nature portfolio

Peer Review File

ROSIE: AI generation of multiplex immunofluorescence staining from histopathology images

Corresponding Author: Mr Eric Wu

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

Version 0:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
This paper presents a simple ConvNeXt based model for virtual multiplex immunofluorescence (mIF) staining of H&E images. Same-section H&E/mIF stained and co-registered tissue microarray dataset is created for training the model.

Unfortunately, the paper is not well-written and a bit convoluted. There are several weaknesses that need to be addressed or re-evaluated:

(1) Which 19 studies are used in the training? What specific disease types and the marker panels? There are different statements saying 1000 or 2000 samples? Which one is which? Most previous works focus on specific sites to allow for more rigorous evaluation and downstream analysis. Just predicting 50 biomarkers and then ignoring the large portion (26 out of 50) due to poor accuracy just represents a poor study design. According to Figure 2, easier markers such as CD3e and PanCK are performing way more poorly than the more complex markers such as macrophages (CD68 or CD20); this seems like a biased dataset issue maybe due to TMAs (and where those TMAs were sampled tumor center or at the boundary, etc)? Isn't it better to focus on few markers and clearly show the advantage with more rigorous quantitative evaluation? In the current form, the manuscript is all over the place and doesn't provide any significant scientific advancement or insights.

(2) There are already works (e.g. SegPath) using more than 1000 TMA H&E/mIF co-stained samples with more than 30 cancer types using models where the labels are transferred from mIF to H&E [1]. SegPath dataset is public as well as other co-stained datasets such as mIHC/mIF HNSCC TCIA dataset [2] and Orion-CRC [3]. Ideally, this study should refrain from predicting a hodgepodge of markers and focus on a specific disease type and go deeper into the analysis while evaluating on these publicly available same-section H&E/mIF datasets.

(3) Why the ConvNeXt model? Since this is paired/co-stained dataset, it will be important to compare against the standard baseline approaches like pix2pix to clearly show why this is helpful. There are well-defined strategies to counter the convergence and mode collapse issues so there is no reason not to compare. And most models use much larger patch sizes. Why 128x128? Most markers will perform poorly due to limited spatial context as shown by HoverNet cell segmentation and classification model which has really poor classification accuracy due to limited training patch resolution. Since the cells were segmented, can't the model be used to produce segmentation as well for more quantitative analysis similar to the standard metrics used in all the virtual staining literature. Comparisons with state-of-the-art cellular phenotyping algorithms e.g. CellVIT [4] and its latest version CellVIT2 will also be important to see where the proposed algorithms performance plateaus; class imbalance is a big problem when resolving lymphocytes and macrophages (no discussion is provided on how this is handled in the proposed algorithm).

(4) Another issue with blindly predicting 50 or even 100 markers is you ignore the underlying biological relationships that are represented by these markers, e.g. CD45 is a superset of CD3 and CD20 and CD3 is a superset of CD8 and FoxP3. How are these incorporated into the model to get the optimal performance out of virtual staining approaches? These are important questions to think about and answer before jumping onto pan-marker prediction for pan-cancer datasets.

(5) [Minor] Co-registration of TMA cores are typically challenging due to their cylinder shape. This should be more thoroughly addressed.

[1] Komura, Daisuke, Takumi Onoyama, Koki Shinbo, Hiroto Odaka, Minako Hayakawa, Mieko Ochi, Ranny Rahaningrum Herdiantoputri et al. "Restaining-based annotation for cancer histology segmentation to overcome annotation-related limitations among pathologists." Patterns 4, no. 2 (2023).

[2] Ghahremani, Parmida, Joseph Marino, Juan Hernandez-Prera, Janis V. de la Iglesia, Robbert JC Slebos, Christine H. Chung, and Saad Nadeem. "An AI-ready multiplex staining dataset for reproducible and accurate characterization of tumor immune microenvironment." In International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 704-713. Cham: Springer Nature Switzerland, 2023.

[3] Lin JR, Chen YA, Campton D, Cooper J, Coy S, Yapp C, Tefft JB, McCarty E, Ligon KL, Rodig SJ, Reese S. High-plex immunofluorescence imaging and traditional histology of the same tissue section for discovering image-based biomarkers. Nature cancer. 2023 Jul;4(7):1036-52.

[4] Hörst, Fabian, Moritz Rempe, Lukas Heine, Constantin Seibold, Julius Keyl, Giulia Baldini, Selma Ugurel et al. "Cellvit: Vision transformers for precise cell segmentation and classification." Medical Image Analysis 94 (2024): 103143.

(Remarks on code availability)
The repository is empty

Reviewer #2

(Remarks to the Author)
Summary:

The paper presents a deep learning-based framework (ROSIE) designed to load the expression and localization of multiple protein biomarkers from hematoxylin and eosin (H&E) histopathology images. The method utilizes a dataset comprising over 1,000 paired and co-registered H&E and multiplex immunofluorescence (mIF) images from 20 tissue types and disease conditions. By training a convolutional neural network (CNN) on these samples, ROSIE aims to predict 50 biomarker expressions, identifying key immune cell phenotypes (e.g., T cells, B cells, and TILs). As a result, the model provides spatially resolved biomarker information, which is crucial for understanding tumor microenvironment and developing personalized medicine strategies. The computational framework has potential applications in cancer research and computational pathology by enabling the extraction of molecular information from routine H&E slides, thus bypassing the need for expensive and time-intensive mIF staining.

Weaknesses:
- The absence of an interpretability method was noted. This approach is particularly important in artificial intelligence applications, as it allows for a better understanding of the model’s decision-making process and increases trust in its predictions. The study does not provide detailed model interpretability analyses.
- Although the training dataset was constructed from paired H&E and CODEX samples, it is unclear how well the model generalizes to other immunofluorescence platforms. Also, it seems that the model's performance on individual biomarkers varies, and some are predicted with significantly lower accuracy. Please clarify.
- How the batch effect variability between dataset from institutions could affect performance ? The authors mention that Vision Transformers (ViTs) did not outperform CNN-based architectures, but the comparison lacks depth.
- In light of advances in transformer-based foundation models for histopathology, could the use of those foundation models result in better performance?

How would ROSIE be incorporated into daily clinical practice, and what adjustments would be necessary for its effective deployment?

Suggestions for Improvement:
- Investigate why certain biomarkers perform worse and apply feature selection or ensembling techniques to improve low-confidence predictions;
- Evaluate ROSIE robustness against domain shifts and batch effects;
- Testing a transformer-based model pre-trained on large pathology datasets might enhance performance.

(Remarks on code availability)

Reviewer #3

(Remarks to the Author)
The ability to infer expression of multiple proteins from a single H&E stain, then use this information to characterise the tumour microenvironment, is potentially game changing and of wide interest. The work demonstrates proof-of concept and, overall, the results are impressive. However, I have some concerns with the manuscript, as outlined below.

The model’s performance, particularly its ability to predict immune cell subtypes appears to be overstated. 19/50 predicted

biomarkers have a Pearson's R < 0.25, including CD3, CD4 and CD8. The corresponding representative CODEX versus ROSIE-generated images for these less concordant markers are visually very different. It is in the ability to detect the expression of less prevalent markers, and thereby define more complex cell phenotypes, that the power of multiplex immunofluorescence lies.

More detail should be provided on the training and evaluation samples. For each study, the disease type, number of patients and biomarkers assessed should be reported. Do the TMA cores represent tumour tissue only, or include other tissue types (e.g. normal epithelium)?

Where the H&E stains performed locally at each study site or centrally using the same protocol and reagents? This is key to assessing generalisation.

It would be useful to provide more information on how ROSIE-generated images could be used in downstream analyses. For example, are a series of single-marker grayscale images generated that can be imported into digital image analysis software? It appears from Fig.S1 that this may be the case.

Any additional validation required to assess utility for other tissue types should be discussed, as should any barriers to incorporation into the clinical workflow (IT requirements, how long the model takes to run etc).

Other comments:
• The reported number of cells, samples, patches etc is inconsistent.
• B and C appear to be the wrong way around in the Figure 1 legend.
• Figure S6 is incorrectly referenced.
• Supplementary tables are labelled as supplementary figures
• The Sample Generation section of methods references 9px strides
• The colorectal & pancreatic datasets are incorrectly labelled and referenced in the discussion. 'Hot' CRC is also largely limited to the MSI+ subset.

(Remarks on code availability)
I can confirm that the code is available on gitlab but I do not have capacity to run/check it.

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
Thank you so much to the authors for their amazing effort. All my concerns have been addressed.

(Remarks on code availability)
The code is there now.

Reviewer #2

(Remarks to the Author)
The Authors have addressed all of my concerns with the original manuscript. The revised manuscript is ready for publication.

(Remarks on code availability)

Reviewer #3

(Remarks to the Author)
The authors have made significant amendments and, on the whole, my comments have been adequately addressed. However, I still have some concerns.

The addition of the Orion-CRC dataset analysis doesn't, in my view, strengthen the results. Instead, this highlights the difficulty in generalising across disease types. Some of the markers most accurately predicted in this dataset are among the least accurate in the combined evaluation dataset and vice versa.

There is still a discrepancy in the description of datasets with Table 1 stating that both the PGC & DLBLC datasets were used in training and evaluation.

Lines 254-255. The typos here have been corrected, but it is not entirely accurate to state that CRC is known to be immunologically 'hot'. Mismatch-repair deficient colorectal tumours are immunologically hot, but these account for only ~15% of CRC cases. That's not to say that the comparison with pancreatic cancer isn't valid, CRC is hotter, but this statement should be modified.

Lines 356-358 added to the discussion should be reworded. I assume this is referring to true co-expression. ‘Signal-sharing’ implies channel cross-talk, something to be avoided in mIF.

Lines 383-388. I agree with this statement but am not convinced ROSIE is sufficiently accurate to do this.

Batch effects introduced by inter-user variability at the same centre using the same protocol, as clarified in the methods (lines 451-456), would be far less significant than between centres using different reagents and protocols. This is another potential limitation for generalisability, which should be acknowledged in the discussion.

(Remarks on code availability)

Version 2:

Reviewer comments:

Reviewer #3

(Remarks to the Author)
Many thanks to the authors for their effort. My comments have all been addressed.

(Remarks on code availability)

REVIEWER COMMENTS

Reviewer #1 (Remarks to the Author):

This paper presents a simple ConvNeXt based model for virtual multiplex immunofluorescence (mIF) staining of H&E images. Same-section H&E/mIF stained and co-registered tissue microarray dataset is created for training the model.

Unfortunately, the paper is not well-written and a bit convoluted. There are several weaknesses that need to be addressed or re-evaluated:

(1) Which 19 studies are used in the training? What specific disease types and the marker panels? There are different statements saying 1000 or 2000 samples? Which one is which?

We have clarified the training and evaluation dataset details:

The training dataset consists of over 16M cells from 18 studies, 1342 samples, and 13 disease types. The evaluation dataset consists of 5M cells from 4 studies, 485 samples, and 4 disease types. One study is split between both training and evaluation datasets.

We have also included a table with the disease type and biomarkers present in each study:

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.

Most previous works focus on specific sites to allow for more rigorous evaluation and downstream analysis. Just predicting 50 biomarkers and then ignoring the large portion (26 out of 50) due to poor accuracy just represents a poor study design. According to Figure 2, easier markers such as CD3e and PanCK are performing way more poorly than the more complex markers such as macrophages (CD68 or CD20); this seems like a biased dataset issue maybe due to TMAs (and where those TMAs were sampled tumor center or at the boundary, etc)? Isn't it better to focus on few markers and clearly show the advantage with more rigorous quantitative evaluation? In the current form, the manuscript is all over the place and doesn't provide any significant scientific advancement or insights.

We have expanded our discussion on our approach of training on a larger set of 50 biomarkers:

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 signal sharing between channels 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.

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.

(2) There are already works (e.g. SegPath) using more than 1000 TMA H&E/mIF co-stained samples with more than 30 cancer types using models where the labels are transferred from mIF to H&E [1]. SegPath dataset is public as well as other co-stained datasets such as mIHC/mIF HNSCC TCIA dataset [2] and Orion-CRC [3]. Ideally, this study should refrain from predicting a hodgepodge of markers and focus on a specific disease type and go deeper into the analysis while evaluating on these publicly available same-section H&E/mIF datasets.

We thank the reviewer for these suggestions. To strengthen our results, we have included the Orion-CRC dataset as an additional evaluation dataset for the ROSIE model since it contains WSIs and a large number of overlapping biomarkers with our set. We have included the results from this experiment in the manuscript.

We additionally validate ROSIE's performance on a colorectal cancer dataset imaged using Orion, a multiplexed immunofluorescence platform [1]. 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. Figure S8 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).

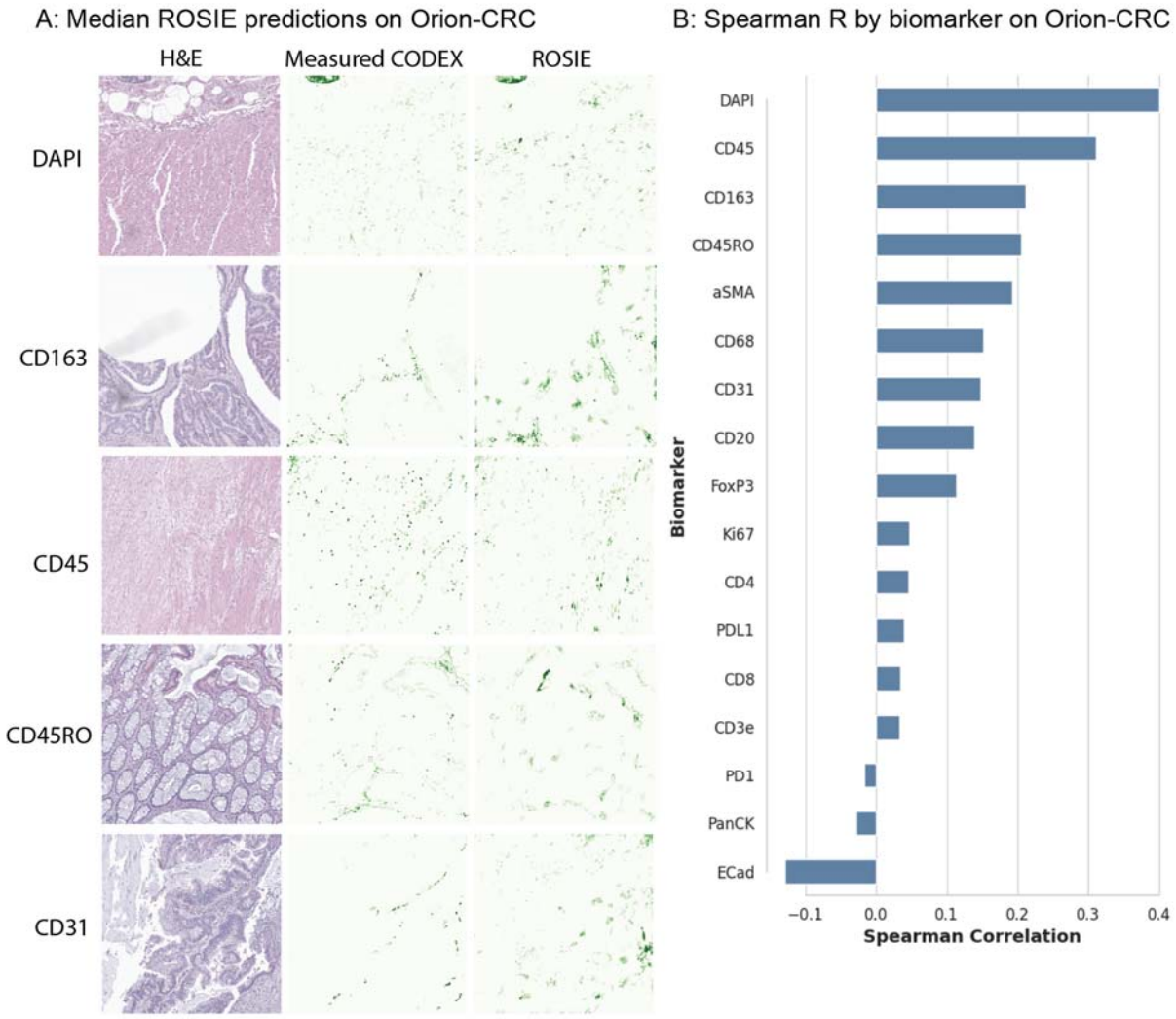

Fig. S8: *Evaluation of ROSIE on Orion-CRC, a colorectal dataset imaged on the Orion platform. (A) visualizes the median ROSIE-predicted sample (by Spearman R) for each of the top 5 biomarkers, along with the input H&E patch and the CODEX-measured ground truth measurement. (B) shows the overall Spearman R for each of the 17 overlapping biomarkers in the Orion-CRC dataset.*

(3) Why the ConvNeXt model? Since this is paired/co-stained dataset, it will be important to compare against the standard baseline approaches like pix2pix to clearly show why this is helpful. There are well-defined strategies to counter the convergence and mode collapse issues so there is no reason not to compare.

We have included a comparison to a baseline approach of using pix2pix to produce in silico staining in the following table and figure:

Stanford-PGC (50 biomarkers)	***Model parameters (#)***	***Pearson R***	***Spearman R***	***C-index (sample)***
- *ConvNext (Ours)*	*50.2M*	*0.319*	*0.386*	*0.694*
- *UNI (VitL-16)*	*304.3M*	*0.278*	*0.337*	*0.619*
- *VitB-16*	*86.6M*	*0.294*	*0.356*	*0.626*
- *Pix2pix (GAN)*	*57.3M*	*0.034*	*0.091*	*0.559*

Table S1: *Comparison of model architectures. Our CNN-based approach outperformed larger Vision Transformer models, histopathology image pre-training, and GAN-based approaches.*

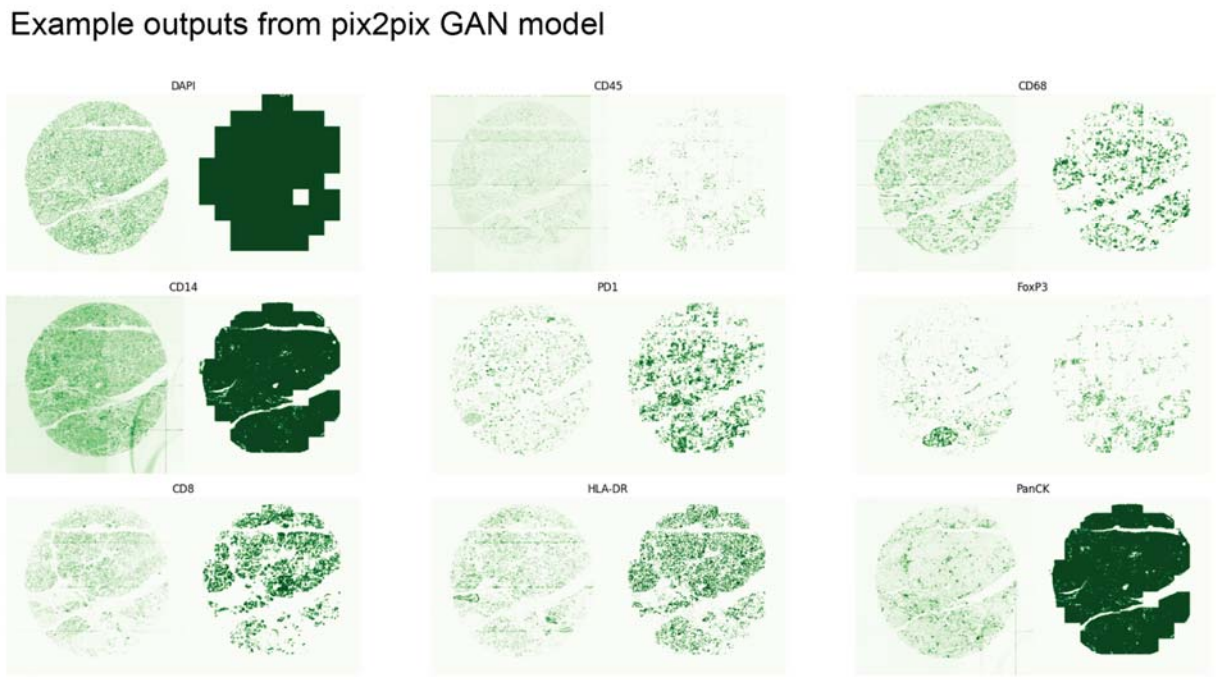

Fig. S11: Example outputs from a GAN model architecture (pix2pix). Each pair of images shows the CODEX-measured ground truth on the left and the GAN-model-generated output on the right. This example shows that GAN model outputs can be unstable and yield degenerate images, especially when jointly training on a large biomarker panel.

Finally, we train and evaluate a generative adversarial network (GAN) using the pix2pix [2] model architecture to predict CODEX expression and show that such an approach significantly underperforms ROSIE (Table S1, Figure S11).
To train the pix2pix GAN model, we use the model architecture described in [2] 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.

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 (Figure S11), 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.

And most models use much larger patch sizes. Why 128x128? Most markers will perform poorly due to limited spatial context as shown by HoverNet cell segmentation and classification model which has really poor classification accuracy due to limited training patch resolution.

We also add discussion on the effect of context size on the input patches:

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. 256x256px) but did not improve model performance. At the standard 40x magnification for H&E scanning, our current 128x128px 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.

Since the cells were segmented, can't the model be used to produce segmentation as well for more quantitative analysis similar to the standard metrics used in all the virtual staining literature.

We use the DAPI channel to perform cell segmentation and perform quantitative analysis at the cellular level for cell phenotyping analyses.

Comparisons with state-of-the-art cellular phenotyping algorithms e.g. CellVIT [4] and its latest version CellVIT2 will also be important to see where the proposed algorithms performance plateaus;

We have included a comparison to the CellViT++ model and demonstrate that our model performs more accurate cell phenotyping:

We also evaluate a top-performing cell phenotyping algorithm, CellViT++ [3], 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 [4] and reconcile the cell types from Lizard and ROSIE into six categories: B cells, connective, epithelial, neutrophils, T cells, and other. Figure S9 shows that ROSIE outperforms CellViT++ in predicting these cell types from 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 [4] 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 220K cells when evaluated on the Stanford-PGC evaluation dataset.

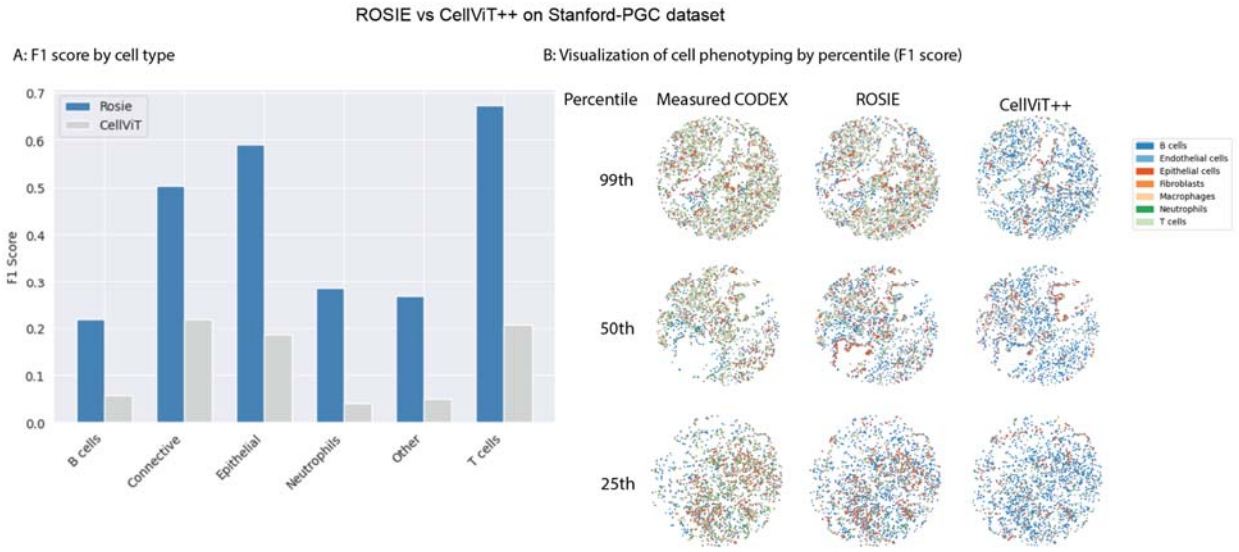

Fig. S9: *Comparison of CellViT++ and ROSIE on cell phenotyping for the Stanford-PGC dataset. A: F1 scores of the two algorithms on 220K cells for identifying six cell types. B: Visualizations of samples from Stanford-PGC, from the 99th, median, and 25th percentile scores by F1 score. The left column shows the cell types derived from the CODEX ground truth measurements, the middle column shows the cell types derived from the ROSIE-generated measurements, and the right column shows the CellViT++-predicted cell types.*

class imbalance is a big problem when resolving lymphocytes and macrophages (no discussion is provided on how this is handled in the proposed algorithm).

We have also added clarification on handling class imbalance in cell phenotyping:

To prevent class imbalance, we sample from each cell label class with equal proportion for the training set.

(4) Another issue with blindly predicting 50 or even 100 markers is you ignore the underlying biological relationships that are represented by these markers, e.g. CD45 is a superset of CD3 and CD20 and CD3 is a superset of CD8 and FoxP3. How are these incorporated into the model to get the optimal performance out of virtual staining approaches? These are important questions to think about and answer before jumping onto pan-marker prediction for pan-cancer datasets.

We have included a discussion section on how the model architecture allows for signal sharing across related biomarkers:

An additional benefit on 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.

(5) [Minor] Co-registration of TMA cores are typically challenging due to their cylinder shape. This should be more thoroughly addressed.

We describe the co-registration algorithm in detail in the Methods section. Despite inherent challenges in the co-registration of sequential slices, the H&E and CODEX staining are performed on the same sample, which yields a much stronger alignment.

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 6N 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 [5] of each image are then found and matched based on the RANSAC algorithm [6] 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 [7]) 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.*

[1] Komura, Daisuke, Takumi Onoyama, Koki Shinbo, Hiroto Odaka, Minako Hayakawa, Mieko Ochi, Ranny Rahaningrum Herdiantoputri et al. "Restaining-based annotation for

cancer histology segmentation to overcome annotation-related limitations among pathologists." Patterns 4, no. 2 (2023).

[2] Ghahremani, Parmida, Joseph Marino, Juan Hernandez-Prera, Janis V. de la Iglesia, Robbert JC Slebos, Christine H. Chung, and Saad Nadeem. "An AI-ready multiplex staining dataset for reproducible and accurate characterization of tumor immune microenvironment." In International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 704-713. Cham: Springer Nature Switzerland, 2023.

[3] Lin JR, Chen YA, Campton D, Cooper J, Coy S, Yapp C, Tefft JB, McCarty E, Ligon KL, Rodig SJ, Reese S. High-plex immunofluorescence imaging and traditional histology of the same tissue section for discovering image-based biomarkers. Nature cancer. 2023 Jul;4(7):1036-52.

[4] Hörst, Fabian, Moritz Rempe, Lukas Heine, Constantin Seibold, Julius Keyl, Giulia Baldini, Selma Ugurel et al. "Cellvit: Vision transformers for precise cell segmentation and classification." Medical Image Analysis 94 (2024): 103143.

Reviewer #1 (Remarks on code availability):

The repository is empty

Our GitLab repository has been updated with the relevant analysis and evaluation code.

Reviewer #2 (Remarks to the Author):

Summary:

The paper presents a deep learning-based framework (ROSIE) designed to load the expression and localization of multiple protein biomarkers from hematoxylin and eosin (H&E) histopathology images. The method utilizes a dataset comprising over 1,000

paired and co-registered H&E and multiplex immunofluorescence (mIF) images from 20 tissue types and disease conditions. By training a convolutional neural network (CNN) on these samples, ROSIE aims to predict 50 biomarker expressions, identifying key immune cell phenotypes (e.g., T cells, B cells, and TILs). As a result, the model provides spatially resolved biomarker information, which is crucial for understanding tumor microenvironment and developing personalized medicine strategies. The computational framework has potential applications in cancer research and computational pathology by enabling the extraction of molecular information from routine H&E slides, thus bypassing the need for expensive and time-intensive mIF staining.

Weaknesses:

- The absence of an interpretability method was noted. This approach is particularly important in artificial intelligence applications, as it allows for a better understanding of the model's decision-making process and increases trust in its predictions. The study does not provide detailed model interpretability analyses.

We have included a Supplementary figure that expands on using a gradient heatmap method for interpreting our CNN model:

To interpret how ROSIE generates biomarker predictions from H&E images, we apply Gradient-weighted Class Activation Mapping (Grad-CAM) [8] 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.

To explain the predictions generated by ROSIE, we apply Grad-CAM [8], 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. Figure S10 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.

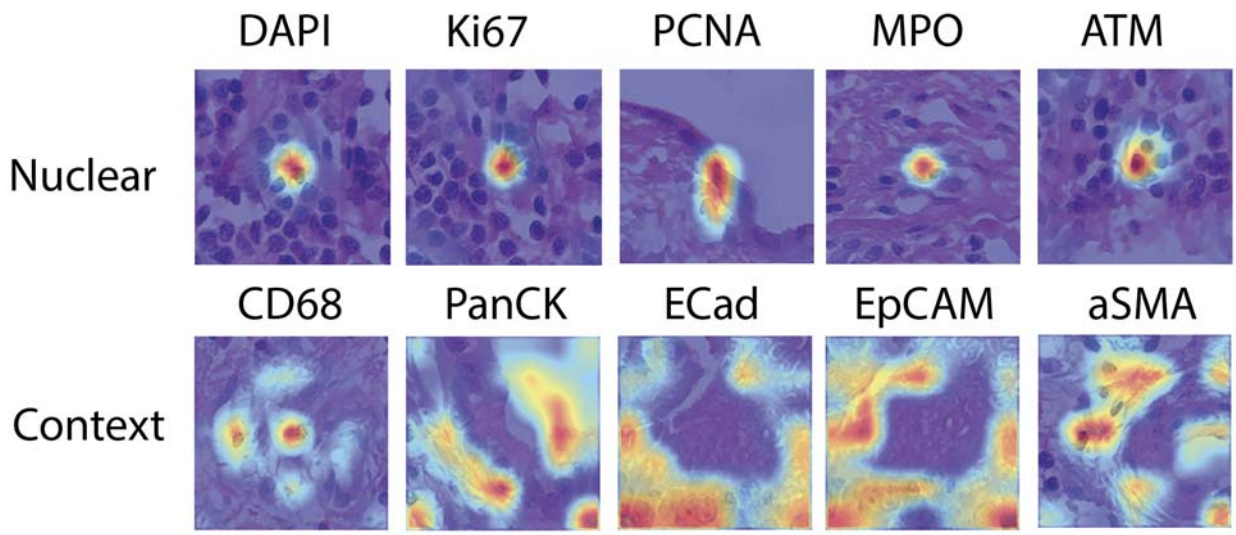

Fig. S10: *Gradient heatmaps created using GRAD-CAM for select biomarkers on a sample from Stanford-PGC. We visualize randomly sampled input H&E patches along with the gradient heatmaps derived from ROSIE, indicating where the model attended to in the image. The top row shows biomarkers that are associated with nuclear morphology, whereas the bottom row shows biomarkers that are typically defined by the surrounding context.*

- Although the training dataset was constructed from paired H&E and CODEX samples, it is unclear how well the model generalizes to other immunofluorescence platforms.

We have added an evaluation the performance of our model on the Orion-CRC dataset, which uses the Orion immunofluorescence platform (16-18 plex IF). We have also added discussion on why certain biomarkers are easier to predict than others.

We additionally validate ROSIE's performance on a colorectal cancer dataset imaged using Orion, a multiplexed immunofluorescence platform[1]. 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. Figure S8 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).

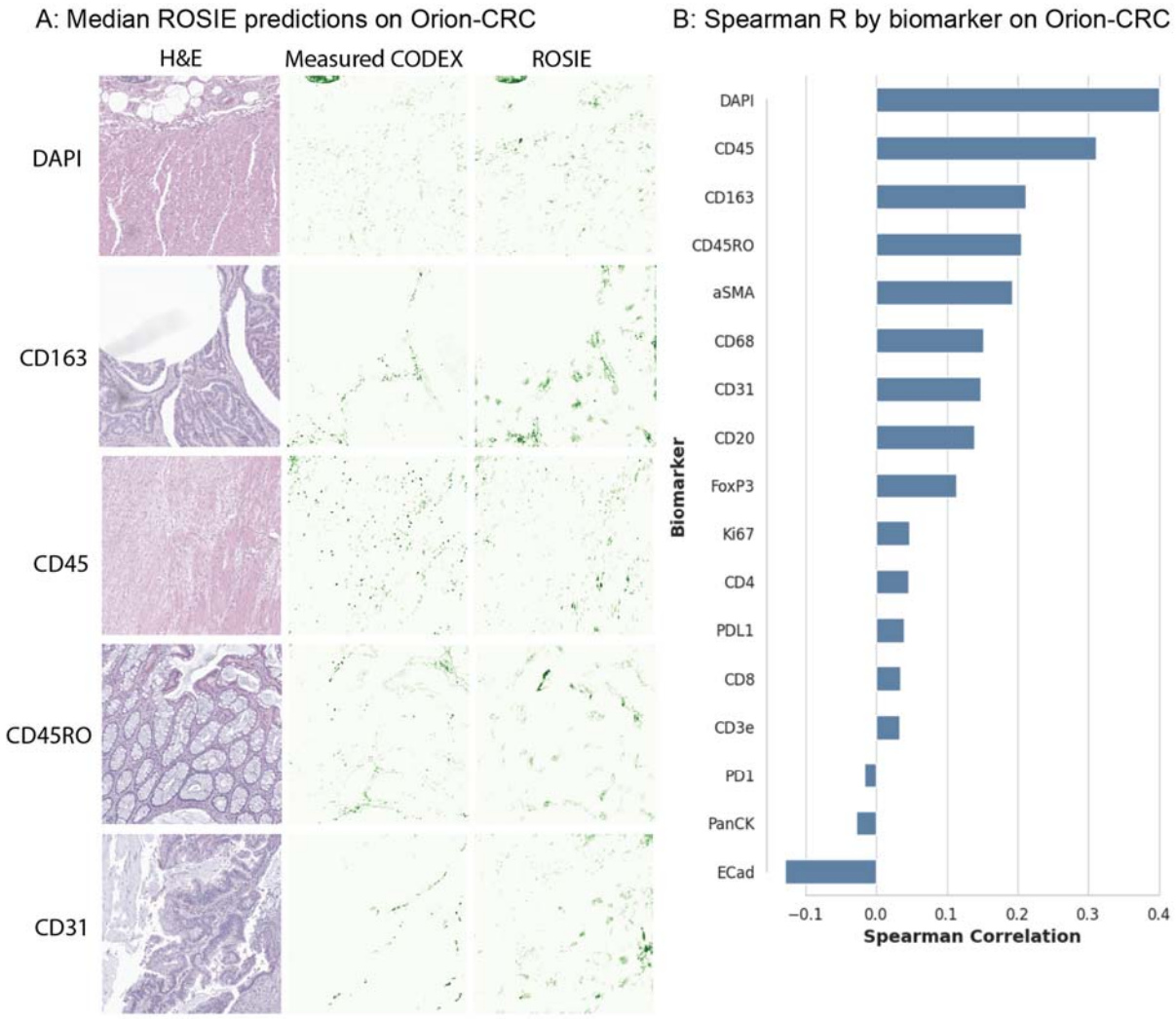

Fig. S8: *Evaluation of ROSIE on Orion-CRC, a colorectal dataset imaged on the Orion platform. (A) visualizes the median ROSIE-predicted sample (by Spearman R) for each of the top 5 biomarkers, along with the input H&E patch and the CODEX-measured ground truth measurement. (B) shows the overall Spearman R for each of the 17 overlapping biomarkers in the Orion-CRC dataset.*

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.

Also, it seems that the model's performance on individual biomarkers varies, and some are predicted with significantly lower accuracy. Please clarify.

We have added additional discussion on this point:

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 signal sharing between channels 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.*

- How the batch effect variability between dataset from institutions could affect performance ? The authors mention that Vision Transformers (ViTs) did not outperform CNN-based architectures, but the comparison lacks depth.

- In light of advances in transformer-based foundation models for histopathology, could the use of those foundation models result in better performance?

We have added an evaluation of CellViT++, a ViT-based Transformer model trained on H&E images, as a comparison for cell phenotyping:

We also evaluate a top-performing cell phenotyping algorithm, CellViT++ [3], 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 [4] and reconcile the cell types from Lizard and ROSIE into six categories: B cells, connective, epithelial, neutrophils, T cells, and other. Figure S9 shows that ROSIE outperforms CellViT++ in predicting these cell types from 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 [4] 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 220K cells when evaluated on the Stanford-PGC evaluation dataset.

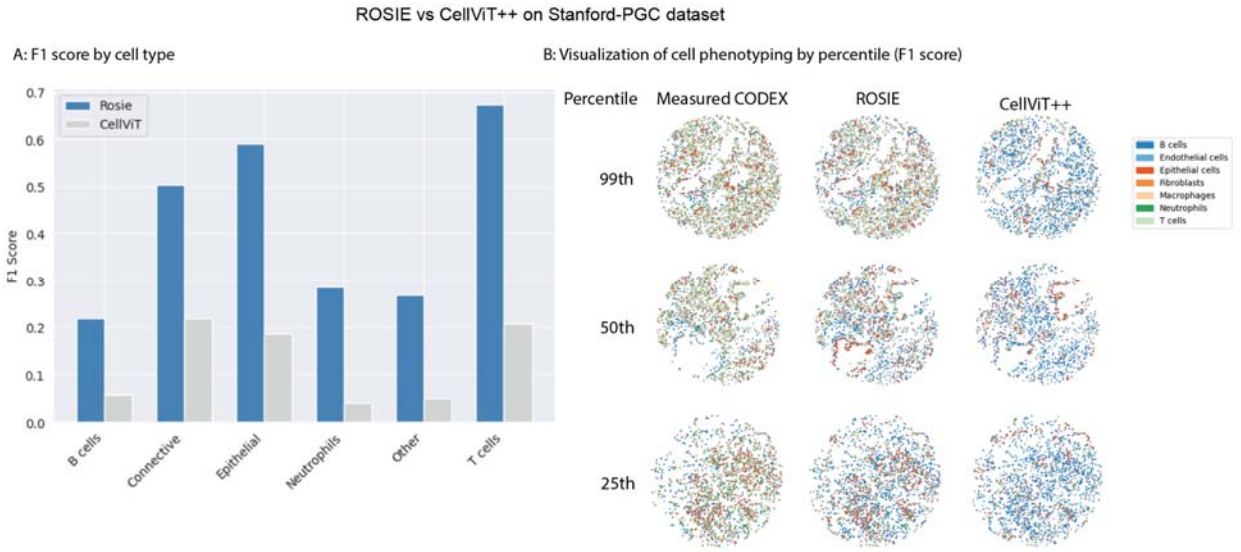

Fig. S9: *Comparison of CellViT++ and ROSIE on cell phenotyping for the Stanford-PGC dataset. A: F1 scores of the two algorithms on 220K cells for identifying six cell types. B: Visualizations of samples from Stanford-PGC, from the 99th, median, and 25th percentile scores by F1 score. The left column shows the cell types derived from the CODEX ground truth measurements, the middle column shows the cell types derived from the ROSIE-generated measurements, and the right column shows the CellViT++-predicted cell types.*

We also highlight our performance by comparing it with several ViTs including UNI, a Transformer-based histopathology foundation model.

Stanford-PGC (50 biomarkers)	**Model parameters (#)**	**Pearson R**	**Spearman R**	**C-index (sample)**
- *ConvNext (Ours)*	*50.2M*	*0.319*	*0.386*	*0.694*
- UNI (VitL-16)	304.3M	0.278	0.337	0.619
- VitB-16	86.6M	0.294	0.356	0.626
- Pix2pix (GAN)	57.3M	0.034	0.091	0.559

Table S1: Comparison of model architectures. Our CNN-based approach outperformed larger Vision Transformer models, histopathology image pre-training, and GAN-based approaches.

We have also expanded our discussion about ViTs and CNNs:

Indeed, a large-scale study comparing foundation and task-specific models [9] found that smaller task-specific models often outperform foundation 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.

How would ROSIE be incorporated into daily clinical practice, and what adjustments would be necessary for its effective deployment?

We have included a discussion about clinical deployment.

In silico staining can reduce the need for costly immunostaining assays, significantly decreasing turnaround time and associated clinical costs. Additionally, this method 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.

Suggestions for Improvement:

- Investigate why certain biomarkers perform worse and apply feature selection or ensembling techniques to improve low-confidence predictions;

We include additional discussion on the variability in biomarker performance:

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.

- Evaluate ROSIE robustness against domain shifts and batch effects;

We include evaluations on datasets from different domains (tissue/disease types) and batches than the training data in Table 1 and Figure 2C.

- Testing a transformer-based model pre-trained on large pathology datasets might enhance performance.

In Table S1, we included results from UNI, a transformer-based pre-trained histopathology foundation model and showed that it does not improve performance.

Reviewer #3 (Remarks to the Author):

The ability to infer expression of multiple proteins from a single H&E stain, then use this information to characterise the tumour microenvironment, is potentially game changing and of wide interest. The work demonstrates proof-of concept and, overall, the results are impressive. However, I have some concerns with the manuscript, as outlined below.

The model's performance, particularly its ability to predict immune cell subtypes appears to be overstated. 19/50 predicted biomarkers have a Pearson's R < 0.25, including CD3, CD4 and CD8. The corresponding representative CODEX versus ROSIE-generated images for these less concordant markers are visually very different. It is in the ability to detect the expression of less prevalent markers, and thereby define more complex cell phenotypes, that the power of multiplex immunofluorescence lies.

We have added the following piece of discussion to address this point:

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.

More detail should be provided on the training and evaluation samples. For each study, the disease type, number of patients and biomarkers assessed should be reported. Do the TMA cores represent tumour tissue only, or include other tissue types (e.g. normal epithelium)?

We have added additional details on the training, validation, and evaluation samples:

The training dataset consists of over 16M cells from 18 studies, 1342 samples, and 13 disease types. The evaluation dataset consists of 5M cells from 4 studies, 485 samples, and 4 disease types. One study is split between both training and evaluation datasets.

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.

The TMA cores are taken from both tumor and normal tissue. The exact number of patients is not available for all studies.

Where the H&E stains performed locally at each study site or centrally using the same protocol and reagents? This is key to assessing generalisation.

We have expanded the methods section to include this detail:

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.

It would be useful to provide more information on how ROSIE-generated images could be used in downstream analyses. For example, are a series of single-marker grayscale images generated that can be imported into digital image analysis software? It appears from Fig.S1 that this may be the case.

We have included the following section in the methods section to clarify the output format of ROSIE:

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 CDOEX data can similarly be performed on ROSIE-generated data.

Any additional validation required to assess utility for other tissue types should be discussed, as should any barriers to incorporation into the clinical workflow (IT requirements, how long the model takes to run etc).

We have included a discussion section on clinical deployment:

In silico staining can reduce the need for costly immunostaining assays, significantly decreasing turnaround time and associated clinical costs. Additionally, this method 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 have also included technical details about evaluating the model:

On the same hardware, evaluation of a single TMA core takes approximately 5 minutes. Given that the ConvNext model is relatively lightweight (~50M parameters), it can also be deployed on a single on-premise, consumer-grade GPU or CPU.

Other comments:

• The reported number of cells, samples, patches etc is inconsistent.

- B and C appear to be the wrong way around in the Figure 1 legend.
- Figure S6 is incorrectly referenced.
- Supplementary tables are labelled as supplementary figures
- The Sample Generation section of methods references 9px strides
- The colorectal & pancreatic datasets are incorrectly labelled and referenced in the discussion. ‘Hot’ CRC is also largely limited to the MSI+ subset.

Thank you for these comments, we have addressed them in the manuscript.

Reviewer #3 (Remarks on code availability):

I can confirm that the code is available on gitlab but I do not have capacity to run/check it.

References:

1. 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).
2. Isola, P., Zhu, J.-Y., Zhou, T. & Efros, A. A. Image-to-image translation with conditional adversarial networks. *arXiv [cs.CV]* (2016).
3. Hörst, F. *et al.* CellViT++: Energy-efficient and adaptive cell segmentation and classification using foundation models. *arXiv [cs.CV]* (2025).
4. Graham, S. *et al.* Lizard: A large-scale dataset for colonic nuclear instance segmentation and classification. *arXiv [cs.CV]* (2021).
5. Lowe, D. G. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis.* **60**, 91–110 (2004).
6. 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).
7. Ruifrok, A. C. & Johnston, D. A. Quantification of histochemical staining by color deconvolution. *Anal. Quant. Cytol. Histol.* **23**, 291–299 (2001).
8. Selvaraju, R. R. *et al.* Grad-CAM: Visual explanations from deep networks via Gradient-based localization. *arXiv [cs.CV]* (2016) doi:10.1007/s11263-019-01228-7.
9. Mulliqi, N. *et al.* Foundation models -- A panacea for artificial intelligence in pathology? *arXiv [cs.CV]* (2025).

REVIEWER COMMENTS

Reviewer #1 (Remarks to the Author):

Thank you so much to the authors for their amazing effort. All my concerns have been addressed.

Reviewer #1 (Remarks on code availability):

The code is there now.

Reviewer #2 (Remarks to the Author):

The Authors have addressed all of my concerns with the original manuscript. The revised manuscript is ready for publication.

Reviewer #3 (Remarks to the Author):

The authors have made significant amendments and, on the whole, my comments have been adequately addressed. However, I still have some concerns.

The addition of the Orion-CRC dataset analysis doesn't, in my view, strengthen the results. Instead, this highlights the difficulty in generalising across disease types. Some of the markers most accurately predicted in this dataset are among the least accurate in the combined evaluation dataset and vice versa.

Thank you for the suggestion, we have incorporated additional discussion to more directly highlight the difficulties of generalization from the Orion-CRC evaluation.

> *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.*

There is still a discrepancy in the description of datasets with Table 1 stating that both the PGC & DLBLC datasets were used in training and evaluation.

Thank you for highlighting this. We have updated the description in the methods:

> *We introduce a first-of-its-kind training and evaluation dataset across 20 studies (Figure 1A, Table 1; see Supplementary Materials for dataset details). The training dataset consists of over 16M cells from 18 studies, 1342 samples, and 13 disease types. The evaluation dataset consists of 5M 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.*

Lines 254-255. The typos here have been corrected, but it is not entirely accurate to state that CRC is known to be immunologically 'hot'. Mismatch-repair deficient colorectal tumours are immunologically hot, but these account for only ~15% of CRC cases. That's not to say that the comparison with pancreatic cancer isn't valid, CRC is hotter, but this statement should be modified.

Thank you, we have modified the manuscript to describe them as "hotter" and "colder":

> *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" [36,37]. Indeed, our results 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).*

Lines 356-358 added to the discussion should be reworded. I assume this is referring to true co-expression. 'Signal-sharing' implies channel cross-talk, something to be avoided in mIF.

Thank you, we have clarified that we are referring to the channels in the neural network:

> *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.*

Lines 383-388. I agree with this statement but am not convinced ROSIE is sufficiently accurate to do this.

We have modified and added a sentence to reflect this:

> *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.*

Batch effects introduced by inter-user variability at the same centre using the same protocol, as clarified in the methods (lines 451-456), would be far less significant than between centres using different reagents and protocols. This is another potential limitation for generalisability, which should be acknowledged in the discussion.

Thank you for this point, we have expanded our discussion to clarify this point:

> *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. We look forward to additional validation of our method as more paired H&E/mIF data is made publicly available.*