nature machine intelligence

ARTICLES

https://doi.org/10.1038/s42256-022-00471-x

Check for updates

Deep learning-inferred multiplex immunofluorescence for immunohistochemical image quantification

Parmida Ghahremani[1], Yanyun Li[2], Arie Kaufman[1], Rami Vanguri[2], Noah Greenwald[3], Michael Angelo[3], Travis J. Hollmann[2]✉ and Saad Nadeem[4]✉

Reporting biomarkers assessed by routine immunohistochemical (IHC) staining of tissue is broadly used in diagnostic pathology laboratories for patient care. So far, however, clinical reporting is predominantly qualitative or semi-quantitative. By creating a multitask deep learning framework, DeepLIIF, we present a single-step solution to stain deconvolution/separation, cell segmentation and quantitative single-cell IHC scoring. Leveraging a unique de novo dataset of co-registered IHC and multiplex immunofluorescence (mpIF) staining of the same slides, we segment and translate low-cost and prevalent IHC slides to more informative, but also more expensive, mpIF images, while simultaneously providing the essential ground truth for the superimposed brightfield IHC channels. A new nuclear-envelope stain, LAP2beta, with high (>95%) cell coverage is also introduced to improve cell delineation/segmentation and protein expression quantification on IHC slides. We show that DeepLIIF trained on clean IHC Ki67 data can generalize to noisy images as well as other nuclear and non-nuclear markers.

The assessment of protein expression using immunohistochemical (IHC) staining of tissue sections on glass slides is critical for guiding clinical decision-making in several diagnostic clinical scenarios, including cancer classification, residual disease detection and even mutation detection (BRAFV600E and NRASQ61R). Standard brightfield chromogenic IHC staining, while high-throughput, has a narrow dynamic range and results in superimposed channels with high chromogen/stain overlap, requiring specialized digital stain deconvolution/separation[1] as an essential preprocessing step in both state-of-the-art research and commercial IHC quantification algorithms. Stain deconvolution is an open problem requiring extensive hyperparameter tuning (on a per-case basis) or (highly error-prone and time-consuming) manual labelling of different cell types[2,3], but still results in sub-optimal colour separation in regions of high chromogen overlap.

As opposed to standard brightfield IHC staining, multiplex immunofluorescence (mpIF) staining provides the opportunity to examine panels of several markers individually (without requiring stain deconvolution) or simultaneously as a composite, permitting accurate co-localization, stain standardization, more objective scoring and cutoffs for all the markers' values (especially in low-expression regions, which are difficult to assess on IHC stained slides and can be misconstrued as negative due to weak staining that can be masked by the haematoxylin counterstain)[4,5]. Moreover, in a recent meta-analysis[6], mpIF was shown to have a higher diagnostic prediction accuracy (on par with multimodal cross-platform composite approaches) than IHC scoring, tumour mutational burden or gene expression profiling. However, mpIF assays are expensive and not widely available. There is thus a unique opportunity to leverage the advantages of mpIF to improve the explainability and interpretability of conventional IHC using recent deep learning breakthroughs.

Current deep learning methods for scoring IHCs rely solely on error-prone manual annotations (which feature unclear cell boundaries, overlapping cells and challenging assessment of low-expression regions) rather than on co-registered high-dimensional imaging of the same tissue samples (which can provide the essential ground truth for the superimposed brightfield IHC channels). In this Article we present a new multitask deep learning algorithm that leverages a unique co-registered IHC and mpIF training data of the same slides to simultaneously translate low-cost/prevalent IHC images to high-cost and more informative mpIF representations (creating a deep-learning-inferred IF image), accurately auto-segment relevant cells and quantify protein expression for more accurate and reproducible IHC quantification. Using multitask learning[7] to train models to perform a variety of tasks rather than one narrowly defined task makes them more generally useful and robust. Specifically, once trained, DeepLIIF takes only an IHC image as input (for example, Ki67 protein IHC as a brown Ki67 stain with standard haematoxylin nuclear counterstain), completely bypassing stain deconvolution, and produces/generates the corresponding haematoxylin, mpIF nuclear (4′,6-diamidino-2-phenylindole (DAPI)), mpIF protein (for example, Ki67) and mpIF LAP2Beta (a new nuclear envelope stain with >95% cell coverage to better separate touching/overlapping cells) channels and segmented/classified cells (for example, $Ki67^+$ and $Ki67^-$ cell masks for estimating the Ki67 proliferation index, which is an important clinical prognostic metric across several cancer types), as shown in Fig. 1 (for the complete pipeline, see Extended Data Figs. 1 and 2). Moreover, DeepLIIF trained just on clean IHC Ki67 images generalizes to more noisy and artefact-ridden images, as well as other nuclear and non-nuclear markers such as CD3, CD8, BCL2, BCL6, MYC, MUM1, CD10 and TP53. Example IHC images stained with different markers along with the DeepLIIF-inferred modalities and

[1]Department of Computer Science, Stony Brook University, Stony Brook, NY, USA. [2]Department of Pathology, Memorial Sloan Kettering Cancer Center, New York, NY, USA. [3]Department of Pathology, Stanford University, Stanford, CA, USA. [4]Department of Medical Physics, Memorial Sloan Kettering Cancer Center, New York, NY, USA. ✉e-mail: hollmant@mskcc.org; nadeems@mskcc.org

segmented/classified nuclear masks are also shown in Fig. 1. In essence, DeepLIIF presents a single-step solution to stain deconvoluion, cell segmentation and quantitative single-cell IHC scoring. Additionally, our co-registered mpIF data, for the first time, creates an orthogonal dataset to confirm and further specify the target brightfield IHC staining characteristics.

Results

In this section we describe the performance of DeepLIIF on cell segmentation and classification tasks. We evaluated the performance of our model and other state-of-the-art methods using pixel accuracy (PixAcc) computed from the number of true positives (TP), false positives (FP) and false negatives (FN) as $\frac{TP}{TP+FP+FN}$, dice score as $\frac{2\times TP}{2\times TP+FP+FN}$ and Intersection over Union (IOU) as the class-wise intersection over the union. We computed these metrics for each class, including negative and positive, and computed the average value of both classes for each metric. A pixel was counted as TP if it was segmented and classified correctly. A pixel was considered FP if it was falsely segmented as the foreground of the corresponding class. A pixel was counted as FN if it was falsely detected as the background of the corresponding class—for example, where the model segments a pixel as a pixel of a negative cell (blue), but in the ground-truth mask it is marked as positive (red). Because there is no corresponding pixel in the foreground of the ground-truth mask of the negative class, it is considered FP for the negative class and FN for the positive class, as there is no marked corresponding pixel in the foreground of the predicted mask of the positive class. We also evaluated our model against other methods using the aggregated Jaccard index (AJI), an object-level metric[8], defined as $\frac{\Sigma_{i=1}^{N}|G_i\cap P_M^i|}{\Sigma_{i=1}^{N}|G_i\cup P_M^i|+\Sigma_{F\in U}|P_F|}$ where G is the ground truth and P is the prediction. Considering that the goal is an accurate interpretation of the IHC staining results, we computed the difference between the IHC quantification percentage of the predicted mask and the real mask, as shown in Fig. 2.

To compare our model with state-of-the-art models, we used three different datasets. (1) We evaluated all models on our internal test set, including 600 images of size 512×512 and ×40 magnification from bladder carcinoma and non-small-cell lung carcinoma slides. (2) We randomly selected and segmented 41 images of size 640×640 from the recently released BC dataset[9], which contains Ki67 stained sections of breast carcinoma from scanned whole-slide images with manual Ki67$^+$ and Ki67$^-$ cell centroid annotations (targeting cell detection as opposed to the cell instance segmentation task), created from a consensus of ten pathologists. We split these tiles into 164 images of size 512×512; the test set varies widely in the density of tumour cells and the Ki67 index. (3) We also tested our model and others on a publicly available CD3 and CD8 IHC NuClick dataset[10]. We used the training set of the BC dataset, which contains 671 IHC patches of size 256×256, extracted from the LYON19 dataset[11]. LYON19[11] originates from a Grand Challenge to provide a dataset and an evolution platform to benchmark existing algorithms for lymphocyte detection in IHC stained specimens. The dataset contains IHC images of breast, colon and prostate stained with an antibody against CD3 or CD8.

Trained on clean lung and bladder images stained with the Ki67 marker, DeepLIIF generalizes well to other markers. We trained state-of-the-art segmentation networks, including FPN[12], LinkNet[13], Mask_RCNN[14], Unet++[15] and nnU-Net[16], on our training set (described in the section Training data), using the IHC images as the input and generating the coloured segmentation mask representing normal cells and lymphocytes. DeepLIIF outperformed previous models trained and tested on the same data on all three metrics. We trained and tested all models on a desktop with an NVIDIA Quadro RTX 6000 graphics processing unit (GPU), which was also used for all implementations.

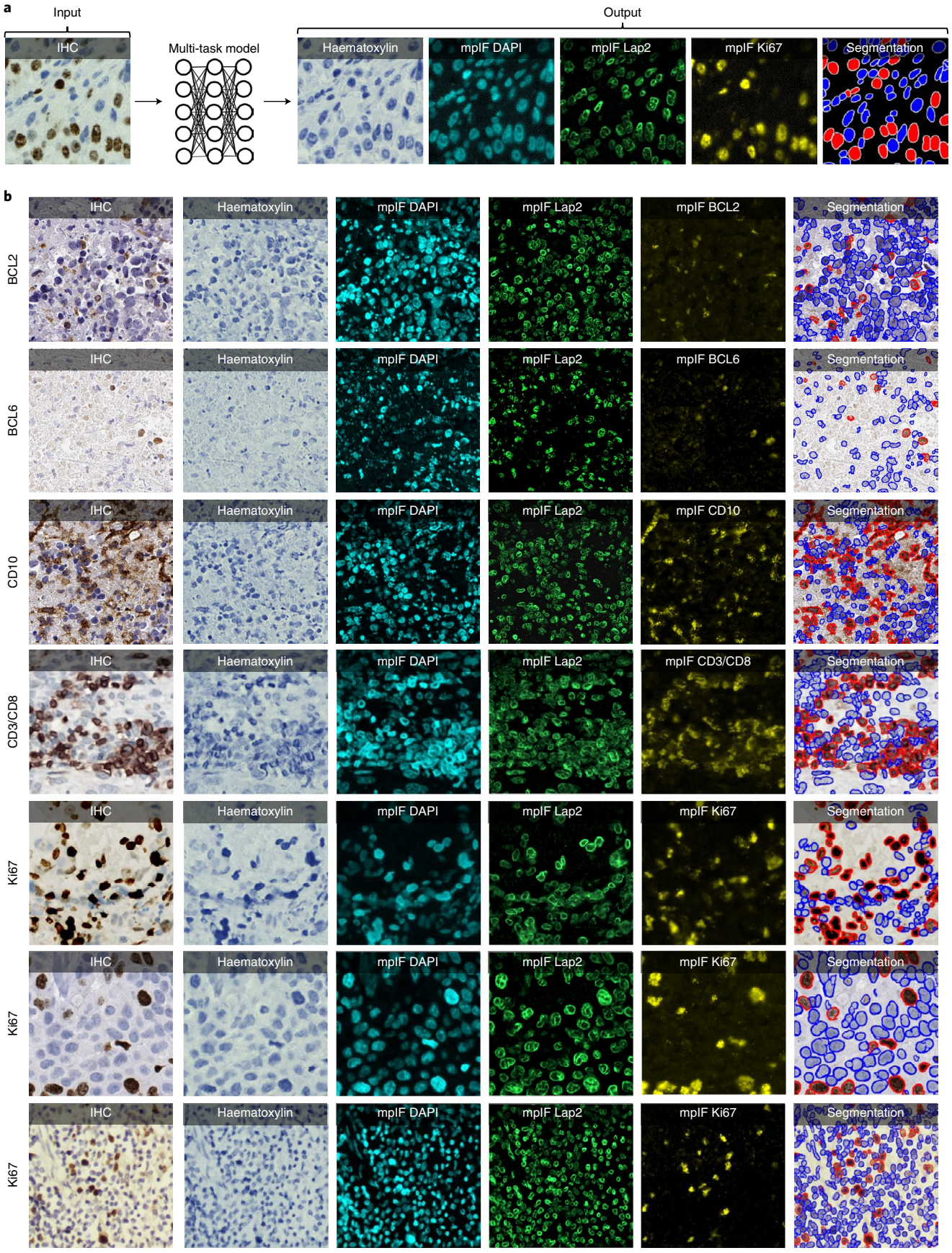
We compared the DeepLIIF model's performance against state-of-the-art models on the test set obtained from BC dataset[9]. The results were analysed both qualitatively and quantitatively, as shown in Fig. 2. All models were trained and validated on the same training set as the DeepLIIF model.

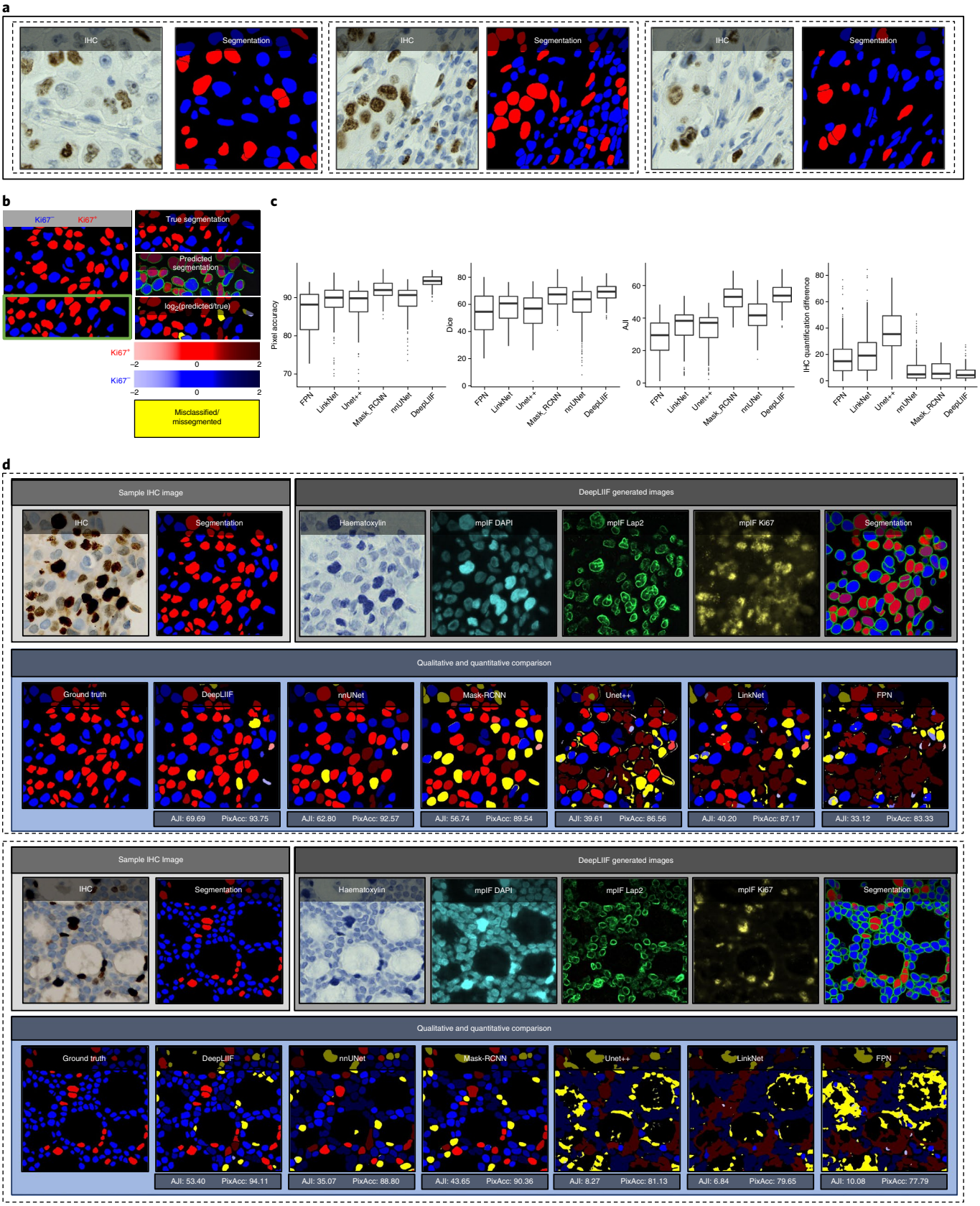
Application of DeepLIIF to the BC dataset resulted in a pixel accuracy of 94.18%, dice score of 68.15%, IOU of 53.20%, AJI of 53.48% and IHC quantification difference of 6.07%, and it outperformed Mask_RCNN (pixel accuracy of 91.95%, IOU of 66.16%, dice score of 51.16%, AJI of 52.36% and IHC quantification difference of 8.42%), nnUnet (pixel accuracy of 89.24%, dice score of 58.69%, IOU of 43.44%, AJI of 41.31% and IHC quantification difference of 9.84%), Unet++ (pixel accuracy of 87.99%, dice score of 54.91%, IOU of 39.47%, AJI of 32.53% and IHC quantification difference of 36.67%), LinkNet (pixel accuracy of 88.59%, dice score of 33.64%, IOU of 41.63%, AJI of 33.64% and IHC quantification difference of 21.57%) and FPN (pixel accuracy of 85.78%, dice score of 52.92%, IOU of 38.04%, AJI of 27.71% and IHC quantification difference of 17.94%), while maintaining a lower standard deviation on all metrics. We also performed a significance test to show that DeepLIIF significantly outperforms other models. As mentioned earlier, all models are trained and tested on exactly the same dataset, meaning that the data are paired. We thus performed a paired Wilcoxon rank-sum test, where a P value of 5% or lower is considered statistically significant. All tests were two-sided, and the assumption of normally distributed data was tested using a Shapiro–Wilk test. The computed P values of all metrics show that DeepLIIF significantly outperforms the state-of-the-art models.

We used pixel-level accuracy metrics for the primary evaluation, as we are formulating the IHC quantification problem as cell instance segmentation/classification. However, because DeepLIIF is capable of separating touching nuclei, we also performed a cell-level analysis of DeepLIIF against cell centroid detection approaches. U-CSRNet[9], for example, detects and classifies cells without performing cell instance segmentation. Most of these approaches use crowd counting techniques to find cell centroids. The major hurdle in evaluating these techniques is the variance in detected cell centroids. We trained FCRN_A[17], FCRN_B[17], Deeplab_Xeption[18], SC_CNN[19], CSRNet[20] and U-CSRNet[9] using our training set (the centroids of our individual cell segmentation masks are used as detection masks). Most of these approaches failed in detecting and classifying cells on the BC dataset testing set, and the rest detected centroids far from the ground-truth centroids. As a result, we resorted to comparing the performance of DeepLIIF (trained on our training set) with these models trained on the training set of the BC dataset and tested on the testing set of the BC dataset. As shown in Extended Data Fig. 3, even though

Fig. 1 | Overview of the DeepLIIF pipeline and sample input IHCs (different brown/DAB markers—BCL2, BCL6, CD10, CD3/CD8, Ki67) with corresponding DeepLIIF-generated haematoxylin/mpIF modalities and classified segmentation masks. a, Overview of DeepLIIF. Given an IHC input, our multitask deep learning framework simultaneously infers the corresponding haematoxylin channel, mpIF DAPI and mpIF protein expression (Ki67, CD3, CD8 and so on), as well as the positive/negative protein cell segmentation, baking explainability and interpretability into the model itself rather than relying on coarse activation/attention maps. In the segmentation mask, the red cells denote cells with positive protein expression (brown/DAB cells in the input IHC), whereas blue cells represent negative cells (blue cells in the input IHC). **b**, Example DeepLIIF-generated haematoxylin/mpIF modalities and segmentation masks for different IHC markers. DeepLIIF, trained on clean IHC Ki67 nuclear marker images, can generalize to noisier as well as other IHC nuclear/cytoplasmic marker images.

our model was trained on a completely different dataset from the testing set, it demonstrated better performance than the detection models that were trained on the same training set of the test dataset. The results show that, unlike DeepLIIF, the detection models are not robust across different datasets, staining techniques and tissue/cancer types.





As was mentioned earlier, our model generalizes well to segment/classify cells stained with different markers, including CD3/CD8. We compared the performance of our trained model against other trained models on the training set of the NuClick dataset[21]. The comparative analysis is shown in Fig. 3. The DeepLIIF model outperformed other models on segmenting and classifying CD3/CD8$^+$ cells (tumour-infiltrating lymphocytes (TILs)) on all three metrics.

Fig. 2 | Qualitative and quantitative analysis of DeepLIIF against state-of-the-art semantic segmentation models tested on the BC dataset. a, Three example pairs of images from our training set. **b**, A segmentation mask showing Ki67$^-$ and Ki67$^+$ cell representation, along with a visual segmentation and classification accuracy. Predicted classes are shown in different colours, with blue representing Ki67$^-$ and red representing Ki67$^+$ cells, and the hue is set using the $\log_2$ of the ratio between the predicted area and ground-truth area. Cells with areas that are too large are shown in dark colours, and cells with too small areas are shown in a light colour. For example, if the model correctly classifies a cell as Ki67$^+$, but the predicted cell area is too large, the cell is coloured in dark red. If there is no cell in the ground-truth mask corresponding to a predicted cell, the predicted cell is shown in yellow, which means that the cell is misclassified (cell segmented correctly but classified wrongly) or missegmented (no cell in the segmented cell area). **c**, The accuracy of the segmentation and classification is measured by obtaining the average of the dice score, pixel accuracy and the absolute value of the IHC quantification difference between the predicted segmentation mask of each class and the ground-truth mask of the corresponding class (0 indicates no agreement and 100 indicates perfect agreement). Evaluation of all scores shows that DeepLIIF outperforms all state-of-the-art models. **c**, The box plots represent the minimum, lower quartile, median (line in the box), upper quartile, and the maximum values. **d**, As mentioned earlier, DeepLIIF generalizes across different tissue types and imaging platforms. Two example images from the BC dataset[9], along with the inferred modalities and generated classified segmentation masks, are shown in the top rows, and the ground-truth mask and segmentation masks of five state-of-the-art models are shown in the second row. The mean IOU and pixel accuracy are given for each model in the box below the image.

We also evaluated the quality of the inferred modalities using mean squared error (m.s.e., the average squared difference between the synthetic image and the actual image) and the structural similarity index (SSIM; the similarity between two image). As shown in Extended Data Fig. 4, based on these metrics, DeepLIIF generates highly realistic images. In this figure we also visualize the first two components of principal component analysis (PCA) applied to the feature vectors of synthetic and real images, calculated by the VGG16 model, and then applied PCA on the calculated feature vectors. The results show that the synthetic image data points have the same distribution as the real image data points, confirming that the images generated by the model have the same characteristics as the real images. Original/real and DeepLIIF-inferred modality images of two samples taken from bladder and lung tissues are also shown side by side the SSIM and m.s.e. scores.

We also tested DeepLIIF on IHC images stained with eight other markers acquired with different scanners and staining protocols. Our testing set includes (1) nine IHC snapshots from a digital microscope stained with Ki67 and PDL1 markers (two examples are shown in Extended Data Fig. 5), (2) the testing set of LYON19[11] containing 441 IHC CD3/CD8 breast, colon and prostate regions of interest (ROIs; no annotations) with various staining/tissue artefacts from eight different institutions (Fig. 3c and Extended Data Fig. 6), (3) the PathoNet IHC Ki67 breast cancer dataset[22], containing manual centroid annotations created from a consensus of multiple pathologists, acquired in low-resource settings with a microscope camera (Extended Data Fig. 7), (4) Human Protein Atlas[23] IHC Ki67 (Fig. 4) and TP53 images (Extended Data Fig. 8) and (5) the DLBCL-Morph dataset[24], containing IHC tissue-microarrays for 209 patients stained with BCL2, BCL6, CD10, MYC and MUM1 markers (Extended Data Fig. 8).

We visualized the structure of the testing dataset by applying *t*-distributed stochastic neighbour embedding (*t*-SNE) to the image styles tested on DeepLIIF in Fig. 5. We first extracted the features from each image using the VGG16 model, and applied PCA to reduce the number of dimensions in the feature vectors. Next, we visualized the image data points based on the extracted feature vectors using *t*-SNE. As shown in Fig. 5, DeepLIIF is able to adapt to images with various resolutions, colour and intensity distributions, and magnifications captured in different clinical settings, and successfully segment and classify the heterogeneous collection of aforementioned testing sets covering eight different IHC markers.

We also evaluated the performance of DeepLIIF with and without LAP2beta and found the segmentation performance of DeepLIIF with LAP2beta better than without LAP2beta (Extended Data Fig. 9). LAP2beta is a nuclear envelope protein that is broadly expressed in normal tissues. In Extended Data Fig. 9, LAP2beta IHC reveals nuclear envelope-specific staining in the majority of cells in spleen (99.98%), colon (99.41%), pancreas (99.50%), placenta (76.47%), testis (95.59%), skin (96.74%), lung (98.57%), liver (98.70%), kidney (95.92%) and lymph node (99.86%). Placenta syncytiotrophoblast does not stain with LAP2beta, and the granular layer of skin does not show LAP2beta expression. However, the granular layer of skin lacks nuclei and is therefore not expected to express nuclear envelope proteins. We also observed a lack of consistent Lap2beta staining in the smooth muscle of blood vessel walls (not shown).

DeepLIIF, solely trained on IHC images stained with Ki67 marker, was also tested on haemotoxylin and eosin (H&E) images from the MonuSeg dataset[8]. As shown in Extended Data Fig. 10, DeepLIIF (out-of-the-box, without being trained on H&E images) was able to infer high-quality mpIF modalities and correctly segment the nuclei in these images.

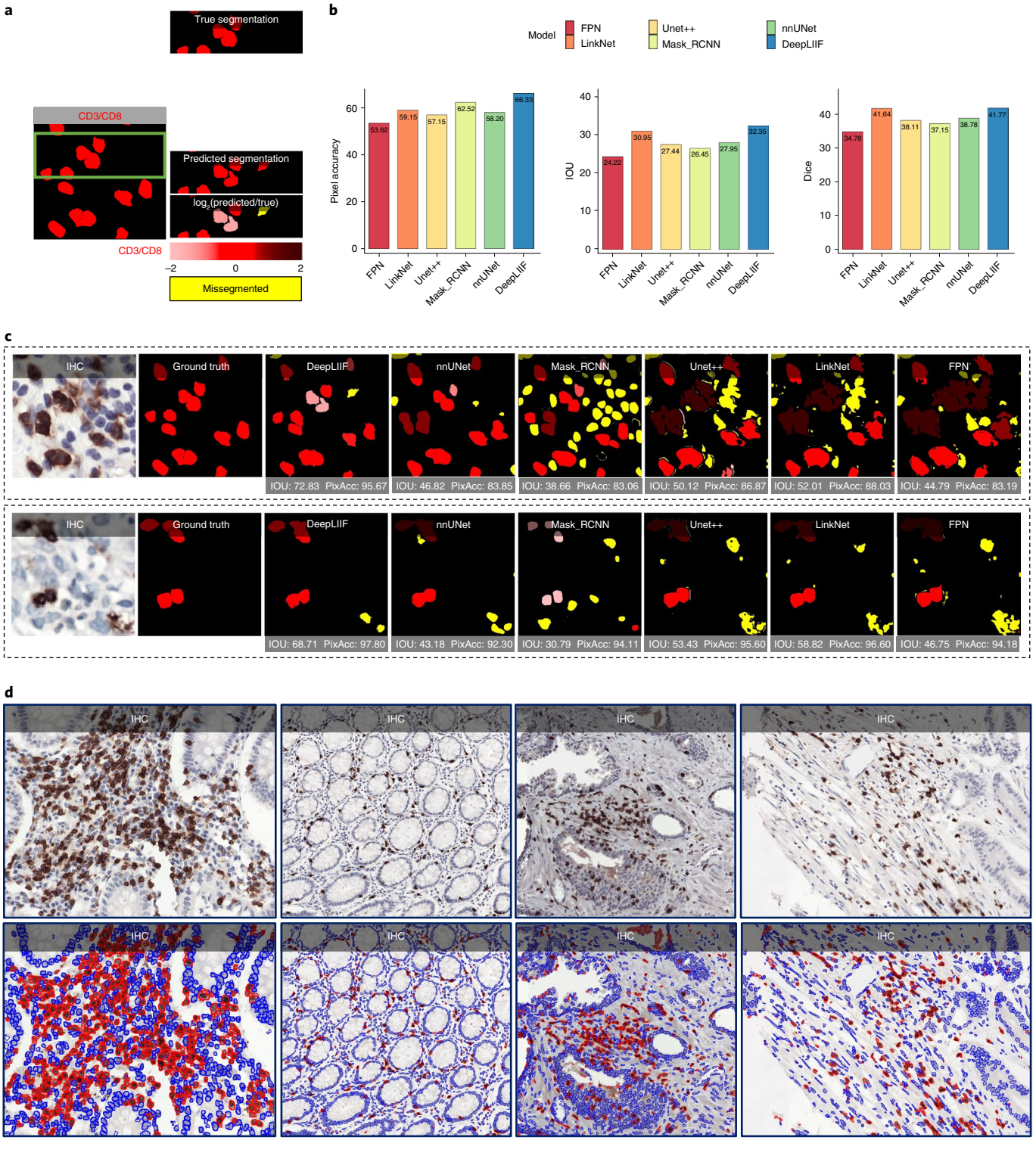
Discussion

Assessing IHC stained tissue sections is a widely utilized technique in diagnostic pathology laboratories, worldwide. IHC-based

Fig. 3 | Qualitative and quantitative analysis of DeepLIIF against state-of-the-art semantic segmentation models tested on the NuClick dataset and four sample images from the LYON19 challenge dataset. a, A segmentation mask showing CD3/CD8$^+$ cells, along with a visual segmentation and classification accuracy. Predicted CD3/CD8$^+$ cells are shown in red colour, and the hue is set using the $\log_2$ of the ratio between the predicted area and ground-truth area. Cells with areas that are too large are shown in dark colours, and cells with too small areas are shown in a light colour. For example, if the model correctly classifies a cell as CD3/CD8$^+$, but the predicted cell area is too large, the cell is coloured in dark red. If there is no cell in the ground-truth mask corresponding to a predicted cell, the predicted cell is shown in yellow, which means that the cell is missegmented (no corresponding ground-truth cell in the segmented cell area). **b**, The accuracy of the segmentation and classification is measured by obtaining the average of the dice score, pixel accuracy and IOU between the predicted segmentation mask of CD3/CD8 and the ground-truth mask of the corresponding cells (0 indicates no agreement and 100 indicates perfect agreement). Evaluation of all scores shows that DeepLIIF outperforms all state-of-the-art models. **c**, As mentioned earlier, DeepLIIF generalizes across different tissue types and imaging platforms. Two example images from the NuClick Dataset[21] along with the modalities and classified segmentation masks generated by DeepLIIF are shown in the top rows and the ground-truth mask and quantitative segmentation masks of DeepLIIF and state-of-the-art models are shown in the second row. The mean IOU and pixel accuracy are given for each generated mask. **d**, Randomly chosen samples from the LYON19 challenge dataset[11]. The top row shows the IHC image, and the bottom row shows the classified segmentation mask generated by DeepLIIF. In the mask, the blue colour shows the boundary of negative cells, and the red colour shows the boundary of positive cells.

protein detection in tissue with microscopic visualization is used for many purposes, including tumour identification, tumour classification, cell enumeration and biomarker detection and quantification. Nearly all IHC stained slides for clinical care are analysed and reported qualitatively or semi-quantitatively by diagnostic pathologists.

Several approaches have been proposed for deep learning-based stain-to-stain translation of unstained (label-free), H&E, IHC and multiplex slides, but relatively few attempts have been made (in limited contexts) at leveraging the translated enriched feature set for cellular-level segmentation, classification or scoring[25,26]. Recently, publicly available fluorescence microscopy and histopathology H&E datasets[27] were used for unsupervised nuclei segmentation in histopathology images by learning from fluorescence microscopy DAPI images. However, their pipeline incorporated CycleGAN, which hallucinated[28] nuclei in the target histopathology domain and hence required segmentation masks in the source domain to remove any redundant or unnecessary nuclei in the target domain. The model was also not generalizable across the two target histopathology datasets due to the stain variations, making this unsupervised



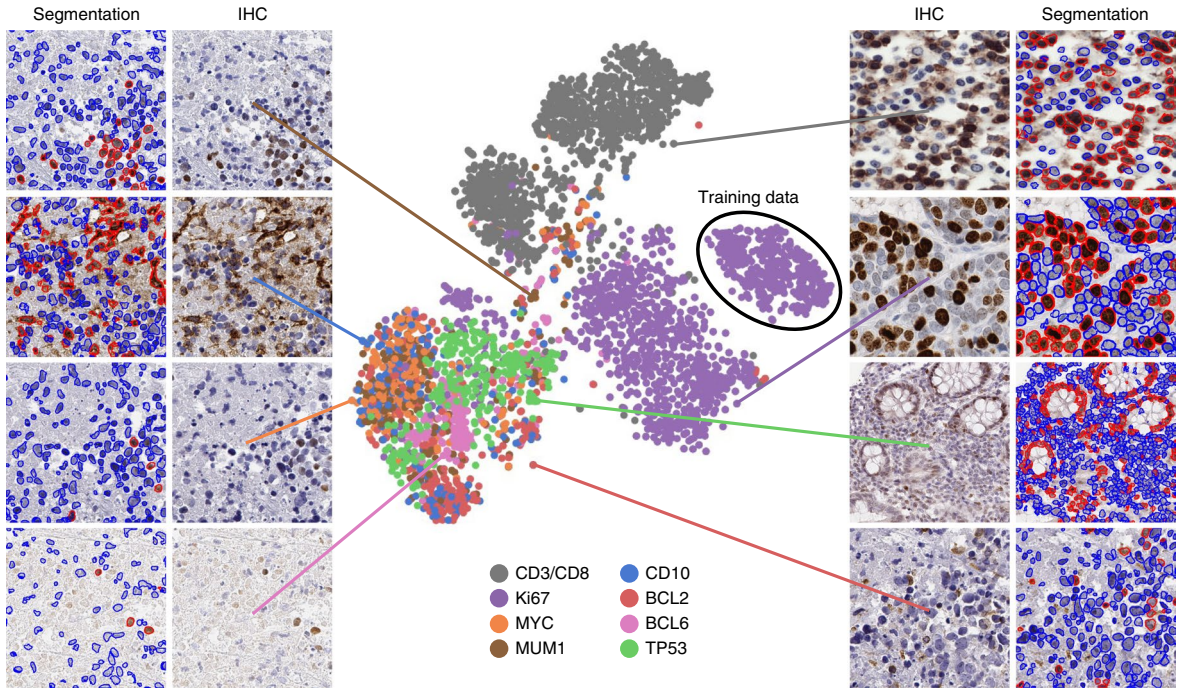


Fig. 4 | IHC quantification of four cancer type images taken from the Protein Atlas IHC Ki67 dataset. In each row, a sample is shown along with the inferred modalities and the classified segmentation mask. The demographic information of the patient and details about the staining, along with the manual protein score and the score predicted by DeepLIIF, are reported next to each sample.

solution less suitable for inferring different cell types from given H&E or IHC images. Reference [29], on the other hand, reports supervised learning trained on H&E and co-registered single-channel pancytokeratin IF for four patients with pancreatic ductal adenocarcinoma (PDAC) to infer pancytokeratin stain for a given PDAC H&E image. Another work[30] used a supervised learning method trained on H&E, and co-registered IHC PHH3 3,3′-diaminobenzidine (DAB) slides for mitosis detection in H&E breast cancer whole slide images. Recently, co-registered H&E and special stains for kidney needle core biopsy sections[31] were used to translate a given H&E image to special stains. In essence, there are methods to translate between H&E and IHC, but none for translating between IHC and mpIF modalities.To focus on immediate clinical application, we want to accentuate/disambiguate the cellular information in low-cost IHCs (using a higher-cost and more informative mpIF representation) to improve the interpretability for pathologists as well as for downstream analysis/algorithms.

By creating a multitask deep learning framework, DeepLIIF, we provide a unified solution to nuclear segmentation and quantification of IHC stained slides. DeepLIIF is automated and does not require annotations. In contrast, most commercial platforms use a time-intensive workflow for IHC quantification, which involves user-guided (1) IHC-DAB deconvolution, (2) nuclei segmentation of the haematoxylin channel, (3) threshold setting for the brown DAB stain and (4) cell classification based on the threshold. We present a simpler workflow, in which, given an IHC input, we generate different modalities along with the segmented/classified cell masks. Our multitask deep learning framework performs IHC quantification in one process and does not require error-prone IHC deconvolution or manual thresholding steps. We use a single optimizer for all generators and discriminators that improves the performance of all tasks simultaneously. Unique to this model, DeepLIIF is trained by generating registered mpIF, IHC and haematoxylin staining data from the same slide, with the inclusion of nuclear envelope staining to assist in accurate segmentation of adjacent and overlapping nuclei.

Formulating the problem as cell instance segmentation/classification rather than a detection problem helps us to move beyond the reliance on crowd counting algorithms and towards more precise boundary delineation (semantic segmentation) and classification algorithms. DeepLIIF was trained for multi-organ, stain-invariant determination of nuclear boundaries and the classification of subsequent single-cell nuclei as positive or negative for Ki67 staining detected with the DAB chromogen. Subsequently, we determined that DeepLIIF accurately classified all tested nuclear antigens as positive or negative.

Surprisingly, DeepLIIF is often capable of accurate cell classification of non-nuclear staining patterns using CD3, CD8, BCL2, PDL1 and CD10. We believe the success of the DeepLIIF classification of non-nuclear markers is at least in part dependent on the location of the chromogen deposition. BCL2 and CD10 protein staining often shows cytoplasmic chromogen deposition close to the nucleus, and CD3 and CD8 most often stain small lymphocytes with scant cytoplasm where the chromogen deposition is physically close to the nucleus. DeepLIIF is slightly less accurate in classifying PDL1 staining (Extended Data Fig. 5); notably, PDL1 staining is more often membranous staining of medium to large cells such as tumour cells and monocyte-derived cell lineages where DAB chromogen deposition is physically further from the nucleus. Because DeepLIIF was not trained for non-nuclear classification, we anticipate that further training using non-nuclear markers will rapidly improve their classification with DeepLIIF.

DeepLIIF handling of H&E images (Extended Data Fig. 10) was the most pleasant surprise, with the model, out-of-the-box, learning to even separate the H&E images into haematoxylin and (instead of mpIF protein marker) eosin stains. The nuclei segmentations were highly precise. This opens up a lot of interesting avenues where we can potentially drive whole-slide image registration of neighbouring H&E and IHC sections[32] by converting these to a common domain (clean mpIF DAPI images) and then performing deformable image registration.

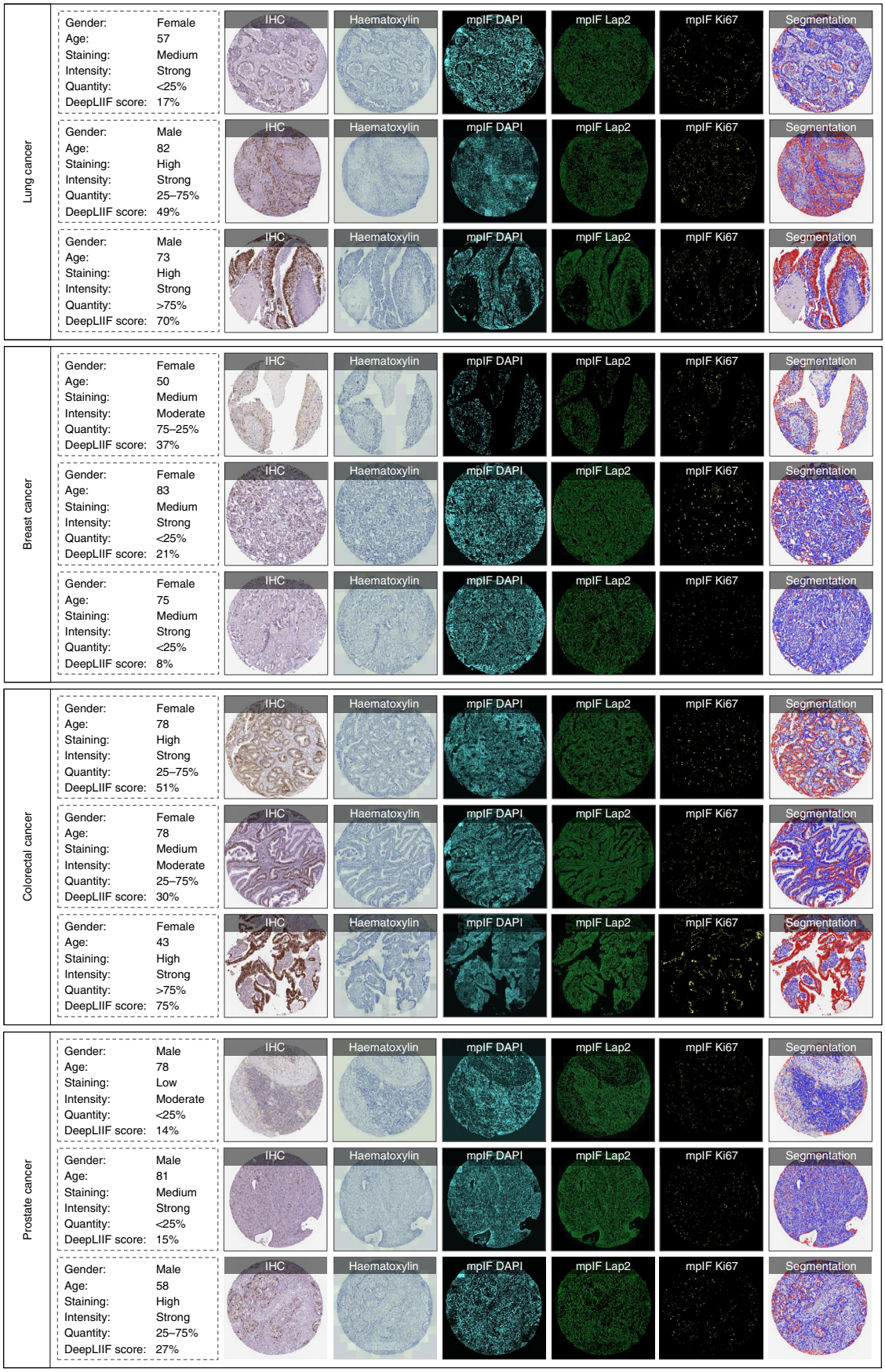

Lung cancer
Gender: Female
Age: 57
Staining: Medium
Intensity: Strong
Quantity: <25%
DeepLIIF score: 17%
Gender: Male
Age: 82
Staining: High
Intensity: Strong
Quantity: 25–75%
DeepLIIF score: 49%
Gender: Male
Age: 73
Staining: High
Intensity: Strong
Quantity: >75%
DeepLIIF score: 70%
Breast cancer
Gender: Female
Age: 50
Staining: Medium
Intensity: Moderate
Quantity: 75–25%
DeepLIIF score: 37%
Gender: Female
Age: 83
Staining: Medium
Intensity: Strong
Quantity: <25%
DeepLIIF score: 21%
Gender: Female
Age: 75
Staining: Medium
Intensity: Strong
Quantity: <25%
DeepLIIF score: 8%
Colorectal cancer
Gender: Female
Age: 78
Staining: High
Intensity: Strong
Quantity: 25–75%
DeepLIIF score: 51%
Gender: Female
Age: 78
Staining: Medium
Intensity: Moderate
Quantity: 25–75%
DeepLIIF score: 30%
Gender: Female
Age: 43
Staining: High
Intensity: Strong
Quantity: >75%
DeepLIIF score: 75%
Prostate cancer
Gender: Male
Age: 78
Staining: Low
Intensity: Moderate
Quantity: <25%
DeepLIIF score: 14%
Gender: Male
Age: 81
Staining: Medium
Intensity: Strong
Quantity: <25%
DeepLIIF score: 15%
Gender: Male
Age: 58
Staining: High
Intensity: Strong
Quantity: 25–75%
DeepLIIF score: 27%
IHC
Haematoxylin
mpIF DAPI
mpIF Lap2
mpIF Ki67
Segmentation

Fig. 5 | A *t*-SNE plot of tested IHC markers on DeepLIIF. The structure of the testing dataset is visualized by applying *t*-SNE to the image styles tested on DeepLIIF. The IHC protein markers in the tested datasets were embedded using *t*-SNE. Each point represents an IHC image of its corresponding marker. Randomly chosen example images of each marker are shown around the *t*-SNE plot. The black oval shows the cluster of training images. The distribution of data points shows that DeepLIIF is able to adapt to images with various resolutions, colour and intensity distributions, and magnifications captured in different clinical settings, and successfully segment and classify the heterogeneous collection of testing sets covering eight different IHC markers.

For the IHC images we purposely assessed the performance of DeepLIIF for the detection of proteins currently reported semi-quantitatively by pathologists with the goal of facilitating the transition to quantitative reporting if deemed appropriate. We anticipate the further extension of this work to assess the usability of Ki67 quantification in tumours with more unusual morphologic features such as sarcomas. The approach will also be extended to handle more challenging membranous/cytoplasmic markers such as PDL1, Her2 and so on, as well as H&E and multiplex IHC staining (without requiring any manual/weak annotations for different cell types[3]). Finally, we will incorporate additional mpIF tumour and immune markers into DeepLIIF for more precise phenotypic IHC quantification such as for distinguishing PDL1 expression within tumour versus macrophage populations.

This work provides a universal, multitask model for both segmenting nuclei in IHC images and recognizing and quantifying positive and negative nuclear staining. Importantly, we describe a modality where training data from higher-cost and higher-dimensional multiplex imaging platforms improves the interpretability of more widely used and lower-cost IHC.

Methods

Training data. To train DeepLIIF we used a dataset of lung and bladder cancer tissues containing IHC, haematoxylin, mpIF DAPI, mpIF Lap2 and mpIF Ki67 of the same tissue, scanned using a ZEISS Axioscan scanner (see Supplementary Information for more details on the staining protocol). Specifically, three patients with lung cancer (two males, one female, ages 45–57 years) and three patients with bladder cancer (three males, ages 52–66) were used for training and two patients with lung cancer (males, ages 48 and 55 years) and two patients with bladder cancer (males, ages 61 and 68 years) for testing. All were Caucasian. The images were scaled and co-registered with the fixed IHC images using affine transformations, resulting in 1,667 registered sets of IHC images and the other modalities of size 512×512. We randomly selected 363 sets for training, 53 sets for validation and 600 sets for testing the model. As described in the Synthetic data generation section, we synthetically generated 250 sets using our synthetic data generation model and added 212 to training and 38 to validation.

Ground-truth classified segmentation mask. To create the ground-truth segmentation mask for training and testing our model, we used our interactive deep learning ImPartial annotations framework[33]. Given mpIF DAPI images and few cell annotations, this framework auto-thresholds and performs cell instance segmentation for the entire image. Using this framework, we generated nuclear segmentation masks for each registered set of images with precise cell boundary delineation. Finally, using the mpIF Ki67 images in each set, we classified the segmented cells in the segmentation mask, resulting in 9,180 $Ki67^+$ cells and 59,000 $Ki67^-$ cells. Examples of classified segmentation masks from the ImPartial framework are shown in Figs. 1 and 2. The green boundaries around the cells are generated by ImPartial, and the cells are classified into red (positive) and blue (negative) using the corresponding mpIF Ki67 image. If a segmented cell has any representation in the mpIF Ki67 image, we classify it as positive (red colour); otherwise, we classify it as negative (blue colour).

Objective. Given a dataset of IHC + Ki67 RGB images, our objective is to train a model $f(\cdot)$ that maps an input image to four individual modalities, haematoxylin channel, mpIF DAPI, mpIF Lap2 and mpIF Ki67 images, and, using the mapped representations, generate the segmentation mask. We present a framework, as shown in Extended Data Fig. 1, that performs two tasks simultaneously. First, the translation task translates the IHC + Ki67 image into four different modalities for clinical interpretability as well as for segmentation. Second, a segmentation task generates a single classified segmentation mask from the IHC input and three of the inferred modalities by applying a weighted average and colouring cell boundaries in green, positive cells in red and negative cells in blue.

We used conditional generative adversarial networks (cGANs) to generate the modalities and the segmentation mask. cGANs are composed of two distinct components, a generator and a discriminator. The generator learns a mapping from the input image x to output image y, $G: x \rightarrow y$. The discriminator learns the paired input and output of the generator from the paired input and ground-truth result. We defined eight generators to produce four modalities and segmentation masks that cannot be distinguished from real images by eight adversarially trained discriminators (trained to detect fake images from the generators).

Translation. Generators G_{t_1}, G_{t_2}, G_{t_3} and G_{t_4} produce haematoxylin, mpIF DAPI, mpIF Lap2 and mpIF Ki67 images from the input IHC image, respectively ($G_{t_i} : x_i \rightarrow y_i$, where $i = 1, 2, 3, 4$). The discriminator D_{t_i} is responsible for discriminating images generated by generators G_{t_i}. The objectives of the cGAN for the image translator tasks are defined as follows:

$$\begin{aligned}\mathcal{L}_{t\mathrm{GAN}}(G_{t_i}, D_{t_i}) = \mathbb{E}_{x,y_i}[\log D_{t_i}(x, y_i)] \\ + \mathbb{E}_{x,y_i}[\log(1 - D_{t_i}(x, G_{t_i}(x)))]\end{aligned} \tag{1}$$

We use smooth L1 loss (Huber loss) to compute the error between the predicted value and the true value, because it is less sensitive to outliers than L2 loss and prevents exploding gradients while minimizing blur[34,35]. This is defined as

$$\mathcal{L}_{\mathrm{L1}}(G) = \mathbb{E}_{x,y}[\mathrm{smooth}_{\mathrm{L1}}(y - G(x))] \tag{2}$$

where

$$\mathrm{smooth}_{\mathrm{L1}}(a) = \begin{cases} 0.5a^2 & \text{if } |a| < 0.5 \\ |a| - 0.5 & \text{otherwise} \end{cases} \tag{3}$$

The objective loss function of the translation task is

$$\mathcal{L}_T(G_t, D_t) = \sum_{i=1\sim 5} \mathcal{L}_{t\mathrm{GAN}}(G_{t_i}, D_{t_i}) + \lambda \mathcal{L}_{\mathrm{L1}}(G_{t_i}) \tag{4}$$

where λ controls the relative importance of two objectives.

Segmentation and classification. The segmentation component consists of five generators G_{S_1}, G_{S_2}, G_{S_3}, G_{S_4} and G_{S_5} producing five individual segmentation masks from the original IHC, inferred haematoxylin image (G_{t_1}), inferred mpIF DAPI (G_{t_2}), inferred mpIF Lap2 (G_{t_3}) and inferred mpIF marker (G_{t_4}), $G_{S_i} =: z_i \rightarrow y_{s_i}$ where $i = 1, 2, 3, 4, 5$. The final segmentation mask is created by averaging the five generated segmentation masks by G_{S_i} using predefined weights $S(z_i) = \sum_{n=1}^{5} ws_i \times G_{S_i}(z_i)$, where ws_i are the predefined weights. The discriminators D_{S_i} are responsible for discriminating the images generated by generators G_{S_i}.

In this task, we use the LSGAN loss function, because it solves the problem of vanishing gradients for the segmented pixels on the correct side of the decision boundary, but far from the real data, resulting in a more stable boundary segmentation learning process. We define the objective of the cGAN for the segmentation/classification task as follows:

$$\begin{aligned}\mathcal{L}_{s\mathrm{GAN}}(D_S) = \sum_{i=1\sim 5} \Big(\tfrac{1}{2}\mathbb{E}_{z_i,y_{s_i}}[(D_{S_i}(z_i, y_{s_i}) - 1)^2] \\ + \tfrac{1}{2}\mathbb{E}_{z_i,y_{s_i}}[(D_{S_i}(z_i, S(z_i)))^2]\Big) \\ \mathcal{L}_{s\mathrm{GAN}}(S) = \sum_{i=1\sim 5} \tfrac{1}{2}\mathbb{E}_{z_i,y_{s_i}}[(D_{S_i}(z_i, S(z_i)) - 1)^2]\end{aligned} \tag{5}$$

For this task, we also use smooth L1 loss. The objective loss function of the segmentation/classification task is

$$\mathcal{L}_S(S, D_S) = \mathcal{L}_{s\mathrm{GAN}}(S, D_S) + \lambda \mathcal{L}_{\mathrm{L1}}(S) \tag{6}$$

Final objective. The final objective is

$$\begin{aligned}\mathcal{L}(G_t, D_t, S, D_S) = \mathcal{L}_T(G_t, D_t) \\ + \mathcal{L}_S(S, D_S)\end{aligned} \tag{7}$$

Generator. We use two different types of generator, the ResNet-9blocks generator for producing modalities and the U-Net generator for creating the segmentation mask.

ResNet-9blocks generator. The generators responsible for generating modalities including haematoxylin, mpIF DAPI and mpIF Lap2 start with a convolution layer and a batch normalization layer, then a rectified linear unit (ReLU) activation

function, two downsampling layers, nine residual blocks, two upsampling layers and a convolutional layer, followed by a tanh activation function. Each residual block consists of two convolutional layers with the same number of output channels. Each convolutional layer in the residual block is followed by a batch normalization layer and a ReLU activation function. These convolution operations are then skipped and the input is directly added before the final ReLU activation function.

U-Net generator. To generate the segmentation masks we use the generator proposed by ref. [35], using the general shape of U-Net[36] with skip connections. The skip connections are added between each layer i and layer $n-i$, where n is the total number of layers. Each skip connection concatenates all channels at layer i with those at layer $n-i$.

Markovian discriminator (PatchGAN). To address high frequencies in the image, we use a PatchGAN discriminator that only penalizes structure at the scale of patches. It classifies each $N \times N$ patch in an image as real or fake. We run this fully convolutional discriminator across the image, averaging all responses to provide the final output of D.

Optimization. To optimize our network, we use the same standard approach as ref. [37], alternating between one gradient descent step on D and one step on G. In all defined tasks (translation, classification and segmentation), the network generates different representations for the same cells in the input, meaning that all tasks have the same endpoint. We therefore use a single optimizer for all generators and a single optimizer for all discriminators. Using this approach, optimizing the parameters of a task with a more clear representation of cells improves the accuracy of other tasks because all these tasks are optimized simultaneously.

Synthetic data generation. We found that our model consistently failed in regions with dense clusters of IHC positive cells due to the absence of similar characteristics in our training data. To infuse more information about the clustered positive cells into our model, we developed a novel GAN-based model for the synthetic generation of IHC images using co-registered data. The model takes as input the haematoxylin channel, an mpIF DAPI image and the segmentation mask and generates the corresponding IHC and marker images (Extended Data Fig. 2). The model converts the haematoxylin channel to greyscale to infer more helpful information (such as the texture) and discard unnecessary information (such as colour). The haematoxylin image guides the network to synthesize the background of the IHC image by preserving the shape and texture of the cells and artefacts in the background. The DAPI image assists the network in identifying the location, shape and texture of the cells to better isolate the cells from the background. The segmentation mask helps the network specify the colour of cells based on the type of cell (positive cell, a brown hue; negative, a blue hue) and in creating the marker image. In the next step, we generated synthetic IHC images with more clustered positive cells. To do so, we changed the segmentation mask by choosing a percentage of random negative cells in the segmentation mask (called Neg-to-Pos), converting these into positive cells. We synthesized new IHC images by setting Neg-to-Pos to 50%, 70% and 90%. DeepLIIF was retrained with the new dataset, which contained original images and the synthesized ones, resulting in an improvement of the dice score by 6.57%, IOU by 7.08%, AJI by 5.53% and pixel accuracy by 2.49%.

Training details. We trained our model from scratch, using a learning rate of 0.0002 for 100 epochs, and linearly decayed the rate to zero over the next 100 epochs. The weights were initialized from a Gaussian distribution, $N(0, 0.02)$. We set $\lambda = 100$ to give more weight to the L1 loss. We used batch normalization in our main model. The Adam solver[38] was used with a batch size of 1. We used the tree-structured Parzen estimator (TPE) for hyperparameter optimization, and chose the L1 loss (least absolute deviations) as the evaluation metric to be minimized. We computed the L1 loss for the segmentation mask generated by the model and tried to minimize the L1 loss using the TPE approach. We optimized various hyperparameters, including the network generator architecture, the discriminator architecture, the number of layers in the discriminator while using the layered architecture, the number of filters in the generator and discriminator, the normalization method, the initialization method, the learning rate and the learning policy λ, the GAN loss function and segmentation mask generator weights, with diverse options for each of them.

Based on the hyperparameter optimization, the following predefined weights (ws_i) were set for the individual modalities to generate the final segmentation mask: weight of the segmentation mask generated by the original IHC image (ws_1) = 0.25, haematoxylin channel (ws_2) = 0.15, mpIF DAPI (ws_3) = 0.25, mpIF Lap2 (ws_4) = 0.1 and mpIF protein marker image (ws_5) = 0.25. The cell type (positive or negative) was classified using the original IHC image (brown cells are positive and blue cells negative) and the mpIF protein marker image (which only shows the positive cells). Therefore, to have enough information on the cell types, these two representations were assigned 50% of the total weight, with an equal contribution. The mpIF DAPI image contained the representation of the cell with the background and artefacts removed. Because this representation has the most useful information on the cell shape, area and boundaries, it was assigned 25% of the total weight in creating the segmentation mask. The mpIF Lap2 image was generated from the mpIF DAPI image and it contained only the boundaries on the cells. Even though it had more than 90% coverage, it still missed out on cells, so 15% of the total weight makes sense. With this weightage, we can be sure that if there is any confusing information in the mpIF DAPI image, it does not get infused into the model by a large weight. Also, by giving less weight to the Lap2, we increase the final segmentation probability of the cells not covered by Lap2. The haematoxylin image has all the information, including the cells with lower intensities, the artefacts and the background. As this image shares the background and artefact information with the IHC image and the cell information with the mpIF DAPI image, it is given less weight to decrease the probability of artefacts being segmented and classified as cells.

One of the challenges of GANs is the instability of their training. We used spectral normalization, a weight normalization technique, to stabilize the training of the discriminator[39]. Spectral normalization stabilizes the training of discriminators in GANs by re-scaling the weight tensor with spectral norm σ of the weight matrix calculated using the power iteration method. If the dimension of the weight tensor is greater than 2, it is reshaped to two dimensions by the power iteration method to achieve the spectral norm. We first trained the model using spectral normalization on the original dataset. The spectral normalization could not significantly improve the performance of the model. The original model achieved a dice score of 61.57%, IOU of 46.12%, AJI of 47.95% and pixel accuracy of 91.69%, whereas the model with spectral normalization achieved a dice score of 61.57%, IOU of 46.17%, AJI of 48.11% and pixel accuracy of 92.09%. In another experiment, we trained the model with spectral normalization on our new dataset containing original as well as the generated synthetic IHC images. The dice score, IOU and pixel accuracy of the model trained using spectral normalization dropped from 68.15% to 65.14%, 53.20% to 51.15% and 94.20% to 94.18%, respectively, while the AJI improved from 53.48% to 56.49%. As the results show, the addition of the synthetic images in training improved the model's performance across all metrics.

To increase the inference speed of the model, we also experimented with a many-to-one approach for the segmentation/classification task to decrease the number of generators to one. In this approach, we still have four generators and four discriminators for inferring the modalities, but use one generator and one discriminator (instead of five) for the segmentation/classification task, trained on the combination of all inferred modalities. We first trained this model with the original dataset. Compared to the original model with five segmentation generators, the dice score, IOU, AJI and pixel accuracy dropped by 12.13%, 10.21%, 12.45% and 3.66%, respectively. In another experiment, we trained the model with one segmentation generator on the new dataset including synthetic images. Similar to the previous experiment, using one generator instead of five independent generators deteriorated the model's performance: the dice score by 7%, IOU by 6.49%, AJI by 3.58% and the pixel accuracy by 0.98%. We observed, that similar to the original model, the addition of synthetic IHC images in the training process with one generator could increase the dice score from 49.44% to 61.13%, the IOU from 35.91% to 46.71%, the AJI from 35.50% to 49.90% and pixel accuracy from 88.03 to 93.22%, while reducing the performance drop, compared to the original model. This was still significantly less than the best performance from the multi-generator configuration, as shown above: dice score of 68.15%, IOU of 53.20%, AJI of 53.48% and pixel accuracy of 94.20%.

Testing details. The inference time of the model for a patch of 512 × 512 is 4 s. To infer the modalities and segment an image larger than 512 × 512, we tiled the image into overlapping patches. The tile size and overlap size can be given by the user as an input to the framework. The patches containing no cells are ignored in this step, improving the inference time. We then ran the tiles through our model. The model resizes the given patches to 512 for inference. In the final step, we stitched the tiles using the given overlap size to create the final inferred modalities and the classified segmentation mask. This takes about 10–25 min (depending on the percentage of cell-containing region, the WSI magnification level, user-selected tile size and overlap size) to infer the modalities and the classified segmentation mask of a WSI with size of 10,000 × 10,000 with ×40 magnification.

Ablation study. DeepLIIF infers four modalities to compute the segmentation/classification mask of an IHC image. We performed an ablation study on each of these four components. The goal of this experiment was to investigate whether the performance improvements are due to the increased ability of each task-specific network to share their respective features. In each experiment, we trained our model with three modalities, each time removing a modality to study the accuracy of the model in the absence of that modality. We tested all models on the BC dataset of 164 images with size 512 × 512. The results show that the original model (with all modalities), with a dice score of 65.14%, IOU of 51.15%, AJI of 56.49% and pixel accuracy of 94.20%, outperforms the model without the haematoxylin modality (dice score of 62.86%, IOU of 47.68%, AJI of 50.10% and pixel accuracy of 92.43%), the model without mpIF DAPI (dice score of 62.45%, IOU of 47.13%, AJI of 50.38% and pixel accuracy of 92.35%), the model without mpIF Lap2 (dice score of 61.07%, IOU of 45.71%, AJI of 49.14% and pixel accuracy of 92.16%) and the model without mpIF protein marker (dice score of 57.92%, IOU of 42.91%, AJI of 47.56% and pixel accuracy of 91.81%). The mpIF Lap2 is important for splitting overlapping cells and detecting boundaries (the model without mpIF

Lap2 has the lowest AJI score). Moreover, mpIF Lap2 is the only modality among the four that clearly outlines the cells in regions with artefacts or noise. The model without mpIF protein marker image has the worst pixel accuracy and dice score, showing its clear importance in cell classification. The mpIF DAPI image guides the model in predicting the location of the cells, given the drop in pixel accuracy and AJI score. The haematoxylin image on the other hand seems to make the least difference when removed, although it helps visually (according to two trained pathologists) by providing a separated haematoxylin channel from the IHC (haematoxylin + DAB) input.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this Article.

Data availability

The complete IHC Ki67 BC Dataset with manual annotations is available at https://sites.google.com/view/bcdataset. The complete lymphocytes detection IHC CD3/CD8 (LYON challenge) dataset is available at https://zenodo.org/record/3385420#.XW-6JygzYuW. The NuClick IHC annotations for crops from the LYON19 dataset can be found at https://warwick.ac.uk/fac/sci/dcs/research/tia/data/nuclick/ihc_nuclick.zip. The DLBCL-Morph dataset with BCL2, BCL6, MUM1, MYC and CD10 IHCs is accessible at https://stanfordmedicine.box.com/s/ub8e0wlhsdenyhdsuuzp6zhj0i82xrb1. The high-resolution tiff images for TP53 IHCs can be downloaded from https://www.proteinatlas.org/ENSG00000141510-TP53. All our internal training and testing data (acquired under IRB protocol approval no. 16-1683) and source data underlying the figures (in excel files), along with the pretrained models, are available at https://zenodo.org/record/4751737#.YV379XVKhH4. Source data are provided with this paper.

Code availability

All code was implemented in Python using PyTorch as the primary deep learning package. All code and scripts to reproduce the experiments of this paper are available at https://github.com/nadeemlab/DeepLIIF and releases are available at https://doi.org/10.5281/zenodo.5553268. For convenience, we have also included docker file as well as Google CoLab Demo project (in case someone does not have access to a GPU and wants to run their images directly via the CoLab project). The Google CoLab project can be accessed at https://colab.research.google.com/drive/12zFfL7rDAtXfzBwArh9hb0jvA38L_ODK?usp=sharing. We have also provided multi-GPU training code as well as highly optimized inference modules implemented in Torchserve as well as Dask and Torchscript. A cloud-native platform with a user-friendly web interface is available at https://deepliif.org for users to upload input images, visualize and download IHC quantification results. The interactive deep learning module for performing multiplex immunofluorescence cell segmentation is available at https://github.com/nadeemlab/impartial.

Received: 13 July 2021; Accepted: 28 February 2022;
Published online: 7 April 2022

References

1. Vahadane, A. et al. Structure-preserving color normalization and sparse stain separation for histological images. *IEEE Trans. Med. Imag.* **35**, 1962–1971 (2016).
2. Abousamra, S. et al. Weakly-supervised deep stain decomposition for multiplex IHC images. In *Proc. 2020 IEEE 17th International Symposium on Biomedical Imaging* (ISBI) 481–485 (IEEE, 2020).
3. Fassler, D. J. et al. Deep learning-based image analysis methods for brightfield-acquired multiplex immunohistochemistry images. *Diagn. Pathol.* **15**, 100 (2020).
4. Chang Colin Tan, W. et al. Overview of multiplex immunohistochemistry/immunofluorescence techniques in the era of cancer immunotherapy. *Cancer Commun.* **40**, 135–153 (2020).
5. Yeong, J. et al. Multiplex immunohistochemistry/immunofluorescence (mIHC/IF) for PD-L1 testing in triple-negative breast cancer: a translational assay compared with conventional IHC. *J. Clin. Pathol.* **73**, 557–562 (2022).
6. Lu, S. et al. Comparison of biomarker modalities for predicting response to PD-1/PD-L1 checkpoint blockade: a systematic review and meta-analysis. *JAMA Oncol.* **5**, 1195–1204 (2019).
7. Caruana, R. Multitask learning. *Mach. Learn.* **28**, 41–75 (1997).
8. Kumar, N. et al. A dataset and a technique for generalized nuclear segmentation for computational pathology. *IEEE Trans. Med. Imag.* **36**, 1550–1560 (2017).
9. Huang, Z. et al. BCData: A large-scale dataset and benchmark for cell detection and counting. In *Proc. Medical Image Computing and Computer Assisted Intervention—MICCAI 2020* (eds Martel, A. L. et al.) 289–298 (Springer, 2020).
10. Koohbanani, N. A., Jahanifar, M., Tajadin, N. Z. & Rajpoot, N. NuClick: a deep learning framework for interactive segmentation of microscopic images. *Med. Image Anal.* **65**, 101771 (2020).
11. Swiderska-Chadaj, Z. et al. Learning to detect lymphocytes in immunohistochemistry with deep learning. *Med. Image Anal.* **58**, 101547 (2019).
12. Kirillov, A., He, K., Girshick, R. & Dollár, P. A unified architecture for instance and semantic segmentation. https://presentations.cocodataset.org/COCO17-Stuff-FAIR.pdf (2017).
13. Chaurasia, A. & Culurciello, E. LinkNet: exploiting encoder representations for efficient semantic segmentation. In *Proc. 2017 IEEE Visual Communications and Image Processing* (VCIP) 1–4 (IEEE, 2017).
14. He, K., Gkioxari, G., Dollár, P. & Girshick, R. Mask R-CNN. In *Proc. IEEE International Conference on Computer Vision* 2961–2969 (IEEE, 2017).
15. Zhou, Z., Rahman Siddiquee, M. M., Tajbakhsh, N. & Liang, J. UNet++: a nested U-Net architecture for medical image segmentation. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support* 3–11 (Springer, 2018).
16. Isensee, F., Jaeger, P. F., Kohl, S. A. A., Petersen, J. & Maier-Hein, K. H. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nat. Methods* **18**, 203–211 (2021).
17. Xie, W., Alison Noble, J. & Zisserman, A. Microscopy cell counting and detection with fully convolutional regression networks. *Comput. Methods Biomech. Biomed. Eng. Imag. Vis.* **6**, 283–292 (2018).
18. Chen, L.-C., Papandreou, G., Schroff, F. & Adam, H. Rethinking atrous convolution for semantic image segmentation. Preprint at https://arxiv.org/abs/1706.05587 (2017).
19. Ram, S. & Rodríguez, J. J. Size-invariant detection of cell nuclei in microscopy images. *IEEE Trans. Med. Imag.* **35**, 1753–1764 (2016).
20. Sirinukunwattana, K. et al. Locality sensitive deep learning for detection and classification of nuclei in routine colon cancer histology images. *IEEE Trans. Med. Imag.* **35**, 1196–1206 (2016).
21. Alemi Koohbanani, N., Jahanifar, M., Zamani Tajadin, N. & Rajpoot, N. NuClick: a deep learning framework for interactive segmentation of microscopic images. *Med. Image Anal.* **65**, 101771 (2020).
22. Negahbani, F. et al. PathoNet introduced as a deep neural network backend for evaluation of Ki-67 and tumor-infiltrating lymphocytes in breast cancer. *Sci. Rep.* **11**, 8489 (2021).
23. Digre, A. & Lindskog, C. The Human Protein Atlas—spatial localization of the human proteome in health and disease. *Protein Sci.* **30**, 218–233 (2021).
24. Vrabac, D. et al. DLBCL-Morph: morphological features computed using deep learning for an annotated digital DLBCL image set. *Sci. Data* **8**, 135 (2021).
25. Tschuchnig, M. E., Oostingh, G. J. & Gadermayr, M. Generative adversarial networks in digital pathology: a survey on trends and future potential. *Pattern* **1**, 100089 (2020).
26. Rivenson, Y., de Haan, K., Wallace, W. D. & Ozcan, A. Emerging advances to transform histopathology using virtual staining. *BME Frontiers* **2020**, 9647163 (2020).
27. Liu, D. et al. Unsupervised instance segmentation in microscopy images via panoptic domain adaptation and task re-weighting. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition* 4243–4252 (IEEE, 2020).
28. Cohen, J. P., Luck, M. & Honari, S. Distribution matching losses can hallucinate features in medical image translation. In *Proc. International Conference on Medical Image Computing and Computer-Assisted Intervention—MICCAI 2018* 529–536 (Springer, 2018).
29. Burlingame, E. A. et al. SHIFT: speedy histological-to-immunofluorescent translation of a tumor signature enabled by deep learning. *Sci. Rep.* **10**, 17507 (2020).
30. Mercan, C. et al. Virtual staining for mitosis detection in breast histopathology. In *Proc. IEEE International Symposium on Biomedical Imaging* (ISBI) 1770–1774 (IEEE, 2020).
31. de Haan, K. et al. Deep learning-based transformation of H&E stained tissues into special stains. *Nat. Commun.* **12**, 4884 (2021).
32. Borovec, J. et al. ANHIR: automatic non-rigid histological image registration challenge. *IEEE Trans. Med. Imag.* **39**, 3042–3052 (2020).
33. Martinez, N., Sapiro, G., Tannenbaum, A., Hollmann, T. J. & Nadeem, S. ImPartial: partial annotations for cell instance segmentation. Preprint at *bioRxiv* https://doi.org/10.1101/2021.01.20.427458 (2021).
34. Girshick, R. Fast R-CNN. In *Proc. IEEE International Conference on Computer Vision* (ICCV) 1440–1448 (IEEE, 2015).
35. Isola, P., Zhu, J.-Y., Zhou, T. & Efros, A. A. Image-to-image translation with conditional adversarial networks. In *Proc. 2017 IEEE Conference on Computer Vision and Pattern Recognition* (CVPR) 5967–5976 (IEEE, 2017).
36. Ronneberger, O., Fischer, P. & Brox, T. U-Net: convolutional networks for biomedical image segmentation. In *Proc. Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015* 234–241 (Springer, 2015).
37. Goodfellow, I. et al. Generative adversarial nets. In *Advances in Neural Information Processing Systems* 2672–2680 (NIPS, 2014).
38. Kingma, D. P. & Ba, J. Adam: a method for stochastic optimization. Preprint at https://arxiv.org/abs/1412.6980 (2014).
39. Miyato, T., Kataoka, T., Koyama, M. & Yoshida, Y. Spectral normalization for generative adversarial networks. In *International Conference on Learning Representations* (2018).

Acknowledgements
This project was supported by an MSK Cancer Center Support Grant/Core Grant (P30 CA008748) and in part by MSK DigITs Hybrid Research Initiative and NSF grants nos. CNS1650499, OAC1919752 and ICER1940302.

Author contributions
S.N., T.J.H. and P.G. conceived the study and designed the experiments. S.N. and P.G. wrote the computer codes and performed the experimental analysis. Y.L. and T.J.H. performed the IHC and multiplex staining. M.A., T.J.H. and N.G. conceived the Lap2BETA idea for nuclear envelope staining. P.G., S.N., A.K. and R.V. analysed the results. S.N., T.J.H. and P.G. prepared the manuscript with input from all co-authors. S.N. supervised the research.

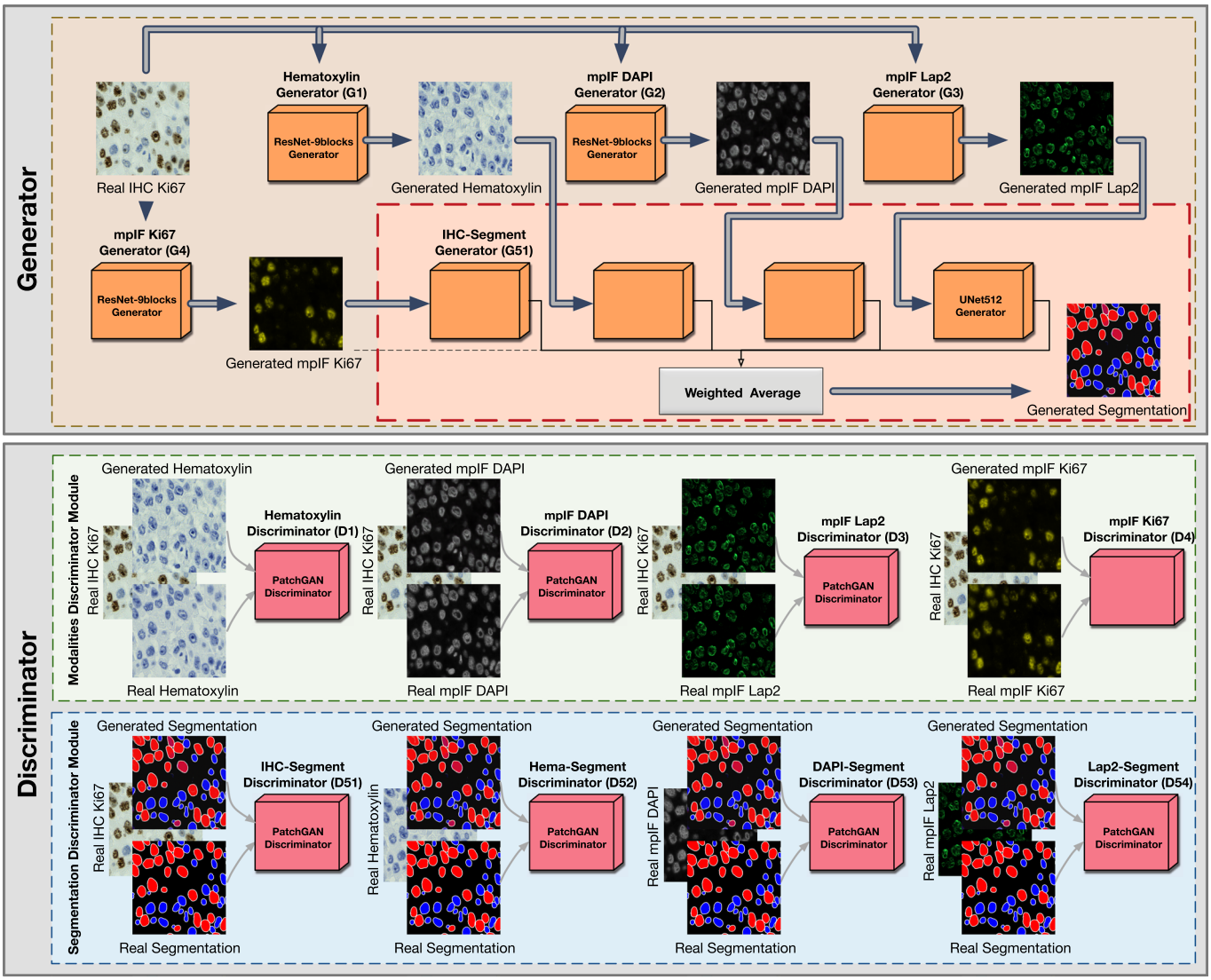
Competing interests
The authors declare no competing interests

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s42256-022-00471-x.

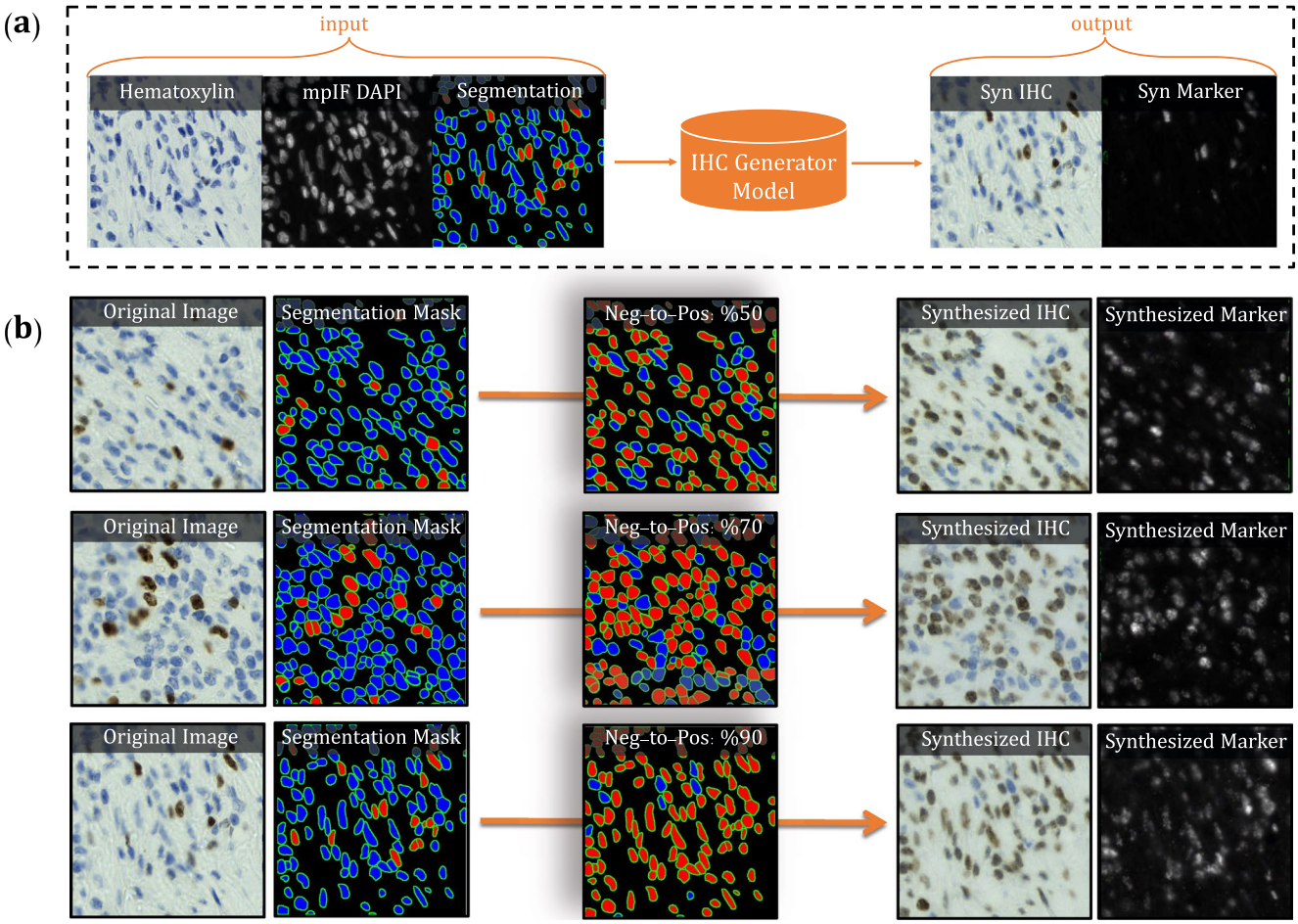
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s42256-022-00471-x.

Correspondence and requests for materials should be addressed to Travis J. Hollmann or Saad Nadeem.

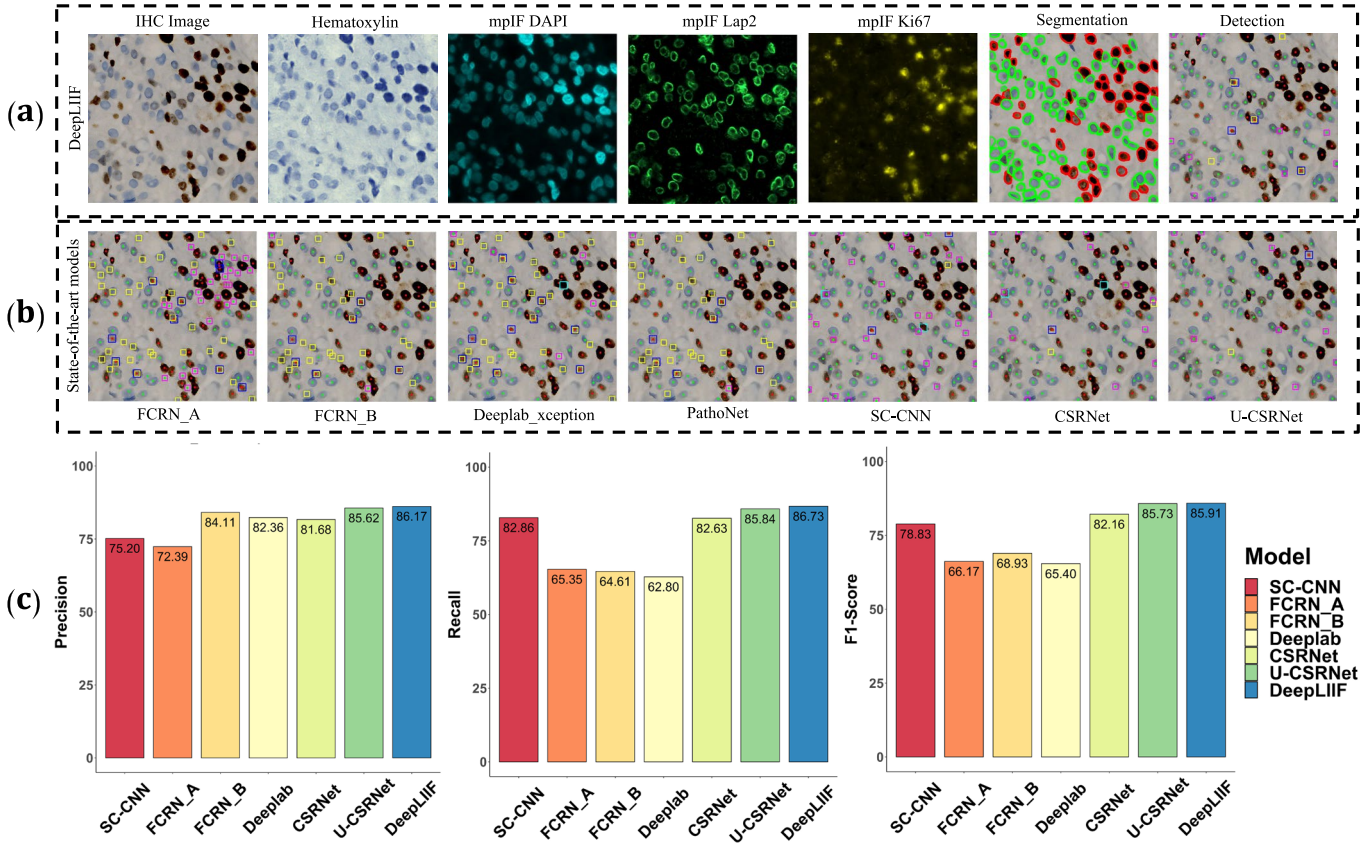
Peer review information *Nature Machine Intelligence* thanks Phedias Diamandis and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.



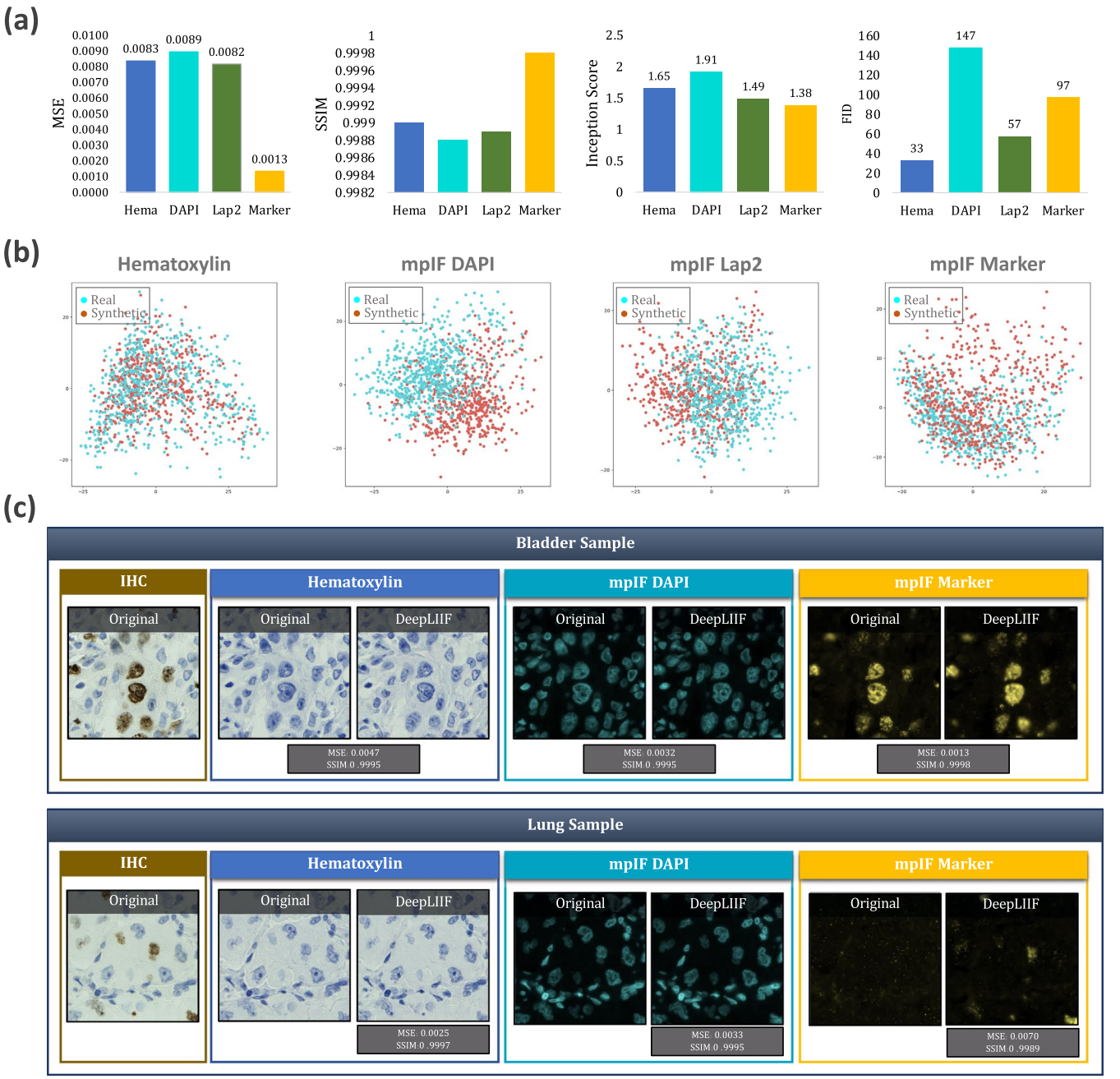
Extended Data Fig. 1 | DeepLIIF architecture diagram. Overview of DeepLIIF. The network consists of a generator and a discriminator component. It uses ResNet-9block generator for generating the modalities including Hematoxylin, mpIF DAPI, mpIF Lap2, and mpIF Ki67 and UNet512 generator for generating the segmentation mask. In the segmentation component, the generated masks from IHC, Hematoxylin, mpIF DAPI, and mpIF Lap2 representations are averaged with predefined weights to create the final segmentation mask. The discriminator component consists of the modalities discriminator module and segmentation discriminator module.



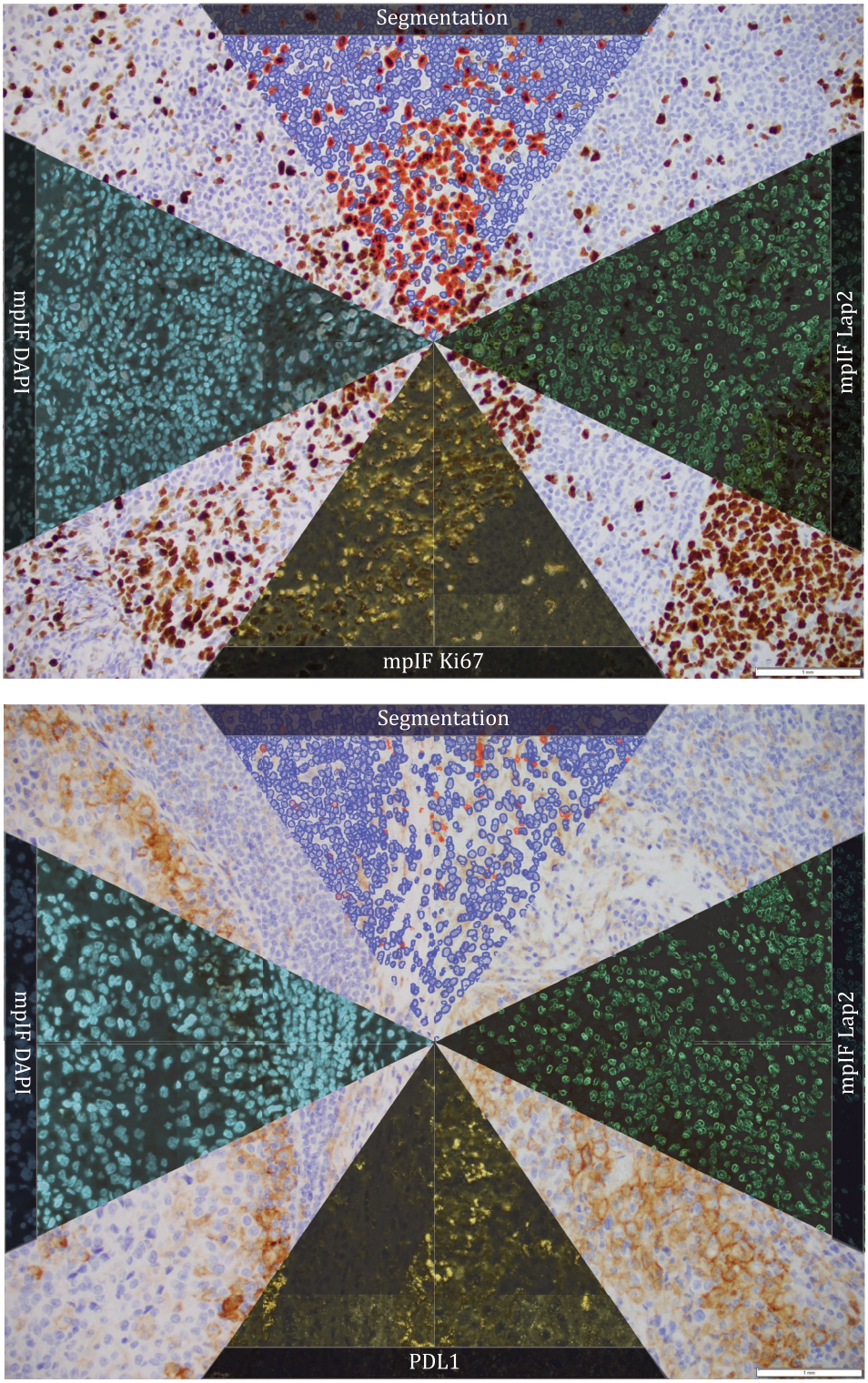
Extended Data Fig. 2 | Synthetic IHC generation pipeline. Overview of synthetic IHC image generation. (a) A training sample of the IHC-generator model. (b) Some samples of synthesized IHC images using the trained IHC-Generator model. The Neg-to-Pos shows the percentage of the negative cells in the segmentation mask converted to positive cells.



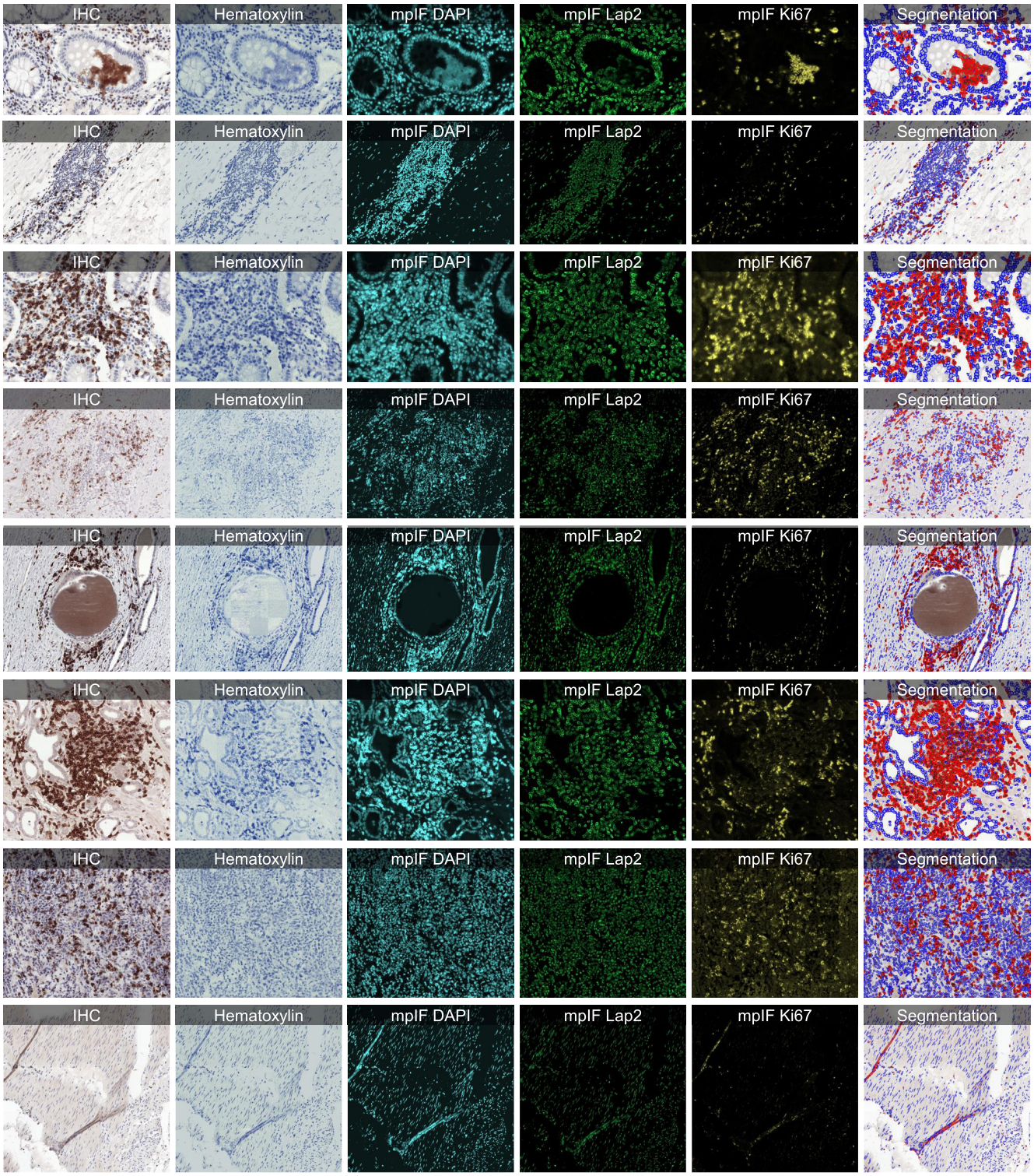
Extended Data Fig. 3 | Qualitative and quantitative analysis of DeepLIIF against detection-only deep learning models. Qualitative and quantitative analysis of DeepLIIF against detection models on the testing set of the BC Data [9]. (a) An example IHC image from the BC Data testing set, the generated modalities, segmentation mask overlaid on the IHC image, and the detection mask generated by DeepLIIF. (b) The detection masks generated by the detection models. In the detection mask, the center of a detected positive cell is shown with red dot and the center of a detected negative cell is shown with blue dot. We show the missing positive cells in cyan bounding boxes, the missing negative cells in yellow bounding boxes, the wrongly detected positive cells in blue bounding boxes, the wrongly detected negative cells in pink bounding boxes. (c) The detection accuracy is measured by getting average of precision ($\frac{TP}{TP+FP}$), recall ($\frac{TP}{TP+FN}$), and f1-score ($\frac{2\times precision\times recall}{precision+recall}$) between the predicted detection mask of each class and the ground-truth mask of the corresponding class. A predicted point is regarded as true positive if it is within the region of a ground-truth point with a predefined radius (we set it to 10 pixels in our experiment which is similar to the predefined radius in[9]). Centers that have been detected more than once are considered as false positive. Evaluation of all scores show that DeepLIIF outperforms all state-of-the-art models.



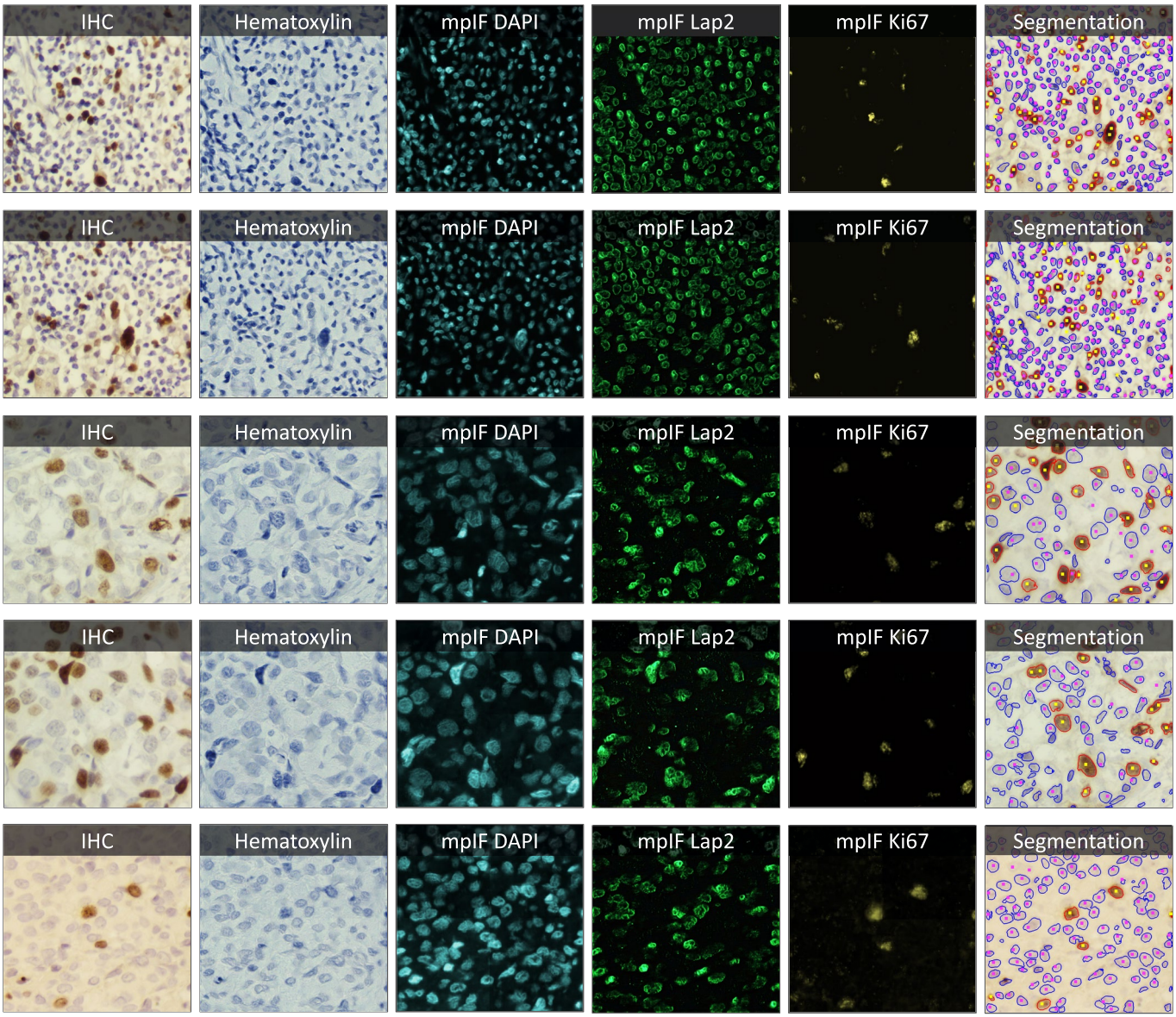
Extended Data Fig. 4 | Quantitative and qualitative analysis of DeepLIIF for modality inference. Quantitative and qualitative analysis of DeepLIIF for modality inference. (a) The Quantitative analysis of the synthetic data against the real data using MSE, SSIM, Inception Score, and FID. The low value of MSE (close to 0) and the high value of SSIM (close to 1) shows that the model generates high quality synthetic images similar to real images. (b) Visualization of first two components of PCA applied to synthetic and real images. We first, calculated a feature vector for each image using VGG16 model and then we applied PCA on the calculated feature vectors and visualized the first two components. As shown in the figure, the synthetic image data points have the same distribution as the real image data points, showing that the generated images by the model have the same characteristics as the real images. (c) The original/real and model-inferred modalities of two samples taken from Bladder and Lung tissues are shown side-by-side.



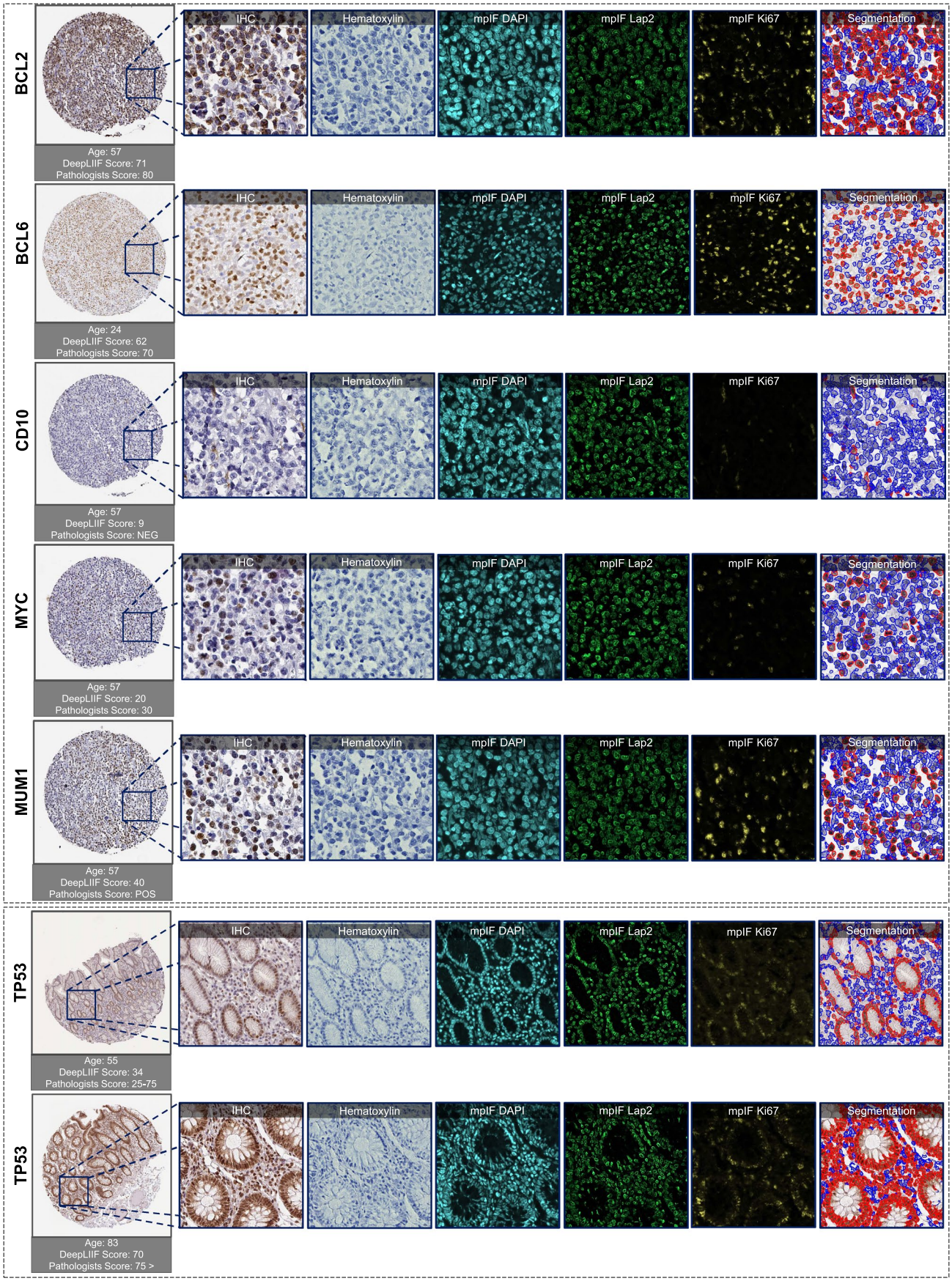
Extended Data Fig. 5 | DeepLIIF results on microscope snapshots. Microscopic snapshots of IHC images stained with two different markers along with inferred modalities and generated classified segmentation mask (top: Microscope Snapshot for **IHC Ki67** with inferred modalities and generated classified segmentation mask. bottom: Microscopic snapshots for **IHC PDL1** with inferred modalities and generated classified segmentation mask).



Extended Data Fig. 6 | DeepLIIF results on public IHC CD3/CD8 dataset. Some examples from LYON19 Challenge Dataset [11]. The generated modalities and classified segmentation mask for each sample are in a separate row.

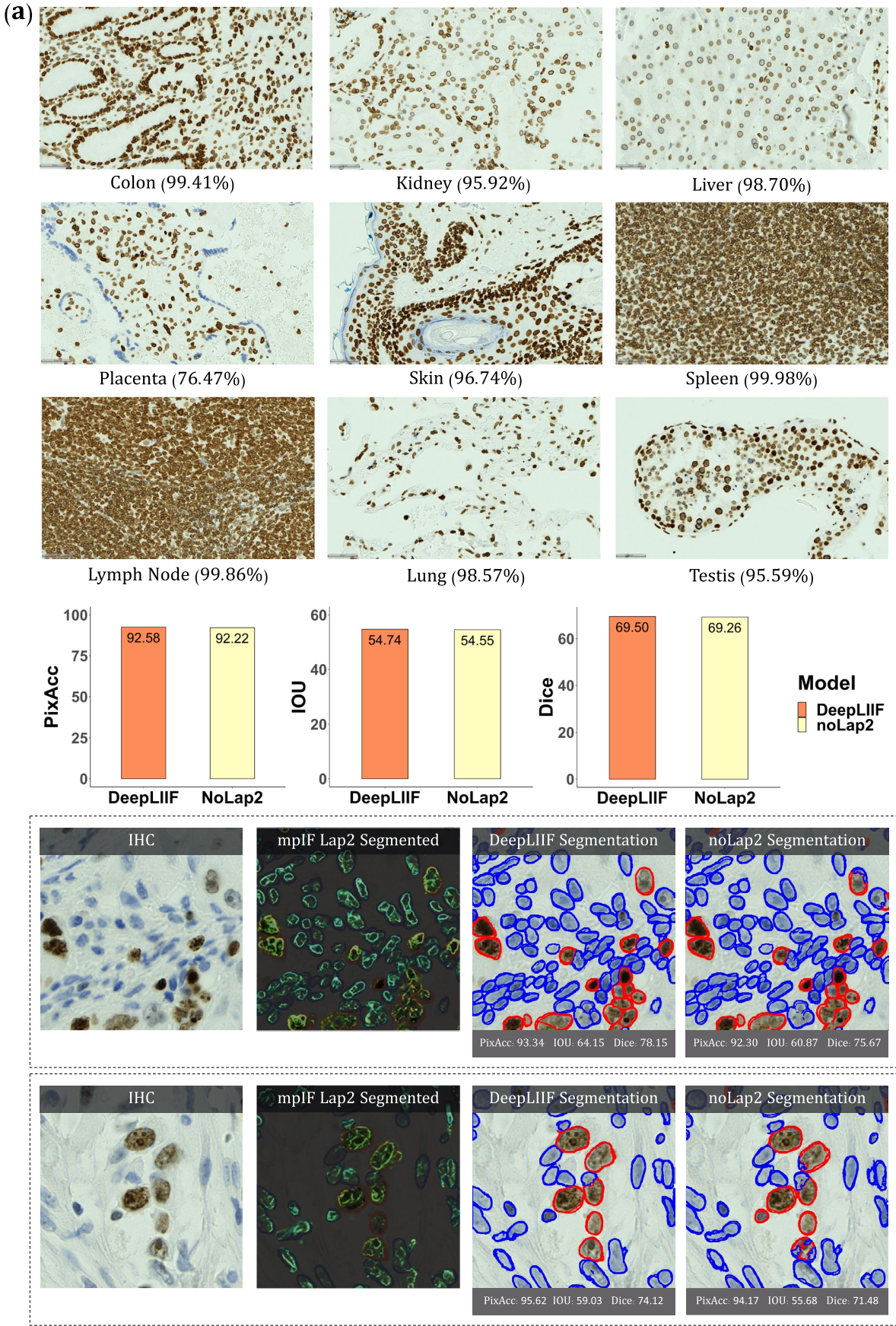


Extended Data Fig. 7 | DeepLIIF results on a different IHC Ki67 dataset with annotations based on consensus of multiple pathologists. Samples taken from the PathoNet IHC Ki67 breast cancer dataset [22] along with the inferred modalities and classified segmentation mask marked by manual centroid annotations created from consensus of multiple pathologists. The IHC images were acquired in low-resource settings with microscope camera. In each row, the sample IHC image along with the inferred modalities are shown. The overlaid classified segmentation mask generated by DeepLIIF with manual annotations are shown in the furthest right column. The blue and red boundaries represent the negative and positive cells predicted by the model, while the pink and yellow dots show the manual annotations of the negative and positive cells, respectively.



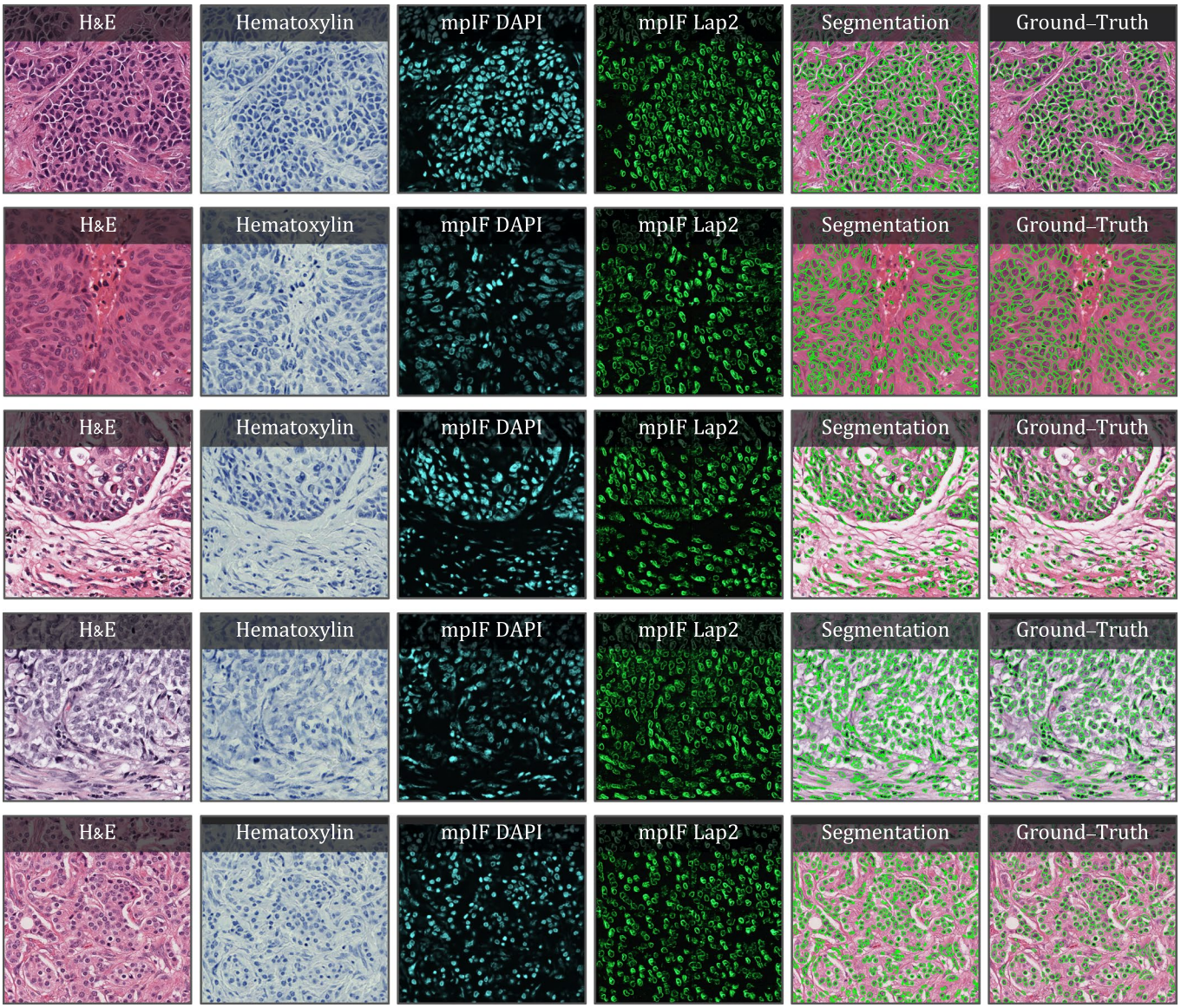
Extended Data Fig. 8 | See next page for caption.

Extended Data Fig. 8 | DeepLIIF results on DLBCL IHC markers. Examples of tissues stained with various markers. The top box shows sample tissues stained with BCL2, BCL6, CD10, MYC, and MUM1 from DLBCL-morph dataset [24]. The bottom box shows sample images stained with TP53 marker from the Human Protein Atlas [23]. In each row, the first image on the left shows the original tissue stained with a specific marker. The quantification score computed by the classified segmentation mask generated by DeepLIIF is shown on the top of the whole tissue image, and the predicted score by pathologists is shown on the bottom. In the following images of each row, the modalities and the classified segmentation mask of a chosen crop from the original tissue are shown.



Extended Data Fig. 9 | See next page for caption.

Extended Data Fig. 9 | Analysis of LAP2Beta effectiveness in DeepLIIF model. Analysis of Lap2beta effectiveness. (a) LAP2beta coverage for normal tissues. LAP2beta immunohistochemistry reveals nuclear envelope-specific staining in the majority of cells in spleen (99.98%), colon (99.41%), pancreas (99.50%), placenta (76.47%), testis (95.59%), skin (96.74%), lung (98.57%), liver (98.70%), kidney (95.92%) and lymph node (99.86%). (b) A qualitative comparison of DeepLIIF against noLap2 model. (c) Some example IHC images. The first image in each row shows the input IHC image. In the second image, the generated mpIF Lap2 image is overlaid on the classified/segmented IHC image. The third and fourth images show the segmentation mask, respectively, generated by DeepLIIF and noLap2.



Extended Data Fig. 10 | DeepLIIF generalizes out-of-the-box to H&E images. Application of DeepLIIF on some H&E sample images taken from MonuSeg Dataset [8]. We tested DeepLIIF, trained solely on IHC images stained with Ki67 marker, on H&E images. In each row, the inferred modalities and the segmentation mask overlaid on the original H&E sample are shown.

nature research

Corresponding author(s): Saad Nadeem and Travis J Hollmann

Last updated by author(s): Feb 9, 2022

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☐	☒	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	All data was generated and collected at MSKCC for the purposes of this study. This is a novel data set in which sequential stacks of biomarkers and traditional histochemical stains were performed and acquired/scanned sequentially on single slides, registered and analyzed. All techniques are described and data provided for broad use.
Data analysis	All code for DeepLIIF is original and provided at submission (https://github.com/nadeemlab/DeepLIIF). We used Python 3 and its libraries and wrote custom code for data analysis and preparation. We used Pytorch>=0.4.0 and torchvision>=0.2.1 for training and testing the designed deep leaning model.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The complete IHC Ki67 BCDataset with manual annotations is available at https://sites.google.com/view/bcdataset. Complete lymphocytes detection IHC CD3/CD8 (LYON challenge) dataset is available at https://zenodo.org/record/3385420#.XW-6JygzYuW. The NuClick IHC annotations for crops from the LYON19 dataset can be found at https://warwick.ac.uk/fac/sci/dcs/research/tia/data/nuclick/ihc_nuclick.zip. DLBCL-Morph dataset with BCL2, BCL6, MUM1, MYC, and CD10 IHCs is accessible at https://stanfordmedicine.box.com/s/ub8e0wlhsdenyhdsuuzp6zhj0i82xrb1. The high-res tiff images for TP53 IHCs can be downloaded from https://

www.proteinatlas.org/ENSG00000141510-TP53. All our internal training and testing data (acquired under the IRB protcol approval #16-1683, and source data for figures 2 and 3 (in excel files) along with the pretrained models are available at https://zenodo.org/record/4751737#.YV379XVKhH4.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	*Describe how sample size was determined, detailing any statistical methods used to predetermine sample size OR if no sample-size calculation was performed, describe how sample sizes were chosen and provide a rationale for why these sample sizes are sufficient.*
Data exclusions	*Describe any data exclusions. If no data were excluded from the analyses, state so OR if data were excluded, describe the exclusions and the rationale behind them, indicating whether exclusion criteria were pre-established.*
Replication	*Describe the measures taken to verify the reproducibility of the experimental findings. If all attempts at replication were successful, confirm this OR if there are any findings that were not replicated or cannot be reproduced, note this and describe why.*
Randomization	*Describe how samples/organisms/participants were allocated into experimental groups. If allocation was not random, describe how covariates were controlled OR if this is not relevant to your study, explain why.*
Blinding	*Describe whether the investigators were blinded to group allocation during data collection and/or analysis. If blinding was not possible, describe why OR explain why blinding was not relevant to your study.*

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☐	☒ Antibodies
☐	☐ Eukaryotic cell lines
☐	☐ Palaeontology and archaeology
☐	☐ Animals and other organisms
☐	☐ Human research participants
☐	☐ Clinical data
☐	☐ Dual use research of concern

Methods

n/a	Involved in the study
☐	☐ ChIP-seq
☐	☐ Flow cytometry
☐	☐ MRI-based neuroimaging

Antibodies

Antibodies used	Lap2-Beta (clone 27/LAP2, Becton Dickinson), Ki67 (clone SP6, Biocare), PanCK (clone AE1/AE3, DAKO)
Validation	Primary antibody staining conditions were optimized using standard immunohistochemical staining on the Leica Bond RX automated research stainer with DAB detection (Leica Bond Polymer Refine Detection DS9800). Using 4 µm formalin-fixed, paraffin-embedded tissue sections and serial antibody titrations across 12 normal tissues, the optimal antibody concentration was determined followed by transition to multiplex assay with equivalency. Optimal primary antibody stripping conditions between rounds in the 3-color assay were performed following 1 cycle of tyramide deposition followed by heat-induced stripping (see below) and subsequent chromogenic development (Leica Bond Polymer Regine Detection DS9800) with visual inspection for chromogenic product with a light microscope. The Ki67 and PanCK clones are commonly used clinical-grade reagents while Lap2-Beta expression is compatible with that described in the Human Protein Atlas with appropriate pattern distribution for a nuclear pore-specific marker.

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)	*State the source of each cell line used.*
Authentication	*Describe the authentication procedures for each cell line used OR declare that none of the cell lines used were authenticated.*
Mycoplasma contamination	*Confirm that all cell lines tested negative for mycoplasma contamination OR describe the results of the testing for mycoplasma contamination OR declare that the cell lines were not tested for mycoplasma contamination.*
Commonly misidentified lines (See ICLAC register)	*Name any commonly misidentified cell lines used in the study and provide a rationale for their use.*

Palaeontology and Archaeology

Specimen provenance	*Provide provenance information for specimens and describe permits that were obtained for the work (including the name of the issuing authority, the date of issue, and any identifying information).*
Specimen deposition	*Indicate where the specimens have been deposited to permit free access by other researchers.*
Dating methods	*If new dates are provided, describe how they were obtained (e.g. collection, storage, sample pretreatment and measurement), where they were obtained (i.e. lab name), the calibration program and the protocol for quality assurance OR state that no new dates are provided.*

☐ Tick this box to confirm that the raw and calibrated dates are available in the paper or in Supplementary Information.

Ethics oversight	*Identify the organization(s) that approved or provided guidance on the study protocol, OR state that no ethical approval or guidance was required and explain why not.*

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Animals and other organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research

Laboratory animals	*For laboratory animals, report species, strain, sex and age OR state that the study did not involve laboratory animals.*
Wild animals	*Provide details on animals observed in or captured in the field; report species, sex and age where possible. Describe how animals were caught and transported and what happened to captive animals after the study (if killed, explain why and describe method; if released, say where and when) OR state that the study did not involve wild animals.*
Field-collected samples	*For laboratory work with field-collected samples, describe all relevant parameters such as housing, maintenance, temperature, photoperiod and end-of-experiment protocol OR state that the study did not involve samples collected from the field.*
Ethics oversight	*Identify the organization(s) that approved or provided guidance on the study protocol, OR state that no ethical approval or guidance was required and explain why not.*

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Human research participants

Policy information about studies involving human research participants

Population characteristics	*Describe the covariate-relevant population characteristics of the human research participants (e.g. age, gender, genotypic information, past and current diagnosis and treatment categories). If you filled out the behavioural & social sciences study design questions and have nothing to add here, write "See above."*
Recruitment	*Describe how participants were recruited. Outline any potential self-selection bias or other biases that may be present and how these are likely to impact results.*
Ethics oversight	*Identify the organization(s) that approved the study protocol.*

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Clinical data

Policy information about clinical studies

All manuscripts should comply with the ICMJE guidelines for publication of clinical research and a completed CONSORT checklist must be included with all submissions.

Clinical trial registration	*Provide the trial registration number from ClinicalTrials.gov or an equivalent agency.*

Study protocol	*Note where the full trial protocol can be accessed OR if not available, explain why.*
Data collection	*Describe the settings and locales of data collection, noting the time periods of recruitment and data collection.*
Outcomes	*Describe how you pre-defined primary and secondary outcome measures and how you assessed these measures.*

Dual use research of concern

Policy information about dual use research of concern

Hazards

Could the accidental, deliberate or reckless misuse of agents or technologies generated in the work, or the application of information presented in the manuscript, pose a threat to:

No	Yes	
☒	☐	Public health
☒	☐	National security
☒	☐	Crops and/or livestock
☒	☐	Ecosystems
☒	☐	Any other significant area

Experiments of concern

Does the work involve any of these experiments of concern:

No	Yes	
☒	☐	Demonstrate how to render a vaccine ineffective
☒	☐	Confer resistance to therapeutically useful antibiotics or antiviral agents
☒	☐	Enhance the virulence of a pathogen or render a nonpathogen virulent
☒	☐	Increase transmissibility of a pathogen
☒	☐	Alter the host range of a pathogen
☒	☐	Enable evasion of diagnostic/detection modalities
☒	☐	Enable the weaponization of a biological agent or toxin
☒	☐	Any other potentially harmful combination of experiments and agents

ChIP-seq

Data deposition

☐ Confirm that both raw and final processed data have been deposited in a public database such as GEO.

☐ Confirm that you have deposited or provided access to graph files (e.g. BED files) for the called peaks.

Data access links *May remain private before publication.*	*For "Initial submission" or "Revised version" documents, provide reviewer access links. For your "Final submission" document, provide a link to the deposited data.*
Files in database submission	*Provide a list of all files available in the database submission.*
Genome browser session (e.g. UCSC)	*Provide a link to an anonymized genome browser session for "Initial submission" and "Revised version" documents only, to enable peer review. Write "no longer applicable" for "Final submission" documents.*

Methodology

Replicates	*Describe the experimental replicates, specifying number, type and replicate agreement.*
Sequencing depth	*Describe the sequencing depth for each experiment, providing the total number of reads, uniquely mapped reads, length of reads and whether they were paired- or single-end.*
Antibodies	*Describe the antibodies used for the ChIP-seq experiments; as applicable, provide supplier name, catalog number, clone name, and lot number.*
Peak calling parameters	*Specify the command line program and parameters used for read mapping and peak calling, including the ChIP, control and index files used.*

Data quality	*Describe the methods used to ensure data quality in full detail, including how many peaks are at FDR 5% and above 5-fold enrichment.*
Software	*Describe the software used to collect and analyze the ChIP-seq data. For custom code that has been deposited into a community repository, provide accession details.*

Flow Cytometry

Plots

Confirm that:

☐ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).

☐ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).

☐ All plots are contour plots with outliers or pseudocolor plots.

☐ A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

Sample preparation	*Describe the sample preparation, detailing the biological source of the cells and any tissue processing steps used.*
Instrument	*Identify the instrument used for data collection, specifying make and model number.*
Software	*Describe the software used to collect and analyze the flow cytometry data. For custom code that has been deposited into a community repository, provide accession details.*
Cell population abundance	*Describe the abundance of the relevant cell populations within post-sort fractions, providing details on the purity of the samples and how it was determined.*
Gating strategy	*Describe the gating strategy used for all relevant experiments, specifying the preliminary FSC/SSC gates of the starting cell population, indicating where boundaries between "positive" and "negative" staining cell populations are defined.*

☐ Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.

Magnetic resonance imaging

Experimental design

Design type	*Indicate task or resting state; event-related or block design.*
Design specifications	*Specify the number of blocks, trials or experimental units per session and/or subject, and specify the length of each trial or block (if trials are blocked) and interval between trials.*
Behavioral performance measures	*State number and/or type of variables recorded (e.g. correct button press, response time) and what statistics were used to establish that the subjects were performing the task as expected (e.g. mean, range, and/or standard deviation across subjects).*

Acquisition

Imaging type(s)	*Specify: functional, structural, diffusion, perfusion.*
Field strength	*Specify in Tesla*
Sequence & imaging parameters	*Specify the pulse sequence type (gradient echo, spin echo, etc.), imaging type (EPI, spiral, etc.), field of view, matrix size, slice thickness, orientation and TE/TR/flip angle.*
Area of acquisition	*State whether a whole brain scan was used OR define the area of acquisition, describing how the region was determined.*

Diffusion MRI ☐ Used ☐ Not used

Preprocessing

Preprocessing software	*Provide detail on software version and revision number and on specific parameters (model/functions, brain extraction, segmentation, smoothing kernel size, etc.).*
Normalization	*If data were normalized/standardized, describe the approach(es): specify linear or non-linear and define image types used for transformation OR indicate that data were not normalized and explain rationale for lack of normalization.*
Normalization template	*Describe the template used for normalization/transformation, specifying subject space or group standardized space (e.g.*

Normalization template

original Talairach, MNI305, ICBM152) OR indicate that the data were not normalized.

Noise and artifact removal

Describe your procedure(s) for artifact and structured noise removal, specifying motion parameters, tissue signals and physiological signals (heart rate, respiration).

Volume censoring

Define your software and/or method and criteria for volume censoring, and state the extent of such censoring.

Statistical modeling & inference

Model type and settings

Specify type (mass univariate, multivariate, RSA, predictive, etc.) and describe essential details of the model at the first and second levels (e.g. fixed, random or mixed effects; drift or auto-correlation).

Effect(s) tested

Define precise effect in terms of the task or stimulus conditions instead of psychological concepts and indicate whether ANOVA or factorial designs were used.

Specify type of analysis: ☐ Whole brain ☐ ROI-based ☐ Both

Statistic type for inference
(See Eklund et al. 2016)

Specify voxel-wise or cluster-wise and report all relevant parameters for cluster-wise methods.

Correction

Describe the type of correction and how it is obtained for multiple comparisons (e.g. FWE, FDR, permutation or Monte Carlo).

Models & analysis

n/a	Involved in the study
☐	☐ Functional and/or effective connectivity
☐	☐ Graph analysis
☐	☐ Multivariate modeling or predictive analysis

Functional and/or effective connectivity

Report the measures of dependence used and the model details (e.g. Pearson correlation, partial correlation, mutual information).

Graph analysis

Report the dependent variable and connectivity measure, specifying weighted graph or binarized graph, subject- or group-level, and the global and/or node summaries used (e.g. clustering coefficient, efficiency, etc.).

Multivariate modeling and predictive analysis

Specify independent variables, features extraction and dimension reduction, model, training and evaluation metrics.