Translation of lung tissue photographs into protein distributions using imaging mass cytometry labels

Mingxuan Zhang
Weill Cornell Medicine
miz4003@med.cornell.edu

Tom Berry
Weill Cornell Medicine
thb4002@med.cornell.edu

Jiwei Yang
Cornell Tech
jy875@cornell.edu

Vaed Prasad
Cornell University
vsp22@cornell.edu

Inferring disease state and drug response from tissue imaging is an important goal of pathology, but is error-prone and requires expensive specialist labor. To remedy these issues, classifiers have been trained from photographs of histological stains to categorical clinical labels, but have not yet achieved widespread use. Here, we leverage image pairs generated by imaging mass cytometry to train a model that can infer the distribution of protein markers throughout tissue from tissue photographs. We expect our model to serve as a feature extractor that feeds into application-specific clinical classifiers. We report the results of training the model using TransUNet, pix2pix, and pix2pixHD architectures, finding pix2pixHD most effective.

1 Introduction

Pathologists identify features of interest in tissue through histological staining, immunohistochemistry (IHC), imaging mass cytometry (IMC), and other methods. The most popular histological stain, hematoxylin and eosin (H&E), contrasts nuclei and cytoplasm, but does not identify cell types (such as fibroblasts or macrophages), which are generally defined by the proteins on their cell surface. The spatial distribution of proteins is critical to understanding biological state. IHC and IMC reveal these proteins, but are expensive and time-consuming to perform. We present here a neural network model that, given a tissue photograph, reveals protein surface markers, effectively performing IHC "in silico".

2 Dataset

The dataset consists of 237 lung tissue regions. For each region, we have a photograph of the tissue and three protein distribution images: collagen, alpha-SMA and keratin. The photographs, which are cheaply obtained, are the inputs to our model. The protein distributions, obtained using resource-intensive IMC, are the targets of our model (Figure 1). Both the input photographs and protein targets are grayscale images. For each example, the input and targets have equal dimensions, but the dimensions vary across examples, from 192x960 to 2525x2144, with an average width of 1165 pixels and an average height of 1079 pixels.

2.1 Preprocessing

Each whole photograph input is composed of several small square 304x304 photographs stitched together, each of which has an identical shadow artifact. We found several small squares with near-empty contents and averaged them together, making an "average artifact". Then, for each whole input, we create a grid of "average artifacts" matching the whole input's dimensions and subtract

32nd Conference on Neural Information Processing Systems (NeurIPS 2018), Montréal, Canada.

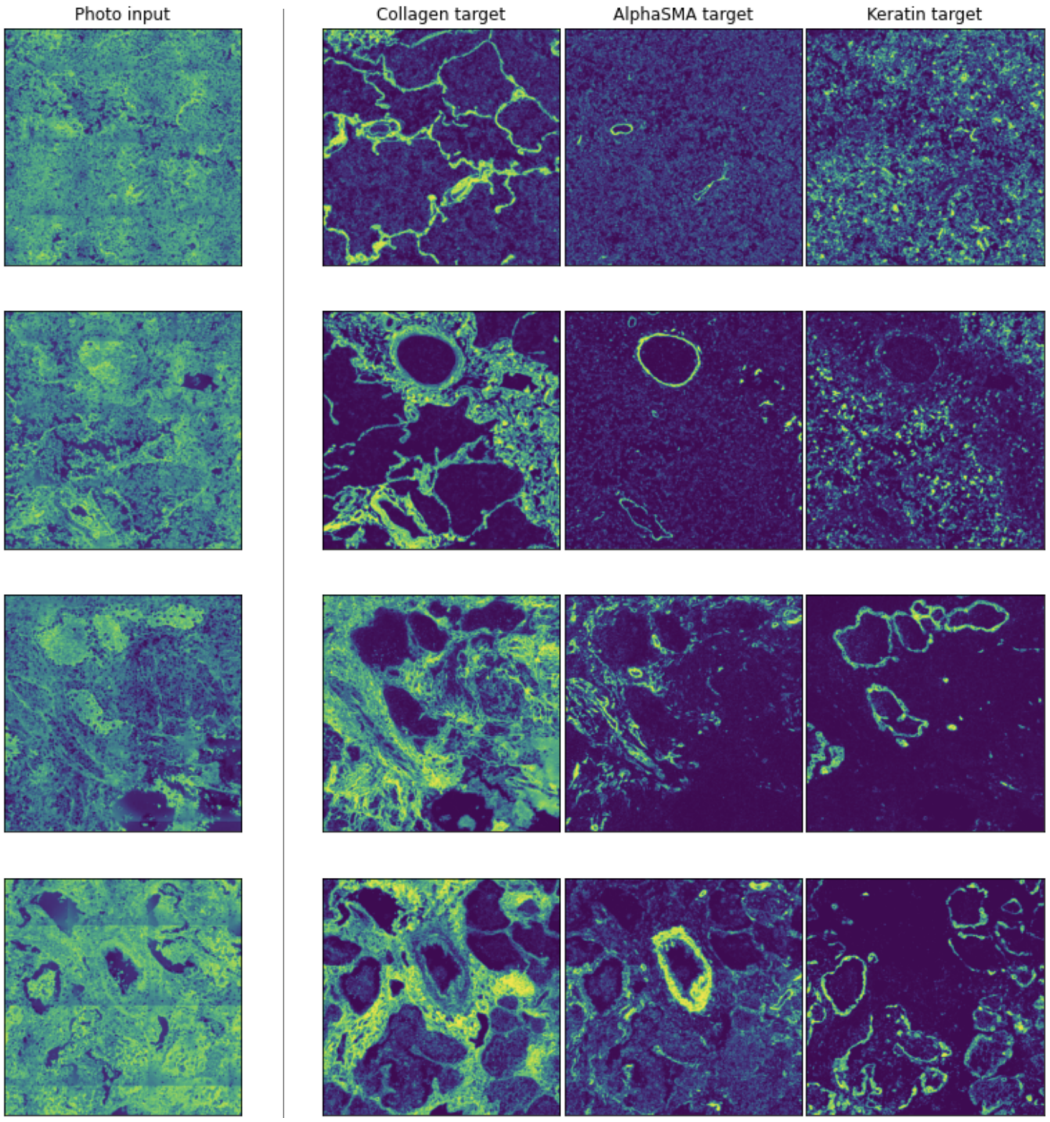


Figure 1: Four selected training examples

the grid from the input, greatly reducing the shadow artifacts (Figure 2). We then divide inputs and targets into the small 304x304 squares for training.

3 Background

Our problem can be framed either as semantic segmentation or paired image translation. Different neural network architectures have been developed for each approach. We explore both approaches to discover which performs better.

3.1 Semantic Segmentation

Semantic segmentation is a supervised learning task that attempts to learn categorical labels for each pixel in an image. One approach to the task, known as a "fully convolutional network", leverages a series of convolutional layers to transform the input image, concluding with a softmax layer that outputs a probability distribution for each pixel over the label categories. The U-Net architecture extends this idea with a symmetric network of contracting layers followed by upsampling

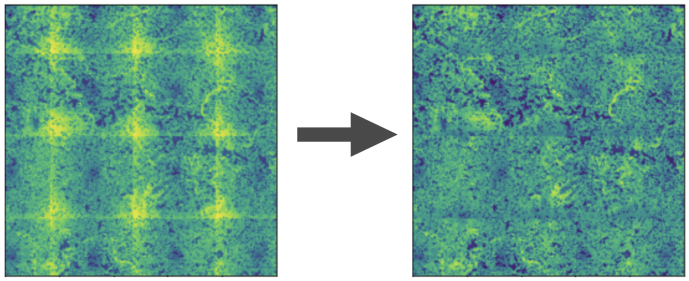


Figure 2: Shadow artifact removal

layers, with skip connections between the two halves to preserve spatial context [9]. The UNet++ architecture extends U-Net with a denser set of skip connections [18], while TransUNet incorporates a self-attention into the encoding process.

While each pixel in our target images is a continuous grayscale intensity, our task – revealing the location of proteins throughout an image – is easy to conceptualize as semantic segmentation. There are two ways we can map our data onto the problem. First, we could select an intensity threshold and learn to classify pixels to the two classes of low or high intensity. Second, we could normalize pixel intensities to the (0, 1) range and treat them as class probabilities. We take the latter approach to avoid the information loss that would result from thresholding.

3.2 Paired image translation

Image translation is the task of mapping an image from one domain to another domain. In paired image translation, we learn from image pairs that are aligned in pixel-space. Because both our inputs and outputs are grayscale images, our data fits this task well.

pix2pix is a GAN-based image translation architecture that is similar to the popular CycleGAN, but that works on image pairs. It combines a generator, implemented as a U-Net, with a learned discriminator. The discriminator combines an L_1 loss function with a GAN loss to synthesize images in the target domain. pix2pix has shown impressive results on image translation tasks similar to ours: for example, pix2pix can translate from satellite imagery to street maps – a translation from a denser domain to a sparser domain, similar to our translation from photographs to protein distributions. We investigate pix2pix and a similar model, pix2pixHD, which combines multiple generators and discriminators trained on varying image scales.

4 Models

4.1 TransUNet

TransUNet is a variant of UNet, which adopts a hybrid architecture combining a CNN and transformers for the encoding process [8]. The CNN first produces embeddings for the input images, as well as high resolution feature maps for skip connection in the decoding process. The embeddings are then fed into an array of transformers for encoding. The decoding process is similar to that of the original UNet. Therefore, TransUNet, which incorporates both Transformers and U-Net, can not only encode strong global context by treating the image features as sequences, but also utilize low-level CNN features via a U-shaped hybrid architectural design (Supplemental Figure 1).

Prediction results Figure 3 shows selected images generated by our TransUNet model. TransUNet achieves the best results on the collagen protein, shown in the leftmost column. The target highlights and general structure are captured well by TransUNet's predictions. However, TransUNet struggles to predict the other proteins (alpha-SMA and keratin), shown in the right two columns. This difference is also reflected by metrics in Table 1.

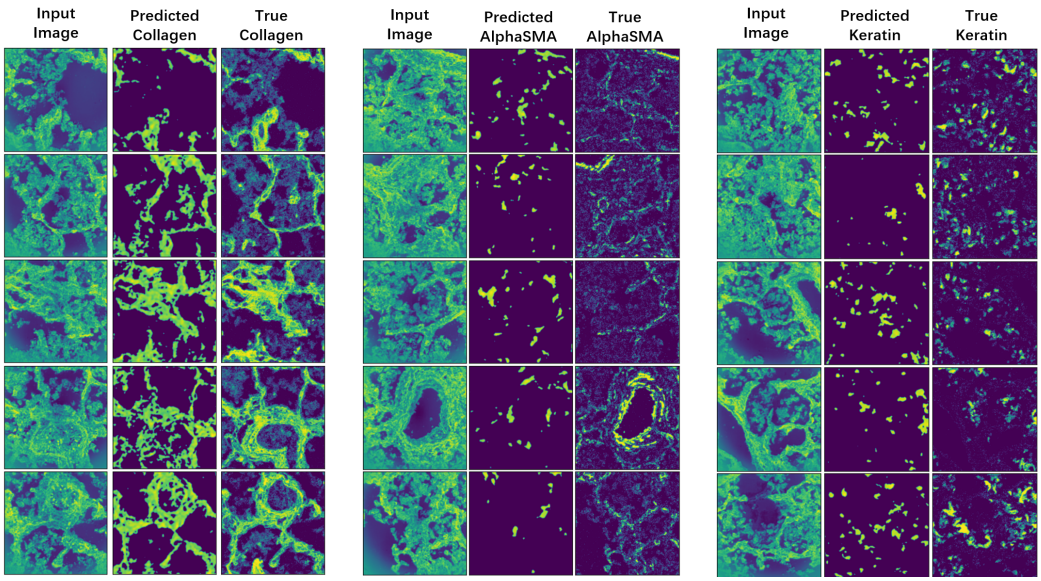


Figure 3: Randomly selected images generated by TransUNet with three different target proteins

Analysis If we compare the true labels of the three proteins, we see that collagen tends to cluster and form large shapes and patterns, while alpha-SMA and keratin tend to scatter across the whole image. This distinction may explain the discrepency in model performance across proteins: since collagen has a connected shape, sub-regions tend to have spatial relationships with each other, which are exactly what TransUNet is good at learning. However, for alpha-SMA and keratin proteins, such clusters are rare. This hypothesis is supported by the fourth alpha-SMA prediction in Figure 3: in the true alpha-SMA, the proteins form an oval-like shape in the middle of the image, and the predicted alpha-SMA partially captures this structure.

	L_1 distance	L_2 distance	IoU
Collagen	0.338	0.155	0.623
Alpha-SMA	0.543	0.337	0.562
Keratin	0.549	0.343	0.439

Table 1: TransUNet performance for three target proteins

4.2 pix2pix

pix2pix is a conditional generative adversarial network designed to learn the translation function between image pairs [15]. In this section, we examine the application of pix2pix to our dataset. Due to the noisy nature of our target distributions, we introduce L_2 loss and design a perceptual loss function to generate accurate translations. We use mean peak signal to noise ratio (PSNR) and mean Frechet inception distance (FID) on holdout samples to measure the quality of images generated by models trained with these different loss landscapes.

Loss landscape of pix2pix The loss function of pix2pix is a combination of an adversarial loss and a reconstruction loss, where the adversarial loss is defined as conditional binary cross entropy and the reconstruction loss as the L_1 loss between the generated and target images. More specifically, the pix2pix loss function is given by [15]:

$$L(G, D) = L_{GAN}(G, D) + \lambda L_{recon}(G) \tag{1}$$

where the terms are defined as

$$L_{GAN}(G, D) = E_{x,y}[logD(x, y)] + E_{x,z}[1 - logD(x, G(x, z))] \tag{2}$$

$$L_{recon}(G) = E_{x,y,z}[||y - G(x,z)||_1] \tag{3}$$

Since the discriminator of pix2pix is designed as a Markovian discriminator which models high frequency structures within the image by classifying patches as real or fake [15], the model relies on L_1 loss to model low frequency structures. However, for target distributions with complex structures such as protein distributions, this objective does not account for all the variability within the signal. In our experiments, we see that with loss function (1) the model generates images that capture structures with the highest frequencies – such as sharp edges and large holes – but generally fails to reconstruct the signal within these regions (Figure 4). This behavior fits our expectations of pix2pix as a model that generates targets like road maps without much fine detail.

Accounting for structural variabilities We hypothesize that the current loss function for pix2pix does not account for complex structural changes in the protein masks. One potential solution is to introduce L_2 loss, which encourages minimzation of pixel-wise differences. To test this, we changed our GAN into a LSGAN by defining the adversarial loss as mean squared error. In our experiment, the model generates blurry images that capture the general structures of protein masks and fill the desired regions with pixels of near uniform intensity (Figure 4). A better way to tackle this problem is to take advantage of the convolutional nature of our generator and look at a group of pixels at the same time, rather than individually as L_2 loss does. We can model the differences between the generated images and the target images with a Gaussian filter to construct a differentiable loss that captures perceptual distances between inputs and targets [16]. Here we choose to use structural similarity between the generated and target image. More specifically, the structural similarity is given by:

$$SSIM(\hat{p}) = \frac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1} \cdot \frac{2\sigma_{xy} + C_2}{\sigma_x^2 + \sigma_y^2 + C_2} \tag{4}$$

where the means and standard deviations are computed with respect to the center pixels $\hat{p}$, with two Gaussian kernels x, y on two images. Here we choose to set the kernel size as 11 by 11. Given (4), we can define structural similarity loss as:

$$L_{SSIM} = 1 - SSIM(\hat{p}) \tag{5}$$

To retain the effect of L_1 loss, we choose to set our new reconstruction loss as:

$$L_{recon}(G) = \alpha L_{SSIM}(G) + (1 - \alpha)L_{L_1}(G) \tag{6}$$

The model trained with this new loss function generates images that capture both high-frequency elements and structurally-complex regions (Figure 4). It also reaches the highest peak noise-to-signal intensity and the lowest Frechet inception distance among the three loss functions explored (Table 2). Therefore, we conclude that the new loss function is well-fit to our task, given targets with abundant signal.

Metrics	GAN loss + L_1	LSGAN loss + L_1	GAN loss + SSIM
FID	13.9745	11.8381	11.5377
PSNR	11.1218	11.3760	12.8364

Table 2: Performance of pix2pix models trained with three different loss landscapes

4.3 pix2pixHD

In order to improve performance on sparser proteins like alpha-SMA and keratin, we trained models on full regions of interest instead of the cropped 304x304 squares. This adjustment to the input space coupled with leveraging the pix2pixHD model for training enabled the model to augment its receptive field with larger spatial areas while maintaining fine-grained accuracy [17].

pix2pixHD is an architecture designed to translate large images without sacrificing small details. As such, it is an appropriate model for learning to translate full, uncropped regions from our dataset. The architecture combines three losses: a GAN loss, which induces the generator to generate images that convincingly belong to the target domain; a feature matching loss, which measures the L_1 distance between discriminator feature maps for a synthesized target vs its corresponding ground truth; and

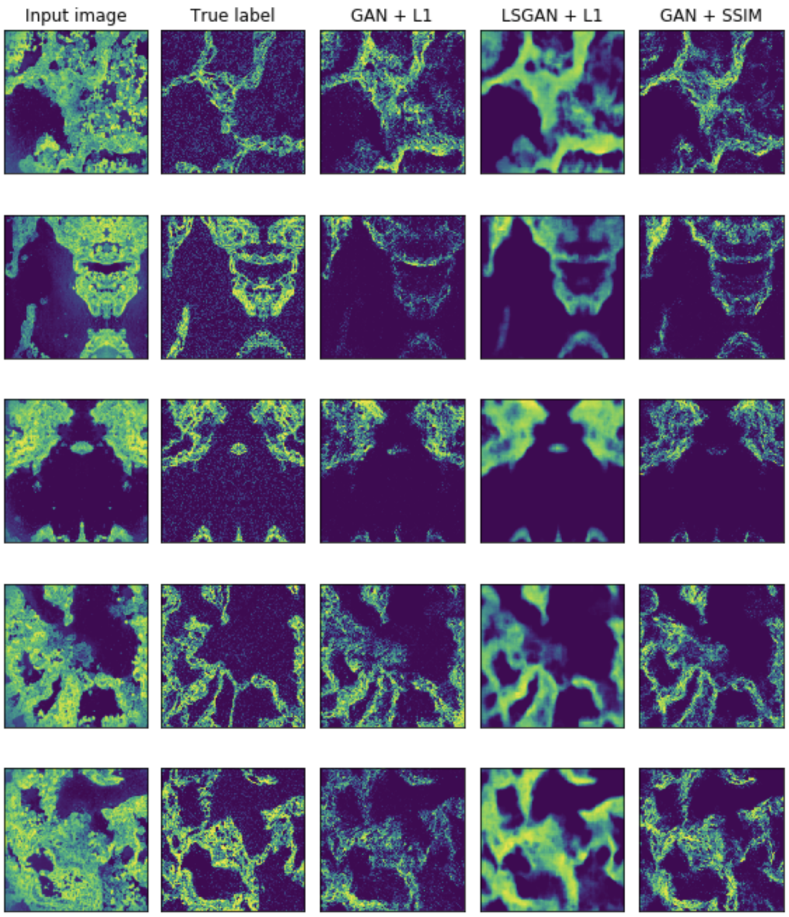


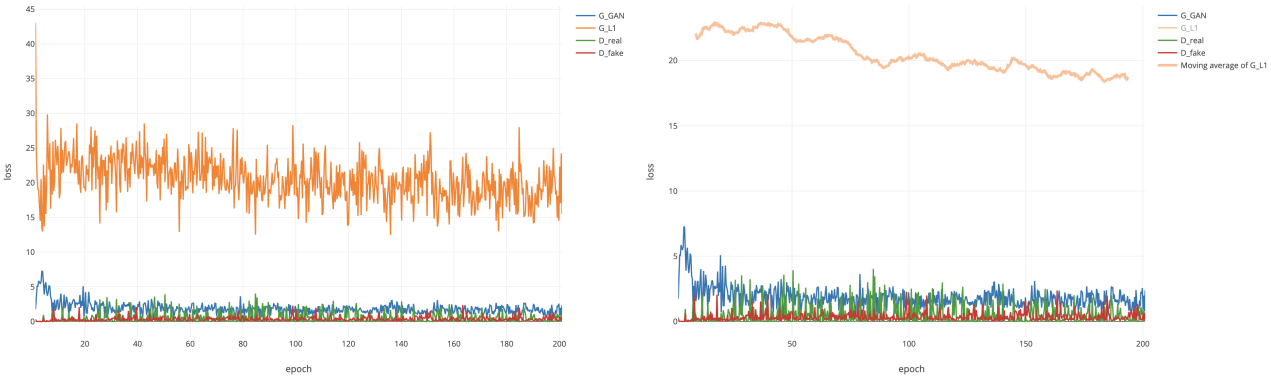
Figure 4: Selected collagen distribution images generated by pix2pix models with three different training objectives. All models are trained with $\lambda = 100$. In addition, $\alpha = 0.9$ was chosen to highlight the effect of the structural similarity-based loss.

a perceptual loss, which measure the L_1 distance between VGG feature maps for a synthesized target vs its corresponding ground truth. Because the VGG network has been trained on real-world, three-color images – a very different domain from our one-color tissue photographs – we exclude the VGG-based perceptual loss when training our model.

Training set performance was monitored over epochs by observing the generator's L_1 loss (Figure 5). While this loss is quite noisy (Figure 5a), by taking a moving average we can see a clear downward trend over time (Figure 5b), reflecting increasingly accurate predictions from our generator. Validation set performance was monitored by reviewing images generated every 10 epochs. We tuned hyperparameters based on validation performance, arriving at a batch size of 8 and a "fine size" – used for crop-based data augmentation by the model – of 608 pixels.

Unimodal prediction results The pix2pixHD model successfully predicted protein distribution on many tissue regions from held-out patients (Figures 6, 7). Due to the high complexity of our protein distribution targets, we determined that no ideal quantitative metric exists to evaluate the performance of our models. We report here performance using L_1 distance, L_2 distance, and intersection over union (IoU) metrics (Table 3).

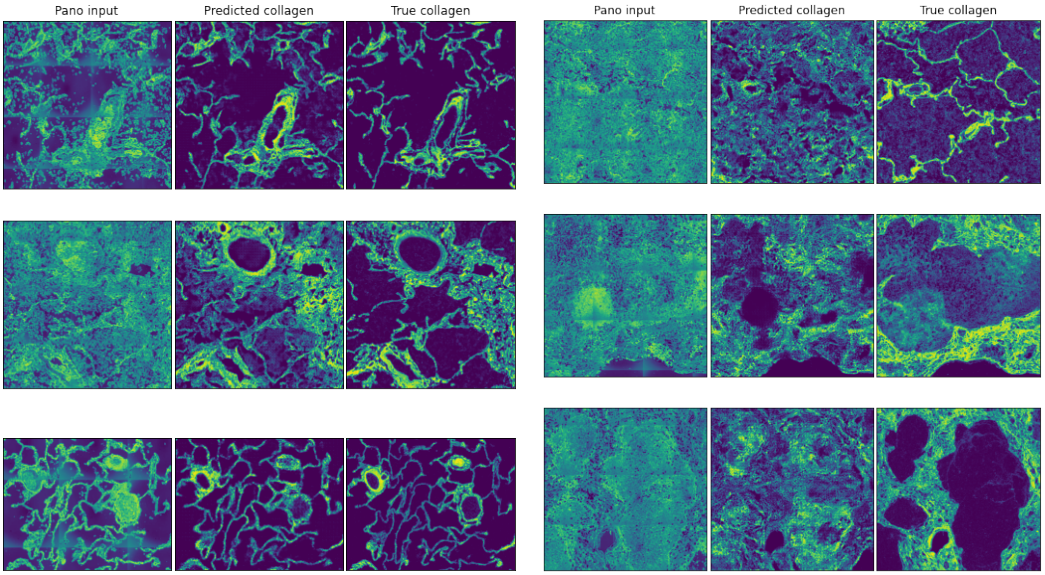
Since the pathology images are large, detailed, and high-frequency, an otherwise-perfect prediction only a few pixels offset from the ground truth in some regions may score poorly on various metrics. However, this prediction would still be highly useful to pathologists or downstream application-specific tissue classifiers. To account for these high frequencies, we smooth the predictions and



(a) Raw generator L_1 loss (orange)

(b) Smoothed generator L_1 loss (orange)

Figure 5: pix2pixHD loss curves



(a) Successful predictions

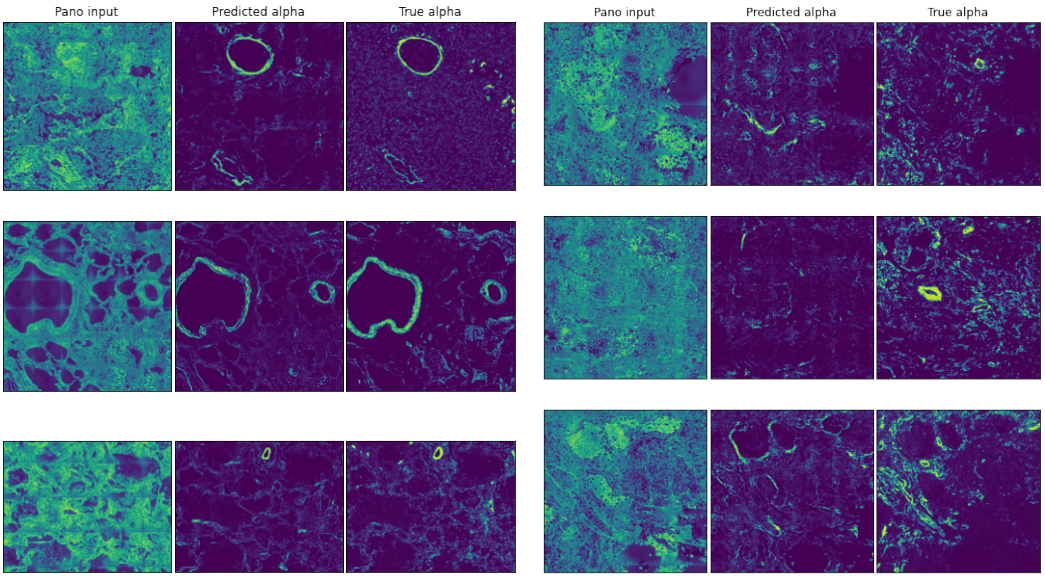
(b) Unsuccessful predictions

Figure 6: Selected pix2pixHD collagen predictions

targets before computing metrics by convolving the images with a uniform filter of kernel size 64 to average each pixel with its neighbors.

For IoU calculations, we converted predictions and targets to categorical data by setting pixels lower in intensity than their image mean to class 0 and the remaining pixels to class 1. By using each image's mean as a threshold for each image, we enable our metric to reflect structural similarity between images, while disregarding general differences in intensity that do not impair the predictions' utility.

Multimodal prediction The distributions of the proteins are highly biologically interrelated. We hypothesized that one model trained to predict several protein channels would generalize better than a model trained to predict just one channel, due to the learning of more robust representations of tissue state. We trained such a model and observed that its performance was slightly inferior to the unimodal models for all three proteins (Table 3). However, this approach may eventually show superior results with more training data.



(a) Successful predictions

(b) Unsuccessful predictions

Figure 7: Selected pix2pixHD alpha-SMA predictions

	Modalities	L_1 distance	L_2 distance	IoU
Collagen	Single	0.092	0.017	0.685
Collagen	Multi	0.089	0.017	0.679
Alpha-SMA	Single	0.072	0.011	0.546
Alpha-SMA	Multi	0.062	0.008	0.544
Keratin	Single	0.076	0.014	0.501
Keratin	Multi	0.069	0.010	0.498

Table 3: Performance of pix2pixHD models

4.4 Comparison of TransUNet, pix2pix, and pix2pixHD

While we explored many quantitative performance metrics for our model, none was superior to human judgment in ranking model results. Consequently, to determine which model was most effective at generating protein distributions, we obtained ratings from an independent human reviewer. The reviewer scored each prediction according to the following scale:

- Score 1.0: The prediction "substantially resembled" the target image
- Score 0.5: The prediction "partially resembled" the target image
- Score 0.0: Otherwise

The mean scores over 20 generated images for each model are shown in Table 4. All models made effective collagen predictions on a majority of tissue regions. TransUNet and pix2pix, however, struggled to predict alpha-SMA – a much sparser protein – effectively. pix2pixHD had similar performance on both proteins.

5 Discussion

We have seen that our model can translate photographs of tissue into protein distribution images. We believe this to be the first model trained on imaging mass cytometry input/output pairs. Our model can be improved by incorporating data from any IMC run, providing us an effectively free, constantly

	TransUNet	pix2pix	pix2pixHD
Collagen	0.60	0.80	0.70
Alpha-SMA	0.05	0.05	0.60

Table 4: Comparison of models by human rating. Scores indicate the degree to which generated images resemble their respective targets.

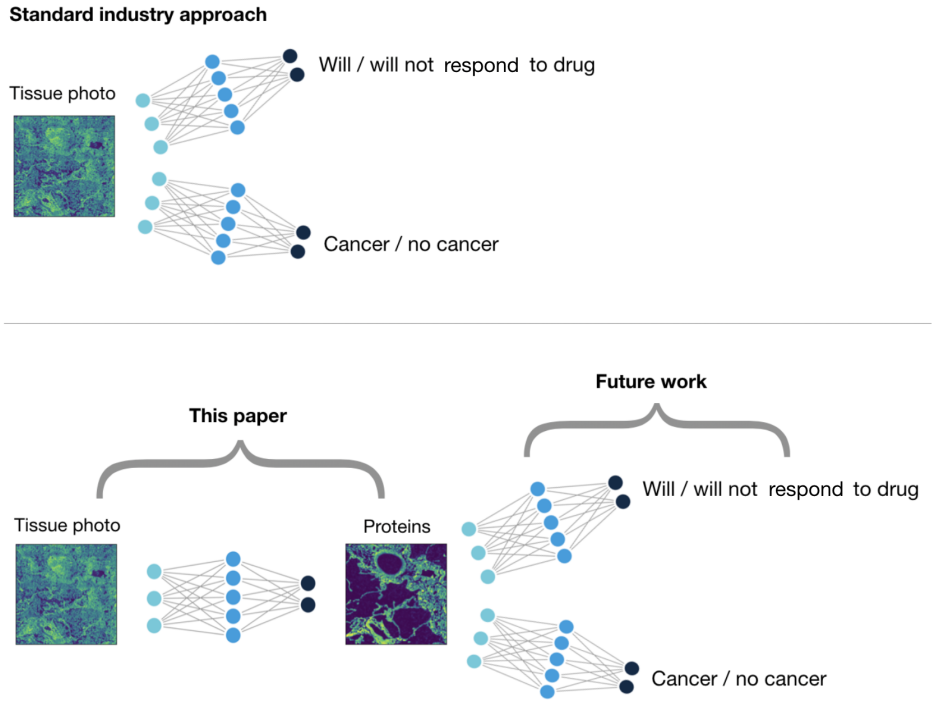


Figure 8: Comparison of the standard approach taken today, work in this paper, and planned work.

growing source of labeled training data (there is no cost to us to leverage IMC data, which is being generated by labs around the world on a regular basis). As a result our model should continue to improve in accuracy over time.

While we have developed here a generic protein prediction model, we plan to apply this model to more specific applications. Our model converts raw photographs into protein distributions. Many models aspire to convert raw photographs into clinical predictions, such as a diagnosis of cancer or an assessment of whether a patient will respond to a certain drug. We believe a two-step model that converts photographs into protein distributions, and then protein distributions into clinical predictions, will perform better than a direct photograph-to-clinical prediction model, by leveraging the rich knowledge of proteins and cell types encoded by our model (Figure 8). Therefore we plan to use our model as a feature extractor that can be trained jointly with various binary neural network classifiers to address important unmet clinical needs.

References

[1] Duggento, Andrea, et al. "Deep computational pathology in breast cancer." Seminars in cancer biology. Academic Press, 2020.

[2] Bulten, Wouter, et al. "Epithelium segmentation using deep learning in H&E-stained prostate specimens with immunohistochemistry as reference standard." Scientific reports 9.1 (2019): 1-10.

[3] Jiang, Jun, et al. "Robust hierarchical density estimation and regression for re-stained histological whole slide image co-registration." Plos one 14.7 (2019): e0220074.

[4] Xu, Zhaoyang, et al. "GAN-based virtual re-staining: a promising solution for whole slide image analysis." arXiv preprint arXiv:1901.04059 (2019).

[5] Masci, Jonathan, Ueli Meier, Dan Cireşan, and Jürgen Schmidhuber. Stacked Convolutional Auto-Encoders for Hierarchical Feature Extraction. *Artificial Neural Networks and Machine Learning – ICANN 2011* 52–59. Springer Berlin Heidelberg, 2011.

[6] Hoo-Chang Shin, Alvin Ihsani, Swetha Mandava, Sharath Turuvekere Sreenivas, Christopher Forster, Jiook Cha, and Alzheimer's Disease Neuroimaging Initiative. GANBERT: Generative Adversarial Networks with Bidirectional Encoder Representations from Transformers for MRI to PET synthesis. *arXiv preprint arXiv:* 2008.04393, 2020.

[7] Sachin Mehta, Ximing Lu, Donald Weaver, Joann G. Elmore, Hannaneh Hajishirzi, and Linda Shapiro. HATNet: An End-to-End Holistic Attention Network for Diagnosis of Breast Biopsy Images. *arXiv preprint arXiv:* 2007.13007, 2020.

[8] Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L. Yuille, and Yuyin Zhou. TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation. *arXiv preprint arXiv:* 2102.04306, 2020.

[9] Ronneberger, Olaf, Philipp Fischer, and Thomas Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015* 234–41. Springer International Publishing, 2015.

[10] Yakubovskiy, Pavel. Segmentation Models Pytorch. *GitHub repository*. GitHub, 2020.

[11] Arvaniti, Eirini. Gleason CNN. *GitHub repository*. GitHub, 2021.

[12] vqdang. Hover Net. *GitHub repository*. GitHub, 2021.

[13] Long, Jonathan, Evan Shelhamer, and Trevor Darrell. 2014. "Fully Convolutional Networks for Semantic Segmentation." *arXiv [cs.CV]. arXiv*. http://arxiv.org/abs/1411.4038.

[14] Rendeiro, André Figueiredo, Hiranmayi Ravichandran, Yaron Bram, Steven Salvatore, Alain Borczuk, Olivier Elemento, and Robert Edward Schwartz. 2020. "The spatio-temporal landscape of lung pathology in SARS-CoV-2 infection." *medrxiv*. https://doi.org/10.1101/2020.10.26.20219584.

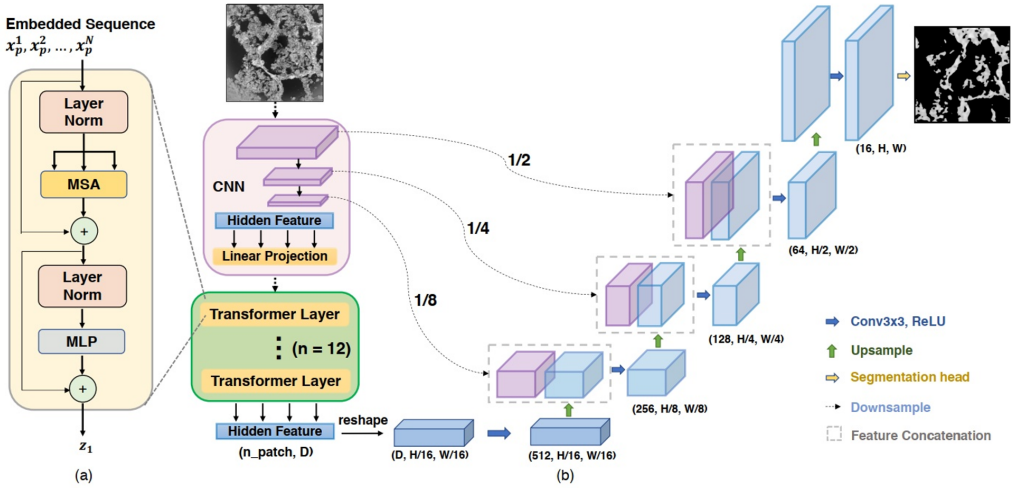
[15] Isola, Phillip, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. 2016. "Image-to-Image Translation with Conditional Adversarial Networks." *arXiv* http://arxiv.org/abs/1611.07004.

[16] Pang, Yingxue, Jianxin Lin, Tao Qin, and Zhibo Chen. 2021. "Image-to-Image Translation: Methods and Applications." *arXiv.* http://arxiv.org/abs/2101.08629.

[17] Wang, Ting-Chun, et al. "High-resolution image synthesis and semantic manipulation with conditional gans." Proceedings of the IEEE conference on computer vision and pattern recognition. 2018.

[18] Zhou, Zongwei, et al. "Unet++: A nested u-net architecture for medical image segmentation." Deep learning in medical image analysis and multimodal learning for clinical decision support. Springer, Cham, 2018. 3-11.

Appendix



Supplemental Figure 1: Architecture of TransUNet [8]