Evaluation of Virtual Stain Multiplexed CD68 for Macrophage Detection in NSCLC PD-L1 Slides

Elad Arbel ELAD.ARBEL@AGILENT.COM
Oded Ben-David
Itay Remer
Amir Ben-Dor
Daniela Rabkin
Agilent Research Laboratories, Tel Aviv, Israel

Sarit Aviel-Ronen
Adelson School of Medicine, Ariel University, Ariel, Israel

Frederik Aidt
Tine Hagedorn-Olsen
Lars Jacobsen
Agilent Technologies, Glostrup, Denmark

Kristopher Kersch
Jim Christian
Quyen Nguyen
Agilent Technologies, Carpinteria, US

Anya Tsalenko
Agilent Research Laboratories, Santa Clara, US

Editors: Under Review for MIDL 2025

Abstract

Manual reading of tissue slides by pathologists serves both as a foundation for clinical decision-making and as a source of ground truth for training artificial intelligence (AI) models. However, challenges such as inter-observer variability, limited tissue availability, and complex annotation tasks often compromise reliability and scalability. This study exemplifies a broader trend in pathology: leveraging virtual staining and other AI-based methodologies to address these challenges. We applied virtual stain multiplexing to a challenging annotation task - macrophage identification in non-small cell lung cancer tissue PD-L1 IHC stains, demonstrating its ability to improve pathologist performance and inter-observer agreement. In six challenging regions selected from 49 curated whole slide images, virtual staining significantly increased macrophage detection consistency, with Fleiss' kappa improving from -0.1 to 0.62, and enhanced overall accuracy, with the F1 score increasing from 0.13 to 0.65. These results highlight the potential use of AI-based virtual staining to assist pathologists reading slides, thereby improving consistency, enhancing accuracy, and alleviating the dependence on additional costly staining. Virtual stain multiplexing demonstrates a generalizable approach to improving pathologist performance through measurement-based AI tools, addressing broader needs for reproducibility and efficiency in diagnostic pathology.

Keywords: Immunohistochemistry, Virtual Stain, Multiplexing, Deep Learning, Macrophages, NSCLC, PD-L1, AI, Agreement, Pathology

1. Introduction

A central challenge in clinical pathology is the need to extract ever-increasing information from progressively smaller tissue samples (Hirsch et al., 2010). This challenge is amplified by the growing demand for personalized medicine (Hirsch et al., 2010), which often necessitates application of a broad range of immunohistochemical (IHC) stains onto multiple consecutive tissue sections from each patient sample. This process is time-consuming and constrained by limited tissue availability, especially in the case of small biopsies, and incurs significant costs in terms of labor and reagents. While research methodologies, such as fluorescent IHC multiplexing (Hoyt, 2021) or mass spectrometry imaging (Keren et al., 2019), offer potential solutions, they are not yet practical for routine diagnostic use.

Digital pathology provides alternative solutions to this challenge, such as virtual staining (VS) (Imran et al., 2024; Latonen et al., 2024), a novel approach leveraging generative adversarial networks (GANs) to infer staining patterns directly from existing hematoxylin and eosin (H&E)-stained tissue without the need for additional physical staining (Owkin, 2020; de Haan et al., 2021; Pati et al., 2024). This method has evolved to allow inference from unstained tissue (Rivenson et al., 2019; Zhang et al., 2020; Pradhan et al., 2021), including virtual HER2 IHC (Bai et al., 2022; Kapil et al., 2021).

Another critical challenge in pathology is low agreement between pathologists on some clinically relevant tasks(Robert et al., 2023; Han et al., 2022). Even for relatively straightforward tasks, such as scoring HER2 IHC 3+ specimens, inter-pathologist concordance is often suboptimal, potentially leading to inconsistent patient treatment decisions (Robbins et al., 2023). Addressing this issue, new quantitative measurement-based methods have been proposed to replace traditional subjective slide interpretations (Li et al., 2024; Aidt et al., 2025). It has been suggested that the deterministic nature of computational image-processing and AI-based methods can help alleviate subjectivity and improve reproducibility in these contexts (Baxi et al., 2022; Iyer et al., 2024; Pulaski et al., 2024).

However, the development of AI models requires reliable, large-scale ground truth data, which can be costly when derived from manual pathologist annotations. Low agreement among annotators often necessitates employing multiple pathologists per sample and using majority vote or consensus as ground truth, further compounding costs. A recent study comparing AI models for HER2 scoring showed that agreement among automated models was similar to agreement among pathologists(McKelvey et al., 2024), potentially due to the models being trained on subjective, manually scored data. Using measurement-based ground truth for training AI models could overcome these limitations and lead to more reliable diagnostic tools.

VS methods trained with measurement-based ground truth offer a potential solution, providing reliable information to improve pathologist accuracy and reproducibility. For instance, in PD-L1 IHC 22C3 pharmDx stained non-small cell lung cancer (NSCLC) tissues, macrophage detection presents a notable challenge, due to their morphological similarity to tumor cells (Paces et al., 2022; Tsao et al., 2018). Beck et al. (Beck et al., 2019) reported similarly low agreement for macrophage detection in urothelial carcinoma, with inter-observer correlation as low as 0.287. The PD-L1 IHC 22C3 pharmDx assay is FDA-approved for detecting PD-L1 protein expression in NSCLC tissue, scored using the tumor proportion score (TPS), which excludes immune cells such as macrophages from the calcula-

tion (Agilent Technologies, 2021). While combined staining of PD-L1 with a tumor marker (Mercier et al., 2023), or with a macrophage marker like CD68 could assist pathologists, it is not routinely available in clinical settings.

To address these challenges, we recently introduced a method for training AI-based virtual stain multiplexing (VSM) models using sequential IHC staining as measurement-based ground truth (Ben-David et al., 2024). In this study, we present an in-depth evaluation of this approach, demonstrating its efficacy in creating VSM CD68 to assist pathologists in detecting macrophages in tumor microenvironment of PD-L1 IHC 22C3 pharmDx-stained NSCLC samples. Our findings show that VSM significantly improves both pathologist performance and inter-pathologist agreement on this challenging task.

2. Methods

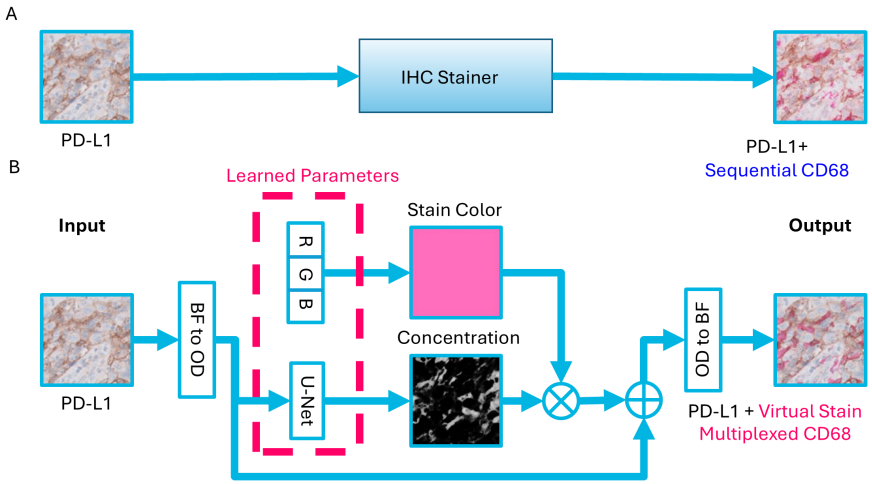

Figure 1: Workflow for Sequential and Virtual Staining Processes. (A) A PD-L1 IHC-stained slide sequentially stained with CD68 IHC to generate a PD-L1 and PD-L1+CD68 image pair. (B) A virtually stained multiplexed PD-L1 + CD68 image is generated from a PD-L1 IHC stained WSI.

Our VSM approach (Ben-David et al., 2024) involves training a neural network to simulate sequential stain multiplexing, enabling virtual multiplexing without physical restaining, as illustrated in Figure 1. The technique models the physical Beer-Lambert law, which governs the optical absorption of stained tissue in scanned histopathological images.

In the sequential multiplexing workflow, illustrated in Figure 1A, a PD-L1 immunohistochemistry (IHC) stained tissue undergoes a sequential staining with an additional CD68 IHC marker. The VSM workflow is illustrated in Figure 1B. A U-Net convolutional neural network (CNN) forms the backbone of our virtual stain generation process. The input to the U-Net is a patch extracted from the PD-L1 stained brightfield (BF) RGB image, converted to optical density (OD) space to align with the Beer-Lambert law. The network

predicts a concentration map of the CD68 stain, which is then multiplied by a learned RGB color vector, generating a virtual CD68 stain patch. To produce the final VSM image, the virtual CD68 patch is combined with the input PD-L1 patch in OD space and converted back to RGB space, resulting in a synthetically stained PD-L1 + virtual CD68 image pair.

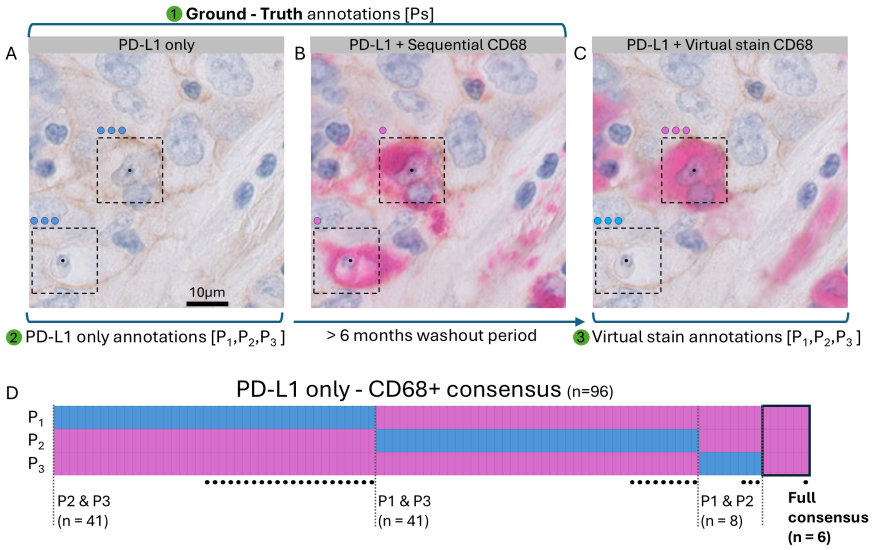

Figure 2: Cell-level annotations, represented by dots on top of squares marking sample cells, with CD68-, CD68+, and unknown classes colored blue, magenta, and black, respectively. Annotations of (A) PD-L1-only WSIs and (C) PD-L1 + VSM CD68, by P1-P3 (B) Ground-truth by the study pathologist, using PD-L1 + sequentially stained CD68 IHC. (D) Agreement analysis of consensus annotated CD68+ cells using PD-L1-only WSIs. Black circles indicate cells with CD68+ ground truth annotations.

The dataset used for training and evaluation is detailed in (Ben-David et al., 2024). Briefly, it consists of 49 curated NSCLC whole-slide images (WSIs). These WSIs were selected to include cases with PD-L1 expression levels near the clinically relevant TPS thresholds. Each WSI was stained using the PD-L1 IHC 22C3 assay, scanned, and subsequently restained with the CD68 PG-M1 (GA613) marker to generate sequentially stained PD-L1 + CD68 WSI pairs. Both WSIs in each pair were aligned to enable pixel-accurate training, with the sequentially stained PD-L1 + CD68 WSI serving as a measurement-based ground truth.

To focus the model on clinically relevant areas, a study pathologist, Ps, annotated tumor regions on each sequentially stained WSI pair for use during model training. In addition, six particularly challenging regions of interest (ROIs, Appendix D) were selected for cell-level evaluation based on the difficulty of distinguishing macrophages from tumor cells or the presence of macrophages infiltrating PD-L1-positive tumor areas, making them critical for assessing the efficacy of the proposed VSM method.

Our purpose was to study the efficacy of the VSM method rather than the performance of a specific model. Considering the challenging nature of the regions chosen for evaluation, we adopted a leave-tissue-out methodology. A separate model was trained for each evaluation region, excluding the entire WSI containing that region from the training data. This prevented any tissue-specific overfitting, while maximizing overall training set diversity.

The training hyperparameters were kept consistent across all models, using the same settings reported in the ablation tests section of (Ben-David et al., 2024), ensuring comparability across results.

As previously reported (Ben-David et al., 2024), inspection of model prediction by Ps revealed the VSM patterns follow closely true sequential staining. However, to quantitatively evaluate macrophage detection performance, the study pathologist provided cell-level annotations for each of the six evaluation regions, marking cells as CD68-positive (CD68+) or CD68-negative (CD68-) based on the paired and aligned PD-L1 and sequentially stained PD-L1 + CD68 WSIs (Figure 2B). These annotations served as the ground-truth reference for model evaluation. Cells where the CD68 staining was insufficient for a definitive classification were annotated as "unknown."

Three additional certified pathologists, P1-P3, independently performed two rounds of cell-level annotation for each region. In the first round, they annotated cells using only the PD-L1-stained WSI (Figure 2A). After a washout period of more than six months to minimize recall bias, they repeated the annotation task using pairs of PD-L1 and VSM CD68 WSIs (Figure 2C).

Since the task was defined as macrophage detection task, cells marked as "unknown" by P1-P3 were treated as CD68-, as the pathologists did not positively identify them as macrophages. To assess the robustness of this assumption, an alternative analysis was conducted (Appendix A) where unknown cells were instead treated as CD68+, yielding qualitatively similar results. Cells annotated as "unknown" by the study pathologist were excluded from macrophage detection performance metrics. An alternative analysis including these cells (set as either CD68+ or CD68-) is provided in Appendix A.

Inter-pathologist agreement was measured using Fleiss's Kappa for overall agreement across the three annotators and Cohen's Kappa for pairwise agreement between each pathologists pair (Fleiss and Cohen, 1973). Performance metrics were computed for each of P1-P3, as well as consensus labels derived from majority voting among P1-P3. Given the class imbalance present in the dataset, F1 scores were used as the primary performance measure, supplemented by precision, recall, and specificity.

3. Experimental Results

The analysis included a total of 2,304 annotated cells, of which 457 were marked as CD68 positive in the ground truth annotations provided by the study pathologist. Figure 2D summarizes the predictions of pathologists P1-P3 for a subset of 96 cells where the consensus PD-L1-only annotation indicated CD68 positivity. Among these 96 cells, only 35 were confirmed as macrophages in the ground truth annotations by the study pathologist, as indicated by dots in the figure. Notably, full consensus among all three pathologists was reached for only six cells, and of these, just one was confirmed as a macrophage by the ground truth; four cells were annotated as CD68-, and one classified as "unknown".

Table 1: Inter-pathologist agreement measured by Fleiss' and Cohen's kappa for PD-L1 and virtual staining multiplexing (VSM).

Method	Fleiss'	Cohen's		
	P1 vs. P2 vs. P3	P1 vs. P2	P1 vs. P3	P2 vs. P3
PD-L1	-0.11	0.07	-0.03	0.01
VSM	0.62	0.79	0.51	0.58

Table 1 presents Fleiss's and Cohen's kappa measures of inter-rater agreement for P1-P3, both with and without the assistance of VSM CD68. Agreement was significantly improved with the aid of VSM: Fleiss' kappa increased from -0.1 (indicating extremely poor agreement) with PD-L1-only annotations to 0.62 (substantial agreement) with VSM.

The number of cells with full consensus annotations as CD68+ also increased markedly, rising to 463 with the use of VSM. Even the pair of P1 and P2, who demonstrated the highest agreement among the pairings, showed only slight agreement with PD-L1-only annotations (Cohen's kappa = 0.07). With VSM, all pairwise agreements improved to at least moderate levels (Cohen's kappa > 0.5).

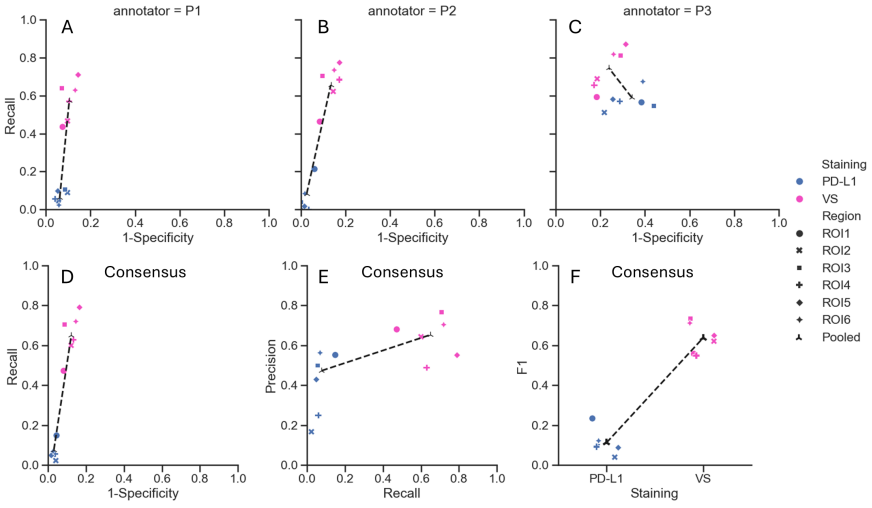

Figure 3: Performance metrics across six evaluation regions with and without assistive virtual staining multiplexing. Changes in pooled performance metrics are indicated by a dashed black line. (A-C) Receiver operating characteristic (ROC) plots for P1-P3. (D) ROC plot for consensus annotations. (E) Precision-recall plot for consensus annotations. (F) F1 scores for consensus annotations.

The effectiveness of pathologists in detecting macrophages was further analyzed using VSM CD68. Figure 3 provides detailed comparisons of performance with and without VSM.

Figure 3A–C present receiver operating characteristic (ROC) plots for P1-P3, displaying recall and 1-specificity across six evaluation regions.

Without the assistance of VSM, P1 and P2 exhibited low recall (0.06, 0.08) and high specificity (0.94, 0.97) across all evaluation regions, while P3 achieved higher recall (0.59) but at the expense of reduced specificity (0.66). The introduction of VSM substantially enhanced recall for P1 and P2 (0.58, 0.65), albeit with some decreases in specificity (0.9, 0.87). For P3, both recall (0.75) and specificity (0.76) improved across all evaluation regions with the aid of VSM.

Figure 3D–F focuses on the performance of consensus annotations. Figure 3D illustrates the ROC plot for consensus annotations, showing a similar trend to the individual performances of P1 and P2, with low recall (0.077) and high specificity (0.97) in the absence of VSM. VSM led to much higher recall (0.65), accompanied by a decrease in specificity (0.88). Figure 3E depicts precision-recall curves, highlighting notably improved pooled recall (0.65) and precision (0.65), with precision increases of 2–3 times for certain regions. Finally, Figure 3F summarizes F1 scores, which consistently improved with VSM. The pooled F1 score for consensus annotations showed a five-fold improvement with the aid of VSM, rising from 0.13 to 0.65.

Comparing majority-vote consensus annotations to full-consensus annotations revealed additional insights. Without VSM, full-consensus performance was notably low, with an F1 score of 0.012, recall of 0.006, and precision of 0.2. VSM significantly improved these metrics, raising them to 0.71, 0.72, and 0.69, respectively. Appendix B provides a comprehensive table of evaluation metrics for all annotators and regions, with and without the use of VSM.

4. Conclusion

In this study, we presented a comprehensive evaluation of VSM for CD68 to assist pathologists in detecting macrophages within the tumor microenvironment of PD-L1-stained NSCLC slides.

Unassisted macrophage detection showed agreement levels even lower than those observed in clinical scoring tasks such as HER2 and PD-L1, where inter-pathologist agreement is already below ideal thresholds(Robbins et al., 2023; Han et al., 2022). Unlike such scoring tasks, where pathologists' annotations often serve as the only ground truth, our study leveraged sequentially stained CD68 IHC to provide an independent and more objective ground truth. Comparisons to this ground truth revealed consistently poor unassisted performance across the curated challenging regions.

The analysis of individual pathologists' annotations further revealed notable variability. Two pathologists exhibited similar overall performance metrics, while the third displayed distinct tendencies, potentially reflecting differences in training or decision-making styles (see Appendix C for annotation distributions). However, even the two pathologists with similar overall metrics showed low agreement on a per-cell basis, as evidenced by their low Cohen's kappa values. These results suggest that individual variability remains a challenge, even among pathologists with ostensibly similar capabilities.

Macrophage detection is explicitly required in clinical TPS assessments to exclude macrophages from scoring. The poor unassisted performance observed here suggests that

macrophage misclassification could affect TPS scoring in challenging cases, underscoring the need for improved tools to assist in these scenarios.

The introduction of VSM CD68 markedly improved pathologists' performance and agreement. Detection errors with VSM are attributable in part to model imperfections, as suggested by the increased agreement between pathologists and anecdotal fault analysis of model predictions (Figure 1). These results indicate that further refinement of the model would be beneficial to reduce such errors. Additionally, with VSM, the annotation statistics of all pathologists became more closely aligned, demonstrating a reduction in inter-pathologist variability.

Consensus annotations provide further insight into the difficulty of macrophage detection. Without VSM, majority-vote consensus annotations failed to mitigate the challenges of unassisted detection. Full-consensus annotations performed even worse, with only six macrophages identified across the regions. However, both majority-vote and full-consensus annotations improved substantially with VSM.

Our findings suggest that tasks with poor agreement among pathologists, such as unassisted macrophage detection, may pose a challenge for the development of AI-based diagnostic tools. Manual annotations from individual pathologists may not provide reliable ground truth in these contexts. Prior studies (Robbins et al., 2023; Han et al., 2022) have shown that requiring agreement among multiple pathologists might reduce the dataset size significantly. This aligns with our observation that without the assistance of VSM the number of full-consensus annotations was drastically limited. Furthermore, when compared to sequential-stain based ground truth, full consensus was found to perform worse than majority vote for unassisted macrophage detection. These results suggest that increasing the number of annotators or requiring unanimity does not address the fundamental challenges of low-agreement tasks.

The scarcity of medical data poses additional challenges for method evaluation. To mitigate overfitting, we employed a leave-tissue-out strategy, training separate models for each tissue in our evaluation set. We note that comparison with our previously reported model (Ben-David et al., 2024), which likely exhibited overfitting due to within-tissue evaluation, reveals a decrease in performance metrics with the current approach. However, the qualitative conclusions remained consistent, indicating that within-tissue evaluation may be sufficient for exploratory, proof-of-concept studies, particularly when time or computational resources are limited.

This in-depth evaluation demonstrates that VSM CD68 significantly improves pathologists' ability to detect macrophages. This capability is particularly important for borderline PD-L1 slide scoring and may offer opportunities for spatial analysis of macrophages in the tumor microenvironment. Additionally, our work demonstrates that VS methods trained using measurement-based ground truth, can effectively address the challenges of low inter-pathologist agreement and the need for extensive ground-truth data, enabling robust performance with minimal reliance on manual annotation.

Regulatory statement: This study is for proof of concept only. This paper does not imply any clinical functionality.

References

Agilent Technologies. PD-L1 IHC 22C3 pharmDx interpretation manual, NSCLC. https://www.agilent.com/cs/library/usermanuals/public/29158_pd-l1-ihc-22C3-pharmdx-nsclc-interpretation-manual.pdf, 2021. Last accessed 2024/09/08.

Frederik Aidt, Elad Arbel, Itay Remer, Oded Ben-David, Amir Ben-Dor, Daniela Rabkin, Kirsten Hoff, Karin Salomon, Sarit Aviel-Ronen, Gitte Nielsen, Jens Mollerup, Lars Jacobsen, and Anya Tsalenko. Quantification of her2-low and ultra-low expression in breast cancer specimens by quantitative ihc and artificial intelligence, 2025. URL https://www.biorxiv.org/content/early/2025/01/25/2025.01.22.634230.

B. Bai, H. Wang, Y. Li, K. de Haan, F. Colonnese, Y. Wan, J. Zuo, N.B. Doan, X. Zhang, Y. Zhang, and J. Li. Label-free virtual HER2 immunohistochemical staining of breast tissue using deep learning. *BME Frontiers*, 2022.

Vipul Baxi, George Lee, Chunzhe Duan, Dimple Pandya, Daniel N Cohen, Robin Edwards, Han Chang, Jun Li, Hunter Elliott, Harsha Pokkalla, et al. Association of artificial intelligence-powered and manual quantification of programmed death-ligand 1 (pd-l1) expression with outcomes in patients treated with nivolumab±ipilimumab. *Modern Pathology*, 35(11):1529–1539, 2022.

A. Beck, B. Glass, H. Elliott, J.K. Kerner, A. Khosla, A. Lahiri, H. Pokkalla, D. Wang, I. Wapinski, G. Lee, and V. Baxi. P730 an empirical framework for validating artificial intelligence–derived PD-L1 positivity predictions applied to urothelial carcinoma. In *Cancer cells*, volume 3728, pages 18–056, Unknown, 2019.

Oded Ben-David, Elad Arbel, Sarit Aviel-Ronen, Frederik Heurlin Aidt, Kristopher Kersch, Tine Hagedorn Olsen, Daniela Rabkin, Itay Remer, Amir Ben-Dor, Lars Jacobsen, et al. Deep-learning based virtual stain multiplexing immunohistochemistry slides–a pilot study. In *MICCAI Workshop on Computational Pathology with Multimodal Data (COMPAYL)*, 2024.

K. de Haan, Y. Zhang, J.E. Zuckerman, T. Liu, A.E. Sisk, M.F. Diaz, K.Y. Jen, A. Nobori, S. Liou, S. Zhang, and R. Riahi. Deep learning-based transformation of h&e stained tissues into special stains. *Nature communications*, 12(1):4884, 2021.

Joseph L Fleiss and Jacob Cohen. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and psychological measurement*, 33(3):613–619, 1973.

Gang Han, Michael J Schell, Emily S Reisenbichler, Bohong Guo, and David L Rimm. Determination of the number of observers needed to evaluate a subjective test and its application in two pd-l1 studies. *Statistics in medicine*, 41(8):1361–1375, 2022.

Fred R Hirsch, Murry W Wynes, David R Gandara, and Paul A Bunn Jr. The tissue is the issue: personalized medicine for non-small cell lung cancer. *Clinical Cancer Research*, 16 (20):4909–4911, 2010.

Clifford C. Hoyt. Multiplex immunofluorescence and multispectral imaging: Forming the basis of a clinical test platform for immuno-oncology. *Frontiers in Molecular Biosciences*, 8, 2021. doi: 10.3389/fmolb.2021.674747. URL https://www.frontiersin.org/journals/molecular-biosciences/articles/10.3389/fmolb.2021.674747.

Muhammad Talha Imran, Imran Shafi, Jamil Ahmad, Muhammad Fasih Uddin Butt, Santos Gracia Villar, Eduardo Garcia Villena, Tahir Khurshaid, and Imran Ashraf. Virtual histopathology methods in medical imaging-a systematic review. *BMC medical imaging*, 24(1):318, 2024.

Janani S Iyer, Dinkar Juyal, Quang Le, Zahil Shanis, Harsha Pokkalla, Maryam Pouryahya, Aryan Pedawi, S Adam Stanford-Moore, Charles Biddle-Snead, Oscar Carrasco-Zevallos, et al. Ai-based automation of enrollment criteria and endpoint assessment in clinical trials in liver diseases. *Nature Medicine*, 30(10):2914–2923, 2024.

Ansh Kapil, Armin Meier, Keith Steele, Marlon Rebelatto, Katharina Nekolla, Alexander Haragan, Abraham Silva, Aleksandra Zuraw, Craig Barker, Marietta L Scott, et al. Domain adaptation-based deep learning for automated tumor cell (tc) scoring and survival analysis on pd-l1 stained tissue images. *IEEE Transactions on Medical Imaging*, 40(9): 2513–2523, 2021.

Leeat Keren, Marc Bosse, Steve Thompson, Tyler Risom, Kausalia Vijayaragavan, Erin McCaffrey, Diana Marquez, Roshan Angoshtari, Noah F Greenwald, Harris Fienberg, et al. Mibi-tof: A multiplexed imaging platform relates cellular phenotypes and tissue structure. *Science advances*, 5(10):eaax5851, 2019.

Leena Latonen, Sonja Koivukoski, Umair Khan, and Pekka Ruusuvuori. Virtual staining for histology by deep learning. *Trends in Biotechnology*, 2024.

Xiaoxian Li, Ji-Hoon Lee, Yuan Gao, Jilun Zhang, Katherine M Bates, David L Rimm, Huina Zhang, Geoffrey Hughes Smith, Diane Lawson, Jane Meisel, et al. Correlation of her2 protein level with mrna level quantified by rnascope in breast cancer. *Modern Pathology*, 37(2):100408, 2024.

Brittany McKelvey, Pedro A. Torres-Saavedra, Jessica Li, Glenn Broeckx, Frederik Deman, Siraj Ali, Hillary Andrews, Salim Arslan, Santhosh Balasubramanian, J. Carl Barrett, Peter Caie, Ming Chen, Daniel Cohen, Tathagata Dasgupta, Brandon Gallas, George Green, Mark Gustavson, Sarah Hersey, Ana Hidalgo-Sastre, Shahanawaz Jiwani, Wonkyung Jung, Kimary Kulig, Vladimir Kushnarev, Xiaoxian Li, Meredith Lodge, Joan Mancuso, Mike Montalto, Satabhisa Mukhopadhyay, Matthew Oberley, Pahini Pandya, Oscar Puig, Edward Richardson, Alexander Sarachakov, Or Shaked, Mark Stewart, Lisa M. McShane, Roberto Salgado, and Jeff Allen. Agreement across 10 artificial intelligence models in assessing her2 in breast cancer whole slide images:findings from the friends of cancer research digital path project. In *San Antonio Breast Cancer Symposium*, San Antonio, Texas, Dec. 10-13 2024. URL https://friendsofcancerresearch.org/wp-content/uploads/SABCS-Poster_Digital-PATH-Project-Findings.pdf.

Anaïs Mercier, Virginie Conan-Charlet, Isabelle Quintin-Roué, Laurent Doucet, Pascale Marcorelles, and Arnaud Uguen. Reproducibility in pd-l1 immunohistochemistry quantification through the tumor proportion score and the combined positive score: Could dual immunostaining help pathologists? *Cancers*, 15(10):2768, 2023.

Owkin. Developing virtual staining at Owkin to enrich digital pathology. https://www.owkin.com/blogs-case-studies/developing-virtual-staining-at-owkin-to-enrich-digital-pathology, 2020. Last accessed 2024/09/08.

W. Paces, E. Ergon, E. Bueche, G.D. Young, V. Adisetiyo, C. Luengo, M. James, C. Caldwell, D. Miller, M. Wambaugh, and G. Metcalf. A digital assay for programmed death-ligand 1 (22C3) quantification combined with immune cell recognition algorithms in non-small cell lung cancer. *Scientific Reports*, 12(1):9745, 2022.

Pushpak Pati, Sofia Karkampouna, Francesco Bonollo, Eva Compérat, Martina Radić, Martin Spahn, Adriano Martinelli, Martin Wartenberg, Marianna Kruithof-de Julio, and Marianna Rapsomaniki. Accelerating histopathology workflows with generative ai-based virtually multiplexed tumour profiling. *Nature Machine Intelligence*, 6(9):1077–1093, 2024.

P. Pradhan, T. Meyer, M. Vieth, A. Stallmach, M. Waldner, M. Schmitt, J. Popp, and T. Bocklitz. Computational tissue staining of non-linear multimodal imaging using supervised and unsupervised deep learning. *Biomedical Optics Express*, 12(4):2280–2298, 2021.

Hanna Pulaski, Stephen A Harrison, Shraddha S Mehta, Arun J Sanyal, Marlena C Vitali, Laryssa C Manigat, Hypatia Hou, Susan P Madasu Christudoss, Sara M Hoffman, Adam Stanford-Moore, et al. Clinical validation of an ai-based pathology tool for scoring of metabolic dysfunction-associated steatohepatitis. *Nature Medicine*, pages 1–8, 2024.

Y. Rivenson, H. Wang, Z. Wei, K. de Haan, Y. Zhang, Y. Wu, H. Günaydın, J.E. Zuckerman, T. Chong, A.E. Sisk, and L.M. Westbrook. Virtual histological staining of unlabeled tissue-autofluorescence images via deep learning. *Nature Biomedical Engineering*, 3(6): 466–477, 2019.

Charles J Robbins, Aileen I Fernandez, Gang Han, Serena Wong, Malini Harigopal, Mirna Podoll, Kamaljeet Singh, Amy Ly, M Gabriela Kuba, Hannah Wen, et al. Multi-institutional assessment of pathologist scoring her2 immunohistochemistry. *Modern Pathology*, 36(1):100032, 2023.

Marie E Robert, Josef Rüschoff, Bharat Jasani, Rondell P Graham, Sunil S Badve, Manuel Rodriguez-Justo, Liudmila L Kodach, Amitabh Srivastava, Hanlin L Wang, Laura H Tang, et al. High interobserver variability among pathologists using combined positive score to evaluate pd-l1 expression in gastric, gastroesophageal junction, and esophageal adenocarcinoma. *Modern Pathology*, 36(5):100154, 2023.

Ming Sound Tsao, Keith M Kerr, Mark Kockx, Mary-Beth Beasley, Alain C Borczuk, Johan Botling, Lukas Bubendorf, Lucian Chirieac, Gang Chen, Teh-Ying Chou, et al.

Pd-l1 immunohistochemistry comparability study in real-life clinical samples: results of blueprint phase 2 project. *Journal of Thoracic Oncology*, 13(9):1302–1311, 2018.

Y. Zhang, K. de Haan, Y. Rivenson, J. Li, A. Delis, and A. Ozcan. Digital synthesis of histological stains using micro-structured and multiplexed virtual staining of label-free tissue. *Light: Science & Applications*, 9(1):78, 2020.

Appendix A. Alternative Analysis

This section contains analysis results with alternative choices for the handling of the "unknown" classes. Table 2 and Figure 4 present results where "unknown" annotations by P1-P3 were set to CD68+.

Table 2: Alternative analysis of inter-pathologist agreement, with "unknown" cells set to CD68+, measured by Fleiss' and Cohen's kappa for PD-L1 and virtual stain multiplexing (VSM).

Method	**Fleiss'**	**Cohen's**		
	P1 vs. P2 vs. P3	P1 vs. P2	P1 vs. P3	P2 vs. P3
PD-L1	0.05	0.14	0.23	0.03
VSM	0.54	0.60	0.56	0.48

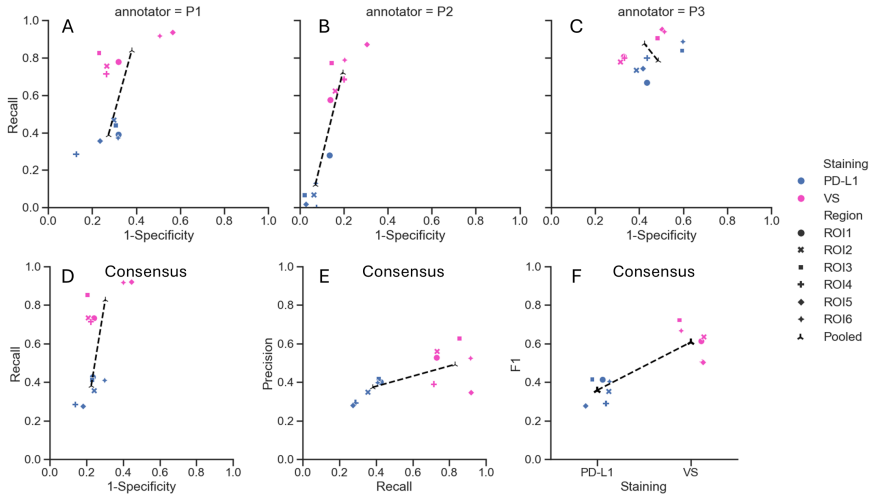

Figure 4: Performance metrics with P1-P3 "unknown" annotations set to CD68+ (A-C) Receiver operating characteristic (ROC) plots for P1-P3. (D) ROC plot for consensus annotations. (E) Precision-recall plot for consensus annotations. (F) F1 scores for consensus annotations.

Figure 5 and Figure 6 present performance metrics where Ps "unknown" annotations were not excluded from the analysis. Instead, both Ps "unknown" annotations and P1-P3 "unknown" annotations were set to CD68- or CD68+, respectively.

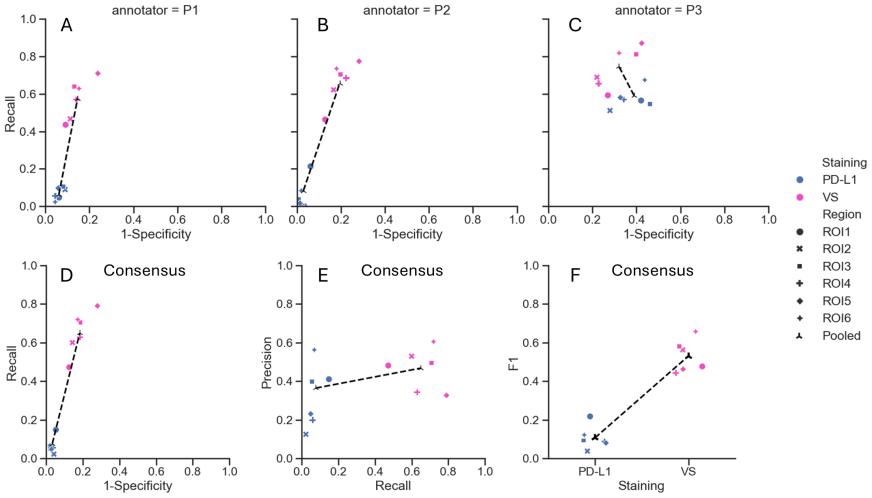

Figure 5: Performance metrics with all "unknown" annotations set to CD68-, Ps annotated cells not omitted from analysis as in main text. (A-C) Receiver operating characteristic (ROC) plots for P1-P3. (D) ROC plot for consensus annotations. (E) Precision-recall plot for consensus annotations. (F) F1 scores for consensus annotations.

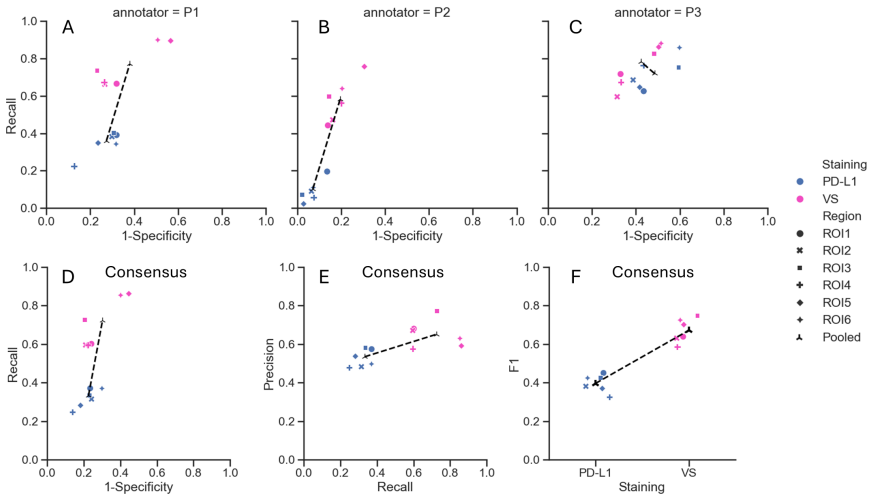

Figure 6: Performance metrics with all "unknown" annotations set to CD68+, Ps annotated cells not omitted from analysis as in main text. (A-C) Receiver operating characteristic (ROC) plots for P1-P3. (D) ROC plot for consensus annotations. (E) Precision-recall plot for consensus annotations. (F) F1 scores for consensus annotations.

Appendix B. Performance Metrics

The below tables present pooled performance metrics as well as metrics for each region separately. Each table contains metrics for consensus annotation and each of P1-P3, with and without the assistance of virtual stain multiplexing.

Table 3: Pooled performance metrics

annotator	Staining	Region	Precision	Recall	F1	Specificity
Consensus	PD-L1	Pooled	0.47	0.08	0.13	0.97
Consensus	VSM	Pooled	0.65	0.65	0.65	0.88
P1	PD-L1	Pooled	0.26	0.06	0.1	0.94
P1	VSM	Pooled	0.66	0.58	0.61	0.9
P2	PD-L1	Pooled	0.53	0.08	0.14	0.97
P2	VSM	Pooled	0.63	0.66	0.64	0.87
P3	PD-L1	Pooled	0.38	0.59	0.46	0.66
P3	VSM	Pooled	0.53	0.75	0.62	0.76

Table 4: Performance metrics for ROI1

annotator	Staining	Region	Precision	Recall	F1	Specificity
Consensus	PD-L1	ROI1	0.55	0.15	0.23	0.96
Consensus	VSM	ROI1	0.68	0.47	0.56	0.92
P1	PD-L1	ROI1	0.24	0.05	0.08	0.95
P1	VSM	ROI1	0.68	0.44	0.53	0.93
P2	PD-L1	ROI1	0.56	0.21	0.31	0.94
P2	VSM	ROI1	0.67	0.46	0.55	0.91
P3	PD-L1	ROI1	0.35	0.56	0.43	0.62
P3	VSM	ROI1	0.54	0.59	0.57	0.82

Table 5: Performance metrics for ROI2

annotator	Staining	Region	Precision	Recall	F1	Specificity
Consensus	PD-L1	ROI2	0.17	0.02	0.04	0.96
Consensus	VSM	ROI2	0.64	0.6	0.62	0.88
P1	PD-L1	ROI2	0.25	0.09	0.13	0.9
P1	VSM	ROI2	0.64	0.47	0.54	0.9
P2	PD-L1	ROI2	0.0	0.0	0.0	0.99
P2	VSM	ROI2	0.61	0.62	0.62	0.85
P3	PD-L1	ROI2	0.46	0.51	0.48	0.78
P3	VSM	ROI2	0.57	0.69	0.63	0.81

Table 6: Performance metrics for ROI3

annotator	Staining	Region	Precision	Recall	F1	Specificity
Consensus	PD-L1	ROI3	0.5	0.05	0.1	0.98
Consensus	VSM	ROI3	0.77	0.71	0.74	0.91
P1	PD-L1	ROI3	0.33	0.11	0.16	0.91
P1	VSM	ROI3	0.79	0.64	0.71	0.93
P2	PD-L1	ROI3	1.0	0.04	0.08	1.0
P2	VSM	ROI3	0.75	0.71	0.73	0.9
P3	PD-L1	ROI3	0.33	0.55	0.41	0.56
P3	VSM	ROI3	0.53	0.81	0.64	0.71

Table 7: Performance metrics for ROI4

annotator	Staining	Region	Precision	Recall	F1	Specificity
Consensus	PD-L1	ROI4	0.25	0.06	0.09	0.97
Consensus	VSM	ROI4	0.49	0.63	0.55	0.87
P1	PD-L1	ROI4	0.22	0.06	0.09	0.96
P1	VSM	ROI4	0.53	0.57	0.55	0.9
P2	PD-L1	ROI4	0.0	0.0	0.0	0.97
P2	VSM	ROI4	0.44	0.69	0.54	0.83
P3	PD-L1	ROI4	0.29	0.57	0.38	0.71
P3	VSM	ROI4	0.43	0.66	0.52	0.83

Table 8: Performance metrics for ROI5

annotator	Staining	Region	Precision	Recall	F1	Specificity
Consensus	PD-L1	ROI5	0.43	0.05	0.09	0.98
Consensus	VSM	ROI5	0.55	0.79	0.65	0.83
P1	PD-L1	ROI5	0.32	0.1	0.15	0.95
P1	VSM	ROI5	0.56	0.71	0.62	0.86
P2	PD-L1	ROI5	0.2	0.02	0.03	0.98
P2	VSM	ROI5	0.53	0.77	0.63	0.83
P3	PD-L1	ROI5	0.37	0.58	0.45	0.74
P3	VSM	ROI5	0.42	0.87	0.56	0.69

Table 9: Performance metrics for ROI6

annotator	Staining	Region	Precision	Recall	F1	Specificity
Consensus	PD-L1	ROI6	0.56	0.07	0.12	0.97
Consensus	VSM	ROI6	0.7	0.72	0.71	0.85
P1	PD-L1	ROI6	0.16	0.02	0.04	0.94
P1	VSM	ROI6	0.7	0.63	0.66	0.87
P2	PD-L1	ROI6	0.69	0.08	0.15	0.98
P2	VSM	ROI6	0.7	0.73	0.72	0.85
P3	PD-L1	ROI6	0.45	0.67	0.54	0.61
P3	VSM	ROI6	0.6	0.82	0.69	0.74

Appendix C. Annotation Distributions

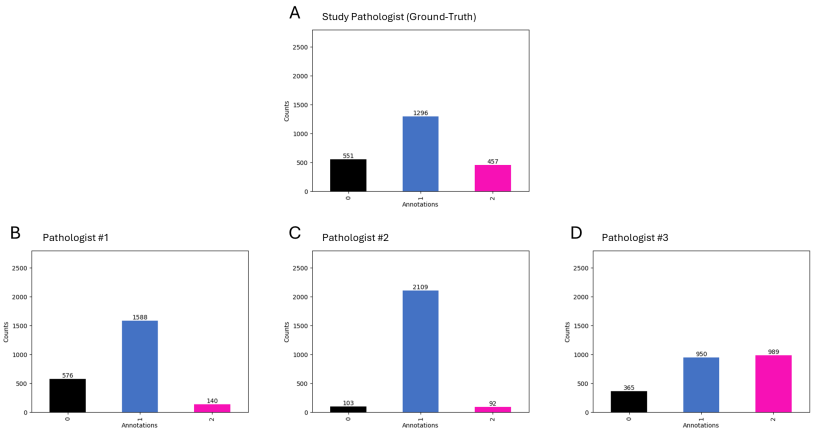

Figure 7: Annotation Distributions. Panel A presents annotations distributions for the study pathologist, using both PD-L1 and PD-L1 + sequential CD68 staining (used as ground-truth annotations for performance analysis). Distributions of P1–P3 annotations given in panels B-D, reveal notable variability. Pathologist P1 and P2 exhibit somewhat similar annotation distributions, seen in panels B and C, respectively (although agreeing on only 14 cells). Pathologist P3 annotation distribution given in panel D is starkly different, potentially reflecting differences in training or decision-making styles.

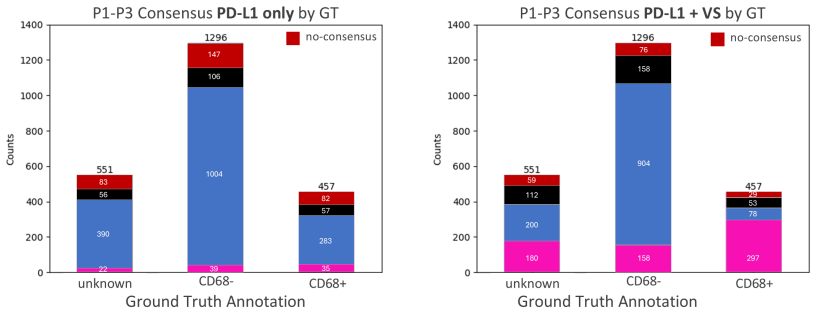

Figure 8: Consensus Annotation Distributions. Stacked consensus annotations by Ps ground-truth annotations. (left) P1-P3 consensus annotations using only PD-L1 stained slides. (right) P1-P3 consensus annotations with the addition of VSM CD68. The increase in overall CD68+ annotations by P1-P3 when assisted by VSM is apparent. In addition, the number of cells where no consensus was attained decreased with the addition of VSM.

Appendix D. Evaluation Regions

The below figures present the ROIs used for macrophage detection evaluation. For each ROI, we give the PD-L1 only, PD-L1 + VSM CD68 and PD-L1 + sequential CD68 IHC.

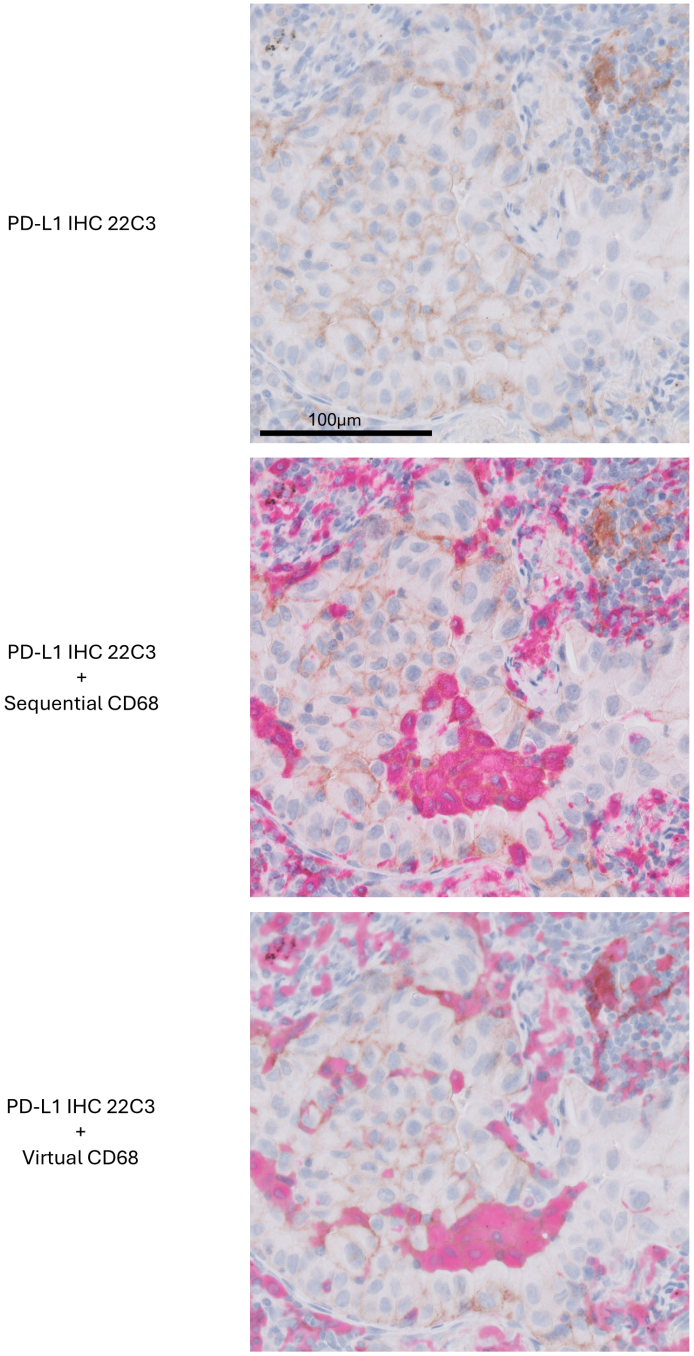

Figure 9: ROI1

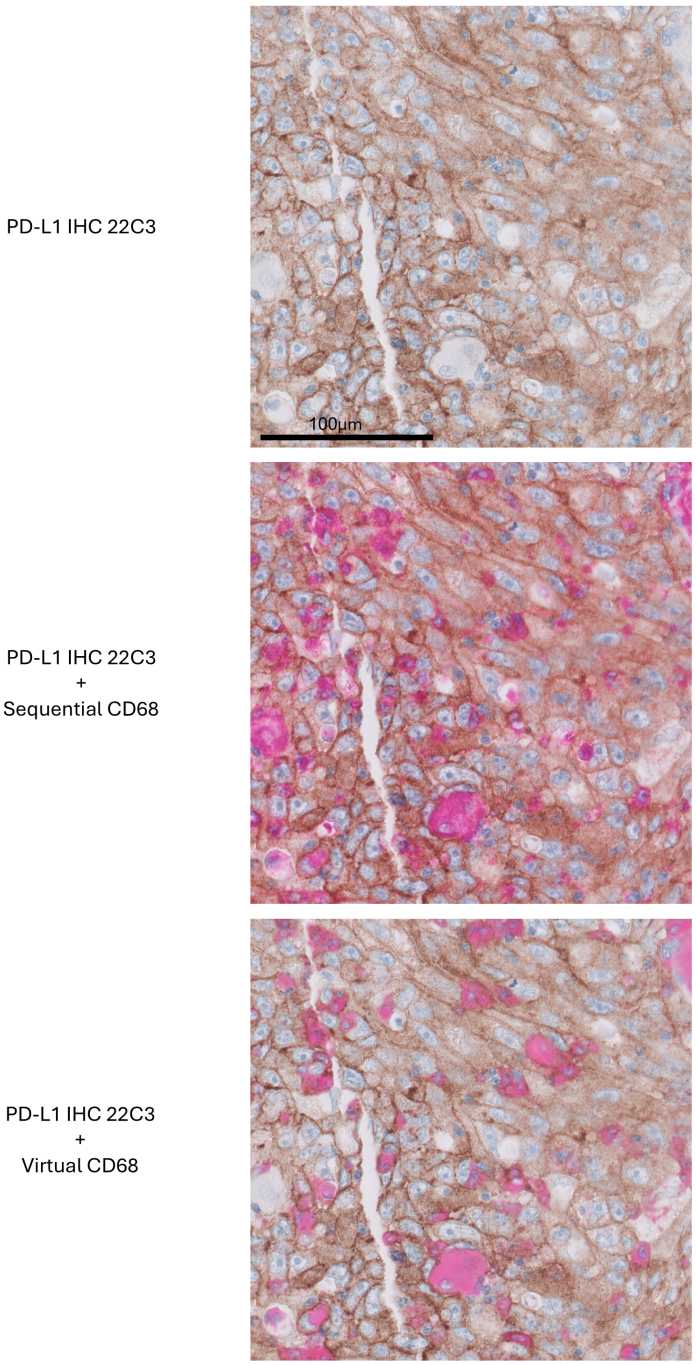

Figure 10: ROI2

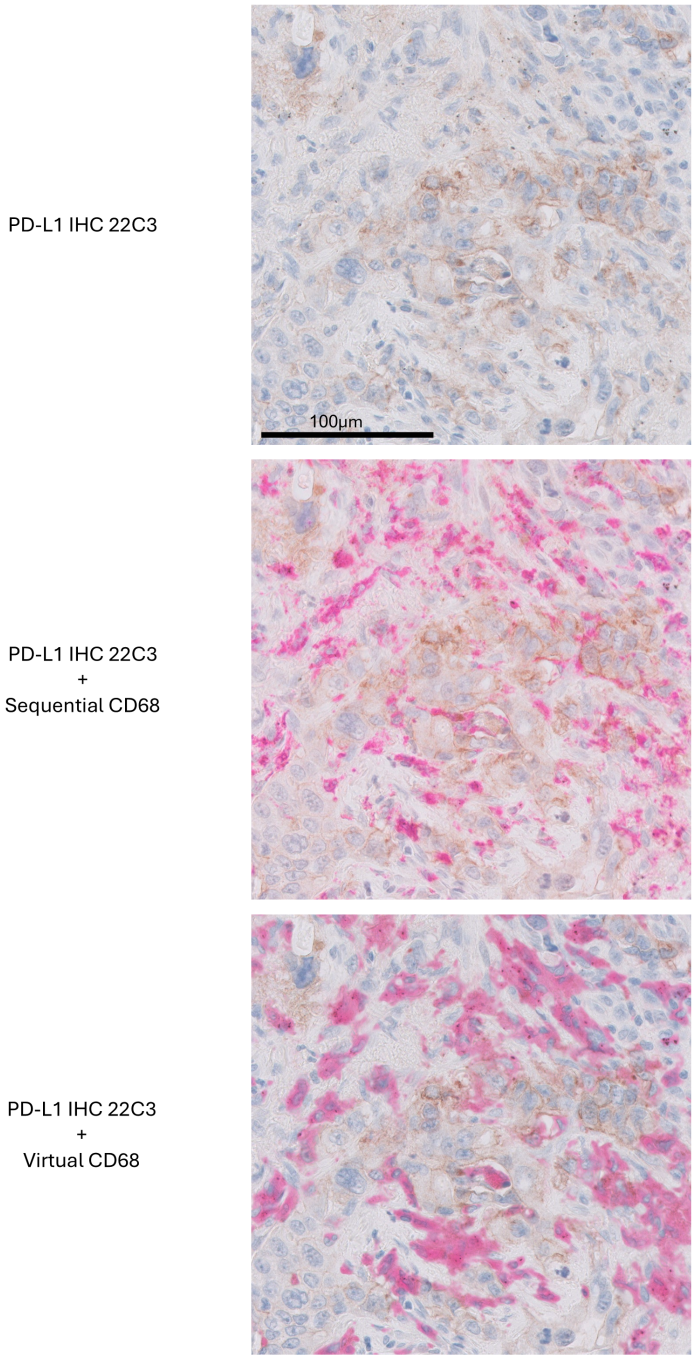

Figure 11: ROI3

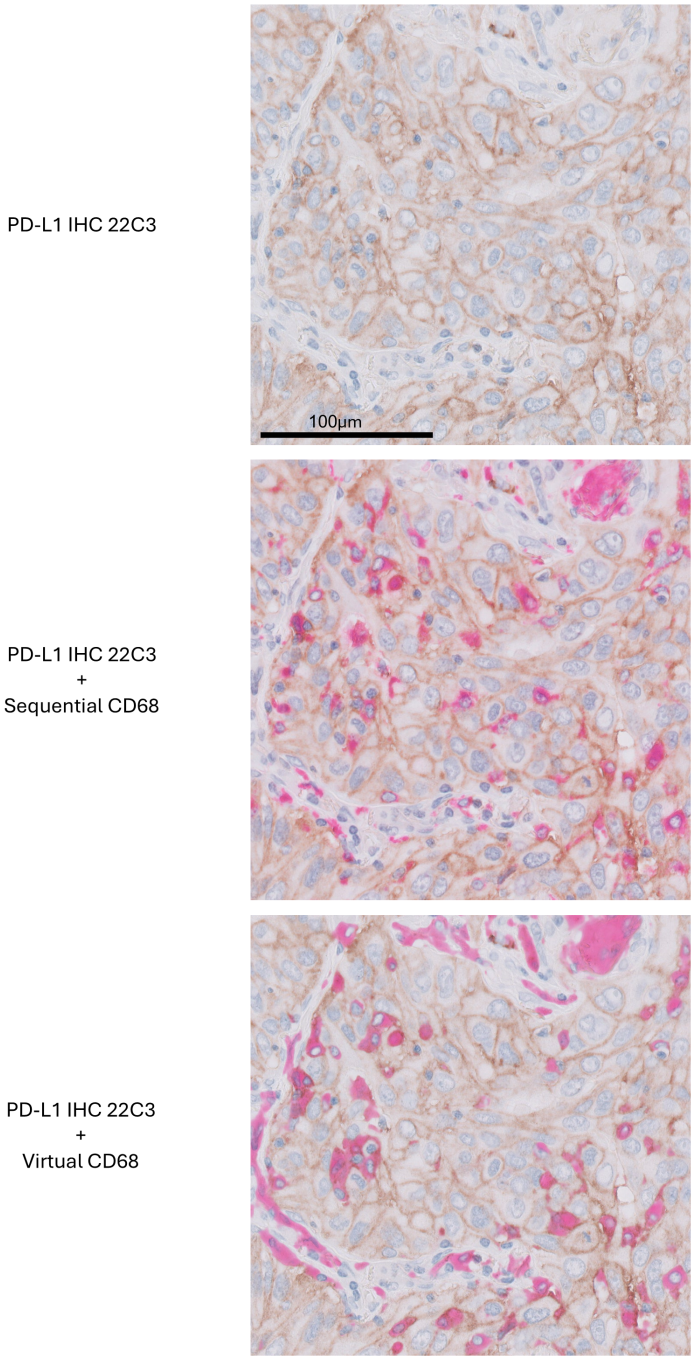

Figure 12: ROI4

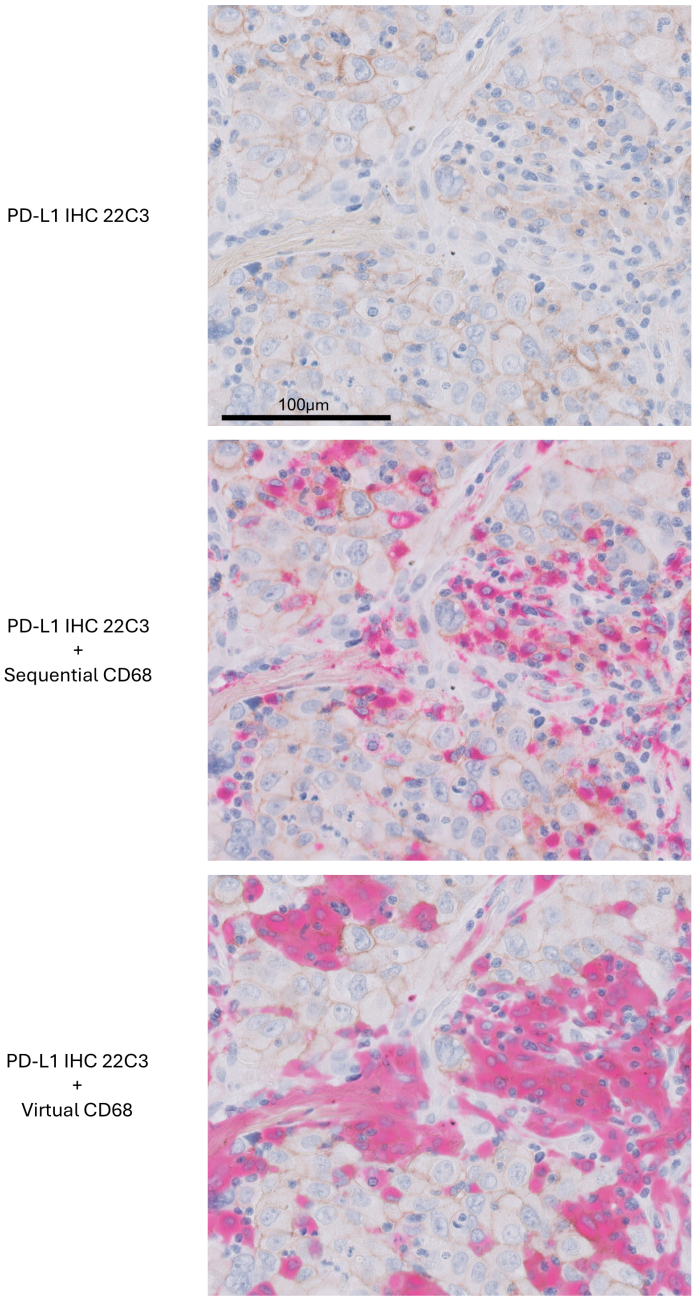

Figure 13: ROI5

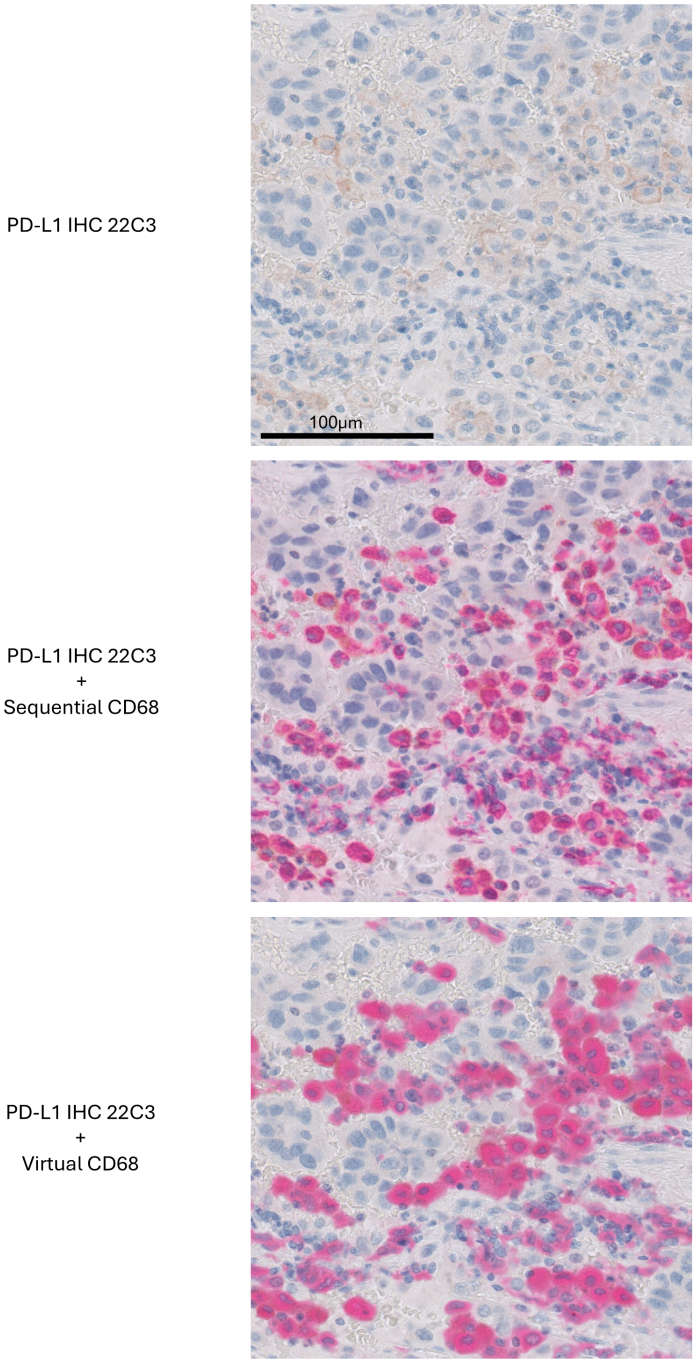

Figure 14: ROI6