Multi-V-Stain: Multiplexed Virtual Staining of Histopathology Whole-Slide Images

Sonali Andani[*1,3] **Boqi Chen**[* 1,2,5] **Joanna Ficek-Pascual**[* 1]
Simon Heinke[1] **Ruben Casanova**[4] **Bettina Sobottka**[3]
Bernd Bodenmiller[4] **Tumor Profiler Consortium**
Viktor H. Koelzer[† 3] **Gunnar Rätsch**[† 1]
[1]Department of Computer Science, ETH Zurich [2]ETH AI Center, ETH Zurich
[3]Department of Pathology and Molecular Pathology, University Hospital Zurich
[4]Department of Quantitative Biomedicine, University of Zurich
[5] Computer Vision Laboratory, ETH Zurich

Abstract

Pathological assessment of Hematoxylin & Eosin (H&E) stained tissue samples is a well-established clinical routine for cancer diagnosis. While providing rich morphological data, it lacks information on protein expression patterns which is crucial for cancer prognosis and treatment recommendations. Imaging Mass Cytometry (IMC) excels in highly multiplexed protein profiling but faces challenges like high operational cost and a restrictive focus on small Regions-of-Interest. Addressing this, we introduce Multi-V-Stain, a novel image-to-image translation method for multiplexed IMC virtual staining. Our method effectively utilizes the rich morphological features from H&E images to predict multiplexed protein expressions at a Whole-Slide Image level. In evaluations using an in-house melanoma dataset, Multi-V-Stain consistently outperforms existing methods in terms of image quality and biological relevance of the generated stains.

1 Introduction

Hematoxylin & Eosin (H&E) staining of tissue sections is a gold standard in the clinical practice for cancer diagnosis. The morphological features in H&E images offer a comprehensive view of tissue organisation, crucial for cancer grading, proliferation, and staging [1]. To determine cancer prognosis, technologies such as Immunohistochemistry (IHC) and Imaging Mass Cytometry (IMC) are valuable tools, allowing for the analysis of protein expression in cancer tissue samples [2]. Compared to IHC, IMC facilitates the simultaneous spatially-resolved quantification of up to 40 proteins. Such highly multiplexed data provides a comprehensive view of the tumor microenvironment (TME) [3, 4], enabling study of protein co-expression and cellular dynamics, which is crucial for treatment decision process [5]. Nevertheless, IMC is limited by low throughput and high costs, and restricted to profiling within small Region-of-Interests (RoIs). Therefore, novel computational methods are needed to generate multiplexed virtual IMC stains from clinically-available diagnostic Whole-Slide Images (WSIs) in a cost and time-effective manner to provide insights into the TME.

The advent of Image-to-Image Translation (I2IT) has extended the frontiers of virtual staining of protein markers from H&E images [6, 7]. Despite the progress, most methods target only one protein marker, necessitating the training of separate models for multiplexed predictions, which is inefficient in time, resources, and modeling inter-marker correlations. Recent advancements in multi-domain I2IT [8–10] have enabled the generation of multiple stains for distinct protein markers using a

[*]Equal contribution [†]Joint supervision

1st Workshop on Deep Generative Models for Health at NeurIPS 2023.

single model [11, 12]. However, these methods consider each protein marker as an independent domain and optimize for individual markers iteratively. We propose that a simultaneous optimization for all markers would more effectively capture their interrelationships, thereby producing more biologically meaningful stains. Moreover, given the scarcity of paired multi-stain data, these methods predominantly rely on unsupervised I2IT. However, it has been empirically shown that unpaired approaches are less effective in generating biologically significant stains compared to supervised translation methods with paired data [7]. Addressing these challenges, we introduce Multi-V-Stain, a virtual staining method for generating multiplexed IMC stains from H&E images. Our approach utilizes direct pixel-level supervision from paired H&E and IMC data and optimizes concurrently for all IMC markers. Additionally, considering that H&E and IMC staining are performed on consecutive slices during data acquisition, leading to positional shifts known as slice-to-slice differences, we introduce a novel Adaptive Contrastive Perceptual (ACP) loss. The ACP harnesses the capabilities of contrastive learning [13] to reduce the need for precisely pixel-wise aligned data in supervised translation. Through rigorous evaluations, Multi-V-Stain demonstrates superior quality in generated IMC images and highlights the biological relevance of the stains.

In summary, our main contributions are: (1) To the best of our knowledge, this is the first study to predict multiplexed IMC stains from H&E images. (2) We introduce a novel ACP loss to effectively address the challenges introduced by slice-to-slice differences. (3) We demonstrate that by learning a joint mapping of multiple IMC markers from H&E, we can exploit underlying inter-marker correlations to generate more biologically meaningful stains.

2 Methods

Problem Definition Given a dataset $\mathcal{D} = \{(x^{(i)}, y^{(i)})\}_{i=1}^{M}$ of M paired samples, where $x \in \mathcal{X}$ and $y = (y_1, \dots, y_N) \in \mathcal{Y}$, the task is to estimate the expectation over the conditional joint distribution $p(y|x) = p(y_1, y_2, \dots, y_N|x)$ of the expression levels of N protein markers from IMC stains given the H&E image. Here, $\mathcal{X} \in \mathbb{R}^{H \times W \times 3}$ represents the domain of all possible H&E images, $\mathcal{Y} \in \mathbb{R}^{H \times W \times N}$ represents the domain of all possible corresponding IMC images, and y_i denotes the i^{th} protein marker. The objective is to learn a mapping function $G : \mathcal{X} \to \mathcal{Y}$, such that $\mathbb{E}(y|x) \approx G(x)$. Note that for existing multi-domain I2IT methods, the problem is essentially formulated as learning N disjoint mappings $\{G_i : \mathbb{R}^{H \times W \times C} \to \mathbb{R}^{H \times W \times 1}\}_{i=1}^{N}$ to approximate expected values over each marginal distributions $p(y_i|x)$.

ACP Loss Despite having paired data, ensuring strict pixel correspondence presents challenges due to slice-to-slice differences between two consecutive tissue cuts used for H&E and IMC staining. To address this, we introduce a contrastive loss, termed ACP loss. This loss aims to maximize mutual information between the predicted and ground-truth (GT) IMC image, based on the principle that higher mutual information correlates with more accurate predictions [14]. The ACP loss employs a VGG-19 [15] network, denoted as F, pretrained on ImageNet [16] for feature embedding. Such a design choice is inspired by the demonstrated effectiveness of pretrained feature spaces in capturing *perceptual* similarity, even in tasks that appear unrelated [17]. The incorporation of this aspect is crucial for enhancing the perceptual realism of the generated IMC images, which is often linked with their biological relevance. Formally, let $\hat{v}_l^p$ represent the features of patch p in the generated IMC image where $p \in P_l$ and P_l is the set of patches at layer l. This $\hat{v}_l^p$ serves as the anchor sample. In contrast, v_l^p denotes the features from the corresponding patch in the GT IMC image, acting as the positive sample. Meanwhile, $\bar{v}_l^p$ represents features from a collection of non-corresponding patches in the GT IMC image, serving as negative samples. Here, $l \in L$ indicates the layer from which these features are extracted. With these notions, the ACP loss is defined as

$$\mathcal{L}_{\text{ACP}} = \mathop{\mathbb{E}}_{\substack{x \sim X \\ y \sim \mathcal{Y}}} \frac{1}{|L|} \frac{1}{|P_l|} \sum_{l \in L} \sum_{p \in P_l} \frac{w_t(\hat{v}_l^p, v_l^p)}{W_{t,l}} \ell_{\text{InfoNCE}}(\hat{v}_l^p, v_l^p, \bar{v}_l^p). \tag{1}$$

In this equation, $w_t(\hat{v}_l^p, v_l^p)$ and W_t^l are the adaptive weight and normalization factor, respectively; and ℓ_{InfoNCE} denotes the InfoNCE loss [18]. For a comprehensive definition of the ACP loss, please refer to Appendix B.1.

Training Objective By combining the ACP with the Pix2pix [19] loss, our objective is to find an optimal generator G, which is trained adversarially with a discriminator D using min-max optimization. The goal is to achieve:

$$G^* = \arg\min_G \min_F \max_D \mathbb{E}_{\substack{x\sim\mathcal{X}\\ y\sim\mathcal{Y}}}\big[\mathcal{L}_{\text{Pix2pix}}(x, G(x), y; G, D) + \lambda_{\text{ACP}}\mathcal{L}_{\text{ACP}}(G(x), y; G, F)\big], \quad (2)$$

where λ_{ACP} is the weight for ACP loss. For the detailed definition of Pix2pix loss, please refer to Appendix B.2. Note that there is a slight abuse of notation: F in Equation 2 represents the composition of the frozen feature extractor F detailed in the previous section and a small trainable Multi-layer Perceptron (MLP), which is shown to improve the performance of contrastive learning [13] .

3 Experiments

Dataset We use an internal dataset of 80 patients with metastatic melanoma. A total of six RoIs of 1 mm^2 are selected on each H&E WSI (scanned at the resolution of 0.25 µm/pixel) based on expert pathologist visual inspection. Corresponding IMC data is generated for the same RoIs on the consecutive tissue sections, with a resolution of 1 µm/pixel. This yields 336 paired H&E and IMC RoI samples, aligned using template matching [20]. The dataset is split into training (231 pairs), validation (38 pairs), and test (67 pairs) sets, stratified by immune phenotype. Further details are provided in Appendix C.

Evaluation Metrics We assess the quality of generated IMC images using Multiscale Structural Similarity Index (MS-SSIM) [21], Fréchet Inception Distance (FID) and Kernelized Inception Distance (KID). For evaluating biological relevance, we measure the protein co-expression similarity, indicated by the normalized mean square error (NMSE) between predicted and GT Spearman's Correlation Coefficient (SCC) for each protein pair, averaged over test RoIs. It is important to note that protein co-expression patterns is preserved across tissue sections, regardless of slicing. This characteristic provides a robust evaluation metric, particularly useful in mitigating the impact of slice-to-slice discrepancies. Further details are provided in Appendix E.

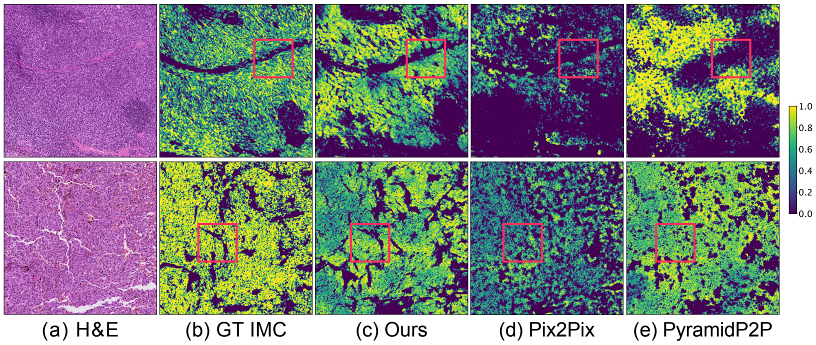
4 Results

We benchmark Multi-V-Stain against Pix2pix [19] and PyramidP2P [7]. All methods are evaluated on two settings: Multiplex (MP) and Singleplex (SP). In MP, a single model is trained to predict all markers simultaneously; whereas in SP, separate models are trained to predict individual markers, which are subsequently stacked together to get a (pseudo-)multiplexed prediction.

Multi-V-Stain captures structural details. The image quality assessment metrics, namely MS-SSIM, FID, and KID, are presented in Table 1. Notably, our method demonstrates superior performance in FID and KID in both settings. In terms of MS-SSIM, our method outperforms all MP baselines and achieves comparable performance with the top-performing method in SP. This indicates our model's capability to generate IMC images with a high level of fidelity to the ground truth (GT). The qualitative evaluation, illustrated in Figures 1 and 5, further supports our model's ability to capture fine-grained structural details from the IMC ground truth (GT), as indicated by the red squares. Interestingly, we observe that models in the SP setting sometimes exhibit superior performance compared to their counterparts in the MP setting in terms of image quality metrics, such as MS-SSIM. This could potentially be due to the relatively easier optimization landscape when modeling marginal distribution for each protein independently in SP setting, as opposed to a more complex landscape to learn joint probability distribution in MP setting. Our objective is to learn biologically relevant markers, and in the following section, we demonstrate the utility of learning markers jointly in the MP setting.

Multi-V-Stain predicts biologically meaningful stains. Table 1 illustrates the biological relevance of IMC images, assessed through protein co-expression patterns. The last column shows the NMSE between predicted and GT SCC, averaged over all protein pairs. Our observations indicate that in the MP setting, our method exhibits the lowest NMSE, signifying superior agreement with the GT compared to baseline methods. Moreover, across all methods, MP consistently outperforms SP.

Figure 2 visualizes the co-expression pattern for each protein pair, highlighting an overall better alignment of MP with GT, particularly for pairs exhibiting high positive correlations (*i.e.*, those on the right of the X-axis). Specifically, MP provides more accurate co-expression patterns for protein pairs CD16:HLA-DR, CD8a:HLA-ABC, CD3:HLA-ABC, CD20:HLA-ABC, and CD16:CD8a. Each



(a) H&E (b) GT IMC (c) Ours (d) Pix2Pix (e) PyramidP2P

Figure 1: Qualitative assessment of Multi-V-Stain and baselines at RoI level. Rows represent two RoIs: top for MelanA and bottom for S100 IMC protein.

Table 1: Quantitative evaluation on an in-house Melanoma dataset for both multiplex (MP) and singleplex (SP). Arrows indicate whether higher (↑) or lower (↓) values are better. Best results are highlighted in **bold** for MP and SP.

Metric	MS-SSIM↑	FID↓	KID↓	NMSE ↓
Pix2pix-MP [19]	0.203	6.588	0.052	0.009
PyramidP2P-MP [7]	0.205	7.415	0.095	0.008
Ours-MP	**0.208**	**5.489**	**0.034**	**0.007**
Pix2pix-SP [19]	**0.274**	35.796	0.102	**0.012**
PyramidP2P-SP [7]	0.206	5.811	0.057	0.013
Ours-SP	0.210	**5.498**	**0.028**	0.016

of these pairs includes at least one CD-based immune marker, which is sparsely represented in our dataset. We hypothesize that in the absence of an abundant marker, SP models for sparse markers receive insufficient structural details of tissue organization, failing to predict the protein expression accurately. Conversely, a model in MP setting captures superior co-expression patterns for two reasons. First, simultaneous training of multiple proteins facilitates learning correlations between markers, contributing to improved predictions of co-expression patterns. Second, for sparse markers, abundant markers provide auxiliary information on tissue morphology, supporting better predictions. Additionally, we observe over-prediction in co-expression patterns in MP and under-prediction in SP for the CD3:CD8a protein pair. CD3 and CD8a are expressed on the surface of T cells, which are a type of white blood cell involved in the immune response. CD3 is universally expressed on all T-cells, while CD8a specifically defines cytotoxic T-cells, a distinct subset of T cells. The different subtypes of T-cell exhibit similar morphological features in H&E images. Consequently, models in the MP setting tend to predict the expression of all plausible cells, leading to false-positive or over-prediction. In contrast, models in SP setting are likely to fail to predict proteins expressed in small subsets of cells due to data sparsity, resulting in false-negative or under-prediction. Despite the over-prediction of certain proteins, our model can facilitate decision-making of the physicians by providing insights into TME. For example, when studying the interaction of immune cells with tumor compartment for designing effective immunotherapeutic strategies, the pathologists focus on global T and B cell-types instead of subtypes of T and B cells. However, for more fine-grained analyses, such as distinguishing between closely related cellular subsets, our model may face limitations. Therefore, it is a priority for future work to refine the model's ability to accurately distinguish between these finer subsets of cells.

Multi-V-Stain scales to WSI level, suggesting clinical applicability. We showcase the clinical utility of our model by generating multiplexed IMC stains on WSIs of up to 100,000×100,000 pixels in $\leq$ 8 minutes on a NVIDIA A100 GPU. By enabling simultaneous visualization of multiple protein markers at WSI level, our model facilitates a comprehensive understanding of the TME. For instance, in Figure 3 (b), we can visualize the interaction between the tumor and its microenvironment. Specifically, for the sample on top, we observe that immune cells (depicted in green and blue) predominantly localize outside the tumor compartment (depicted in red). This spatial segregation suggests limited interaction between the tumor and neighboring immune components, typically

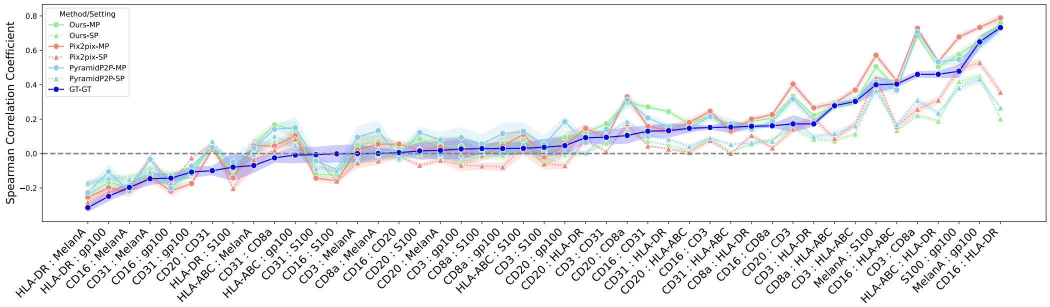
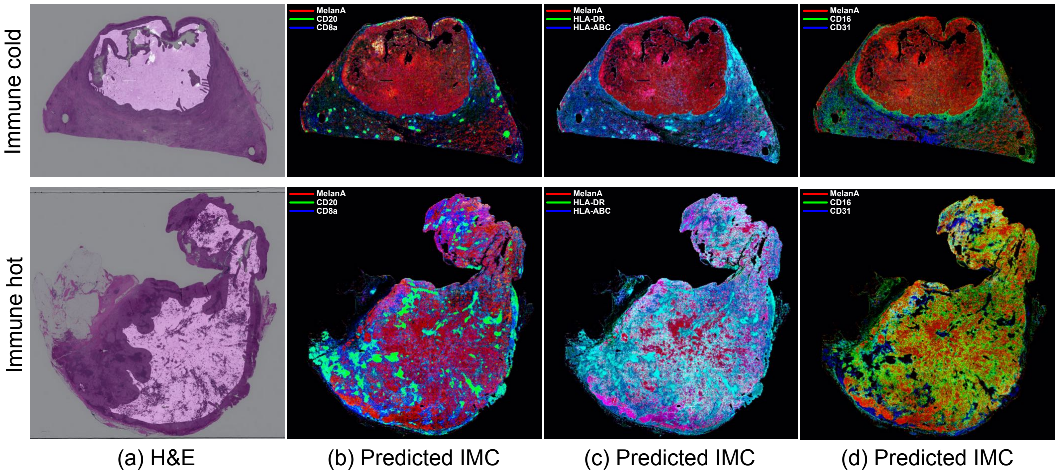


Figure 2: Mean Spearman Correlation Coefficient (SCC) between protein pairs across test ROIs for GT (blue), ours (green), PyramidP2P (pink), and Pix2pix (orange). Multiplex (MP) predicted IMC is denoted with dots, while singleplex (SP) stacked IMC is represented by triangles. The Y-axis and X-axis represent the SCC and protein pairs, respectively.



(a) H&E (b) Predicted IMC (c) Predicted IMC (d) Predicted IMC

Figure 3: Multiplexed IMC prediction at WSI level on an immune cold (top) and immune hot (bottom) sample. (a) H&E input with overlay of tumor center. (b-d) Predicted IMC markers visualized in RGB: (b) MelanA (tumor marker), CD20 (B cell marker), and CD8a (cytotoxic T cell marker); (c) MelanA, HLA-DR (interacts with CD4+ T cells), and HLA-ABC (interacts with CD8+ T cells); (d) MelanA, CD16 (NK cell marker), and CD31 (endothelial marker).

classified as *immune cold*. This observation aligns with the label assigned by the expert pathologist. For the second sample in bottom in Figure 3 (b), we observe the infiltration of immune cells (depicted in green and blue) within the tumor compartment (depicted in red). This observation aligns with the clinically known *immune hot* case and corresponds to the label assigned by our expert. To conclude, we have demonstrated the potential of WSI level predictions from Multi-V-Stain serving as a biomarker for assessing the response to immune therapy. Additional examples are provided in Figure 6.

5 Conclusion

We introduce Multi-V-Stain, a novel virtual staining method for generating multiplexed IMC stains. Our proposed ACP loss enables effective utilization of pixel-level supervision from paired H&E and IMC data while adjusting for inter-slice variations, which is demonstrated to help better capture fine-grained structural details. Evaluations on an in-house melanoma dataset validate Multi-V-Stain's ability to generate faithful and biologically meaningful IMC stains, outperforming existing I2IT techniques. In addition, we show that Multi-V-Stain can efficiently scale to the WSI level, revealing its potential to augment clinical diagnostics by creating a comprehensive protein landscape.

References

[1] A. H. Fischer, K. A. Jacobson, J. Rose, R. Zeller, Hematoxylin and eosin staining of tissue and cell sections, Cold spring harbor protocols 2008 (5) (2008) pdb–prot4986.

[2] J. Duraiyan, R. Govindarajan, K. Kaliyappan, M. Palanisamy, Applications of immunohistochemistry, Journal of pharmacy & bioallied sciences 4 (Suppl 2) (2012) S307.

[3] H. W. Jackson, J. R. Fischer, V. R. T. Zanotelli, H. R. Ali, R. Mechera, S. D. Soysal, H. Moch, S. Muenst, Z. Varga, W. P. Weber, B. Bodenmiller, The single-cell pathology landscape of breast cancer, Nature 578 (7796) (2020) 615–620. `doi:10.1038/s41586-019-1876-x`.
URL `https://doi.org/10.1038/s41586-019-1876-x`

[4] J. Ptacek, D. Locke, R. Finck, M.-E. Cvijic, Z. Li, J. G. Tarolli, M. Aksoy, Y. Sigal, Y. Zhang, M. Newgren, J. Finn, Multiplexed ion beam imaging (mibi) for characterization of the tumor microenvironment across tumor types, Laboratory Investigation 100 (8) (2020) 1111–1123.

[5] M.-Z. Jin, W.-L. Jin, The updated landscape of tumor microenvironment and drug repurposing, Signal Transduction and Targeted Therapy 5 (1) (2020) 166. `doi:10.1038/s41392-020-00280-x`.
URL `https://doi.org/10.1038/s41392-020-00280-x`

[6] E. A. Burlingame, M. McDonnell, G. F. Schau, G. Thibault, C. Lanciault, T. Morgan, B. E. Johnson, C. Corless, J. W. Gray, Y. H. Chang, Shift: speedy histological-to-immunofluorescent translation of a tumor signature enabled by deep learning, Scientific Reports 10 (1) (2020) 17507. `doi:10.1038/s41598-020-74500-3`.
URL `https://doi.org/10.1038/s41598-020-74500-3`

[7] S. Liu, C. Zhu, F. Xu, X. Jia, Z. Shi, M. Jin, Bci: Breast cancer immunohistochemical image generation through pyramid pix2pix, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 1815–1824.

[8] X. Huang, M.-Y. Liu, S. Belongie, J. Kautz, Multimodal unsupervised image-to-image translation, in: Proceedings of the European conference on computer vision (ECCV), 2018, pp. 172–189.

[9] M.-Y. Liu, X. Huang, A. Mallya, T. Karras, T. Aila, J. Lehtinen, J. Kautz, Few-shot unsupervised image-to-image translation, in: Proceedings of the IEEE/CVF international conference on computer vision, 2019, pp. 10551–10560.

[10] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, J. Choo, Stargan: Unified generative adversarial networks for multi-domain image-to-image translation, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 8789–8797.

[11] Y. Lin, B. Zeng, Y. Wang, Y. Chen, Z. Fang, J. Zhang, X. Ji, H. Wang, Y. Zhang, Unpaired multi-domain stain transfer for kidney histopathological images, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 36, 2022, pp. 1630–1637.

[12] M. Berijanian, N. S. Schaadt, B. Huang, J. Lotz, F. Feuerhake, D. Merhof, Unsupervised many-to-many stain translation for histological image augmentation to improve classification accuracy, Journal of Pathology Informatics 14 (2023) 100195.

[13] T. Chen, S. Kornblith, M. Norouzi, G. Hinton, A simple framework for contrastive learning of visual representations, in: International conference on machine learning, PMLR, 2020, pp. 1597–1607.

[14] A. Andonian, T. Park, B. Russell, P. Isola, J.-Y. Zhu, R. Zhang, Contrastive feature loss for image prediction, in: Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 1934–1943.

[15] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, arXiv preprint arXiv:1409.1556 (2014).

[16] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: A large-scale hierarchical image database, in: 2009 IEEE conference on computer vision and pattern recognition, Ieee, 2009, pp. 248–255.

[17] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, O. Wang, The unreasonable effectiveness of deep features as a perceptual metric, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 586–595.

[18] A. v. d. Oord, Y. Li, O. Vinyals, Representation learning with contrastive predictive coding, arXiv preprint arXiv:1807.03748 (2018).

[19] P. Isola, J.-Y. Zhu, T. Zhou, A. A. Efros, Image-to-image translation with conditional adversarial networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 1125–1134.

[20] G. Bradski, The OpenCV Library, Dr. Dobb's Journal of Software Tools (2000). doi:10.1038/s41374-020-0417-4.
URL https://doi.org/10.1038/s41374-020-0417-4

[21] Z. Wang, E. P. Simoncelli, A. C. Bovik, Multiscale structural similarity for image quality assessment, in: The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003, Vol. 2, Ieee, 2003, pp. 1398–1402.

[22] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, in: International Conference on Medical image computing and computer-assisted intervention, Springer, 2015, pp. 234–241.

[23] F. Li, Z. Hu, W. Chen, A. Kak, Adaptive supervised patchnce loss for learning h&e-to-ihc stain translation with inconsistent groundtruth image pairs, arXiv preprint arXiv:2303.06193 (2023).

[24] X. Mao, Q. Li, H. Xie, R. Y. Lau, Z. Wang, S. Paul Smolley, Least squares generative adversarial networks, in: Proceedings of the IEEE international conference on computer vision, 2017, pp. 2794–2802.

[25] A. Karnewar, O. Wang, Msg-gan: Multi-scale gradients for generative adversarial networks, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 7799–7808.

[26] L. Mescheder, A. Geiger, S. Nowozin, Which training methods for gans do actually converge?, in: International conference on machine learning, PMLR, 2018, pp. 3481–3490.

[27] C. McQuin, A. Goodman, V. Chernyshev, L. Kamentsky, B. A. Cimini, K. W. Karhohs, M. Doan, L. Ding, S. M. Rafelski, D. Thirstrup, et al., Cellprofiler 3.0: Next-generation image processing for biology, PLoS biology 16 (7) (2018) e2005970.

[28] H. L. Crowell, S. Chevrier, A. Jacobs, S. Sivapatham, B. Bodenmiller, M. D. Robinson, T. P. Consortium, et al., An r-based reproducible and user-friendly preprocessing pipeline for cytof data, F1000Research 9 (1263) (2020) 1263.

[29] N. Otsu, A threshold selection method from gray-level histograms, IEEE transactions on systems, man, and cybernetics 9 (1) (1979) 62–66.

[30] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, et al., Pytorch: An imperative style, high-performance deep learning library, Advances in neural information processing systems 32 (2019).

[31] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, arXiv preprint arXiv:1412.6980 (2014).

[32] K. I. Bland, M. M. Konstadoulakis, M. P. Vezeridis, H. J. Wanebo, Oncogene protein co-expression. value of ha-ras, c-myc, c-fos, and p53 as prognostic discriminants for breast carcinoma., Annals of surgery 221 (6) (1995) 706.

[33] K. Tanimoto, Y. Yakushijin, H. Fujiwara, M. Otsuka, K. Ohshima, A. Sugita, A. Sakai, T. Hato, H. Hasegawa, M. Yasukawa, Clinical significance of co-expression of cd21 and lfa-1 in b-cell lymphoma, International journal of hematology 89 (2009) 497–507.

[34] S. Graham, Q. D. Vu, S. E. A. Raza, A. Azam, Y. W. Tsang, J. T. Kwak, N. Rajpoot, Hover-net: Simultaneous segmentation and classification of nuclei in multi-tissue histology images, Medical image analysis 58 (2019) 101563.

A Network Architecture

The *translator* based on a fully convolutional U-Net [22] has an encoder and a decoder. The encoder uses six downsampling blocks, each with a convolution layer of stride 2. The original decoder uses deconvolution layers but they often lead to checkerboard effects caused by pixel replication. To generate realistic looking Imaging Mass Cytometry (IMC) images, we adapt the decoder by first up-scaling the input using nearest-neighbour interpolation and then using convolution layer with stride 1. Inspired by [19], each layer in the encoder and decoder is followed by a batch-norm layer and ReLU activation. Further, to account for the difference in resolution of Hematoxylin & Eosin (H&E) and ground-truth (GT) IMC images, we adapt the U-Net architecture to allow for input and output images of any resolution. The *discriminator* has six blocks each with convolution layer with spectral normalisation layer and ReLU activation.

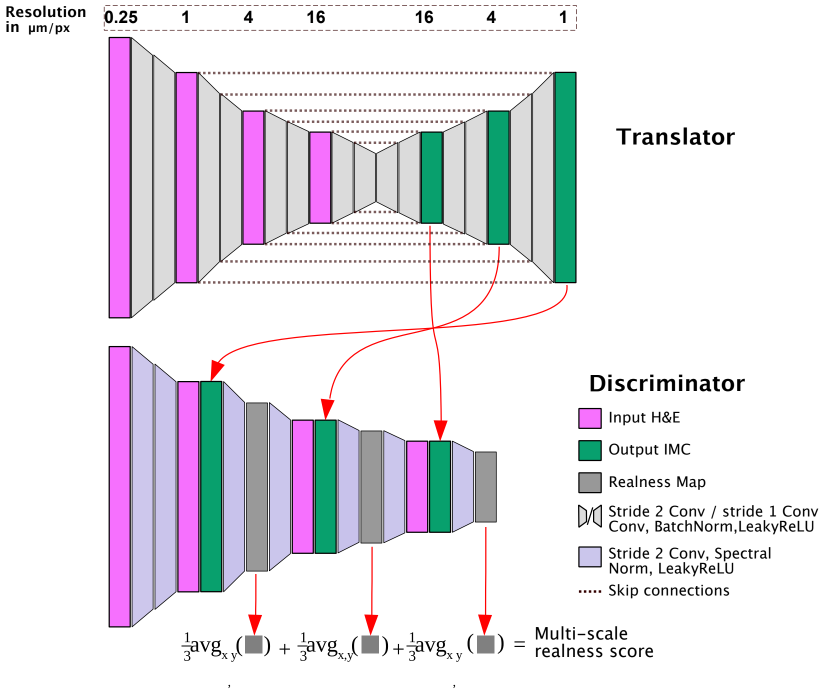

Resolution in μm/px
0.25
1
4
16
16
4
1
Translator
Discriminator
Input H&E
Output IMC
Realness Map
Stride 2 Conv / stride 1 Conv Conv, BatchNorm,LeakyReLU
Stride 2 Conv, Spectral Norm, LeakyReLU
Skip connections
$\frac{1}{3}\text{avg}_{x,y}(\ ) + \frac{1}{3}\text{avg}_{x,y}(\ ) + \frac{1}{3}\text{avg}_{x,y}(\ )$ = Multi-scale realness score

Figure 4: Detailed architecture of Multi-V-Stain. A U-Net architecture with multiple inputs (H&E in pink) and outputs (IMC in green) is used for the *Translator* (top). The *Discriminator* (bottom) outputs three realness maps (dark gray), conditioned on the input H&E (pink). The realness maps are averaged to obtain a final realness score.

B Loss

B.1 Adaptive Contrastive Perceptual (ACP) Loss

Consider the mapping function $G : \mathcal{X} \rightarrow \mathcal{Y}$, which transforms images from the H&E domain $\mathcal{X} \in \mathbb{R}^{H\times W\times 3}$ into the IMC domain $\mathcal{Y} \in \mathbb{R}^{H\times W\times N}$, yielding a predicted image representation $\hat{y}$ for a given input x. Furthermore, we define a sampling function, denoted as $S : \mathcal{Y} \times \mathcal{L} \times \mathcal{P}_l \rightarrow \mathcal{P}$. This function facilitates random sampling of patches from an image in $\mathcal{Y}$, conditioned on a layer index from the set $\mathcal{L}$ and a patch index from $\mathcal{P}_l$. The resultant patch resides in the domain $\mathcal{P}$. With these definitions, given an input H&E image x, the predicted corresponding IMC image $\hat{y}$ is given by

$$\hat{y} = G(x). \tag{3}$$

For each specified layer l and patch p, the patches sampled from the original and predicted images at the same spatial locations are given by

$$\begin{aligned} s_{l,p} &= S(y, l, p), \\ \hat{s}_{l,p} &= S(\hat{y}, l, p). \end{aligned} \tag{4}$$

The randomly sampled patches are then subjected to the feature extraction via a pretrained encoder F. To enrich the expressiveness of these features, a transformation via a small projection head, denoted as H, is applied following the feature extraction. We adopt a two-layer Multi-layer Perceptron as our projection head, which is shown to improve the performance of contrastive learning [13]. The combined operation for both GT and predicted IMC images can be represented as:

$$\begin{aligned} v_l^p &= H(F(s_{l,p})), \\ \hat{v}_l^p &= H(F(\hat{s}_{l,p})). \end{aligned} \tag{5}$$

Note that there is a slight abuse of notation in the main paper, where we use F to represent the combined operation of F and H.

Then, we can obtain two feature sets by aggregating the features from each layer, giving

$$\begin{aligned} V_l &= \{v_l^p\}_{p=1}^{|P_l|}, \\ \hat{V}_l &= \{\hat{v}_l^p\}_{p=1}^{|P_l|}. \end{aligned} \tag{6}$$

Our ACP loss critically relies on the contrastive objective, which in this context adopts the form of the InfoNCE loss [18]. The formal expression for the InfoNCE loss is given as

$$\ell_{\text{InfoNCE}}(v, v^+, v^-) = -\log \frac{\exp(v \cdot v^+/\tau)}{\exp\left(v \cdot v^+/\tau\right) + \sum_{n=1}^{N} \exp\left(v \cdot v_n^-\right)/\tau)}, \tag{7}$$

where v, v^+ and v^- are the embeddings of the anchor, positive and negative samples, respectively.

Let $\hat{v}_l^p$ denote the feature from l^{th} layer of p^{th} patch, serving as the anchor sample, while v_l^p and $\bar{v}_l^p \in \mathbb{R}^{(|P_l|-1)\times D_l}$ denote the feature from the corresponding patch and the collection of features from non-corresponding patches, serving as positive and negative samples, respectively, with D_l signifying the feature dimension; our ACP loss is formulated as:

$$\mathcal{L}_{\text{ACP}} = \mathop{\mathbb{E}}_{\substack{x\sim X \\ y\sim\mathcal{Y}}} \frac{1}{|L|} \frac{1}{|P_l|} \sum_{l\in L} \sum_{p\in P_l} \frac{w_t(\hat{v}_l^p, v_l^p)}{W_{t,l}} \ell_{\text{InfoNCE}}(\hat{v}_l^p, v_l^p, \bar{v}_l^p), \tag{8}$$

where

$$w_t(\hat{v}_l^p, v_l^p) = \left(1 - g\left(\frac{t}{T}\right)\right) \times 1.0 + g\left(\frac{t}{T}\right) \times h(\hat{v}_l^p, v_l^p) \tag{9}$$

is the adaptive patch weight [23]. Here, t and T denote the current and total training steps, respectively. The weight combines a weighting function h based on prediction-GT similarity and a scheduling function g to initially equalize patch pair weights, mitigating bias in early training. The normalization factor $W_t^l = \sum_{p\in P_l} w_{t,l}^p$ maintains loss magnitude after applying the weights.

While our objective is to model the joint distribution $\mathcal{Y}$ of the expression of N IMC markers, we encounter a practical limitation when employing pre-trained feature encoders, which are often designed for 3-channel images. To circumvent this, we decompose the N-channel image into N binary images, each corresponding to one protein marker. We then assume these protein markers to be independent and pass them to the feature encoder separately. This enables us to approximate $\mathcal{Y}$ by the product of its marginal distributions $\mathcal{Y}_i$, *i.e.*, $\mathcal{Y} = \mathcal{Y}_1 \times \mathcal{Y}_2 \times \ldots \times \mathcal{Y}_N$. Our ACP loss can then be formulated as

$$\mathcal{L}_{\text{ACP}} = \mathop{\mathbb{E}}_{\substack{x\sim X \\ y_i\sim\mathcal{Y}_i}} \frac{1}{|L|} \frac{1}{|P_l|} \sum_{i\in N} \sum_{l\in L} \sum_{p\in P_l} \frac{w_t(\hat{v}_{l,i}^p, v_{l,i}^p)}{W_{t,l}} \ell_{\text{InfoNCE}}(\hat{v}_{l,i}^p, v_{l,i}^p, \bar{v}_{l,i}^p). \tag{10}$$

It is important to note that the assumption of independence among the protein markers is a strong one, and it contradicts the biological understanding that these markers are often interrelated. This simplification is a trade-off for the practical benefit of using pre-trained feature encoders. However, we conjecture that training a specialized feature encoder capable of handling multi-channel IMC images as input would provide a better feature representation and thereby improve the performance.

B.2 Pix2pix Loss

We use the least square loss proposed in LSGAN [24] as the adversarial loss. To account for slice-to-slice discrepancy [23], we adopt the multi-scale gradient approach [25], which allows gradient propagation at multiple scales simultaneously. More specifically, the translator predicts target images at multiple resolutions, and the discriminator produces an average realness score of generated images across all scales. For a set of scales S,

$$\begin{aligned}\mathcal{L}_G^{\text{adv}} &= \frac{1}{|S|}\sum_{s\in S}\mathbb{E}_{x\sim\mathcal{X}}\left[\left(D(G^{(s)}(x)|x)-1\right)^2\right],\\ \mathcal{L}_D^{\text{adv}} &= \frac{1}{|S|}\sum_{s\in S}\left[\mathbb{E}_{\substack{x\sim\mathcal{X}\\ y\sim\mathcal{Y}}}\left[(D(y|x)-1)^2\right]+\mathbb{E}_{x\sim\mathcal{X}}\left[(D(G^{(s)}(x)|x))^2\right]\right].\end{aligned} \tag{11}$$

where $|\cdot|$ denotes the cardinality.

The $\mathcal{L}_1$ loss is computed only on the final prediction at the highest resolution. To relax the constraint on pixel-to-pixel alignment, we employ a Gaussian pyramid-based $\mathcal{L}_1$ loss, utilizing the multi-resolution representations of the predicted and GT images [7]. A Gaussian pyramid is constructed through iterative Gaussian smoothing and downsampling. Each level of resolution, termed an "octave", comprises a series of images with increasing degrees of smoothness. Transition between resolutions is achieved by downsampling the image at the highest smoothness level of the current octave to initiate the next. The $\mathcal{L}_1$ loss is computed on the primary layer of each octave as follows:

$$\mathcal{L}_1^r = \mathbb{E}_{\substack{x\sim\mathcal{X}\\ y\sim\mathcal{Y}}}\left\|y_1^r - \hat{y}_1^r\right\|_1, \tag{12}$$

where r denotes the resolution level. The final $\mathcal{L}_1$ loss is a weighted sum across all resolutions:

$$\mathcal{L}_1 = \sum_{r\in R} w_r \mathcal{L}_1^r, \tag{13}$$

where w_r is the weight for resolution r. The scale transformation within an octave and the transition between octaves are formalized as:

$$\begin{aligned} y_{j+1}^r &= K * y_j^r, \qquad & \hat{y}_{j+1}^r &= K * \hat{y}_j^r, \qquad & \forall j \in J, r \in R\\ y_1^{r+1} &= M(y_{|J|}^r), \qquad & \hat{y}_1^{r+1} &= M(\hat{y}_{|J|}^r), \qquad & \forall r \in R\end{aligned} \tag{14}$$

with K being the Gaussian kernel, M the downsampling function, R and J the sets of resolution and smoothness levels, respectively. And $*$ denotes the convolution operation.

The losses for the translator G and discriminator D are then formulated as,

$$\begin{aligned}\mathcal{L}_G &= \mathcal{L}_G^{adv} + \lambda_{\mathcal{L}_1}\mathcal{L}_1\\ \mathcal{L}_D &= \mathcal{L}_D^{adv} + \lambda_{R_1} R_1\end{aligned} \tag{15}$$

where R_1 is the gradient penalty $\mathbb{E}_{\substack{x\sim\mathcal{X}\\ y\sim\mathcal{Y}}} \frac{1}{|S|}\sum_{s\in S}\left\|\nabla_y D(y^{(s)}|x)\right\|_2^2$ [26], and $\lambda_{\mathcal{L}_1}$ and λ_{R_1} are the weights for $\mathcal{L}_1$ loss and gradient penalty, respectively.

C Dataset

IMC profiling was performed using a panel of 40 antibodies, from which 10 have been selected for this study based on the biological function of the corresponding proteins as well as high signal–to–noise ratio. The proteins targeted by the 10 antibodies include cell-type markers, such as tumor markers (MelanA, gp100, S100), lymphocyte markers (CD20, CD16, CD3, CD8a) and endothelial marker (CD31). Moreover, two functional markers corresponding to proteins involved in antigen presentation (HLA-ABC, HLA-DR) are included in the protein set. The functional markers allow for differentiation between fine-grained cell states, beyond cell types. The raw IMC images have been processed with CellProfiler software for cell segmentation [27]. The protein counts extracted from the images have been first clipped to 99.9% per protein to exclude outliers ad then transformed using the $arcsinh$-function with cofactor one [28]. In order to exclude background noise, we apply OTSU thresholding [29] with kernel size three and sigma three and the threshold, separating signal from background, determined per sample using all available Region-of-Interests (RoIs). The resulting data per protein is first centered and standardized and then subjected to min-max-transformation, all using data statistics based on the train set only.

Table 2: Summary of hyperparameters for each method.

Method	Parameter	Value	Description
Pix2pix [19]	$\lambda_{\mathcal{L}_1}$	1	L_1 loss
	λ_{R_1}	1	R_1 regularization
PyramidP2P [7]	$\lambda_{\mathcal{L}_{\text{GP}}}$	1	Gaussian pyramid loss
	λ_{R_1}	1	R_1 regularization
Multi-V-Stain (*Ours*)	λ_{GP}	5	Gaussian pyramid loss
	λ_{ACP}	1	Adaptive perceptual loss
	λ_{R_1}	1	R_1 regularization

D Implementation Details

All methods are implemented using PyTorch [30] and trained with with random 1024×1024 crops for H&E and 256×256 for IMC images. The experiments are run on a single NVIDIA A100 GPU with a batch size of 16. We use the Adam optimizer [31] and a learning rate of 2.5×10^{-4} with for both the translator and discriminator. The hyperparameters for each method are summarized in Table 2.

E Evaluation Metrics

We conduct a quantitative evaluation of the generated IMC stains on both pixel and (pseudo-)cell levels. While the pixel-based metrics offer an initial assessment that aligns with visual inspections, the cell-level metrics provide a more nuanced evaluation, particularly useful for accounting for the slice-to-slice discrepancies between the H&E and GT IMC images. The two-level assessment obviates the need for exact pixel-level correspondence, thereby enhancing the robustness of our evaluation.

E.1 Pixel-based Evaluation

Pixel-level evaluations are conducted on individual proteins. For each protein, we calculate Multiscale Structural Similarity Index (MS-SSIM), Fréchet Inception Distance (FID) and Kernelized Inception Distance (KID) score between GT and predicted abundance across each RoI. While all three scores measure similarity, FID and KID measure the feature distribution similarity whereas MS-SSIM focuses on measuring the structural information and details, as perceived by the human visual system. For this reason, in some cases, there might be a discordance between the the evaluation metrics.

E.2 Co-expression Analysis

Co-expression analysis goes beyond pixel-level evaluation and is well-suited in our data setting with slice-to-slice discrepancies. It quantifies the correlation between two or more proteins that are simultaneously expressed within a specific region. This is of biological importance as co-expression of protein markers can be a significant variable for overall survival [32, 33].

It is based on cell-level read-outs that requires getting average expression for each cell by first detecting cell coordinates and then averaging the protein expression for an approximate cell diameter. To enable this, we use CellProfiler software to detect cell centroids for GT IMC. We then get the pseudo-cells by assuming a circle of diameter 10 µm around the cell centroids. Expression of each protein, one at a time, is then averaged across pixels belonging to a given pseudo-cell. For each pair, we calculate Spearman's Correlation Coefficient (SCC) between the two protein expression vectors across pseudo-cells in a given RoI, which we further refer to as *co-expression values*. Next, we repeat the above steps to get co-expression values for generated IMC by using the cell centroids detected by using HoVer-Net [34] on H&E images.

Once we have the co-expression values for both GT and generated IMC, for each protein pair, we calculate the normalized mean square error (NMSE) between them, averaged over each protein-pair. This comparison measures the extent to which the predicted multiplex recapitulates protein co-expression patterns.

F Additional Results

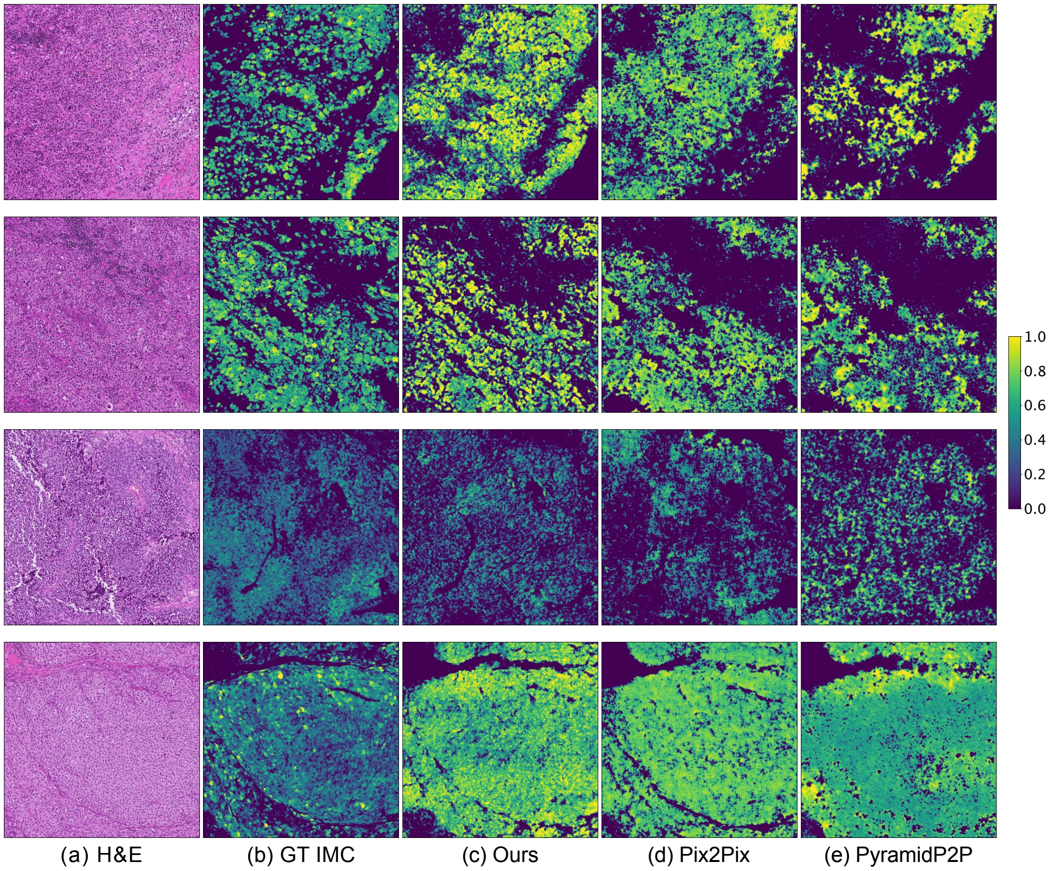
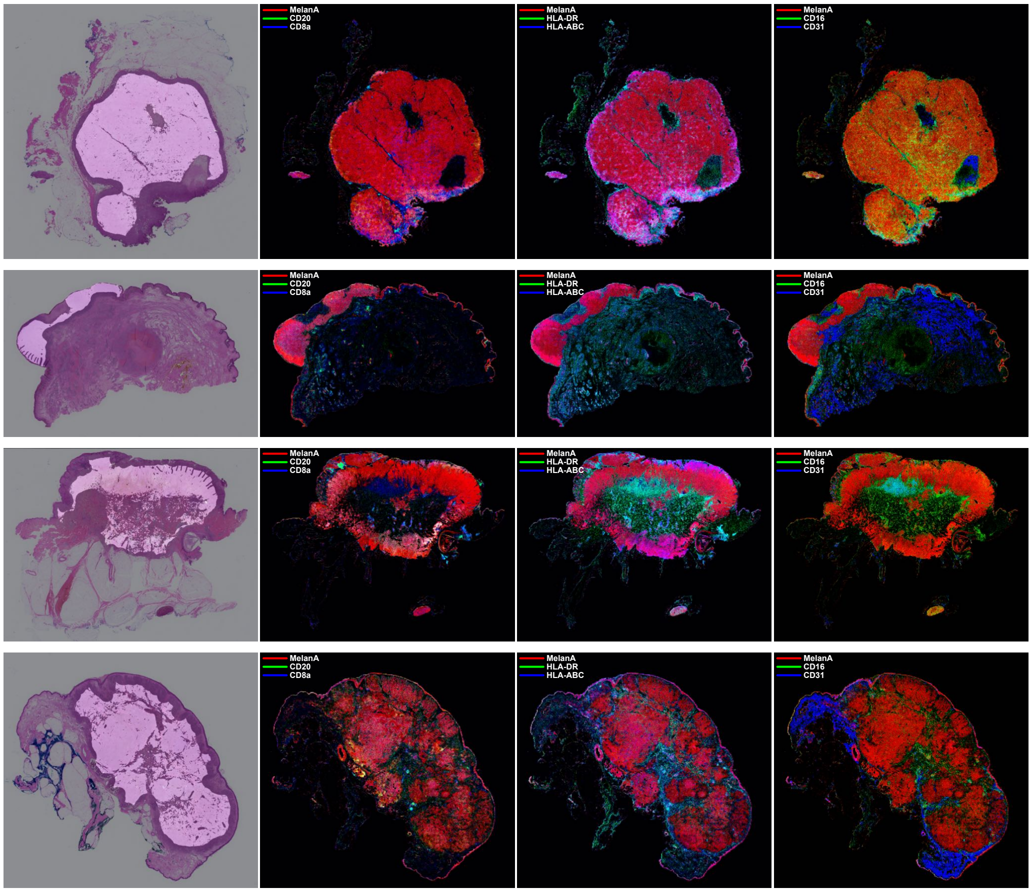


Figure 5: Qualitative assessment of Multi-V-Stain and baselines at RoI level. Rows represent four RoIs for S100 protein.



(a) H&E (b) Predicted IMC (c) Predicted IMC (d) Predicted IMC

Figure 6: Multiplexed IMC prediction at Whole-Slide Image (WSI) level on four immune cold samples. (a) H&E input with overlay of tumor center. (b-d) Predicted IMC markers visualized in RGB: (b) MelanA (tumor marker), CD20 (B cell marker), and CD8a (cytotoxic T cell marker); (c) MelanA, HLA-DR (interacts with CD4+ T cells), and HLA-ABC (interacts with CD8+ T cells); (d) MelanA, CD16 (NK cell marker), and CD31 (endothelial marker).