SHIFT: speedy histological-to-immunofluorescent translation of whole slide images enabled by deep learning

Erik A. Burlingame,^{1,2} Mary McDonnell,² Geoffrey F. Schau,^{1,2} Guillaume Thibault,² Christian Lanciault,³ Terry Morgan,³ Brett E. Johnson,² Christopher Corless,^{4,5} Joe W. Gray,^{2,5} Young Hwan Chang^{1,2,6}

KEY POINTS

- Spatially-resolved molecular profiling is an essential complement to histopathological evaluation of cancer tissues.
- Information obtained by immunofluorescence imaging is encoded by features in histological images.
- SHIFT leverages previously unappreciated features in histological images to facilitate virtual immunofluorescence staining.
- Feature representations of images guide sample selection, improving model generalizability.

KEY WORDS

Digital pathology, deep learning, generative adversarial networks, hematoxylin and eosin, immunofluorescence

ABSTRACT

Spatially-resolved molecular profiling by immunostaining tissue sections is a key feature in cancer diagnosis, subtyping, and treatment, where it complements routine histopathological evaluation by clarifying tumor phenotypes. In this work, we present a deep learning-based method called speedy histological-to-immunofluorescent translation (SHIFT) which takes histologic images of hematoxylin and eosin-stained tissue as input, then in near-real time returns inferred virtual immunofluorescence (IF) images that accurately depict the underlying distribution of phenotypes without requiring immunostaining of the tissue being tested. We show that deep learning-extracted feature representations of histological images can guide representative sample selection, which improves SHIFT generalizability. SHIFT could serve as an efficient preliminary, auxiliary, or substitute for IF by delivering multiplexed virtual IF images for a fraction of the cost and in a fraction of the time required by nascent multiplexed imaging technologies.

INTRODUCTION

Physicians depend on histopathology—the visualization and pathological interpretation of tissue biopsies—to diagnose cancer. Hematoxylin and eosin (H&E)-stained histologic sections (3µm-thick formalin-fixed paraffin-embedded tissue biopsies) are the standard of care routinely employed by pathologists to make diagnoses. However, in challenging cases with indeterminate histology, or tumor differentiation, antibody labeling of tumor cells by a molecular imaging technique like immunofluorescence (IF) provides further characterization. It is

¹Computational Biology Program, Department of Biomedical Engineering, Oregon Health & Science University, Portland, OR, USA

²OHSU Center for Spatial Systems Biomedicine, Department of Biomedical Engineering, Oregon Health & Science University, Portland, OR, USA

³Department of Pathology, Oregon Health & Science University, Portland, OR, USA

⁴Knight Diagnostic Laboratories, Oregon Health & Science University, Portland, OR, USA

⁵Knight Cancer Institute, Oregon Health & Science University, Portland, OR, USA

⁶Lead Contact, chanyo@ohsu.edu

becoming increasingly apparent that determining the spatially-resolved molecular profile of a cancer is important for disease subtyping and choosing a patient's course of treatment (Duraiyan et al., 2012). Despite its clinical value, IF is time- and resource-intensive and requires expensive reagents and hardware, so assessment is typically limited to a small representative section of a tumor, which may not be fully representative of the neoplasm, which can be the case in areas of squamous differentiation in an adenocarcinoma (Hester et al., 2018). Also, the cost associated with IF may in some cases limit its use to within highly-developed clinical laboratories, further widening the quality-of-care gap between high- and low-income communities. The gaps between H&E and IF technologies highlights the broader need for automated tools that leverage information attained by a low-cost technology to infer information typically attained by a high-cost technology.

Recent advances in digital pathology and deep learning (DL) have made it possible to automatically extract valuable, human-imperceptible information from H&E-stained histology images (Campanella et al., 2019; Chen et al., 2018; Liu et al., 2017; Rivenson et al., 2019b, 2019a). Apart from histology images, Christiansen et al. and Ounkomol et al. each described supervised DL-based methods for inferring fluorescence images from transmitted light images of unlabeled human or rat cell lines or cell cultures, but not in complex, associated human tissues. Their methods were also based on strictly supervised learning frameworks, which are known to produce incoherent or discontinuous patterns in the virtual stains for some markers, though some of the authors suggest that an adversarial learning framework could address the problem (Christiansen et al., 2018). We previously introduced an adversarial DL-based method called speedy histological-to-IF translation (SHIFT) and demonstrated its ability to infer fluorescence images from images of H&E-stained tissue from a single patient with pancreatic ductal adenocarcinoma (PDAC) (Burlingame et al., 2018). In the current study, we test the generalizability of virtual IF staining by SHIFT through model validation on a limited but morphologically heterogeneous PDAC dataset comprised of samples from multiple patients.

DL models require a large amount of heterogeneous training data to generalize well across the population from which the training data was drawn. Since data limitations are common to many biomedical data domains (Campanella et al., 2019; Costa et al., 2018; Udrea and Mitra, 2017; Xiao et al., 2018), we begin by exploring the possibility that the choice of training samples can be optimized by selecting the few samples that are most representative of the population of samples at our disposal. Some DL-based applications have been proposed for histological image comparison and retrieval (Hegde et al., 2019; Otálora et al., 2018), but to the best of our knowledge none have been proposed for the express purpose of training set selection in a data-limited biomedical imaging domain. We describe the use of a data-driven DL-based method to select samples that optimizes the morphological heterogeneity of the dataset and promotes SHIFT model generalizability. As a proof of concept, we objectively measure the ability of SHIFT models to infer the spatial distribution of a pan-cytokeratin (panCK) antibody which marks cancer cells. We also show preliminary results for inference on the

stromal marker α -smooth muscle actin (α -SMA), a first step toward the development of a generalized platform for multiplexed virtual IF imaging in human tissues. The advantages of a validated SHIFT model over traditional IF include (1) eliminating the need for expensive imaging hardware, bulk reagents, and technical undertaking of IF protocols; (2) virtual IF images can be generated in near-real time; and (3) the portability of SHIFT allows it to be integrated into existing imaging workflows with minimal effort.

RESULTS

Building a Dataset of Spatially-registered H&E and IF Images

SHIFT requires spatially-registered pairs of H&E and IF whole slide images (WSIs) for model training and testing (**Figure 1A**). Such data would usually be acquired by processing two adjacent tissue sections, one stained by H&E and another stained by IF, then spatially registering the images into the same coordinate system based on their shared features (Chang et al., 2017). Unfortunately, this can lead to inconsistencies between H&E and IF image contents when high-frequency cellular features differ between adjacent sections, even when the sections are as few as 5 μ m apart. To alleviate this issue, we developed a protocol that allows for H&E and IF staining in the same section of tissue (see **METHODS**). Clinical samples of PDAC from four patients (Samples A, B, C and D) were chosen via pathological review of archival H&E images as exemplifying a spectrum of both histological differentiation and heterogeneity (**Figure 1B**). Chosen samples were sectioned, processed, and stained with DAPI nuclear stain and panCK monoclonal antibody; the staining was confirmed and the slides were scanned. After scanning, the coverslips were removed and the slides were stained with the designed modified H&E protocol, permanently cover slipped and then scanned again. Nuclear information from the hematoxylin and DAPI stains in pairs of H&E and IF images were used to register images in a common coordinate system. Images were then pre-processed to minimize noise and account for technical variability in staining and image acquisition. To exclude regions of autofluorescence that greatly diminished the signal-to-noise ratio of the real IF images, images from samples B and D were subdivided into image subsets {B1, B2, B3} and {D1, D2, D3, D4, D5}.

Feature-guided Identification of Representative Histological Samples

For a SHIFT model to generalize well across the population of PDAC samples, it must be trained on a representative subset of the population, which motivated the development of a means to quantitatively compare images (see **METHODS**). In particular, we wished to learn which sample—or sequence of samples—should be selected to build a training set that is most representative of the population of samples. As a consequence of their large dimensions, direct comparison between gigapixel H&E images is intractable, so we decomposed each image into sets of non-overlapping 256 $\times$ 256 pixel tiles (**Figure 1C**). Even the small 256 $\times$ 256 pixel H&E tiles contain 196,608 (256 $\times$ 256 $\times$ 3 channels = 196,608) pixel values each and are difficult to compare directly. To establish a more compact but still expressive representation of the H&E tiles, we trained a variational autoencoder (VAE) (Kingma and Welling, 2013)—an unsupervised DL-based method for

representation learning and feature extraction--to learn 16-dimensional feature representations of each tile (see **METHODS**), which makes comparing tiles more tractable (**Figure 1D**). Using the dimensionality reduction algorithm *t*-SNE (van der Maaten and Hinton, 2008), we embed all the H&E tiles into a 2D map to visualize the learned feature space over which the H&E tiles are distributed, which provides a visual aid to our selection of the most representative set of samples (**Figures 1D**).

Using the VAE which we pre-trained on all samples, we extracted features from each H&E tile (**Table S1**) and assessed how each feature was distributed across samples, finding that several of the features discriminated between samples (**Figure 2A**). In particular, the bimodal distribution of some features suggested two sample clusters, one formed by samples A and B, and another formed by samples C and D. These clusters were corroborated by visualization of the sample tiles embedded in the reduced-dimension feature space generated by *t*-SNE (**Figure 2B** and **Table S1**). Since the feature distributions of the H&E tiles highlighted the redundancy between clustered samples, we reasoned that a balanced selection of samples from each cluster would yield a more representative training set and ultimately improve SHIFT model generalizability. Using the full 16-dimensional feature representations of the H&E tiles and an information-theoretic framework for representative sample selection (Feng Pan et al., 2005; see **METHODS**), we were able to quantitatively identify sample B and the duo of samples B and D as the single and two most representative samples, respectively, and were then considered for training sets in subsequent experiments. **Figure S1** illustrates the feature distributions of several sample combinations in comparison to that of the full dataset.

Virtual IF Staining in Histological Samples

SHIFT models are built on an adversarial image-to-image translation framework (Isola et al., 2016), with a regularization strategy designed to improve inference on sparse IF images (**Figures 1A and S2**; see **METHODS**). Adversarial learning frameworks compute their losses over images, in contrast to strictly supervised learning frameworks where losses are computed over pixels, which is likely to improve model inference in many virtual staining tasks (Christiansen et al., 2018). Having identified the most representative samples in our dataset, we next tested whether or not a SHIFT model could learn a correspondence between H&E and IF images that generalizes across samples.

We hypothesized that if tissue and cell morphologies observed in H&E-stained tissue are a function of a given marker, then it should be possible to infer the spatial distribution of that marker based on the H&E-stained tissue alone; that is, H&E-to-IF translation should be learnable and generalizable such that a model can be extended to samples from patients that were not included in the training set. To test this hypothesis, we trained SHIFT models to generate virtual IF images of the cancer marker panCK conditioned on input H&E images alone (**Figure 1A**). To simultaneously assess the utility of our sample selection method, we trained models using different combinations of sample subsets in the training set. Training sets consisted of paired H&E and IF

image tiles from either sample subset B1 (from the most representative single sample), sample subsets B1 and D5 (from the most representative duo of samples), or sample subsets A1 and B1 (from a less representative duo of samples as counterexample). Sample subsets B1 and D5 were selected because they contained a similar number of tiles, providing a balance between the sample clusters. Model performance was quantified by measuring the structural similarity (SSIM) (Rivenson et al., 2019b; Wang et al., 2004), a widely used measure of image similarity as perceived by the human visual system, between corresponding virtual and real IF images from samples left out of the model's training set. The SSIM between two images is calculated over pixel neighborhoods in the images and provides a more coherent measure of image similarity than pixel-wise measures like Pearson's r . Considering the performance of the model trained on B1 as the baseline, we see a marked improvement in model generalizability across samples when the more representative sample subsets B1 and D5 are used for training than when the less representative sample subsets A1 and B1 are used (**Figure 2C**). By stitching together the virtual IF tiles from a given sample in the test set, we were able to make large-scale comparisons between real and virtual panCK IF images (**Figure S3**). We also experimented with SHIFT inference of the stromal marker α -SMA (**Figure S4**).

SHIFT is not the only virtual staining method to have been recently proposed. Label-Free Determination (LFD) (Ounkomol et al., 2018) is a supervised DL-based virtual staining method which produces models that were shown to have learned the relationship between images of cell cultures visualized by transmitted light or fluorescence, where sub-cellular structures have been labeled with genetically-encoded fluorescent tags. Because the SHIFT generator G and the LFD are both based on the popular U-net architecture (Ronneberger et al., 2015), we compared these models that generate images using a similar architecture, but have differing training formulae and loss functions (see **METHODS**). To make a fair comparison between the adversarial SHIFT and supervised LFD models, we trained a LFD using the representative sample subsets B1 and D5, matching the number of optimization steps taken by the SHIFT model that was trained using the same training set (**Figure S5**). In addition to the performance of independent SHIFT and LFD models, we also considered the ensemble result, taken as the average image of the SHIFT and LFD output images. Across all samples in the test set, either SHIFT alone or the ensemble of SHIFT and LFD tended to perform better than LFD alone (**Figure 2C**).

DISCUSSION

Spatially-resolved molecular profiling of cancer tissues by technologies like IF provides more information than routine H&E histology alone. However, the rich information obtained from IF comes at significant expense in time and resources, restricting IF access and use. Here, we present and extend the validation of SHIFT, a DL-based method which takes standard H&E-stained histology images as input and returns virtual IF images of inferred biomarker distributions. Using a limited but heterogeneous dataset, we demonstrated that SHIFT models are able to generalize across samples drawn from different PDAC patients, even for training sets that

are over an order of magnitude smaller than the test set (train $n = 665$ and test $n = 9693$ for models trained on sample subset B1 only). Results from our sampling experiments are consistent with the expectation that an automated and quantitative method for representative sample selection will be critical to the effective development and deployment of DL models on large-scale digital pathology datasets. Finally, we compared the adversarial SHIFT method with an alternative, supervised virtual staining method and found that the virtual staining task tends to be best accomplished by the ensemble of both methods. Based on the success of DL-based ensemble methods in other biomedical domains (Codella et al., 2017; Xiao et al., 2018), we expect ensemble methods to become increasingly relevant to the development of virtual staining applications.

While we have demonstrated the utility of SHIFT for only a limited number of virtual stains, there are emerging opportunities to test the extensibility of our approach. Following recent advances in multiplexed immunofluorescence and immunohistochemistry (mIF/IHC) technology (Goltsev et al., 2018; Lin et al., 2016; Reiß et al., 2019; Tsujikawa et al., 2017), it is now possible to visualize tens or hundreds of distinct biomarkers in a single tissue section. On their own, these technologies promise a more personalized medicine through a more granular definition of disease subtypes, and will undoubtedly broaden our understanding of cellular heterogeneity and interaction within the tumor microenvironment, both of which play increasingly important roles in the development and selection of effective treatments (Lu et al., 2019; Yuan, 2016). With a paired H&E and mIF/IHC dataset that encompasses the expression of hundreds of markers within serial tissue sections, we could begin to quantify the mutual information between histology and expression of any biomarker of interest.

Though our results suggest broad applicability, there are limitations and challenges to both the feature-guided sampling and virtual staining methods we present here. Both methods assume a relationship between H&E and IF representations of tissue, so they may fail to maximize representativeness and make incorrect inference in the IF domain if there is no association between histology and a given biomarker. Even when there is an association, finding a meaningful way to compare real and virtual images remains a challenge, as we have experienced in experiments modeling biomarkers with fine, high-frequency distributions like α -SMA (**Figure S4**). Like other virtual staining methods that have been deployed on whole human tissues (Rivenson et al., 2019b, 2019a), we used SSIM as a measure of image similarity between real and virtual IF images. This classical perceptual measure is used in many imaging domains, but we found that it is sensitive to perturbations commonly associated with image registration and technical or instrumentation noise (**Figure S6**). In light of this, we advocate for the development and use of perceptual measures that are more aware of such perturbations and better correlate with human perception of image similarity or quality (Azulay and Weiss, 2018; Patel et al., 2019). It must be restated that our results are supported by a dataset comprised of samples from just four patients, so some fluctuation in performance between samples should be expected, and indeed was observed (**Figure S3C**). Our goal in choosing a relatively small dataset was to demonstrate that, even

when limited, SHIFT could learn a general relationship between H&E and IF tissue representations, and we believe that the fluctuation in performance between samples could be addressed by increasing the sample size. With the emergence of digital pathology datasets containing tens of thousands of whole slide images (Campanella et al., 2019), the opportunities to improve virtual staining technologies are only becoming more numerous.

Since SHIFT can infer virtual IF images as H&E-stained tissue section are imaged, SHIFT could provide pathologists with near-real-time interpretations based on standard H&E-stained tissue. Therefore, SHIFT could serve as an efficient preliminary, auxiliary, or substitute technology for traditional IF in both research and clinical settings by delivering comparable virtual IF images for a fraction of the cost and in a fraction of the time required by traditional IF or mIF/IHC imaging. As such, we see SHIFT as an opportunity to simultaneously economize and democratize advanced imaging technologies in histopathology workflows, with implications for multiplexed virtual imaging. Further, we see our method of optimal selection of representative histological images, which promotes morphological heterogeneity in the training set, as a complement to data augmentation, transfer learning, and other means of addressing the problem of limited training data. Moreover, this method will contribute to saving resources and minimizing unnecessary efforts to acquire additional staining or manual annotation for DL applications in biomedical imaging.

AUTHOR CONTRIBUTIONS

Conceptualization, E.A.B. and Y.H.C.; Methodology, E.A.B, M.M., B.E.J., and Y.H.C.; Software, E.A.B., G.F.S., G.T., and Y.H.C.; Validation, E.A.B., C.L., T.M., and C.C.; Formal Analysis, E.A.B., C.L., T.M., and Y.H.C.; Investigation, E.A.B., M.M., C.L., T.M., B.E.J., C.C., and Y.H.C.; Resources, E.A.B., B.E.J., J.W.G., and Y.H.C.; Data Curation, E.A.B., M.M., and Y.H.C.; Writing--Original Draft, E.A.B., M.M., G.F.S., T.M., and Y.H.C. ; Writing--Review & Editing, E.A.B., M.M., G.F.S., G.T., T.M., B.E.J., C.C., J.W.G. and Y.H.C.; Visualization, E.A.B and M.M.; Supervision, B.E.J., C.C., J.W.G. and Y.H.C.; Project Administration, E.A.B., M.M., B.E.J., and Y.H.C.; Funding Acquisition, E.A.B., G.F.S., Y.H.C., B.E.J., and J.W.G.

ACKNOWLEDGMENTS

This work was supported in part by the National Cancer Institute (U54CA209988, U2CCA233280), the OHSU Center for Spatial Systems Biomedicine, the Brenden-Colson Center for Pancreatic Care, and a Biomedical Innovation Program Award from the Oregon Clinical & Translational Research Institute. We acknowledge expert technical assistance by staff in the Advanced Multiscale Microscopy Shared Resource, supported by the OHSU Knight Cancer Institute (NIH P30 CA069533) and the Office of the Senior Vice President for Research. Equipment purchases included support by the OHSU Center for Spatial Systems Biomedicine, the MJ Murdock Charitable Trust, and the Collins Foundation. We acknowledge the Histopathology Shared Resource for pathology studies, which is supported in part by the University Shared Resource Program at Oregon Health

and Sciences University and the Knight Cancer Institute (P30 CA069533, and P30 CA069533 13S5). The resources of the Exacloud high performance computing environment developed jointly by OHSU and Intel and the technical support of the OHSU Advanced Computing Center are gratefully acknowledged. E.A.B. receives support from a scholar award provided by the ARCS Foundation Oregon. B.E.J. was supported by a Postdoctoral Fellowship, PF-17-031-01-DDC, from the American Cancer Society.

DECLARATION OF INTERESTS

The authors declare no competing interests.

REFERENCES

- Azulay, A., Weiss, Y., 2018. Why do deep convolutional networks generalize so poorly to small image transformations? ArXiv180512177 Cs.
- Bay, H., Tuytelaars, T., Van Gool, L., 2006. SURF: Speeded Up Robust Features, in: Leonardis, A., Bischof, H., Pinz, A. (Eds.), Computer Vision – ECCV 2006. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 404–417. https://doi.org/10.1007/11744023_32
- Bokeh Development Team, 2019. Bokeh: Python library for interactive visualization.
- Bousmalis, K., Silberman, N., Dohan, D., Erhan, D., Krishnan, D., 2016. Unsupervised Pixel-Level Domain Adaptation with Generative Adversarial Networks. ArXiv161205424 Cs.
- Burlingame, E.A., Margolin, A.A., Gray, J.W., Chang, Y.H., 2018. SHIFT: speedy histopathological-to-immunofluorescent translation of whole slide images using conditional generative adversarial networks. Proc. SPIE-- Int. Soc. Opt. Eng. 10581. <https://doi.org/10.1117/12.2293249>
- Campanella, G., Hanna, M.G., Geneslaw, L., Mirafior, A., Silva, V.W.K., Busam, K.J., Brogi, E., Reuter, V.E., Klimstra, D.S., Fuchs, T.J., 2019. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. Nat. Med. 1. <https://doi.org/10.1038/s41591-019-0508-1>
- Chang, Y.H., Thibault, G., Madin, O., Azimi, V., Meyers, C., Johnson, B., Link, J., Margolin, A., Gray, J.W., 2017. Deep learning based Nucleus Classification in pancreas histological images, in: 2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). Presented at the 2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 672–675. <https://doi.org/10.1109/EMBC.2017.8036914>
- Chen, P.-H.C., Gadepalli, K., MacDonald, R., Liu, Y., Nagpal, K., Kohlberger, T., Dean, J., Corrado, G.S., Hipp, J.D., Stumpe, M.C., 2018. Microscope 2.0: An Augmented Reality Microscope with Real-time Artificial Intelligence Integration. ArXiv181200825 Cs.
- Christiansen, E.M., Yang, S.J., Ando, D.M., Javaherian, A., Skibinski, G., Lipnick, S., Mount, E., O'Neil, A., Shah, K., Lee, A.K., Goyal, P., Fedus, W., Poplin, R., Esteva, A., Berndl, M., Rubin, L.L., Nelson, P., Finkbeiner, S., 2018. In Silico Labeling: Predicting Fluorescent Labels in Unlabeled Images. Cell. <https://doi.org/10.1016/j.cell.2018.03.040>
- Codella, N.C.F., Nguyen, Q., Pankanti, S., Gutman, D.A., Helba, B., Halpern, A.C., Smith, J.R., 2017. Deep learning ensembles for melanoma recognition in dermoscopy images. IBM J. Res. Dev. 61, 5:1-5:15. <https://doi.org/10.1147/JRD.2017.2708299>
- Costa, P., Galdran, A., Meyer, M.I., Niemeijer, M., Abràmoff, M., Mendonça, A.M., Campilho, A., 2018. End-to-End Adversarial Retinal Image Synthesis. IEEE Trans. Med. Imaging 37, 781–791. <https://doi.org/10.1109/TMI.2017.2759102>
- Duraiyan, J., Govindarajan, R., Kaliyappan, K., Palanisamy, M., 2012. Applications of immunohistochemistry. J. Pharm. Bioallied Sci. 4, S307–S309. <https://doi.org/10.4103/0975-7406.100281>
- Feng Pan, Wei Wang, Tung, A.K.H., Jiong Yang, 2005. Finding Representative Set from Massive Data, in: Fifth IEEE International Conference on Data Mining (ICDM'05). Presented at the Fifth IEEE International Conference on Data Mining (ICDM'05), IEEE, Houston, TX, USA, pp. 338–345. <https://doi.org/10.1109/ICDM.2005.69>
- Goltsev, Y., Samusik, N., Kennedy-Darling, J., Bhate, S., Hale, M., Vazquez, G., Black, S., Nolan, G.P., 2018.

- Deep Profiling of Mouse Splenic Architecture with CODEX Multiplexed Imaging. *Cell* 174, 968-981.e15. <https://doi.org/10.1016/j.cell.2018.07.010>
- Goodfellow, I., Pouget-Abadie, J., Mirza, M., 2014. Generative Adversarial Networks. *ArXiv Prepr. ArXiv* ... 1–9.
- Hegde, N., Hipp, J.D., Liu, Y., Emmert-Buck, M., Reif, E., Smilkov, D., Terry, M., Cai, C.J., Amin, M.B., Mermel, C.H., Nelson, P.Q., Peng, L.H., Corrado, G.S., Stumpe, M.C., 2019. Similar image search for histopathology: SMILY. *Npj Digit. Med.* 2, 56. <https://doi.org/10.1038/s41746-019-0131-z>
- Hester, C.A., Augustine, M.M., Choti, M.A., Mansour, J.C., Minter, R.M., Polanco, P.M., Porembka, M.R., Wang, S.C., Yopp, A.C., 2018. Comparative outcomes of adenocarcinoma of the pancreas: An analysis of the National Cancer Database. *J. Surg. Oncol.* 118, 21–30. <https://doi.org/10.1002/jso.25112>
- Isola, P., Zhu, J.-Y., Zhou, T., Efros, A.A., 2016. Image-to-Image Translation with Conditional Adversarial Networks. *ArXiv161107004 Cs*.
- Jones, E., Oliphant, T., Peterson, P., others, 2001. SciPy: Open source scientific tools for Python.
- Jung, Alexander B., 2018. *imgaug*.
- Kingma, D.P., Welling, M., 2013. Auto-Encoding Variational Bayes. *ArXiv13126114 Cs Stat*.
- Kluyver, T., Ragan-Kelley, B., Pérez, F., Granger, B., Bussonnier, M., Frederic, J., Kelley, K., Hamrick, J., Grout, J., Corlay, S., Ivanov, P., Avila, D., Abdalla, S., Willing, C., 2016. Jupyter Notebooks – a publishing format for reproducible computational workflows, in: Loizides, F., Schmidt, B. (Eds.), *Positioning and Power in Academic Publishing: Players, Agents and Agendas*. IOS Press, pp. 87–90.
- Lin, J.-R., Fallahi-Sichani, M., Chen, J.-Y., Sorger, P.K., 2016. Cyclic Immunofluorescence (CycIF), A Highly Multiplexed Method for Single-cell Imaging. *Curr. Protoc. Chem. Biol.* 8, 251–264. <https://doi.org/10.1002/cpch.14>
- Liu, Y., Gadepalli, K., Norouzi, M., Dahl, G.E., Kohlberger, T., Boyko, A., Venugopalan, S., Timofeev, A., Nelson, P.Q., Corrado, G.S., Hipp, J.D., Peng, L., Stumpe, M.C., 2017. Detecting Cancer Metastases on Gigapixel Pathology Images. *ArXiv170302442 Cs*.
- Lu, S., Stein, J.E., Rimm, D.L., Wang, D.W., Bell, J.M., Johnson, D.B., Sosman, J.A., Schalper, K.A., Anders, R.A., Wang, H., Hoyt, C., Pardoll, D.M., Danilova, L., Taube, J.M., 2019. Comparison of Biomarker Modalities for Predicting Response to PD-1/PD-L1 Checkpoint Blockade: A Systematic Review and Meta-analysis. *JAMA Oncol.* <https://doi.org/10.1001/jamaoncol.2019.1549>
- Macenko, M., Niethammer, M., Marron, J.S., Borland, D., Woosley, J.T., Guan, X., Schmitt, C., Thomas, N.E., 2009. A method for normalizing histology slides for quantitative analysis. *Proc. - 2009 IEEE Int. Symp. Biomed. Imaging Nano Macro ISBI 2009* 1107–1110. <https://doi.org/10.1109/ISBI.2009.5193250>
- Mirza, M., Osindero, S., 2014. Conditional Generative Adversarial Nets. *ArXiv14111784 Cs Stat*.
- Otálora, S., Schaer, R., Jimenez-del-Toro, O., Atzori, M., Müller, H., 2018. Deep Learning Based Retrieval System for Gigapixel Histopathology Cases and Open Access Literature. *bioRxiv* 408237. <https://doi.org/10.1101/408237>
- Ounkomol, C., Seshamani, S., Maleckar, M.M., Collman, F., Johnson, G.R., 2018. Label-free prediction of three-dimensional fluorescence images from transmitted-light microscopy. *Nat. Methods* 15, 917–920. <https://doi.org/10.1038/s41592-018-0111-2>
- Ouyang, W., Aristov, A., Lelek, M., Hao, X., Zimmer, C., 2018. Deep learning massively accelerates super-resolution localization microscopy. *Nat. Biotechnol.* 36, 460–468. <https://doi.org/10.1038/nbt.4106>
- Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., Lerer, A., 2017. Automatic differentiation in PyTorch. *NIPS Autodiff Workshop*.
- Patel, Y., Appalaraju, S., Manmatha, R., 2019. Deep Perceptual Compression. *ArXiv190708310 Cs Eess*.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, É., 2011. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* 12, 2825–2830.
- Reiß, S., Tomiuk, S., Kollet, J., Drewes, J., Brück, W., Jungblut, M., Bosio, A., 2019. Abstract 245: Characterization and classification of glioblastoma multiforme using the novel multiparametric cyclic immunofluorescence analysis system MACSima. *Cancer Res.* 79, 245–245. <https://doi.org/10.1158/1538-7445.SABCS18-245>
- Rivenson, Y., Liu, T., Wei, Z., Zhang, Y., Haan, K. de, Ozcan, A., 2019a. PhaseStain: the digital staining of

- label-free quantitative phase microscopy images using deep learning. *Light Sci. Appl.* 8, 23. <https://doi.org/10.1038/s41377-019-0129-y>
- Rivenson, Y., Wang, H., Wei, Z., Haan, K. de, Zhang, Y., Wu, Y., Günaydin, H., Zuckerman, J.E., Chong, T., Sisk, A.E., Westbrook, L.M., Wallace, W.D., Ozcan, A., 2019b. Virtual histological staining of unlabelled tissue-autofluorescence images via deep learning. *Nat. Biomed. Eng.* 1. <https://doi.org/10.1038/s41551-019-0362-y>
- Ronneberger, O., Fischer, P., Brox, T., 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. *ArXiv150504597 Cs*.
- Stevens, J.-L., Rudiger, P., Bednar, J., 2015. HoloViews: Building Complex Visualizations Easily for Reproducible Science. Presented at the Python in Science Conference, Austin, Texas, pp. 59–66. <https://doi.org/10.25080/Majora-7b98e3ed-00a>
- Tsujikawa, T., Kumar, S., Borkar, R.N., Azimi, V., Thibault, G., Chang, Y.H., Balter, A., Kawashima, R., Choe, G., Sauer, D., El Rassi, E., Clayburgh, D.R., Kulesz-Martin, M.F., Lutz, E.R., Zheng, L., Jaffee, E.M., Leyshock, P., Margolin, A.A., Mori, M., Gray, J.W., Flint, P.W., Coussens, L.M., 2017. Quantitative Multiplex Immunohistochemistry Reveals Myeloid-Inflamed Tumor-Immune Complexity Associated with Poor Prognosis. *Cell Rep.* 19, 203–217. <https://doi.org/10.1016/j.celrep.2017.03.037>
- Udrea, A., Mitra, G.D., 2017. Generative Adversarial Neural Networks for Pigmented and Non-Pigmented Skin Lesions Detection in Clinical Images, in: 2017 21st International Conference on Control Systems and Computer Science (CSCS). Presented at the 2017 21st International Conference on Control Systems and Computer Science (CSCS), pp. 364–368. <https://doi.org/10.1109/CSCS.2017.56>
- van der Maaten, L., Hinton, G., 2008. Visualizing Data using t-SNE. *J. Mach. Learn. Res.* 9, 2579–2605.
- van der Walt, S., Schönberger, J.L., Nunez-Iglesias, J., Boulogne, F., Warner, J.D., Yager, N., Gouillart, E., Yu, T., 2014. scikit-image: image processing in Python. *PeerJ* 2, e453. <https://doi.org/10.7717/peerj.453>
- Wang, Z., Bovik, A.C., Rahim Sheikh, H., Simoncelli, E.P., 2004. Image Quality Assessment: From Error Visibility to Structural Similarity. *IEEE Trans. IMAGE Process.* 13. <https://doi.org/10.1109/TIP.2003.819861>
- Xiao, Y., Wu, J., Lin, Z., Zhao, X., 2018. A deep learning-based multi-model ensemble method for cancer prediction. *Comput. Methods Programs Biomed.* 153, 1–9. <https://doi.org/10.1016/j.cmpb.2017.09.005>
- Yuan, Y., 2016. Spatial Heterogeneity in the Tumor Microenvironment. *Cold Spring Harb. Perspect. Med.* 6, a026583. <https://doi.org/10.1101/cshperspect.a026583>
- Zhou Wang, Bovik, A.C., Sheikh, H.R., Simoncelli, E.P., 2004. Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process.* 13, 600–612. <https://doi.org/10.1109/TIP.2003.819861>

Figure 1. Overview of virtual IF staining with SHIFT and feature-guided H&E sample selection.

(A) Schematic of SHIFT modeling for training and testing phases. The generator network G generates virtual IF tiles conditioned on H&E tiles. The discriminator network D learns to discriminate between real and generated image pairs. See also **Figure S2**.

(B) Four heterogeneous samples of H&E-stained PDAC biopsy tissue used in the current study. Pathologist annotations indicate regions that are benign (green), grade 1 PDAC (black), grade 2/3 PDAC (blue), and grade 2/3 adenosquamous (red).

(C) Making direct comparisons between H&E whole slide images (WSIs) is intractable because each WSI can contain billions of pixels. By decomposing WSIs into sets of non-overlapping 256×256 pixel tiles, we can make tractable comparisons between the feature-wise distribution of tile sets.

(D) Schematic of feature-guided H&E sample selection. First, H&E samples are decomposed into 256×256 pixel tiles. Second, all H&E tiles are used to train a variational autoencoder (VAE) to learn feature representations for all tiles; for each 196,608-pixel H&E tile in the dataset, the encoder $\mathcal{E}$ learns a compact but expressive feature representation that maximizes the ability of the decoder $\mathcal{D}$ to reconstruct the original tile from its feature representation (see **METHODS**). Third, the tile feature representations are used to determine which samples are most representative of the whole dataset.

Figure 2. Feature-guided H&E sample selection and virtual IF staining with SHIFT.

(A) Distribution of the 16 latent features (L1-L16) extracted by VAE from sample H&E tiles.

(B) *t*-SNE embedding of latent feature representations of sample H&E tiles, faceted by sample identity. Each point in each plot represents a single H&E tile. Contour lines indicate point density.

(C) SHIFT model test performance for optimal and non-optimal training set sample compositions. The paired H&E and IF images from samples B and D were subdivided into smaller images $B = \{B1, B2\}$ and $D = \{D1, D2, D3, D4, D5\}$ to avoid regions of IF that exhibited substantial autofluorescence. The x-axis labels indicate sample identity, where each letter corresponds to a unique sample and each number corresponds to a subset of that sample. Each n denotes the number of image tiles that were extracted from that sample. Plots for sample subsets are not shown if that sample subset was a component of a model's training set. * $p < .05$; for three group comparisons we used the Friedman test with Nemenyi post-hoc test; for two group comparisons we used the Wilcoxon signed-rank test. White dots in violin plots represent distributional medians.

(D) Visual comparison of virtual staining methods. The ensemble results are attained by averaging the output images of SHIFT and Label-Free Determination models. See also **Figure S2**.

(E) Test performance comparison of virtual staining methods. The x-axis labels indicate sample identity, where each letter corresponds to a unique sample and each number corresponds to a subset of that sample. Each n denotes the number of image tiles that were extracted from that sample. Plots for sample subsets B1 and D5 are not shown because those sample subsets were components of the models' training sets. * $p < .05$; Friedman test with Nemenyi post-hoc test. White dots in violin plots represent distributional medians.

METHODS

Resources

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Antibodies		
Alpha-smooth muscle actin monoclonal antibody (1A4 (ASM-1))	ThermoFisher Scientific	https://www.thermofisher.com/antibody/product/Alpha-Smooth-Muscle-Actin-Antibody-clone-1A4-asm-1-Monoclonal/MA5-11547
Goat anti-mouse IgG2a cross-adsorbed secondary antibody, AlexaFluor® 555	ThermoFisher Scientific	https://www.thermofisher.com/antibody/product/Goat-anti-Mouse-IgG2a-Cross-Adsorbed-Secondary-Antibody-Polyclonal/A-21137
Pan-cytokeratin monoclonal antibody (AE1/AE3), AlexaFluor® 488, EBioscience	ThermoFisher Scientific	https://www.thermofisher.com/antibody/product/Pan-Cytokeratin-Antibody-clone-AE1-AE3-Monoclonal/53-9003-82
Ki-67 (D3B5) monoclonal antibody, AlexaFluor® 647, Conjugate #12075	Cell Signaling Technology	https://www.cellsignal.com/products/antibody-conjugates/ki-67-d3b5-rabbit-mab-alexa-fluor-647-conjugate/12075?N=4294960176&Nrp=100&fromPage=plp
Biological Samples		
Human PDAC tumor samples	Oregon Pancreas Tissue Registry	IRB: 3609
Chemicals, Peptides, and Recombinant Proteins		
SlowFade Gold Antifade Mountant with DAPI	ThermoFisher Scientific	https://www.thermofisher.com/order/catalog/product/S36938
Normal goat serum blocking solution	Vector laboratories	https://vectorlabs.com/normal-goat-serum-blocking-solution.html
Deposited Data		
Imaging data	This manuscript	https://gitlab.com/eburling/shift/
Software and Algorithms		
Python	Python Software Foundation	RRID:SCR_008394

PyTorch	Paszke et al., 2017	https://pytorch.org/
imgaug	Jung, 2018	https://github.com/aleju/imgaug
Netron	https://www.lutzroeder.com/ai/	https://github.com/lutzroeder/netron
HoloViews	Stevens et al., 2015	https://holoviews.org/
Bokeh	Bokeh Development Team, 2019	https://bokeh.pydata.org/
scikit-image	Walt et al., 2014	https://scikit-image.org/
scikit-learn	Pedregosa et al., 2011	https://scikit-learn.org
SciPy	Jones et al., 2001	https://www.scipy.org/
Jupyter	Kluyver et al., 2016	https://jupyter.org/

Human tissue samples

This work was conducted as part of the Oregon Pancreas Tissue Registry (IRB#3609), which was approved by the Institutional Review Board at Oregon Health and Science University (OHSU). Informed consent was obtained from all subjects. Four cases of pancreatic ductal adenocarcinoma (PDAC) diagnosed as moderately differentiated adenocarcinoma were retrieved from the OHSU Surgical Pathology Department. Sample A was from a male aged 83 at diagnosis; sample B was from a female aged 74 at diagnosis; sample C was from a female aged 57 at diagnosis; and sample D was from a female aged 73 at diagnosis. H&E-stained sections were secondarily reviewed by two board-certified surgical pathologists tasked to identify and classify areas of tumor heterogeneity in representative sections from each case. Discrepancies between pathologists were ameliorated by consensus review. Samples were chosen via pathological review as exemplifying a spectrum of both histological differentiation and morphology.

Pathological evaluation of human tissue samples

Gold standard review of histologic sections by pathologists tasked with identifying heterogeneous differences in PDAC tumor morphology and grade revealed interobserver agreement in the identification of areas of squamous differentiation in one case and various tumor grades within neoplasms in the other three cases. All four cases were predominantly grade 2 adenocarcinoma and there was no disagreement evaluating marked regions of interest. The case with areas of squamous differentiation did not clearly meet the 30% threshold for

adenosquamous classification. The other three cases were predominantly grade 2 with foci of grade 1 and others with grade 3.

Immunofluorescence staining

Preparation of tissue for immunofluorescence staining

Formalin-fixed paraffin-embedded tissue blocks were serially sectioned by the OHSU Histopathology Shared Resource. From each block, three sections were cut in order to generate a standard H&E for pathological review and downstream analysis, a second serial section of tissue for immunofluorescence staining/post-immunofluorescence H&E staining, and a third section for secondary only control. After sectioning, the second serial tissue section was immediately baked at 55°C for 12 hours and subjected to standard deparaffinization; the slides underwent standard antigen retrieval processing, washing, and blocking. Upon completion, primary antibodies were diluted and applied.

Application of antibodies

Alpha-Smooth Muscle Actin (Mouse monoclonal antibody, IgG2a, Clone: 1A4; Pierce/Invitrogen, cat#MA5-11547) was diluted to 1:200 with Ki-67 (D3B5), (Rabbit monoclonal antibody, IgG, Alexa Fluor® 647 Conjugate; Cell Signaling Technology, cat#12075S) diluted to 1:400, along with Pan Cytokeratin (AE1/AE3) (Mouse monoclonal antibody, IgG1, Alexa Fluor® 488 Conjugate; ThermoFisher, cat#53-9003-82), which was diluted to 1:200 in 10% Normal Goat Serum in 1% Bovine Serum Albumin in Phosphate Buffered Saline. Primary antibodies were diluted and incubated overnight at 4°C. After incubation, secondary antibody (Goat anti-mouse monoclonal antibody, IgG2A, Alexa Fluor® 555 Conjugate; Life Technologies, cat# A21137), at 1:200 dilution was applied to the slides and incubated at room temperature for one hour. After incubation slides were washed and mounted with Slowfade Gold Antifade Mountant with DAPI (Fisher Scientific, cat#S36936) in preparation for image acquisition.

Post-IF H&E staining of tissue samples

After the IF stained slides were scanned and the immunofluorescence staining verified, the glass coverslips were removed and the slides were immediately processed for post-IF H&E staining. Post-IF H&E staining was performed with the Leica Autostainer XL staining system at the OHSU Histopathology Shared Resource with the modified staining protocol described in the table below:

Hematoxylin	10 minutes
Wash in water	1 minute
Acid alcohol (.5% HCl in 70% Ethanol)	8 seconds
Wash in water	25 seconds

Bluing solution	2 minutes
Wash in water	20 seconds
80% Ethanol/water	25 seconds
Eosin	10 seconds
80% Ethanol/water	25 seconds
95% Ethanol/water	20 seconds
100% Ethanol (two times)	25 seconds
Xylene (five times)	25 seconds

Image acquisition and presentation

Slides were scanned with the Zeiss Axio Scan.Z1 slide scanner of the OHSU Advanced Multiscale Microscopy Shared Resource with the 20X objective in both brightfield and immunofluorescence scanning. Carl Zeiss Images (CZI) were acquired using Zeiss Zen software. CZI images from the Zeiss Axioscan Slide Scanner were processed with the Zeiss Blue Zen Lite microscope software package. All brightfield and immunofluorescence images were manually annotated and exported as TIFF files for downstream image processing.

Image pre-processing

Raw H&E and IF whole slide images (WSIs) must be pre-processed to remove technical noise, account for between-sample intensity variation, and align paired H&E and IF WSIs in a shared coordinate system. To do so, we use the following pipeline:

1. Downscaling: 20X WSI are downscaled by a factor of 2 in x and y dimensions to generate 10X WSIs. We experimented with using either 20X or 10X images and found that models performed best when using 10X images.
2. Registration: H&E and IF WSIs are spatially registered using SURF features (Bay et al., 2006; Chang et al., 2017) extracted from hematoxylin and DAPI binary masks of nuclei generated by Otsu's method, respectively.
3. Technical noise reduction: IF WSIs are median filtered with a 5-pixel-radius disk structuring element.
4. Intensity normalization: H&E WSI pixel intensities are normalized as previously described (Macenko et al., 2009). Following Christiansen et al., IF WSI pixel intensities are normalized to have a fixed mean=0.25 and standard deviation=0.125, then clipped to fall within [0,1].

5. Image tiling: WSIs are tiled into non-overlapping 256×256 pixel tiles, such that each H&E tile has a corresponding spatially-registered IF tile. H&E tiles that contained more than 50% background were removed along with the corresponding IF tiles. Each 10X WSI is comprised of hundreds or thousands of non-overlapping 256×256 pixel tiles.

Feature-guided training set selection

Although DL approaches like SHIFT and Label-Free Determination require substantial training data to be robust and generalizable, due to resource constraints we hope that a small number of paired H&E and IF image samples is required for model training. Typically, archival WSIs of H&E-stained tissue sections exist on-hand for each sample, which allows for the screening of samples to identify the minimal number of samples that maximally represent the morphological spectrum of the disease being considered. The recent works of Hegde et al. and Otálora et al. demonstrate that DL systems are well-suited for image retrieval tasks in digital pathology, wherein a pathologist submits a query image or region of interest and the DL system returns similar images based on their DL-defined feature representations. We seek to solve the inverse task of heterogeneous training set selection in digital pathology, though our approach could be extended to any data-limited biomedical imaging domain.

Since PDAC is a morphologically heterogeneous disease, building a representative training set is crucial to the design of a model that will generalize across heterogeneous biopsy samples after deployment. In order to minimize the required resources for acquiring paired H&E and IF images but still cover a broad spectrum of heterogeneous morphological features in the selected H&E samples, we propose a clustering method to learn a heterogeneous representation of H&E sample images. To assess the morphological features of each sample, we use a variational autoencoder (VAE) (Kingma and Welling, 2013) to extract 16-dimensional feature vectors from each H&E tile to establish comparisons between samples. Since texture and morphological features on H&E tiles in each cluster of samples will be comparatively more similar than those of the other cluster, we only select representative H&E samples from each cluster for our training dataset. We also tried using other feature vector sizes for representation learning, e.g. 2, 4, 8, 32, but found that a feature vector size of 16 yielded the lowest reconstruction losses.

For example, if there are four samples being considered for IF staining, but resources limit the number of samples that can be stained to two, a decision must be made about which samples should be selected. For the four samples, we aggregate their archival H&E WSIs, extract features from H&E tiles for each sample using a VAE, and quantitatively determine the samples needed to maximally cover the feature space over which the H&E tile set is distributed. By screening and selecting samples in this data-driven fashion, we exclude homogeneous or redundant samples that would not contribute to model generalizability. This maximizes model performance by ensuring that our training dataset is representative of the disease being modeled, thus

minimizing cost through the selection of the fewest samples required to do so. Perhaps more importantly, when we fail to generate reliable virtual IF images for certain tissue samples or IF markers, this framework will be useful to examine whether or not their morphological features are well presented in the training dataset, which can guide how we select additional samples when updating our dataset.

To identify the sequence of samples that should be selected, we adapt a sample selection algorithm (Feng Pan et al., 2005) parameterized using the following notation:

Parameter	Description
X	Complete tile set of all samples, $X = \{x_1, x_2, \dots, x_n\}$
x_i	Single tile, $x_i \in X$
X_i	Subset of X corresponding to the i th sample, $X_i \subset X$
F	Complete VAE-learned feature set, $F = \{f_1, f_2, \dots, f_m\}$
f_i	Single feature, $f_i \in F$
A	Random variable defined over F
T	Random variable defined over X

We begin with a tiles $\times$ features table, where we set $m = 16$ for our experiments:

	f_1	f_2	...	f_m
x_1	-1.64266	1.36952	...	1.23509
x_2	-0.792104	-0.481497	...	1.07938
...	...	...	...	...
x_n	0.00163981	-0.0162441	...	-0.95883

We normalize across rows of the table, such that each tile is now represented as a probability distribution over the feature domain:

	f_1	f_2	...	f_m	Sum
x_1	0.00418311	0.0850498	...	0.0743519	1
x_2	0.0148384	0.0202433	...	0.0964193	1
...	...	...	...	...	1
x_n	0.0208721	0.0205021	...	0.00798802	1

We define the random variables T and A over tile domain X and the feature domain F , respectively, such that $P(A = f_1 | T = x_1) = 0.418311$, $P(A = f_2 | T = x_2) = 0.0202433$, and so on. With this conditional probability table,

we can define probability distributions for each subset X_i : $P(A|X_i) = \frac{1}{|X_i|} \sum_{x \in X_i} P(A|x)$. To measure the representativeness of sample X_i to the full dataset X , we compute the Kullback-Leibler (KL) divergence between $P(A|X_i)$ and $P(A|X)$: $KL(P(A|X_i)||P(A|X)) = \sum_{f \in F} P(f | X_i) \log \frac{P(f | X_i)}{P(f | X)}$. We then weight this divergence by the proportion of X that X_i comprises, $\frac{|X_i|}{|X|}$, to prioritize subsets that contribute many tiles to X . We define the single most representative sample as $\widehat{X}_1 = \min_{X_i \subset X} \left(\frac{|X_i|}{|X|} KL(P(A|X_i)||P(A|X)) \right)$, the most representative duo of samples as $\widehat{X}_2 = \widehat{X}_1 + \min_{X_i \subset X - \widehat{X}_1} \left(\frac{|X_i|}{|X|} KL(P(A|X_i + \widehat{X}_1)||P(A|X)) \right)$, the most representative trio of samples as $\widehat{X}_3 = \widehat{X}_2 + \min_{X_i \subset X - \widehat{X}_2} \left(\frac{|X_i|}{|X|} KL(P(A|X_i + \widehat{X}_2)||P(A|X)) \right)$, and so on. In this way, we define the sequence of samples that should be chosen to optimally increase the representativeness of the training set.

Model architectures

Conditional generative adversarial networks

Image-to-image translation—the mapping of pixels from one scene representation to pixels of another representation of the same scene—is a fundamental image processing problem. Conditional generative adversarial networks (cGANs) (Goodfellow et al., 2014; Mirza and Osindero, 2014) are a compelling deep learning-based solution to the image-to-image translation problem which have been deployed for many tasks, including detection of skin lesions (Udrea and Mitra, 2017), retinal image synthesis (Costa et al., 2018), super-resolution fluorescence image reconstruction (Ouyang et al., 2018), and virtual H&E staining (Rivenson et al., 2019b). To approach the problem of translating H&E images to their IF counterparts, SHIFT adopts the cGAN-driven architecture *pix2pix* (Isola et al., 2016), which benefits from its bipartite formulation of generator and discriminator. Like other methods proposed for image-to-image translation, cGANs learn a functional mapping from input images x to ground truth target images $\hat{y}$, but, unique to a cGAN architecture, it is the task of a generator network G to generate images $\hat{y}$ conditioned on x , i.e. $G(x) = \hat{y}$, that fool an adversarial discriminator network D , which is in turn trained to tell the difference between real and generated images. What ensues from this two-network duel is a G that generates realistic images that are difficult to distinguish from real images, some GAN-generated images being sufficiently realistic to be considered as a proxy for the ground truth when labeled data are scarce or prohibitively expensive (Bousmalis et al., 2016). Concretely, the cGAN objective is posed as a binary cross-entropy loss:

$$\mathcal{L}_{\text{cGAN}}(G, D) = \mathbb{E}_{x, y \sim p_{\text{data}}(x, y)} \left[\log D(x, y) \right] + \mathbb{E}_{x \sim p_{\text{data}}(x)} \left[\log (1 - D(x, G(x))) \right]$$

where G seeks to minimize the objective and thus minimize the distinguishability of generated and real images, while D seeks the opposite. In addition to the task of fooling D , G is also encouraged to generate images that resemble real images through incorporation of an L1 reconstruction loss term:

$$\mathcal{L}_{L1}(G) = \mathbb{E}_{x,y \sim p_{data}(x,y)} [\|y - G(x)\|_1]$$

The full cGAN objective is:

$$G^* = \arg \min_G \max_D \mathcal{L}_{cGAN}(G, D) + \lambda \mathcal{L}_{L1}(G)$$

where the L1 tuning parameter λ is adapted according to the IF stain prevalence in the current batch of IF tiles (Burlingame et al., 2018). Training data consist of spatially registered pairs of H&E image tiles (x) and IF image tiles (y), while the test data consist of H&E and IF image pairs withheld from the training data. Models were trained using the Adam optimizer with a learning rate of 0.002 for 500 epochs. Training batch sizes were set to 64. The first layers of both the generator and discriminator networks were 128 filters deep (see **Figure S2** for additional architectural details). Full model details are available at <https://gitlab.com/eburling/shift/>.

Variational autoencoders

Variational autoencoder (VAE) architecture (Kingma and Wellington, 2013) is designed to elucidate salient features of data in a data-driven and unsupervised manner. A VAE model seeks to train a pair of complementary networks: an encoder network θ that seeks to model an input x_i as a hidden latent representation z_i , and a decoder network ϕ that seeks to reconstitute x_i from its latent representation z_i . The VAE cost function shown below penalizes model training with an additional Kullback-Leibler (KL) divergence term that works to conform the distribution of z with respect to a given prior, which in our case is the standard normal distribution:

$$\mathcal{L}_i(x_i, \theta, \phi) = -\mathbb{E}_{z \sim q_\theta(z|x_i)} [\log p_\phi(x_i|z)] + \text{KL}(q_\theta(z|x_i) \| p(z))$$

where

$$p(z) = \mathcal{N}(0, 1)$$

By specifying a latent dimension z less than the input dimension of x , a VAE model learns a pair of optimal encoding and decoding functions that enable reconstruction of an input sample subject to capacity constraints of the latent feature space within the model. In general, this formulation learns encoding functions that compress the information content in the high-dimensional input into a low-dimensional embedding space that learns dataset features sufficient to reconstitute the original input sample while preserving an expected distribution over the learned features. This interpretation enables a specified selection criteria function designed to sample whole slide images whose constituent tiles maximally cover the entire learned feature space with a minimal number of samples.

Model ensembles

In addition to testing the ability of independent SHIFT and LFD models to generate virtual IF images, we also tested model ensembles. Ensemble images were generated by simply averaging the virtual IF image outputs of SHIFT and LFD models trained to generate the same stain using the same training set.

Imaging data augmentation

To boost the effective number of images in our training sets and improve model robustness against expected types of technical noise, we apply image augmentations to each image in each training batch using the Python library `imgaug` (Jung, 2018). We apply Gaussian blur, flipping, affine geometric transformation, Gaussian noise, Poisson noise, rotation, and add to hue and saturation in each channel. The implementation of our imaging data augmentation can be viewed at <https://gitlab.com/eburling/shift>.

Image processing in figures

The IF images in Figure 1 are contrast enhanced by saturating the top 1% and bottom 1% of pixel intensities. All other images are processed as described in the image pre-processing section above.

Image comparisons

To compare real and virtual IF images, we measure their structural similarity (Zhou Wang et al., 2004) using the `compare_ssim` function implemented in the Python library `scikit-learn` (Pedregosa et al., 2011). We calculate the SSIM between 11-pixel windows of the real and virtual IF image tiles. The SSIM between two windows x and y is defined as:

$$\text{SSIM}(x, y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)}$$

where μ_x and μ_y are the mean intensities of x and y , σ_{xy} is the covariance of x and y , σ_x^2 and σ_y^2 are the variances of x and y , and c_1 and c_2 are stabilizing constants. The SSIM between real and virtual IF images is the SSIM averaged over their windows. Based on our observation that SSIM is sensitive to simulations of technical noise which are impossible for SHIFT models to infer (**Figure S6**), we apply Gaussian filtering to real and virtual IF images tiles before calculating SSIM using the gaussian function implemented in the Python library `scikit-image` with sigma set to 3. We also measured the Pearson's correlation coefficient between images in **Figure S6** using the `pearson_r` function implemented in the Python library `SciPy` (Jones et al., 2001).

Code availability

Code for training and inference can be found at <https://gitlab.com/eburling/shift>.

SUPPLEMENTARY MATERIALS

Table S1. Variational autoencoder-learned representations and t-SNE coordinates for H&E tiles.

Figure S1. H&E tile feature distributions of experiment sample combinations. The VAE-learned 16-dimensional feature vector representations for each H&E are embedded into 2 dimensions using *t*-SNE. Each point in each plot represents a single H&E tile. Contour lines indicate point density. Sample B is the single sample that is most representative of the whole dataset. Samples B and D are the duo of samples that are most representative of the whole dataset. Samples A and B are a duo of samples that poorly represent the whole dataset.

Figure S2. Schematics of cGAN architecture used by SHIFT. The cGAN architecture used by SHIFT is based on the *pix2pix* framework (Isola et al., 2016). Schematics were generated with the Netron network viewer tool (<https://github.com/lutzroeder/netron>). Full implementation details available on GitLab (<https://gitlab.com/eburling/shift>).

(A) Architecture of generator network G which is based on the U-net architecture (Ronneberger et al., 2015).

(B) Architecture of discriminator network D .

Figure S3. Large-scale comparison of real and virtual panCK staining generated by SHIFT.

SHIFT images were generated by a model trained on sample subsets B1 and D5. Results show are from the test set. Tiles were excluded if they contained more than 50% background in the H&E representation (black tiles).

(A) Representative images taken from sample A.

(B) Representative images taken from sample B.

(C) Representative images taken from sample C.

(D) Representative images taken from sample D.

Figure S4. Large-scale comparison of real and virtual α -SMA staining generated by SHIFT.

SHIFT images were generated by a model trained on sample subsets B1 and D5. Results shown are from the test set. Tiles were excluded if they contained more than 50% background in the H&E representation (black tiles).

Figure S5. Training losses for virtual staining models.

(A) L1 training loss for SHIFT models for each training set composition.

(B) GAN training loss for SHIFT models for each training set composition.

(C) Mean squared error (MSE) training loss for Label-Free Determination model.

Figure S6. Metric sensitivity to perturbations. Structural similarity (SSIM) and Pearson's r are two commonly used metrics for image comparison, both having been used to make comparisons between real and generated biological images in recent related work (Christiansen et al., 2018; Ounkomol et al., 2018; Rivenson et al., 2019b). When considering a stereotypical panCK IF tile, both measures are found to be sensitive to rotation (A) and translation (B), perturbations common to image registration. Grey dotted lines indicate the parameter selected to generate each transformed image. We also observe that both measures are sensitive to simulations of technical noise (C). By applying a Gaussian filter with variance (sigma) set to 3, we recover the SSIM between real and perturbed IF images without sacrificing global image details.