

# ROSIE: AI generation of multiplex immunofluorescence staining from histopathology images

Received: 18 December 2024

Accepted: 18 July 2025

Published online: 16 August 2025



Eric Wu<sup>1,2</sup>✉, Matthew Bieniosek<sup>1</sup>, Zhenqin Wu<sup>1</sup>, Nitya Thakkar<sup>3</sup>, Gregory W. Charville<sup>4</sup>, Ahmad Makky<sup>5</sup>, Christian M. Schürch<sup>5,6</sup>, Jeroen R. Huyghe<sup>7</sup>, Ulrike Peters<sup>7,8</sup>, Christopher I. Li<sup>7,8</sup>, Li Li<sup>9</sup>, Hannah Giba<sup>10,11</sup>, Vivek Behera<sup>10,12</sup>, Arjun Raman<sup>10,11,13</sup>, Alexandro E. Trevino<sup>1</sup>✉, Aaron T. Mayer<sup>1</sup>✉ & James Zou<sup>1,2,14</sup>✉

Hematoxylin and eosin (H&E) is a common and inexpensive histopathology assay. Though widely used and information-rich, it cannot directly inform about specific molecular markers, which require additional experiments to assess. To address this gap, we present ROSIE, a deep-learning framework that computationally imputes the expression and localization of dozens of proteins from H&E images. Our model is trained on a dataset of over 1300 paired and aligned H&E and multiplex immunofluorescence (mIF) samples from over a dozen tissues and disease conditions, spanning over 16 million cells. Validation of our *in silico* mIF staining method on held-out H&E samples demonstrates that the predicted biomarkers are effective in identifying cell phenotypes, particularly distinguishing lymphocytes such as B cells and T cells, which are not readily discernible with H&E staining alone. Additionally, ROSIE facilitates the robust identification of stromal and epithelial microenvironments and immune cell subtypes like tumor-infiltrating lymphocytes (TILs), which are important for understanding tumor-immune interactions and can help inform treatment strategies in cancer research.

H&E staining is ubiquitously used in clinical histopathology due to its affordability, accessibility, and effectiveness for discerning clinically relevant features. While H&E readily identifies nuclear and cytoplasmic morphology, its utility is limited in revealing more complex molecular information associated with modern precision medicine<sup>1</sup>. Pathologists can identify diverse cell types from H&E

staining alone; however, computational approaches to annotating H&E images only distinguish a few broad cell categories, such as endothelial, epithelial, stromal, and immune cells<sup>2–4</sup>. These methods are valuable for detecting tumors and identifying basic structural features but are limited in revealing detailed aspects of the cellular microenvironment, such as protein expression profiles,

<sup>1</sup>Enable Medicine, Menlo Park, CA, USA. <sup>2</sup>Department of Electrical Engineering, Stanford University, Stanford, CA, USA. <sup>3</sup>Department of Computer Science, Stanford University, Stanford, CA, USA. <sup>4</sup>Department of Pathology, Stanford University, Stanford, CA, USA. <sup>5</sup>Department of Pathology and Neuropathology, University Hospital and Comprehensive Cancer Center Tübingen, Tübingen, Germany. <sup>6</sup>Cluster of Excellence iFIT (EXC 2180) “Image-Guided and Functionally Instructed Tumor Therapies”, University of Tübingen, Tübingen, Germany. <sup>7</sup>Public Health Sciences Division, Fred Hutchinson Cancer Center, Seattle, WA, USA. <sup>8</sup>Department of Epidemiology, University of Washington, Seattle, WA, USA. <sup>9</sup>Ochsner Health, New Orleans, LA, USA. <sup>10</sup>Duchossois Family Institute, University of Chicago, Chicago, IL 60637, USA. <sup>11</sup>Department of Pathology, University of Chicago, Chicago, IL 60637, USA. <sup>12</sup>Department of Medicine, Section of Hematology/Oncology, University of Chicago, Chicago, IL 60637, USA. <sup>13</sup>Center for the Physics of Evolving Systems, University of Chicago, Chicago, IL 60637, USA. <sup>14</sup>Department of Biomedical Data Science, Stanford University, Stanford, CA, USA. ✉e-mail: [wue@stanford.edu](mailto:wue@stanford.edu); [alex@enablemedicine.com](mailto:alex@enablemedicine.com); [aaron@enablemedicine.com](mailto:aaron@enablemedicine.com); [jamesz@stanford.edu](mailto:jamesz@stanford.edu)

disease signatures, or the specific identity of immune cells like lymphocytes.

In contrast, multi-plex immunofluorescence (mIF) imaging techniques such as Co-Detection by Indexing (CODEX) and immunohistochemistry (IHC) enable in situ detection of dozens of proteins simultaneously. This capability allows for the exploration of richer tissue microenvironments, offering insights that are unattainable through H&E staining alone<sup>5–8</sup>. However, the application of CODEX and similar mIF techniques is limited by high costs, time-intensive

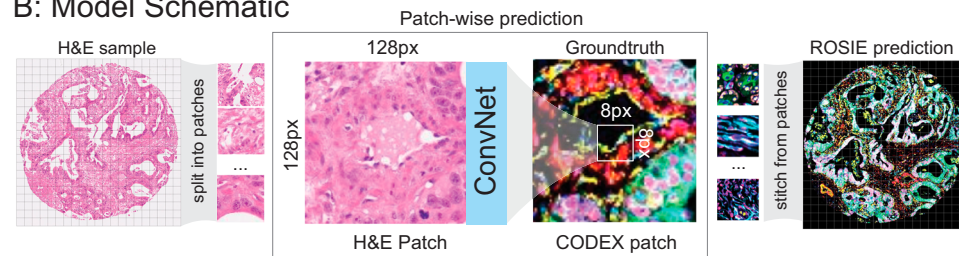
protocols, and lack of adoption in clinical labs, making them less feasible for routine use<sup>9</sup>.

In this work, we present ROSIE (**RO**bst in **S**ilico **I**mmunofluorescence from H&E images), a framework for in silico mIF staining based on an H&E-stained input image. We train a deep learning model on a dataset of over 1,000 tissue samples co-stained with H&E and CODEX. This dataset, comprising nearly 30 million cells, is the largest of its kind to date and significantly surpasses the scale of previous studies, which typically focus on data from a single clinical site or a

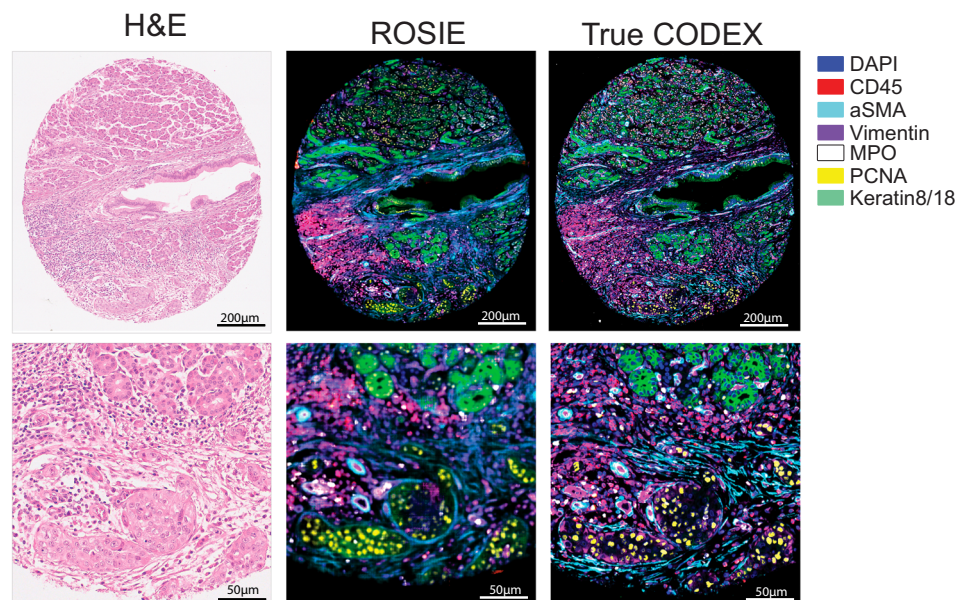
## A: Data Description



## B: Model Schematic



## C: Example of ROSIE (in silico staining)



**Fig. 1 | Overview of ROSIE. A** Our training dataset consists of 18 studies and over 16 M cells. Each tissue sample is co-stained with H&E and CODEX. 16 disease types and 10 body areas are represented in this dataset. The overall distribution of represented tissue types across training and evaluation datasets is shown on the right. **B** A schematic of model training and inference is shown. Given an H&E sample, the image is split into patches of size 128px by 128px. The model is trained to predict the average expressions of the center 8px by 8px patch in the

corresponding CODEX image. After the model is trained, a predicted CODEX image is generated by aggregating all of the generated patches into a single image.

**C** Given an H&E-stained image, ROSIE predicts the pixel-level expression of 50 biomarkers. An exemplary image (with the highest Pearson R score) is visualized, where seven representative biomarkers are colored and shown alongside the true CODEX image. While the generated images used in our analyses are produced with 8px striding, this image is produced using 1px striding for greater visual clarity.

**Table 1 | Description of evaluation datasets**

Evaluation Datasets	In training data?		# Samples	# Slides	# Cells
	Study	Disease			
Stanford-PGC	Yes	Yes	149	1	817,765
Ochsner-CRC	No	No	94	1	635,649
Tuebingen-GEJ	No	No	240	1	365,734
UChicago-DLBCL	Yes	Yes	2	1	3,099,419
Total	-	-	485	8	4,918,567

We describe the four evaluation datasets used in this study. Samples from Stanford-PGC and UChicago-DLBCL are divided into training and test splits; on the other hand, no samples and disease types from Ochsner-CRC and Tuebingen-GEJ are used in the model's training data. UChicago-DLBCL contains two full tissue samples, whereas the rest of the datasets consist of TMA core samples.

limited number of stains. Our findings demonstrate that the proposed method can robustly predict and spatially resolve dozens of proteins from H&E stains alone.

We validate the biological accuracy of these *in silico-generated* protein expressions by employing them in detailed cell phenotyping and the discovery of tissue structures such as stromal and epithelial tissues. Our approach enables the identification of immune cell subtypes, including B cells and T cell subtypes that are not discernible by H&E staining alone, thus offering a powerful tool for enhancing the diagnostic and research potential of standard histopathological practices.

Recent advancements in training histopathology foundation models<sup>10–12</sup> have demonstrated that models trained on large, diverse sets of histology images in an unsupervised manner can yield strong performance when adapted to downstream tasks like predicting tissue types and disease diagnosis and prognosis. While foundation models can learn intricate biological features within the distribution of H&E images, they still need to be explicitly trained on other imaging modalities and molecular information to be adapted for generative methods like *in silico* staining.

Previous works in predicting immunostains from H&E have typically focused on small paired or unpaired datasets and imputing up to several biomarkers at once. To start, VirtualMultiplexer is a GAN-based method for predicting 6-plex IHC stains<sup>13</sup> using unpaired H&E and IHC samples. The limitation of using unpaired samples is that validation of predictions is limited to qualitative or visual assessments. Several methods have been trained on paired (adjacent slice) datasets, such as: Multi-V-Stain<sup>14</sup>, which predicts a 10-plex mIMC panel on 336 melanoma samples; DeepLIIF<sup>15</sup> which predicts 3-plex mIHC on one sample; and other GAN-based methods for predicting single or several biomarkers<sup>16–22</sup> predict two IHC biomarkers on brain tissue. Compared with paired samples, co-stained (or same slice) samples allow for direct pixel-level alignment and prediction from H&E to immunostain; to this end, HEMIT (3-plex mIHC)<sup>23</sup> and vIHC (1-plex mIHC) both are trained on co-stained samples but are also limited to evaluation on a single sample<sup>24</sup>. focus on predicting a transcriptomics panel (1000 genes) using 4 co-stained samples. Specific to multiplexed immunofluorescence, 7-UP<sup>25</sup> used a small 7-plex panel to predict over 30 biomarkers<sup>26</sup>. use autofluorescence and DAPI channels to infer seven biomarkers.

Our method improves upon this previous body of work in several ways. First, we train and evaluate the largest co-stained H&E and immunostaining dataset with over 1300 samples. Second, whereas previous datasets were limited to one or several tissue types, our dataset spans ten body areas and disease types. Third, whereas previous works focus on visual or quantitative metrics of expression prediction only, ours demonstrates the usefulness of the predicted expressions for cell phenotyping and tissue structure

**Table 2 | Main evaluation results**

Evaluation Datasets (50 biomarkers)	Pearson R	Spearman R	C-index (sample)
Stanford-PGC			
<i>H&amp;E expression</i>	0.007	0.013	0.504
<i>Cell morphology</i>	0.072	0.081	0.574
ROSIE (our method)	0.319	0.386	0.694
ROSIE evaluated on:			
Ochsner-CRC	0.218	0.276	0.597
Tuebingen-GEJ	0.265	0.289	0.668
UChicago-DLBCL	0.254	0.327	0.820
Average (ROSIE)	0.246	0.297	0.695

Our method is evaluated using three metrics: Pearson R, Spearman R, and C-index (sample), which refers to the concordance index computed on the sample level using the 75th percentile of expression values in a sample as a threshold. The performance of our method on the primary dataset (Stanford-PGC) is reported along with two baseline methods (shown in italics): H&E expression, which uses the mean cell-wise H&E pixel value as a proxy for protein expression, and cell morphology, which uses features derived from the cell outlines based on DAPI expression as well as the RGB pixel values as input to a neural network to predict protein expression. We also report the performance of our method on the other three evaluation datasets.

discovery. Finally, instead of using difficult-to-train adversarial methods, we use a straightforward single MSE objective for training our model.

**Results**

**A comprehensive, diverse dataset of co-stained tissue samples**

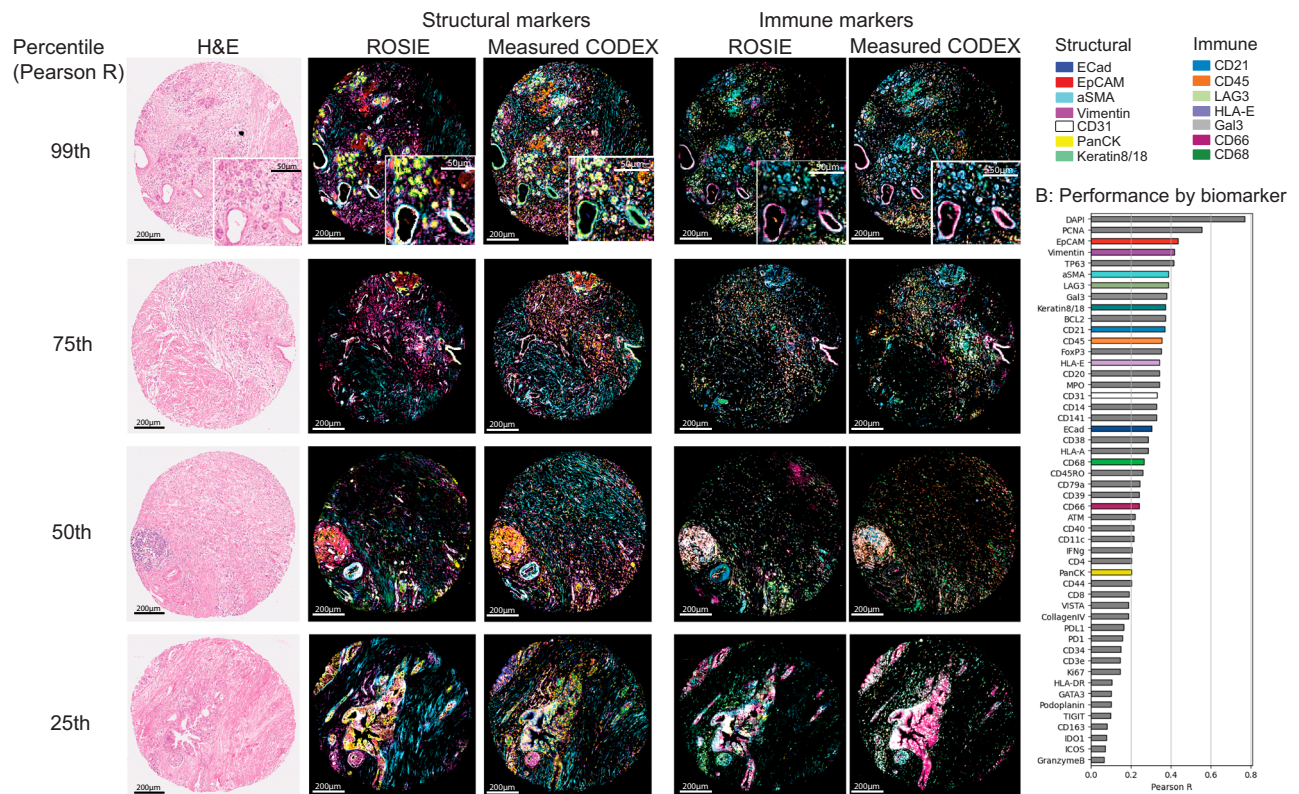
We introduce a training and evaluation dataset across 20 studies (Fig. 1A, Table 1; see Supplementary Materials for dataset details). The training dataset consists of over 16 M cells from 18 studies, 1342 samples, and 13 disease types. The evaluation dataset consists of 5 M cells from 4 studies, 485 samples, and 4 disease types. Two studies (Stanford-PGC and UChicago-DLBCL) are split between both training and evaluation datasets. All studies are on tissue samples with H&E and CODEX co-staining on the exact same samples. All datasets consist of tissue microarray (TMA) cores (average 10 K cells per core) except UChicago-DLBCL, which contains full slide samples (average 1.5 M cells per slide). Stanford-PGC is a study of patients with pancreatic and gastrointestinal cancer from Stanford Healthcare. Ochsner-CRC is a study of patients with colorectal cancer from Ochsner Medical Center. Tuebingen-GEJ is a study of patients with cancer in the gastroesophageal junction from Tübingen University Hospital. UChicago-DLBCL is a study of patients with diffuse large B-cell lymphoma from the University of Chicago Medical Center.

**Generative deep learning model for inferring protein expression from H&E stains**

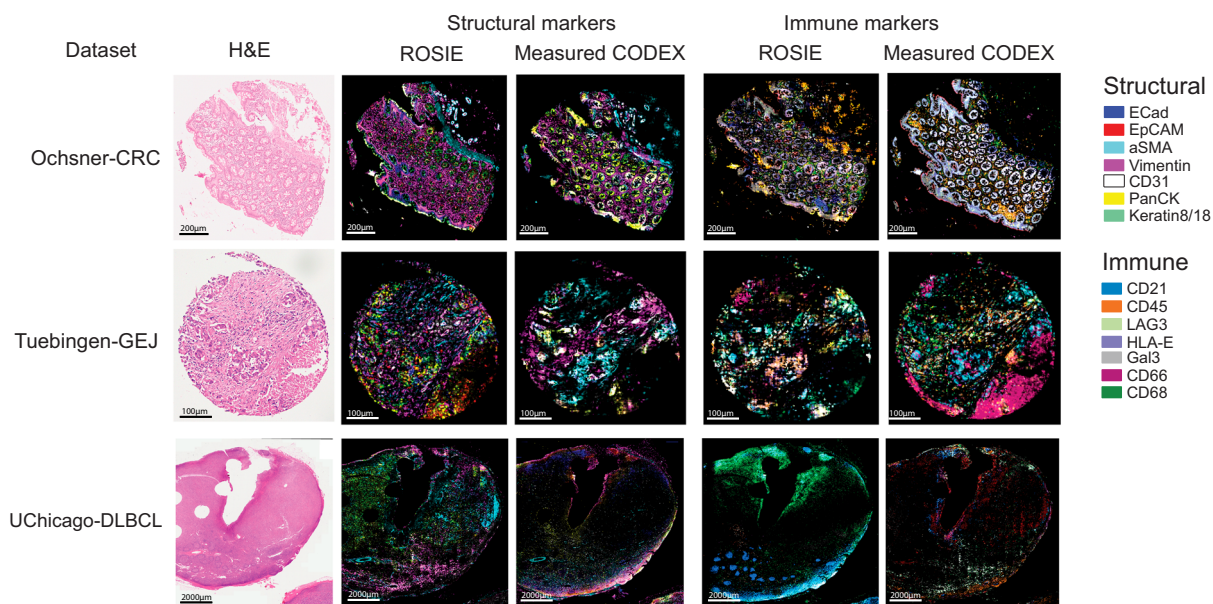
ROSIE is a framework for *in silico* staining on a sample based on an H&E image. Using a ConvNext<sup>27</sup> convolutional neural network (CNN) architecture, ROSIE operates on the patch level: given an input 128 × 128 pixel patch, it produces a prediction for the average expressions of the biomarker panel across the center 8 × 8 pixels (Fig. 1B). Using a sliding window with an 8px step size, we iteratively generate predictions on all 8 × 8 pixel patches within a sample, then stitch together predictions to produce a whole, contiguous image. ROSIE can be run with smaller sliding window sizes, down to 1px for native resolution outputs (see Supplementary Fig. 5 for examples). Due to computational tradeoffs, analyses performed in this paper are on the standard 8px sliding window setting. Although Vision Transformer (ViT) models<sup>28</sup> have recently gained traction as top-performing histopathology foundation models, we find that ConvNext outperforms several ViT models despite its smaller size (Supplementary Table 1).



## A: Examples of predictions from ROSIE, Stanford-PGC



## C: Examples of predictions from ROSIE, additional datasets



**Fig. 2 | Visualization of predictions from ROSIE. A** Predicted and measured CODEX samples along with the co-stained H&E images. The 99th, 75th, 50th, and 25th percentile (by Pearson R) samples are shown, colored with seven structural (Keratin8/18, EpCAM, Vimentin, ECad, aSMA, CD31, PanCK) and immune (HLA-E, Gal3, CD45, CD21,

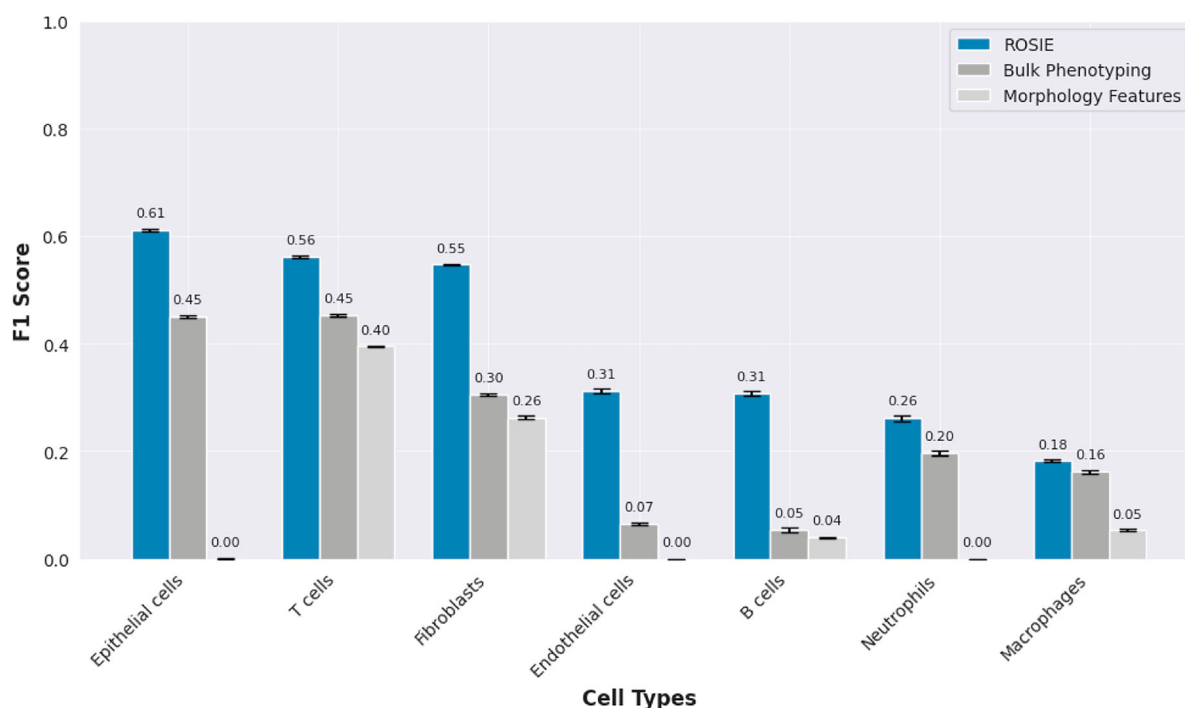
LAG3, CD66, CD68) markers. **B** Pearson R correlation on all evaluation datasets for 50 biomarkers (visualized biomarkers are colored). Supplementary Fig. 5 visualizes zoomed-in patches of predictions. **C** Visualizations of median samples (by Pearson R) from three additional datasets: Ochsner-CRC, Tuebingen-GEJ, and UChicago-DLBCL.

A total of 148 unique biomarkers are represented across all studies. We constrain our method to predict the top 50 biomarkers by prevalence. While all evaluation studies are stained with these 50 biomarkers, some studies used in training are not; in these cases, only the subset of biomarkers present in this set are used. The full biomarker set that the model is trained to predict includes (in order of prevalence):

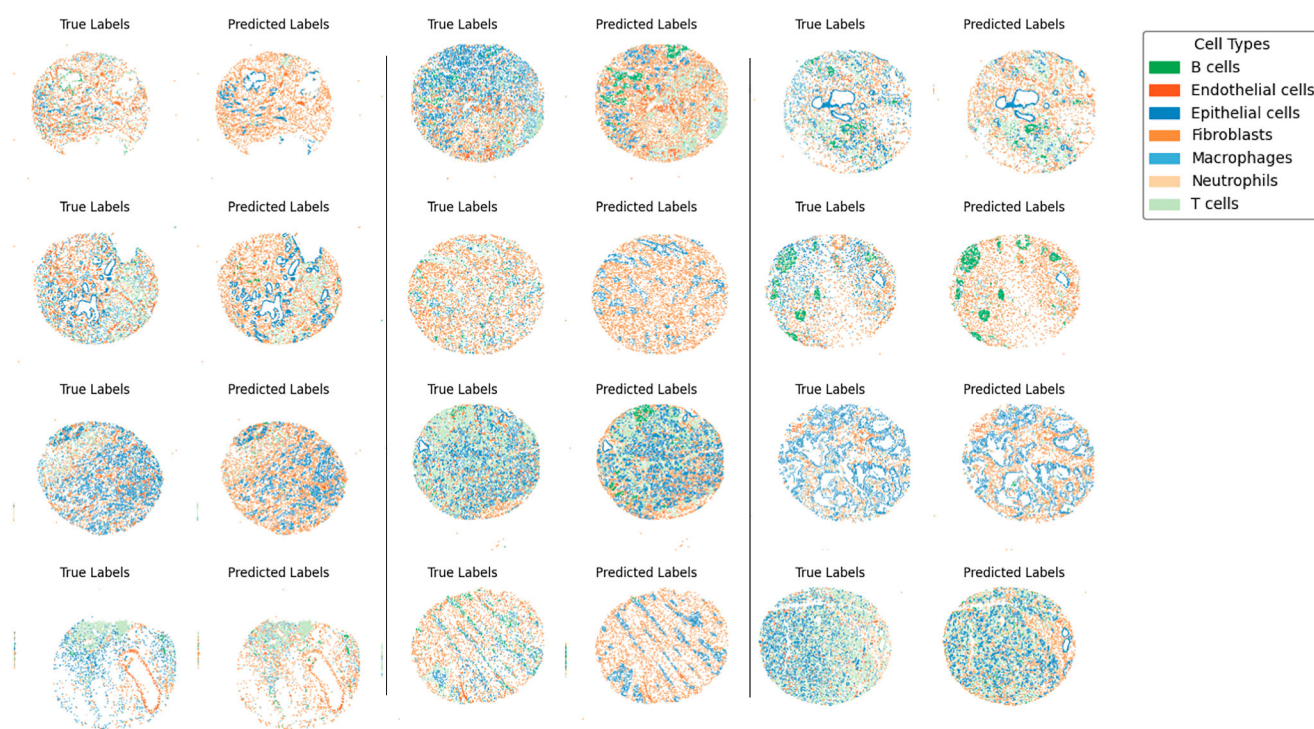
DAPI, CD45, CD68, CD14, PD1, FoxP3, CD8, HLA-DR, PanCK, CD3e, CD4, aSMA, CD31, Vimentin, CD45RO, Ki67, CD20, CD11c, Podoplanin, PDL1, GranzymeB, CD38, CD141, CD21, CD163, BCL2, LAG3, EpCAM, CD44, ICOS, GATA3, Gal3, CD39, CD34, TIGIT, ECad, CD40, VISTA, HLA-A, MPO, PCNA, ATM, TP63, IFN $\gamma$ , Keratin8/18, IDO1, CD79a, HLA-E, CollagenIV, CD66.



## A: Cell phenotype predictions from ROSIE, F1 score



## B: Examples of cell phenotype predictions from ROSIE



**Fig. 3 | Cell type predictions using ROSIE.** **A** F1 scores ( $N = 817,765$  cells) on the primary Stanford-PGC dataset, comparing the performance of ROSIE to two base-lines: *bulk phenotyping*, which randomly assigns cell types based on sample-level cell type proportions, and *morphology features*, which uses a three-layer neural

network to classify cells based on morphology features and the H&E RGB channels. Data are presented as mean values with error bars as the 95% bootstrapped confidence intervals. **B** Visualization of cell phenotype predictions from twelve median samples by Pearson R.

### ROSIE accurately predicts protein biomarker expressions

When applying ROSIE to the four evaluation datasets, we report a Pearson R correlation of 0.285, a Spearman R correlation of 0.352, and a sample-level C-index of 0.706 when comparing the ground

truth and computationally generated expressions across all 50 biomarkers in all four datasets (Table 2). Whereas the Pearson correlation indicates a linear predictive relationship, the Spearman R and C-index indicate the usefulness of the predicted expressions in

clinical tasks that involve ordering cells or samples by expressing a certain biomarker (e.g., identifying immune markers in a cancer patient cohort). C-index refers to the concordance index computed on the sample level using the 75th percentile expression value as a threshold. A C-index of 0.5, for instance, indicates random chance. We show that our method significantly outperforms two baseline methods: *H&E expression*, which uses the average intensity across RGB channels as a direct proxy for protein expression for every biomarker, is intended to test whether our predictive accuracy is simply due to recapitulating the staining signal; and *cell morphology*, which uses morphology features derived from the cell segmentations in addition to the three RGB channels as inputs to a multi-layer perceptron (MLP) neural network trained to predict protein expression, is intended as a representative machine learning model that uses common H&E-derived features as input. Both baseline methods reported near-at-random performance based on the three evaluation metrics. Figure 1C visualizes an exemplar predicted sample with a representative seven biomarker panel. Finally, we train and evaluate a generative adversarial network (GAN) using the pix2pix<sup>29</sup> model architecture to predict CODEX expression and show that such an approach significantly underperforms ROSIE (Supplementary Table 1, Supplementary Fig. 11).

ROSIE generates highly accurate full-sample CODEX images and recapitulates salient visual features across representative immune and structural biomarker panels (Fig. 2A). To illustrate the robustness of ROSIE across a range of predictions, including relatively low-performing ones, we display side-by-side predicted and ground truth samples drawn from the 99th, 75th, 50th, and 25th percentiles of performance (by Pearson R) in the Stanford-PGC dataset. Figure 2B shows the Pearson R score for each of the 50 predicted biomarkers averaged across all evaluation datasets. We also visualize every single biomarker individually (Supplementary Fig. 1) from Stanford-PGC (median samples by Pearson R), along with their distributions by Pearson R (Supplementary Fig. 2) and scores on rank tests (Spearman R and C-index, Supplementary Fig. 3).

To explain the predictions generated by ROSIE, we apply Grad-CAM<sup>30</sup>, a visual interpretability technique for CNNs, on randomly sampled input H&E image patches from the Stanford-PGC dataset to see where in the image the model pays attention when making predictions. Supplementary Fig. 10 visualizes the gradient heatmaps generated for 10 biomarkers from the randomly sampled image patches. We observe that for biomarkers that are primarily determined by the nuclear protein expression (e.g. DAPI, Ki67, PCNA), the heatmap values are located around the center of the patch; conversely, for biomarkers that may be more context-dependent (e.g. CD68, PanCK, ECad), the heatmap values are more diffuse and located around the cellular neighborhood.

### Generalization to unseen studies and disease types

ROSIE performs robustly even when evaluated on samples from clinical sites and disease types never seen during training. When evaluated on Ochsner-CRC and Tübingen-GEJ, two studies whose samples and disease types do not appear in the training dataset, ROSIE reports comparable average performance to Stanford-PGC: a Pearson R of 0.241 (vs. 0.319 on Stanford-PGC), Spearman R of 0.283 (vs. 0.386 on Stanford-PGC), and a sample-level C-index of 0.633 (vs. 0.694 on Stanford-PGC) across all 50 biomarkers (see Table 2). We confirm these results visually in Fig. 2C, which contains median samples (by Pearson R) from three other evaluation datasets (Ochsner-CRC, Tübingen-GEJ, and UChicago-DLBCL) with the same representative immune and structural biomarker panels.

We additionally validate ROSIE's performance on a colorectal cancer dataset imaged using Orion, a multiplexed immunofluorescence platform<sup>31</sup>. This dataset ("Orion-CRC") consists of co-stained, paired H&E and mIF whole slide images (WSIs). In our analysis,

we sample fifty 3000px by 3000px samples from 5 WSIs and evaluate ROSIE on its prediction performance of 17 overlapping protein biomarkers measuring using Spearman rank correlation. Supplementary Fig. 8 shows that on a dataset produced from a different mIF imaging platform than CODEX, ROSIE performs well on a subset of the biomarkers but also struggles in several biomarkers (e.g., ECad).

### A simple metric for postprocessing and filtering in silico staining quality

Batch effects, or variations across histopathology samples due to factors like staining quality, tissue type, and artifacts, can significantly affect deep learning model generalizability<sup>32,33</sup>. It is desirable, therefore, to be able to predict which samples might be of lower quality due to batch effects and exclude them from downstream analyses.

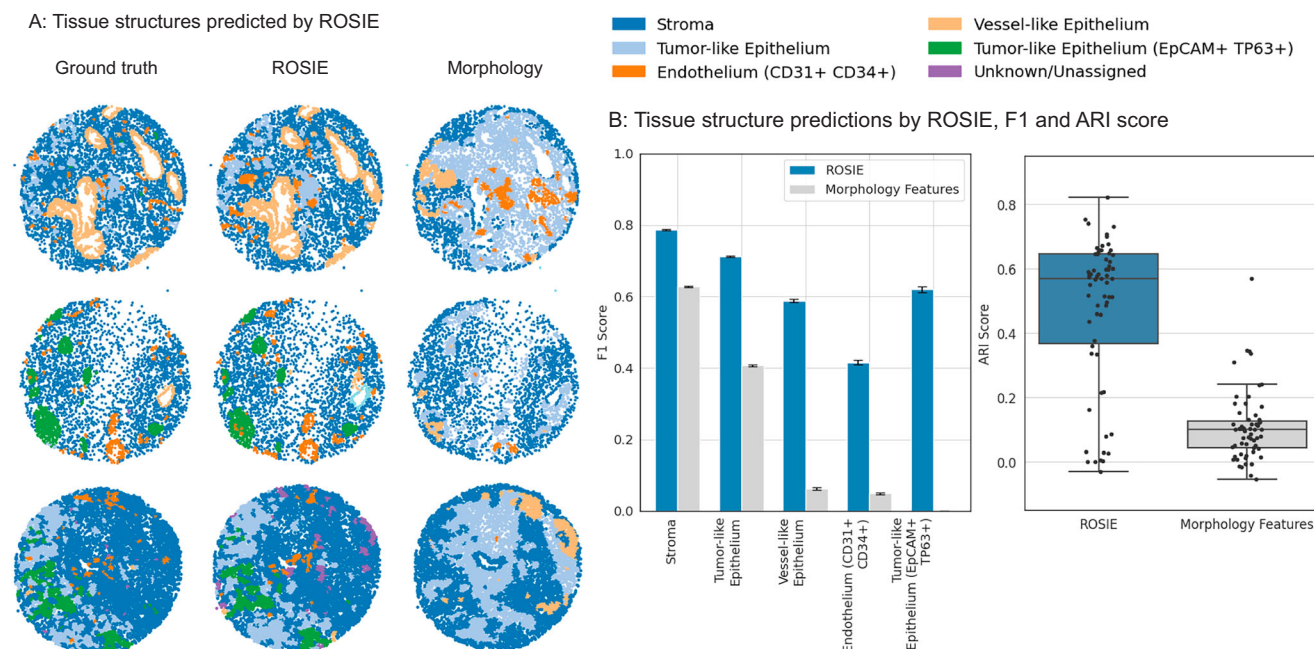
To this end, we introduce two simple but effective heuristics for scoring predicted samples on staining quality: dynamic range, which is a measure of the difference between the 99th and 1st percentile values in a biomarker stain; and W1 distance, which is the average Wasserstein distance between a test H&E image's histogram distribution and all histogram distributions from the training H&E image dataset. Supplementary Fig. 6 shows the effect of applying each quality filter to the four evaluation datasets: using the median as a cutoff for out-of-distribution samples, the average Pearson R score increases from 0.285 to 0.312 using W1 distance and to 0.336 using the dynamic range. Panel A of Supplementary Fig. 6 illustrates the relationship between the dynamic range and Pearson R score. Panel B likewise shows the relationship between the predicted Wasserstein distance and Pearson R score.

### Biomarker predictions are useful for cell and tissue phenotyping

Given that the protein biomarker expressions generated by ROSIE are highly correlated with ground truth measurements, we validate their biological and clinical usefulness by using them in phenotyping cells. To do this, we first train a nearest-neighbor algorithm to predict annotated cell labels on the ground truth CODEX biomarker expressions. Then, we input the biomarker expressions generated by ROSIE into the algorithm to produce cell type predictions. ROSIE can predict seven cell types (B cells, Endothelial cells, Epithelial cells, Fibroblasts, Macrophages, Neutrophils, and T cells) significantly better than a model using cell morphology and H&E RGB channels as inputs (*cell morphology*) or randomly assigning cell types according to average sample proportions (*bulk phenotyping*) (Fig. 3). Further analyses show strong B and T cell differentiation from the cell type classification confusion matrix (Supplementary Fig. 4). Furthermore, we perform cell phenotyping on the Ochsner-CRC dataset and find that the labels produced by the ROSIE-generated biomarkers performed comparably to Stanford-PGC (average F1 of 0.411 vs. 0.507, respectively) (Supplementary Table 2 and Supplementary Fig. 4).

We also evaluate a top-performing cell phenotyping algorithm, CellViT + +<sup>34</sup>, and compare its performance against ROSIE. CellViT + + is a Vision Transformer model trained on H&E images to predict cell phenotypes. To perform a side-by-side evaluation, we use the model checkpoint that has been trained on the Lizard dataset<sup>35</sup> and reconcile the cell types from Lizard and ROSIE into six categories: B cells, connective, epithelial, neutrophils, T cells, and other. Supplementary Fig. 9 shows that ROSIE outperforms CellViT + + in predicting these cell types from the Stanford-PGC dataset.

We are interested in whether these predictions validate therapeutically relevant distinctions between tissues. Different tissue types are typically known to be immunologically "hotter", indicating greater immune cell presence and infiltration, or "colder", implying less immune cell activity. For instance, pancreatic cancer (e.g., Stanford-PGC dataset) is known to be "colder" while colorectal cancer (e.g., Ochsner-CRC dataset) is known to be "hotter"<sup>36,37</sup>. Indeed, our results



**Fig. 4 | Tissue structure discovery by ROSIE.** Discovery of tissue structures using biomarkers generated by ROSIE on the Stanford-PGC test dataset. Five tissue structures are identified using a graph partitioning algorithm that clusters cells based on their expression profiles and neighboring cells. This algorithm is performed on both the ground truth measured and ROSIE-generated biomarker expressions and then reconciled to a common label set. **A** Visualizes several representative samples of tissue structures discovered using the ground truth CODEX measurements, ROSIE-generated expressions, and the morphology baseline

method. **B** Left: We report the F1 score ( $N = 635,649$  cells) by comparing the structures discovered using ground truth, ROSIE-generated biomarkers, and morphology features. Data are presented as mean values with error bars as the 95% bootstrapped confidence intervals. Right: ARI score is also reported by comparing the unlabeled discovered clusters, where each dot is a sample. Box plots show the median (center line), 25th and 75th percentiles (box edges), and the minimum and maximum values.

reflect this immunological validity check, with the immunologically “colder” Stanford-PGC having a lower predicted average proportion of T cells per sample of 20.0% (vs. 21.1% ground truth) and the immunologically “hotter” Ochsner-CRC having 40.1% T cells per sample (vs. 30.6% ground truth).

We also extend our method beyond single-cell phenotyping and demonstrate its effectiveness in identifying tissue structures within a sample. We use a top-performing tissue structure identification algorithm, SCGP<sup>38</sup>, which projects the acquired tissue sample into a graph, with nodes as cells and edges as neighboring cell pairs. Using this graph structure, the algorithm performs unsupervised clustering to discover tissue structures based on the ground truth and ROSIE-generated biomarker expressions separately. Figure 4 shows structures discovered in several samples and the reported Adjusted Rand Index (ARI) and F1 scores by comparing the ground truth and generated expressions. Our method achieves average ARI and F1 scores of 0.475 and 0.624, respectively, and is significantly higher than a baseline method of expressions generated from a three-layer neural network using cell morphology features derived from cell segmentations and mean H&E RGB values as inputs (ARI of 0.105 and F1 of 0.229). Full scores are reported in Supplementary Table 3.

Additionally, we use the ROSIE-generated expressions to identify two cell neighborhood phenotypes of interest: tumor-infiltrating lymphocytes (TILs) and lymphocyte neighboring epithelial cells (LNEs). These cell types are defined by their cellular niche and their biomarker expression profile: TILs are lymphocytes that reside in epithelial tissues, and LNEs are epithelial cells that neighbor lymphocytes. Figure 5 shows that the predicted proportions of TILs and LNEs per sample in the Stanford-PGC test dataset are highly correlated with the ground truth-derived proportions (Pearson R of 0.805 and 0.598 and Spearman R of 0.329 and 0.575 for TILs and LNEs, respectively), suggesting that the ROSIE-generated expressions may be useful for

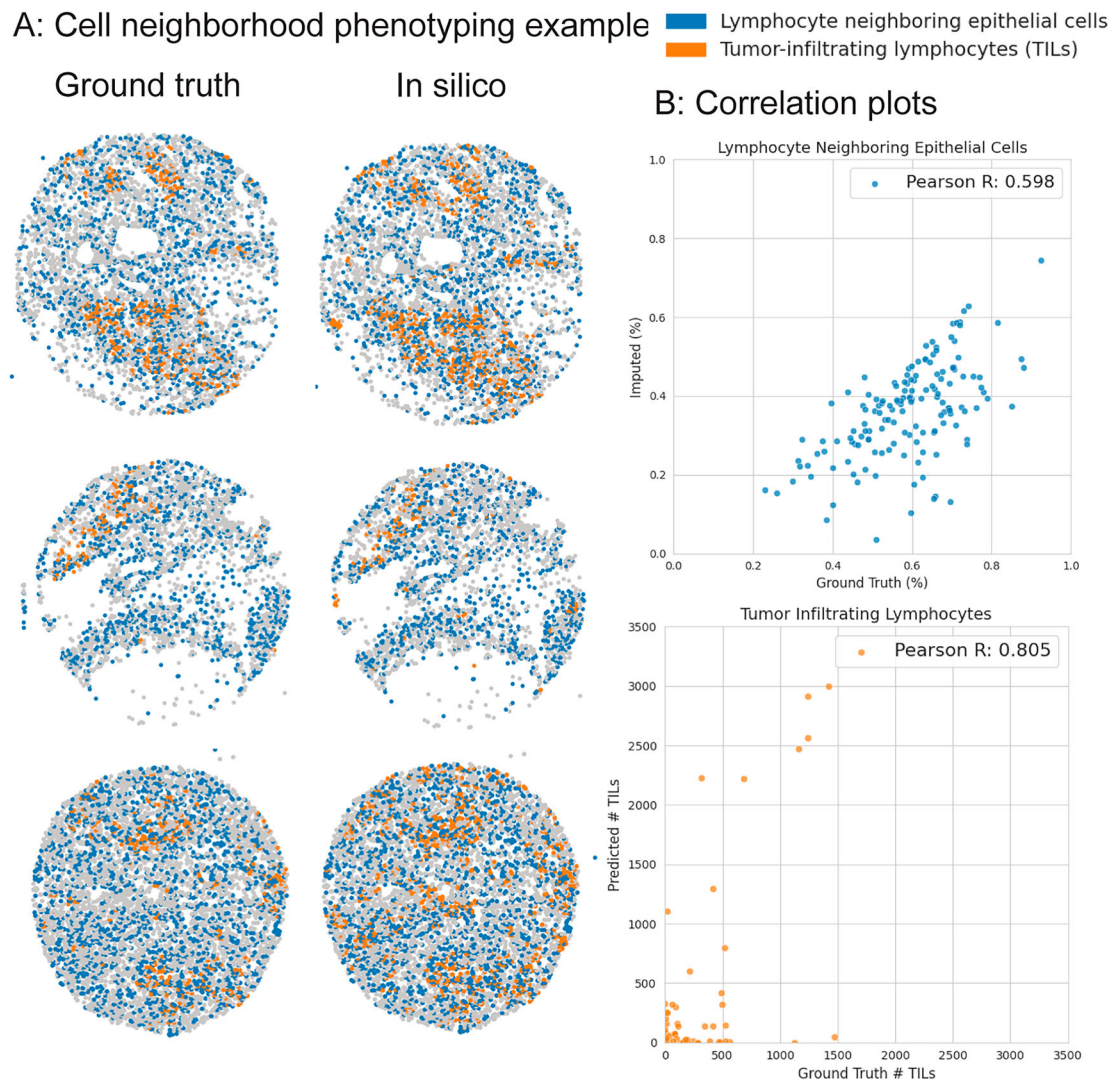
clinical tasks that involve estimating or ordering samples in a patient cohort by the presence of specific biomarkers, cells, or cell interactions.

## Discussion

Our study aims to bridge the gap between abundant, inexpensive H&E staining and the rich but more costly molecular information provided by multiplex immunofluorescence (mIF) staining. The primary question we shed light on is the extent to which H&E stains embed molecular hallmarks and features that could be computationally extracted. Our results indicate that although H&E staining has been traditionally limited in identifying few cell phenotypes, its structural and morphological features indeed contain significant information about protein expressions when analyzed with deep learning generative AI methods. This suggests that H&E staining has unrealized potential for use in clinical decisions that traditionally require more complex and expensive assays. Although prior work has focused on imputing up to several markers at a time, our study is the first to use deep learning to learn the relationship between H&E and up to 50 protein biomarkers at once. This setting now enables a more comprehensive view of the tissue microenvironment and offers more nuanced insights into specific tumor and immune cellular phenotypes.

Recent advances in Transformer architecture-based foundation models trained on histopathology images<sup>10,12,39</sup> show promise when adapted and fine-tuned on a wide range of clinical downstream tasks. However, our results indicate that such models still underperform our significantly smaller convolutional neural network. For instance, a ViT-L/16 Transformer model with over 300 M parameters pre-trained on histopathology images still did not perform as well as ConvNext, a 50 M parameter CNN pre-trained on non-pathology images. Indeed, a large-scale study comparing foundation and task-specific models<sup>40</sup> found that smaller task-specific models often outperform foundation





**Fig. 5 | Cell neighborhood phenotyping by ROSIE.** We identify two cellular neighborhood phenotypes of interest: tumor-infiltrating lymphocytes (TILs) and lymphocyte neighboring epithelial cells (LNEs). TILs are defined as cells that are labeled as lymphocytes (B and T cells) and reside in epithelial tissue (using the graph partitioning algorithm). LNEs are epithelial cells that have at least one

lymphocyte as a neighbor. TILs are measured as the raw count per sample, whereas LNEs are measured as the proportion of epithelial cells with lymphocyte neighbors. **A** We visualize three samples (by median Pearson R) of TILs and LNEs based on ground truth-derived and ROSIE-predicted expressions. **B** Scatter plots of predicted and ground truth measurements, where each dot represents a sample.

models, especially in scenarios with sufficient labeled training data. Additionally, Transformer-based foundation models are significantly more computationally expensive to train due to their larger model sizes. We offer several potential explanations for this counterintuitive finding: First, the inductive biases of convolutional operations in CNNs are better suited for extracting local contextual features from patch-based histology imaging; second, larger foundation models are harder to train and may overfit easier to the training data—indeed, we observed that the performance of the larger Vision Transformer models plateaus earlier in the process than compared to a CNN. This result suggests the importance of future research in analyzing when large foundation models are more appropriate than smaller CNNs.

We further demonstrated that our pixel-level CNN-based approach outperforms prior GAN-based methods in both accuracy and training stability. While GANs have been used for in silico staining, they face several limitations: (1) evaluation in prior literature has focused primarily on visual metrics such as SSIM rather than biologically meaningful accuracy, (2) instability during training due to the minimax optimization process, especially when jointly predicting a large panel of biomarkers (Supplementary Fig. 11), and (3) border artifacts

introduced by the patch-based processing of H&E inputs. In contrast, our model trains with a single, stable MSE loss and scales effectively to a 50-plex biomarker panel.

One focus of our study is on the rank correlation of our generated expressions and cellular phenotypes. Strong rank correlations suggest that the generated biomarkers are useful in clinical settings where the relative ordering of the presence of a biomarker or phenotype is important, e.g., finding patients that are the most receptive to a therapy or predicting patient prognosis based on a specific biomarker. For instance, we observe that while our model can effectively predict cell neighborhood phenotypes (TILs and lymphocyte neighboring epithelial cells), these predictions also exhibit biases in over- or under-predicting the proportion or counts of these phenotypes. Despite these biases, the relative ordering of these phenotypes is still largely maintained and thus is still useful in the settings mentioned above.

Inter-batch effects due to staining technique, quality, and machinery are known to cause variations in the image statistics of H&E stains. Since predictions generated on H&E stains that significantly deviate from the training data are expected to perform worse, we propose two methods for quantifying the quality of generated mIF

stains. We demonstrate that computing the dynamic range of predicted expressions and calculating the Wasserstein distance between training and generated data image histograms both correlate well with the empirical prediction accuracy. We believe these can be valuable tools that accompany our deep learning framework, allowing users to determine their clinically acceptable range of stain generation quality.

We acknowledge several limitations with our data and framework. Though we perform quality control on each sample (see Supplementary Methods), the alignment of H&E and CODEX images is susceptible to artifacts such as fraying, which can lead to misalignments and affect the accuracy of predictions. Second, not every biomarker is equally represented in the training data—this data imbalance is one reason why our performance on certain biomarkers is poor. We attempted to mitigate this by oversampling underrepresented biomarkers but did not observe significant improvement over equal sampling. As a result, we focus our phenotyping analysis on using the top 24 biomarkers, where performance is the most robust. Nonetheless, we observed experimental benefit from training on the full 50-plex panel versus a smaller panel (e.g., 10 and 24) even when evaluating only on the smaller panels. We hypothesize that training on a larger set encourages information sharing between the neural network layers and thus improves the overall predictive performance on the analyzed subset. An additional benefit of this approach is that in the ConvNeXt model architecture, all biomarkers are predicted from a shared embedding space (penultimate layer of the model), so biomarkers that are related to or a subset of other ones share the same representations in the model.

Another potential limitation in our approach is that there are limited context windows around cells. Given the model constraints, we experimented with larger context windows (i.e.  $256 \times 256$ px) but did not improve model performance. At the standard 40x magnification for H&E scanning, our current  $128 \times 128$ px window typically includes around thirty cells or a 3-hop neighborhood of cells, which would include information about not only the center cell but its surrounding niche. Additionally, all data collected and imaged in our study was performed in-house on the same experimental setup (e.g., H&E scanner, PhenoCycler Fusion). Due to this uniformity, we have limited experimental evidence demonstrating our model's robustness on data sourced from significantly different environments, as batch effects from different reagents and protocols would be expected to be more significant. Our evaluation on the Orion-CRC dataset highlights the performance of ROSIE when pushed to the boundaries of generalizability (different imaging platform, disease type, and clinical site): while we observe robust prediction on some biomarkers (e.g., CD45), other biomarkers underperform (e.g., ECad). These results imply fundamental limitations of ROSIE to easily generalize across multiple sources of variability without additional adaptation such as re-training. We look forward to additional validation of our method as more paired H&E/mIF data is made publicly available.

Due to the variability in biomarker performance across the full panel, ROSIE is not an appropriate replacement for a full-panel CODEX experiment. The primary aim of our method of training on the full 50-biomarker panel is to 1. maximize the amount of signal available to the model via information sharing from different markers, and 2. determine in an unsupervised manner which markers are the most predictive. The inherent limitations of H&E staining mean that resolving certain protein expressions (e.g., immune markers) are difficult without additional assaying. However, we demonstrate that ROSIE can still reveal significantly more information compared to the H&E stain alone.

Our study demonstrates a method for extracting multi-plex spatially resolved protein expression from H&E stains. Given the ubiquity of H&E staining in clinical workflows, a framework for enabling *in silico* staining of dozens of spatially resolved protein biomarkers offers enormous potential for improving clinical workflows and decision-making. *In silico* staining can reduce the need for costly

immunostaining assays, significantly decreasing turnaround time and associated clinical costs. Additionally, it can serve as a screening tool using existing H&E slides from patient cohorts, enabling clinicians to prioritize patients based on predicted therapeutic responses. Finally, in resource-limited settings, *in silico* staining can democratize access to advanced biomarker assays that would otherwise be unavailable. We hope that future validation and continued improvements to ROSIE will enable these use cases to be realized in real-world clinical settings.

## Methods

Patient samples and data were obtained using institutional protocols.

### CODEX data collection

All samples are prepared, stained, and acquired following CODEX User Manual Rev C.

**Coverslip preparation.** Coverslips are coated with 0.1% poly-L-lysine solution to enhance adherence of tissue sections prior to mounting. The prepared coverslips are washed and stored according to the guidelines in the CODEX User Manual.

**Tissue sectioning.** formaldehyde-fixed paraffin-embedded (FFPE) samples are sectioned at a thickness of 3–5  $\mu$ m on the poly-L-lysine coated glass coverslips.

**Antibody conjugation.** Custom conjugated antibodies are prepared using the CODEX Conjugation Kit, which includes the following steps: (1) the antibody is partially reduced to expose thiol ends of the antibody heavy chains; (2) the reduced antibody is conjugated with a CODEX barcode; (3) the conjugated antibody is purified; (4) Antibody Storage Solution is added for antibody stabilization for long term storage. Post-conjugated antibodies are validated by SDS-polyacrylamide gel electrophoresis (SDS-PAGE) and quality control (QC) tissue testing, where immunofluorescence images are stained and acquired following standard CODEX protocols, then evaluated by immunologists.

**Staining.** CODEX multiplexed immunofluorescence imaging was performed on FFPE patient biopsies using the Akoya Biosciences PhenoCycler platform (also known as CODEX). 5  $\mu$ m thick sections were mounted onto poly-L-lysine-treated glass coverslips as tumor microarrays. Samples were pre-treated by heating on a 55°C hot plate for 25 minutes and cooled for 5 minutes. Each coverslip was hydrated using an ethanol series: two washes in HistoChoice Clearing Agent, two in 100% ethanol, one wash each in 90%, 70%, 50%, and 30% ethanol solutions, and two washes in deionized water (ddH<sub>2</sub>O). Next, antigen retrieval was performed by immersing coverslips in Tris-EDTA pH 9.0 and incubating them in a pressure cooker for 20 minutes on the High setting, followed by 7 minutes to cool. Coverslips were washed twice for two minutes each in ddH<sub>2</sub>O, then washed in Hydration Buffer (Akoya Biosciences) twice for two minutes each. Next, coverslips were equilibrated in Staining Buffer (Akoya Biosciences) for 30 minutes. The conjugated antibody cocktail solution in Staining Buffer was added to coverslips in a humidity chamber and incubated for 3 hours at room temperature or 16 hours at 4°C. After incubation, the sample coverslips were washed and fixed following the CODEX User Manual. Additional antibody information including clone name, barcode number, and reporter name is available at: <https://gitlab.com/enable-medicine-public/rosie/-/blob/main/Antibody%20Information.xlsx>.

**Data acquisition.** Sample coverslips are mounted on a microscope stage. Images are acquired using a Keyence microscope that is configured to the PhenoCycler Instrument at a 20X objective. All of the sample collections were approved by institutional review boards.

To correct for possible autofluorescence, “blank” images were acquired in each microscope channel during the first cycle of CODEX and during the last. For these images, no fluorophores were added to the tissue. These images were used for background subtraction. Typically, autofluorescence will decrease over the course of a CODEX experiment (due to repeated exposures). Thus, to correct each cycle, our method determines the extent of subtraction needed by interpolating between the first and last “blank” images.

### Sample preprocessing

Samples are first stained and imaged using CODEX antibodies on the Akoya Biosciences PhenoCycler platform and then stained and imaged with H&E on a MoticEasyScan Pro 6 N scanner with default settings and magnification (40x). The CODEX DAPI channel and a grayscale version of the H&E image are used to perform image registration. Both images have their contrast enhanced using contrast-limited adaptive histogram equalization. The SIFT features<sup>41</sup> of each image are then found and matched based on the RANSAC algorithm<sup>42</sup> to find an image transformation from the H&E image to the CODEX coordinate space. The transformation was limited to a partial affine transformation (combinations of translation, rotation, and uniform scaling). If the initial alignment based on the grayscale H&E image was not successful, the process was repeated using the nuclear channel of the deconvolved H&E image (using a predetermined optical density matrix<sup>43</sup>) or individual color channels of the H&E image. The H&E stains were performed centrally on-site using standard staining protocols. However, inter-user variability such as staining durations and slide handling techniques may have contributed to batch effects. Such variability introduces potential sources of technical heterogeneity, which could impact the consistency of tissue staining quality and consequently influence downstream analytical results.

### Training details

The training process is visualized in Fig. 1B. Each sample was first split into patches. In our standard approach, we subdivided a sample into non-overlapping 8x8px patches. The average expression within each patch is computed for every CODEX biomarker. Then, a 128 by 128px H&E patch is extracted centered on the 8x8px patch, which is used as input to the model. Thus, for instance, a 1024px by 1024px sample would yield a prediction of 128px by 128px resolution, as each 8px by 8px patch is represented by a single pixel prediction. Our model is based on a ConvNext-Small architecture with 50 M parameters. The model is pre-trained on the ImageNet image dataset. During training, random augmentations are performed: horizontal and vertical flipping, brightness, contrast, saturation, hue jittering, and normalization are all performed. The training task consists of the model predicting the mean expression of the center 8 × 8px of a 128 × 128px patch for each biomarker. Thus, the model performs multitask regression (i.e., a 50-length vector) for a given patch. Model validation during training is performed using two metrics: Pearson R and SSIM (Structural Similarity Index Measure). Pearson R is a correlation metric used to assess the similarity between the measured and predicted expressions; SSIM is a similarity metric that assesses the qualitative similarity between two samples. Pearson R is computed across patches, while SSIM is computed on the reconstructed samples. This is to evaluate the model's predictions both in terms of biological accuracy and visual similarity.

Each study has a different biomarker panel, so to account for missing biomarkers, we used a masked mean squared error loss where only the loss over present biomarkers is computed.

Training is performed on 4 V100 GPUs with a batch size of 256 and a learning rate of 1e-4 with the Adam<sup>44</sup> optimizer on a schedule that reduces by half every 30 K iterations. Models are trained until no improvement in this metric is observed for 75 K steps. On the same hardware, evaluation of a single TMA core takes approximately

5 minutes. Given that the ConvNext model is relatively lightweight (~50 M parameters), it can also be deployed on a single on-premise, consumer-grade GPU or CPU. Model training and evaluation are performed with Python using the torch<sup>45</sup>, pytorch-geometric<sup>46</sup>, scikit-learn<sup>47</sup>, and scipy<sup>48</sup> packages.

A description of the different studies comprising the full training dataset, including disease type, number of cells and samples, and biomarker panel, are available at <https://gitlab.com/enable-medicine-public/rosie/-/blob/main/Training%20Datasets.csv>.

### Sample generation

Inference is performed on the foreground H&E patches and then stitched into the predicted sample. In the standard analyses presented, a stride of 8px is used, which produces a predicted image that is 8x downsampled from the original image size. This produces predictions at a resolution of 3.02 microns per pixel. Using this image, we then produce cell-level expression predictions by upsampling the predicted image to the native resolution and then computing the average expression per cell based on the cell segmentation mask. The ROSIE-generated images are saved in TIFF image format, which are identical to the CODEX-measured images. Thus, any downstream analyses that can be performed on CODEX data can similarly be performed on ROSIE-generated data.

To produce higher-resolution images, we also demonstrate the predictions using 1px strides. In this setting, overlapping patches are generated at 64x the number of total predictions. This setting produces predictions at a resolution of 0.3775 microns per pixel. The resulting images are demonstrated and compared to the standard 8px setting in Supplementary Fig. 5.

### Quality control

We introduce two quantitative metrics for determining whether an H&E sample is in distribution and a high-quality generation. First, we measure the deviation between the image intensity distributions of a test H&E sample and the H&E samples in the training data. For a given test and training image pair, we extract 256 histogram bins from the image to obtain discrete distributions and then compute the Wasserstein (or Earth Mover's) distance (also called W1 distance) between the two distributions. The quality metric for a given test image, then, is computed as the average W1 distance across all training images. By using this metric, we can a priori determine whether a test sample is in distribution and appropriate for evaluation.

Additionally, at the biomarker channel level, we use the dynamic range as a simple proxy for estimating the quality of a predicted sample. Since channels with very low maximum expression correlate with poor quality acquisitions (due to artifacts, staining issues, etc.), we set a threshold below which we exclude biomarkers from evaluation. Dynamic range is computed as the difference between the 99th and 1st percentile values in a generated biomarker stain. Since the dynamic range is only a function of the predicted image, it does not depend on having ground truth CODEX measurements for an H&E sample.

### Expression metrics

We report three primary evaluation metrics for patch-level predictions: Pearson R (sklearn.metrics.pearsonr), Spearman R (sklearn.metrics.spearmanr), and concordance index, or C-index (lifelines.utils.concordance\_index). Pearson correlation is a measure of the linear relationship between the ground truth and predicted biomarker expressions. Additionally, we report two rank metrics (Spearman R and C-index) to assess the model predictions' usefulness in clinical tasks that rely on ordering patient samples by a specific biomarker expression or cell type count. Pearson and Spearman correlations are calculated as the average correlations across all ground truth and predicted CODEX patches. C-index is computed on the 75th percentile values for both ground truth and predicted CODEX for each biomarker and



across all samples. Only during training, SSIM is computed across ground truth and predicted images. To ensure that the metrics are calculated on valid data points, we exclude patches with a lower groundtruth expression value than the 90th percentile value of background noise for each biomarker.

### Baseline methods

We also introduce several baseline methods for comparison to ROSIE:

First, we use the H&E sample alone to predict CODEX expression (called *H&E expression*). In this method, we apply a simple threshold ( $> 50$ ), averaged across the three color channels, and then use the intensity to predict each biomarker. This is to evaluate the similarity of the hematoxylin and eosin stains to the CODEX stains and to validate that the model is not simply recapitulating stain intensity in its CODEX predictions.

Second, we compute morphology statistics based on the segmentation masks calculated from the DAPI channel (called *cell morphology*). Cell segmentation is performed using the DeepCell algorithm<sup>49</sup>. We use the HistomicsTK compute\_morphometry\_features function and extract 19 features in total: Orientation, Area, Convex Hull Area, Major Axis Length, Minor Axis Length, Perimeter, Circularity, Eccentricity, Equivalent Diameter, Extent, Minor to Major Axis Ratio, Solidity, Hu Moments (1st to 7th). In addition, we compute the average across each of the RGB channels and include these as three additional features. We train a three-layer multi-layer perceptron neural network using these features as input to predict the expression of 50 protein biomarkers. Each layer has 100 nodes followed by a ReLU activation function and is trained with a  $1e-4$  learning rate, mean squared error loss, and Adam optimizer. The weights that generated the best validation accuracy after 50 epochs are used. This is a stronger baseline that is intended to represent typical features (morphology and intensity) derived from H&E images. Finally, as an additional baseline for cell phenotyping, we assign cell labels randomly according to the average ground truth cell label proportions across all samples (called *bulk phenotyping*). This method approximates estimating cell type proportions through a cheaper, more readily available phenotyping technique than CODEX (like flow cytometry) and then using them to infer spatially located cell types.

To train the pix2pix GAN model, we use the pix2pix model architecture<sup>29</sup> and software package and randomly sample 256px by 256px H&E and CODEX patches from the Stanford-PGC training dataset. The model is trained to predict the 50-channel CODEX output image based on the H&E input. The model is trained for 50 epochs with 100 patches randomly chosen from each sample. The full implementation and code are provided in the Gitlab repository.

### Cell phenotyping

Cell phenotyping metrics are computed using the following steps: First, cell clusters are produced using Leiden clustering based on the cell-level CODEX measured expressions. These clusters are identified and merged based on cell expression within each cluster to produce manually annotated cell labels. Then, we trained a k-nearest neighbors algorithm (where  $k = 100$ ) to generate a graph based on these clusters, which is used to automatically generate the reference cell labels. To prevent class imbalance, we sample from each cell label class with equal proportion for the training set. We use the same kNN and the predicted expressions to generate the predicted cell labels. We report the F1 scores, which are relative to the cell typing determined using clustering on the CODEX measurements. In this analysis, we use only the top 24 biomarkers by Pearson correlation as input features to the kNN algorithm. Supplementary Fig. 7 shows the mean biomarker expressions for each defined cell type. For cell phenotyping performed on Ochsner-CRC, we similarly train a kNN on manually annotated cell labels and use it to generate reference cell labels. For uniformity of comparison, we define the same cell phenotypes as in the Stanford-PGC dataset.

To perform a cell phenotyping comparison with CellViT ++, we use the model checkpoint that has been trained on the Lizard dataset<sup>35</sup> since the cell definitions most closely match the set derived in our analyses. We reconcile the cell types from Lizard and ROSIE into six categories: B cells, connective, epithelial, neutrophils, T cells, and other. Given that the CellViT ++ algorithm jointly performs segmentation and phenotyping, we also reconcile the cell segmentations produced by the algorithm and our DAPI-derived segmentations by only including the intersecting set of cells from both sets. This yields a total of 220 K cells when evaluated on the Stanford-PGC evaluation dataset.

### Tissue structure discovery

The Spatial Cellular Graph Partitioning (SCGP) framework<sup>38</sup> is summarized in the following steps:

1. Construct a graph with cells as nodes. Spatial edges are added between neighboring cell pairs, and feature edges are added for cell pairs with similar expression profiles.
2. Partitions are detected by community detection algorithms such as the Leiden algorithm<sup>50</sup>.
3. Each partition is manually annotated based on its underlying expression profile and cell morphology.

The above steps are performed independently on the ground truth and H&E imputed mIF samples, and a mapping from the imputed partitions to the ground truth partitions is calculated. Finally, we compute the adjusted Rand Index (ARI) and F1 scores. ARI measures the similarity between the ground truth and imputation-derived partitions and does not require cluster labeling; the F1 score is computed over manually annotated labels. As a baseline, we perform the partitioning over morphology features extracted from the cell segmentations, as well as adding in the average RGB expressions per cell. Supplementary Fig. 7 shows the mean biomarker expressions for each defined tissue structure.

### Cell neighborhood phenotyping

To define cell neighbors, we first perform cell segmentation on the DAPI channel for each sample. Based on the computed cell centroids, we then construct a Delaunay triangulation and Voronoi diagram, from which we then construct a graph with cells as nodes and Delaunay neighbors as edges. To define lymphocyte neighboring epithelial cells, we identify epithelial cells and then find the subset of these cells that share an edge with a lymphocyte (B cell or T cell). The reported percentage of LNEs is defined as the proportion of epithelial cells in a sample that are LNEs. Additionally, we define tumor-infiltrating lymphocytes (TILs) as lymphocytes embedded in tumor regions. To identify TILs, we find lymphocytes that are assigned to epithelial tissue structures. For each sample, we report the raw count of TILs.

### Gradient heatmaps

To interpret how ROSIE generates biomarker predictions from H&E images, we apply Gradient-weighted Class Activation Mapping (Grad-CAM)<sup>30</sup> to visualize the spatial regions within an input patch that influence the model's predictions. Specifically, we use the gradients of each biomarker's output with respect to the final convolutional feature maps to compute a weighted sum of activations, producing a heatmap that localizes the regions most relevant to the prediction. These gradients are globally average-pooled across the spatial dimensions to obtain a set of weights, which are then multiplied by the corresponding activation maps, summed, and upsampled to produce a localization map.

### Statistics and reproducibility

No statistical method was used to predetermine sample size. No data were excluded from the analysis after initial aggregation of the full

training and evaluation datasets. Data splits (e.g., training, validation, testing) were performed on the coverslip level, with splits randomly chosen among coverslips. The investigators were not blinded to allocation during experiments and outcome assessment.

### Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

### Data availability

All data are available in the main text or the supplementary materials. The raw training data are not available due to licensing and data privacy restrictions. Model weights are available upon request by contacting [eric@enablemedicine.com](mailto:eric@enablemedicine.com). Source data are provided with this paper.

### Code availability

The full code repository, including files for evaluating ROSIE, are available at <https://gitlab.com/enable-medicine-public/rosie>.

### References

- Couture, H. D. Deep learning-based prediction of molecular tumor biomarkers from H&E: A practical review. *J. Pers. Med.* **12**, 2022 (2022).
- Graham S., et al. HoVer-net: Simultaneous segmentation and classification of nuclei in multi-tissue histology images [Internet]. arXiv [cs.CV]. 2018; Available from: <http://arxiv.org/abs/1812.06499>.
- Gamper J., Alemi Koohbanani N., Benet K., Khuram A., Rajpoot N. PanNuke: An Open Pan-Cancer Histology Dataset for Nuclei Instance Segmentation and Classification. In: Digital Pathology. Cham: Springer International Publishing; 2019. p. 11–19.
- Amgad M., et al. NuCLS: A scalable crowdsourcing approach and dataset for nucleus classification and segmentation in breast cancer. *Gigascience* [Internet] 2021;11. Available from: <https://academic.oup.com/gigascience/article/doi/10.1093/gigascience/gjac037/6586817>.
- Hickey, J. W., Tan, Y., Nolan, G. P. & Goltsev, Y. Strategies for accurate cell type identification in CODEX multiplexed imaging data. *Front Immunol.* **12**, 727626 (2021).
- Goltsev, Y. et al. Deep Profiling of Mouse Splenic Architecture with CODEX Multiplexed Imaging. *Cell* **174**, 968–81.e15 (2018).
- Wu, Z. et al. Graph deep learning for the characterization of tumour microenvironments from spatial protein profiles in tissue specimens. *Nat. Biomed. Eng.* **6**, 1435–1448 (2022).
- Phillips, D. et al. Highly multiplexed phenotyping of immunoregulatory proteins in the tumor microenvironment by CODEX tissue imaging. *Front Immunol.* **12**, 687673 (2021).
- Black, S. et al. CODEX multiplexed tissue imaging with DNA-conjugated antibodies. *Nat. Protoc.* **16**, 3802–3835 (2021).
- Huang, Z., Bianchi, F., Yuksekgonul, M., Montine, T. J. & Zou, J. A visual-language foundation model for pathology image analysis using medical Twitter. *Nat. Med.* **29**, 2307–2316 (2023).
- Lu, M. Y. et al. A visual-language foundation model for computational pathology. *Nat. Med.* **30**, 863–874 (2024).
- Chen, R. J. et al. Towards a general-purpose foundation model for computational pathology. *Nat. Med.* **30**, 850–862 (2024).
- Pati P., et al. Multiplexed tumor profiling with generative AI accelerates histopathology workflows and improves clinical predictions [Internet]. bioRxiv. 2023 [cited 2024 May 20];2023.11.29.568996. Available from: <https://www.biorxiv.org/content/10.1101/2023.11.29.568996v1>.
- Andani S., et al. Multi-V-Stain: Multiplexed Virtual Staining of Histopathology Whole-Slide Images. Available from: <https://doi.org/10.1101/2024.01.26.24301803>.
- Ghahremani, P. et al. Deep Learning-Inferred Multiplex Immunofluorescence for Immunohistochemical Image Quantification. *Nat. Mach. Intell.* **4**, 401–412 (2022).
- Burlingame, E. A. et al. SHIFT: speedy histological-to-immunofluorescent translation of a tumor signature enabled by deep learning. *Sci. Rep.* **10**, 17507 (2020).
- Bouteldja, N. et al. Tackling stain variability using CycleGAN-based stain augmentation. *J. Pathol. Inf.* **13**, 100140 (2022).
- Bouteldja, N., Klinkhammer, B. M., Schlaich, T., Boor, P. & Merhof, D. Improving unsupervised stain-to-stain translation using self-supervision and meta-learning. *J. Pathol. Inf.* **13**, 100107 (2022).
- Wieslander H., Gupta A., Bergman E., Hallström E., Harrison P. J. Learning to see colours: generating biologically relevant fluorescent labels from bright-field images [Internet]. bioRxiv. 2021 [cited 2024 May 20];2021.01.18.427121. Available from: <https://www.biorxiv.org/content/10.1101/2021.01.18.427121v3>.
- Cetin O., Chen M., Ziegler P., Wild P., Koeppl H. Deep learning-based restaining of histopathological images. In: 2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). p. 1467–1474. (IEEE, 2022).
- de Haan, K. et al. Deep learning-based transformation of H&E stained tissues into special stains. *Nat. Commun.* **12**, 4884 (2021).
- He, B. et al. AI-enabled in silico immunohistochemical characterization for Alzheimer's disease. *Cell Rep. Methods* **2**, 100191 (2022).
- Bian C., Philips B., Cootes T., Fergie M. HEMIT: H&E to Multiplex-immunohistochemistry Image Translation with Dual-Branch Pix2pix Generator [Internet]. arXiv [eess.IV]. 2024. <http://arxiv.org/abs/2403.18501>.
- Srinivasan G., et al. Potential to Enhance Large Scale Molecular Assessments of Skin Photoaging through Virtual Inference of Spatial Transcriptomics from Routine Staining. bioRxiv [Internet] 2023. <https://doi.org/10.1101/2023.07.30.551188>.
- Wu, E. et al. 7-UP: Generating in silico CODEX from a small set of immunofluorescence markers. *PNAS Nexus* **2**, gad171 (2023).
- Zhou, Z. et al. Virtual multiplexed immunofluorescence staining from non-antibody-stained fluorescence imaging for gastric cancer prognosis. *EBioMedicine* **107**, 105287 (2024).
- Liu Z., et al. A ConvNet for the 2020s [Internet]. arXiv [cs.CV]. 2022. <http://arxiv.org/abs/2201.03545>.
- Dosovitskiy A., et al. An image is worth 16x16 words: Transformers for image recognition at scale [Internet]. arXiv [cs.CV]. 2020 [cited 2024 Sep 5]. <http://arxiv.org/abs/2010.11929>.
- Isola P., Zhu J.-Y., Zhou T., Efros A. A. Image-to-image translation with conditional adversarial networks [Internet]. arXiv [cs.CV]. 2016 [cited 2025 Apr 8]. <http://arxiv.org/abs/1611.07004>.
- Selvaraju R. R., et al. Grad-CAM: Visual explanations from deep networks via Gradient-based localization [Internet]. arXiv [cs.CV]. 2016 [cited 2025 Apr 7]. <https://doi.org/10.1007/s11263-019-01228-7>.
- Lin, J.-R. et al. High-plex immunofluorescence imaging and traditional histology of the same tissue section for discovering image-based biomarkers. *Nat. Cancer* **4**, 1036–1052 (2023).
- Howard, F. M. et al. The impact of site-specific digital histology signatures on deep learning model accuracy and bias. *Nat. Commun.* **12**, 4423 (2021).
- Schmitt, M. et al. Hidden variables in deep learning digital pathology and their potential to cause batch effects: Prediction model study. *J. Med Internet Res* **23**, e23436 (2021).
- Hörst F. et al. CellViT++: Energy-efficient and adaptive cell segmentation and classification using foundation models [Internet]. arXiv [cs.CV]. 2025 [cited 2025 Apr 3]. <http://arxiv.org/abs/2501.05269>.
- Graham S. et al. Lizard: A large-scale dataset for colonic nuclear instance segmentation and classification [Internet]. arXiv [cs.CV]. 2021 [cited 2025 Apr 3]. <http://arxiv.org/abs/2108.11195>.

36. Ouyang, P. et al. Overcoming cold tumors: a combination strategy of immune checkpoint inhibitors. *Front Immunol.* **15**, 1344272 (2024).
37. Hartupée, C. et al. Pancreatic cancer tumor microenvironment is a major therapeutic barrier and target. *Front Immunol.* **15**, 1287459 (2024).
38. Wu, Z. et al. Discovery and generalization of tissue structures from spatial omics data. *Cell Rep. Methods* **4**, 100838 (2024).
39. Wang X. et al. A pathology foundation model for cancer diagnosis and prognosis prediction. *Nature* 2024;1–9.
40. Mulliqi N. et al. Foundation models -- A panacea for artificial intelligence in pathology? [Internet]. arXiv [cs.CV]. 2025. <http://arxiv.org/abs/2502.21264>.
41. Lowe, D. G. Distinctive image features from scale-invariant key-points. *Int J. Comput. Vis.* **60**, 91–110 (2004).
42. Fischler, M. A. & Bolles, R. C. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM* **24**, 381–395 (1981).
43. Ruifrok, A. C. & Johnston, D. A. Quantification of histochemical staining by color deconvolution. *Anal. Quant. Cytol. Histol.* **23**, 291–299 (2001).
44. Kingma D. P., Ba J. Adam: A method for stochastic optimization [Internet]. arXiv [cs.LG]. 2014 [cited 2024 Sep 10]. <http://arxiv.org/abs/1412.6980>.
45. Paszke A., et al. PyTorch: An imperative style, high-performance deep learning library. *Neural Inf Process Syst* [Internet] 2019;abs/1912.01703. <https://papers.nips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library>.
46. Fey M., Lenssen J. E. Fast graph representation learning with PyTorch Geometric [Internet]. arXiv [cs.LG]. 2019 [cited 2025 Jun 11]. <http://arxiv.org/abs/1903.02428>.
47. Pedregosa, F. et al. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
48. Virtanen, P. et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods* **17**, 261–272 (2020).
49. Greenwald, N. F. et al. Whole-cell segmentation of tissue images with human-level performance using large-scale data annotation and deep learning. *Nat. Biotechnol.* **40**, 555–565 (2022).
50. Traag, V. A., Waltman, L. & van Eck, N. J. From Louvain to Leiden: guaranteeing well-connected communities. *Sci. Rep.* **9**, 5233 (2019).

## Acknowledgements

We thank Rosalyn Grace Wu and Susan Wu for their helpful support. Funding: National Cancer Institute grant P20CA252733 (JRH, UP, LL, CIL); National Cancer Institute grant P50CA285275 (JRH, UP, LL, CIL); National Institutes of Health grant R01CA280639 (JRH, UP); NSF CAREER award 194292 (JZ).

## Author contributions

Conceptualization was led by E.W., A.E.T., A.T.M. and J.Z. Methodology, investigation, and visualization were carried out by E.W., A.E.T., Z.W. and N.T. G.W.C., A.M., C.M.S., J.R.H., U.P., C.I.L., L.L., H.G., V.B. and A.R. provided resources. A.E.T., A.T.M. and J.Z. supervised the project. The original draft was written by E.W. and A.E.T. and the manuscript was reviewed and edited by M.B., Z.W., A.T.M. and J.Z.

## Competing interests

Several authors are affiliated with Enable Medicine as employees (M.B., Z.W., A.E.T., A.T.M.), consultants (E.W.), or scientific advisor (J.Z., C.M.S.). The remaining authors declare no competing interests.

## Ethical statement

Patient samples and data were obtained using institutional protocols. Sample collections were approved by institutional review boards at the clinical sites where the data were collected. UChicago - IRB protocol (13-1297) was approved by U of Chicago BSD/UCMC Institutional Review Board. Informed consent was not required in this retrospective study of anonymized samples; Stanford - IRB protocol (IRB-8) was approved by the Stanford University Institutional Review Board. Informed consent was not required in this retrospective study of anonymized samples; Ochsner - IRB protocol (10515) was approved by the Fred Hutchinson Cancer Center Institutional Review Board. Informed consent was not required in this retrospective study of anonymized samples; Tuebingen - Patients gave written informed consent and the use of their tissue samples and data was approved by the local Ethics Committee of the Canton of Bern (KEK 200/2014).

## Additional information

**Supplementary information** The online version contains supplementary material available at <https://doi.org/10.1038/s41467-025-62346-0>.

**Correspondence** and requests for materials should be addressed to Eric Wu, Alexandro E. Trevino, Aaron T. Mayer or James Zou.

**Peer review information** *Nature Communications* thanks Saad Nadeem, Israel da Silva, and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. A peer review file is available.

**Reprints and permissions information** is available at <http://www.nature.com/reprints>

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025