Virtual immune cell prediction from H&E Whole Slide Images

Kyung A Kim[1], Jongwoo Yoo[1], and Hyo Sup Shim[1,*]

[1]Department of Pathology, Yonsei University College of Medicine, Seoul 03722, South Korea.

[*]Corresponding author: shimhs@yuhs.ac

Abstract. Immunotherapy has become part of standard care in lung cancer and other solid tumors, however only a subset of patients derives durable benefit. This variability has increased interest in the tumor immune microenvironment (TIME), but large-scale single-cell spatial proteomics is rarely available in routine practice. Multiplexed immunofluorescence (mIF) can resolve the TIME at single-cell resolution, but it is costly and hard to incorporate into standard pathology workflows. We introduce histology-derived virtual immune composition as a scalable alternative to experimental mIF. Using MACSima mIF, we phenotype 1,042,624 single cells, and co-register H&E to mIF at single-cell resolution. From these paired data, we train models to predict cell-type composition from H&E alone and show that the resulting virtual immune profiles outperform existing H&E-based approaches on mIF-derived ground truth. On our cohort, foundation-based virtual cells show substantially higher agreement with mIF-based ground truth, with the largest gains for macrophage, plasma cell, and tumor compartments. They further support patch-level detection of TLS-like patterns and tumor-infiltrating lymphocytes (TILs). Beyond this experimental reference, we extend the virtual immune composition to TCGA lung, melanoma, and head and neck cancer whole-slide images, where the same signatures show robust associations with bulk RNA-seq immune transcripts and consistently stratify overall survival across these immunotherapy-relevant cancers. Taken together, these results suggest that virtual immune profiling from routine histology can bring single-cell-level immune context closer to cohort-scale H&E datasets, while still revealing residual gaps relative to experimental spatial proteomics.

1 Introduction

Tumor immune microenvironment (TIME) shapes prognosis and response to immune checkpoint inhibitors across multiple solid tumors, including lung cancer, melanoma, and head and neck cancer [6,9,7,4]. Spatial organization is central to this biology: TIME is defined not only by which cells are present, but also by how they are arranged in tissue, and spatial cellular patterns such as T- and B-cell

localization, macrophage infiltration, and epithelial–immune interfaces provide information that complements bulk molecular profiling [12,1]. mIF and related spatial proteomics assays can measure these patterns at single-cell resolution, but they remain expensive and difficult to deploy across large archival cohorts [12].

H&E slides are inexpensive, routinely acquired, and available at scale, motivating computational pathology models that infer molecular and cellular information directly from tissue morphology [13,5]. Much of the broader virtual cell literature has focused on single-cell transcriptomic data and perturbation responses in dissociated cells [10,8]. These approaches do not directly capture spatial protein states within their native tissue context. Clinically relevant processes are often shaped by protein-level changes, cell–cell interactions, and tissue architecture that are not fully represented by transcriptomic measurements alone [12,1]. This motivates histology-derived virtual immune composition as a tissue-level step toward virtual cell prediction: instead of modelling full cellular dynamics, we ask when H&E-derived virtual immune composition can approximate cell-level spatial information that would otherwise require experimental mIF, and when it carries clinically relevant signal in immunotherapy-relevant cancers.

Pathology foundation models now provide strong frozen representations across a wide range of downstream tasks [2,14,3,11]. More recently, H&E-to-mIF translation model GigaTIME demonstrated the feasibility of generating virtual mIF from routine histology. However, marker recovery remains uneven, particularly in immune and cytoplasmic compartments critical for microenvironment analysis [13]. We therefore use a standardized linear probe and common cell-level mIF ground truth to benchmark pathology foundation encoders for virtual immune composition, and assess whether the resulting scores support downstream TIL, TLS-like, molecular, and survival analyses.

Contributions.

1. A single-cell resolution mIF–H&E reference for lung cancer, yielding over one million single-cell phenotypes for key immune and epithelial markers.
2. Evidence that a deliberately simple frozen-encoder probe recovers marker-resolved immune composition at finer lineage resolution than existing H&E classifiers and a virtual mIF baseline.
3. Evidence that H&E-derived virtual tumor microenvironment composition provides molecularly validated, independently prognostic TIL and TLS-like pattern readouts in immunotherapy-relevant TCGA cancers.

2 Methodology

2.1 mIF and cell-based ground truth

We used MACSima multiplex immunofluorescence images from a lung adenocarcinoma tissue microarray. DAPI-based co-registration aligned staining cycles to a common frame, nuclei were segmented with Cellpose, and per-cell marker

intensities were measured over a dilated nuclear region to capture membranous and cytoplasmic signal.

For phenotyping, we selected pan-cytokeratin and canonical lymphoid and myeloid markers (CD3, CD4, CD8α, CD20, CD68, CD163, and CD11c) from the multiplex panel to characterize TILs and TLS-associated immune composition. Pan-cytokeratin was used to separate tumor epithelium from non-epithelial cells. For each immune marker, positivity was defined using the 98th percentile of marker intensity among pan-cytokeratin–positive tumor cells as a negative control. A deterministic hierarchical marker-based ruleset then assigned each cell to a tumor epithelial, B, T, macrophage, dendritic, or plasma-cell phenotype, or left it unassigned.

H&E sections were registered to the MACSima frame by aligning H&E and DAPI nuclei-density maps with an affine transform. Registration quality was assessed by normalized cross-correlation (NCC; 0.66–0.93 among retained cores) together with visual inspection of anatomical alignment; one of 18 cores was excluded because it failed this quality-control assessment. Patches with ≤ 5 detected cells were excluded. Across the 17 retained cores, we generated 8,318 patches (256×256 pixels at $20\times$); 8,117 eligible patches were included in 17-fold leave-one-core-out cross-validation, with one core held out for testing in each fold.

2.2 Virtual immune cell prediction and baselines

On this patch-level regression task, we estimated virtual immune composition directly from H&E by feeding features from a frozen pathology foundation encoder (H-optimus-1) into a standardized linear probe with separate ridge-regression heads for each phenotype, keeping the encoder fixed so that only the probe was trained. For each phenotype $c \in \{\text{tumor}, \text{B}, \text{T}, \text{macrophage}, \text{plasma}, \text{DC}\}$, we independently fit a ridge-regression probe:

$$\mathcal{L}_c(\boldsymbol{\beta}_c, b_c) = \|\mathbf{y}_{c,\text{train}} - (\mathbf{X}_{\text{train}}\boldsymbol{\beta}_c + b_c\mathbf{1})\|_2^2 + \alpha\,\|\boldsymbol{\beta}_c\|_2^2\,, \qquad \alpha = 500. \tag{1}$$

Here, $\mathbf{X}_{\text{train}}$ contains z-standardized frozen encoder features using training-fold statistics, $\mathbf{y}_{c,\text{train}}$ contains mIF-derived patch fractions, and b_c is an unpenalized intercept. With

$$\hat{\tilde{\boldsymbol{\beta}}}_c = \Big(\tilde{\mathbf{X}}_{\text{train}}^\top \tilde{\mathbf{X}}_{\text{train}} + \alpha\,\text{diag}(1, \ldots, 1, 0)\Big)^{-1} \tilde{\mathbf{X}}_{\text{train}}^\top \mathbf{y}_{c,\text{train}}, \qquad \tilde{\boldsymbol{\beta}}_c = [\boldsymbol{\beta}_c^\top, b_c]^\top. \tag{2}$$

To isolate the effect of representation, we reused the same standardized linear probe across multiple frozen encoders, including recent pathology foundation models and an ImageNet-pretrained CNN, varying only the encoder backbone (Table 1).

We evaluated patch-level agreement with mIF-derived cell-type fractions using Spearman correlation, framed TLS recognition as a patch-level detection

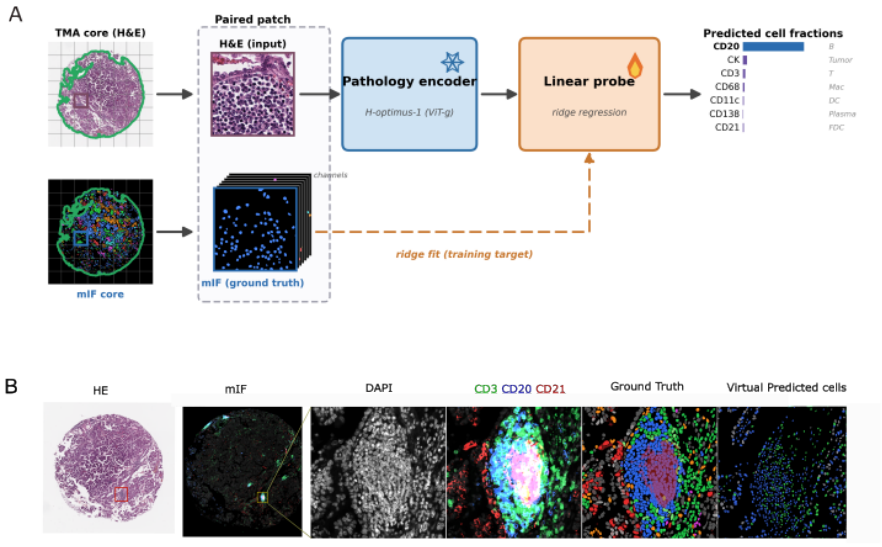

Figure 1. Virtual immune cell prediction from H&E using co-registered mIF. A. Overview of the virtual immune cell prediction pipeline. B. TMA core and co-registered ROI, from left to right: H&E of the TMA core, corresponding mIF and DAPI, individual CD3 (T cells, green), CD20 (B cells, blue), and CD21 (follicular dendritic cells, red) channels, ground-truth cell segmentation, and the H&E-based virtual-cell prediction highlighting $CD20^+$, $CD3^+$, and dim $CD21^+$ regions.

problem assessed by AUROC, and quantified uncertainty by core-level bootstrapping to capture dependence among patches within each core. Figure 1 summarizes the ground-truth construction, H&E–mIF registration, and the virtual immune prediction pipeline.

2.3 TILs and TLS-like definitions

TIL were quantified as an intratumoral score that weights the predicted pan-CD3 T-cell signal by local tumor content. For each slide, we computed the intratumoral TIL as

$$\mathrm{TIL}_{\mathrm{intra}} = \frac{\sum_p f_p^{\mathrm{T}} \, f_p^{\mathrm{tumor}}}{\sum_p f_p^{\mathrm{tumor}}}, \tag{3}$$

averaged per patient to obtain a continuous intratumoral infiltration value used in TCGA prognostic analyses; this definition emphasises lymphocytes in tumor-rich regions and downweights those in tumor-free stroma. As sensitivity analyses, we also considered a thresholded intratumoral variant (mean predicted T-cell fraction over patches with predicted tumor fraction ≥ 0.10, requiring at least ten such patches per slide) and the unrestricted whole-slide T-cell mean, and

observed prognostic associations that were concordant in direction and magnitude across all three definitions. TLS-like patterns were defined as mIF-derived B/T co-enrichment to approximate a key lymphoid component of TLS: patches with both B-cell and T-cell fractions > 0.05 were considered TLS-like. Patch-level AUROC was evaluated against pathologist-annotated TLS ground truth in TLS-bearing cores. In this way, the TIL score captures dispersed intratumoral T-cell infiltration, whereas the TLS-related scores capture organized B–T co-localization in follicular and T-zone structures.

2.4 TCGA validation cohorts

For external validation, we used several immunotherapy-relevant TCGA cohorts with matched diagnostic H&E whole-slide images, bulk RNA-seq, and clinical follow-up. H&E slides were tiled, passed through quality control, and processed with the same virtual immune cell pipeline. Per-slide fractions were obtained by averaging patch predictions and were aggregated per patient. Molecular validation used patients with matched RNA-seq, whereas survival analyses used primary-tumor cases with matched clinical data and positive follow-up time. Within each cohort, we assessed cross-modal consistency by computing Spearman's rho between predicted per-patient immune cell fractions and bulk expression of corresponding marker genes, with 2,000-resample bootstrap 95% confidence intervals; because RNA-seq was not normalized across cohorts and fractions were only interpreted within-cohort rank associations. Clinical relevance was evaluated via Cox proportional-hazards models for overall survival (hazard ratio per 1 standard deviation of each predicted fraction) and median-split Kaplan–Meier analyses with log-rank tests, with multivariable models adjusted for age (per 10 years), sex, and stage (III–IV vs I–II). Tumor purity and stromal/immune scores were obtained using ESTIMATE, which infers stromal and immune content from bulk gene expression profiles [15].

3 Results

3.1 Representation drives virtual immune prediction

Under a standardized linear probe, encoder identity dominated patch-level performance for virtual immune composition (Table 1). Pathology-pretrained encoders consistently outperformed a generic ImageNet-pretrained CNN, with non-overlapping bootstrap CIs for the best-recovered phenotypes (e.g., macrophage ρ 0.44 [0.33, 0.54] vs 0.07 [−0.08, 0.20]). These patterns indicate that pathology pretraining itself, rather than any particular task-specific prediction head, is the main driver of virtual immune performance. H-optimus-1 provided the best overall agreement and was used thereafter. Follicular dendritic cells, a diffuse meshwork without a clear nuclear correlate, were not recovered by any encoder.

Table 1. Encoder ablation: patch-level Spearman ρ against mIF-derived cell-based ground truth ; only the frozen encoder differs. Bold represents best marker.

Encoder	Macroph.	Plasma	Tumor	B	T	DC
H-optimus-1 (ViT-g, ours)	**0.44**	**0.22**	**0.65**	0.17	**0.28**	0.25
Gigapath (ViT-g)	0.37	0.20	0.63	0.15	0.23	0.20
UNI (ViT-L)	0.35	0.19	0.62	**0.18**	0.20	0.20
CONCH-v1.5 (ViT-L)	0.22	0.21	0.62	0.14	0.21	**0.28**
ResNet-50 (ImageNet)	0.07	0.11	0.42	0.09	0.12	0.10

Table 2. Methods for TLS-like pattern detection: mean AUROC over 15 TLS-bearing cores and output type. HoVer-Net/HoVer-Next are reported over the 14 cores with completed nuclei inference.

Method	Mean TLS AUROC	Output type	Training source
Ours	$\mathbf{0.76 \pm 0.10}$	Cell fractions	mIF (cells)
HoVer-Next	0.74 ± 0.12	Nuclei classes	H&E (WSI)
HoVer-Net	0.68 ± 0.21	Nuclei classes	H&E (WSI)
GigaTIME	0.55 ± 0.09	Marker masks	mIF (pixels)

3.2 TILs and TLS-like pattern detection and comparison with H&E classifier

Framed as patch-level detection, the H-optimus-1 virtual-cell model localized TILs and pathologist-annotated TLS regions on cores using only H&E input. A simple TLS score based on B- and T-cell co-localization recovered B-follicle, composite TLS, and T-zone patches with higher pooled AUROC than the virtual-staining baseline GigaTIME and achieved a mean TLS-like AUROC of 0.76±0.10 across 15 TLS-bearing cores compared with 0.55 ± 0.09 for GigaTIME (Table 2). This improvement over GigaTIME was statistically significant across cores (paired Wilcoxon signed-rank test, Holm-adjusted $p < 0.001$; median ΔAUROC 0.20, 95% CI 0.14–0.27). When benchmarked against H&E nuclei classifiers HoVer-Net and HoVer-Next at matched magnification, our virtual-cell model matched or exceeded their TLS-detection AUROC while uniquely providing cell-level, marker-resolved, mIF-validated composition (Table 2). Differences from HoVer-Net and HoVer-Next were not statistically significant (Holm-adjusted $p > 0.6$); our model is thus statistically comparable to these classifiers on coarse TLS detection while additionally resolving multi-lineage, marker-level composition (B, T, macrophage, dendritic, plasma, tumor) that they do not provide. We visualized TLS-related patch scores as heatmaps across models to compare their spatial localization on cores (Figure 2).

3.3 Virtual immune markers generalize across TCGA cohorts

We applied the frozen-encoder virtual-cell model to 2,168 diagnostic WSIs from TCGA-LUAD, -LUSC, -SKCM, -BLCA, and -HNSC and evaluated molecular

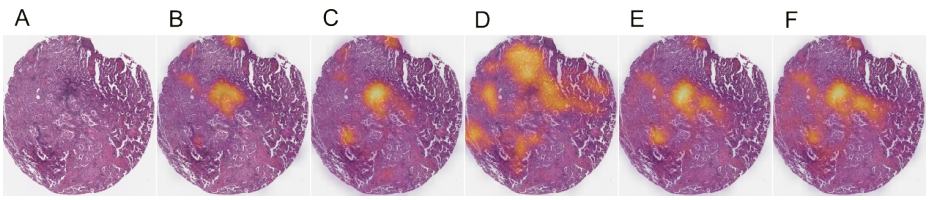

Figure 2. TLS detection heatmaps. A: H&E, B: pathologist-annotated TLS ground truth on mIF, C: ours, D: GigaTIME, E: HoVer-Net, and F: HoVer-Next predictions.

validity and overall survival. Predicted immune composition reproduced matched bulk RNA-seq across cohorts: in two independent lung cohorts, predicted per-lineage fractions correlated with cell-type-specific transcript signatures for CD8 T cells, pan-T cells, B cells, and macrophages (Spearman ρ in the range 0.15–0.39, all significant after Benjamini–Hochberg correction), at magnitudes comparable to the internal mIF benchmark (Figure 3C), providing surrogate-level support that this composition transfers across cohorts. Predicted tumor fraction, whose bulk epithelial-gene reference is confounded by normal epithelium, instead correlated positively with ESTIMATE-derived tumor purity and negatively with immune and stromal scores, and the TLS score based on B/T co-localization aligned with an established 12-chemokine TLS signature. For prognosis, higher intratumoral TIL—the predicted T-cell signal weighted by local tumor content—was consistently protective and remained independently associated with longer OS in LUAD and BLCA after adjustment for age, sex, and stage, with a concordant trend in HNSC and no clear association in LUSC or the primary-tumor SKCM (Figure 3D–K). Estimates were stable across whole-slide, threshold-based, and tumor-weighted TIL definitions.

4 Discussion

We benchmarked whether immune cell composition can be recovered from routine H&E under a standardized linear probe and common mIF ground truth. Our cell-level mIF–H&E reference in lung adenocarcinoma suggests that immune signal in H&E, although only modestly correlated with mIF-derived fractions for some phenotypes at the patch level, can carry enough information to support practically relevant downstream analyses. Accordingly, the modest B- and T-cell agreement supports their use as spatially aggregated surrogates rather than absolute per-patch cell counts. These include TLS/TIL-related detection and surrogate molecular and survival studies. In this view, the model acts less as a simple architecture and more as a practical bridge. The mIF–H&E reference marks where H&E carries usable immune context and where spatial proteomics remains essential. TCGA applications provide surrogate-level evidence that virtual immune markers can extend single-cell immune context to large slide archives. Prior work has shown that deep learning on H&E can recover immune patterns associated with molecular features and survival, and recent virtual staining and "vir-

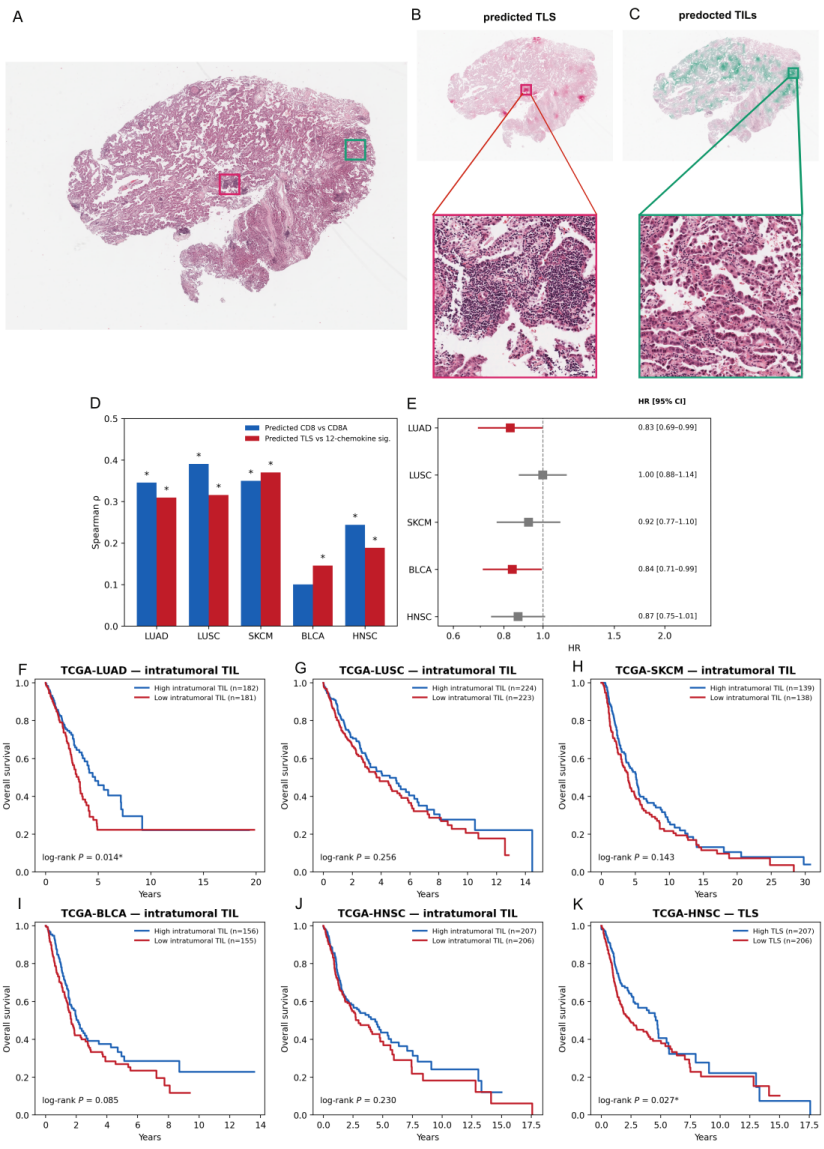

Figure 3. Virtual cells in five TCGA cohorts: molecular validation and survival. Frozen-encoder model applied to diagnostic WSIs from TCGA-LUAD ($n = 425$), LUSC ($n = 476$), SKCM ($n = 432$), BLCA ($n = 386$), and HNSC ($n = 449$); survival analyses included 413, 463, 277, 311, and 413 patients, respectively. (A) Representative TCGA-LUAD H&E whole-slide image. Predicted TLS-like pattern (B) and intratumoral TIL (C) heatmap. (D) Within-cohort Spearman correlations between predicted CD8 T cells and CD8A, and between the TLS-like score and a bulk TLS signature. (E) Multivariable hazard ratios per standard deviation of intratumoral TIL for overall survival, adjusted for age, sex, and stage. (F–K) Kaplan–Meier overall survival by median intratumoral TIL for each cohort, and by median predicted TLS-like score in HNSC.

tual spatial proteomics" models, including GigaTIME, predict multiplex protein markers directly from histopathology and improve prognostic and immunotherapy response modeling.[13] Our study follows this line of work but focuses on cell-resolved immune composition with an explicit mIF–H&E reference. In this sense, virtual immune markers are best viewed as surrogate readouts that help prioritize cases and hypotheses rather than as replacements for high-plex imaging. At the same time, virtual composition is directly evaluated only within a single lung TMA, and TCGA validation relies on bulk RNA-seq and survival surrogates rather than mIF ground truth. Differences in staining, scanning, and immune context between sites and cancer types may further limit generalization. Interpretations of cross-site and cross-cancer transfer should therefore be made with appropriate caution, and future work with multi-center mIF–H&E references will be required to more fully establish where virtual immune composition is reliable and where it breaks down.

5 Conclusion

We introduced a cell-level mIF–H&E reference for lung adenocarcinoma and used it to conduct a controlled comparison of pathology representations for virtual immune profiling from routine H&E. Under leave-one-core-out evaluation, encoder pretraining was the main determinant of H&E-derived immune-composition performance. The resulting scores supported virtual TIL and TLS-like readouts at TCGA scale, with molecular and survival associations. Rather than presenting a new architecture, this work delivers a mIF-anchored reference and a scoped benchmark that clarify where routine H&E can act as an accessible surrogate for high-plex spatial proteomics in retrospective and archival cohorts. Extending the reference to multi-center, multi-cancer mIF–H&E data will be essential to establish where such virtual immune markers can safely inform translational and clinical decisions.

Acknowledgements

This study was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean Government (MSIT) (RS-2025-00518452) and the Bio & Medical Technology Development Program of the NRF Korean Government (MSIT) (RS-2023-00261820).

Disclosure of Interests The authors have no competing interests to declare.

Ethical Approval and Informed Consent This study received approval from the Institutional Review Board of Severance Hospital (IRB No. 4-2020-0945). Written informed consent was obtained from all participants.

References

1. Bifulco, C., et al.: Large-scale spatial profiling of the tumor microenvironment. Cell (2026). https://doi.org/10.1016/j.cell.2025.12.018
2. Bioptimus: H-optimus-0: an open-source foundation model for histology (2024), model release. https://huggingface.co/bioptimus/H-optimus-0
3. Chen, R.J., Ding, T., Lu, M.Y., Williamson, D.F., Jaume, G., Song, A.H., Chen, B., Zhang, A., Shao, D., Shaban, M., et al.: Towards a general-purpose foundation model for computational pathology. Nature Medicine **30**, 850–862 (2024)
4. Chen, W., et al.: Advances in tumor immune microenvironment of head and neck squamous cell carcinoma: A review of literature. Medicine **103**(9), e37387 (2024). https://doi.org/10.1097/MD.0000000000037387
5. Graham, S., Vu, Q.D., Raza, S.E.A., Azam, A., Tsang, Y.W., Kwak, J.T., Rajpoot, N.M.: Hover-net: Simultaneous segmentation and classification of nuclei in multi-tissue histology images. Medical Image Analysis **58**, 101563 (2019). https://doi.org/10.1016/j.media.2019.101563
6. Hanna, N.H., Schneider, B.J., Temin, S., Baker, S., Brahmer, J., Ellis, P.M., Giaccone, G., Hegde, P., Jain, N., Jaiyesimi, I., Johnson, M., Levy, B., Lilenbaum, R., Liu, S.V., McGettigan, S., Otterson, G.A., Patel, J., Redman, M.W., Reckamp, K.L., Riess, J.W., Schenk, E., Spigel, D., Waterhouse, D., West, H., Horn, L.: Society for immunotherapy of cancer (sitc) clinical practice guideline on immunotherapy for the treatment of lung cancer. Journal for ImmunoTherapy of Cancer **10**(5), e003956 (2022). https://doi.org/10.1136/jitc-2021-003956
7. Huang, A.C., Zappasodi, R.: A decade of checkpoint blockade immunotherapy in melanoma: Understanding the molecular basis for immune sensitivity and resistance. Nature Immunology **23**(5), 660–670 (2022). https://doi.org/10.1038/s41590-022-01141-1
8. Kim, C., et al.: Single-cell foundation models: bringing artificial intelligence into cell biology. Experimental & Molecular Medicine (2025). https://doi.org/10.1038/s12276-025-01547-5
9. Li, Y., et al.: Lung cancer immunotherapy in 2025: where we stand and what comes next. Frontiers in Immunology **16**, 1728163 (2026). https://doi.org/10.3389/fimmu.2025.1728163
10. Lotfollahi, M., et al.: Closing the loop: Teaching single-cell foundation models to learn from perturbations. Cell (2025). https://doi.org/10.1016/j.cell.2025.06.019
11. Lu, M.Y., Chen, B., Williamson, D.F., Chen, R.J., Liang, I., Ding, T., Jaume, G., Odintsov, I., Le, L.P., Gerber, G., et al.: A visual-language foundation model for computational pathology. Nature Medicine **30**, 863–874 (2024)
12. Lundberg, E., et al.: An update on spatial proteomics. Nature Methods (2026). https://doi.org/10.1038/s41592-026-03046-5
13. Valanarasu, J.M.J., Xu, H., Usuyama, N., Kim, C., Wong, C., Argaw, P., Ben Shimol, R., Crabtree, A., Matlock, K., Bartlett, A.Q., Bagga, J., Gu, Y., Zhang, S., Naumann, T., Fox, B.A., Wright, B., Robicsek, A., Piening, B., Bifulco, C., Wang, S., Poon, H.: Multimodal ai generates virtual population for tumor microenvironment modeling. Cell **189**(2), 386–400.e19 (2026). https://doi.org/10.1016/j.cell.2025.11.016
14. Xu, H., Usuyama, N., Bagga, J., Zhang, S., Rao, R., Naumann, T., Wong, C., Gero, Z., González, J., Gu, Y., et al.: A whole-slide foundation model for digital pathology from real-world data. Nature **630**, 181–188 (2024)

15. Yoshihara, K., Shahmoradgoli, M., Martinez, E., Vegesna, R., Kim, H.I., Torres-Garcia, W., Treviño, V., Shen, H., Laird, P.W., Levine, D.A., et al.: Inferring tumour purity and stromal and immune cell admixture from expression data. Nature Communications **4**, 2612 (2013). https://doi.org/10.1038/ncomms3612