Brief report

MuTILs: A multiresolution deep-learning model for interpretable scoring of Tumor-Infiltrating Lymphocytes in breast carcinomas using clinical guidelines

Mohamed Amgad[1], Roberto Salgado[2,3], Lee A.D. Cooper[1]*

[1] Department of Pathology, Northwestern University, Chicago, IL, USA [2] Department of Pathology, GZA-ZNA Ziekenhuizen, Antwerp, Belgium, [3] Division of Research, Peter MacCallum Cancer Centre, Melbourne, Australia.

*** Address correspondence to:** ***lee.cooper@northwestern.edu***

Keywords: Tumor-Infiltrating Lymphocytes, TILs, Deep Learning, Breast Cancer, Explainable AI, Histopathology.

Abstract

Tumor-Infiltrating Lymphocytes (TILs) have strong prognostic and predictive value in breast cancer, but their visual assessment is subjective. We present MuTILs, a convolutional neural network architecture for the panoptic segmentation of tissue regions and cell nuclei in whole-slide images of breast cancer, specifically optimized for assessing TILs in accordance with clinical scoring recommendations. MuTILs is a concept bottleneck model designed to be interpretable and to encourage sensible predictions at multiple resolutions. Our computational scores match visual scores and have independent prognostic value in invasive breast cancers from the TCGA dataset.

Introduction

Advances in slide scanners, machine learning, and computational efficiency have increased interest in histology as a source of data in cancer studies [1,2]. Tissue morphology contains essential prognostic and diagnostic information and reflects underlying molecular and biological processes. This work presents approaches for the computational discovery of interpretable predictive histologic biomarkers, focusing on invasive breast carcinomas. Histopathology is a medical field where medical experts (i.e., pathologists) examine stained microscopic tissue sections to make diagnostic decisions, most often from tumor biopsies. While much of clinical medicine relies on the clinical examination of patients, histopathology is an imaging-focused field, like radiology, where much of the focus is on visual pattern recognition.

The term biomarker refers to a biological feature that we can use to indicate a clinical outcome. For example, prognostic biomarkers are biological features associated with good (or bad) prognosis, while predictive biomarkers predict response to therapy in randomized controlled trials [3]. Typically, when a histologic trait is related to outcomes in cancer, it is incorporated into the grading criteria, though this is not always the case. For example, there has been a strong focus on tumor-infiltrating lymphocytes (TILs) as a prognostic and predictive biomarker in breast cancer and other solid tumors in recent years [4]. This is because TILs infiltration can be a somewhat direct visualization of how well the host (patient) body can respond to the growing tumor by immune cells.

The majority of breast cancers are carcinomas. Based on morphology, breast carcinomas include many variants; the most common are infiltrating ductal carcinoma (which originates from breast duct epithelium) and infiltrating lobular carcinoma (from breast acini/glands) [5,6]. There are numerous morphological elements within a single breast cancer slide. Integrative genomic analysis of breast cancer identified four main subtypes, including Luminal-A, Luminal-B, Her2-Enriched/Her2+, and Basal [7]. These subtypes have distinct alterations and are associated with distinct patient survival prospects [8]. TILs are particularly prognostic and predictive of therapeutic response in basal and Her2+ breast carcinomas [9].

The stromal TILs score is the fraction of stroma within the tumor bed occupied by lymphoplasmacytic infiltrates (Fig. 1). TILs are assessed visually by pathologists through examination of formalin-fixed paraffin-embedded, hematoxylin and eosin (FFPE H&E) stained slides from tumor biopsies or resections. They are subject to considerable inter- and intraobserver variability, and hence a set of standardized recommendations was developed by the International Immuno-Oncology Working Group [10,11]. Nevertheless, observer variability remains a critical limiting factor in the widespread clinical adoption of TILs in research and clinical settings. Therefore, a set of recommendations was published for developing computational tools for TILs assessment [12]. A number of existing algorithms have been developed to score TILs. However, most diverge from clinical scoring recommendations, as summarized by Amgad *et al.* [12]. This report describes MuTILs, an interpretable deep-learning model for the panoptic segmentation of breast cancer WSIs, with a special emphasis on evaluating TILs.

Methods

MuTILs jointly segments tissue regions and cell nuclei and extends our earlier work on this topic (Fig. 2) [13]. It acts as a panoptic segmentation algorithm; that is, it detects all tissue regions and nuclei within a slide to enable a holistic, context-aware assessment of TILs infiltration [14]. MuTILs comprises two parallel U-Nets (each with a depth of 5) for segmenting regions and nuclei at 1 and 0.5 microns-per-pixel (MPP), respectively [15]. Inspired by the HookNet architecture, information is passed from the region branch down to the nucleus branch, by concatenation, to provide low-power context [16]. Additionally, region predictions from the low-resolution branch are upsampled and used to constrain the nucleus predictions in the high-resolution branch. Thus, region prediction is used to infer class-specific attention maps, which are derived by modeling the nucleus class prior probability as a linear combination of the corresponding region probability vector. User-defined manual compatibility kernels mask out incompatible predictions. This constraint promotes compatible, biologically sensible predictions. The model was trained using a multi-task loss that gives equal weight to Regions of Interest (ROI) ROI and High-Power Field (HPF) region predictions, unconstrained HPF nuclear predictions, and region-constrained nuclear predictions.

We relied on images from 125 infiltrating ductal breast carcinoma patients from the BCSS and NuCLS datasets [17,18]. Additionally, we supplemented the training set with annotations from 85 slides from the Cancer Prevention Study II cohort [19]. The slides were separated into training and testing sets using 5-fold internal-external cross-validation, using the same folds as the NuCLS modeling paper [18,20]. For training, we extrapolated the nuclear labels from the small ~256x256 pixel high-power fields to large 1024x1024 pixel ROIs by using NuCLS models to perform inference on the same slides they were trained on to obtain bootstrapped “weak” labels. Generalization results presented here use manual labels (Fig. 3). Each high-power field from the pathologist-corrected single-rater NuCLS dataset was padded to 1024x1024 at 0.5 MPP resolution (20x objective). As a result, each ROI had region segmentation for the entire field (from the BCSS dataset) and nucleus segmentation and classification for the central portion (from the NuCLS dataset). Note that the nucleus ground truth contains a mixture of bounding boxes and segmentation. The fields shown in Fig. 3 are from the testing sets.

For whole-slide image (WSI) inference, we relied on data from 305 breast carcinoma patients for validation, 269 of whom were infiltrating ductal carcinomas, and 156 were Her2+. Visual scores were assessed by one pathologist (RS) and used as the baseline. The WSI accession and tiling workflow used the *histolab* and *large_image* packages and included: 1. Tissue detection; 2. Detection and exclusion of empty space and markers/inking; 3. Tiling the slide and scoring tiles at a very low resolution (2 MPP); 4. Analyzing the top 300 tiles [21,22]. Fixing the number of analyzed ROIs ensured a near-constant run time of less than two hours per slide. Low-resolution tiles with a high composition of cellular (hematoxylin-rich) and acellular (eosin-rich) regions received a higher informativeness score. This favored tiles with more peritumoral stroma. Color deconvolution was performed using the Macenko method from the HistomicsTK package [23,24]. Each of the top informative tiles was assigned one of the trained MuTILs models in a grid-like fashion. This scheme acted as a form of ensembling without increasing the overall inference time.

Trained MuTILs models were then used to segment tissue and nuclear components. A Euclidean distance transform was applied to detect stroma within 32 microns from the tumor boundary. The fraction of image pixels occupied by this peritumoral stroma was considered a saliency score. We assessed the following variants of the TILs score (Fig. 1):

1. Number of TILs / Stromal area (nTSa)
2. Number of TILs / Number of cells in stroma (nTnS)
3. Number of TILs / Total Number of cells (nTnA)

We obtained these score variants both globally (aggregating region and nuclear counts from all ROIs) and through saliency-weighted averaging of scores obtained for each ROI independently. A simple linear calibration was then used to ensure the scores occupied a similar range as the visual scores.

Results

MuTILs has a strong emphasis on explainability; it segments individual regions and nuclei, which are then used to calculate the computational scores. Table 1 shows the region segmentation and nucleus classification accuracy on the testing sets. Results are on testing sets from the internal-external 5-fold cross-validation scheme (separation by hospital). Fold 1 contributed to hyperparameter tuning, so it is not included in the mean and standard deviation calculation. MuTILs achieves a high classification performance for components of the computational TILs score, including stromal region segmentation (DICE=80.8±0.4) as well as the classification of fibroblasts (AUROC=91.0±3.6), lymphocytes (AUROC=93.0±1.1), and plasma cells (AUROC=81.6±6.6). Region segmentation performance is variable and class-dependent, with the predominant classes (cancer, stroma, and empty) being the most accurate. The region constraint improves nuclear classification accuracy by ~2-3% overall, mainly by reducing the misclassification of immature fibroblasts and large TILs/plasma cells as cancer.

The generalization accuracy of MuTILs predictions is also supported by a qualitative examination of model predictions on the ROIs from BCSS and NuCLS datasets (**Fig. 3**) and the full WSI (**Fig. 4**). Note that in Fig. 4, the predictions show full WSI inference for illustration. Our analysis, however, only admitted the 300 most informative ROIs to the MuTILs model to ensure a constant run time of less than two hours per slide for practical applicability. ROI "informativeness" was measured at a very low resolution (2 MPP) during WSI tiling and favored ROIs with more peritumoral stroma.

Computational TILs score variants had a modest-to-high correlation with the visual scores (Spearman R ranges between 0.55 - 0.58) (**Fig. 5**). Visual scores were obtained from one pathologist using clinical scoring recommendations from the TILs Working Group. Two variants of computational scores were obtained: either the number of stromal TILs was divided by the stromal region area, or the number of TILs was divided by the total number of cells within the stromal region. We then calibrated these numbers to the visual scores for easy comparison. Note that while the scatter plots in Fig. 5 show the calibrated scores, the correlation coefficients were obtained using the raw scores to avoid optimistic results. In that figure, points in red are outliers that contributed to the correlation metric but not to the calibration. Some slides were outliers with discrepant visual and computational scores; the causes for this discrepancy are discussed below. Both global and ROI saliency-weighted scores were significantly correlated with the visual scores ($p<0.001$).

We examined the prognostic value of MuTILs on infiltrating ductal carcinomas and Her2+ carcinomas. While we had access to visual scores from the basal cohort, the number of outcomes was limited, and neither visual nor computational scores had prognostic value. Progression-free interval (PFI) is the endpoint used per recommendations from Liu *et al.* for TCGA, with progression events including local and distant spread, recurrence, or death [25]. First, we examined the Kapan-Meier curves for patient subgroups using a TILs-score threshold of 10% for stromal TILs score and the median value for the nTnA computational score variant (**Fig. 6**). A threshold of 10% was used for visual and calibrated computational scores consistent with some of the research literature. Note that there is no recommended threshold for stromal TILs scoring, and so these results should be considered along with continuous results used in Cox regression modeling. For comparison, we also included a metric that looks into the predictive value of TILs when the denominator includes all cells, not just those in the stromal compartment. All metrics in Fig. 6 were obtained by weighted averaging of computational scores from 300 ROIs. Both visual and computational scores had good separation within the infiltrating ductal cohort, although only the nTnS and nTnA computational scores had significant log-rank p-values ($p=0.009$ and $p=0.006$, respectively). Within the Her2+ cohort, all metrics had good separation on the Kaplan-Meier, although the visual score had a borderline p-value. All computational scores were significant within this cohort ($p=0.018$ for nTSa, $p=0.002$ for nTnS, and $p=0.006$ for nTnA).

We also examined the prognostic value of the continuous (untresholded) TILs scores using Cox proportional hazards regression, with and without controlling for clinically-salient covariates, including patient age, AJCC pathologic stage, histologic subtype, and basal status (**Table 2**). The analysis was restricted to slides where visual TILs scores were available for a fair comparison. In the multivariable setting, each metric was part of an independent model along with clinically-salient covariates. We controlled all multivariable models for patient age and AJCC pathologic stage I and II status. Additionally, we controlled models using the infiltrating ductal carcinoma subset for basal genomic subtype status, and we controlled models using the Her2+ subset for infiltrating ductal histologic subtype status. Within the infiltrating ductal cohort, the only metric with significant independent prognostic value on multivariable analysis was the nTnS computational score. Within the Her2+ cohort, the visual score was not independently prognostic ($p=0.158$), while the computational scores all had independent prognostic value, with the most prognostic being the nTnS variant ($p=0.003$, $HR<0.001$). Saliency-weighted ROI scores almost always had better prognostic value than global computational scores.

Discussion

One of the difficulties facing widespread adoption of state-of-the-art DL in medical domains is their opacity. There is a broad consensus that explainability is critical to trustworthiness, especially in clinical applications [1,12,26–28]. The standard application of DL models in histopathology involves the direct prediction of targets from the raw images. For example, we may predict patient survival given a WSI scan [29]. However, an alternative paradigm is beginning to emerge that combines the strong predictive power of opaque DL models and the interpretable nature of handcrafted features, a technique called Concept bottleneck modeling [30]. The fundamental idea is simple: 1. Use DL to delineate various tissue compartments and cells; 2. Extract handcrafted features that make sense to a pathologist; 3. Learn to predict the target variable, say patient survival, using an interpretable ML model that takes handcrafted features as its input. Hence, the most challenging task is handled using powerful DL models, while the terminal prediction task uses highly interpretable models.

MuTILs is a concept bottleneck model; it learns to predict the individual components that contribute to the TILs score (i.e., peritumoral stroma and TILs cells)

and uses those to make the final predictions [30]. This setup makes its predictions explainable and helps identify sources of error. The region constraint helped provide context for the nuclear predictions at high resolution, which helped reduce the misclassification of immature fibroblasts and plasma cells as cancer (**Fig. 7**). The training dataset contained several subclassifications for region and nuclear data with unreliable or variable ground truth. Hence, we assessed performance at the level of grouped classes with reliable ground truth (tumor, stroma, TILs) at evaluation. The low representativeness of normal breast acini in training makes raw MuTILs predictions unreliable for differentiating normal and cancerous epithelial tissue (Fig. 7, bottom row). This issue can be mitigated by expanding the training set or downstream modeling of architectural patterns, which is beyond the scope of this work. Note how the region constraint improves nuclear classifications (Fig. 7, third vs fourth column). This improvement is most notable for large TILs (Fig 7, first row) and immature fibroblasts (Fig. 7, second row), which are misclassified as cancer without the region constraint.

A qualitative examination of slides with discrepant visual and computational TILs scores shows there are three major contributors to discrepancies:

1. Misclassifications of some benign or low-grade tumor nuclei as TILs.

2. Variations in TILs density in different areas within the slide, which cause inconsistencies in visual scoring. This phenomenon is also a well-known contributor to inter-observer variability in visual TILs scoring [11].

3. Variable influence of tertiary lymphoid structures on the WSI-level score.

Our results show that the most prognostic TILs score variant (nTnS) is derived from dividing the number of TILs cells by the total number of cells within the stromal region. The visual scoring guidelines rely on the nTSa, which is reflected in the slightly higher correlation of the nTSa variant with the visual scores compared to nTnS [10]. So why is nTnS more prognostic than nTSa? There are two potential explanations. First, it may be that nTnS is better controlled for stromal cellularity since it would be the same in low- vs. high-cellularity stromal regions as long as the proportion of stromal cells that are TILs is the same. Second, nTnS may be less noisy since it relies entirely on nuclear assessment at 20x objective, while stromal regions are segmented at half that resolution.

Finally, we note that this validation was done only using the TCGA cohort, and future work will include validation on more breast cancer cohorts. In addition, we note that MuTILs cannot distinguish cancer from normal breast tissue at low resolution, which may necessitate manual curation of the analysis region, especially for low-grade cases.

Conclusion

MuTILs is a lightweight deep-learning model for reliable assessment of TILs in breast carcinomas. It jointly classifies tissue regions and cell nuclei at different resolutions and uses these predictions to derive patient-level scores. We show that MuTILs can produce predictions with good generalization for the predominant tissue and cell classes relevant for TILs scoring. Furthermore, computational scores correlate significantly with visual assessment and have strong independent prognostic value in infiltrating ductal carcinoma and Her2+ cancer.

Data Availability Statement

This study relies on The Cancer Genomic Atlas (TCGA) dataset for scanner slides, along with corresponding annotations from the Breast Cancer Semantic Segmentation dataset and the NuCLS dataset. Each of these datasets is publicly available.

Conflicts of Interest

None to disclose.

Author Contributions

MA: Idea conception, model implementation, validation, and manuscript review. **RS**: manual scoring of TILs, manuscript writing. **LAD**: Idea conception, manuscript writing.

Acknowledgments

This work was supported by the U.S. NIH NCI grants U01CA220401 and U24CA19436201. We acknowledge support from Dr. David Gutman and the American Cancer Society, including Dr. Mia M. Gaudet, Dr. Samantha Puvanesarajah, Dr. Lauren Teras, James Hodge, and Elizabeth Bain.

References

1. *Abels E, Pantanowitz L, Aeffner F, Zarella MD, Laak J, Bui MM, et al. Computational pathology definitions, best practices, and recommendations for regulatory guidance: a white paper from the Digital Pathology Association. J Pathol. 2019 Nov 3;249(3):286–94.*
2. *van der Laak J, Litjens G, Ciompi F. Deep learning in histopathology: the path to the clinic. Nat Med. 2021 May 14;27(5):775–84.*
3. *Ballman K v. Biomarker: Predictive or Prognostic? Journal of Clinical Oncology. 2015 Nov 20;33(33):3968–71.*
4. *Savas P, Salgado R, Denkert C, Sotiriou C, Darcy PK, Smyth MJ, et al. Clinical relevance of host immunity in breast cancer: from TILs to the clinic.*

Nat Rev Clin Oncol. 2016 Apr 15;13(4):228–41.

5. *Molavi DW. The Practice of Surgical Pathology: A Beginner's Guide to the Diagnostic Process. 2nd ed. 2017.*
6. *Amin MB, Greene FL, Edge SB, Compton CC, Gershenwald JE, Brookland RK, et al. The Eighth Edition AJCC Cancer Staging Manual: Continuing to build a bridge from a population-based to a more "personalized" approach to cancer staging. CA Cancer J Clin. 2017 Mar;67(2):93–9.*
7. *Comprehensive molecular portraits of human breast tumours. Nature. 2012 Oct 23;490(7418):61–70.*
8. *Fallahpour S, Navaneelan T, De P, Borgo A. Breast cancer survival by molecular subtype: a population-based analysis of cancer registry data. CMAJ Open. 2017 Sep 25;5(3):E734–9.*
9. *Savas P, Salgado R, Denkert C, Sotiriou C, Darcy PK, Smyth MJ, et al. Clinical relevance of host immunity in breast cancer: from TILs to the clinic. Nat Rev Clin Oncol. 2016 Apr 15;13(4):228–41.*
10. *Salgado R, Denkert C, Demaria S, Sirtaine N, Klauschen F, Pruneri G, et al. The evaluation of tumor-infiltrating lymphocytes (TILs) in breast cancer: recommendations by an International TILs Working Group 2014. Annals of Oncology. 2015 Feb;26(2):259–71.*
11. *Kos Z, Roblin E, Kim RS, Michiels S, Gallas BD, Chen W, et al. Pitfalls in assessing stromal tumor infiltrating lymphocytes (sTILs) in breast cancer. NPJ Breast Cancer. 2020 May 12;6(1):17.*
12. *Amgad M, Stovgaard ES, Balslev E, Thagaard J, Chen W, Dudgeon S, et al. Report on computational assessment of Tumor Infiltrating Lymphocytes from the International Immuno-Oncology Biomarker Working Group. NPJ Breast Cancer. 2020 May 12;6(1):16.*
13. *Amgad M, Sarkar A, Srinivas C, Redman R, Ratra S, Bechert CJ, et al. Joint region and nucleus segmentation for characterization of tumor infiltrating lymphocytes in breast cancer. In: Tomaszewski JE, Ward AD, editors. Medical Imaging 2019: Digital Pathology. SPIE; 2019. p. 20.*
14. *Kirillov A, He K, Girshick R, Rother C, Dollar P. Panoptic Segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2019. p. 9404–13.*
15. *Ronneberger O, Fischer P, Brox T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In 2015. p. 234–41.*
16. *van Rijthoven M, Balkenhol M, Siliņa K, van der Laak J, Ciompi F. HookNet: Multi-resolution convolutional neural networks for semantic segmentation in histopathology whole-slide images. Med Image Anal. 2021 Feb;68:101890.*
17. *Amgad M, Elfandy H, Hussein H, Atteya LA, Elsebaie MAT, Abo Elnasr LS, et al. Structured crowdsourcing enables convolutional segmentation of histology images. Bioinformatics. 2019 Sep 15;35(18):3461–7.*
18. *Amgad M, Atteya LA, Hussein H, Mohammed KH, Hafiz E, Elsebaie MAT, et al. NuCLS: A scalable crowdsourcing approach and dataset for nucleus classification and segmentation in breast cancer. Gigascience. 2022 May 17;11.*
19. *Calle EE, Rodriguez C, Jacobs EJ, Almon ML, Chao A, McCullough ML, et al. The American Cancer Society Cancer Prevention Study II Nutrition Cohort. Cancer. 2002 May 1;94(9):2490–501.*
20. *Steyerberg EW, Harrell FE. Prediction models need appropriate internal, internal–external, and external validation. J Clin Epidemiol. 2016 Jan;69:245–7.*
21. *Marcolini A, Bussola N, Arbitrio E, Amgad M, Jurman G, Furlanello C. histolab: A Python library for reproducible Digital Pathology preprocessing with automated testing. SoftwareX. 2022 Dec;20:101237.*
22. *Kitware Inc. large_image. https://github.com/girder/large_image.*
23. *Gutman DA, Khalilia M, Lee S, Nalisnik M, Mullen Z, Beezley J, et al. The Digital Slide Archive: A Software Platform for Management, Integration, and Analysis of Histology for Cancer Research. Cancer Res. 2017 Nov 1;77(21):e75–8.*
24. *Macenko M, Niethammer M, Marron JS, Borland D, Woosley JT, Xiaojun Guan, et al. A method for normalizing histology slides for quantitative analysis. In: 2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro. IEEE; 2009. p. 1107–10.*
25. *Liu J, Lichtenberg T, Hoadley KA, Poisson LM, Lazar AJ, Cherniack AD, et al. An Integrated TCGA Pan-Cancer Clinical Data Resource to Drive High-Quality Survival Outcome Analytics. Cell. 2018 Apr;173(2):400-416.e11.*
26. *Amgad M, Atteya LA, Hussein H, Mohammed KH, Hafiz E, Elsebaie MAT, et al. Explainable nucleus classification using Decision Tree Approximation of Learned Embeddings. Bioinformatics. 2022 Jan 3;38(2):513–9.*
27. *Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell. 2019 May 13;1(5):206–15.*
28. *Kundu S. AI in medicine must be explainable. Nat Med. 2021 Aug;27(8):1328.*
29. *Mobadersany P, Yousefi S, Amgad M, Gutman DA, Barnholtz-Sloan JS, Velázquez Vega JE, et al. Predicting cancer outcomes from histology and genomics using convolutional networks. Proceedings of the National Academy of Sciences. 2018 Mar 27;115(13).*
30. *Wei Koh P, Nguyen T, Siang Tang Y, Mussmann S, Pierson E, Kim B, et al. Concept Bottleneck Models. In: Proceedings of the 37th International Conference on Machine Learning, PMLR . 2020. p. 5338–48.*

Table 1. Generalization accuracy for region segmentation and nucleus classification using manual ground truth. Results are on testing sets from the internal-external 5-fold cross-validation scheme (separation by hospital). Fold 1 contributed to hyperparameter tuning, so it is not included in the mean and standard deviation calculation. MuTILs achieves a high classification performance for components of the computational TILs score. Region segmentation performance is variable and class-dependent, with the predominant classes (cancer, stroma, and empty) being the most accurate. The region constraint improves nuclear classification accuracy by ~2-3% overall, mainly by reducing the misclassification of immature fibroblasts and large TILs/plasma cells as cancer (see qualitative examination figure).

* Classes that contribute to the computational TILs score.

† Performance for Necrosis/Debris and TILs-dense regions is modest, primarily because of the inherent subjectivity of the task and variability in the ground truth. For example, how dense should the infiltrate be to be considered "dense"? Necrotic regions also often have TILs infiltrates at the margin or adjacent areas of fibrosis, which are inconsistently labeled as necrosis, stroma, or TILs-dense in the ground truth. Nonetheless, classifying cells/material that comprise necrotic regions (neutrophils, apoptotic bodies, debris, etc.) is reasonable at higher magnification.

‡ From the table, it is clear that the model essentially fails to segment normal breast acini at 10x magnification. This failure is likely caused by: 1. The low representation of normal breast tissue in the validation data from NuCLS and BCSS datasets; 2. Inconsistency in defining "normal," which is sometimes used in the sense of "non-cancer" (including benign proliferation), and sometimes only refers to terminal ductal and lobular units (TDLUs). At high resolution, the distinction between cancer versus normal/benign epithelial nuclei is reasonable.

	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Mean	Std
Regions at 10x objective (DICE)							
Cancer	84.4	82.1	83	82.8	82.8	82.7	0.4
Normal ‡	1.6	2.3	2.1	2.3	2.3	2.3	0.1
Stroma *	81.3	80.2	81	80.8	81	80.8	0.4
TILs-dense †	64.8	64	65.3	65.6	65.6	65.1	0.8
Necrosis/Debris †	64.1	55.6	56.7	57.3	57.1	56.7	0.8
Empty	83.5	83.5	84	84.2	84.3	84.0	0.4
Nuclei at 20x objective (AUROC)							
Cancer	96.5	97.2	98	97.4	91.1	95.9	3.2
Normal ‡		84.6	89.3	80	74.7	82.2	6.3
Fibroblast *	90.4	93	91.8	93.5	85.8	91.0	3.6
Lymphocyte *	93.3	92.3	93.6	91.9	94.2	93.0	1.1
Plasma Cell *	80.9	73.5	88	78.9	85.8	81.6	6.6
Debris †	82.8	84.9	80.1	93.9	57.1	79.0	15.7
Micro-avg.	91.9	92.2	95.6	93.5	88.9	92.6	2.8
Macro-avg.	85.4	83.9	86.3	85.2	75.3	82.7	5.0
Nuclei without region constraint (AUROC)							
Micro-avg.	90.5	91.1	95.4	91.9	86.2	91.2	3.8
Macro-avg.	84.5	78.1	86.9	81.5	73.1	79.9	5.8

Table 2. Cox regression survival analysis of the predictive value of visual and computational TILs scores for breast cancer progression. The analysis was restricted to slides where visual TILs scores were available for a fair comparison. In the multivariable setting, each metric was part of an independent model along with clinically-salient covariates. We controlled all multivariable models for patient age and AJCC pathologic stage I and II status. Additionally, we controlled models using the infiltrating ductal carcinoma subset for basal genomic subtype status, and we controlled models using the Her2+ subset for infiltrating ductal histologic subtype status. Significant p-values are outlined in bold, using a significance threshold of 0.05. The * symbol indicates values < 0.001. Abbreviations used: HR, Hazard Ratio; 95%CI, upper and lower bounds of the 95% confidence interval; C-index, concordance index; No., number; Avg, weighted average.

Metric	Type	Univariable					Multivariable				
		HR	95% CI		P-value	C-index	HR	95% CI		P-value	C-index
Infiltrating ductal carcinoma (N=269)											
Visual score		0.466	0.074	2.951	0.418	0.520	0.334	0.039	2.881	0.318	0.681
No of TILs / Stromal area	Global	*	*		0.287	0.548	*	*	*	0.321	0.667
No of TILs / No of cells in stroma	Global	0.098	0.004	2.711	0.170	0.546	0.081	0.002	3.428	0.188	0.670
No of TILs / Total No of cells	Global	0.078	*	16.98	0.353	0.526	0.073	*	29.87	0.393	0.667
No of TILs / Stromal area	ROI avg.	*	*		0.159	0.577	*	*	*	0.192	0.668
No of TILs / No of cells in stroma	ROI avg.	0.005	*	0.832	**0.042**	0.600	0.002	*	0.722	**0.038**	0.675
No of TILs / Total No of cells	ROI avg.	0.001	*	11.56	0.151	0.579	0.001	*	18.33	0.164	0.679
Her2+ carcinoma (N=156)											
Visual score		0.073	0.001	3.919	0.198	0.581	0.029	*	3.952	0.158	0.725
No of TILs / Stromal area	Global	*	*		**0.039**	0.644	*	*	*	**0.011**	0.816
No of TILs / No of cells in stroma	Global	*	*	0.201	**0.015**	0.673	*	*	0.057	**0.007**	0.813
No of TILs / Total No of cells	Global	*	*	0.719	**0.045**	0.621	*	*	0.001	**0.007**	0.800
No of TILs / Stromal area	ROI avg.	*	*		**0.020**	0.679	*	*	*	**0.010**	0.837
No of TILs / No of cells in stroma	ROI avg.	*	*	0.010	**0.005**	0.704	*	*	0.002	**0.003**	0.837
No of TILs / Total No of cells	ROI avg.	*	*	0.014	**0.021**	0.660	*	*	*	**0.006**	0.833

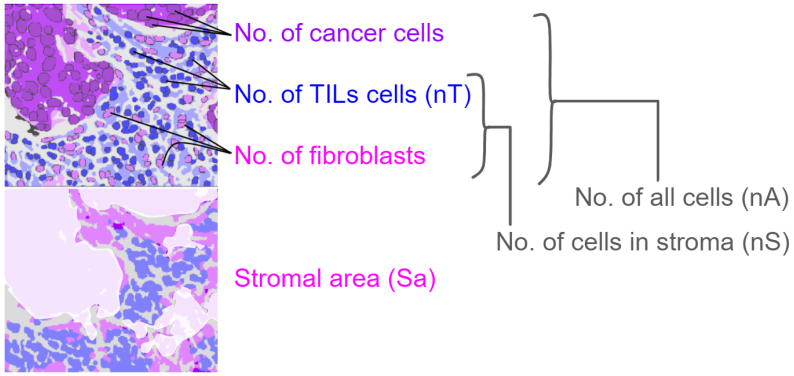

Fig 1. Components of various variants of the computational TILs score.

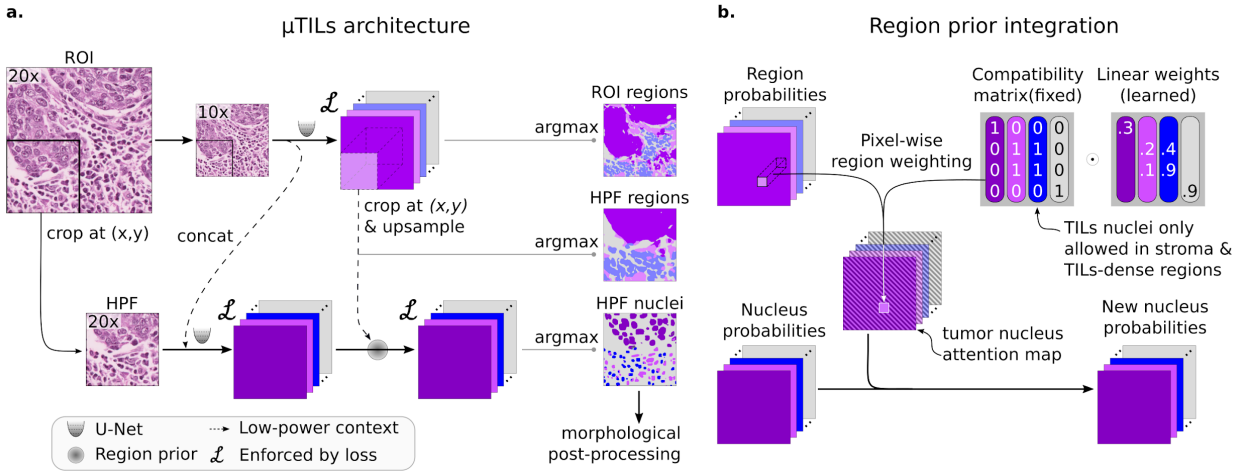

Fig 2. MuTILs model architecture. a. The MuTILs architecture utilizes two parallel U-Net models to segment regions at 1 MPP and nuclei at a 0.5 MPP resolution. Inspired by HookNet, we passed information down from the low-resolution branch to the high-resolution branch by concatenation. Additionally, region predictions from the low-resolution branch are upsampled and used to constrain the nucleus predictions in the high-resolution branch. The model was trained using a multi-task loss that gives equal weight to ROI and HPF region predictions, unconstrained HPF nuclear predictions, and region-constrained nuclear predictions. **b.** Region predictions are used to constrain nucleus predictions to enforce compatible cell predictions through class-specific attention maps. Attention maps are derived by modeling the nucleus class prior probability as a linear combination of the corresponding region probability vector. User-defined manual compatibility kernels mask out incompatible predictions.

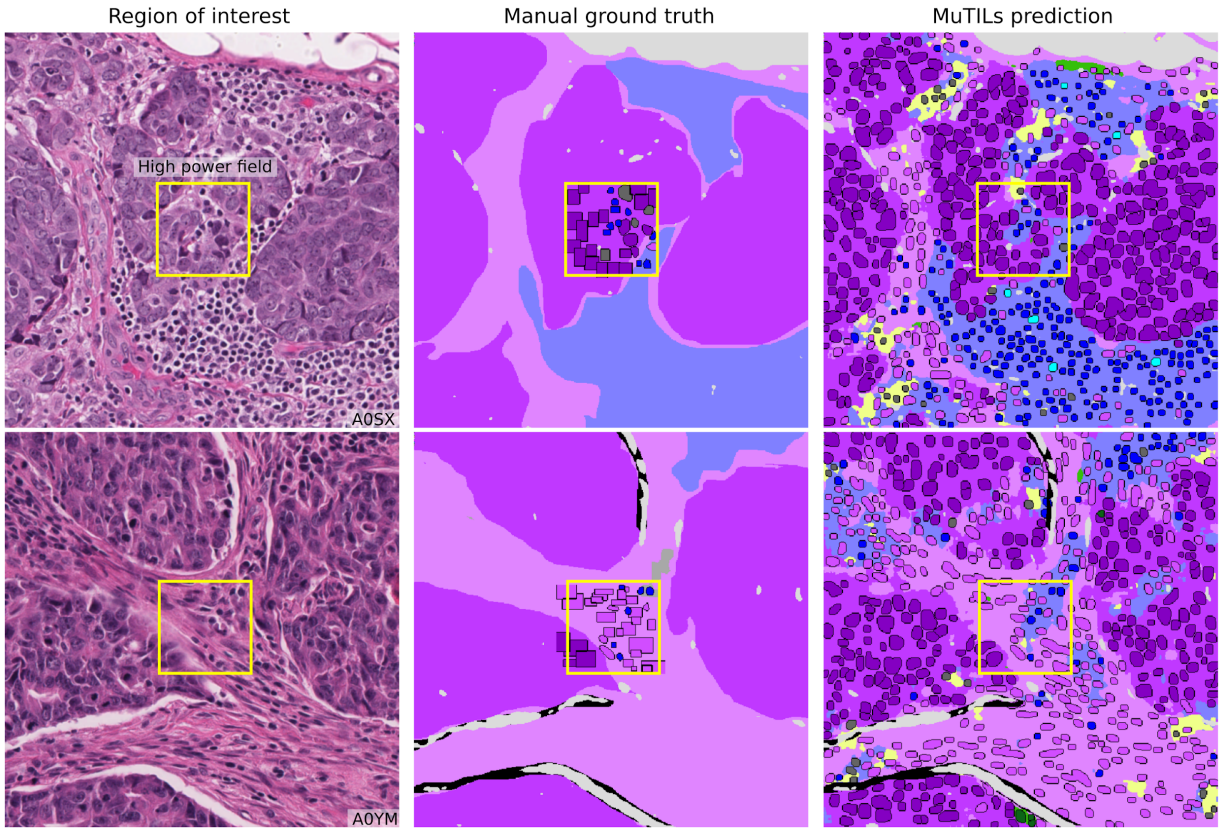

Fig 3. Reconciliation of manual region and nucleus ground truth for model validation. Each high power field from the pathologist-corrected single-rater NuCLS dataset was padded to 1024x1024 at 0.5 MPP resolution (20x objective). As a result, each ROI had region segmentation for the entire field (from the BCSS dataset) and nucleus segmentation and classification for the central portion (from the NuCLS dataset). Note that the nucleus ground truth contains a mixture of bounding boxes and segmentation. The fields shown here are from the testing sets.

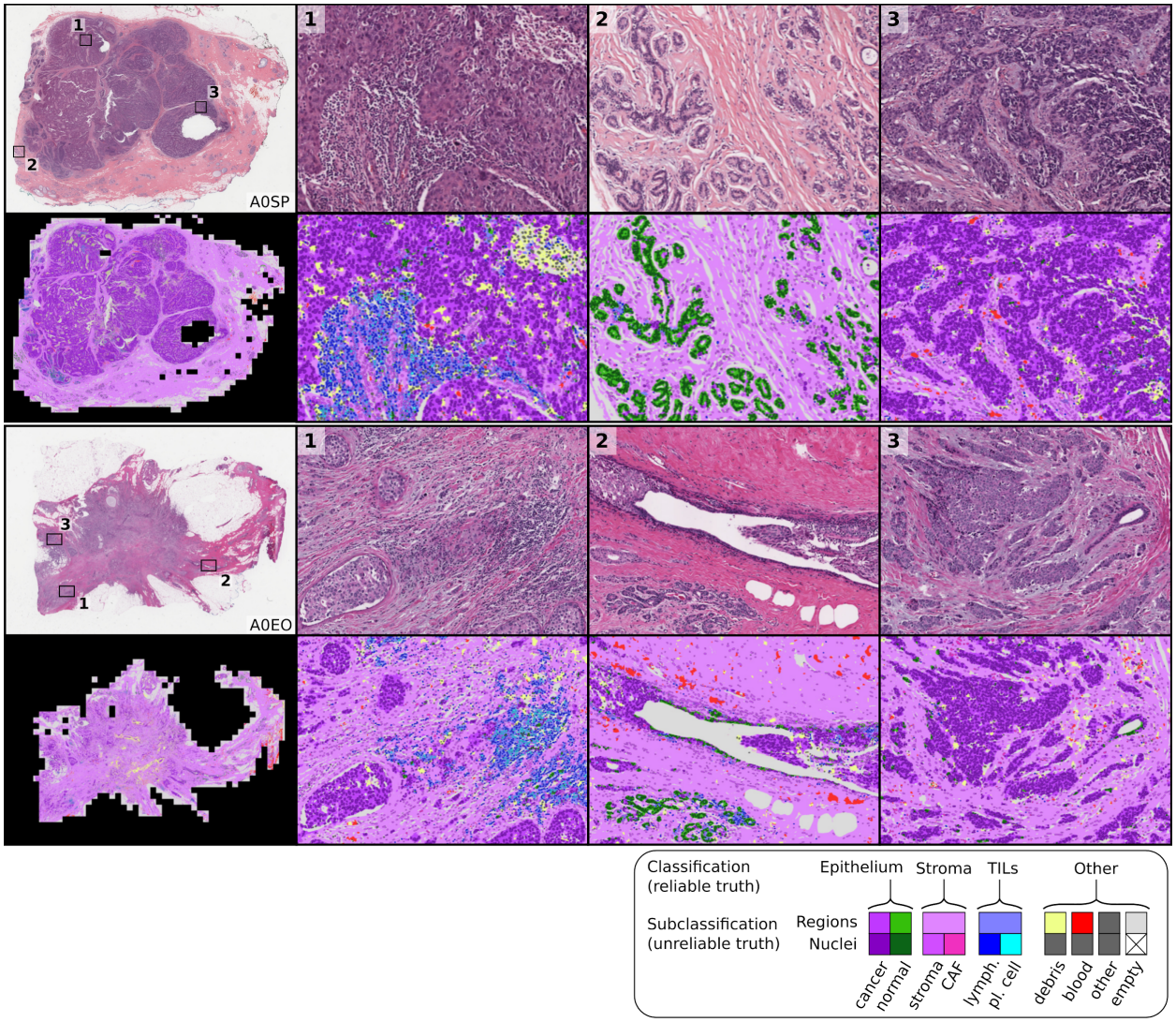

Fig 4. Sample whole-slide predictions from trained MuTILs models. The predictions show full WSI inference for illustration. Our analysis, however, only admitted the 300 most informative ROIs to the MuTILs model to ensure a constant run time of less than two hours per slide for practical applicability. ROI "informativeness" was measured at a very low resolution (2 MPP) during WSI tiling and favored ROIs with more peritumoral stroma.

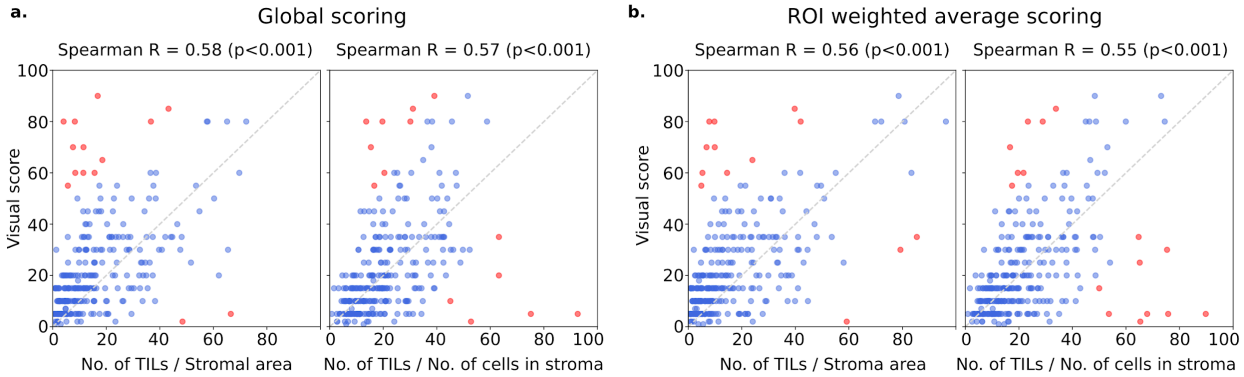

Fig 5. Correlation between visual and computational TILs assessment scores. Visual scores were obtained from one pathologist using clinical scoring recommendations from the TILs Working Group. MuTILs is a concept bottleneck model with a strong emphasis on explainability; it segments individual regions and nuclei, which are then used to calculate the computational scores. Two variants of computational scores were obtained: either the number of stromal TILs was divided by the stromal region area, or the number of TILs was divided by the total number of cells within the stromal region. We then calibrated these numbers to the visual scores for easy comparison. While this scatter plot shows the calibrated scores, the correlation coefficients were obtained using the raw scores to avoid optimistic results. Each point represents a single patient. Points in red are outliers that contributed to the correlation metric but not to the calibration. a. Computational scores are computed globally by aggregating data from all ROIs. b. Computational scores are computed independently for each ROI, and the slide-level score is calculated by weighted averaging.

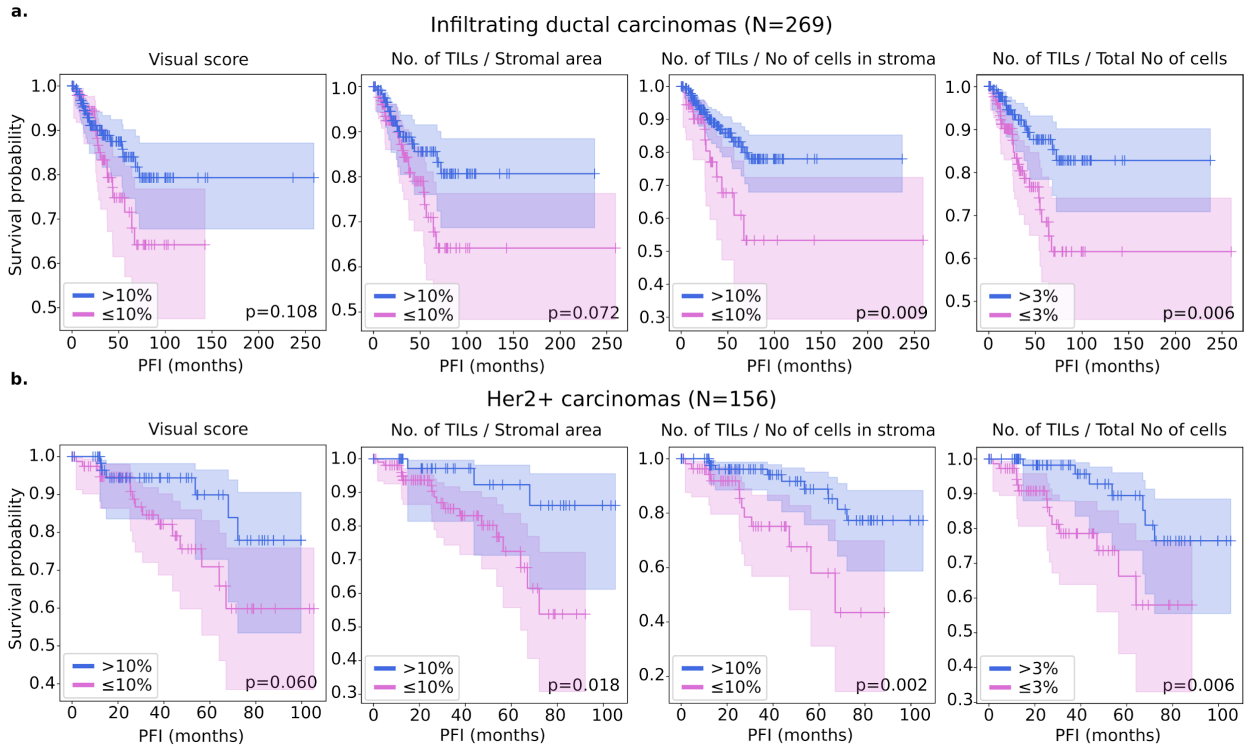

Fig 6. Kaplan-Meier analysis of visual and computational TILs assessment in predicting breast cancer progression. A threshold of 10% was used for visual and calibrated computational scores consistent with some of the research literature. Note that there is no recommended threshold for stromal TILs scoring, and so these results should be considered along with continuous results used in Cox regression modeling. For comparison, we also included a metric that looks into the predictive value of TILs when the denominator includes all cells, not just those in the stromal compartment. All metrics were obtained by weighted averaging of computational scores from 300 ROIs.

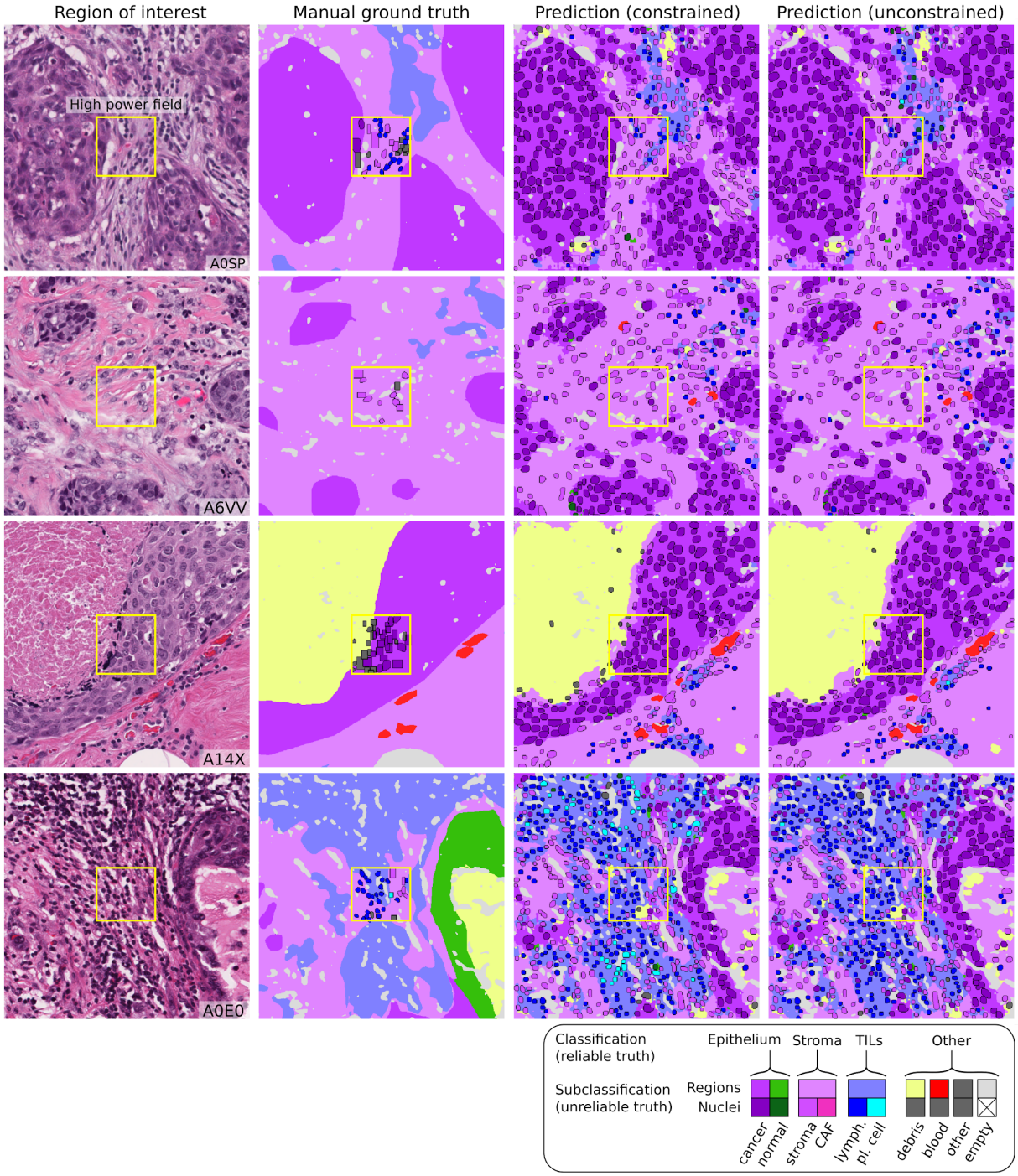

Fig 7. Qualitative examination of sample testing set predictions and sources of misclassification. The training dataset contained several subclassifications for region and nuclear data with unreliable or variable ground truth. Hence, we assessed performance at the level of grouped classes with reliable ground truth (tumor, stroma, TILs) at evaluation. The low representativeness of normal breast acini in training makes raw MuTILs predictions unreliable for differentiating normal and cancerous epithelial tissue (bottom row). This issue can be mitigated by expanding the training set or downstream modeling of architectural patterns, which is beyond the scope of this work. Note how the region constraint improves nuclear classifications (third vs fourth column). This improvement is most notable for large TILs (first row) and immature fibroblasts (second row), which are misclassified as cancer without the region constraint.