0.777	0.709	0.068
BCL2	0.626	0.538	0.088	0.672	0.664	0.008
CCR7	0.546	0.701	0.155	0.729	0.725	0.004
CD14	0.563	0.521	0.042	0.660	0.725	0.065
CD163	0.449	0.684	0.235	0.757	0.730	0.027
CD19	0.633	0.683	0.050	0.744	0.733	0.011
CD27	0.516	0.734	0.218	0.628	0.744	0.116
CD274	0.658	0.594	0.064	0.656	0.628	0.028
CD3E	0.811	0.734	0.077	0.837	0.682	0.155
CD4	0.619	0.754	0.135	0.761	0.706	0.055
CD40	0.726	0.665	0.061	0.769	0.658	0.111
CD68	0.466	0.489	0.023	0.643	0.741	0.098

continued on next page

Table A.5: Validation Moran’s I values for the proposed model. True and predicted marker maps are compared in each transfer direction; absolute error is computed as the absolute difference between those two Moran’s I values. (continued).

Protein	F true	F pred.	F err.	R true	R pred.	R err.
CD8A	0.732	0.672	0.061	0.774	0.681	0.093
CEACAM8	0.570	0.760	0.190	0.671	0.624	0.047
CR2	0.538	0.657	0.119	0.794	0.735	0.059
CXCR5	0.727	0.764	0.037	0.682	0.733	0.050
EPCAM	0.629	0.518	0.111	0.642	0.611	0.031
FCGR3A	0.268	0.800	0.532	0.436	0.742	0.306
HLA-DRA	0.625	0.732	0.107	0.848	0.721	0.127
ITGAM	0.623	0.602	0.021	0.648	0.728	0.080
ITGAX	0.609	0.669	0.060	0.818	0.741	0.077
KRT5	0.871	0.800	0.070	0.799	0.622	0.177
MS4A1	0.657	0.693	0.037	0.780	0.729	0.051
PAX5	0.593	0.712	0.119	0.834	0.732	0.102
PCNA	0.603	0.566	0.037	0.715	0.701	0.014
PDCD1	0.656	0.763	0.108	0.705	0.740	0.035
PECAM1	0.597	0.773	0.176	0.679	0.732	0.053
SDC1	0.583	0.710	0.127	0.773	0.629	0.145
VIM	0.765	0.671	0.094	0.814	0.699	0.115
mouse_IgG1k	0.221	0.600	0.379	0.591	0.685	0.094
mouse_IgG2a	0.114	0.667	0.554	0.658	0.690	0.032
mouse_IgG2bk	0.100	0.631	0.532	0.699	0.615	0.084
rat_IgG2a	0.217	0.596	0.379	0.623	0.666	0.043

Table A.6: Held-out per-protein validation metrics for all compared models and both transfer directions. This table is derived from the validation rows of `per_protein_metrics_long.csv`; train and pooled rows are omitted because the main text uses held-out transfer as the primary benchmark.

Model	Dir.	Protein	PCC	Spearman	SSIM	RMSE	Moran err.
<i>Proposed, forward</i>							
Proposed	F	ACTA2	0.533	0.558	-0.354	0.776	0.132
Proposed	F	BCL2	0.480	0.432	0.466	0.466	0.088
Proposed	F	CCR7	0.380	0.354	0.373	0.386	0.155
Proposed	F	CD14	0.076	0.027	0.035	0.897	0.042
Proposed	F	CD163	0.264	0.281	0.174	0.628	0.235
Proposed	F	CD19	0.457	0.445	0.443	0.265	0.050

continued on next page

Table A.6: Held-out per-protein validation metrics for all compared models and both transfer directions. This table is derived from the validation rows of `per_protein_metrics_long.csv`; train and pooled rows are omitted because the main text uses held-out transfer as the primary benchmark. (continued).

Model	Dir.	Protein	PCC	Spearman	SSIM	RMSE	Moran err.
Proposed	F	CD27	0.266	0.273	0.233	0.498	0.218
Proposed	F	CD274	0.519	0.403	0.391	0.532	0.064
Proposed	F	CD3E	0.539	0.540	0.530	0.683	0.077
Proposed	F	CD4	0.508	0.533	0.499	0.351	0.135
Proposed	F	CD40	-0.142	-0.157	-0.117	0.619	0.061
Proposed	F	CD68	0.301	0.286	0.279	0.368	0.023
Proposed	F	CD8A	0.479	0.483	0.312	1.121	0.061
Proposed	F	CEACAM8	0.567	0.481	0.539	0.671	0.190
Proposed	F	CR2	0.361	0.410	0.269	0.728	0.119
Proposed	F	CXCR5	0.701	0.724	-0.615	0.612	0.037
Proposed	F	EPCAM	0.086	0.093	-0.042	0.753	0.111
Proposed	F	FCGR3A	0.288	0.486	0.195	1.256	0.532
Proposed	F	HLA-DRA	0.399	0.396	0.391	0.417	0.107
Proposed	F	ITGAM	0.447	0.350	0.240	0.659	0.021
Proposed	F	ITGAX	0.231	0.226	0.233	0.448	0.060
Proposed	F	KRT5	0.660	0.570	0.440	1.454	0.070
Proposed	F	MS4A1	0.437	0.444	0.410	0.436	0.037
Proposed	F	PAX5	0.523	0.546	0.475	0.704	0.119
Proposed	F	PCNA	0.341	0.267	0.291	0.434	0.037
Proposed	F	PDCD1	0.635	0.678	-0.377	0.776	0.108
Proposed	F	PECAM1	0.359	0.378	-0.004	0.939	0.176
Proposed	F	SDC1	-0.070	-0.122	-0.047	0.568	0.127
Proposed	F	VIM	0.308	0.344	0.227	0.509	0.094
Proposed	F	mouse_IgG1k	0.108	0.125	0.074	0.586	0.379
Proposed	F	mouse_IgG2a	0.105	0.131	0.081	0.759	0.554
Proposed	F	mouse_IgG2bk	0.011	0.116	0.006	0.853	0.532
Proposed	F	rat_IgG2a	0.207	0.267	0.087	0.773	0.379
<i>Proposed, reverse</i>							
Proposed	R	ACTA2	0.329	0.344	-0.261	1.185	0.068
Proposed	R	BCL2	0.088	0.105	0.026	0.711	0.008
Proposed	R	CCR7	0.045	0.033	0.026	0.432	0.004
Proposed	R	CD14	-0.318	-0.369	-0.039	0.787	0.065
Proposed	R	CD163	-0.104	-0.103	-0.038	0.603	0.027
Proposed	R	CD19	0.398	0.469	0.252	0.330	0.011
Proposed	R	CD27	0.095	0.127	0.058	0.532	0.116
Proposed	R	CD274	0.274	0.247	0.083	0.523	0.028
Proposed	R	CD3E	0.323	0.299	0.136	1.264	0.155
Proposed	R	CD4	0.355	0.419	0.097	0.377	0.055
Proposed	R	CD40	-0.198	-0.252	-0.018	0.746	0.111
Proposed	R	CD68	0.362	0.349	0.102	0.560	0.098
Proposed	R	CD8A	0.348	0.340	0.042	1.822	0.093
Proposed	R	CEACAM8	0.248	0.118	0.049	1.114	0.047
Proposed	R	CR2	0.507	0.499	0.233	0.599	0.059
Proposed	R	CXCR5	0.400	0.450	-0.268	0.869	0.050

continued on next page

Table A.6: Held-out per-protein validation metrics for all compared models and both transfer directions. This table is derived from the validation rows of `per_protein_metrics_long.csv`; train and pooled rows are omitted because the main text uses held-out transfer as the primary benchmark. (continued).

Model	Dir.	Protein	PCC	Spearman	SSIM	RMSE	Moran err.
Proposed	R	EPCAM	0.113	0.152	-0.013	0.521	0.031
Proposed	R	FCGR3A	0.121	0.137	0.098	0.666	0.306
Proposed	R	HLA-DRA	-0.120	-0.064	-0.045	0.775	0.127
Proposed	R	ITGAM	0.232	0.297	0.145	0.665	0.080
Proposed	R	ITGAX	-0.157	-0.194	-0.030	0.725	0.077
Proposed	R	KRT5	0.211	0.151	0.093	1.603	0.177
Proposed	R	MS4A1	0.054	0.055	0.026	0.596	0.051
Proposed	R	PAX5	0.408	0.464	0.142	1.112	0.102
Proposed	R	PCNA	-0.031	-0.017	-0.012	0.519	0.014
Proposed	R	PDCD1	0.559	0.628	-0.220	0.705	0.035
Proposed	R	PECAM1	0.242	0.252	-0.060	1.171	0.053
Proposed	R	SDC1	-0.144	-0.184	-0.041	0.464	0.145
Proposed	R	VIM	0.440	0.460	0.234	0.493	0.115
Proposed	R	mouse_IgG1k	-0.212	-0.227	-0.014	0.528	0.094
Proposed	R	mouse_IgG2a	0.215	0.232	0.040	0.505	0.032
Proposed	R	mouse_IgG2bk	0.161	0.235	0.020	0.806	0.084
Proposed	R	rat_IgG2a	0.239	0.206	0.027	0.637	0.043
<i>DeepPT, forward</i>							
DeepPT	F	ACTA2	0.317	0.331	-0.066	0.975	0.167
DeepPT	F	BCL2	0.318	0.292	0.216	0.500	0.089
DeepPT	F	CCR7	0.227	0.237	0.156	0.318	0.187
DeepPT	F	CD14	-0.093	-0.107	-0.037	0.774	0.134
DeepPT	F	CD163	0.183	0.186	0.113	0.512	0.198
DeepPT	F	CD19	0.192	0.195	0.087	0.349	0.025
DeepPT	F	CD27	-0.003	0.047	0.007	0.550	0.247
DeepPT	F	CD274	0.361	0.217	0.193	0.544	0.078
DeepPT	F	CD3E	0.475	0.419	0.262	0.876	0.034
DeepPT	F	CD4	0.327	0.294	0.305	0.275	0.120
DeepPT	F	CD40	-0.143	-0.144	-0.042	0.645	0.078
DeepPT	F	CD68	0.003	-0.004	0.007	0.468	0.217
DeepPT	F	CD8A	0.452	0.419	0.042	1.460	0.000
DeepPT	F	CEACAM8	0.553	0.460	0.478	0.653	0.264
DeepPT	F	CR2	0.083	0.065	0.069	0.552	0.188
DeepPT	F	CXCR5	0.567	0.580	-0.353	0.629	0.096
DeepPT	F	EPCAM	-0.230	-0.222	0.010	0.603	0.045
DeepPT	F	FCGR3A	0.123	0.300	0.039	1.260	0.526
DeepPT	F	HLA-DRA	0.155	0.138	0.125	0.588	0.032
DeepPT	F	ITGAM	0.246	0.222	0.086	0.588	0.043
DeepPT	F	ITGAX	0.026	0.017	0.021	0.544	0.043
DeepPT	F	KRT5	0.632	0.518	0.335	1.522	0.036
DeepPT	F	MS4A1	0.182	0.168	0.102	0.476	0.006
DeepPT	F	PAX5	0.325	0.360	0.086	1.062	0.090
DeepPT	F	PCNA	0.170	0.171	0.102	0.436	0.057
DeepPT	F	PDCD1	0.363	0.361	-0.046	0.662	0.112

continued on next page

Table A.6: Held-out per-protein validation metrics for all compared models and both transfer directions. This table is derived from the validation rows of `per_protein_metrics_long.csv`; train and pooled rows are omitted because the main text uses held-out transfer as the primary benchmark. (continued).

Model	Dir.	Protein	PCC	Spearman	SSIM	RMSE	Moran err.
DeepPT	F	PECAM1	0.368	0.386	-0.060	1.071	0.147
DeepPT	F	SDC1	-0.174	-0.190	-0.089	0.447	0.051
DeepPT	F	VIM	-0.110	-0.110	-0.066	0.456	0.117
DeepPT	F	mouse_IgG1k	0.105	0.104	0.070	0.452	0.476
DeepPT	F	mouse_IgG2a	-0.006	-0.016	0.005	0.546	0.532
DeepPT	F	mouse_IgG2bk	0.061	0.019	0.025	1.007	0.561
DeepPT	F	rat_IgG2a	0.009	0.021	0.013	0.643	0.451
<i>DeepPT, reverse</i>							
DeepPT	R	ACTA2	0.245	0.267	-0.140	0.817	0.045
DeepPT	R	BCL2	0.290	0.305	0.095	0.641	0.071
DeepPT	R	CCR7	0.246	0.278	0.060	0.395	0.017
DeepPT	R	CD14	-0.057	-0.048	-0.006	0.770	0.047
DeepPT	R	CD163	0.150	0.153	0.047	0.522	0.026
DeepPT	R	CD19	0.068	0.148	0.040	0.337	0.024
DeepPT	R	CD27	0.024	0.043	0.016	0.525	0.069
DeepPT	R	CD274	0.328	0.238	0.190	0.474	0.056
DeepPT	R	CD3E	0.351	0.369	0.180	1.039	0.082
DeepPT	R	CD4	0.311	0.326	0.103	0.344	0.060
DeepPT	R	CD40	-0.108	-0.092	-0.018	0.678	0.138
DeepPT	R	CD68	-0.020	-0.007	-0.000	0.593	0.010
DeepPT	R	CD8A	0.253	0.265	0.038	1.550	0.022
DeepPT	R	CEACAM8	0.537	0.350	0.272	0.963	0.069
DeepPT	R	CR2	-0.005	0.031	0.008	0.571	0.132
DeepPT	R	CXCR5	0.507	0.483	-0.289	0.781	0.038
DeepPT	R	EPCAM	-0.108	-0.139	0.015	0.586	0.109
DeepPT	R	FCGR3A	0.319	0.340	0.176	0.581	0.243
DeepPT	R	HLA-DRA	0.219	0.209	0.042	0.766	0.159
DeepPT	R	ITGAM	0.113	0.103	0.025	0.678	0.049
DeepPT	R	ITGAX	0.256	0.228	0.032	0.778	0.167
DeepPT	R	KRT5	0.482	0.381	0.410	1.241	0.046
DeepPT	R	MS4A1	0.247	0.230	0.048	0.602	0.093
DeepPT	R	PAX5	0.289	0.303	0.053	1.085	0.147
DeepPT	R	PCNA	0.212	0.225	0.060	0.489	0.046
DeepPT	R	PDCD1	0.344	0.381	-0.105	0.668	0.003
DeepPT	R	PECAM1	0.306	0.302	-0.045	1.039	0.031
DeepPT	R	SDC1	-0.027	-0.089	-0.013	0.460	0.026
DeepPT	R	VIM	-0.047	-0.033	-0.024	0.522	0.053
DeepPT	R	mouse_IgG1k	0.050	0.091	0.016	0.475	0.056
DeepPT	R	mouse_IgG2a	-0.038	-0.084	0.001	0.520	0.076
DeepPT	R	mouse_IgG2bk	0.043	0.066	0.018	0.842	0.031
DeepPT	R	rat_IgG2a	0.012	0.025	0.011	0.627	0.079
<i>CatBoost, forward</i>							
CatBoost	F	ACTA2	0.555	0.592	-0.312	0.820	0.127
CatBoost	F	BCL2	0.497	0.447	0.312	0.449	0.122

continued on next page

Table A.6: Held-out per-protein validation metrics for all compared models and both transfer directions. This table is derived from the validation rows of `per_protein_metrics_long.csv`; train and pooled rows are omitted because the main text uses held-out transfer as the primary benchmark. (continued).

Model	Dir.	Protein	PCC	Spearman	SSIM	RMSE	Moran err.
CatBoost	F	CCR7	0.327	0.321	0.036	0.332	0.098
CatBoost	F	CD14	0.114	0.110	0.015	0.794	0.228
CatBoost	F	CD163	0.142	0.182	0.019	0.505	0.026
CatBoost	F	CD19	0.464	0.459	0.415	0.262	0.031
CatBoost	F	CD27	0.239	0.245	0.080	0.529	0.194
CatBoost	F	CD274	0.562	0.430	0.320	0.526	0.026
CatBoost	F	CD3E	0.576	0.588	0.534	0.777	0.068
CatBoost	F	CD4	0.469	0.524	0.058	0.310	0.011
CatBoost	F	CD40	-0.167	-0.171	-0.142	0.546	0.032
CatBoost	F	CD68	0.319	0.300	0.186	0.341	0.036
CatBoost	F	CD8A	0.510	0.512	0.245	1.254	0.062
CatBoost	F	CEACAM8	0.547	0.489	0.419	0.667	0.138
CatBoost	F	CR2	0.298	0.310	0.025	0.612	0.005
CatBoost	F	CXCR5	0.704	0.732	-0.222	0.511	0.038
CatBoost	F	EPCAM	-0.050	-0.032	-0.002	0.619	0.227
CatBoost	F	FCGR3A	0.263	0.484	0.098	1.231	0.476
CatBoost	F	HLA-DRA	0.355	0.350	0.311	0.392	0.082
CatBoost	F	ITGAM	0.444	0.347	0.212	0.579	0.074
CatBoost	F	ITGAX	0.208	0.205	0.190	0.391	0.041
CatBoost	F	KRT5	0.671	0.576	0.476	1.404	0.081
CatBoost	F	MS4A1	0.401	0.416	0.344	0.392	0.001
CatBoost	F	PAX5	0.502	0.529	0.379	0.821	0.109
CatBoost	F	PCNA	0.225	0.152	0.086	0.427	0.029
CatBoost	F	PDCD1	0.552	0.617	-0.015	0.743	0.024
CatBoost	F	PECAM1	0.333	0.363	-0.011	0.953	0.164
CatBoost	F	SDC1	0.058	-0.006	0.008	0.409	0.150
CatBoost	F	VIM	0.254	0.284	0.033	0.407	0.140
CatBoost	F	mouse_IgG1k	0.141	0.157	0.014	0.486	0.156
CatBoost	F	mouse_IgG2a	0.118	0.158	0.015	0.537	0.358
CatBoost	F	mouse_IgG2bk	0.065	0.121	0.026	0.826	0.533
CatBoost	F	rat_IgG2a	0.162	0.222	0.014	0.655	0.173
<i>CatBoost, reverse</i>							
CatBoost	R	ACTA2	0.337	0.336	-0.019	0.814	0.166
CatBoost	R	BCL2	0.124	-0.040	0.004	0.656	0.279
CatBoost	R	CCR7	0.234	0.235	0.025	0.393	0.156
CatBoost	R	CD14	-0.136	-0.112	-0.025	0.780	0.084
CatBoost	R	CD163	-0.080	-0.120	0.007	0.527	0.172
CatBoost	R	CD19	0.464	0.485	0.231	0.315	0.028
CatBoost	R	CD27	-0.011	0.012	0.001	0.496	0.092
CatBoost	R	CD274	0.195	0.223	0.014	0.499	0.261
CatBoost	R	CD3E	0.327	0.313	0.014	1.041	0.224
CatBoost	R	CD4	0.301	0.317	0.019	0.372	0.247
CatBoost	R	CD40	-0.020	-0.005	0.001	0.699	0.288
CatBoost	R	CD68	0.411	0.400	0.230	0.506	0.017

continued on next page

Table A.6: Held-out per-protein validation metrics for all compared models and both transfer directions. This table is derived from the validation rows of `per_protein_metrics_long.csv`; train and pooled rows are omitted because the main text uses held-out transfer as the primary benchmark. (continued).

Model	Dir.	Protein	PCC	Spearman	SSIM	RMSE	Moran err.
CatBoost	R	CD8A	0.353	0.352	0.004	1.504	0.182
CatBoost	R	CEACAM8	0.211	0.073	0.004	1.050	0.241
CatBoost	R	CR2	0.374	0.393	0.274	0.572	0.242
CatBoost	R	CXCR5	0.387	0.397	-0.021	0.844	0.074
CatBoost	R	EPCAM	0.339	0.420	-0.035	0.551	0.095
CatBoost	R	FCGR3A	0.200	0.165	0.098	0.593	0.257
CatBoost	R	HLA-DRA	-0.151	-0.131	-0.001	0.776	0.324
CatBoost	R	ITGAM	0.279	0.275	0.147	0.651	0.041
CatBoost	R	ITGAX	-0.022	0.040	0.001	0.745	0.278
CatBoost	R	KRT5	0.264	0.154	0.057	1.377	0.284
CatBoost	R	MS4A1	0.176	0.128	0.123	0.583	0.118
CatBoost	R	PAX5	0.438	0.439	0.090	1.075	0.125
CatBoost	R	PCNA	0.120	-0.029	0.023	0.488	0.234
CatBoost	R	PDCD1	0.532	0.582	-0.019	0.706	0.043
CatBoost	R	PECAM1	0.330	0.304	-0.021	1.042	0.036
CatBoost	R	SDC1	-0.056	0.165	0.004	0.429	0.376
CatBoost	R	VIM	0.345	0.394	0.067	0.474	0.172
CatBoost	R	mouse_IgG1k	0.022	0.089	0.010	0.494	0.505
CatBoost	R	mouse_IgG2a	-0.093	-0.008	-0.038	0.510	0.416
CatBoost	R	mouse_IgG2bk	0.097	0.323	0.027	0.824	0.204
CatBoost	R	rat_IgG2a	-0.015	0.121	-0.001	0.640	0.345
<i>Ridge, forward</i>							
Ridge	F	ACTA2	0.532	0.554	-0.331	0.809	0.135
Ridge	F	BCL2	0.471	0.426	0.459	0.472	0.103
Ridge	F	CCR7	0.364	0.325	0.346	0.442	0.136
Ridge	F	CD14	0.099	0.034	0.053	0.918	0.009
Ridge	F	CD163	0.304	0.324	0.186	0.653	0.169
Ridge	F	CD19	0.413	0.411	0.369	0.283	0.018
Ridge	F	CD27	0.245	0.248	0.210	0.532	0.211
Ridge	F	CD274	0.546	0.419	0.420	0.534	0.057
Ridge	F	CD3E	0.580	0.578	0.549	0.722	0.073
Ridge	F	CD4	0.523	0.559	0.501	0.394	0.122
Ridge	F	CD40	-0.162	-0.169	-0.133	0.566	0.058
Ridge	F	CD68	0.302	0.297	0.290	0.375	0.027
Ridge	F	CD8A	0.560	0.554	0.238	1.260	0.062
Ridge	F	CEACAM8	0.570	0.481	0.537	0.782	0.149
Ridge	F	CR2	0.312	0.360	0.234	0.713	0.105
Ridge	F	CXCR5	0.723	0.746	-0.270	0.510	0.044
Ridge	F	EPCAM	0.059	0.051	-0.041	0.792	0.083
Ridge	F	FCGR3A	0.287	0.477	0.200	1.247	0.524
Ridge	F	HLA-DRA	0.378	0.369	0.368	0.407	0.062
Ridge	F	ITGAM	0.433	0.343	0.245	0.632	0.023
Ridge	F	ITGAX	0.257	0.259	0.257	0.442	0.036
Ridge	F	KRT5	0.682	0.594	0.506	1.360	0.078

continued on next page

Table A.6: Held-out per-protein validation metrics for all compared models and both transfer directions. This table is derived from the validation rows of `per_protein_metrics_long.csv`; train and pooled rows are omitted because the main text uses held-out transfer as the primary benchmark. (continued).

Model	Dir.	Protein	PCC	Spearman	SSIM	RMSE	Moran err.
Ridge	F	MS4A1	0.410	0.409	0.390	0.408	0.029
Ridge	F	PAX5	0.508	0.540	0.414	0.786	0.133
Ridge	F	PCNA	0.243	0.153	0.199	0.516	0.082
Ridge	F	PDCD1	0.597	0.636	-0.330	0.773	0.089
Ridge	F	PECAM1	0.333	0.364	-0.012	0.955	0.166
Ridge	F	SDC1	-0.171	-0.212	-0.116	0.601	0.093
Ridge	F	VIM	0.254	0.285	0.195	0.557	0.083
Ridge	F	mouse_IgG1k	0.112	0.125	0.090	0.527	0.314
Ridge	F	mouse_IgG2a	0.118	0.137	0.093	0.755	0.470
Ridge	F	mouse_IgG2bk	0.018	0.119	0.010	0.831	0.492
Ridge	F	rat_IgG2a	0.166	0.212	0.085	0.805	0.340
<i>Ridge, reverse</i>							
Ridge	R	ACTA2	0.251	0.318	-0.249	1.024	0.108
Ridge	R	BCL2	0.172	0.204	0.101	0.706	0.212
Ridge	R	CCR7	0.204	0.269	0.179	0.587	0.047
Ridge	R	CD14	0.011	0.110	0.012	0.846	0.036
Ridge	R	CD163	0.050	-0.005	0.044	0.592	0.059
Ridge	R	CD19	0.284	0.416	0.244	0.393	0.110
Ridge	R	CD27	-0.086	-0.053	-0.064	0.473	0.009
Ridge	R	CD274	0.140	0.269	0.129	0.604	0.054
Ridge	R	CD3E	0.333	0.315	0.198	1.473	0.175
Ridge	R	CD4	0.391	0.492	0.366	0.607	0.013
Ridge	R	CD40	-0.052	-0.046	-0.034	0.712	0.097
Ridge	R	CD68	0.440	0.471	0.348	0.484	0.062
Ridge	R	CD8A	0.342	0.346	0.049	1.914	0.196
Ridge	R	CEACAM8	0.296	0.168	0.152	1.054	0.188
Ridge	R	CR2	0.420	0.449	0.320	0.626	0.065
Ridge	R	CXCR5	0.410	0.439	-0.348	0.937	0.016
Ridge	R	EPCAM	0.292	0.471	-0.054	0.650	0.072
Ridge	R	FCGR3A	0.172	0.168	0.173	0.791	0.224
Ridge	R	HLA-DRA	-0.047	-0.109	-0.025	0.847	0.314
Ridge	R	ITGAM	0.096	0.176	0.080	0.787	0.037
Ridge	R	ITGAX	0.102	0.031	0.053	0.827	0.113
Ridge	R	KRT5	0.209	0.218	0.113	2.127	0.057
Ridge	R	MS4A1	0.186	0.271	0.164	0.596	0.097
Ridge	R	PAX5	0.257	0.371	0.130	1.237	0.168
Ridge	R	PCNA	0.208	0.314	0.196	0.597	0.047
Ridge	R	PDCD1	0.539	0.617	-0.206	0.812	0.029
Ridge	R	PECAM1	0.220	0.265	-0.067	1.171	0.022
Ridge	R	SDC1	-0.144	-0.221	-0.097	0.536	0.149
Ridge	R	VIM	0.505	0.514	0.364	0.434	0.100
Ridge	R	mouse_IgG1k	0.174	0.365	0.148	0.564	0.119
Ridge	R	mouse_IgG2a	0.351	0.464	0.229	0.572	0.043
Ridge	R	mouse_IgG2bk	0.409	0.521	0.360	0.643	0.024

continued on next page

Table A.6: Held-out per-protein validation metrics for all compared models and both transfer directions. This table is derived from the validation rows of `per_protein_metrics_long.csv`; train and pooled rows are omitted because the main text uses held-out transfer as the primary benchmark. (continued).

Model	Dir.	Protein	PCC	Spearman	SSIM	RMSE	Moran err.
Ridge	R	rat.IgG2a	0.265	0.453	0.123	0.748	0.131
<i>No aug + tune, forward</i>							
No aug + tune	F	ACTA2	0.566	0.585	-0.270	0.874	0.139
No aug + tune	F	BCL2	0.487	0.445	0.473	0.489	0.025
No aug + tune	F	CCR7	0.343	0.302	0.315	0.431	0.133
No aug + tune	F	CD14	0.038	-0.011	0.024	0.891	0.031
No aug + tune	F	CD163	0.175	0.190	0.112	0.616	0.127
No aug + tune	F	CD19	0.413	0.414	0.351	0.270	0.040
No aug + tune	F	CD27	0.134	0.163	0.107	0.623	0.198
No aug + tune	F	CD274	0.545	0.439	0.393	0.577	0.012
No aug + tune	F	CD3E	0.607	0.618	0.514	0.923	0.070
No aug + tune	F	CD4	0.552	0.594	0.520	0.410	0.122
No aug + tune	F	CD40	-0.144	-0.171	-0.103	0.575	0.113
No aug + tune	F	CD68	0.351	0.356	0.333	0.360	0.053
No aug + tune	F	CD8A	0.590	0.589	0.159	1.377	0.053
No aug + tune	F	CEACAM8	0.576	0.470	0.575	0.709	0.193
No aug + tune	F	CR2	0.354	0.400	0.248	0.708	0.129
No aug + tune	F	CXCR5	0.696	0.707	-0.658	0.662	0.037
No aug + tune	F	EPCAM	0.258	0.290	-0.093	0.688	0.080
No aug + tune	F	FCGR3A	0.271	0.435	0.167	1.226	0.440
No aug + tune	F	HLA-DRA	0.294	0.278	0.276	0.437	0.016
No aug + tune	F	ITGAM	0.362	0.257	0.301	0.609	0.022
No aug + tune	F	ITGAX	0.146	0.146	0.147	0.486	0.072
No aug + tune	F	KRT5	0.679	0.569	0.617	1.260	0.078
No aug + tune	F	MS4A1	0.393	0.381	0.340	0.408	0.014
No aug + tune	F	PAX5	0.430	0.453	0.301	0.883	0.107
No aug + tune	F	PCNA	0.402	0.334	0.317	0.423	0.156
No aug + tune	F	PDCD1	0.627	0.671	-0.384	0.828	0.082
No aug + tune	F	PECAM1	0.337	0.326	-0.059	0.994	0.082
No aug + tune	F	SDC1	0.015	-0.030	0.016	0.515	0.046
No aug + tune	F	VIM	0.381	0.386	0.279	0.447	0.083
No aug + tune	F	mouse.IgG1k	0.174	0.193	0.115	0.497	0.297
No aug + tune	F	mouse.IgG2a	0.113	0.132	0.084	0.645	0.562
No aug + tune	F	mouse.IgG2bk	0.044	0.127	0.020	0.874	0.496
No aug + tune	F	rat.IgG2a	0.229	0.307	0.098	0.741	0.380
<i>No aug + tune, reverse</i>							
No aug + tune	R	ACTA2	0.167	0.294	-0.066	0.795	0.045
No aug + tune	R	BCL2	0.045	0.162	0.024	0.714	0.037
No aug + tune	R	CCR7	0.078	0.143	0.022	0.383	0.012
No aug + tune	R	CD14	0.001	0.078	0.007	0.790	0.207
No aug + tune	R	CD163	-0.033	0.020	-0.001	0.498	0.003
No aug + tune	R	CD19	0.162	0.307	0.157	0.330	0.022
No aug + tune	R	CD27	0.110	0.046	0.101	0.473	0.131
No aug + tune	R	CD274	0.052	0.237	0.051	0.668	0.108

continued on next page

Table A.6: Held-out per-protein validation metrics for all compared models and both transfer directions. This table is derived from the validation rows of `per_protein_metrics_long.csv`; train and pooled rows are omitted because the main text uses held-out transfer as the primary benchmark. (continued).

Model	Dir.	Protein	PCC	Spearman	SSIM	RMSE	Moran err.
No aug + tune	R	CD3E	0.014	0.067	0.007	1.119	0.208
No aug + tune	R	CD4	-0.005	0.156	0.004	0.449	0.005
No aug + tune	R	CD40	-0.036	-0.059	-0.014	0.757	0.003
No aug + tune	R	CD68	0.018	0.111	0.018	0.605	0.133
No aug + tune	R	CD8A	0.006	-0.056	0.001	1.596	0.217
No aug + tune	R	CEACAM8	0.093	0.231	0.033	1.086	0.095
No aug + tune	R	CR2	-0.003	0.153	0.006	0.688	0.037
No aug + tune	R	CXCR5	0.061	0.400	-0.056	1.175	0.071
No aug + tune	R	EPCAM	-0.021	-0.007	-0.001	0.585	0.076
No aug + tune	R	FCGR3A	-0.007	0.244	-0.006	0.945	0.320
No aug + tune	R	HLA-DRA	-0.002	0.074	0.001	0.856	0.073
No aug + tune	R	ITGAM	0.139	0.194	0.030	0.697	0.066
No aug + tune	R	ITGAX	0.102	0.136	0.008	0.737	0.243
No aug + tune	R	KRT5	-0.030	0.199	-0.026	2.063	0.068
No aug + tune	R	MS4A1	-0.021	0.097	-0.008	0.696	0.021
No aug + tune	R	PAX5	0.141	0.337	0.064	1.019	0.085
No aug + tune	R	PCNA	0.043	0.011	0.026	0.506	0.047
No aug + tune	R	PDCD1	0.045	0.310	-0.018	0.903	0.045
No aug + tune	R	PECAM1	0.212	0.330	-0.024	1.012	0.071
No aug + tune	R	SDC1	-0.091	-0.129	-0.074	0.522	0.045
No aug + tune	R	VIM	-0.091	-0.069	-0.020	0.494	0.329
No aug + tune	R	mouse_IgG1k	-0.013	0.073	0.006	0.510	0.120
No aug + tune	R	mouse_IgG2a	-0.038	-0.019	0.000	0.529	0.007
No aug + tune	R	mouse_IgG2bk	0.032	0.196	0.023	0.774	0.045
No aug + tune	R	rat_IgG2a	0.030	0.121	0.016	0.714	0.138
<i>2x aug + frozen, forward</i>							
2x aug + frozen	F	ACTA2	0.247	0.230	-0.003	0.996	0.010
2x aug + frozen	F	BCL2	0.379	0.390	0.034	0.472	0.202
2x aug + frozen	F	CCR7	0.185	0.244	0.021	0.337	0.220
2x aug + frozen	F	CD14	-0.292	-0.285	0.005	0.802	0.198
2x aug + frozen	F	CD163	0.124	0.115	0.017	0.503	0.295
2x aug + frozen	F	CD19	-0.224	-0.236	0.009	0.325	0.099
2x aug + frozen	F	CD27	-0.172	-0.212	0.004	0.541	0.298
2x aug + frozen	F	CD274	0.358	0.256	0.014	0.581	0.126
2x aug + frozen	F	CD3E	0.494	0.428	0.019	0.952	0.027
2x aug + frozen	F	CD4	0.396	0.535	0.042	0.315	0.200
2x aug + frozen	F	CD40	0.213	0.192	0.017	0.707	0.007
2x aug + frozen	F	CD68	-0.121	-0.109	0.003	0.398	0.244
2x aug + frozen	F	CD8A	0.538	0.502	0.003	1.472	0.070
2x aug + frozen	F	CEACAM8	0.472	0.365	0.042	0.867	0.281
2x aug + frozen	F	CR2	-0.052	-0.049	0.011	0.609	0.227
2x aug + frozen	F	CXCR5	0.581	0.568	-0.032	0.849	0.125
2x aug + frozen	F	EPCAM	-0.167	-0.171	-0.002	0.616	0.068
2x aug + frozen	F	FCGR3A	0.210	0.401	0.004	1.249	0.592

continued on next page

Table A.6: Held-out per-protein validation metrics for all compared models and both transfer directions. This table is derived from the validation rows of `per_protein_metrics_long.csv`; train and pooled rows are omitted because the main text uses held-out transfer as the primary benchmark. (continued).

Model	Dir.	Protein	PCC	Spearman	SSIM	RMSE	Moran err.
2x aug + frozen	F	HLA-DRA	0.067	0.076	0.007	0.617	0.162
2x aug + frozen	F	ITGAM	-0.035	-0.023	0.002	0.615	0.113
2x aug + frozen	F	ITGAX	0.012	0.007	0.005	0.573	0.082
2x aug + frozen	F	KRT5	0.641	0.545	0.035	1.690	0.037
2x aug + frozen	F	MS4A1	0.134	0.126	0.006	0.503	0.137
2x aug + frozen	F	PAX5	0.151	0.167	0.003	0.989	0.019
2x aug + frozen	F	PCNA	0.113	0.109	0.007	0.438	0.117
2x aug + frozen	F	PDCD1	0.359	0.459	-0.010	0.747	0.158
2x aug + frozen	F	PECAM1	0.015	0.085	-0.002	1.054	0.171
2x aug + frozen	F	SDC1	-0.380	-0.330	-0.001	0.411	0.122
2x aug + frozen	F	VIM	-0.277	-0.280	0.001	0.413	0.068
2x aug + frozen	F	mouse_IgG1k	0.041	0.048	0.011	0.486	0.550
2x aug + frozen	F	mouse_IgG2a	-0.073	-0.093	0.009	0.526	0.636
2x aug + frozen	F	mouse_IgG2bk	0.152	0.034	0.004	1.015	0.651
2x aug + frozen	F	rat_IgG2a	-0.063	-0.006	0.009	0.651	0.570
<i>2x aug + frozen, reverse</i>							
2x aug + frozen	R	ACTA2	0.106	0.142	-0.016	0.797	0.118
2x aug + frozen	R	BCL2	0.218	0.219	0.036	0.645	0.006
2x aug + frozen	R	CCR7	0.197	0.284	0.017	0.386	0.069
2x aug + frozen	R	CD14	-0.177	-0.260	-0.006	0.812	0.012
2x aug + frozen	R	CD163	0.084	0.128	0.013	0.515	0.140
2x aug + frozen	R	CD19	0.032	0.126	0.017	0.337	0.053
2x aug + frozen	R	CD27	-0.208	-0.152	-0.015	0.515	0.085
2x aug + frozen	R	CD274	0.234	0.177	0.054	0.489	0.035
2x aug + frozen	R	CD3E	0.272	0.270	0.053	1.001	0.159
2x aug + frozen	R	CD4	0.349	0.372	0.052	0.361	0.062
2x aug + frozen	R	CD40	0.132	0.134	0.016	0.716	0.035
2x aug + frozen	R	CD68	-0.147	-0.162	-0.002	0.562	0.001
2x aug + frozen	R	CD8A	0.169	0.158	0.012	1.473	0.080
2x aug + frozen	R	CEACAM8	0.534	0.348	0.085	0.996	0.013
2x aug + frozen	R	CR2	0.085	0.155	0.013	0.615	0.078
2x aug + frozen	R	CXCR5	0.508	0.469	-0.133	0.806	0.007
2x aug + frozen	R	EPCAM	-0.136	-0.228	0.007	0.626	0.039
2x aug + frozen	R	FCGR3A	0.400	0.358	0.105	0.587	0.272
2x aug + frozen	R	HLA-DRA	0.127	0.098	0.012	0.782	0.162
2x aug + frozen	R	ITGAM	0.030	0.113	0.003	0.689	0.065
2x aug + frozen	R	ITGAX	0.156	0.097	0.006	0.738	0.216
2x aug + frozen	R	KRT5	0.487	0.412	0.190	1.308	0.103
2x aug + frozen	R	MS4A1	0.056	0.067	0.006	0.604	0.102
2x aug + frozen	R	PAX5	0.107	0.218	0.005	1.093	0.141
2x aug + frozen	R	PCNA	0.055	0.153	0.008	0.494	0.027
2x aug + frozen	R	PDCD1	0.322	0.375	-0.028	0.692	0.010
2x aug + frozen	R	PECAM1	0.009	0.109	-0.002	1.042	0.033
2x aug + frozen	R	SDC1	0.016	-0.033	0.009	0.433	0.087

continued on next page

Table A.6: Held-out per-protein validation metrics for all compared models and both transfer directions. This table is derived from the validation rows of `per_protein_metrics_long.csv`; train and pooled rows are omitted because the main text uses held-out transfer as the primary benchmark. (continued).

Model	Dir.	Protein	PCC	Spearman	SSIM	RMSE	Moran err.
2x aug + frozen	R	VIM	-0.124	-0.158	-0.024	0.490	0.144
2x aug + frozen	R	mouse_IgG1k	0.144	0.162	0.025	0.494	0.089
2x aug + frozen	R	mouse_IgG2a	-0.072	-0.071	0.003	0.542	0.012
2x aug + frozen	R	mouse_IgG2bk	0.077	0.155	0.017	0.832	0.011
2x aug + frozen	R	rat_IgG2a	-0.077	-0.024	0.006	0.677	0.096
<i>No aug + frozen, forward</i>							
No aug + frozen	F	ACTA2	0.316	0.339	-0.017	1.000	0.044
No aug + frozen	F	BCL2	0.347	0.325	0.136	0.492	0.182
No aug + frozen	F	CCR7	0.128	0.116	0.043	0.358	0.226
No aug + frozen	F	CD14	-0.260	-0.207	-0.020	0.825	0.191
No aug + frozen	F	CD163	0.101	0.016	0.022	0.516	0.115
No aug + frozen	F	CD19	0.293	0.280	0.031	0.314	0.022
No aug + frozen	F	CD27	-0.178	-0.246	-0.023	0.559	0.296
No aug + frozen	F	CD274	0.346	0.299	0.080	0.572	0.126
No aug + frozen	F	CD3E	0.499	0.444	0.098	0.972	0.012
No aug + frozen	F	CD4	0.381	0.534	0.163	0.332	0.191
No aug + frozen	F	CD40	0.158	0.091	0.023	0.703	0.196
No aug + frozen	F	CD68	0.038	0.185	0.014	0.382	0.277
No aug + frozen	F	CD8A	0.482	0.452	0.013	1.500	0.071
No aug + frozen	F	CEACAM8	0.479	0.374	0.206	0.871	0.259
No aug + frozen	F	CR2	0.073	0.223	0.027	0.631	0.250
No aug + frozen	F	CXCR5	0.569	0.567	-0.160	0.871	0.108
No aug + frozen	F	EPCAM	-0.358	-0.207	0.011	0.641	0.091
No aug + frozen	F	FCGR3A	0.183	0.414	0.020	1.238	0.561
No aug + frozen	F	HLA-DRA	-0.197	-0.180	-0.025	0.602	0.005
No aug + frozen	F	ITGAM	-0.119	-0.078	-0.004	0.616	0.153
No aug + frozen	F	ITGAX	-0.038	-0.044	0.003	0.551	0.098
No aug + frozen	F	KRT5	0.633	0.545	0.193	1.560	0.048
No aug + frozen	F	MS4A1	0.019	0.025	0.006	0.488	0.059
No aug + frozen	F	PAX5	0.200	0.245	0.014	0.966	0.160
No aug + frozen	F	PCNA	0.151	0.204	0.021	0.435	0.060
No aug + frozen	F	PDCD1	0.368	0.529	-0.055	0.772	0.142
No aug + frozen	F	PECAM1	0.014	0.089	-0.002	1.033	0.208
No aug + frozen	F	SDC1	-0.471	-0.440	-0.067	0.426	0.149
No aug + frozen	F	VIM	-0.320	-0.164	-0.026	0.419	0.051
No aug + frozen	F	mouse_IgG1k	0.058	0.141	0.019	0.501	0.551
No aug + frozen	F	mouse_IgG2a	-0.029	-0.014	0.006	0.538	0.637
No aug + frozen	F	mouse_IgG2bk	0.177	0.085	0.013	1.001	0.616
No aug + frozen	F	rat_IgG2a	-0.027	0.095	0.007	0.670	0.547
<i>No aug + frozen, reverse</i>							
No aug + frozen	R	ACTA2	0.187	0.298	-0.078	0.782	0.014
No aug + frozen	R	BCL2	0.117	0.213	0.038	0.644	0.137
No aug + frozen	R	CCR7	0.138	0.253	0.020	0.388	0.424
No aug + frozen	R	CD14	-0.044	0.006	-0.011	0.829	0.083

continued on next page

Table A.6: Held-out per-protein validation metrics for all compared models and both transfer directions. This table is derived from the validation rows of `per_protein_metrics_long.csv`; train and pooled rows are omitted because the main text uses held-out transfer as the primary benchmark. (continued).

Model	Dir.	Protein	PCC	Spearman	SSIM	RMSE	Moran err.
No aug + frozen	R	CD163	0.074	0.192	0.018	0.515	0.506
No aug + frozen	R	CD19	-0.019	0.078	0.004	0.350	0.507
No aug + frozen	R	CD27	0.027	-0.041	0.015	0.523	0.057
No aug + frozen	R	CD274	0.043	0.075	0.023	0.508	0.039
No aug + frozen	R	CD3E	0.206	0.251	0.094	0.987	0.228
No aug + frozen	R	CD4	0.062	0.189	0.028	0.369	0.339
No aug + frozen	R	CD40	0.068	0.178	0.024	0.699	0.347
No aug + frozen	R	CD68	-0.021	-0.098	-0.004	0.571	0.125
No aug + frozen	R	CD8A	0.126	0.100	0.020	1.448	0.091
No aug + frozen	R	CEACAM8	0.338	0.398	0.222	0.962	0.060
No aug + frozen	R	CR2	-0.064	-0.025	-0.001	0.619	0.461
No aug + frozen	R	CXCR5	0.309	0.494	-0.153	0.784	0.361
No aug + frozen	R	EPCAM	0.023	0.023	-0.005	0.634	0.008
No aug + frozen	R	FCGR3A	0.279	0.468	0.147	0.611	0.114
No aug + frozen	R	HLA-DRA	0.159	0.267	0.025	0.766	0.638
No aug + frozen	R	ITGAM	-0.051	0.062	-0.010	0.694	0.025
No aug + frozen	R	ITGAX	0.110	0.157	0.028	0.723	0.075
No aug + frozen	R	KRT5	0.329	0.351	0.164	1.357	0.478
No aug + frozen	R	MS4A1	0.108	0.187	0.019	0.596	0.425
No aug + frozen	R	PAX5	-0.018	0.170	-0.002	1.127	0.234
No aug + frozen	R	PCNA	0.030	-0.078	0.014	0.501	0.001
No aug + frozen	R	PDCD1	0.120	0.270	-0.044	0.681	0.144
No aug + frozen	R	PECAM1	-0.068	0.064	0.006	1.067	0.059
No aug + frozen	R	SDC1	0.068	0.040	0.033	0.437	0.232
No aug + frozen	R	VIM	-0.001	-0.075	0.004	0.484	0.481
No aug + frozen	R	mouse.IgG1k	-0.041	-0.099	0.001	0.493	0.287
No aug + frozen	R	mouse.IgG2a	-0.000	0.026	0.009	0.542	0.307
No aug + frozen	R	mouse.IgG2bk	-0.020	0.079	-0.001	0.823	0.130
No aug + frozen	R	rat.IgG2a	-0.007	0.017	0.007	0.678	0.023

Table A.7: Proposed-model per-protein metrics for the forward direction across train, validation, and pooled splits. The table is derived from `forward_proposed/per_protein_metrics.csv` and keeps the core association, error, and spatial-autocorrelation diagnostics.

Split	Protein	PCC	Spearman	SSIM	RMSE	MAE	Moran err.
<i>Training section</i>							
Train	ACTA2	0.912	0.900	0.905	0.237	0.177	0.012
Train	BCL2	0.955	0.952	0.952	0.175	0.131	0.019
Train	CCR7	0.934	0.931	0.932	0.136	0.101	0.062
Train	CD14	0.876	0.897	0.859	0.142	0.075	0.153
Train	CD163	0.887	0.888	0.871	0.140	0.107	0.070

continued on next page

Table A.7: Proposed-model per-protein metrics for the forward direction across train, validation, and pooled splits. The table is derived from `forward_proposed/per_protein_metrics.csv` and keeps the core association, error, and spatial-autocorrelation diagnostics. (continued).

Split	Protein	PCC	Spearman	SSIM	RMSE	MAE	Moran err.
Train	CD19	0.885	0.899	0.885	0.132	0.090	0.065
Train	CD27	0.866	0.881	0.852	0.154	0.109	0.175
Train	CD274	0.921	0.912	0.914	0.196	0.146	0.041
Train	CD3E	0.966	0.947	0.961	0.238	0.187	0.001
Train	CD4	0.940	0.935	0.933	0.109	0.079	0.044
Train	CD40	0.899	0.890	0.887	0.125	0.093	0.015
Train	CD68	0.852	0.826	0.830	0.282	0.213	0.084
Train	CD8A	0.936	0.950	0.841	0.284	0.201	0.044
Train	CEACAM8	0.896	0.803	0.889	0.435	0.318	0.002
Train	CR2	0.940	0.925	0.937	0.098	0.069	0.018
Train	CXCR5	0.925	0.950	0.913	0.214	0.135	0.034
Train	EPCAM	0.867	0.877	0.835	0.149	0.076	0.183
Train	FCGR3A	0.783	0.777	0.749	0.376	0.257	0.281
Train	HLA-DRA	0.923	0.933	0.912	0.248	0.177	0.023
Train	ITGAM	0.916	0.916	0.906	0.274	0.187	0.081
Train	ITGAX	0.932	0.929	0.880	0.230	0.179	0.034
Train	KRT5	0.913	0.902	0.898	0.591	0.449	0.016
Train	MS4A1	0.945	0.943	0.789	0.203	0.158	0.038
Train	PAX5	0.939	0.957	0.926	0.283	0.180	0.006
Train	PCNA	0.906	0.888	0.898	0.211	0.156	0.035
Train	PDCD1	0.910	0.908	0.902	0.118	0.084	0.043
Train	PECAM1	0.902	0.910	0.880	0.184	0.132	0.106
Train	SDC1	0.852	0.933	0.836	0.228	0.125	0.136
Train	VIM	0.930	0.909	0.928	0.178	0.136	0.012
Train	mouse_IgG1k	0.805	0.798	0.779	0.184	0.110	0.229
Train	mouse_IgG2a	0.887	0.905	0.879	0.145	0.093	0.222
Train	mouse_IgG2bk	0.878	0.835	0.861	0.207	0.144	0.147
Train	rat_IgG2a	0.866	0.874	0.850	0.138	0.084	0.206
<i>Held-out validation section</i>							
Val.	ACTA2	0.533	0.558	-0.354	0.776	0.635	0.132
Val.	BCL2	0.480	0.432	0.466	0.466	0.381	0.088
Val.	CCR7	0.380	0.354	0.373	0.386	0.325	0.155
Val.	CD14	0.076	0.027	0.035	0.897	0.867	0.042
Val.	CD163	0.264	0.281	0.174	0.628	0.586	0.235
Val.	CD19	0.457	0.445	0.443	0.265	0.203	0.050
Val.	CD27	0.266	0.273	0.233	0.498	0.413	0.218
Val.	CD274	0.519	0.403	0.391	0.532	0.419	0.064
Val.	CD3E	0.539	0.540	0.530	0.683	0.549	0.077
Val.	CD4	0.508	0.533	0.499	0.351	0.302	0.135
Val.	CD40	-0.142	-0.157	-0.117	0.619	0.542	0.061
Val.	CD68	0.301	0.286	0.279	0.368	0.284	0.023
Val.	CD8A	0.479	0.483	0.312	1.121	0.963	0.061
Val.	CEACAM8	0.567	0.481	0.539	0.671	0.469	0.190
Val.	CR2	0.361	0.410	0.269	0.728	0.690	0.119

continued on next page

Table A.7: Proposed-model per-protein metrics for the forward direction across train, validation, and pooled splits. The table is derived from `forward_proposed/per_protein_metrics.csv` and keeps the core association, error, and spatial-autocorrelation diagnostics. (continued).

Split	Protein	PCC	Spearman	SSIM	RMSE	MAE	Moran err.
Val.	CXCR5	0.701	0.724	-0.615	0.612	0.520	0.037
Val.	EPCAM	0.086	0.093	-0.042	0.753	0.692	0.111
Val.	FCGR3A	0.288	0.486	0.195	1.256	0.867	0.532
Val.	HLA-DRA	0.399	0.396	0.391	0.417	0.318	0.107
Val.	ITGAM	0.447	0.350	0.240	0.659	0.492	0.021
Val.	ITGAX	0.231	0.226	0.233	0.448	0.356	0.060
Val.	KRT5	0.660	0.570	0.440	1.454	1.102	0.070
Val.	MS4A1	0.437	0.444	0.410	0.436	0.343	0.037
Val.	PAX5	0.523	0.546	0.475	0.704	0.541	0.119
Val.	PCNA	0.341	0.267	0.291	0.434	0.321	0.037
Val.	PDCD1	0.635	0.678	-0.377	0.776	0.725	0.108
Val.	PECAM1	0.359	0.378	-0.004	0.939	0.830	0.176
Val.	SDC1	-0.070	-0.122	-0.047	0.568	0.437	0.127
Val.	VIM	0.308	0.344	0.227	0.509	0.400	0.094
Val.	mouse_IgG1k	0.108	0.125	0.074	0.586	0.521	0.379
Val.	mouse_IgG2a	0.105	0.131	0.081	0.759	0.707	0.554
Val.	mouse_IgG2bk	0.011	0.116	0.006	0.853	0.508	0.532
Val.	rat_IgG2a	0.207	0.267	0.087	0.773	0.744	0.379
<i>Pooled train and validation</i>							
All	ACTA2	0.674	0.705	0.379	0.592	0.424	0.024
All	BCL2	0.755	0.650	0.744	0.362	0.266	0.003
All	CCR7	0.552	0.448	0.551	0.298	0.222	0.041
All	CD14	0.063	-0.024	0.036	0.666	0.502	0.207
All	CD163	0.171	0.074	0.144	0.471	0.366	0.132
All	CD19	0.690	0.703	0.680	0.214	0.151	0.051
All	CD27	0.469	0.464	0.407	0.380	0.273	0.165
All	CD274	0.675	0.606	0.624	0.413	0.293	0.056
All	CD3E	0.769	0.776	0.763	0.527	0.383	0.004
All	CD4	0.583	0.515	0.571	0.268	0.199	0.027
All	CD40	0.542	0.479	0.436	0.462	0.335	0.100
All	CD68	0.696	0.596	0.678	0.331	0.251	0.089
All	CD8A	0.661	0.644	0.489	0.845	0.612	0.093
All	CEACAM8	0.777	0.653	0.766	0.574	0.399	0.050
All	CR2	0.233	0.166	0.194	0.539	0.404	0.207
All	CXCR5	0.795	0.790	-0.004	0.472	0.343	0.077
All	EPCAM	0.096	-0.067	-0.067	0.562	0.408	0.169
All	FCGR3A	0.377	0.497	0.297	0.957	0.586	0.247
All	HLA-DRA	0.801	0.737	0.792	0.349	0.253	0.021
All	ITGAM	0.678	0.617	0.545	0.518	0.351	0.016
All	ITGAX	0.759	0.698	0.758	0.364	0.275	0.037
All	KRT5	0.742	0.715	0.632	1.141	0.801	0.049
All	MS4A1	0.764	0.721	0.705	0.348	0.258	0.053
All	PAX5	0.762	0.766	0.729	0.551	0.375	0.057
All	PCNA	0.669	0.581	0.647	0.350	0.245	0.028

continued on next page

Table A.7: Proposed-model per-protein metrics for the forward direction across train, validation, and pooled splits. The table is derived from `forward_proposed/per_protein_metrics.csv` and keeps the core association, error, and spatial-autocorrelation diagnostics. (continued).

Split	Protein	PCC	Spearman	SSIM	RMSE	MAE	Moran err.
All	PDCD1	0.388	0.359	-0.302	0.575	0.430	0.151
All	PECAM1	0.519	0.517	-0.139	0.701	0.509	0.150
All	SDC1	0.407	0.371	0.394	0.445	0.294	0.026
All	VIM	0.648	0.565	0.628	0.393	0.278	0.004
All	mouse_IgG1k	0.263	0.157	0.219	0.448	0.331	0.301
All	mouse_IgG2a	0.079	-0.094	0.076	0.566	0.424	0.228
All	mouse_IgG2bk	0.548	0.631	0.394	0.642	0.340	0.295
All	rat_IgG2a	0.030	-0.081	0.024	0.575	0.440	0.231

Table A.8: Proposed-model per-protein metrics for the reverse direction across train, validation, and pooled splits. The table is derived from `reverse_proposed/per_protein_metrics.csv` and keeps the core association, error, and spatial-autocorrelation diagnostics.

Split	Protein	PCC	Spearman	SSIM	RMSE	MAE	Moran err.
<i>Training section</i>							
Train	ACTA2	0.694	0.714	0.580	0.591	0.461	0.004
Train	BCL2	0.596	0.531	0.472	0.285	0.207	0.160
Train	CCR7	0.419	0.406	0.271	0.217	0.145	0.240
Train	CD14	0.377	0.399	0.184	0.207	0.160	0.256
Train	CD163	0.413	0.445	0.304	0.213	0.159	0.345
Train	CD19	0.609	0.636	0.500	0.222	0.171	0.102
Train	CD27	0.388	0.368	0.235	0.299	0.230	0.206
Train	CD274	0.518	0.333	0.383	0.500	0.387	0.209
Train	CD3E	0.771	0.774	0.705	0.421	0.335	0.045
Train	CD4	0.456	0.525	0.293	0.185	0.140	0.220
Train	CD40	0.290	0.204	0.158	0.220	0.173	0.114
Train	CD68	0.200	0.282	0.094	0.355	0.277	0.290
Train	CD8A	0.779	0.784	0.726	0.434	0.338	0.089
Train	CEACAM8	0.564	0.488	0.424	0.653	0.438	0.289
Train	CR2	0.550	0.547	0.387	0.213	0.148	0.189
Train	CXCR5	0.812	0.832	0.725	0.323	0.246	0.062
Train	EPCAM	0.683	0.584	0.591	0.219	0.161	0.240
Train	FCGR3A	0.391	0.606	0.231	1.168	0.661	0.500
Train	HLA-DRA	0.464	0.439	0.359	0.358	0.268	0.086
Train	ITGAM	0.616	0.406	0.493	0.482	0.335	0.082
Train	ITGAX	0.342	0.357	0.170	0.342	0.269	0.160
Train	KRT5	0.889	0.786	0.844	0.833	0.683	0.005
Train	MS4A1	0.597	0.569	0.500	0.340	0.264	0.065
Train	PAX5	0.608	0.659	0.510	0.489	0.358	0.130
Train	PCNA	0.578	0.596	0.431	0.362	0.258	0.171
Train	PDCD1	0.742	0.779	0.659	0.251	0.176	0.100

continued on next page

Table A.8: Proposed-model per-protein metrics for the reverse direction across train, validation, and pooled splits. The table is derived from `reverse_proposed/per_protein_metrics.csv` and keeps the core association, error, and spatial-autocorrelation diagnostics. (continued).

Split	Protein	PCC	Spearman	SSIM	RMSE	MAE	Moran err.
Train	PECAM1	0.618	0.641	0.535	0.365	0.281	0.140
Train	SDC1	0.707	0.648	0.597	0.298	0.206	0.285
Train	VIM	0.747	0.746	0.684	0.277	0.214	0.096
Train	mouse_IgG1k	0.174	0.182	0.057	0.288	0.182	0.634
Train	mouse_IgG2a	0.178	0.269	0.061	0.288	0.156	0.764
Train	mouse_IgG2bk	0.157	0.030	0.022	0.696	0.357	0.698
Train	rat_IgG2a	0.188	0.217	0.055	0.232	0.135	0.573
<i>Held-out validation section</i>							
Val.	ACTA2	0.329	0.344	-0.261	1.185	1.081	0.068
Val.	BCL2	0.088	0.105	0.026	0.711	0.544	0.008
Val.	CCR7	0.045	0.033	0.026	0.432	0.371	0.004
Val.	CD14	-0.318	-0.369	-0.039	0.787	0.740	0.065
Val.	CD163	-0.104	-0.103	-0.038	0.603	0.532	0.027
Val.	CD19	0.398	0.469	0.252	0.330	0.264	0.011
Val.	CD27	0.095	0.127	0.058	0.532	0.454	0.116
Val.	CD274	0.274	0.247	0.083	0.523	0.392	0.028
Val.	CD3E	0.323	0.299	0.136	1.264	1.085	0.155
Val.	CD4	0.355	0.419	0.097	0.377	0.315	0.055
Val.	CD40	-0.198	-0.252	-0.018	0.746	0.698	0.111
Val.	CD68	0.362	0.349	0.102	0.560	0.434	0.098
Val.	CD8A	0.348	0.340	0.042	1.822	1.684	0.093
Val.	CEACAM8	0.248	0.118	0.049	1.114	0.735	0.047
Val.	CR2	0.507	0.499	0.233	0.599	0.556	0.059
Val.	CXCR5	0.400	0.450	-0.268	0.869	0.739	0.050
Val.	EPCAM	0.113	0.152	-0.013	0.521	0.465	0.031
Val.	FCGR3A	0.121	0.137	0.098	0.666	0.500	0.306
Val.	HLA-DRA	-0.120	-0.064	-0.045	0.775	0.583	0.127
Val.	ITGAM	0.232	0.297	0.145	0.665	0.505	0.080
Val.	ITGAX	-0.157	-0.194	-0.030	0.725	0.590	0.077
Val.	KRT5	0.211	0.151	0.093	1.603	1.232	0.177
Val.	MS4A1	0.054	0.055	0.026	0.596	0.461	0.051
Val.	PAX5	0.408	0.464	0.142	1.112	0.959	0.102
Val.	PCNA	-0.031	-0.017	-0.012	0.519	0.389	0.014
Val.	PDCD1	0.559	0.628	-0.220	0.705	0.669	0.035
Val.	PECAM1	0.242	0.252	-0.060	1.171	1.097	0.053
Val.	SDC1	-0.144	-0.184	-0.041	0.464	0.363	0.145
Val.	VIM	0.440	0.460	0.234	0.493	0.360	0.115
Val.	mouse_IgG1k	-0.212	-0.227	-0.014	0.528	0.463	0.094
Val.	mouse_IgG2a	0.215	0.232	0.040	0.505	0.419	0.032
Val.	mouse_IgG2bk	0.161	0.235	0.020	0.806	0.698	0.084
Val.	rat_IgG2a	0.239	0.206	0.027	0.637	0.591	0.043
<i>Pooled train and validation</i>							
All	ACTA2	0.355	0.347	-0.134	0.914	0.747	0.004
All	BCL2	0.150	0.146	0.087	0.526	0.362	0.005

continued on next page

Table A.8: Proposed-model per-protein metrics for the reverse direction across train, validation, and pooled splits. The table is derived from `reverse_proposed/per_protein_metrics.csv` and keeps the core association, error, and spatial-autocorrelation diagnostics. (continued).

Split	Protein	PCC	Spearman	SSIM	RMSE	MAE	Moran err.
All	CCR7	0.091	0.047	0.054	0.334	0.250	0.010
All	CD14	0.308	0.304	0.063	0.556	0.427	0.207
All	CD163	-0.042	-0.030	-0.014	0.438	0.330	0.148
All	CD19	0.505	0.570	0.366	0.277	0.214	0.028
All	CD27	0.229	0.279	0.116	0.423	0.333	0.162
All	CD274	0.378	0.200	0.260	0.511	0.389	0.084
All	CD3E	0.271	0.270	0.205	0.912	0.681	0.041
All	CD4	0.113	0.117	0.061	0.290	0.220	0.007
All	CD40	-0.234	-0.342	-0.051	0.531	0.415	0.142
All	CD68	0.267	0.290	0.095	0.461	0.350	0.153
All	CD8A	0.110	0.128	0.068	1.277	0.959	0.051
All	CEACAM8	0.283	0.185	0.166	0.895	0.575	0.133
All	CR2	0.347	0.262	0.156	0.436	0.336	0.195
All	CXCR5	0.421	0.404	0.003	0.635	0.474	0.076
All	EPCAM	0.610	0.523	0.322	0.389	0.301	0.155
All	FCGR3A	0.326	0.456	0.210	0.970	0.586	0.285
All	HLA-DRA	0.169	0.176	0.093	0.588	0.413	0.026
All	ITGAM	0.451	0.365	0.330	0.573	0.413	0.021
All	ITGAX	0.063	0.063	0.023	0.553	0.417	0.106
All	KRT5	0.667	0.470	0.559	1.249	0.936	0.052
All	MS4A1	0.334	0.302	0.231	0.475	0.355	0.032
All	PAX5	0.426	0.511	0.258	0.836	0.635	0.029
All	PCNA	0.323	0.304	0.221	0.441	0.318	0.002
All	PDCD1	0.385	0.279	0.205	0.513	0.403	0.139
All	PECAM1	0.163	0.147	0.078	0.839	0.657	0.106
All	SDC1	0.419	0.202	0.316	0.383	0.278	0.029
All	VIM	0.541	0.572	0.450	0.392	0.281	0.026
All	mouse_IgG1k	-0.180	-0.242	-0.028	0.416	0.312	0.253
All	mouse_IgG2a	0.366	0.422	0.081	0.402	0.277	0.240
All	mouse_IgG2bk	0.365	0.460	0.044	0.749	0.514	0.250
All	rat_IgG2a	0.390	0.409	0.044	0.465	0.345	0.205

Table A.9: Complete evaluation-metric taxonomy for virtual spatial omics. The applicability column is removed, and measurement focus is merged with the metric description.

Metric	Focus and definition	Reference
<i>Category: Correlation-based</i>		
Pearson Correlation Coefficient (PCC / r)	Linear relationship between predicted and measured expression: Linear correlation between predicted and ground-truth expression vectors, usually computed per spot, marker, or gene.	(Wang et al., 2025; Li et al., 2026b)

continued on next page

Table A.9: Complete evaluation-metric taxonomy (continued).

Metric	Focus and definition	Reference
Spearman Correlation (Spearman's r)	Rank-order consistency and monotonic relationship: A non-parametric measure of rank correlation used to quantify spatial co-localization patterns or clinical utility in ordering patient cohorts. Robust against non-linear data distributions.	(Li et al., 2026b; Hoang et al., 2024)
<i>Category: Structural Similarity</i>		
Structural Similarity Index Measure (SSIM)	Spatial pattern similarity and structure preservation: A perception-based model that evaluates similarity in luminance, contrast, and structural information. It quantifies spatial coherence beyond pixel-wise error.	(Wang et al., 2004; Andani et al., 2025)
<i>Category: Error-based</i>		
Mean Squared Error (MSE)	Average squared difference between predicted and ground truth: Averages squared prediction errors; heavily penalizes large deviations and is commonly used as a regression loss.	(He et al., 2020; Wang et al., 2025)
Mean Absolute Error (MAE)	Average magnitude of errors in predictions: Measures the sample-wise average of the absolute differences between predictions and ground truth. Provides a direct interpretation of error in the same units as the data without considering direction.	(Wang et al., 2025; Li et al., 2026b)
Root Mean Squared Error (RMSE)	Deviation between predicted and measured intensities: The square root of the average of squared differences. Used to assess reconstruction accuracy, prediction fidelity, and model robustness in both pixel-centric and cell-centric evaluations.	(Zhang et al., 2024; Han et al., 2026)
<i>Category: Clinical Outcomes</i>		
Concordance Index (C-index)	Predictive accuracy for survival analysis and ranking: Measures the rank correlation between predicted risk scores and observed survival times. A value of 0.5 indicates random prediction; values near 0.7-0.9 indicate strong clinical relevance and discriminatory power for risk stratification.	(Hoang et al., 2024)
Recurrence-Free Survival (RFS)	Disease progression and recurrence risk: A clinical outcome metric representing the time until cancer recurrence; used to assess the translational value of AI-predicted biomarkers via Kaplan-Meier or Cox models.	(Hoang et al., 2024)
Odds Ratio (OR)	Association between prediction and response: The ratio of the odds of responding to a treatment among predicted responders versus predicted non-responders in precision oncology contexts.	(Hoang et al., 2024)
<i>Category: Clustering and biological consistency</i>		

continued on next page

Table A.9: Complete evaluation-metric taxonomy (continued).

Metric	Focus and definition	Reference
Adjusted Rand Index (ARI)	Similarity between predicted clusters and expert annotations: Measures the agreement between two data partitions, such as model-derived clusters and pathological labels, while correcting for random agreement.	(Wang et al., 2025; Li et al., 2026b)
<i>Category: Graph and spatial-context consistency</i>		
Neighbor Consistency (NC)	Local spatial context and topological consistency: Quantifies local organization by measuring whether neighboring cells or spots share labels or expression similarity, usually over a graph neighborhood.	(Liu et al., 2026; Comiter, 2024)
Moran's I	Spatial autocorrelation over a neighborhood graph: Measures whether nearby spots have similar values for a marker or embedding dimension. In this thesis it is computed on an 8-nearest-neighbor coordinate graph, separately for measured and predicted marker maps.	(Moran, 1950)
<i>Category: Classification and discrimination</i>		
Area Under the Curve (AUC)	Discriminative performance for binary states: Calculated from the receiver operating characteristic curve. Measures the ability to distinguish between binary classes, such as responder versus non-responder or expressed versus non-expressed.	(Hoang et al., 2024; Li et al., 2026b)
<i>Category: Statistical Alignment</i>		
Cosine Similarity	Directional alignment of expression vectors: Assesses vector-space concordance by calculating the dot product between normalized predicted and ground-truth expression vectors.	(Xie et al., 2023; Chen et al., 2025a)
<i>Category: Information-theoretic agreement</i>		
Mutual Information (MI / NMI)	Statistical dependence and shared information: Measures the amount of information shared between predicted and observed distributions. Normalized mutual information scales results between no shared information and perfect correlation.	(Wang et al., 2025)
<i>Category: Distribution-based</i>		
Earth Mover's Distance (EMD)	Similarity between continuous distributions: Also known as Wasserstein distance; calculates the minimum cost of transforming the inferred expression distribution into the ground-truth distribution.	(Yang and Uhler, 2018)
Jensen-Shannon Divergence (JSD)	Similarity between probability distributions: A symmetric and finite measure of similarity between probability distributions, such as predicted versus ground-truth cell type proportions.	(Comiter, 2024; Zhang et al., 2025)
<i>Category: Model Reliability</i>		

continued on next page

Table A.9: Complete evaluation-metric taxonomy (continued).

Metric	Focus and definition	Reference
Predictive Quality Score (PQS / GDMS)	Internal model confidence and predicted variation: Defined from variation in a gene’s predicted expression over cells. Used to identify genes with reliable predicted spatial distributions.	(Comiter, 2024)
<i>Category: Spatial Concordance</i>		
Dice Score	Fine-grained spatial overlap: Evaluates spatial concordance by calculating the overlap between predicted positive pixels and ground-truth activations in segmented maps.	(Valanarasu et al., 2026)
<i>Category: Diffusion-based</i>		
Diffusion Time (td)	Degree of spatial randomness vs. structure: Measures time until convergence in a simulated diffusion process to assess whether a transcript profile is spatially organized or random.	(Andersson and Lundeberg, 2021)
<i>Category: Network-based</i>		
Correlation Matrix Distance (CMD)	Difference between correlation matrices: Compares predicted and measured correlation matrices so lower values indicate better preservation of gene-gene or cell-cell network structure.	(Liu et al., 2026)
<i>Category: Clustering Quality</i>		
Silhouette Score	Separation and batch effect mitigation: Evaluates how similar an observation is to its own cluster compared with other clusters; higher scores indicate better separation.	(Zhang et al., 2026)
<i>Category: Information-theoretic predictability</i>		
Expression Predictability Score (EPS)	Theoretical upper bound of predictability: Information-theoretic upper-bound score derived from expression variation over an image-embedding neighborhood graph.	(Xia et al., 2025)
<i>Category: Concordance-based</i>		
Lin’s Concordance Correlation Coefficient (CCC)	Agreement between ground truth and predicted expression: A concordance metric used as an alternative to Pearson correlation for identifying well-predicted genes by measuring agreement relative to the 45-degree line.	(Hoang et al., 2024)
<i>Category: Image-quality error metrics</i>		
Peak Signal to Noise Ratio (PSNR)	Pixel-level difference in synthesized images: Based on the mean squared error between corresponding pixels; a standard objective index for generated or translated images.	(Liu et al., 2022)

continued on next page

Table A.9: Complete evaluation-metric taxonomy (continued).

Metric	Focus and definition	Reference
<i>Category: Classification and discrimination</i>		
F1-score	Balance between precision and recall: Harmonic mean of precision and recall. Used to benchmark performance in identifying complex categories such as MSI/MSS status or immune cell types.	(Chen et al., 2025a)

References

- 10x Genomics (2023a). Visium CytAssist Gene and Protein Expression Library of Human Tonsil, H&E, 6.5 mm (FFPE). <https://www.10xgenomics.com/datasets/gene-protein-expression-library-of-human-tonsil-cytassist-ffpe-2-standard>. Accessed: 2026-04-27.
- 10x Genomics (2023b). Visium CytAssist Gene and Protein Expression Library of Human Tonsil with Add-on Antibodies, H&E, 6.5 mm (FFPE). <https://www.10xgenomics.com/datasets/visium-cytassist-gene-and-protein-expression-library-of-human-tonsil-with-add-on-antibodies-h-e-6-5-mm-ffpe-2-standard>. Accessed: 2026-04-27.
- Aitchison, J. (1982). The statistical analysis of compositional data. *Journal of the Royal Statistical Society: Series B (Methodological)*, 44(2):139–160.
- Andani, S., Chen, B., Ficek-Pascual, J., Heinke, S., Casanova, R., Hild, B. F., Sobottka, B., Bodenmiller, B., Koelzer, V. H., et al. (2025). Histopathology-based protein multiplex generation using deep learning. *Nature Machine Intelligence*, 7(8):1292–1307.
- Andersson, A. and Lundeberg, J. (2021). sepal: Identifying transcript profiles with spatial patterns by diffusion-based modeling. *Bioinformatics*, 37(17):2644–2650.
- Avşar, G. and Pir, P. (2023). A comparative performance evaluation of imputation methods in spatially resolved transcriptomics data. *Molecular Omics*, 19(2):162–173.
- Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical

- and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300.
- Bonev, B., Castelo-Branco, G., Chen, F., Codeluppi, S., Corces, M. R., Fan, J., Heiman, M., Harris, K., Inoue, F., Kellis, M., et al. (2024). Opportunities and challenges of single-cell and spatially resolved genomics methods for neuroscience discovery. *Nature neuroscience*, 27(12):2292–2309.
- Brachtel, E., Washiyama, M., Johnson, G., Tenner-Racz, K., Racz, P., and MacLennan, I. (1996). Differences in the germinal centres of palatine tonsils and lymph nodes. *Scandinavian journal of immunology*, 43(3).
- Chen, K. H., Boettiger, A. N., Moffitt, J. R., Wang, S., and Zhuang, X. (2015). Spatially resolved, highly multiplexed rna profiling in single cells. *Science*, 348(6233):aaa6090.
- Chen, W., Zhang, P., Tran, T. N., Xiao, Y., Li, S., Shah, V. V., Cheng, H., Brannan, K. W., Youker, K., Lai, L., et al. (2025a). A visual–omics foundation model to bridge histopathology with spatial transcriptomics. *Nature Methods*, 22(7):1568–1582.
- Chen, Y., Fan, X., Shi, C., Shi, Z., and Wang, C. (2025b). A joint analysis of single cell transcriptomics and proteomics using transformer. *npj Systems Biology and Applications*, 11(1):1.
- Cho, W. J., Yoon, H., Jeong, D., Lim, H., and Chong, Y. (2025). MVHybrid: Improving spatial transcriptomics prediction with hybrid state space-vision transformer backbone in pathology vision foundation models. *arXiv preprint arXiv:2508.00383*.
- Chu, P. G. and Weiss, L. (2002). Keratin expression in human tissues and neoplasms. *Histopathology*, 40(5):403–439.
- Chung, J.-Y., Braunschweig, T., Williams, R., Guerrero, I., Hoffmann, K. M., Kwon, M., Song, Y. K., Libutti, S. K., and Hewitt, S. M. (2008). Factors in tissue handling

- and processing that impact rna obtained from formalin-fixed, paraffin-embedded tissue. *Journal of Histochemistry & Cytochemistry*, 56(11):1033–1042.
- Comiter, C. (2024). Inference of single cell profiles from histology stains with the single-cell omics from histology analysis framework (schaf).
- Curtis, C., Shah, S. P., Chin, S.-F., Turashvili, G., Rueda, O. M., Dunning, M. J., Speed, D., Lynch, A. G., Samarajiwa, S., Yuan, Y., et al. (2012). The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature*, 486(7403):346–352.
- Dang, T. M., Li, H., Guo, Y., Ma, H., Jiang, F., Miao, Y., Zhou, Q., Gao, J., and Huang, J. (2025). Hage: Hierarchical alignment gene-enhanced pathology representation learning with spatial transcriptomics. pages 228–238.
- Dawood, M., Branson, K., Rajpoot, N. M., and Minhas, F. u. A. A. (2021). All you need is color: image based spatial gene expression prediction using neural stain learning. pages 437–450.
- Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. pages 248–255.
- Doroshow, D. B. and Doroshow, J. H. (2019). Genomics and the history of precision oncology. *Surgical oncology clinics of North America*, 29(1):35.
- Ekvall, M., Bergenstråhle, L., Andersson, A., Czarnewski, P., Olegård, J., Käll, L., and Lundeberg, J. (2024). Spatial landmark detection and tissue registration with deep learning. *Nature Methods*, 21(4):673–679.
- Fang, D., Li, Y., Wang, Z., Zeng, Y., and Min, W. (2026). Adapting a pre-trained single-cell foundation model to spatial gene expression generation from histology images. *arXiv preprint arXiv:2603.19766*.

- Fischer, A. H., Jacobson, K. A., Rose, J., and Zeller, R. (2008). Hematoxylin and eosin staining of tissue and cell sections. *Cold spring harbor protocols*, 2008(5):pdb-prot4986.
- Förster, R., Mattis, A. E., Kremmer, E., Wolf, E., Brem, G., and Lipp, M. (1996). A putative chemokine receptor, blr1, directs b cell migration to defined lymphoid organs and specific anatomic compartments of the spleen. *Cell*, 87(6):1037–1047.
- Fu, X., Cao, Y., Bian, B., Wang, C., Graham, D., Pathmanathan, N., Patrick, E., Kim, J., and Yang, J. Y. H. (2025). Spatial gene expression at single-cell resolution from histology using deep learning with ghist. *Nature methods*, 22(9):1900–1910.
- Gal, Y. and Ghahramani, Z. (2016). Dropout as a bayesian approximation: Representing model uncertainty in deep learning. pages 1050–1059.
- Gayoso, A., Steier, Z., Lopez, R., Regier, J., Nazor, K. L., Streets, A., and Yosef, N. (2021). Joint probabilistic modeling of single-cell multi-omic data with totalvi. *Nature methods*, 18(3):272–282.
- Ghose, A. and Poupart, P. (2023). Contrastive deterministic autoencoders for language modeling. pages 8458–8476.
- Giesen, C., Wang, H. A., Schapiro, D., Zivanovic, N., Jacobs, A., Hattendorf, B., Schüffler, P. J., Grolmund, D., Buhmann, J. M., Brandt, S., et al. (2014). Highly multiplexed imaging of tumor tissues with subcellular resolution by mass cytometry. *Nature methods*, 11(4):417–422.
- Goltsev, Y., Samusik, N., Kennedy-Darling, J., Bhate, S., Hale, M., Vazquez, G., Black, S., and Nolan, G. P. (2018). Deep profiling of mouse splenic architecture with codex multiplexed imaging. *Cell*, 174(4):968–981.
- Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., et al. (1999). Molecular classification

- of cancer: class discovery and class prediction by gene expression monitoring. *science*, 286(5439):531–537.
- Gong, D., Arbesfeld-Qiu, J. M., Perrault, E., Bae, J. W., and Hwang, W. L. (2024). Spatial oncology: translating contextual biology to the clinic. *Cancer Cell*, 42(10):1653–1675.
- Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. (2012). A kernel two-sample test. *The journal of machine learning research*, 13(1):723–773.
- Han, A., Li, B., Mah, C. Y., Logan, J., Wang, Y., Liu, N., Ryan, F., Lynn, D., Foreman, D., O’Leary, J., et al. (2026). Predicting gene expression from whole slide images in prostate cancer using deep learning.
- Hanahan, D. (2022). Hallmarks of cancer: new dimensions. *Cancer discovery*, 12(1):31–46.
- He, B., Bergenstråhle, L., Stenbeck, L., Abid, A., Andersson, A., Borg, Å., Maaskola, J., Lundeberg, J., and Zou, J. (2020). Integrating spatial gene expression and breast tumour morphology via deep learning. *Nature biomedical engineering*, 4(8):827–834.
- He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. pages 770–778.
- Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., and Lerchner, A. (2017). beta-VAE: Learning basic visual concepts with a constrained variational framework.
- Hoang, D.-T., Dinstag, G., Shulman, E. D., Hermida, L. C., Ben-Zvi, D. S., Elis, E., Caley, K., Sammut, S.-J., Sinha, S., Sinha, N., et al. (2024). A deep-learning framework to predict cancer treatment response from histopathology images through imputed transcriptomics. *Nature cancer*, 5(9):1305–1317.
- Hoang, D.-T., Shulman, E. D., Dhruva, S. R., Nair, N. U., Barman, R. K., Cantore, T., Biswas, S., Lalchungnunga, H., Singh, O., Chung, Y., et al. (2025). Path2omics enhances

- transcriptomic and methylation prediction accuracy from tumor histopathology. *Cancer research*, 85(24):5098–5112.
- Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67.
- Huang, T., Liu, T., Babadi, M., Jin, W., and Ying, R. (2025). Scalable generation of spatial transcriptomics from histology images via whole-slide flow matching. *arXiv preprint arXiv:2506.05361*.
- Ilse, M., Tomczak, J., and Welling, M. (2018). Attention-based deep multiple instance learning. pages 2127–2136.
- Janesick, A., Shelansky, R., Gottscho, A. D., Wagner, F., Williams, S. R., Rouault, M., Beliakoff, G., Morrison, C. A., Oliveira, M. F., Sicherman, J. T., et al. (2023). High resolution mapping of the tumor microenvironment using integrated single-cell, spatial and in situ analysis. *Nature communications*, 14(1):8353.
- Jaume, G., Doucet, P., Song, A. H., Lu, M. Y., Almagro-Pérez, C., Wagner, S. J., Vaidya, A. J., Chen, R. J., Williamson, D. F., Kim, A., et al. (2024). Hest-1k: A dataset for spatial transcriptomics and histology image analysis. *Advances in Neural Information Processing Systems*, 37:53798–53833.
- Jia, Y., Liu, J., Chen, L., Zhao, T., and Wang, Y. (2024). Thitogene: a deep learning method for predicting spatial transcriptomics from histological images. *Briefings in Bioinformatics*, 25(1):bbad464.
- Keir, M. E., Butte, M. J., Freeman, G. J., and Sharpe, A. H. (2008). Pd-1 and its ligands in tolerance and immunity. *Annu. Rev. Immunol.*, 26:677–704.
- Kingma, D. P. and Welling, M. (2013). Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.
- Lai, Z. W., Weisser, J., Nilse, L., Costa, F., Keller, E., Tholen, M., Kizhakkedathu, J. N.,

- Biniossek, M., Bronsert, P., and Schilling, O. (2016). Formalin-fixed, paraffin-embedded tissues (ffpe) as a robust source for the profiling of native and protease-generated protein amino termini. *Molecular & Cellular Proteomics*, 15(6):2203–2213.
- Li, S., Gai, K., Dong, K., Zhang, Y., and Zhang, S. (2024). High-density generation of spatial transcriptomics with stage. *Nucleic Acids Research*, 52(9):4843–4856.
- Li, W., Li, F., Ge, X., and Jiang, Y. (2026a). Reframing spatial transcriptomics prediction: From regression to classification using swin transformer. *ENAR Proceedings*.
- Li, Z., Li, Y., Xiang, J., Wang, X., Yang, S., Zhang, X., Eweje, F., Chen, Y., Luo, X., Li, Y., et al. (2026b). Ai-enabled virtual spatial proteomics from histopathology for interpretable biomarker discovery in lung cancer. *Nature Medicine*, pages 1–14.
- Liu, S., Zhu, C., Xu, F., Jia, X., Shi, Z., and Jin, M. (2022). Bci: Breast cancer immunohistochemical image generation through pyramid pix2pix. pages 1815–1824.
- Liu, X., Jiang, Z., Zhu, P., Liu, H., and Li, C. (2025). Hadm-st: Histology-assisted differential modeling for spatial transcriptomics generation. pages 129–138.
- Liu, Y., Wang, C., Wang, Z., Chen, L., Li, Z., Song, J., Zou, Q., Gao, R., Qian, B.-Z., Feng, X., et al. (2026). High-parameter spatial multi-omics through histology-anchored integration. *Nature Methods*, 23(2):373–386.
- Min, W., Shi, Z., Zhang, J., Wan, J., and Wang, C. (2024). Multimodal contrastive learning for spatial gene expression prediction using histology images. *Briefings in Bioinformatics*, 25(6):bbae551.
- Monjo, T., Koido, M., Nagasawa, S., Suzuki, Y., and Kamatani, Y. (2022). Efficient prediction of a spatial transcriptomics profile better characterizes breast cancer tissue sections without costly experimentation. *Scientific reports*, 12(1):4133.
- Moran, P. A. (1950). Notes on continuous stochastic phenomena. *Biometrika*, 37(1/2):17–23.

- Nave, H., Gebert, A., and Pabst, R. (2001). Morphology and immunology of the human palatine tonsil. *Anatomy and embryology*, 204(5):367–373.
- Nonchev, K., Manaiev, G., Koelzer, V. H., and Räscher, G. (2025). Deepspot2cell: Predicting virtual single-cell spatial transcriptomics from h&e images using spot-level supervision. *bioRxiv*, pages 2025–09.
- Nutt, S. L., Heavey, B., Rolink, A. G., and Busslinger, M. (1999). Commitment to the b-lymphoid lineage depends on the transcription factor pax5. *Nature*, 401(6753):556–562.
- Oord, A. v. d., Li, Y., and Vinyals, O. (2018). Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.
- Pang, M., Su, K., and Li, M. (2021). Leveraging information in spatial transcriptomics to predict super-resolution gene expression from histology images in tumors. *BioRxiv*, pages 2021–11.
- Piwecka, M., Rajewsky, N., and Rybak-Wolf, A. (2023). Single-cell and spatial transcriptomics: deciphering brain complexity in health and disease. *Nature Reviews Neurology*, 19(6):346–362.
- Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., and Gulin, A. (2018). Catboost: unbiased boosting with categorical features. 31.
- Qian, J., Fang, Y., Weng, Z., Wang, X., Cooper, L. A., and Zhou, B. (2025). Learning from gene names, expression values and images: Contrastive masked text-image pretraining for spatial transcriptomics representation learning. *arXiv preprint arXiv:2509.16892*.
- Qiu, P. (2020). Embracing the dropouts in single-cell rna-seq analysis. *Nature communications*, 11(1):1169.
- Saltz, J., Gupta, R., Hou, L., Kurc, T., Singh, P., Nguyen, V., Samaras, D., Shroyer, K. R., Zhao, T., Batiste, R., et al. (2018). Spatial organization and molecular correlation of

- tumor-infiltrating lymphocytes using deep learning on pathology images. *Cell reports*, 23(1):181–193.
- Sanian, M. V., Hemmat, A., Vahidi, A., Maaskola, J., Lee, J. T. H., Makarchuk, S., Demirci, Y., Chipampe, N.-J., Haniffa, M., Bayraktar, O., et al. (2025). 3d-guided scalable flow matching for generating volumetric tissue spatial transcriptomics from serial histology. *arXiv preprint arXiv:2511.14613*.
- Sato, S., Steeber, D. A., and Tedder, T. F. (1995). The cd19 signal transduction molecule is a response regulator of b-lymphocyte differentiation. *Proceedings of the National Academy of Sciences*, 92(25):11558–11562.
- Schürch, C. M., Bhate, S. S., Barlow, G. L., Phillips, D. J., Noti, L., Zlobec, I., Chu, P., Black, S., Demeter, J., McIlwain, D. R., et al. (2020). Coordinated cellular neighborhoods orchestrate antitumoral immunity at the colorectal cancer invasive front. *Cell*, 182(5):1341–1359.
- Seferbekova, Z., Lomakin, A., Yates, L. R., and Gerstung, M. (2023). Spatial biology of cancer evolution. *Nature Reviews Genetics*, 24(5):295–313.
- Shi, Z., Xue, S., Zhu, F., and Min, W. (2024). High-resolution spatial transcriptomics from histology images using histosge. pages 2402–2407.
- Ståhl, P. L., Salmén, F., Vickovic, S., Lundmark, A., Navarro, J. F., Magnusson, J., Giacomello, S., Asp, M., Westholm, J. O., Huss, M., et al. (2016). Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science*, 353(6294):78–82.
- Tan, W. C. C., Nerurkar, S. N., Cai, H. Y., Ng, H. H. M., Wu, D., Wee, Y. T. F., Lim, J. C. T., Yeong, J., and Lim, T. K. H. (2020). Overview of multiplex immunohistochemistry/immunofluorescence techniques in the era of cancer immunotherapy. *Cancer Communications*, 40(4):135–153.

- Tedder, T. F. and Engel, P. (1994). Cd20: a regulator of cell-cycle progression of b lymphocytes. *Immunology today*, 15(9):450–454.
- Tolstikhin, I., Bousquet, O., Gelly, S., and Schoelkopf, B. (2017). Wasserstein auto-encoders. *arXiv preprint arXiv:1711.01558*.
- Valanarasu, J. M. J., Xu, H., Usuyama, N., Kim, C., Wong, C., Argaw, P., Shimol, R. B., Crabtree, A., Matlock, K., Bartlett, A. Q., et al. (2026). Multimodal ai generates virtual population for tumor microenvironment modeling. *Cell*, 189(2):386–400.
- Wang, C., Chan, A. S., Fu, X., Ghazanfar, S., Kim, J., Patrick, E., and Yang, J. Y. (2025). Benchmarking the translational potential of spatial gene expression prediction from histology. *Nature Communications*, 16(1):1544.
- Wang, C., Wang, X., Liu, D., Su, F., Sun, L., Zhao, Z., and Meng, Z. (2026). Chrep: Cross-modal histology representation and post-hoc calibration for spatial gene expression prediction. *arXiv preprint arXiv:2604.21573*.
- Wang, Z., Bovik, A. C., Sheikh, H. R., and Simoncelli, E. P. (2004). Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612.
- Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics bulletin*, 1(6):80–83.
- Wu, E., Bieniosek, M., Wu, Z., Thakkar, N., Charville, G. W., Makky, A., Schürch, C. M., Huyghe, J. R., Peters, U., Li, C. I., et al. (2025). Rosie: Ai generation of multiplex immunofluorescence staining from histopathology images. *Nature Communications*, 16(1):7633.
- Wu, Y., Koyuncu, C. F., Toro, P., Corredor, G., Feng, Q., Buzzy, C., Old, M., Teknos, T., Connelly, S. T., Jordan, R. C., et al. (2022). A machine learning model for separating

- epithelial and stromal regions in oral cavity squamous cell carcinomas using h&e-stained histology images: a multi-center, retrospective study. *Oral Oncology*, 131:105942.
- Xia, C.-R., Yao, J.-W., and Gao, G. (2025). Quantifying predictability of gene expression from histology image. *bioRxiv*, pages 2025–11.
- Xiao, X., Kong, Y., Li, R., Wang, Z., and Lu, H. (2024). Transformer with convolution and graph-node co-embedding: An accurate and interpretable vision backbone for predicting gene expressions from local histopathological image. *Medical Image Analysis*, 91:103040.
- Xie, R., Pang, K., Chung, S., Perciani, C., MacParland, S., Wang, B., and Bader, G. (2023). Spatially resolved gene expression prediction from histology images via bi-modal contrastive learning. *Advances in Neural Information Processing Systems*, 36:70626–70637.
- Xie, Z., Lin, Y., Yao, Z., Zhang, Z., Dai, Q., Cao, Y., and Hu, H. (2021). Self-supervised learning with swin transformers. *arXiv preprint arXiv:2105.04553*.
- Xu, Y. and McCord, R. P. (2022). Diagonal integration of multimodal single-cell data: potential pitfalls and paths forward. *Nature Communications*, 13(1):3505.
- Yang, K. D. and Uhler, C. (2018). Scalable unbalanced optimal transport using generative adversarial networks. *arXiv preprint arXiv:1810.11447*.
- Yang, Y., Hossain, M. Z., Stone, E., and Rahman, S. (2024). Spatial transcriptomics analysis of gene expression prediction using exemplar guided graph neural network. *Pattern Recognition*, 145:109966.
- Yang, Y., Hossain, M. Z., Stone, E. A., and Rahman, S. (2023). Exemplar guided deep neural network for spatial transcriptomics analysis of gene expression prediction. pages 5039–5048.
- Yao, S., Schroeder, A., Jiang, S., Im, S., Park, J. H., Dumoulin, B., Hwang, T. H.,

- Susztak, K., and Li, M. (2026). Pixel2gene enables histology-guided reconstruction and prediction of spatial gene expression. *bioRxiv*, pages 2026–02.
- Yuan, M., Chen, L., and Deng, M. (2022). Clustering cite-seq data with a canonical correlation-based deep learning method. *Frontiers in Genetics*, 13:977968.
- Zeng, Y., Wei, Z., Yu, W., Yin, R., Yuan, Y., Li, B., Tang, Z., Lu, Y., and Yang, Y. (2022). Spatial transcriptomics prediction from histology jointly through transformer and graph neural networks. *Briefings in Bioinformatics*, 23(5):bbac297.
- Zhang, D. and Li, M. (2025). Inferring super-resolution tissue architecture by integrating spatial transcriptomics with histology. US Patent App. 18/764,858.
- Zhang, D., Schroeder, A., Yan, H., Yang, H., Hu, J., Lee, M. Y., Cho, K. S., Susztak, K., Xu, G. X., Feldman, M. D., et al. (2024). Inferring super-resolution tissue architecture by integrating spatial transcriptomics with histology. *Nature biotechnology*, 42(9):1372–1377.
- Zhang, P., Chen, W., Tran, T. N., Zhou, M., Carter, K. N., Kandel, I., Li, S., Hoi, X. P., Sun, Y., Lai, L., et al. (2025). Thor: a platform for cell-level investigation of spatial transcriptomics and histology. *Nature Communications*, 16(1):7178.
- Zhang, P., Gao, C., Hua, K., Zhang, Z., and Li, S. (2026). Systematically decoding pathological morphologies and molecular profiles with unified multimodal embedding. *Nature Methods*, pages 1–6.
- Zhao, C., Xu, Z., Wang, X., Tao, S., MacDonald, W. A., He, K., Poholek, A. C., Chen, K., Huang, H., and Chen, W. (2024). Innovative super-resolution in spatial transcriptomics: a transformer model exploiting histology images and spatial gene expression. *Briefings in bioinformatics*, 25(2):bbae052.
- Zhao, S., Song, J., and Ermon, S. (2019). Infovae: Balancing learning and inference in variational autoencoders. 33(01):5885–5892.

- Zhong, Y., Zhang, J., and Ren, X. (2024). Spatial transcriptomics prediction from histology images at single-cell resolution using redehist. *bioRxiv*, pages 2024–06.
- Zhu, S., Zhu, Y., Tao, M., and Qiu, P. (2025). Diffusion generative modeling for spatially resolved gene expression inference from histology images.