nature machine intelligence

Article

https://doi.org/10.1038/s42256-024-00889-5

Accelerating histopathology workflows with generative AI-based virtually multiplexed tumour profiling

Received: 4 December 2023

Accepted: 29 July 2024

Published online: 09 September 2024

Check for updates

Pushpak Pati[1], Sofia Karkampouna[2,3], Francesco Bonollo[2], Eva Compérat[4], Martina Radić[2], Martin Spahn[5,6], Adriano Martinelli[1,7,8], Martin Wartenberg[9], Marianna Kruithof-de Julio[2,3,10] ✉ & Marianna Rapsomaniki[1,8,11] ✉

Understanding the spatial heterogeneity of tumours and its links to disease initiation and progression is a cornerstone of cancer biology. Presently, histopathology workflows heavily rely on hematoxylin and eosin and serial immunohistochemistry staining, a cumbersome, tissue-exhaustive process that results in non-aligned tissue images. We propose the VirtualMultiplexer, a generative artificial intelligence toolkit that effectively synthesizes multiplexed immunohistochemistry images for several antibody markers (namely AR, NKX3.1, CD44, CD146, p53 and ERG) from only an input hematoxylin and eosin image. The VirtualMultiplexer captures biologically relevant staining patterns across tissue scales without requiring consecutive tissue sections, image registration or extensive expert annotations. Thorough qualitative and quantitative assessment indicates that the VirtualMultiplexer achieves rapid, robust and precise generation of virtually multiplexed imaging datasets of high staining quality that are indistinguishable from the real ones. The VirtualMultiplexer is successfully transferred across tissue scales and patient cohorts with no need for model fine-tuning. Crucially, the virtually multiplexed images enabled training a graph transformer that simultaneously learns from the joint spatial distribution of several proteins to predict clinically relevant endpoints. We observe that this multiplexed learning scheme was able to greatly improve clinical prediction, as corroborated across several downstream tasks, independent patient cohorts and cancer types. Our results showcase the clinical relevance of artificial intelligence-assisted multiplexed tumour imaging, accelerating histopathology workflows and cancer biology.

Tissues are spatially organized ecosystems, where cells of diverse phenotypes, morphologies and molecular profiles coexist with non-cellular compounds and interact to maintain homeostasis[1]. Several tissue staining technologies are used to interrogate this intricate tissue architecture. Among these, hematoxylin and eosin (H&E) is the undisputed workhorse, routinely used to assess aberrations in tissue morphology linked to disease in histopathology workflows[2]. A notable example is cancer, where H&E staining can reveal abnormal cell proliferation, lymphovascular invasion and immune cell infiltration, among others. Complementary to the morphological information available via H&E, immunohistochemistry (IHC)[3] can detect and quantify the distribution and localization of specific markers within cell compartments and within their proper histological context, crucial for tumour subtyping, prognosis and personalized treatment selection. As tissue restaining in conventional IHC is limited, repeated serial sections stained with different antibodies are required for in-depth tumour profiling,

A full list of affiliations appears at the end of the paper. ✉e-mail: marianna.kruithofdejulio@unibe.ch; marianna.rapsomaniki@unil.ch

a time-consuming and tissue-exhaustive process, prohibitive in cases of limited tissue availability. Additionally, serial IHC staining yields unaligned, non-multiplexed images occasionally of suboptimal quality due to artefacts, and tissue unavailability may lead to missing stainings (Fig. 1a). Recently, multiplexed imaging technologies[4–6] have enabled the simultaneous quantification of dozens of markers on the same tissue, revolutionizing spatial biology[7]. Still, their high cost, cumbersome experimental process, tissue-destructive nature and need for specialized equipment severely limit clinical adoption.

Virtual staining—that is, artificially staining tissue images using generative artificial intelligence (AI)—has emerged as a promising cost-effective, accessible and rapid alternative that addresses the above limitations[8,9]. A virtual staining model is trained on two sets of images—a source and a target set—and learns the source-to-target appearance mapping[10,11] so as to simulate the target staining on the source, ultimately producing at inference time a virtual target image. Initial virtual staining models were based on different flavours of generative adversarial networks (GANs) operating under a paired setting: that is, they depended on precisely aligned source and target images, which allowed them to directly optimize a pixel-wise loss between the virtual and real images[12]. Successful examples of paired models include translating label-free microscopy to H&E and specific stainings[13–16], H&E to special stains[17,18], H&E to IHC[19,20] and IHC to multiplex immunofluorescence[21]. However, as tissue restaining is not routinely done, paired models depend on aligning tissue slices via image registration, a time-consuming and error-prone process, often infeasible in practice because of substantial discrepancies even between consecutive slices. Additionally, as tissue architecture largely alters after the first slices, retrospective addition of new markers is impossible. To circumvent these limitations, unpaired stain-to-stain (S2S) translation models have recently emerged, with early applications in translating from H&E to IHC[22–26] and special staining[27,28] and from cryosections to formalin-fixed paraffin-embedded (FFPE) sections[29]. The vast majority of unpaired models are inspired by CycleGAN[30]; they depend on an adversarial loss to preserve the source content and a cycle consistency loss to preserve the target style. Some employ additional constraints: for example, domain-invariant content and domain-variant style[22], perceptual embeddings[24] or structural similarity[25].

An important limitation of CycleGAN-based models is that cycle consistency assumes a bijective mapping between the source and target domains[30], which does not hold for many S2S translation tasks. As a result, a persistent problem is staining unreliability, observed as incorrect mappings across domains: for example, positive signals from the source domain are mapped to negative signals from the target domain. To account for staining unreliability, recent works guide the translations via expert annotations: ref. 26 translates H&E to cytokeratin-stained IHC using expert annotations of positive and negative metastatic regions on the H&E images, and ref. 25 translated H&E to Ki67-stained IHC by leveraging cancer and normal region annotations in both H&E and IHC images. Although these approaches show promising results for these specific translation tasks, acquiring such annotations is impractical when translating to several IHC markers and infeasible even for experienced pathologists for specialized tasks (for example, identifying $p53^+$ cells in H&E images). To circumvent the annotation challenge, ref. 31 recently introduced a semisupervised approach, which, however, again depends on image registration. Consequently, there is a great need for unpaired S2S translation models that preserve staining consistency without needing consecutive tissue sections, image registration or extensive annotations on the source domain.

Regardless of the underlying modelling assumptions, another important limitation of S2S translation methods concerns evaluation. As ground-truth and virtually generated images are not pixel-wise aligned, S2S translation quality is typically quantified at a high feature level using inception-based scores[32]. However, these scores do not guarantee accurate preservation of complex and biologically meaningful patterns[9]. To alleviate these concerns, some studies employ qualitative assessment through pathological examination of the virtual images[22,24]. Still, a persistent concern is the presence of hallucinations in virtual images[33] that might otherwise appear realistic even to experienced pathologists. Ultimately, to ensure that virtual images not only visually appear realistic but also are useful from a clinical standpoint, using them as input to downstream models that predict clinical endpoints could provide an unbiased, convincing validation[9].

Here, we propose the VirtualMultiplexer, a generative toolkit that translates H&E images to matching IHC images for a variety of markers (one IHC marker at a time) (Fig. 1a,b). The VirtualMultiplexer is inspired by contrastive unpaired translation (CUT)[34], an appealing alternative to CycleGAN that achieves content preservation by maximizing the mutual information between target and source domains. Our toolkit does not necessitate pixel-wise aligned H&E and IHC images and, in contrast to existing approaches, requires minimal expert annotations only on the IHC domain. To ensure biological consistency, the VirtualMultiplexer introduces an architecture with multiscale constraints at the single-cell, cell-neighbourhood and whole-image level that closely mimics human expert evaluation. We trained the VirtualMultiplexer on a prostate cancer tissue microarray (TMA) containing unpaired H&E and IHC images for six clinically relevant nuclear, cytoplasmic and membrane-targeted markers. We evaluated the generated images using quantitative fidelity metrics, expert pathological assessment and visual Turing tests and assessed their clinical relevance by predicting clinical endpoints (Fig. 1c). We successfully transferred the model across tissue image scales and out-of-distribution patient cohorts and demonstrated its potential to transfer across tissue types (Fig. 1d). Our results suggest that the VirtualMultiplexer generates realistic, indistinguishable from real, multiplexed IHC images of high quality, outperforming existing methods. Using the virtually multiplexed datasets improves the prediction of clinical endpoints not only in the training cohort but also in two independent prostate cancer patient cohorts and a pancreatic ductal adenocarcinoma (PDAC) cohort, with important implications in histopathology.

Results

VirtualMultiplexer is a virtually multiplexed staining toolkit

The VirtualMultiplexer is a generative toolkit for unpaired S2S translation, trained on unpaired real H&E (source) and IHC (target) images (Fig. 2; detailed description in Methods). During training, each image is split into patches that are fed into a generator network G that conditions on input H&E and IHC and learns to transfer the staining pattern, as captured by IHC, to the tissue morphology, as captured by H&E. The generated IHC patches are stitched together to create a virtual IHC image (Fig. 2a). We train an independent one-to-one VirtualMultiplexer model for each IHC marker at a time. To ensure staining reliability, we propose a multiscale approach, designed to accurately learn staining specificity at a single-cell level and content and style preservation at a cell-neighbourhood and whole-image level, which involves jointly optimizing three distinct loss functions (Fig. 2b). The neighbourhood loss (1) ensures that generated IHC patches are indistinguishable from real IHC patches and consists of an adversarial and a multilayer contrastive loss (Fig. 2b), adopted from CUT[34]. The adversarial loss $\mathcal{L}_{\text{adv}}$ (1a) is a standard GAN loss[35], where real and virtual IHC patches are used as input to patch discriminator D, which attempts to classify them as either real or virtual, eliminating style differences. The multilayer contrastive loss (1b) is based on a patch-level noise contrastive estimation (NCE) loss[34] $\mathcal{L}_{\text{contrastive}}$ that ensures that the content of corresponding real H&E and virtual IHC patches is preserved across multiple layers of G_{enc}: that is, the encoder of the generator G. The VirtualMultiplexer introduces two losses: a global consistency loss and a local consistency loss (Fig. 2b). The global consistency loss (2) uses a feature extractor F and enforces content consistency between real H&E and virtual IHC images ($\mathcal{L}_{\text{content}}$) and style consistency between real and virtual IHC images ($\mathcal{L}_{\text{style}}$) across multiple layers of F. The local consistency loss

https://doi.org/10.1038/s42256-024-00889-5

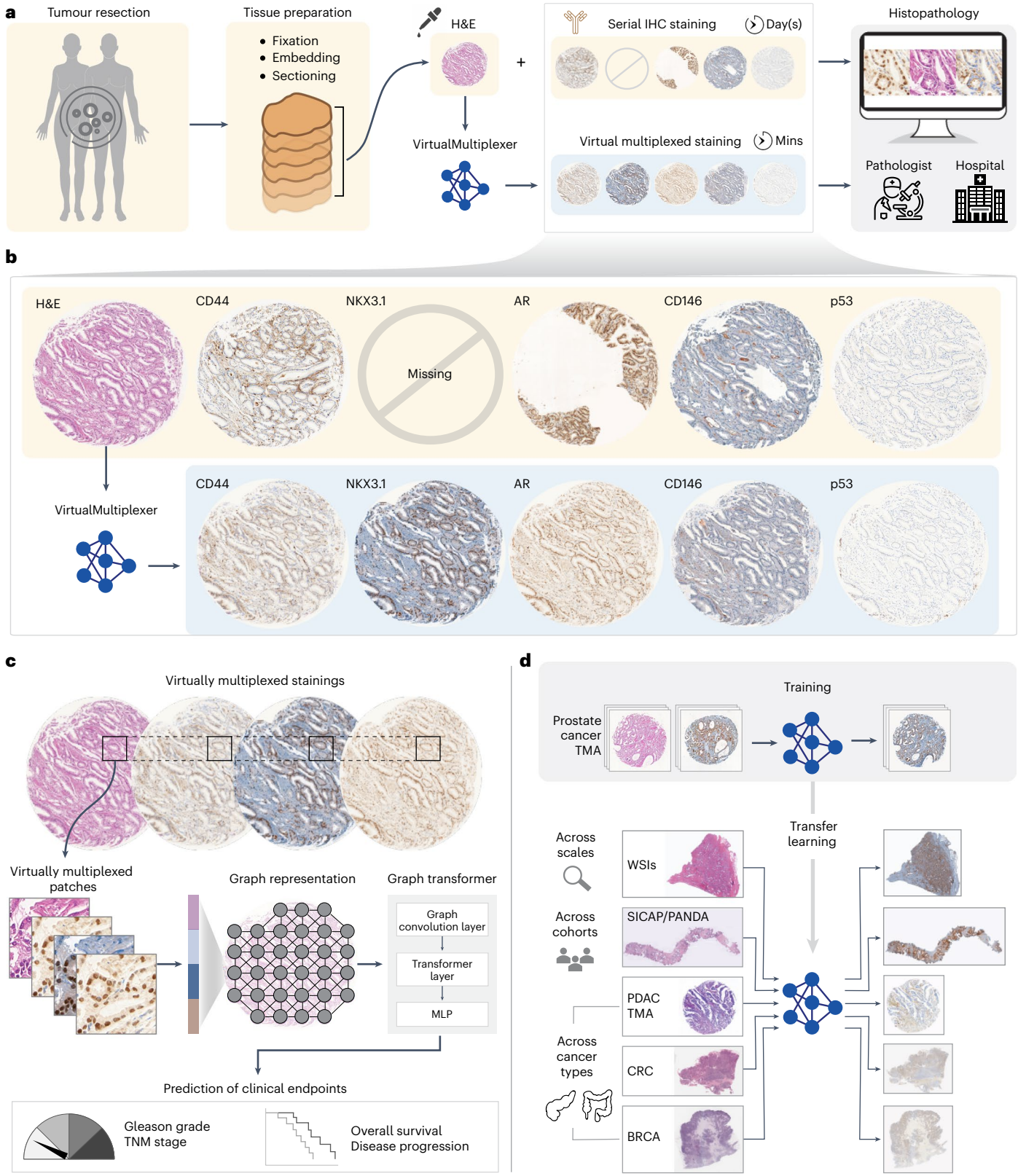


Fig. 1 | VirtualMultiplexer is a generative toolkit for synthesizing virtual multiplexed staining. a, In a typical histopathology workflow, serial tissue sections from a tumour resection are stained with H&E and IHC to highlight tissue morphology and molecular expression of several markers of interest. This time-consuming and tissue-exhaustive process yields unpaired tissue slides that bear the technical risk of suboptimal quality in terms of missing stainings, tissue artefacts and unaligned tissues. **b**, To mitigate these issues, the VirtualMultiplexer uses generative AI to rapidly render, from a real input H&E image, consistent, reliable and pixel-wise aligned IHC stainings. **c**, As the generated images are now virtually multiplexed, they are further exploited to train early fusion graph transformers able to predict several clinically relevant endpoints. **d**, The VirtualMultiplexer was successfully transferred across image scales and patient cohorts and showed potential in being transferred to other tissue types, accelerating clinical applications and discovery.

https://doi.org/10.1038/s42256-024-00889-5

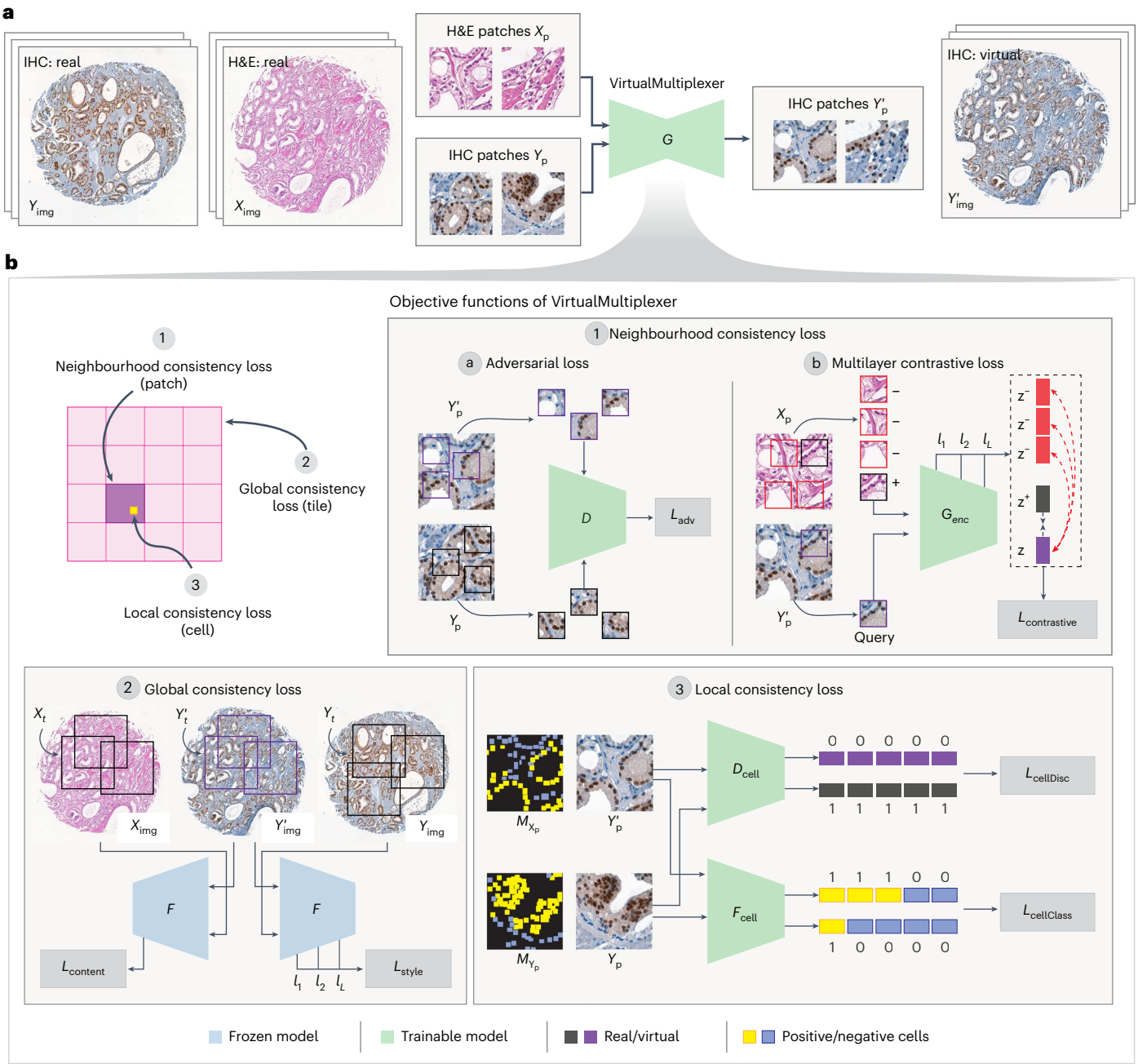


Fig. 2 | Overview of the VirtualMultiplexer architecture. a, The VirtualMultiplexer consists of a generator *G* that takes as input real unpaired H&E and IHC images and is trained to perform S2S translation by mapping the staining distribution of IHC onto H&E while preserving tissue morphology, ultimately generating virtually multiplexed synthetic IHC images only from input H&E images. **b**, During training, the VirtualMultiplexer optimizes several losses that enforce consistent S2S translation at multiple scales, including (1) a neighbourhood consistency loss that ensures indistinguishable translations at a neighbourhood (patch) level, (2) a global consistency loss that ensures that the model accurately captures content and style constraints at a global tile level and (3) a local consistency loss that encodes biological priors on cell type classification and discriminator constraints at a cellular level.

(3) enables the model to capture a realistic appearance and staining pattern at the cellular level while alleviating the multi-subdomain mappings. This is achieved by leveraging prior knowledge on staining status via expert annotations and training two separate networks: a cell discriminator D_{cell} that eliminates differences in the style of real and virtual cells ($\mathcal{L}_{cellDisc}$) and a cell classifier F_{cell} that predicts the staining status and thus enforces staining consistency at a cell level ($\mathcal{L}_{cellClass}$).

Performance assessment of the VirtualMultiplexer

We trained the VirtualMultiplexer on a prostate cancer cohort from the European Multicenter Prostate Cancer Clinical and Translational Research Group (EMPaCT) TMA[36–38] (Methods). The cohort contained unpaired H&E and IHC images from 210 patients with four cores per patient for six clinically relevant markers: androgen receptor (AR), NK3 Homeobox 1 (NKX3.1), CD44, CD146, p53 and ERG. The VirtualMultiplexer generated virtual IHC images that preserved the tissue morphology of the real H&E image and the staining pattern of the real IHC image (Fig. 3a–c; additional examples in Extended Data Fig. 1). We benchmarked the VirtualMultiplexer with four state-of-the-art unpaired S2S translation methods: CycleGAN[30], CUT[34], CUT with kernel instance normalization (KIN)[39] and AI-FFPE[29] using the Fréchet inception distance (FID), an established metric used to assess the quality of

https://doi.org/10.1038/s42256-024-00889-5

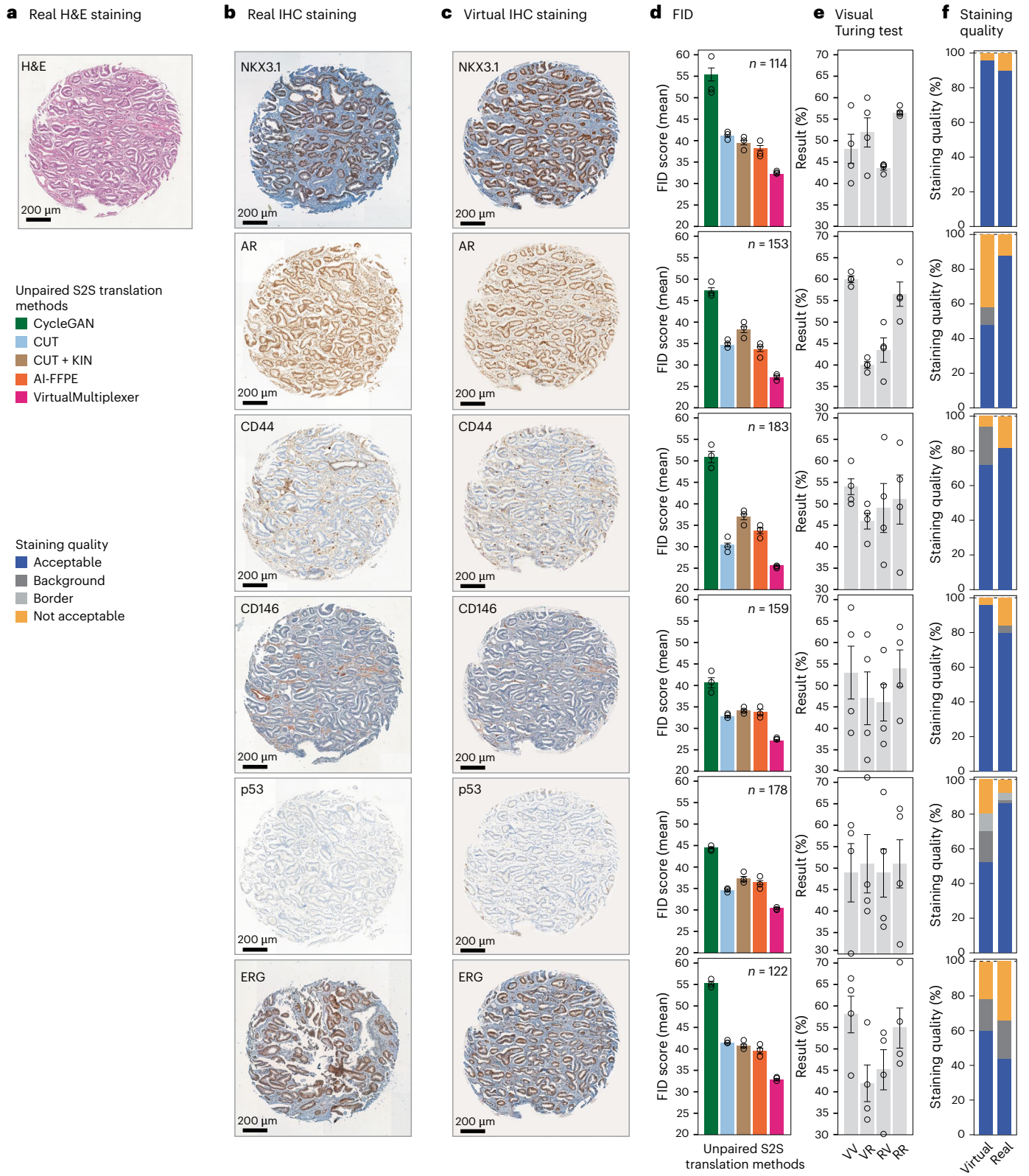


Fig. 3 | Performance evaluation of the VirtualMultiplexer. a, Example H&E core from the EMPaCT TMA. **b**, Real, unpaired IHC-stained cores for different antibody markers corresponding to the H&E core in **a**. **c**, Virtually stained IHC cores, now paired with the H&E core in **a**. **d**, Comparison of the VirtualMultiplexer with state-of-the-art S2S models. Barplots and error bars indicate the mean and standard deviation of the FID score from three independent runs of each model. Number of test samples used varies per marker and is reported in each subplot. **e**, Results of the visual Turing test, where circles indicate results of the guess of each one of the $n = 4$ experts, and barplots and error bars indicate the corresponding mean and standard variation. **f**, Assessment of staining quality of the virtual and real stainings, performed on 50 real and 50 virtual images. RR, real as real; RV, real as virtual; VR, virtual as real; VV, virtual as virtual.

https://doi.org/10.1038/s42256-024-00889-5

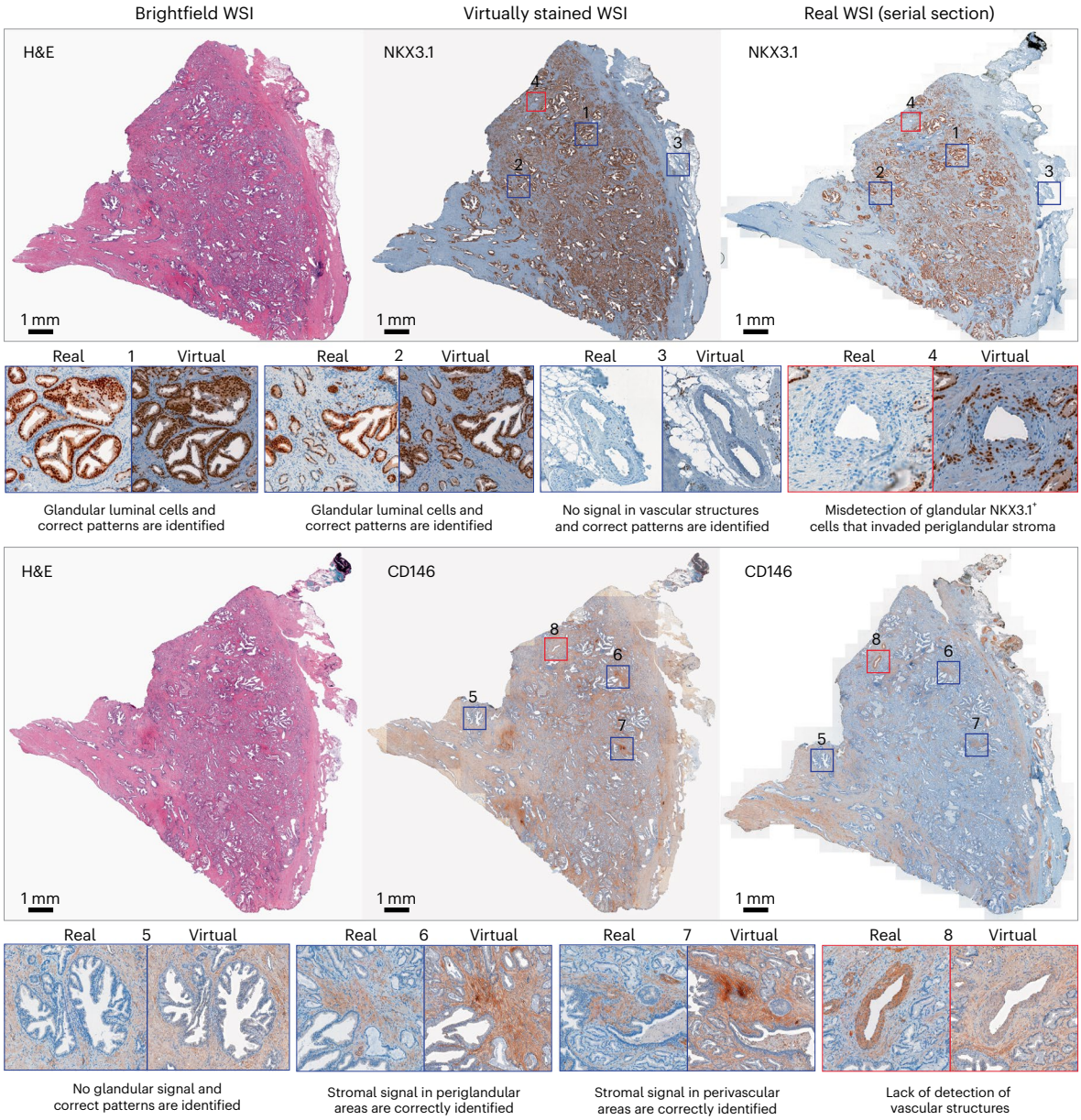


Fig. 4 | Transfer learning from TMA to WSIs of prostate cancer tissue. Example of H&E (left), virtual IHC (middle) and real IHC (right) staining for NKX3.1 (top) and CD146 (bottom) of prostate cancer tissue WSIs. Blue-framed zoomed-in regions display accurate staining pattern. Red-framed zoomed-in regions display examples of virtual staining mispredictions.

AI-generated images[40] (Methods). The VirtualMultiplexer resulted in the lowest FID score across all markers (Fig. 3d), with an average value of 29.2 (±3), consistently lower than CycleGAN (49 ± 6), CUT (35.8 ± 4.5), CUT with KIN (37.8 ± 2.3) and AI-FFPE (35.9 ± 2.6). We also used the contrast-structure similarity score, a variant of the structural similarity score that computes contrast and structure preservation[25], where again the VirtualMultiplexer surpassed all other models in performance (Supplementary Table 1). These results indicated that virtual images generated by the VirtualMultiplexer were closer to the real ones in terms of distribution than any of the competing methods.

To further quantify the indistinguishability of real and virtual images, we conducted a visual Turing test: three experts in prostate histopathology and one board-certified pathologist were shown 100 randomly selected patches per marker, with 50 of them originating from real and 50 from virtual IHC images, and were asked to classify each patch as virtual or real. Our model was able to trick the experts, as they achieved a close-to-random average sensitivity of 52.1% and specificity of 54.1% across all markers (Fig. 3e). Last, we performed a staining quality assessment: we gave the pathologist 50 real and 50 virtual images per marker, revealing which were real and virtual; the pathologist performed a qualitative assessment of the staining, as judged by overall expression levels, background, staining pattern, cell type specificity and subcellular localization (Fig. 3f; detailed annotations in Supplementary Data 1). Across all markers, on average 70.7% of the virtual images reached an acceptable staining quality, as opposed to 78.3% of the real images. The results varied depending on the marker, with virtual NKX3.1 and CD146 images achieving the highest quality of 96%, surpassing even real images. Conversely, virtual AR images had the lowest score of 46%, with an additional 10% exhibiting accurate staining but high background, and the remaining 42% rejected mostly due to heterogeneous staining or falsely unstained cells. Background presented a challenge with CD44 and p53; the latter appeared to be further affected by border artefacts—that is, the presence of abnormally highly stained cells only in the core border—also occasionally present in

real images. ERG achieved a higher staining quality in virtual than in real images, which both often faced background issues. We concluded that for most markers, the staining quality scores and the number of cores with staining artefacts were comparable in virtual versus real images.

Following these observations, we carefully examined if virtual images capture accurate staining patterns. Overall, for all markers, we observed similar patterns, correct cell types and subcellular distributions (Extended Data Fig. 2a). Certain discrepancies were also found, such as systematic lack of recognition of CD146$^+$ vascular structures (Extended Data Fig. 2b). Nonetheless, the more pathologically relevant paterns, crucial for diagnostic applicability, were correctly reconstructed. We also compared the staining intensity of positive and negative cells and observed high concordance between class-wise intensity distributions and separability for both real and virtual images, confirming that the virtual images faithfully capture the staining intensity for both cell classes (Extended Data Fig. 3). Finally, we performed an ablation study demonstrating the effects of different components of the VirtualMultiplexer loss (Extended Data Fig. 4). The mere imposition of the neighbourhood consistency (the primary objective in competing methods) leads to obvious staining unreliability: for example, swapping of staining patterns between positive and negative cells. Our global consistency clearly mitigates this, and our local consistency further optimizes the virtual staining at the cell level.

Transferring from TMAs to WSIs

To assess how well the model can be transferred across imaging scales, we fed the TMA-trained VirtualMultiplexer with five out-of-distribution H&E-stained prostate whole-slide images (WSIs) and generated virtual IHC images for NKX3.1, AR and CD146. We then stained for the same markers by IHC on the direct serial sections, thus generating ground-truth and directly comparable WSIs to visually validate the model predictions (Methods). For NKX3.1 (Fig. 4), the virtual images largely captured the staining appearance of the real ones, both in terms of specific glandular luminal cell identification (positive signal) (examples 1 and 2 in Fig. 4 and Extended Data Fig. 5) and accurate non-annotation of stromal or vascular structures (absence of signal) (example 3 in Fig. 4 and Extended Data Fig. 5). In minority, virtual images did not highlight the rarer NKX3.1$^+$ cell population that are not part of the epithelial gland, but rather in the periglandular stroma (example 4 in Fig. 4 and Extended Data Fig. 5). For CD146 and AR, we observed intensity discrepancies between virtual and real images, more striking for CD146 where the overall signal intensity and background are higher in virtual versus real images (Fig. 4 and Extended Data Fig. 5). These discrepancies can be attributed to the fact that the training set TMA images have a different staining distribution than the WSIs. Although this might lead to false interpretation of marker expression levels at a first inspection, when evaluating at higher magnification, the staining pattern in the matching real and virtual regions was effectively correct: for example, no glandular signal (example 5 in Fig. 4) and appropriate stromal localization of CD146 (examples 6 and 7 in Fig. 4) and nuclear localization of AR in luminal epithelial cells (example 5 in Extended Data Fig. 5). Lack of detection of vascular structures for CD146 was evident in both TMA cores and WSI (example 8 in Fig. 4).

The VirtualMultiplexer improves clinical predictions

We then assessed the utility of the generated stainings in augmenting the performance of AI models when predicting clinically relevant endpoints. Specifically, we encoded the real H&E, real IHC or virtual IHC images as tissue-graph representations and employed a graph transformer (GT)[41] to map the representations to downstream class labels (Fig. 5a,b and Methods). We trained the GT model under three settings (Fig. 5c): (1) a unimodal setting, where independent GT models were trained for each H&E and IHC marker; (2) a multimodal late fusion setting, where the outputs of independent GT models were fused at the last embedding stage, and (3) a multimodal early fusion setting, where the patch features were combined early in the tissue graph and fed into the GT model. Whereas the unimodal setting resulted in a separate prediction per marker, both multimodal settings combined the patch features, resulting in a single prediction. In contrast to the late fusion multimodal setting, in the early fusion case only one model that learns from the joint spatial distribution across all markers was trained, mimicking a multiplexed imaging scenario. With the exception of the early fusion setting that is only feasible for virtual images, we tested all three settings with both real and virtual images as input, resulting in a total of five different combinations (Fig. 5d, legend).

We applied these settings to the EMPaCT dataset to predict patient overall survival status and disease progression (Fig. 5d and Methods). As small discrepancies in the number of real IHC images available were present due to missing stainings, we matched the number of virtual IHC images to the number of available real IHC images to ensure a fair comparison between real and virtual unimodal models (dark and light blue barplots in Fig. 5d, respectively). As H&E images were always available, the unimodal model trained on H&E had a slight advantage over all other models in terms of number of samples used. To compare all multimodal models, we again matched the number of virtual images to the available real data, and thus the last three bars in Fig. 5d are also directly comparable. We observed that the unimodal–virtual settings are on par with the unimodal–real for both tasks, with variations in prediction performance depending on the marker. When predicting overall survival status, two interesting exceptions concern CD146 and p53: for CD146, the unimodal–virtual setting outperformed the unimodal–real, in accordance with the previous observation that virtual CD146 images achieved a higher-quality score than real ones (Fig. 3f). The opposite is true for p53: virtual p53 images were of lower quality than real p53 images, and the corresponding unimodal–virtual setting achieved a lower performance than the unimodal–real one. However, these observations were not replicated for disease progression prediction, which appeared to be an overall harder task. In both tasks, all multimodal settings outperformed the unimodal ones, including the H&E, indicating the utility of combining information from complementary markers. Furthermore, the multimodal early fusion model trained with virtual images achieved the best weighted F_1 score of 82.9% and 74.8% for overall survival status and disease progression, respectively. We also performed a marker-level interpretability analysis, pointing to markers of high importance inline with the unimodal high and low weighted F_1 scores (Extended Data Fig. 6). Overall, our results establish the potential of virtual multiplexed images in augmenting the efficacy of AI models in the prediction of clinical endpoints.

Transferring across patient cohorts and cancer types

We then assessed the model's ability to generalize to out-of-distribution data using two independent prostate cancer cohorts, SICAP[42] and prostate cancer grade assessment (PANDA)[43], containing H&E-stained needle biopsies with associated Gleason scores (Methods). We used the pretrained VirtualMultiplexer to generate IHC images for four markers relevant towards Gleason score prediction: NKX3.1, CD146, AR and ERG (Fig. 6a; additional examples in Extended Data Fig. 7). We observed that the virtual staining patterns of the IHC markers were overall correct and specific in terms of cell type and subcellular localization, with the only exception being the occasional aspecific AR signal in the extracellular matrix areas. Other inconsistencies include weak staining of interstitial tissue for CD146 and heterogeneous gland staining for ERG. We also observed some recurring issues as in the EMPaCT TMA (Fig. 3): background (for example, occasional stromal background in NKX3.1 and ERG) and border and tiling artefacts (for example, CD146). Subsequently, we trained GT models under the previous settings to predict Gleason grade (Fig. 6b,c, respectively). We observed that the predictive performance of the unimodal–virtual settings was close to or superior to the model using standalone H&E images for both datasets. Further improvement was attained by the multimodal–virtual settings, with

https://doi.org/10.1038/s42256-024-00889-5

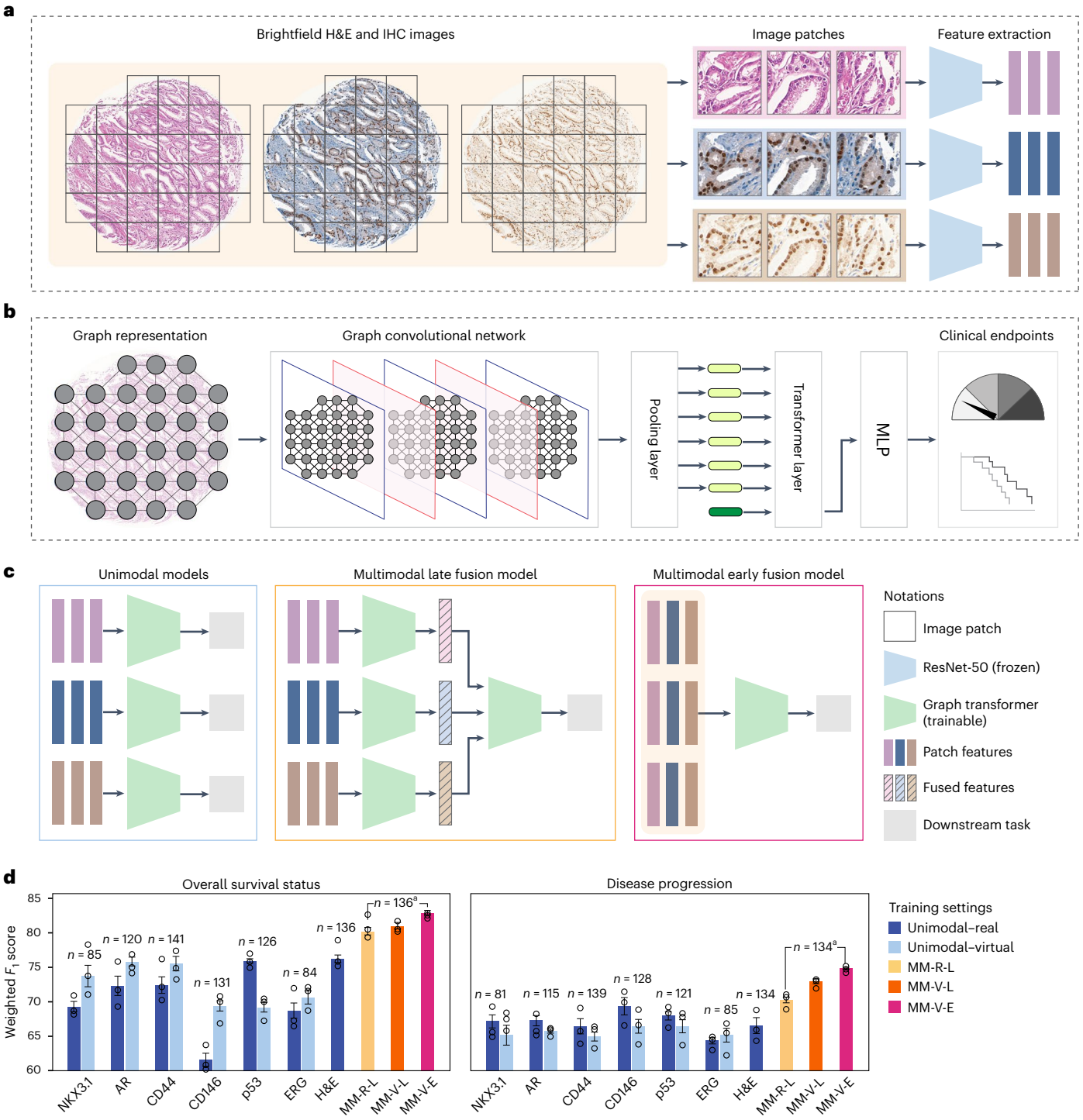


Fig. 5 | Prediction of clinically relevant downstream tasks with virtually multiplexed data. a, Patch extraction and computation of patch features with a frozen ResNet-50 model (blue trapezoid). **b**, Overview of the GT model, implemented by first constructing a patch-level graph representation, followed by a transformer that processes the graph representation to predict clinically relevant endpoints. **c**, Training of GT models (green trapezoid) under three different settings, depending on the integration strategy. **d**, Prediction results of overall survival status (left: 0, alive/censored; 1, prostate cancer related death) and disease progression (right: 0, no recurrence; 1, recurrence). Barplot colours indicate one of the five combinations of training setting and input data used (see legend). For each combination, barplot heights and error bars indicate the mean and standard deviation of the weighted F_1 score, as computed in the held-out test set from three independent runs with different initializations. The exact number of training samples used in each cases is given on the top of the barplots. [a]For all multimodal models, the reported number refers to the union across all markers. MM-R-L, multimodal–real–late fusion; MM-V-E, multimodal–virtual–early fusion; MM-V-L, multimodal–virtual–late fusion.

the early fusion model achieving the highest weighted F_1 score (SICAP, 61.4%; PANDA, 72.3%), which not only outperformed the H&E unimodal counterparts, but also WholeSIGHT[44], the previous top performing model on these datasets that achieved a weighted F_1 score of 58.6% and 67.9% on SICAP and PANDA, respectively. Finally, as for both SICAP and PANDA, ground-truth region-level annotations of Gleason scores exist,

we performed a region-based interpretability analysis and observed that the salient tissue regions contributing to model predictions coincided with the ground-truth annotations (Extended Data Fig. 8).

Finally, we evaluated the generalization ability of the VirtualMultiplexer on other cancer types. We applied the EMPaCT-pretrained VirtualMultiplexer to a PDAC TMA and generated virtual IHC stainings for CD44, CD146 and p53 (Fig. 6d), three markers with expected expression in pancreatic tissue. The generated images appeared overall realistic, with no means of discriminating whether they were virtually or actually stained. We observed that the CD44 and CD146 staining pattern in the virtual images was allocated, as expected, to the extracellular matrix of presented tissue spots, without major staining in the epithelial tissue part. For p53, we again observed overall proper staining allocation to the nuclei of epithelial cells with expected distribution, with no major staining of other compartments. To quantify the utility of the virtual stainings for downstream applications, we followed the same process as before to predict PDAC tumour, node and metastasis (TNM) stage, leading, again, to increased performance of models trained with virtually multiplexed data, concluding that virtually multiplexed data offers a performance advantage to prediction models. We also applied the pretrained VirtualMultiplexer to generate virtual IHC images for CD44 and CD146 from colorectal[45] and breast cancer[46] H&E-stained WSIs from The Cancer Genome Atlas (TCGA) at www.cancer.gov/tcga. Although the lack of normal tissue limited our ability to evaluate the staining quality in the generated images, we again observed an overall realistic virtual staining (Extended Data Fig. 9).

Lastly, we performed a runtime estimation of our framework (Extended Data Fig. 10) and concluded that it leads to substantial time gains when compared to a typical IHC staining, greatly accelerating histopathology workflows.

Discussion

We proposed the VirtualMultiplexer, a generative model that translates H&E to several IHC markers using a multiscale architecture with biological priors that ensures biological consistency on a cellular, neighbourhood and global whole-image scale without requiring image registration or extensive annotations. The VirtualMultiplexer consistently outperformed state-of-the-art methods in image fidelity scores. Detailed evaluation suggested that the virtual IHC images were indistinguishable from real ones to the expert eye, with a staining quality on par with or even exceeding that of real images and occasional staining artefacts largely comparable for three of the six markers. A thorough ablation study demonstrated that our multiscale loss mitigates staining unreliability, as opposed to competing methods that solely use adversarial and contrastive objectives. We also found that the model generalized well to unseen datasets of different image scales without any fine-tuning.

Although our results demonstrate a clear potential, several limitations remain, to be addressed in future extensions. First, we occasionally observed elevated background, especially for markers with faint staining. More pronounced background was present when transferring to prostate cancer WSIs, which was expected considering that this dataset was generated in different institutions using different staining protocols. Second, the patch-wise processing occasionally induced tiling artefacts more pronounced at the core border, a well-known limitation of S2S translation approaches[24,39,47,48]. One possible underlying cause is that as the model has only seen tissue-full patches during training, when it receives as input a patch with little tissue, the losses 'force' it to stain with higher intensity to match the distribution of a full patch. Previous attempts to address the tiling artefact[24,39] have been suggested to cause less efficient translations[49]. As in our case the tiling artefact is limited to edge cases, a straightforward solution is discarding a narrow border surrounding the tissue, as empirically done in actual IHC when border artefacts are present. Alternatively, more sophisticated extensions, such as the bidirectional feature-fusion GAN proposed by ref. 48 could be exploited. Third, discrepancies in staining specificity were occasionally observed (for example, failing to stain $CD146^+$ vascular structures and glandular $NKX3.1^+$ cells invading periglandular stroma), as these patterns were rarely observed in the training images and can be mitigated by ensuring the inclusion of adequate representative examples in the training set.

Importantly, despite their limitations, the generated images enabled the training of early fusion GT models, which consistently improved the prediction of clinical endpoints not only in the training dataset across two prediction tasks but also in both out-of-distribution prostate cancer cohorts and the PDAC TMA cohort. In our experiments, we ensured that the multimodal early fusion models did not have a numerical advantage over models trained with real data and also had a much smaller parameter space in comparison to late fusion ones, suggesting that improved performance is not a mere outcome of higher sample size or model complexity. A potential explanation of the observed improvement is that virtual images are not affected by artefacts occasionally found in real images, corroborated by the fact that for markers where virtual images were of higher quality than real, the corresponding unimodal–virtual models outperformed the unimodal–real ones and vice versa. Another explanation could be that as multimodal early fusion models could learn from the joint spatial distribution of several markers on the same tissue, they managed to pick up single-cell multimodal spatial relationships, mimicking data generated by advanced multiplexed technologies. This is further supported by the fact that in the early fusion case, a single GT model proved to have more learning capacity than the integration of several equivalently potent ones. However, the superior performance of models trained with virtual data could be unrelated to a potential higher quality of the generated images and could be a direct outcome of the fact that the VirtualMultiplexer potentially picks up the most consistent patterns and eliminates a lot of the noise and artefacts in the data, making the prediction task easier. This is further supported by other works that have reported competitive performance using models trained on other spatial features extracted from the tissue images[50,51].

In conclusion, the current work establishes the potential of virtual multiplexed staining, with important implications towards AI-assisted histopathology. For example, the VirtualMultiplexer could be directly used for data inpainting—that is, filling missing regions in an image—or for sample imputation—that is, generating missing samples from scratch. As IHC marker panels are not standardized across labs, filling in the gaps via virtual multiplexing could harmonize datasets within or across research labs, particularly important in cases of limited sample availability[52,53]. This could lead to the generation of harmonized and comprehensive patient cohorts, further used for clinically relevant predictions. An equally important application of our work concerns prehistopathological experimental design: generating a large collection of IHC stains in silico and training AI models could support marker selection for actual experimentation, reducing costs and preserving precious tissue. To reach its full potential, future work will be needed to validate the VirtualMultiplexer in real-world settings. From a technical standpoint, virtually multiplexed stainings can augment existing datasets and enable the development of foundational models for IHC, paving the way for multimodal tissue characterization. Interestingly, virtual multiplexed staining can be exploited as biologically conditioned data augmentations to boost the development and predictive performance of foundational models in histopathology. Our preliminary results on PDAC and TCGA images indicate that our model has the potential to generalize to tissues of different origins. However, more thorough evaluations are needed to solidify these encouraging early results. Finally, as our method is stain-agnostic, straightforward adaptations for S2S translation across multiplexed imaging technologies could substantially reduce costs via antibody panel optimization. Our vision is that future extensions of our work could lead to an ever-growing and readily available dictionary of virtual stainers for IHC

https://doi.org/10.1038/s42256-024-00889-5

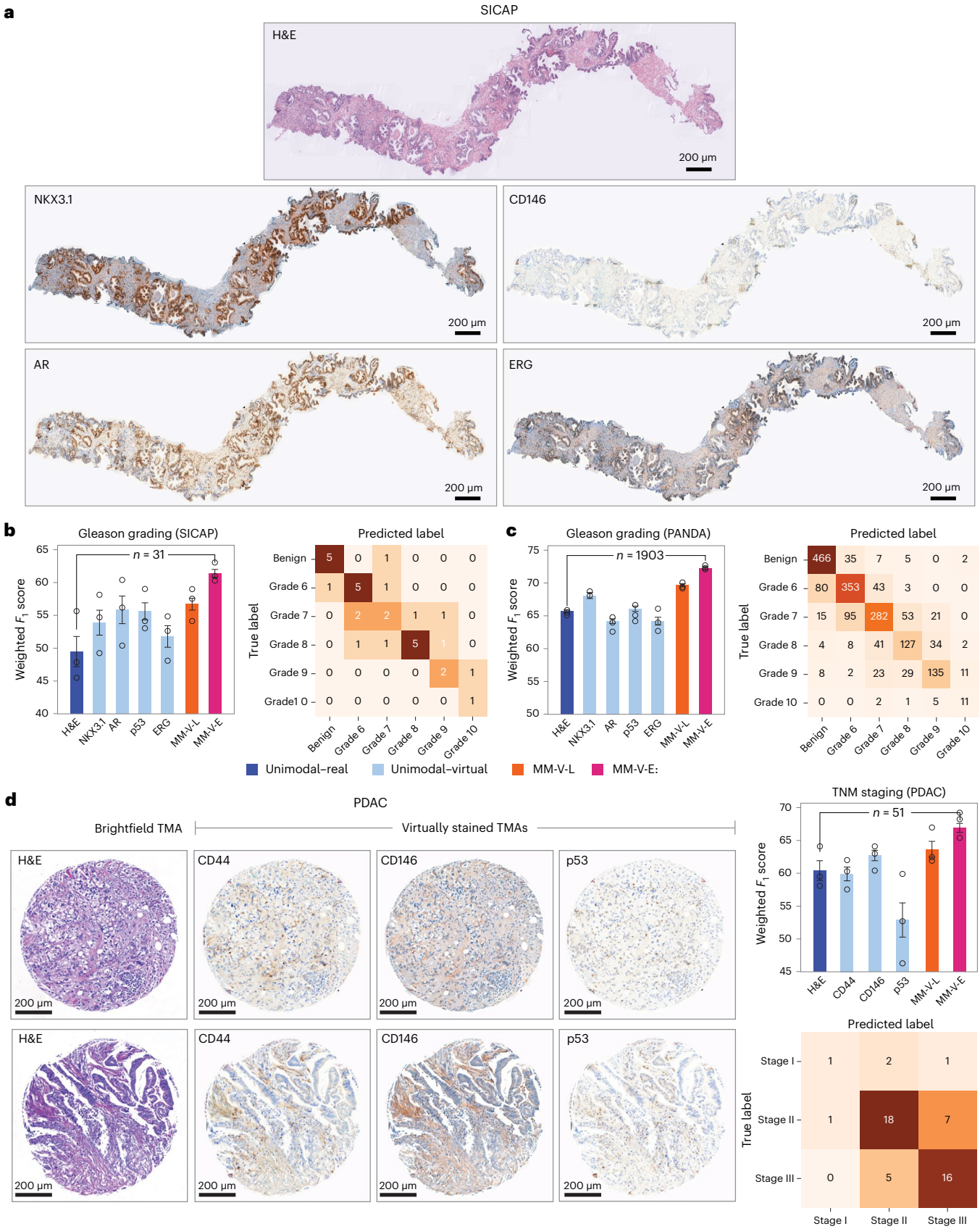


Fig. 6 | Transfer learning across scales, cohorts and cancer types. a, Top, real H&E needle biopsy of the SICAP dataset. Bottom, matching virtual IHC stainings across four IHC markers, as generated from the EMPaCT-trained VirtualMultiplexer. **b**, Prediction results of Gleason grading for the SICAP test set in terms of weighted F_1 score and confusion matrix. Note that the setting unimodal–real (dark blue barplot) only includes training the model on H&E, as no real IHC data are available here. **c**, Same as in **b**, but for the PANDA dataset. **d**, Virtual IHC staining of a PDAC TMA dataset with corresponding prediction of TNM staging. In **b**–**d**, barplots and error bars are as in Fig. 3 and confusion matrices correspond to the multimodal–virtual early fusion model.

and beyond, surpassing in multiplexing even the most cutting-edge technologies and accelerating spatial biology.

Methods

VirtualMultiplexer architecture

The VirtualMultiplexer is a generative AI toolkit that performs unpaired H&E-to-IHC translation. An overview of the model's architecture is shown in Fig. 2a. The VirtualMultiplexer is trained using two sets of images: source H&E images, denoted as $X_{\text{img}} = \{x \in \mathcal{X}\}$, and target IHC images, denoted as $Y_{\text{img}} = \{y \in \mathcal{Y}\}$. X_{img} and Y_{img} are unpaired images that originate from different sections of the same TMA core and thus belong to the same patient, but are pixel-wise unaligned and thus unpaired. We train an independent one-to-one VirtualMultiplexer model for each IHC marker at a time. To train the VirtualMultiplexer, we use patches $X_{\text{p}} = \{x_{\text{p}} \in X_{\text{img}}\}$ and $Y_{\text{p}} = \{y_{\text{p}} \in Y_{\text{img}}\}$ extracted from a pair of images X_{img} and Y_{img}, respectively. The backbone of the VirtualMultiplexer is a GAN-based generator G, specifically a CUT[34] model that consists of two sequential components: an encoder G_{enc} and a decoder G_{dec}. Upon training, the generator takes as input a patch x_{p} and generates a virtual patch y'_{p}: that is, $y'_{\text{p}} = G(x_{\text{p}}) = G_{\text{dec}}(G_{\text{enc}}(x_{\text{p}}))$. The virtually generated patches are stitched together to produce a final virtual image $Y'_{\text{img}} = \{y' \in \mathcal{Y}'\}$. The VirtualMultiplexer is trained under the supervision of three levels of consistency objectives: local, neighbourhood and global consistency (Fig. 2b). The neighbourhood consistency enforces effective staining translation at a patch level, where a patch captures the neighbourhood of a cell. We introduce additional global and local consistency objectives, operating at an image level and cell level, respectively, to further constrain the unpaired S2S translation and alleviate the stain-specific inconsistencies.

Neighbourhood consistency. The neighbourhood objective is a combination of an adversarial loss and a patch-wise multilayer contrastive loss, implemented as previously described in CUT[34] (Fig. 2b, panel 1). Briefly, the adversarial loss dictates the model to learn to eliminate style differences between real and virtual patches, and the multilayer contrastive loss guarantees the content preservation at a patch level[54]. The adversarial loss is a standard GAN min–max loss[35], where the discriminator D takes as input real IHC patches Y_{p} and IHC patches Y'_{p} virtually generated by generator G and attempts to classify them as either real or virtual (Fig. 2b, panel 1a). It is calculated as follows:

$$\mathcal{L}_{\text{adv}}(G, D, X_{\text{p}}, Y_{\text{p}}) = \mathbb{E}_{y_{\text{p}} \sim Y_{\text{p}}} \log D(y_{\text{p}}) + \mathbb{E}_{x_{\text{p}} \sim X_{\text{p}}} \log(1 - D(G(x_{\text{p}}))). \quad (1)$$

The patch-wise multilayer contrastive loss follows a NCE concept as presented in refs. 54,55 and reused in refs. 29,34. Specifically, it aims to maximize the resemblance between input H&E patch $x_{\text{p}} \in X_{\text{p}}$ and corresponding virtually synthesized IHC patch $y'_{\text{p}} \in Y'_{\text{p}}$ (Fig. 2b, panel 1b). We first extract a query subpatch y'_{sp} of size 64 × 64 from the target IHC domain patch y'_{p} (purple square in Fig. 2b, panel 1b) and match it to the corresponding subpatch x_{sp}: that is, a subpatch at the same spatial location as y'_{sp} but from the H&E source domain patch x_{p} (black square in Fig. 2b, panel 1b). Because both subpatches originate from the exact same tissue neighbourhood, we expect that x_{sp} and y'_{sp} form a positive pair. We also sample N subpatches $\{x^-_{\text{sp}}\}$ at different spatial locations from x_{p} (red squares in Fig. 2b, panel 1b) and expect that they form dissimilar, negative pairs with x_{sp}. In a standard contrastive learning scheme, we would map y_{sp}, x_{sp} and $\{x^-_{\text{sp}}\}$ to a d-dimensional embedding space $\mathbb{R}^d$ via G_{enc} and project them to a unit sphere, resulting in v, v^+ and , respectively, and then estimate the probability of a positive pair (v, v^+) selected over negative pairs $(v, v^-_n), \forall n \in N$ as a cross-entropy loss with a temperature scaling parameter τ:

$$\mathcal{L}(v, v^+, v^-) = -\log\left[\frac{\exp(vv^+/\tau)}{\exp(vv^+/\tau) + \sum_{n=1}^{N} \exp(vv^-_n/\tau)}\right] \quad (2)$$

Here, we use a variation of the loss in equation (2), specifically a patch-wise multilayer contrastive loss that extends $\mathcal{L}(v, v^+, v^-)$ by computing it for feature maps extracted from L-layers of G_{enc}[29,34]. This is achieved by passing the L feature maps of x_{p} and y'_{p} through a two-layer multilayer perceptron (MLP) H_l, resulting in a stack of features $\{z_l\}_L = \{H_l(G^l_{\text{enc}}(x_{\text{p}}))\}_L$ and $\{z'_l\}_L = \{H_l(G^l_{\text{enc}}(y'_{\text{p}}))\}_L = \{H_l(G^l_{\text{enc}}(G(x_{\text{p}})))\}_L$, $\forall l \in \{1, 2, \cdots, L\}$, respectively. We also iterate over each spatial location $s \in \{1, \cdots, S_l\}$, and we leverage all $S_l \backslash s$ patches as negatives, ultimately resulting in $z'_{l,s}$, $z_{l,s}$ and $z_{l,S_l \backslash s}$ for the query, positive and negative subpatches, respectively (purple, black and red boxes in Fig. 2b, panel 1b). The final patch-wise multilayer contrastive loss is computed as

$$\mathcal{L}_{\text{contrastive}}(G, H, X_{\text{p}}) = \mathbb{E}_{x_{\text{p}} \sim X_{\text{p}}} \sum_{l=1}^{L} \sum_{s=1}^{S_l} \mathcal{L}\left(z'_{l,s}, z_{l,s}, z_{l,S_l \backslash s}\right) \quad (3)$$

We also employ contrastive loss $\mathcal{L}_{\text{contrastive}}(G, H, Y_{\text{p}})$ on patches $y_{\text{p}} \in Y_{\text{p}}$, a domain-specific version of the identity loss[56,57] that prevents the generator G from making unnecessary changes as proposed in ref. 34. Finally, the overall neighbourhood consistency objective is computed as a weighted sum of the adversarial loss {equation (1)) and the multilayer contrastive loss (equation (3)) with regularization hyperparameter λ_{NCE}:

$$\mathcal{L}_{\text{neighbourhood}} = \mathcal{L}_{\text{adv}}(G, D, X_{\text{p}}, Y_{\text{p}}) + \lambda_{\text{NCE}} \times (\mathcal{L}_{\text{contrastive}}(G, H, X_{\text{p}}) + \mathcal{L}_{\text{contrastive}}(G, H, Y_{\text{p}})) \quad (4)$$

Global consistency. Inspired by seminal work in neural style transfer[58], this objective consists of two loss functions: a content loss $\mathcal{L}_{\text{content}}$ and a style loss $\mathcal{L}_{\text{style}}$ that together enforce biological consistency in terms of both tissue composition and staining pattern at the image (tile) level (Fig. 2b, panel 2). Because the generated IHC images should be virtually paired to their corresponding input H&E image in terms of tissue composition, the content loss aims to penalize the loss in content between H&E and IHC images at a tile level. First, real patches X_{p} and synthesized patches Y'_{p} are stitched to create images X_{img} and Y'_{img}, respectively, and corresponding tiles of size 1,024 × 1,024 are extracted (boxes in Fig. 2b, panel 2), denoted as $X_{\text{t}} = \{x_{\text{t}} \in X_{\text{img}}\}$ and $Y'_t = \{y'_t \in Y'_{\text{img}}\}$, respectively. Then the tiles are encoded by a pretrained feature extractor F, specifically VGG16 (ref. 59) pretrained on ImageNet[60]. The tile-level content loss at layer l of F is calculated as

$$\mathcal{L}^l_{\text{content}}\left(F, X_{\text{p}}, Y'_{\text{p}}\right) = \frac{\sum ||F^l(x_t) - F^l(y'_t)||^2}{h^l \cdot w^l \cdot c^l} \quad (5)$$

where h, w and c are the height, width and channel dimensions of the feature map at the lth layer, respectively.

The style loss utilizes the synthesized image Y'_{img} and the available real image Y_{img} to match the style or overall staining distribution between real and virtual IHC images. Because Y'_{img} and Y_{img} do not have pixel-wise correspondence, large tiles $Y'_t = \{y'_t \in Y_{\text{img}}\}$ and $Y_t = \{y_t \in Y_{\text{img}}\}$ are extracted at random such that each tile incorporates a sufficient staining distribution. Next, Y'_t and Y_t are processed by F to produce feature maps across multiple layers. The style loss is computed as

$$\mathcal{L}^l_{\text{style}}\left(F, Y_{\text{p}}, Y'_{\text{p}}\right) = \frac{\sum ||\mathcal{G} \circ F^l(y_t) - \mathcal{G} \circ F^l(y'_t)||^2}{||\mathcal{G} \circ F^l(y_t)||^2 + ||\mathcal{G} \circ F^l(y'_t)||^2} \quad (6)$$

where $\mathcal{G}$ is the Gram matrix that measures the correlation between all the styles in a feature map. The denominator is a normalization term that compensates for the under- or overstylization of the tiles in a batch[61]. The overall global consistency loss is computed as

$$\mathcal{L}_{\text{global}} = \lambda_{\text{content}} \times \sum_{l}^{L_{\text{content}}} \mathcal{L}^l_{\text{content}}\left(F, X_{\text{p}}, Y'_{\text{p}}\right) + \lambda_{\text{style}} \times \sum_{l}^{L_{\text{style}}} \mathcal{L}^l_{\text{style}}\left(F, Y_{\text{p}}, Y'_{\text{p}}\right) \quad (7)$$

where L_{content} and L_{style} are the lists of the content and style layers of F, respectively, used to extract the feature matrices, and λ_{content} and λ_{style} are regularization hyperparameters for the respective loss terms.

Local consistency. The local consistency objective aims to enforce biological consistency at a local cell level and consists of two loss terms: a cell discriminator loss ($\mathcal{L}_{\text{cellDisc}}$) and a cell classification loss ($\mathcal{L}_{\text{cellClass}}$) (Fig. 2b, panel 3). The cell discriminator loss is inspired by ref. 26 and uses the cell discriminator D_{cell} to identify whether a cell is real or virtual, in the same way that the patch discriminator of equation (1) attempts to classify patches as real or virtual. $\mathcal{L}_{\text{cellDisc}}$ takes as input a real (Y_{p}) and a virtual (Y'_{p}) target patch and their corresponding cell masks ($M_{Y_{\text{p}}}$ and $M_{Y'_{\text{p}}}$, respectively), which include bounding-box demarcation around the cells (Fig. 2b, panel 3). D_{cell} comprises a feature extractor followed by a RoIAlign layer[62] and a final discriminator. The goal of D_{cell} is to output $D_{\text{cell}}(Y_{\text{p}}, M_{Y_{\text{p}}}) \rightarrow 1$ and $D_{\text{cell}}(Y'_{\text{p}}, M_{X_{\text{p}}}) \rightarrow 0$, where 1 and 0 indicate real and virtual cells (indicated in black and purple, respectively, in Fig. 2b, panel 3). The cell discriminator loss is defined as

$$\begin{aligned}\mathcal{L}_{\text{cellDisc}}(D_{\text{cell}}, Y_{\text{p}}, Y'_{\text{p}}, M_{Y_{\text{p}}}, M_{Y'_{\text{p}}}) = \frac{1}{2}\,\mathbb{E}_{y_{\text{p}} \in Y_{\text{p}}}\left(D_{\text{cell}}(y_{\text{p}}, M_{y_{\text{p}}}) - 1\right)^2 \\ + \frac{1}{2}\,\mathbb{E}_{y'_{\text{p}} \in Y'_{\text{p}}}\left(D_{\text{cell}}\left(y'_{\text{p}}, M_{y'_{\text{p}}}\right)\right)^2\end{aligned} \tag{8}$$

Although D_{cell} aims to enforce the generation of realistically looking cells, it is agnostic to their marker expression, as it does not explicitly capture which cells have a positive or a negative staining status. To account for this, we introduce an additional loss via a classifier F_{cell} that is trained to explicitly predict the cell staining status. This is achieved with the help of cell labels $C_{Y'_{\text{p}}}$ and $C_{Y_{\text{p}}}$: that is, binary variables depicting the positive or negative staining status of a cell (indicated as 1: yellow and 0: blue boxes in Fig. 2b, panel 3). The computation of cell masks and labels is described in detail in the section 'Cell masking and labelling of IHC images'. The cell-level classification loss can be easily computed as cross-entropy loss, calculated as

$$\begin{aligned}&\mathcal{L}_{\text{cellClass}}\left(F_{\text{cell}}, Y_{\text{p}}, Y'_{\text{p}}, M_{Y_{\text{p}}}, M_{Y'_{\text{p}}}, C_{Y_{\text{p}}}, C_{Y'_{\text{p}}}\right) \\ &= \frac{-1}{|C_{y_{\text{p}}}|}\sum_{\substack{j=1 \\ i\in\{0,1\}}}^{|C_{y_{\text{p}}}|} \mathbb{1}_{C^j_{y_{\text{p}}}=i} \times \log\left(p\left(M^j_{y_{\text{p}}} \times y_{\text{p}}\right)\right) \\ &+ \frac{-1}{|C_{y'_{\text{p}}}|}\sum_{\substack{j=1 \\ i\in\{0,1\}}}^{|C_{y'_{\text{p}}}|} \mathbb{1}_{C^j_{y'_{\text{p}}}=i} \times \log\left(p\left(M^j_{y'_{\text{p}}} \times y'_{\text{p}}\right)\right)\end{aligned} \tag{9}$$

where $|C_{y_{\text{p}}}|$ and $|C_{y'_{\text{p}}}|$ are the number of cells in y_{p} and y'_{p}, respectively, $\mathbb{1}_{(.)}$ is the indicator function and $p(.)$ is the cell-level probabilities predicted by F_{cell}.

The overall local consistency loss is computed as

$$\begin{aligned}\mathcal{L}_{\text{local}} = \lambda_{\text{cellDisc}} \times \mathcal{L}_{\text{cellDisc}}\left(D_{\text{cell}}, Y_{\text{p}}, Y'_{\text{p}}, M_{Y_{\text{p}}}, M_{Y'_{\text{p}}}\right) \\ + \lambda_{\text{cellClass}} \times \mathcal{L}_{\text{cellClass}}\left(F_{\text{cell}}, Y_{\text{p}}, Y'_{\text{p}}, M_{Y_{\text{p}}}, M_{Y'_{\text{p}}}, C_{Y_{\text{p}}}, C_{Y'_{\text{p}}}\right)\end{aligned} \tag{10}$$

where $\lambda_{\text{cellDisc}}$ and $\lambda_{\text{cellClass}}$ are the regularization hyperparameters for the cell discriminator and classification loss terms, respectively. Importantly, the local consistency loss can be easily generalized to any other cellular or tissue component (for example, nuclei, glands) that might be relevant to other S2S translation problems, provided that corresponding masks and labels are available.

The complete objective function for optimizing VirtualMultiplexer is given as

$$\mathcal{L}_{\text{VirtualMultiplexer}} = \mathcal{L}_{\text{neighbourhood}} + \mathcal{L}_{\text{global}} + \mathcal{L}_{\text{local}} \tag{11}$$

Cell masking and labelling of IHC images. As already discussed, the local consistency loss of equation (11) needs as input cell masks $M_{X_{\text{p}}}, M_{Y_{\text{p}}}$ and cell labels $C_{X_{\text{p}}}, C_{Y_{\text{p}}}$. However, acquiring these inputs manually for all patches across all antibodies is practically prohibitive, even for relatively small datasets. Automatic nuclei segmentation/detection using pretrained models (for example, HoVerNet[63]) is a standard task for H&E images, but no such model exists for IHC images. To circumvent this challenge, we use an attractive property of the VirtualMultiplexer: its ability to synthesize virtual images that are pixel-wise aligned in any direction between the source and target domain. Specifically, we train a separate instance of the VirtualMultiplexer that performs IHC → H&E translation. The VirtualMultiplexer$_{\text{IHC}\rightarrow\text{H\&E}}$ is trained using neighbourhood consistency and global consistency objectives, as previously described. Once trained, it is used to synthesize a virtual H&E image X'_{img} from a real IHC image Y_{img}. At this point, we can leverage HoVerNet[63] to detect cell nuclei on real and virtual H&E images (X_{img} and X'_{img}) and simply transfer the corresponding cell masks ($M_{X_{\text{img}}}$ and $M_{X'_{\text{img}}}$) to their pixel-wise aligned IHC counterparts (Y'_{img} and Y_{img}, respectively) to acquire $M_{Y'_{\text{img}}}$ and $M_{Y_{\text{img}}}$. This 'trick' eliminates the need to train individual cell detection models for each IHC antibody and fully automates the cell masking process in the IHC domain. To acquire cell labels $C_{Y'_{\text{img}}}$ and $C_{Y_{\text{img}}}$, we use only region annotations in Y_{img}, where the experts partially annotated areas as positive or negative stainings in a few representative images. Because IHC stainings are specialized in delineating positive or negative staining status, the annotation was easy and fast and required approximately 2–3 minutes per image and per antibody marker. We also train cell detectors for the source and target domain: that is, $D^{\text{source}}_{\text{cell}}$ and $D^{\text{target}}_{\text{cell}}$, respectively. Provided with the annotations, $D^{\text{target}}_{\text{cell}}$ is trained as a CNN patch classifier. The classifier predictions on Y_{img} combined with $M_{Y_{\text{p}}}$ result in $C_{Y_{\text{p}}}$. The above region predictions on Y_{img} are transferred on to X'_{img}. Afterwards, X'_{img} and the transferred annotations are used to train $D^{\text{source}}_{\text{cell}}$ as a CNN patch classifier. The classifier predictions on X_{img} combined with $M_{X_{\text{p}}}$ result in $C_{X_{\text{p}}}$.

Implementation and training details. The architectural choices of the VirtualMultiplexer were set as follows: G is a ResNet[64] with nine residual blocks, D is a PatchGAN discriminator[12], D_{cell} includes four stride-2 feature convolutions followed by a RoIAlign layer and a discrimination layer and F_{cell} includes four stride-2 feature convolutions and a two-layer MLP. We use Xavier weight initialization[65], instance normalization[66] and a batch size of one image. We use least square GAN loss[67] for $\mathcal{L}_{\text{adv}}$. The model hyperparameters for the loss terms of the VirtualMultiplexer are set as follows: λ_{NCE} is 1 with temperature τ equal to 0.08, $\lambda_{\text{content}} \in \{0.01, 0.1\}$, $\lambda_{\text{style}} \in \{5, 10\}$, $\lambda_{\text{cellDisc}} \in \{0.5, 1\}$ and $\lambda_{\text{cellClass}} \in \{0.1, 0.5\}$. VirtualMultiplexer is optimized for 125 epochs using the Adam optimizer[68] with momentum parameters $\beta_1 = 0.5$ and $\beta_2 = 0.999$. Different learning rates (lr) are employed for different consistency objectives: that is, for neighbourhood consistency, lr_G and lr_D are set to 0.0002; for global consistency, learning rate lr_G is chosen from {0.0001, 0.0002}; and for local consistency, learning rates $\text{lr}_{D_{\text{cell}}}$ and $\text{lr}_{F_{\text{cell}}}$ are chosen from {0.00001, 0.0001, 0.0002}. Among other hyperparameters, the number of tiles extracted per image to compute $\mathcal{L}_{\text{content}}$ and $\mathcal{L}_{\text{style}}$ is set to eight; the content layer in F is relu2_2; the style layers are relu1_2, relu2_2, relu3_3, relu4_3; and the number of cells per patch to compute $\mathcal{L}_{\text{cellDisc}}$ is set to eight.

GT architecture

The GT architecture, proposed by ref. 41, fuses a graph neural network and a vision transformer (ViT) to process histopathology images. The graph neural network operates on a graph-structured representation of a histopathology image, where the nodes and edges of the graph denote patches and interpatch spatial connectivity, and the nodes encode patch features extracted from a pretrained ResNet-50 network[64]. The graph representation underwent graph convolutions to

contextualize the node features of the local tissue neighbourhood. Specifically, the GT employs a graph convolution layer[69] to learn contextualized node embeddings through propagating and aggregating neighbourhood node information. Subsequently, a ViT layer operates on the contextualized node features, leverages self-attention to weigh the importance of the nodes and aggregates the node information to render an image-level feature representation. Finally, an MLP maps the image-level features to a downstream image label. Note that histopathology images can have different spatial dimensions; therefore, their graph representations can have varying number of nodes. Also, the number of nodes can be very high when operating on gigapixel-sized WSIs. These two factors can potentially hinder the integration of the graph convolution layer to the ViT layer. To address these challenges, GT introduces a mincut pooling layer[70], which reduces the number of nodes to a fixed number of tokens while preserving the local neighbourhood information of the nodes.

Implementation and training details. The architecture of the GT follows the official implementation on GitHub (https://github.com/vkola-lab/tmi2022). Each input image was cropped to create a bag of 256 × 256 non-overlapping patches at ×10 magnification, and background patches with non-tissue area greater than 10% were discarded. The patches were encoded using the ResNet-50[64] model pretrained on the ImageNet dataset[60]. A graph representation was constructed using the patches with an eight-node connectivity pattern. The GT network consisted of one graph convolutional layer, and the ViT layer configurations were set as follows: number of ViT blocks = 3, MLP size = 128, embedding dimension of each patch = 32 and number of multihead attention = 8. The model hyperparameters were set as follows: number of clusters in mincut pooling = {50, 100}, Adam optimizer with initial learning rate of {0.0001, 0.00001}, a cosine annealing scheme for scheduling and a mini-batch size of eight. The GT models were trained for 400 epochs with early stopping.

Datasets

The VirtualMultiplexer was trained using the EMPaCT TMA dataset; an independent subset of EMPaCT was used for internal testing. The VirtualMultiplexer was further evaluated in a zero-shot fashion—that is, without any retraining or fine-tuning—on three external prostate cancer datasets (prostate cancer WSIs, SICAP[42] and PANDA[43] needle biopsies), on an independent PDAC dataset (PDAC TMAs) and on TCGA data from breast and colorectal cancer. In all cases, independent GTs are trained and tested for individual datasets by using both real and virtually stained samples to address various downstream classification tasks. Details on all datasets used follow.

EMPaCT. The dataset contains TMAs from 210 primary prostate tissues as part of EMPaCT and the Institute of Tissue Pathology in Bern. The study followed the guidelines of the World Medical Association Declaration of Helsinki 1964, updated in October 2013, and was conducted after approval by the Ethics Committees of Bern (CEC ID2015-00128). For each patient, four cores were selected, with two of them representing a low Gleason pattern and the other two a high Gleason pattern. Consecutive slices from each core were stained with H&E and IHC using multiple antibodies against nuclear markers NKX3.1 and AR, tumour markers p53 and ERG, and membrane markers CD44 and CD146. TMA FFPE sections of 4 µm were deparaffinized and used for heat-mediated antigen retrieval (citrate buffer, pH 6, Vector Labs; or Tris-HCl, pH 9). Sections were blocked for 10 min in 3% H_2O_2, followed by 30 min room temperature incubation in 1% bovine serum albumin in phosphate-buffered saline–0.1% Tween 20. The following antibodies were used: anti-AR (Dako Agilent, catalogue no. M3562, AR441, 1:100 dilution), anti-NKX3.1 (Athena Enzyme Systems, catalogue no. 314, lot 18025, 1:200), anti-p53 (Dako Agilent, catalogue no. M7001, DO-7, 1:800), anti-CD44 (Abcam, catalogue no. ab16728, 156-3C11, 1:2000), anti-ERG (Abcam, catalogue no. ab133264, EPR3864(2), 1:500) and anti-CD146 (Abcam, catalogue no. ab75769, EPR3208, 1:500). Images were acquired using a 3D Histech Panoramic Flash II 250 scanner at ×20 magnification (resolution 0.24 µm per pixel). The cores were annotated at patient level by expert uro-pathologists with binary labels for overall survival status (0, alive/censored; 1, prostate-cancer-related death) and disease progression status (0, no recurrence; 1, recurrence). Clinical follow-up was recorded on a per-patient basis, with a maximum follow-up time of up to 12 years. For both the survival and disease progression clinical endpoints, the available data were imbalanced in terms of class distributions. Access information is possible upon request to the corresponding authors. The distribution of cores per clinical endpoint for the EMPaCT dataset is summarized in Supplementary Table 2.

Prostate cancer WSIs. Primary stage prostate cancer FFPE tissue sections (4 µm) were deparaffinized and used for heat-mediated antigen retrieval (citrate buffer, pH 6, Vector Labs). Sections were blocked for 10 min in 3% H_2O_2, followed by 30 min room temperature incubation in 1% bovine serum albumin in phosphate-buffered saline–0.1% Tween 20. The following primary antibodies were used: anti-CD146 (Abcam, catalogue no. ab75769, EPR3208, 1:500), anti-AR (Abcam, catalogue no. ab133273, EPR1535, 1:100) and anti-NKX3.1 (Cell Signaling, catalogue no. 83700T, D2Y1A, 1:200). Secondary anti-rabbit antibody Envision horseradish peroxidase (DAKO, Agilent Technologies, catalogue no. K400311-2, undiluted) was incubated for 30 min, and signal detection was done using 3-amino-9-ethylcarbazole substrate (DAKO, Agilent Technologies). Sections were counterstained with hematoxylin and mounted with aquatex. Images were acquired using a 3D Histech Panoramic Flash II 250 scanner at ×20 magnification (resolution 0.24 µm per pixel).

SICAP. The dataset contains 155 H&E-stained WSIs from needle biopsies taken from 95 patients, split in 18,783 patches of size 512 × 512 (ref. 42). The WSIs were reconstructed by stitching the patches. The WSIs were scanned at ×40 magnification by a Ventana iScan Coreo scanner and downsampled to ×10 magnification. The WSIs were annotated by expert uro-pathologists for Gleason grades at the Hospital Clínico of Valencia, Spain.

PANDA. The dataset includes 5,759 H&E-stained needle biopsies from 1,243 patients at the Radboud University Medical Center, Netherlands[71] and 5,662 H&E-stained needle biopsies from 1,222 patients at various hospitals in Stockholm, Sweden[72]. The slides from Radboud were scanned with a 3D Histech Panoramic Flash II 250 scanner at ×20 magnification (resolution 0.24 µm per pixel) and were downsampled to ×10. The slides from Sweden were scanned with a Hamamatsu C9600-12 and an Aperio Scan Scope AT2 scanner at ×10 magnification with a pixel resolution of 0.45202 µm and 0.5032 µm, respectively. The Gleason grades of the biopsies were annotated by expert uro-pathologists and were released as part of the PANDA challenge[43]. We removed the noisy and inconspicuously labelled biopsies from the dataset, resulting in 4,564 and 4,988 biopsies from the Radboud and the Swedish cohorts, respectively (9,552 biopsies in total). The distribution of WSIs across Gleason grades for both SICAP and PANDA datasets is shown in Supplementary Table 3.

PDAC. The PDAC TMA contained cancer tissue of 117 (50 female, 67 male) PDAC cases resected in a curative setting at the Department of Visceral Surgery of Inselspital Bern and diagnosed at the Institute of Tissue Medicine and Pathology (ITMP) of the University of Bern between the years 2014 and 2020. The study followed the guidelines of the World Medical Association Declaration of Helsinki 1964, updated in October 2013, and was conducted after approval by the Ethics Committees of Bern (CEC ID2020-00498). All participants provided written general consent. The TMA contained three spots from each case (tumour front, tumour centre, tumour stroma), leading to a total

number of 351 tissue spots. Thirteen of these 117 cases were treated by neoadjuvant chemotherapy followed by surgical resection and adjuvant therapy, and the majority of the cases (104) were resected curatively and received adjuvant therapy. All cases were characterized comprehensively clinico-pathologically, including TNM stage, during a master's thesis of student Jessica Lisa Rohrbach at ITMP, supervised by Martin Wartenberg. All cases were Union for International Cancer Control (UICC) tumour stage I, stage II or stage III cases on pathologic examination, according to the UICC *TNM Classification of Malignant Tumours*, 8th edition[73]; the TMA did not include UICC tumour stage IV cases. In all of our analysis, including the TNM prediction (Fig. 6d), we excluded the 13 neoadjuvant cases and considered only the 104 cases that received adjuvant therapy. The distribution of cores across the three TNM stages is reported in Supplementary Table 4.

TCGA. The dataset includes example H&E WSIs from breast cancer (BRCA) and colorectal cancer (CRC) from The TCGA, available at the GDC data portal (https://portal.gdc.cancer.gov) as diagnostic slides under project IDs TCGA-BRCA and TCGA-CRC, respectively.

Data preprocessing

For all datasets used, we followed a tissue region detection and patch extraction preprocessing procedure. Specifically, the tissue region was segmented using the preprocessing tools in the HistoCartography library[74]. A binary tissue mask denoting the tissue and non-tissue regions was computed for each downsampled input image by iteratively applying Gaussian smoothing and Otsu thresholding until the mean of non-tissue pixels was below a threshold. The estimated contours of the denoted tissue and the cavities of tissue were then filtered depending on their area to generate the final segmentation mask. Subsequently, non-overlapping patches of size 256 × 256 were extracted from ×10 magnification using the segmentation contours. The extracted H&E and IHC patches of the EMPaCT dataset were used for training and internal validation of the VirtualMultiplexer. For the unseen datasets (prostate cancer WSIs, SICAP, PANDA, PDAC, TCGA), the images were first stain-normalized to mitigate the staining appearance variability with respect to the EMPaCT TMAs, and then H&E patches were extracted. Specifically, for the SICAP, PANDA and PDAC datasets, we used the Vahadane stain normalization method[75], from the HistoCartography library[74], on the entire images. We masked out the blank regions by applying a threshold on the Lab colour space and computed the stain-density maps using only the tissue regions. Afterwards, the target stain-density maps are combined with the reference colour appearance matrix to produce normalized images, as proposed by the Vahadane method. Supplementary Fig. 1 presents a sample unnormalized WSI from the PANDA dataset and the corresponding stain-normalized WSI based on the reference EMPaCT TMA. For the prostate cancer and TCGA WSIs, we followed the same procedure but with stain-density maps extracted from a lower magnification (×2.5) for computational efficiency. Note that the VirtualMultiplexer is independent of the stain normalization method and can be trained using H&E images normalized by other advanced stain normalization algorithms: for example, deep learning-based methods[76].

Method evaluation

Patch-level evaluation. We use the FID score[77] to compare the distribution of the virtual IHC patches with the distribution of the real IHC patches, as shown in Fig. 3. The computation begins with projecting the virtual and the real IHC patches to an embedding space using the InceptionV3 (ref. 77) model, pretrained on ImageNet[60]. The extracted embeddings are used to estimate multivariate normal distributions $\mathcal{N}(\mu_r, \Sigma_r)$ for real data and $\mathcal{N}(\mu_s, \Sigma_s)$ for virtual data. Finally, the FID score is computed as

$$\mathrm{FID} = ||\mu_r - \mu_v||^2 + \mathrm{Tr}\left(\Sigma_r + \Sigma_v - 2(\Sigma_r\Sigma_v)^{\frac{1}{2}}\right) \quad (12)$$

where μ_r and μ_v are the feature-wise mean of the real and virtual patches, Σ_r and Σ_v are covariance matrices for the real and virtual embeddings, and Tr is the trace function. A lower FID score indicates a lower disparity between the two distributions and thereby a higher staining efficacy of the VirtualMultiplexer. To ensure reproducibility, we ran each model three times with three independent initializations and computed the mean and standard deviation for each model (barplot height and error bar in Fig. 3). We used a 70%:30% ratio to split the data into train and test sets, respectively. As for each marker a different number of IHC stainings were available in the EMPaCT data, the exact number of cores used per marker are given in Supplementary Table 5.

Image-level evaluation. We used a number of downstream classification tasks to assess the discriminative ability of the virtually stained IHC images on the EMPaCT, SICAP, PANDA and PDAC datasets. We further used these tasks to depict the utility of leveraging virtually multiplexed staining in comparison to standalone real H&E, real IHC and virtual IHC staining. Specifically, provided the aforementioned images, we constructed graph representations as described in Section GT architecture. Subsequently, GTs[41] were trained under unimodal and multimodal settings using both real and virtually stained images and evaluated on a held-out independent test dataset. The final classification scores were reported using a weighted F_1 metric, where a higher score depicts a better classification performance and thereby higher discriminative power of the utilized images. As before, we ran each model three times with three independent initializations and computed the mean and standard deviation for each model (barplot heights and error bars in Figs. 5 and 6). In all cases, we used a 60%:20%:20% ratio to split the data into train, validation and test sets, respectively. The exact number of train, validation and test samples used per task, marker and training setting in the EMPaCT dataset are given in Supplementary Table 6.

For the SICAP, PANDA and PDAC datasets, the exact number of samples used in the train, validation and test splits coincide for all unimodal and multimodal models of Fig. 6 and are reported in Supplementary Table 7.

Computational hardware and software

The image datasets were preprocessed on POWER9 central processing units and one NVIDIA Tesla A100 graphics processing unit using the Histocartography library[74]. The deep learning models were trained on NVIDIA Tesla P100 graphics processing units using PyTorch (v.1.13.1) (ref. 78) and PyTorch Geometric (v.2.3.0) (ref. 79). The entire pipeline was implemented in Python (v.3.9.1).

Reporting summary

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

Data availability

The main dataset used to support this study (EMPaCT) has been deposited in Zenodo, together with the prostate cancer WSIs[80]. The SICAP dataset is available at Mendeley data[81]. The PANDA dataset is available at the Kaggle website (https://www.kaggle.com/c/prostate-cancer-grade-assessment/data). The TCGA WSIs from breast and colorectal tissue are available as diagnostic slides under project IDs TCGA-BRCA and TCGA-CRC, respectively, at the GDC data portal (https://portal.gdc.cancer.gov). The PDAC dataset is available for academic research purposes upon request via e-mail to M.W. (martin.wartenberg@unibe.ch) or the Translational Research Unit Platform of ITMP of the University of Bern (tru.igmp@unibe.ch). All clinical data associated with the EMPaCT and PDAC patient cohorts cannot be shared owing to patient-confidentiality obligations.

Code availability

All source code of the VirtualMultiplexer is available under an open-source license at https://github.com/AI4SCR/VirtualMultiplexer and via Zenodo at https://doi.org/10.5281/zenodo.11941982 (ref. 82).

References

1. Kashyap, A. et al. Quantification of tumor heterogeneity: from data acquisition to metric generation. *Trends Biotechnol.* **40**, 647–676 (2022).
2. Chan, J. K. The wonderful colors of the hematoxylin–eosin stain in diagnostic surgical pathology. *Int. J. Surgical Pathol.* **22**, 12–32 (2014).
3. De Matos, L. L., Trufelli, D. C., De Matos, M. G. L. & da Silva Pinhal, M. A. Immunohistochemistry as an important tool in biomarkers detection and clinical practice. *Biomark. Insights* **5**, BMI–S2185 (2010).
4. Giesen, C. et al. Highly multiplexed imaging of tumor tissues with subcellular resolution by mass cytometry. *Nat. Methods* **11**, 417–422 (2014).
5. Goltsev, Y. et al. Deep profiling of mouse splenic architecture with codex multiplexed imaging. *Cell* **174**, 968–981 (2018).
6. Angelo, M. et al. Multiplexed ion beam imaging of human breast tumors. *Nat. Med.* **20**, 436–442 (2014).
7. Lewis, S. M. et al. Spatial omics and multiplexed imaging to explore cancer biology. *Nat. Methods* **18**, 997–1012 (2021).
8. Pillar, N. & Ozcan, A. Virtual tissue staining in pathology using machine learning. *Expert Rev. Mol. Diagnostics* **22**, 987–989 (2022).
9. Bai, B. et al. Deep learning-enabled virtual histological staining of biological samples. *Light.: Sci. Appl.* **12**, 57 (2023).
10. Tschuchnig, M. E., Oostingh, G. J. & Gadermayr, M. Generative adversarial networks in digital pathology: a survey on trends and future potential. *Patterns* **1**, 100089 (2020).
11. Jose, L., Liu, S., Russo, C., Nadort, A. & Di Ieva, A. Generative adversarial networks in digital pathology and histopathological image processing: a review. *J. Pathol. Inform.* **12**, 43 (2021).
12. Isola, P., Zhu, J.-Y., Zhou, T. & Efros, A. A. Image-to-image translation with conditional adversarial networks. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 5967–5976 (IEEE, 2017).
13. Li, J. et al. Biopsy-free in vivo virtual histology of skin using deep learning. *Light Sci. Appl.* **10**, 233 (2021).
14. Rivenson, Y. et al. PhaseStain: the digital staining of label-free quantitative phase microscopy images using deep learning. *Light Sci. Appl.* **8**, 23 (2019).
15. Rivenson, Y. et al. Virtual histological staining of unlabelled tissue-autofluorescence images via deep learning. *Nat. Biomed. Eng.* **3**, 466–477 (2019).
16. Rana, A. et al. Use of deep learning to develop and analyze computational hematoxylin and eosin staining of prostate core biopsy images for tumor diagnosis. *JAMA Netw. Open* **3**, e205111 (2020).
17. de Haan, K. et al. Deep learning-based transformation of H&E stained tissues into special stains. *Nat. Commun.* **12**, 4884 (2021).
18. Zhang, Y. et al. Digital synthesis of histological stains using micro-structured and multiplexed virtual staining of label-free tissue. *Light Sci. Appl.* **9**, 78 (2020).
19. Liu, S. et al. BCI: breast cancer immunohistochemical image generation through pyramid pix2pix. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 1815–1824 (IEEE, 2022).
20. Xie, W. Prostate cancer risk stratification via non-destructive 3D pathology with deep learning-assisted gland analysis. *Cancer Res.* **82**, 334 (2022).
21. Ghahremani, P. et al. Deep learning-inferred multiplex immunofluorescence for immunohistochemical image quantification. *Nat. Mach. Intell.* **4**, 401–412 (2022).
22. Zhang, R. et al. MVFStain: multiple virtual functional stain histopathology images generation based on specific domain mapping. *Med. Image Anal.* **80**, 102520 (2022).
23. Mercan, C. et al. Virtual staining for mitosis detection in breast histopathology. In *Proc. 17th International Symposium on Biomedical Imaging (ISBI)* 1770–1774 (IEEE, 2020).
24. Lahiani, A., Klaman, I., Navab, N., Albarqouni, S. & Klaiman, E. Seamless virtual whole slide image synthesis and validation using perceptual embedding consistency. *IEEE J. Biomed. Health Inform.* **25**, 403–411 (2020).
25. Liu, S. et al. Unpaired stain transfer using pathology-consistent constrained generative adversarial networks. *IEEE Trans. Med. Imaging* **40**, 1977–1989 (2021).
26. Boyd, J. et al. Region-guided CycleGANs for stain transfer in whole slide images. In *Proc. Medical Image Computing and Computer Assisted Intervention (MICCAI)* 356–365 (Springer, 2022).
27. Lin, Y. et al. Unpaired multi-domain stain transfer for kidney histopathological images. In *Proc. AAAI Conference on Artificial Intelligence*. 1630–1637 (AAAI, 2022).
28. Bouteldja, N., Klinkhammer, B. M., Schlaich, T., Boor, P. & Merhof, D. Improving unsupervised stain-to-stain translation using self-supervision and meta-learning. *J. Pathol. Inform.* **13**, 100107 (2022).
29. Ozyoruk, K. B. et al. A deep-learning model for transforming the style of tissue images from cryosectioned to formalin-fixed and paraffin-embedded. *Nat. Biomed. Eng.* **6**, 1407–1419 (2022).
30. Zhu, J.-Y., Park, T., Isola, P. & Efros, A. A. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proc. IEEE International Conference on Computer Vision (ICCV)* 2242–2251 (IEEE, 2017).
31. Zeng, B. et al. Semi-supervised PR virtual staining for breast histopathological images. In *Proc. Medical Image Computing and Computer Assisted Intervention (MICCAI)*. 232–241 (Springer, 2022).
32. Borji, A. Pros and cons of GAN evaluation measures: new developments. *Computer Vis. Image Underst.* **215**, 103329 (2022).
33. Cohen, J. P., Luck, M. & Honari, S. Distribution matching losses can hallucinate features in medical image translation. In *Proc. Medical Image Computing and Computer Assisted Intervention (MICCAI)* 529–536 (Springer, 2018).
34. Park, T., Efros, A. A., Zhang, R. & Zhu, J.-Y. Contrastive learning for unpaired image-to-image translation. In *In Proc. European Conference on Computer Vision (ECCV)* 319–345 (Springer, 2020).
35. Goodfellow, I. J. et al. Generative adversarial nets. In *Proc. 27th International Conference on Neural Information Processing Systems*. 2672-2680 (2014).
36. Briganti, A. et al. Identifying the best candidate for radical prostatectomy among patients with high-risk prostate cancer. *Eur. Urol.* **61**, 584–592 (2012).
37. Kneitz, B. et al. Survival in patients with high-risk prostate cancer is predicted by mir-221, which regulates proliferation, apoptosis, and invasion of prostate cancer cells by inhibiting IRF2 and SOCS3. *Cancer Res.* **74**, 2591–2603 (2014).
38. Tosco, L. et al. The EMPaCT classifier: a validated tool to predict postoperative prostate cancer-related death using competing-risk analysis. *Eur. Urol. Focus* **4**, 369–375 (2018).
39. Ho, M.-Y., Wu, M.-S. & Wu, C.-M. Ultra-high-resolution unpaired stain transformation via kernelized instance normalization. In *Proc. European Conference on Computer Vision (ECCV)* 490–505 (Springer, 2022).

40. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B. & Hochreiter, S. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In *Proc. 31st International Conference on Neural Information Processing Systems*. 6629–6640 (ACM, 2017).
41. Zheng, Y. et al. A graph-transformer for whole slide image classification. *IEEE Trans. Med. Imaging* **41**, 3003–3015 (2022).
42. Silva-Rodríguez, J., Colomer, Adrián, Sales, María, Molina, R. & Naranjo, V. Going deeper through the Gleason scoring scale: an automatic end-to-end system for histology prostate grading and cribriform pattern detection. *Comput. Methods Programs Biomed.* **195**, 105637 (2020).
43. Bulten, W. et al. Artificial intelligence for diagnosis and Gleason grading of prostate cancer: the PANDA challenge. *Nat. Med.* **28**, 154–163 (2022).
44. Pati, P. et al. Weakly supervised joint whole-slide segmentation and classification in prostate cancer. *Med. Image Anal.* **89**, 102915 (2023).
45. The Cancer Genome Atlas Network. Comprehensive molecular characterization of human colon and rectal cancer. *Nature* **487**, 330 (2012).
46. The Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. *Nature* **490**, 61–70 (2012).
47. de Bel, T., Hermsen, M., Kers, J., van der Laak, J. & Litjens, G. Stain-transforming cycle-consistent generative adversarial networks for improved segmentation of renal histopathology. In *Proc. 2nd International Conference on Medical Imaging with Deep Learning*. 151–163 (PMLR, 2019).
48. Sun, K. et al. Bi-directional feature fusion generative adversarial network for ultra-high resolution pathological image virtual re-staining. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 3904–3913 (IEEE, 2023).
49. Siller, M. et al. On the acceptance of 'fake' histopathology: a study on frozen sections optimized with deep learning. *J. Pathol. Inform.* **13**, 100168 (2022).
50. Liang, J. et al. Deep learning supported discovery of biomarkers for clinical prognosis of liver cancer. *Nat. Mach. Intell.* **5**, 408–420 (2023).
51. Wang, S. et al. Deep learning of cell spatial organizations identifies clinically relevant insights in tissue images. *Nat. Commun.* **14**, 7872 (2023).
52. Nan, Y. et al. Data harmonisation for information fusion in digital healthcare: a state-of-the-art systematic review, meta-analysis and future research directions. *Inf. Fusion* **82**, 99–122 (2022).
53. Vert, J. P. How will generative AI disrupt data science in drug discovery? *Nat. Biotechnol.* **41**, 750–751 (2023).
54. Gutmann, M. & Hyvärinen, A. Noise-contrastive estimation: a new estimation principle for unnormalized statistical models. In *Proc. 13th International Conference on Artificial Intelligence and Statistics* 297–304 (PMLR, 2010).
55. van den Aaron, O., Li, Y. & Vinyals, O. Representation learning with contrastive predictive coding. Preprint at https://arXiv.org/abs/1807.03748 (2018).
56. Taigman, Y., Polyak, A. & Wolf, L. Unsupervised cross-domain image generation. In *Proc. 4th International Conference on Learning Representations (ICLR)* 1441–1455 (ICLR, 2017).
57. Zhang, L., Zhang, L., Mou, X. & Zhang, D. FSIM: a feature similarity index for image quality assessment. *IEEE Trans. Image Process.* **20**, 2378–2386 (2011).
58. Gatys, L. A., Ecker, A. S. & Bethge, M. Image style transfer using convolutional neural networks. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 2414–2423 (IEEE, 2016).
59. Simonyan, K. & Zisserman, A. Very deep convolutional networks for large-scale image recognition. Preprint at https://arXiv.org/abs/1409.1556 (2014).
60. Deng, J. et al. Imagenet: A large-scale hierarchical image database. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 248–255 (IEEE, 2009).
61. Cheng, J., Jaiswal, A., Wu, Y., Natarajan, P. & Natarajan, P. Style-aware normalized loss for improving arbitrary style transfer. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 134–143 (IEEE, 2021).
62. He, K., Gkioxari, G., Dollár, P. & Girshick, R. Mask r-cnn. In *Proc. IEEE International Conference on Computer Vision (ICCV)* 2980–2988 (IEEE, 2017).
63. Graham, S. et al. Hover-Net: simultaneous segmentation and classification of nuclei in multi-tissue histology images. *Med. Image Anal.* **58**, 101563 (2019).
64. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* II-718–II-725 (IEEE, 2016).
65. Glorot, X. & Bengio, Y. Understanding the difficulty of training deep feed-forward neural networks. In *Proc. 13th International Conference on Artificial Intelligence and Statistics* 249–256 (PMLR, 2010).
66. Ulyanov, D., Vedaldi, A. & Lempitsky, V. Instance normalization: the missing ingredient for fast stylization. Preprint at https://arXiv.org/607.08022 (2016).
67. Mao, X. et al. Least squares generative adversarial networks. In *Proc. IEEE International Conference on Computer Vision (ICCV)* 2813–2821 (IEEE, 2017).
68. Kingma, D. P. & Ba, J. Adam: A method for stochastic optimization. Preprint at https://arXiv.org/1412.6980 (2014).
69. Kipf, T. N. & Welling, M. Semi-supervised classification with graph convolutional networks. Preprint at https://arXiv.org/abs/1609.02907, (2017).
70. Bianchi, F. M., Grattarola, D. & Alippi, C. Spectral clustering with graph neural networks for graph pooling. In *Proc. 37th International Conference on Machine Learning (ICML)* 874–883 (PMLR, 2020).
71. Bulten, W. et al. Automated deep-learning system for Gleason grading of prostate cancer using biopsies: a diagnostic study. *Lancet Oncol.* **21**, 233–241 (2020).
72. Ström, P. et al. Pathologist-level grading of prostate biopsies with artificial intelligence. Preprint at https://arXiv.org/1907.01368 (2019).
73. Brierley, J. D., Gospodarowicz, M. K. & Wittekind, C.*TNM Classification of Malignant Tumours* (Wiley, 2017).
74. Jaume, G., Pati, P., Anklin, V., Foncubierta, A. & Gabrani, M. Histocartography: a toolkit for graph analytics in digital pathology. In *Proc. MICCAI Workshop on Computational Pathology* 117–128 (PMLR, 2021).
75. Vahadane, A. et al. Structure-preserving color normalization and sparse stain separation for histological images. *IEEE Trans. Med. Imaging* **35**, 1962–1971 (2016).
76. Voon, W. et al. Evaluating the effectiveness of stain normalization techniques in automated grading of invasive ductal carcinoma histopathological images. *Sci. Rep.* **13**, 20518 (2023).
77. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J. & Wojna, Z. Rethinking the inception architecture for computer vision. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 2818–2826 (IEEE, 2016).
78. Paszke, A. et al. PyTorch: an imperative style, high-performance deep learning library. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)* 8024–8035 (ACM, 2019).
79. Fey, M. & Lenssen, J. E. Fast graph representation learning with Pytorch Geometric. Preprint at https://arXiv.org/abs/1903.02428 (2019).
80. Karkampouna, S. & Kruithof-de Julio, M. Dataset EMPaCT TMA. *Zenodo* https://doi.org/10.5281/zenodo.10066853 (2023).

81. Silva-Rodríguez, J. SICAPv2-prostate whole slide images with gleason grades annotations. *Mendeley Data* https://doi.org/10.17632/9xxm58dvs3.1 (2020).
82. Pati, P. VirtualMultiplexer code. *Zenodo* https://doi.org/10.5281/zenodo.11941982 (2024).
83. Lundberg, S. M. & Lee, S.-I. A unified approach to interpreting model predictions. In *Proc. 31st International Conference on Neural Information Processing Systems*. 4768–4777 (2017).
84. Tang, F. et al. Chromatin profiles classify castration-resistant prostate cancers suggesting therapeutic targets. *Science* **376**, eabe1505 (2022).
85. Blank, A., Dawson, H., Hammer, C., Perren, A. & Lugli, A. Lean management in the pathology laboratory. *Der Pathol.* **38**, 540–544 (2017).

Acknowledgements

We would like to thank G. Jaume, J. Born and M. Graziani for constructive comments, discussions and suggestions. The results published here are in part based upon data generated by the TCGA Research Network at https://www.cancer.gov/tcga. This work was supported by the Swiss National Science Foundation Sinergia grant no. 202297 to M.R. and M.K.-d.J. The PDAC TMA construction took place at the Translational Research Unit Platform of the ITMP of the University of Bern (https://www.ngtma.com/) in the setting of a grant by the Foundation for Clinical-Experimental Cancer Research Bern to M.W. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Author contributions

P.P. conceived and implemented the model. P.P., A.M. and M. Rapsomaniki designed and performed computational analyses. S.K., F.B. and M. Radić performed experiments. S.K., F.B., E.C., M.W. and M.K.-d.J. performed all qualitative assessments. P.P., S.K., F.B., A.M. and M.R. compiled the figures. M.S., M.W. and M.K.-d.J. contributed materials for the experiments. P.P., S.K., F.B. and M. Rapsomaniki wrote the paper with inputs from all authors. M.K.-d.J. and M. Rapsomaniki were responsible for the overall planning and supervision of the project.

Competing interests

The authors declare no competing interests.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s42256-024-00889-5.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s42256-024-00889-5.

Correspondence and requests for materials should be addressed to Marianna Kruithof-de Julio or Marianna Rapsomaniki.

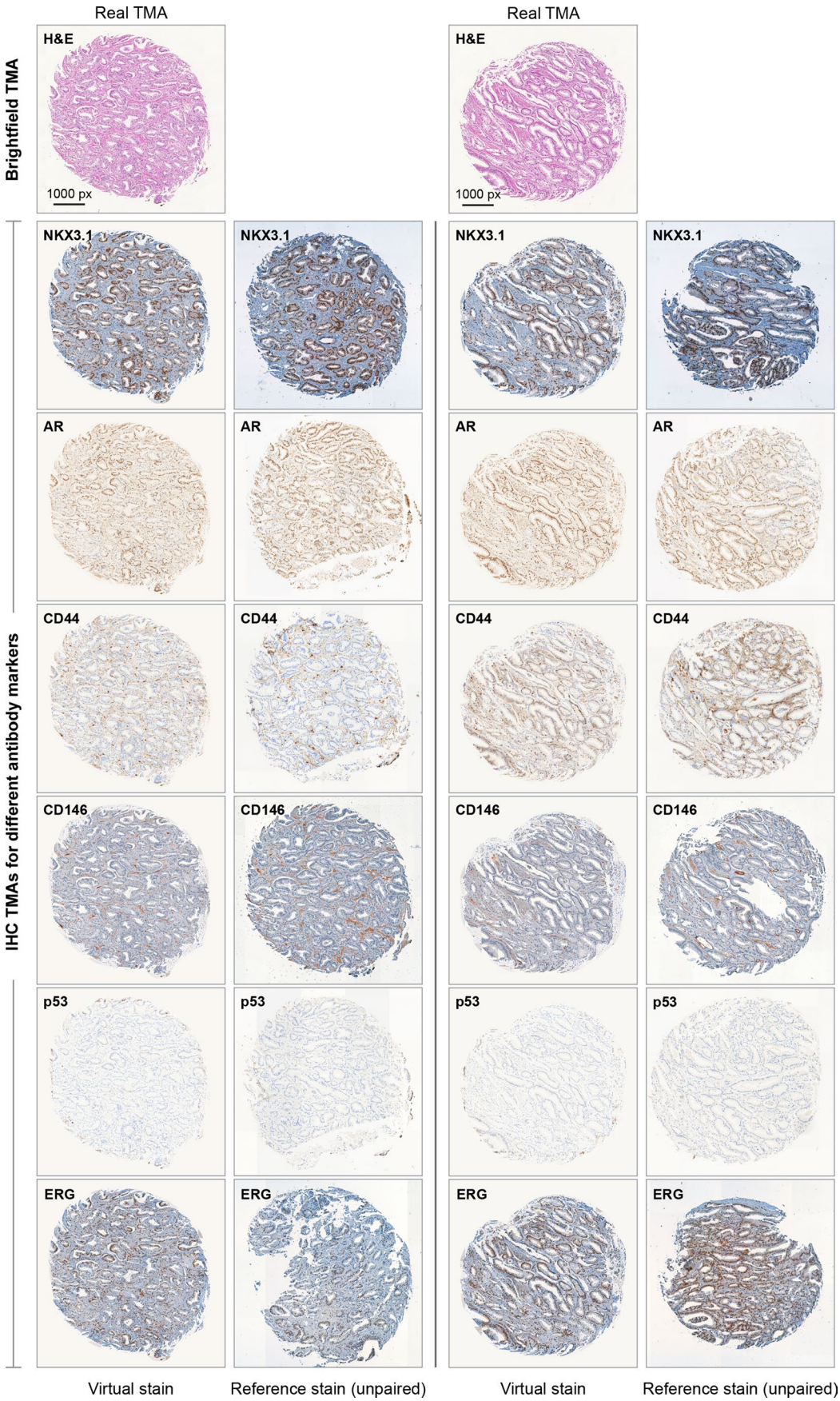
Peer review information *Nature Machine Intelligence* thanks the anonymous reviewers for their contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

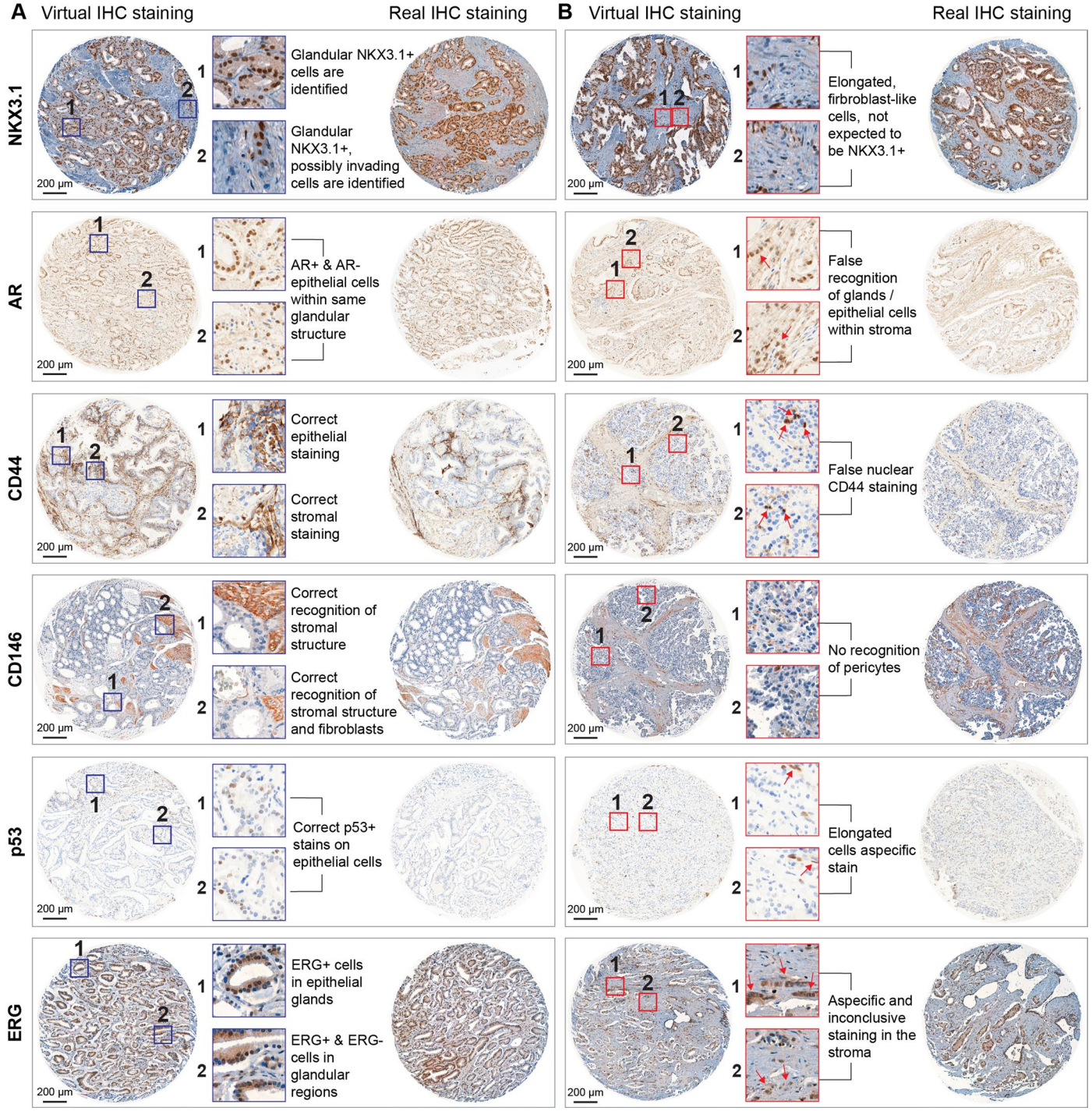
[1]IBM Research Europe, Rüschlikon, Switzerland. [2]Urology Research Laboratory, Department for BioMedical Research, University of Bern, Bern, Switzerland. [3]Department of Urology, Inselspital, Bern University Hospital, University of Bern, Bern, Switzerland. [4]Department of Pathology, Medical University of Vienna, Vienna, Austria. [5]Department of Urology, Lindenhofspital Bern, Bern, Switzerland. [6]Department of Urology, University Duisburg-Essen, Essen, Germany. [7]ETH Zürich, Zürich, Switzerland. [8]Biomedical Data Science Center, Lausanne University Hospital, Lausanne, Switzerland. [9]Institute of Tissue Medicine and Pathology, University of Bern, Bern, Switzerland. [10]Translational Organoid Resource, Department for BioMedical Research, University of Bern, Bern, Switzerland. [11]Faculty of Biology and Medicine, University of Lausanne, Lausanne, Switzerland. ✉e-mail: marianna.kruithofdejulio@unibe.ch; marianna.rapsomaniki@unil.ch

https://doi.org/10.1038/s42256-024-00889-5



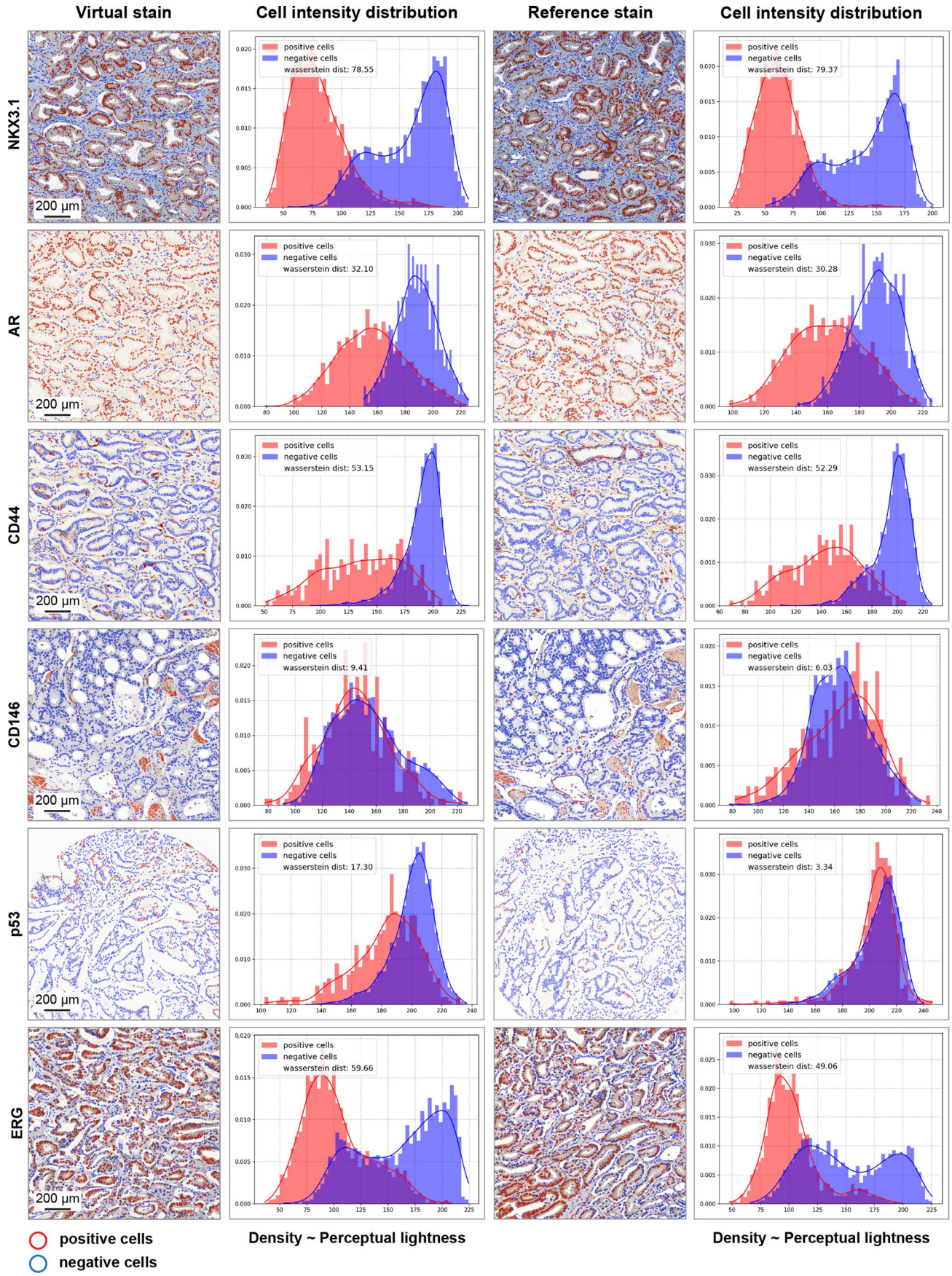
Extended Data Fig. 1 | Qualitative evaluation of the VirtualMultiplexer for two TMA cores in the EMPaCT dataset. Additional examples to the ones presented in Fig. 3. Columns one and three present two H&E stained TMA cores and corresponding virtually stained images for six IHC markers. Columns two and four present reference IHC images for the same core.

https://doi.org/10.1038/s42256-024-00889-5



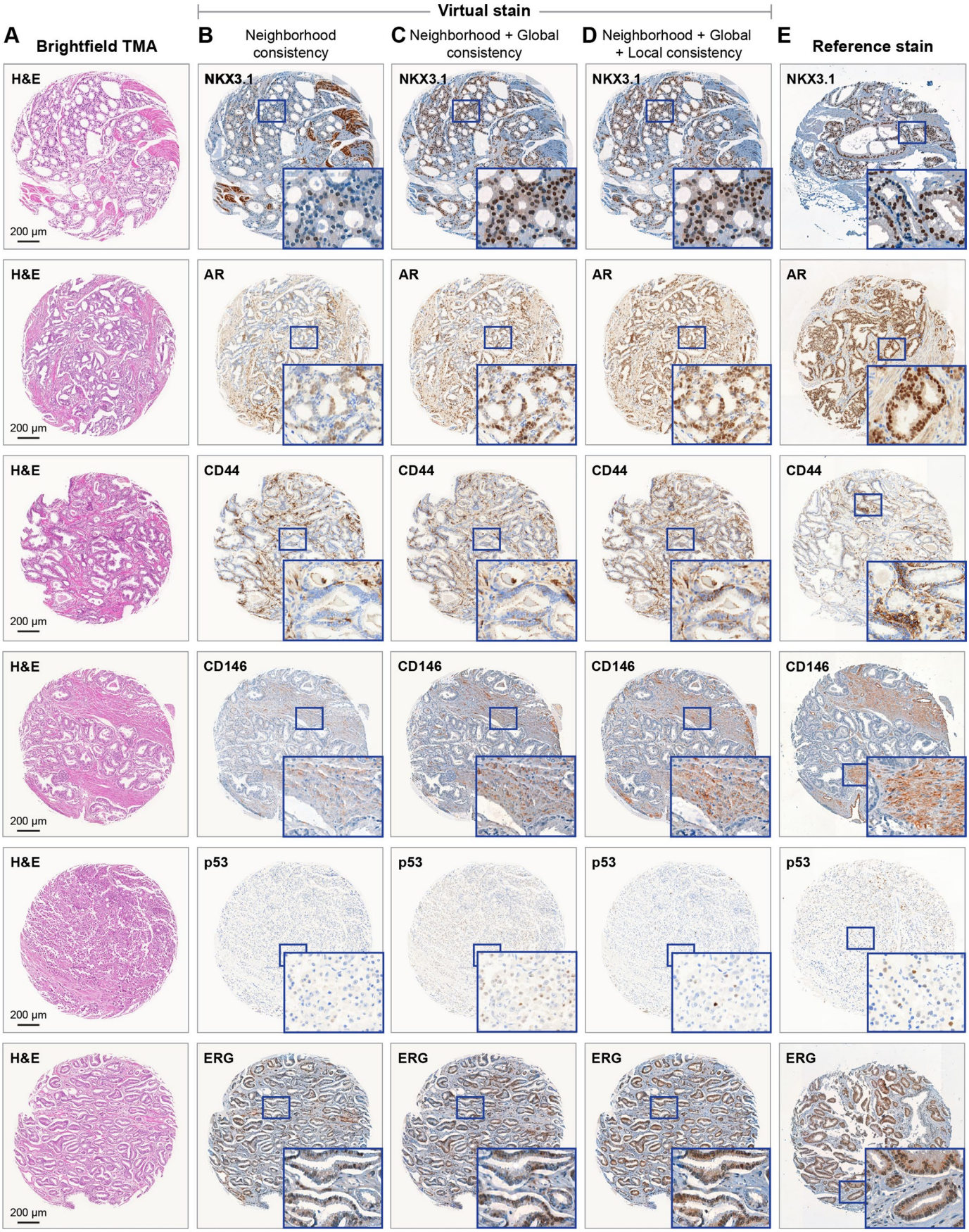
Extended Data Fig. 2 | Visual quality assessment of virtually stained IHC images of the EMPaCT prostate cancer TMA. (A) Example virtual TMA cores across all six markers (left column) and selected zoomed in regions (middle column) that highlight accurate staining patterns. Real reference IHC images for each marker are given on the right column. We observed that AR+ and NKX3.1+ cells exhibited correct distribution in the luminal epithelial compartment of the prostatic glands and nuclear localization. Furthermore, a few NKX3.1+ cells in stromal regions (possibly stroma-invading tumor cells) were correctly predicted. Similarities in specific, matched areas between virtual and real IHC images were mainly assessed for staining pattern and overall intensity levels: we observed that the expression of markers indicative of tumor-specific molecular profile, such as loss of TP53 and ERG overexpression, did not largely deviate between virtual and real images at a TMA core level, which would be crucial for diagnostic applicability. **(B)** Same as (A) but highlighting regions with inaccurate or inconclusive staining. We observed non-specific signal in extra-cellular-matrix/stroma regions (NKX3.1, p53, ERG), occasional false nuclear expression (CD44), and systematic lack of recognition of CD146+ vascular structures.

https://doi.org/10.1038/s42256-024-00889-5



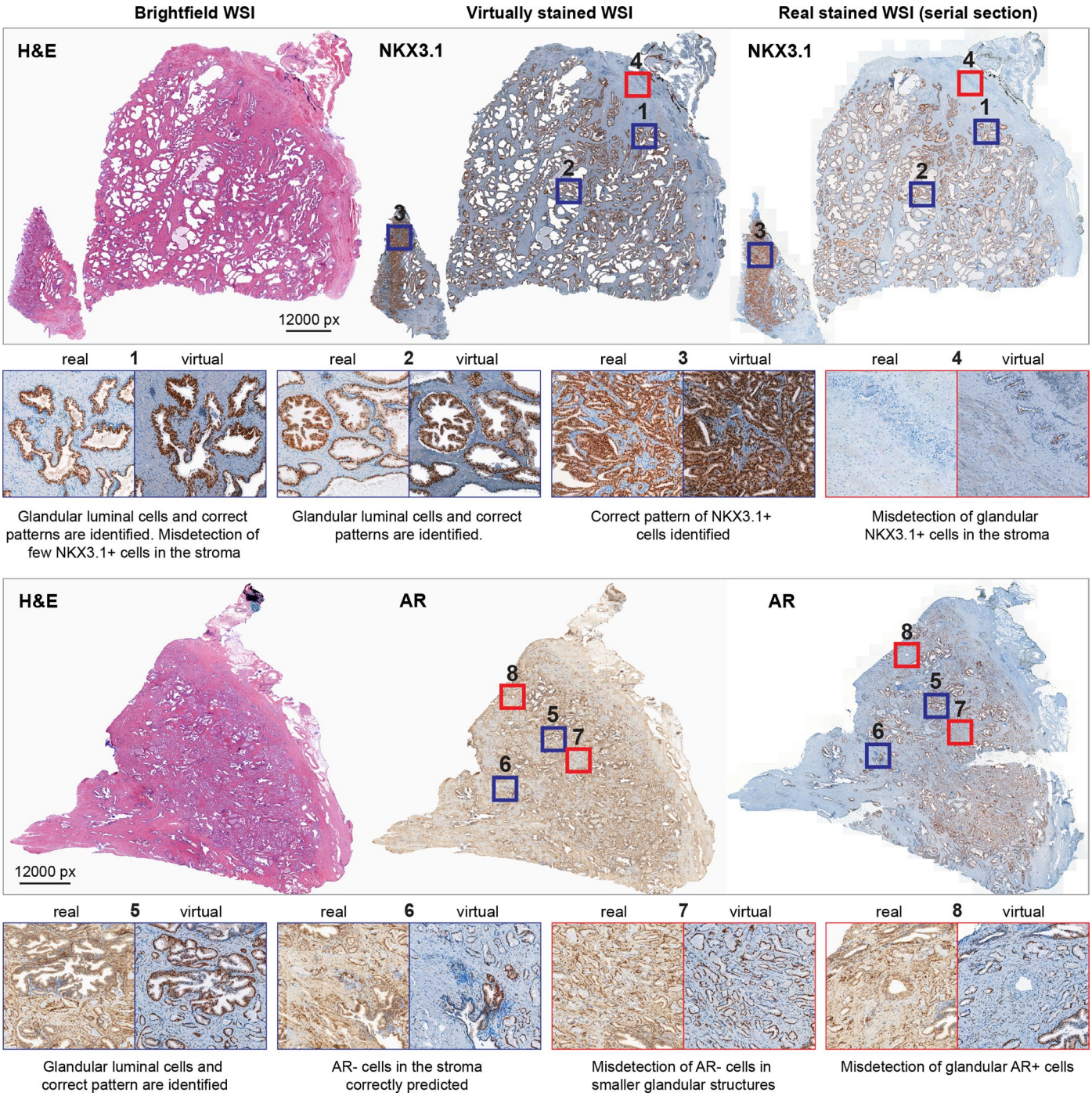
Extended Data Fig. 3 | Intensity distribution of positive and negative cells for real and virtual IHC images. Cell segmentation and classification is performed using DeepLIIF[21]. Intensity of a cell is measured as the average of pixel values in the perceptual lightness (L) channel of Lab colorspace. The Wasserstein distance between the positive and negative cell distributions is computed to quantify the cell-class separability.

https://doi.org/10.1038/s42256-024-00889-5



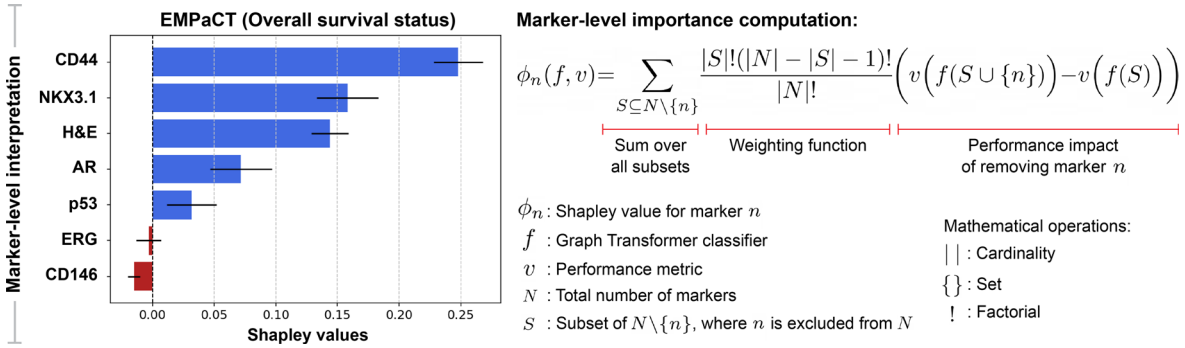
Extended Data Fig. 4 | Ablation study. Qualitative evaluation of the impact of multi-scale consistency objectives on the virtual staining quality of the VirtualMultiplexer across six IHC markers, presented in each row. **(A)** Sample H&E cores from the EMPaCT dataset. Corresponding virtually stained IHC cores for training the VirtualMultiplexer with neighbourhood consistency (**B**), neighborhood and global consistencies (**C**), and neighborhood, global, and local consistencies (**D**). The bounding boxes highlight zoomed-in regions in the IHC cores. **(E)** Reference real IHC cores corresponding to the cores in **(A)**.

https://doi.org/10.1038/s42256-024-00889-5



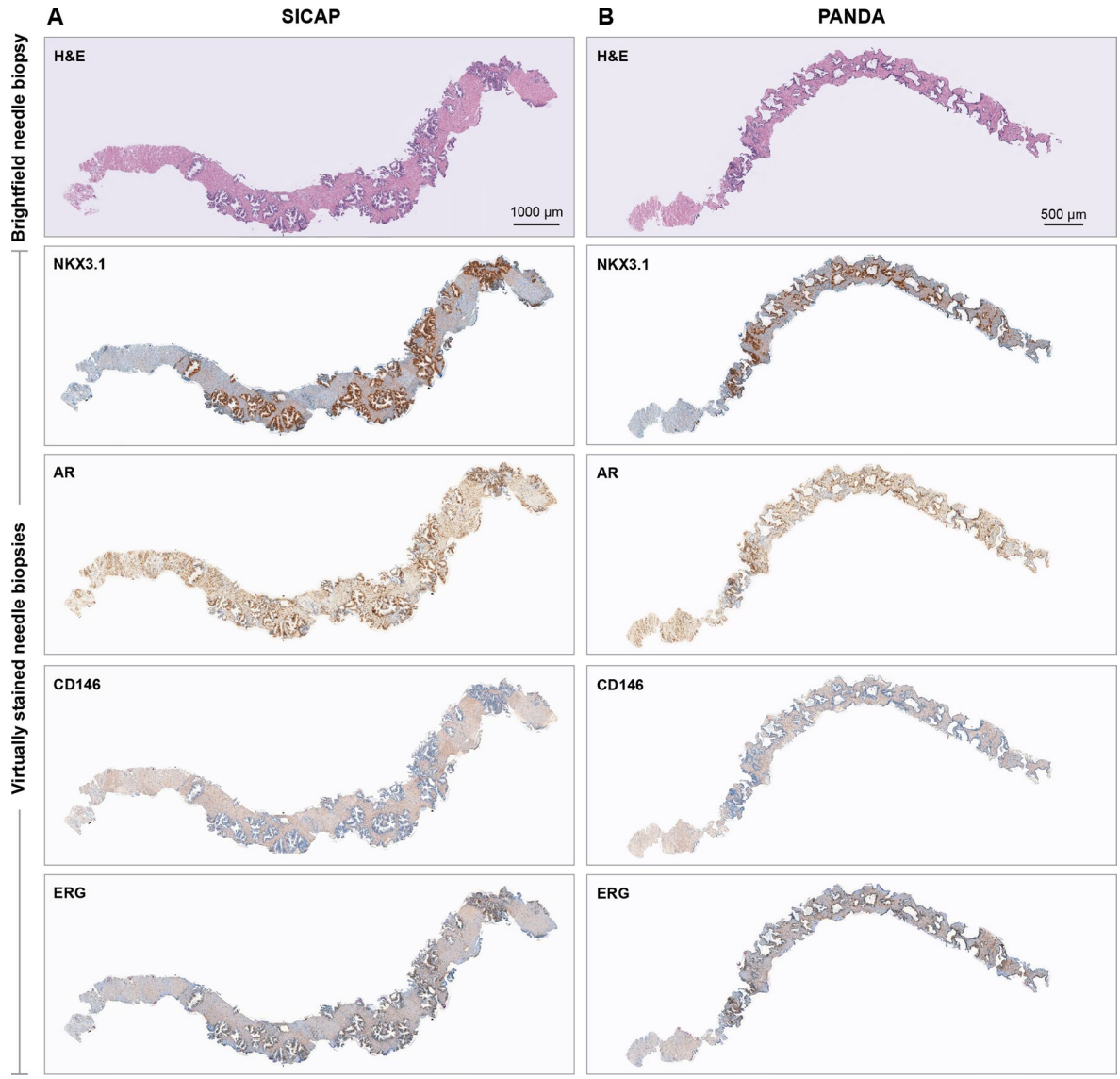
Extended Data Fig. 5 | Transfer learning from TMAs to WSIs of prostate cancer tissue. Additional examples to the ones presented in Fig. 4. Example of H&E (left image), virtual IHC (middle image), and real IHC (right image) staining for NKX3.1 (top) and AR (bottom) of prostate cancer tissue WSIs. Blue-framed zoomed-in regions display accurate staining pattern. Red-framed zoomed-in regions display examples of virtual staining mispredictions.

https://doi.org/10.1038/s42256-024-00889-5



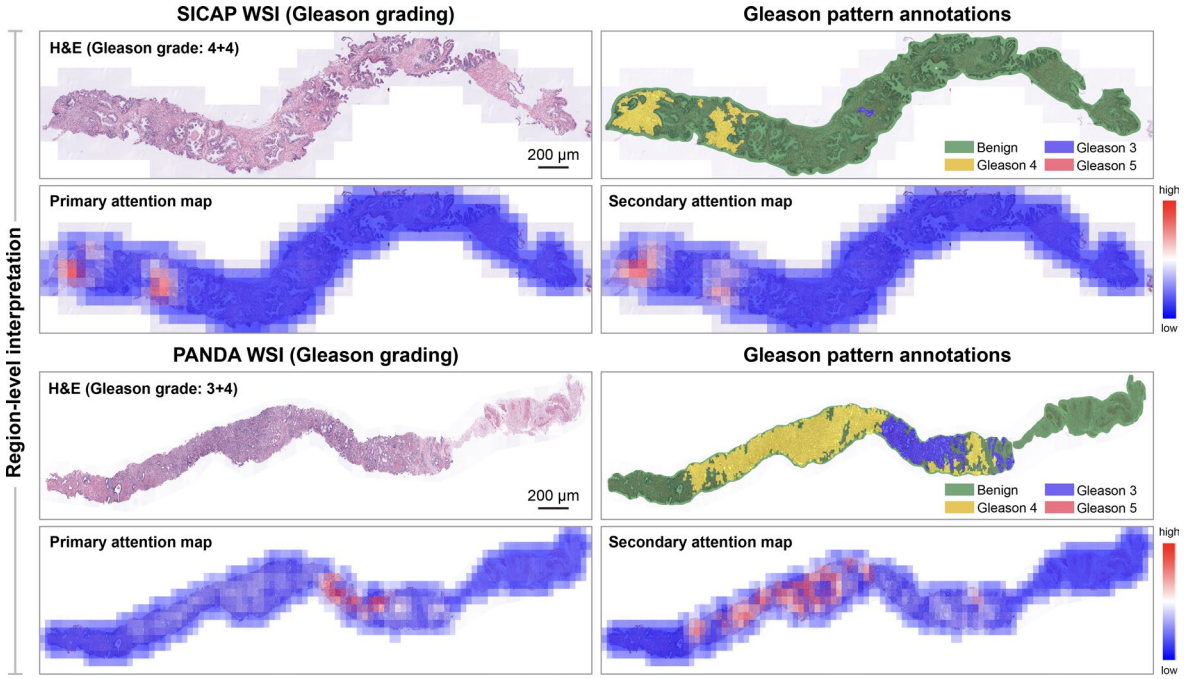
Extended Data Fig. 6 | Marker-level interpretation of Graph-Transformer-based survival prediction classification. For the modality-level interpretation, we performed Shapley Additive Explanations (SHAP)[83] analysis for the overall survival prediction task on EMPaCT (see the relevant computation for reference). We systematically dropped the modalities during inference and measured the change in classification weighted F1 scores, inline with the SHAP algorithm to compute modality-level importance. Here, the barplots and errorbars indicate the mean and the standard deviation, respectively, of the estimated Shapley values across all 134 test images for $n = 3$ Graph-Transformer classifiers. In the absence of ground truth marker importance, we used biological knowledge for qualitative analysis. NKX3.1 and AR were identified as crucial, which is sensible as they both express specific patterns in luminal epithelial cells in prostate and aid in distinguishing normal from carcinoma. High importance of CD44 could be linked to its heterogeneous pattern and pleiotropic effects found in tumor microenvironment[84]. Conversely, CD146's relevance lies in highlighting vascular or fibroblast changes, rendering it less diagnostically informative. Notably, the high importance of CD44 and NKX3.1, and the low importance of CD146 and ERG, are inline with the unimodal high and low weighted F1 scores in Fig. 6, respectively.

https://doi.org/10.1038/s42256-024-00889-5



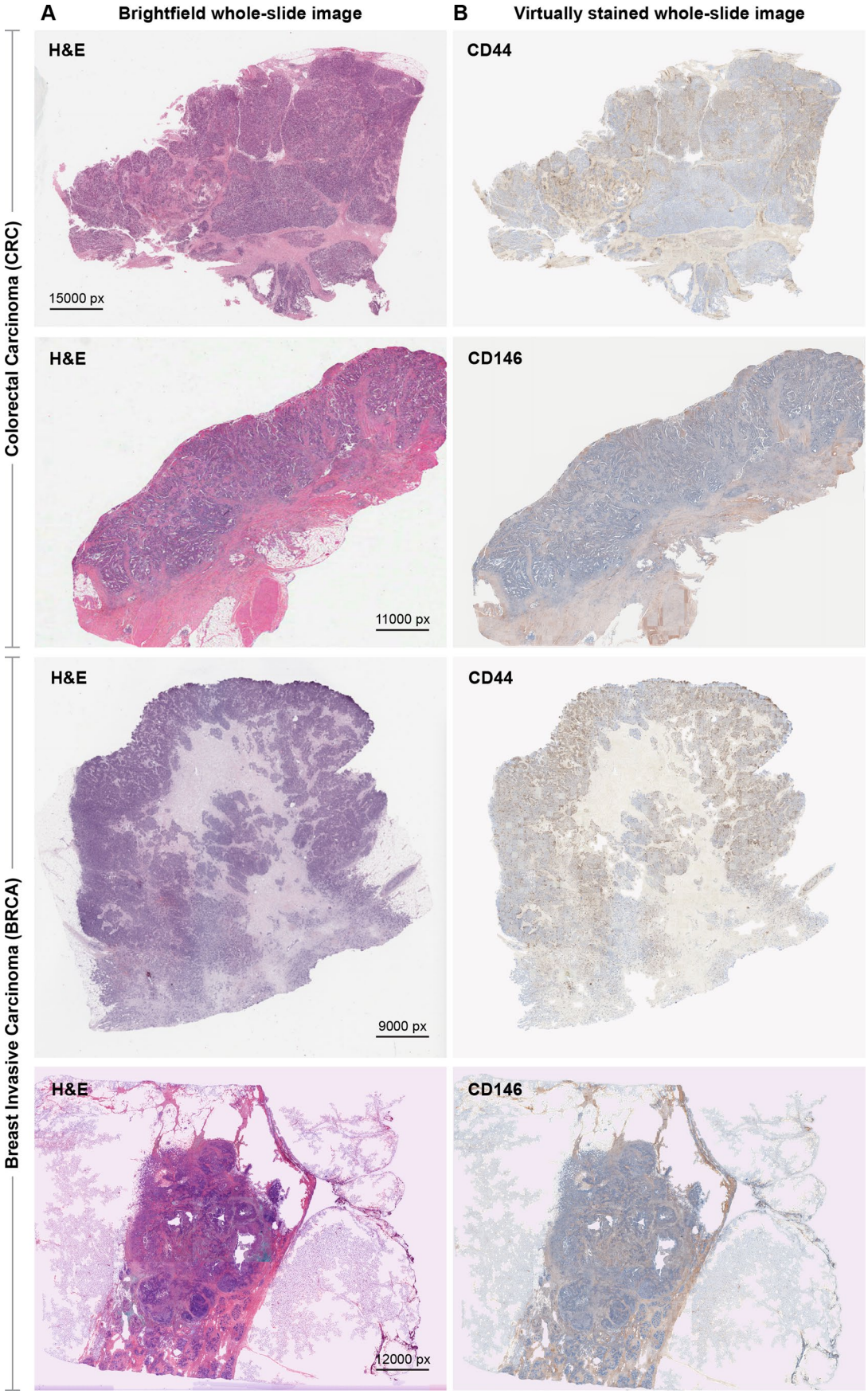
Extended Data Fig. 7 | Transfer learning from TMAs to needle biopsies of prostate cancer tissue. Additional examples to the qualitative samples presented in Fig. 6. **(A)** and **(B)** present H&E biopsies from SICAP and PANDA datasets, respectively, and corresponding virtually stained IHC biopsies for six markers.

https://doi.org/10.1038/s42256-024-00889-5



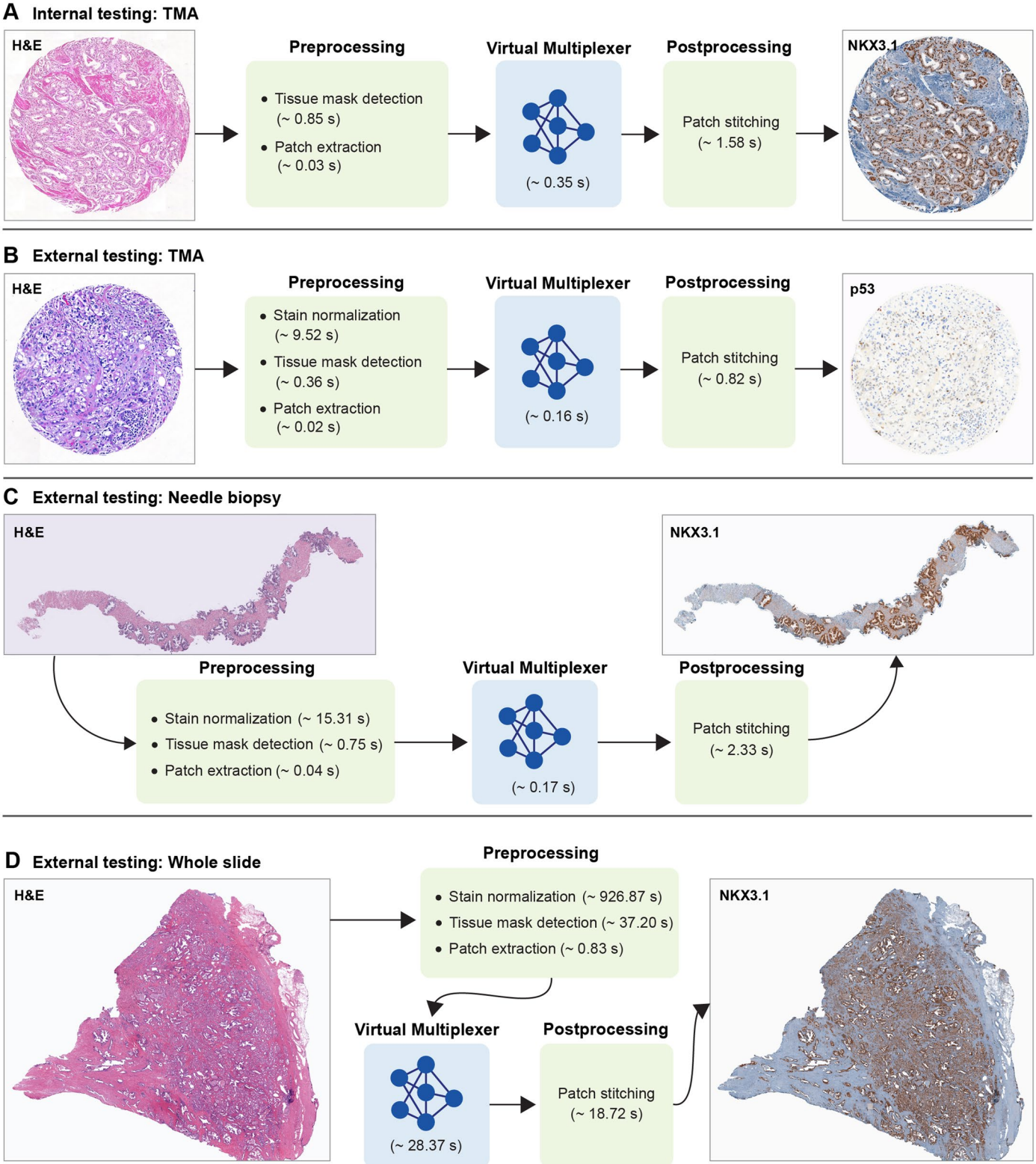
Extended Data Fig. 8 | Region-level interpretation of Graph-Transformer-based Gleason grade classification. Results for sample WSIs from the SICAP[42] (top) and PANDA[71] (bottom) datasets for interpreting the Gleason grading outcome of our Graph-Transformer, with accompanying ground truth annotations of Gleason scores. The model was trained using virtual images under early fusion setting. We used the GraphCAM method from[41] to produce attention maps corresponding to salient tissue regions contributing to model predictions. We observe a great overlap between the identified salient regions and the ground-truth Gleason pattern annotations for both primary and secondary class predictions in both datasets.

https://doi.org/10.1038/s42256-024-00889-5



Extended Data Fig. 9 | Transfer learning from TMAs to WSIs of different tissue types from TCGA cohort. (A) H&E WSIs and **(B)** corresponding virtually stained IHC WSIs from colorectal carcinoma (top two rows) and breast invasive carcinoma (bottom two rows). For both the tissue types, the virtual stainings are produced for relevant CD44 and CD146 IHC markers.

https://doi.org/10.1038/s42256-024-00889-5



Extended Data Fig. 10 | See next page for caption.

https://doi.org/10.1038/s42256-024-00889-5

Extended Data Fig. 10 | The VirtualMultiplexer can greatly accelerate histopathology workflows. We performed a runtime estimation of all components of the VirtualMultiplexer framework across imaging datasets of different scales: an in-domain TMA from the EMPaCT dataset **(A)**, an out-of-domain TMA from the PDAC dataset **(B)**, an out-of-domain needle biopsy from the SICAP dataset **(C)**, and an out-of-domain WSI from the in-house dataset **(D)**. We calculated that applying the trained VirtualMultiplexer on a single EMPaCT TMA core (6000 × 6000 pixels at 20X magnification-0.24μm/pixel) for one marker resulted in a total runtime of 2.81 seconds, and the same process for an out-of-distribution TMA core resulted in a runtime of 10.88 seconds, with the increase attributed to stain normalization. However, the stain normalization step is crucial as it alleviates the appearance disparity between the training and the out-of-distribution samples (Supplementary Fig. 1), and allows for a faithful application of the VirtualMultiplexer to unseen datasets. The above result implies that virtual staining of a hypothetical TMA slide containing 250 out-of-distribution TMA cores for 6 markers would be feasible in ≈ 65.8 minutes (preprocessing: ≈ 9.9 seconds per core, virtual staining and post-processing: ≈ 0.98 seconds per core and marker). Conversely, performing the IHC staining for the same hypothetical TMA for 6 IHC markers could take an estimated time of approximately 1 day, when applied in a cutting-edge pathology laboratory using the latest protocols[85]. When applied in a biology lab that does not specialize in pathology, however, IHC staining could take up to 5 days per marker (sectioning: 1 day, staining: 2 days, slide drying: 1 day, imaging: 1 day), leading to a minimum of 5 days, if done simultaneously for all 6 markers, and more than 10 days, if performed mostly sequentially. Importantly, as our method scales linearly with the size of the tissue (TMA to WSI) and with the number of markers, similar time gains would be feasible for virtually staining needle biopsies and WSIs.

nature portfolio

Corresponding author(s): Marianna Kruithof-de Julio, Maria Anna Rapsomaniki

Last updated by author(s): June 27th, 2024

Reporting Summary

Nature Portfolio 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 Portfolio 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	not applicable
Data analysis	All code used in this work is available at https://github.com/orgs/AI4SCR/VirtualMultiplexer. The architecture of the Graph Transformer follows the official implementation on GitHub (https://github.com/vkola-lab/tmi2022). The image datasets were preprocessed using the Histocartography library (https://github.com/BiomedSciAI/histocartography) version 0.2.1. The deep learning models were developed using PyTorch (version 1.13.1) and PyTorch Geometric (version 2.3.0). The entire pipeline was implemented in Python (version 3.9.1).

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 Portfolio 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 description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

The main dataset used to support this study (EMPaCT) has been deposited in Zenodo (doi:10.5281/zenodo.10066853) together with the prostate cancer WSIs. The

SICAP data is available at Mendeley data (doi: 10.17632/9xxm58dvs3.1). The PANDA dataset is available at the Kaggle website (https://www.kaggle.com/c/prostate-cancer-grade-assessment/data). The TCGA WSIs from breast and colorectal tissue are available as Diagnostic Slides under Project IDs TCGA-BRCA and TCGA-CRC , respectively, at the GDC data portal (https://portal.gdc.cancer.gov). The PDAC dataset is available upon reasonable request from the authors. All clinical data associated with the EMPaCT and PDAC cohorts cannot be shared owing to patient-confidentiality obligations.

Human research participants

Policy information about studies involving human research participants and Sex and Gender in Research.

Reporting on sex and gender	For the prostate cancer datasets (EMPaCT, SICAP, PANDA) the biological sex of all patients is male. For the PDAC dataset, patient distribution in terms of biological sex is reported in Methods-Datasets
Population characteristics	Not relevant
Recruitment	No patient recruitment took place in the context of this study. All histology images analyzed were previously acquired, as described in the manuscript.
Ethics oversight	For the PDAC data, the study followed the guidelines of the World Medical Association Declaration of Helsinki1964, updated in October 2013, and was conducted after approval by the Ethics Committees of Bern (CEC ID2020-00498). All participants provided written general consent. For the EMPaCT data the study was approved by the Ethics Committees of Bern (CEC ID2015-00128).

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

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	Exact sample sizes for all datasets are given in the corresponding Tables in Methods. No sample size determination method was used, we leveraged instead the entirety of the EMPaCT dataset that contains a large number of images per IHC marker with rich clinical metadata to train the model.
Data exclusions	For all datasets, we excluded samples based on image quality, for cases when the image included damaged or missing tissue.
Replication	All experiments were performed in three runs with random initializations, and means/std of all outcomes are reported in all figures.
Randomization	In the context of training the models, dataset division in training, test and validation splits was performed randomly (for EMPaCT and PDAC: at patient level, for SICAP and PANDA: at sample level). Exact sizes of these data splits for all datasets are given in Methods, Tables 4-6.
Blinding	During the visual Turing test, the experts were blinded to the label of the patch (real/virtual). Blinding was not relevant for any other experiment in the paper.

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
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

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

Antibodies

Antibodies used

Antibodies used:
AR: M3562, clone AR441, Dako Agilent,
AR: ab133273, clone EPR1535, Abcam,
NKX3.1: #314, lot 18025, Athena Enzyme Systems,
NKX3.1: 83700T, clone D2Y1A, Cell Signaling,
CD44: ab16728, clone 156-3C11, Abcam,
p53: M7001, clone DO-7, Dako Agilent,
ERG: ab133264, clone EPR3864(2), Abcam,
CD146: ab75769, clone EPR3208, Abcam,
Secondary anti-rabbit antibody Envision HRP : K400311-2, Dako Agilent.

Validation

The concentration and method specifications (e.g. antigen retrieval) used were based on the antibody manufacturer`s datasheet and the standard-operating procedures (S.O.Ps) established at Institute of Tissue Medicine of Pathology, Bern. Isotype controls same as the primary antibodies were used as negative controls.

AR: (M3562, AR441, Dako Agilent). Reacts with human. The antibody was validated by Western blotting (WB) analysis, showing specific identification of a 110 and 112 kD doublet in extracts of the metastatic prostate cancer cell line LNCap and in extracts of cells transfected with the gene for androgen receptor (positive control).

AR: (ab133273, EPR1535, Abcam). Human Androgen Receptor aa 1-100 (N terminal). Reacts with human, mouse and rat. Specificity tested using positive control LnCaP and 22Rv1 cell lysates and rat and mouse prostate lysates by WB. IHC-P: Human prostate, prostatic adenocarcinoma, prostatic hyperplasia tissues. PMID: 35385726.

NKX3.1: antibody #314, lot 18025, Athena Enzyme Systems. Reacts with human. Specificity demonstrated by WB on 22Rv1 prostatic cell line extracts. PMID: 17108105, 18077445.

NKX3.1: 83700T, D2Y1A, Cell Signaling. Reacts with human. Specificity was tested using extracts from human prostate cell lines of both positive (e.g. LNCaP) and negative NKX3.1 expression (DND-41 cell line), as reported in the manufacturer's website. Specificity was additionally validated using chromatin immunoprecipitation on known NKX3.1 target genes.

CD44: ab16728, clone 156-3C11, Abcam. Reacts with human and baboon. Tissue specificity: Isoform 10 (epithelial isoform) is expressed by cells of epithelium and highly expressed by carcinomas. PMID: 34824203.

p53: M7001, clone DO-7, Dako Agilent. Reacts with human. SDS-PAGE analysis of immunoprecipitates formed between lysate of the BT474 breast cancer cell line and the antibody shows reaction with a 53 kDa protein corresponding to p53 (1). In IHC, the antibody labels mutant-type p53 in the A431 cell line and wild-type p53 in the SVK14 cell line (SV40-transformed keratinocyte line). PMID: 7514027. The staining performance of all FLEX RTU antibodies has been defined, tested and approved through collaboration with leading, international pathology experts.

ERG: ab133264, clone EPR3864(2), Abcam. Reacts with human and mouse. Immunohistochemical staining of positive control human colonic carcinoma and Western blotting on HEK293 cell extracts confirmed specificity.

CD146: ab75769, clone EPR3208, Abcam. Reacts with human and mouse. For validation by WB, extracts from HeLa, A375, HUVEC and B16-F0 cell lysate were used; and by IHC-P: Melanoma, breast carcinoma vessel, urinary bladder transitional carcinoma vessel, glioma vessel, normal tonsil and normal spleen tissue were used.