bioRxiv preprint doi: https://doi.org/10.1101/2023.11.29.568996; this version posted December 1, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Multiplexed tumor profiling with generative AI accelerates histopathology workflows and improves clinical predictions

Pushpak Pati[1], Sofia Karkampouna[2], Francesco Bonollo[2], Eva Compérat[3], Martina Radic[2], Martin Spahn[4, 5], Adriano Martinelli[1,6], Martin Wartenberg[7], Marianna Kruithof-de Julio[2,8*], Maria Anna Rapsomaniki[1*]

[1]IBM Research Europe, Säumerstrasse 4, Rüschlikon, 8803, Switzerland.
[2]Department for BioMedical Research, Urology Research Laboratory, University of Bern, Murtenstrasse 24, Bern, 3008, Switzerland.
[3]Department of Pathology, Medical University of Vienna, Währinger Gürtel 18-20, Vienna, 1090, Austria.
[4]Department of Urology, Lindenhofspital Bern, Bremgartenstrasse 117, Bern, 3012, Switzerland.
[5]Department of Urology, University Duisburg-Essen, Universitätsstrasse 2, Essen, 45141, Germany.
[6]ETH Zürich, Rämistrasse 101, Zürich, 3010, Switzerland.
[7]Institute of Tissue Medicine and Pathology, University of Bern, Murtenstrasse 31, Bern, 3008, Switzerland.
[8]Department of Urology, Inselspital Bern, Freiburgstrasse 37, Bern, 3010, Switzerland.

*Corresponding author(s). E-mail(s):
marianna.kruithofdejulio@unibe.ch; aap@zurich.ibm.com;
Contributing authors: pus@zurich.ibm.com;
sofia.karkampouna@unibe.ch; francesco.bonollo@unibe.ch;
eva.comperat@meduniwien.ac.at; martina.radic@unibe.ch;
martin.spahn@hin.ch; art@zurich.ibm.com; martin.wartenberg@unibe.ch;

bioRxiv preprint doi: https://doi.org/10.1101/2023.11.29.568996; this version posted December 1, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Abstract

Understanding the spatial heterogeneity of tumors and its links to disease is a cornerstone of cancer biology. Emerging spatial technologies offer unprecedented capabilities towards this goal, but several limitations hinder their clinical adoption. To date, histopathology workflows still heavily depend on hematoxylin & eosin (H&E) and serial immunohistochemistry (IHC) staining, a cumbersome and tissue-exhaustive process that yields unaligned tissue images. We propose the VirtualMultiplexer, a generative AI toolkit that translates real H&E images to matching IHC images for several markers based on contrastive learning. The VirtualMultiplexer learns from unpaired H&E and IHC images and introduces a novel multi-scale loss to ensure consistent and biologically reliable stainings. The virtually multiplexed images enabled training a Graph Transformer that simultaneously learns from the joint spatial distribution of several markers to predict clinically relevant endpoints. Our results indicate 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. We successfully employed transfer learning to generate realistic virtual stainings across tissue scales, patient cohorts, and cancer types with no need for model fine-tuning. Crucially, the generated images are not only realistic but also clinically relevant, as they greatly improved the prediction of different clinical endpoints across patient cohorts and cancer types, speeding up histopathology workflows and accelerating spatial biology.

Keywords: virtual staining, stain-to-stain translation, multiplexed imaging, generative modeling

Introduction

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 and identify morphological and molecular patterns linked to disease. Among these technologies, H&E staining is the undisputed workhorse, routinely used to assess aberrations in tissue morphology in histopathology workflows across diseases [2]. A notable example is cancer, where H&E staining is routinely used to reveal abnormal cell proliferation, nuclear shape, lymphovascular invasion and immune cell infiltration, among others. Complementary to the morphological information available *via* H&E staining, IHC [3], another staining technique routinely applied in histopathology laboratories, exploits antigen-antibody binding to detect and quantify the abundance of a single protein marker *in situ*. IHC visualizes the distribution and localization of specific markers within cell compartments (membranous, cytoplasmic and/or nuclear) and within their proper histological context, which is crucial for tumor subtyping, prognosis, and personalized treatment selection. As tissue re-staining in conventional IHC is limited, repeated serial tissue sections stained with different antibodies are required for in-depth tumor profiling. However, this is

bioRxiv preprint doi: https://doi.org/10.1101/2023.11.29.568996; this version posted December 1, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

a time-consuming and tissue-exhaustive process, prohibitive in cases of limited tissue availability. At the same time, serial IHC staining yields unaligned, non-multiplexed tissue images occasionally of suboptimal quality in terms of tissue artifacts, and tissue unavailability may lead to missing stainings (Figure 1A). Recently, multiplexed imaging technologies (*e.g.*, imaging mass cytometry - IMC [4], co-detection by indexing - CODEX [5], multiplexed ion beam imaging - MIBI [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, long turnaround time, and need for specialized personnel and equipment severely limit their clinical adoption.

Virtual staining, *i.e.*, the generation of artificially stained tissue images using generative AI models, has emerged as a promising cost-effective, easily accessible and rapid alternative that addresses above limitations [8, 9]. A virtual staining model exploits two sets of images - a *source* set and a *target* set - to learn the source-to-target appearance mapping [10, 11]. During inference, the trained model takes as input a source image and produces a virtually stained target image by simulating the target staining on the source. Initial virtual staining models were based on different flavors of generative adversarial networks (GANs) and operated under a *paired* setting, *i.e.*, 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 tissue images 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 re-staining is not routinely done in most cases, training paired models depends on aligning tissue slices via image registration, a time-consuming and error-prone process, which is often infeasible in practice because of substantial discrepancies even between consecutive slices. Additionally, as tissue architecture largely alters after the first set of slices, retrospective addition of new markers or multiplexing of several markers on a specific area/focal plane of interest 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 S2S translation models are inspired by CycleGAN [30]; they depend on an adversarial loss to preserve the source content (tissue architecture), and a cycle consistency loss to preserve the target style (staining pattern). Some employ additional constraints, *e.g.*, domain-invariant content and domain-variant style [22], perceptual embeddings [24] or structural similarity [25].

However, 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 necessarily hold for many S2S translation tasks. As a result, a persistent problem is staining unreliability, observed as incorrect mappings across domains, *e.g.*, a positive signal from the source domain can get mapped to a negative signal from the target domain. To account for staining unreliability, recent works exploit expert annotations in the source and target domains to guide the translation. For example, Boyd *et al.* [26] translated H&E to Cytokeratin (CK) stained IHC using a

bioRxiv preprint doi: https://doi.org/10.1101/2023.11.29.568996; this version posted December 1, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

region-based CycleGAN and expert annotations of positive and negative metastatic regions on the H&E images. Similarly, Liu *et al.* [25] translated H&E to Ki67 stained IHC by leveraging cancer or normal region annotations in both the H&E and IHC images. Although these approaches show promising results for these specific translation problems, acquiring such annotations is impractical when translating to several IHC markers, and infeasible even for experienced pathologists for more specialized translation tasks (*e.g.*, identifying p53+ cells in H&E images). To circumvent the annotation challenge, Zheng *et al.* [31] recently introduced a semi-supervised approach, which however, again, depends on consecutive tissue sections and image registration. Consequently, there is a great need for S2S translation models that can learn from unpaired source and target images and preserve staining consistency without the need for consecutive tissue sections, image registration or extensive expert annotations on the source domain.

Regardless of the underlying modeling assumptions, an important limitation of the above S2S translation studies concerns model 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 that the generated images accurately preserve complex and biologically meaningful spatial patterns [9]. To alleviate these concerns, a step in the right direction is expert qualitative assessment though pathological examination of the virtual images, employed by some studies [22, 24]. Still, as with all generative AI approaches, a persistent concern is the presence of hallucinations in virtually stained images that might otherwise appear realistic even to experienced pathologists. Ultimately, to ensure that virtually stained images do not only appear visually realistic but are also useful and relevant from a clinical standpoint, using them as input to downstream deep learning models to predict diagnostic or prognostic endpoints could provide an unbiased and convincing validation [9].

Here, we propose the VirtualMultiplexer, a generative toolkit that translates H&E images (source) to matching IHC images (target) for a variety of molecular markers, generating virtually multiplexed tissue images (Figure 1B). The VirtualMultiplexer is inspired by contrastive unpaired translation (CUT) [33], an appealing contrastive learning alternative to CycleGAN that achieves content preservation by maximizing the mutual information between target and source domains. Our toolkit does not depend on pixel-wise aligned H&E and IHC images during training and, in contrast to existing approaches, requires minimal expert annotations only on the target IHC domain. To ensure consistent and biologically relevant virtual stainings, VirtualMultiplexer introduces a novel architecture based on multi-scale constraints at the single-cell, cell-neighborhood and whole-image level that closely mimic human expert evaluation. Specifically, our multi-scale approach is designed to accurately capture the staining specificity at the individual cell level, while also ensuring content and style preservation at a cell neighborhood level and a global image level. We trained the VirtualMultiplexer on a prostate cancer tissue microarray (TMA) dataset containing unpaired H&E and IHC stainings for six clinically relevant nuclear, cytoplasmic, and membrane-targeted markers, and evaluated the generated images using quantitative image fidelity metrics, expert pathological assessment and visual Turing tests.

bioRxiv preprint doi: https://doi.org/10.1101/2023.11.29.568996; this version posted December 1, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

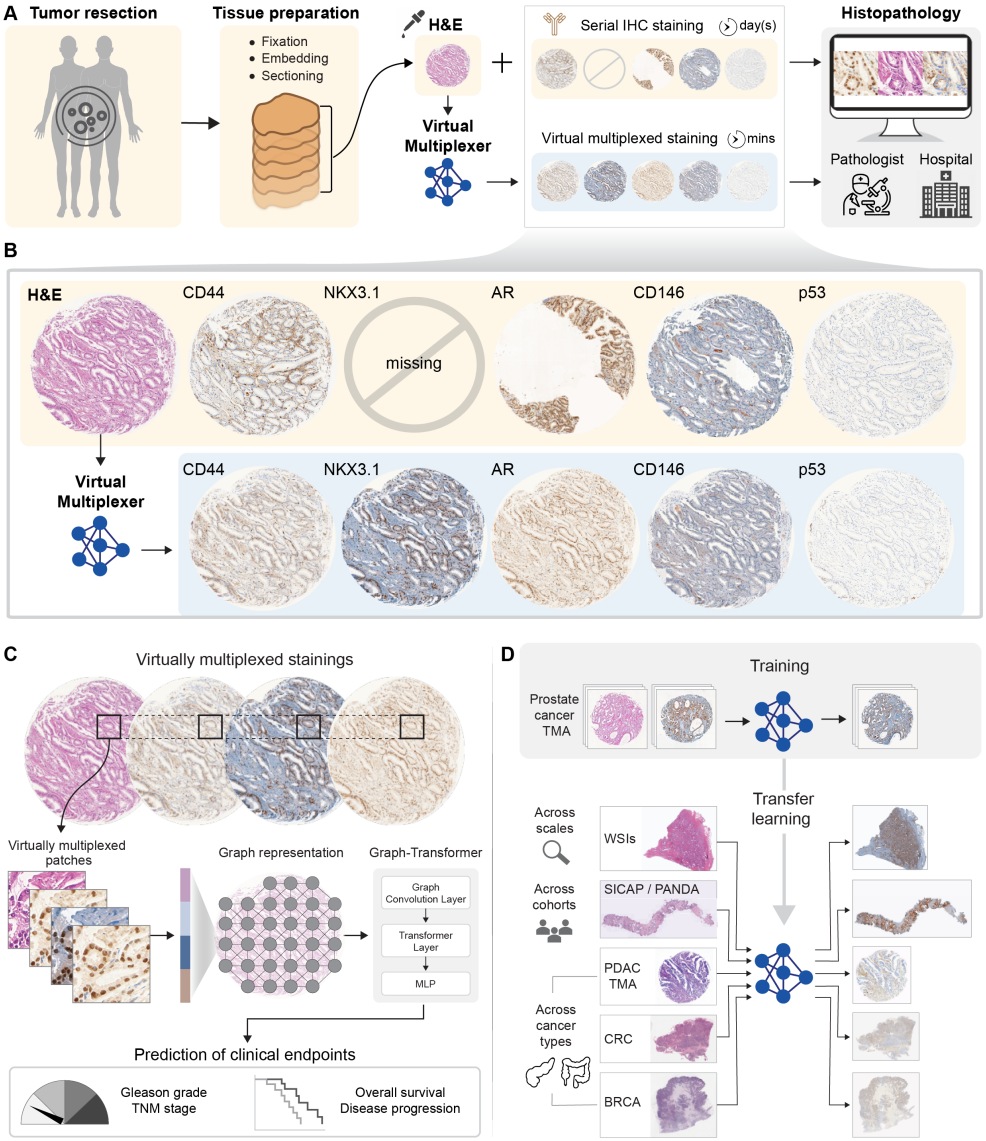

Fig. 1 VirtualMultiplexer: a generative toolkit for synthesizing virtual multiplexed staining. **(A)** In a typical histopathology workflow, serial tissue sections from a tumor resection are stained with H&E and IHC to highlight tissue morphology and molecular expression of several markers of interest. 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 artifacts, 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 is transferable across image scales, patient cohorts and tissue types, accelerating clinical applications and discovery.

bioRxiv preprint doi: https://doi.org/10.1101/2023.11.29.568996; this version posted December 1, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Importantly, to evaluate the clinical relevance of the generated images, we devised a Graph Transformer model that combined a Graph Neural Network and Vision Transformer (ViT) to simultaneously learn from the joint spatial distribution of several markers and predict clinically relevant endpoints (Figure 1C). As our toolkit is not limited by tissue availability or time constraints, we successfully transferred it across tissue image scales (TMAs to whole slide images (WSIs)), across two additional out-of-distribution large prostate cancer patient cohorts, and across tissue types (from prostate to pancreatic, breast and colorectal tissue) (Figure 1C). Our results suggest that the VirtualMultiplexer generates realistic multiplexed IHC stainings of high staining quality which are indistinguishable from real IHC stainings, outperforming existing methods. Importantly, using the generated datasets to train early fusion Graph Transformers surpassed in performance models trained with real unaligned data when predicting clinically relevant endpoints (*e.g.*, survival status, tumor grade and stage) not only in the training cohort, but also in two independent prostate cancer patient cohorts and a pancreatic ductal adenocarcinoma (PDAC) cohort. Overall, the VirtualMultiplexer is a cost-effective, rapid, and easily accessible toolkit that can be readily used to generate virtual multiplexed imaging datasets of high quality, alleviate issues caused by missing modalities and tissue artifacts, improve the prediction of clinical endpoints and generalize across image scales, patient cohorts, and cancer types, with important implications in histopathology.

Results

VirtualMultiplexer: a generative toolkit for virtually multiplexed staining

The VirtualMultiplexer is a generative toolkit for unpaired S2S translation, trained on unpaired real H&E (source) and IHC (target) images. An overview of the model is presented in Figure 2 and a detailed description of its architecture and objective functions is provided in the Methods (Section 8). During training, each image is split into patches of size 256×256 pixels at 10× resolution, which are in turn fed into a generator network G that conditions on input H&E and IHC and learns to transfer the staining pattern, as captured from IHC images, to the tissue morphology, as captured by the H&E images. The generated IHC patches are then stitched together to create a final virtually stained IHC image (Figure 2A). During the S2S translation, we aim to ensure that the virtually stained IHC images are consistent with the real IHC images in terms of appearance, from the macroscopic to the microscopic scale. This implies that the generated IHC images should preserve the staining distribution of real IHC at a patch and tile scale, but also accurately learn the staining specificity of the different markers at the cellular scale. To achieve this, the model jointly optimizes three distinct loss functions, each one designed to preserve image consistency across different scales (Figure 2B). The neighborhood loss (1) ensures that the generated IHC patches are indistinguishable from real IHC patches and consists of an adversarial and a multilayer contrastive loss (Figure 2B), adopted from CUT [33]. The adversarial loss $\mathcal{L}_{\text{adv}}$ (1a) is a standard GAN loss [34], where real and virtual IHC patches are used as input to a convolutional neural network (CNN) patch discriminator network D

bioRxiv preprint doi: https://doi.org/10.1101/2023.11.29.568996; this version posted December 1, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

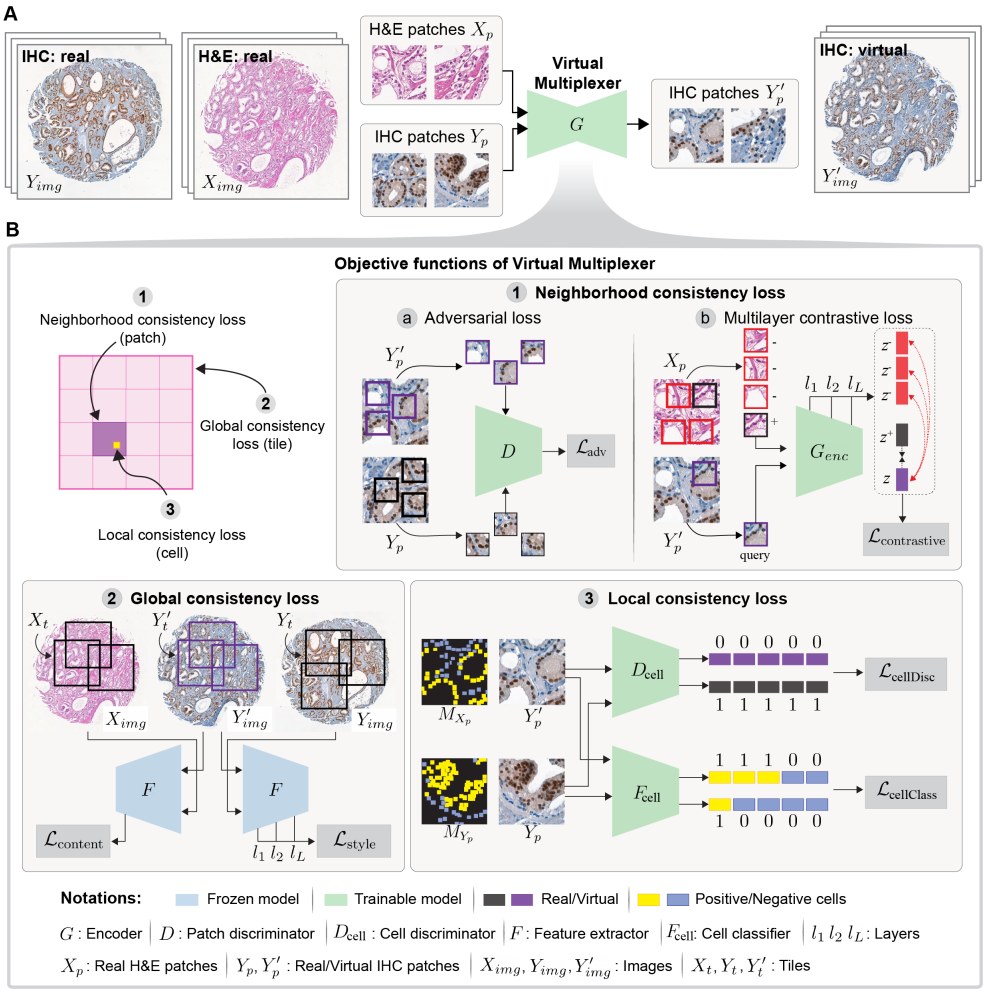

Fig. 2 Overview of VirtualMultiplexer architecture. (A) The VirtualMultiplexer consists of a generator G that takes as input real unpaired H&E and IHC images and is trained to perform S2S translation by mapping the staining distribution of IHC onto H&E while preserving tissue morphology, ultimately generating 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 a neighborhood consistency loss (1) that ensures indistinguishable translations at a neighborhood (patch) level, a global consistency loss (2) that ensures that the model accurately captures content and style constraints at a global tile-level, and a local consistency loss (3) that encodes biological priors on cell type classification and discriminator constraints at a cellular level.

which attempts to classify them as either real or virtual, eliminating style differences between real and virtual patches. The multilayer contrastive loss (1b) is based on a patch-level noise contrastive estimation loss [33] $\mathcal{L}_{\text{contrastive}}$ that aims to ensure that the content of corresponding real H&E and virtual IHC patches is preserved across multiple layers of G_{enc}, *i.e.*, the encoder of the generator G. The VirtualMultiplexer

bioRxiv preprint doi: https://doi.org/10.1101/2023.11.29.568996; this version posted December 1, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

introduces two novel losses, namely a global and a local consistency loss (Figure 2B). The global consistency loss (2) uses a feature extractor F, a pre-trained VGG network [35], and enforces content consistency between the real H&E and the virtual IHC image ($\mathcal{L}_{\text{content}}$), and style consistency between the real and virtual IHC images ($\mathcal{L}_{\text{style}}$) at a tile level across multiple layers of F. In this way the model leverages a high-level sample information, i.e., the correspondence between real H&E and real IHC pairs, to ensure global tissue composition consistency and mitigate the macroscopic appearance variability. Finally, the local consistency loss (3) consists of a cell discriminator loss $\mathcal{L}_{\text{cellDisc}}$ and a cell classification loss ($\mathcal{L}_{\text{cellClass}}$) that together enable the model to capture a realistic appearance and staining pattern at the cellular level while alleviating the multi sub-domain mapping issue. This is achieved by leveraging prior knowledge on staining status via expert annotations, and training two separate networks: a cell discriminator D_{cell} that attempts to eliminate differences in the style of real and virtual cells, and a cell classifier F_{cell} that predicts the staining status and thus enforces staining consistency at a cell-level.

Performance assessment of the VirtualMultiplexer

We first assessed the performance of VirtualMultiplexer in terms of the quality of the generated IHC stainings. We trained the model on a cohort of prostate cancer TMAs from the European Multicenter Prostate Cancer Clinical and Translational Research Group (EMPaCT) [36–38] (Methods). The cohort contained unpaired H&E and IHC stainings from 210 patients with 4 cores per patient for six clinically relevant markers, namely androgen receptor (AR), NK3 Homeobox 1 (NKX3.1), CD44, CD146, p53 and ERG. The VirtualMultiplexer was used to generate a series of virtual IHC stainings (Figure 3C) that preserved the tissue morphology as seen in the real H&E image (Figure 3A) and the staining pattern as seen in the real IHC image (Figure 3B). Additional examples for two more TMA cores and all IHC markers are presented in the Appendix Figure A1. We quantitatively compared the results of the VirtualMultiplexer with four state-of-the-art unpaired S2S translation methods, namely CycleGAN [30], CUT [33], CUT with kernel instance normalization (KIN) [39], and AI-FFPE [29] using the Fréchet Inception Distance (FID), an established metric used to assess the proximity of images created by a generative model to a set of target domain images [40]. The VirtualMultiplexer resulted in the lowest FID score across all markers (Figure 3D), with an average value of 29.2 ($\pm$3), which was consistently lower than the competing methods CycleGAN (49 $\pm$ 6), CUT (35.8 $\pm$ 4.5), CUT with KIN (37.8 $\pm$ 2.3), and AI-FFPE (35.9 $\pm$ 2.6). This result indicated that virtual stainings generated by the VirtualMultiplexer were the closest to the real IHC stainings in terms of distribution than any of the competing methods.

To further quantify the indistinguishability of real and synthetically stained images, we conducted a visual Turing test as follows: three independent evaluators with expertise in prostate tissue histopathology and one board-certified pathologist were shown 100 randomly selected patches per marker, with 50 of them originating from real and 50 from virtually stained IHC images, and were asked to classify each patch as virtual or real. We observed that our model was able to trick the experts, as it achieved an average sensitivity of 52.1% and specificity of 54.1% across all six markers (Figure 3E)

bioRxiv preprint doi: https://doi.org/10.1101/2023.11.29.568996; this version posted December 1, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

(note that a random classification into real or virtual would correspond to a sensitivity and specificity of 50%). Last, we performed a staining quality assessment: we provided the pathologist 50 real and 50 virtual images from 50 TMA cores per IHC marker, revealing which are real and which are virtual, who in turn performed a qualitative assessment of the staining, as judged by overall expression levels, overall background, staining pattern, cell type specificity and subcellular localization (Figure 3F). Across all six markers, on average 70.7% of the virtually stained images reached an acceptable staining quality, as opposed to 78.3% of the real images. The results varied depending on the marker, with cores virtually stained for NKX3.1 and CD146 achieving the highest staining quality of 96%, surpassing even the real images. Conversely, virtually stained AR images had the lowest score of 46%, with an additional 10% exhibiting accurate staining but high background, and the remaining 42% rejected due to mostly heterogeneous staining, or falsely unstained cells. Background staining presented a challenge with CD44 and p53; the latter appeared to be further affected by border artifacts, *i.e.*, presence of abnormally highly stained cells only in the border of the core, an artifact also occasionally present in real images. ERG achieved a higher staining quality assessment in virtual than in the real images, which both appeared to often face background issues. We concluded that, for most of the markers, the staining quality scores and the number of cores with staining artefacts (high background and/or border artefacts) were comparable in virtual versus real images.

Following these observations, we carefully examined the virtually generated images and assessed to which extend the VirtualMultiplexer captured the staining patterns of the real ones. The virtual images showed similar pattern and signal distribution to the real images for all six markers, with correct cell type and subcellular distribution (Figure 4A). For instance, AR+ and NKX3.1+ cells were evaluated as having 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 in the virtually stained cores. Similarities in specific, matched areas between virtual and real IHC images were also assessed mainly for staining pattern and overall intensity levels. To this end, we specifically assessed that the virtual expression of markers indicative of tumor-specific molecular profile, such as loss of TP53 and ERG overexpression, did not largely deviate from the real IHC at the overall TMA core level (Figure 4A), which would be crucial for diagnostic applicability. Certain discrepancies in virtual versus real images were also found, such as non-specific signal in extra-cellular-matrix/stroma regions (NKX3.1, p53, ERG), false nuclear expression (CD44), and systematic lack of recognition of CD146+ vascular structures (Figure 4B). Nonetheless, the more pathologically relevant staining patterns were correctly reconstructed, as described in Figure 4A. We also performed an ablation study demonstrating the effects of training with different components of the objective function of the VirtualMultiplexer (Appendix Figure A2). Indeed, we observe that incorporating multi-scale loss terms, from the whole image to the cellular level, allows us to capture biologically consistent staining patterns that resemble the real IHC stainings. The mere imposition of the neighborhood consistency, the primary objective employed in competing methods, produces virtually stained tissue regions mimicking the overall staining distribution of real target IHC stainings,

bioRxiv preprint doi: https://doi.org/10.1101/2023.11.29.568996; this version posted December 1, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

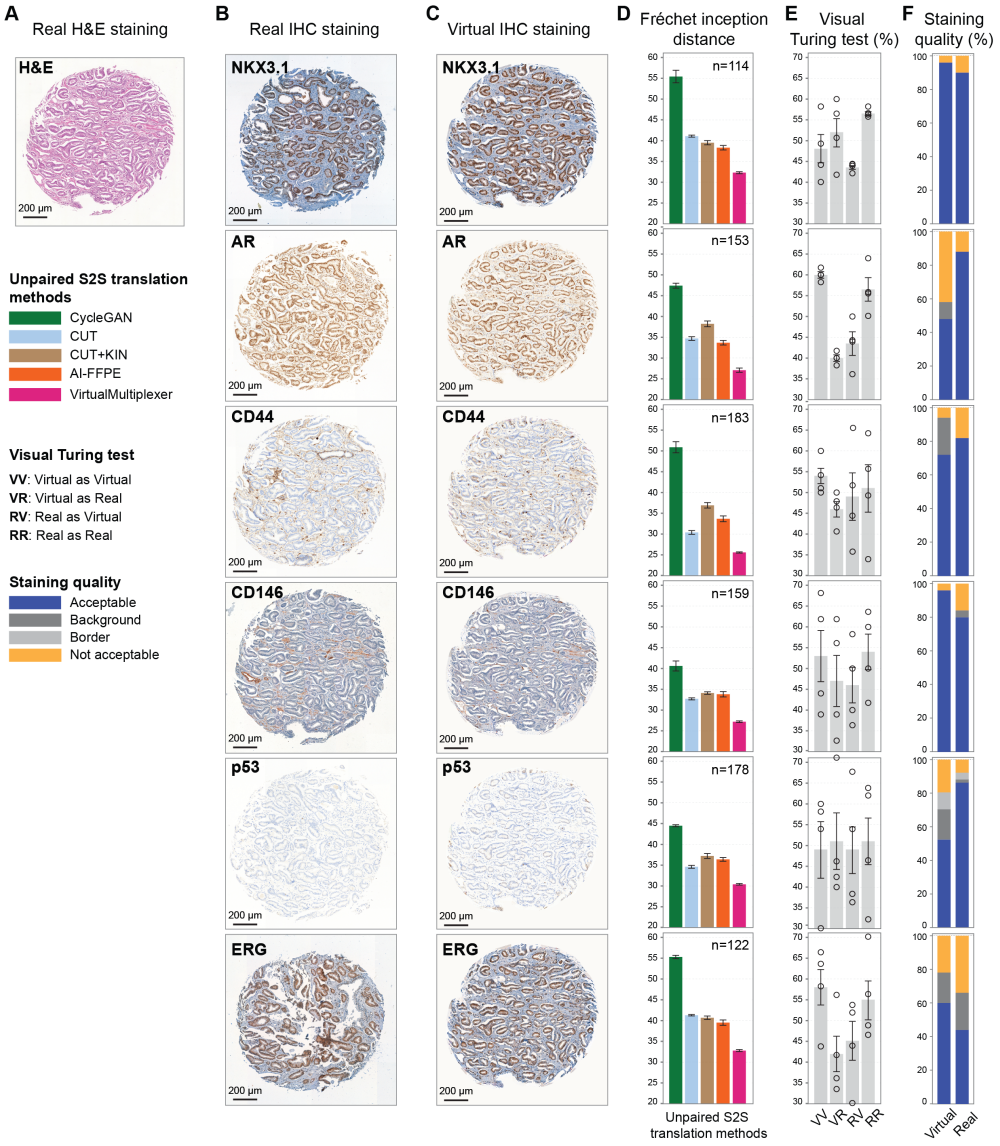

Fig. 3 Performance evaluation of VirtualMultiplexer. **(A)** Example H&E core from the EMPaCT TMA. **(B)** Real, unpaired IHC-stained cores for different antibody markers corresponding to the H&E core in (A). **(C)** Virtually stained IHC cores, now paired with the H&E core in (A). **(D)** Comparison of the VirtualMultiplexer with state-of-the-art S2S models. Barplots and errorbars indicate the mean and standard deviation of the FID score from 3 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 the guess of each one of the four experts, and barplots and errorbars 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.

bioRxiv preprint doi: https://doi.org/10.1101/2023.11.29.568996; this version posted December 1, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

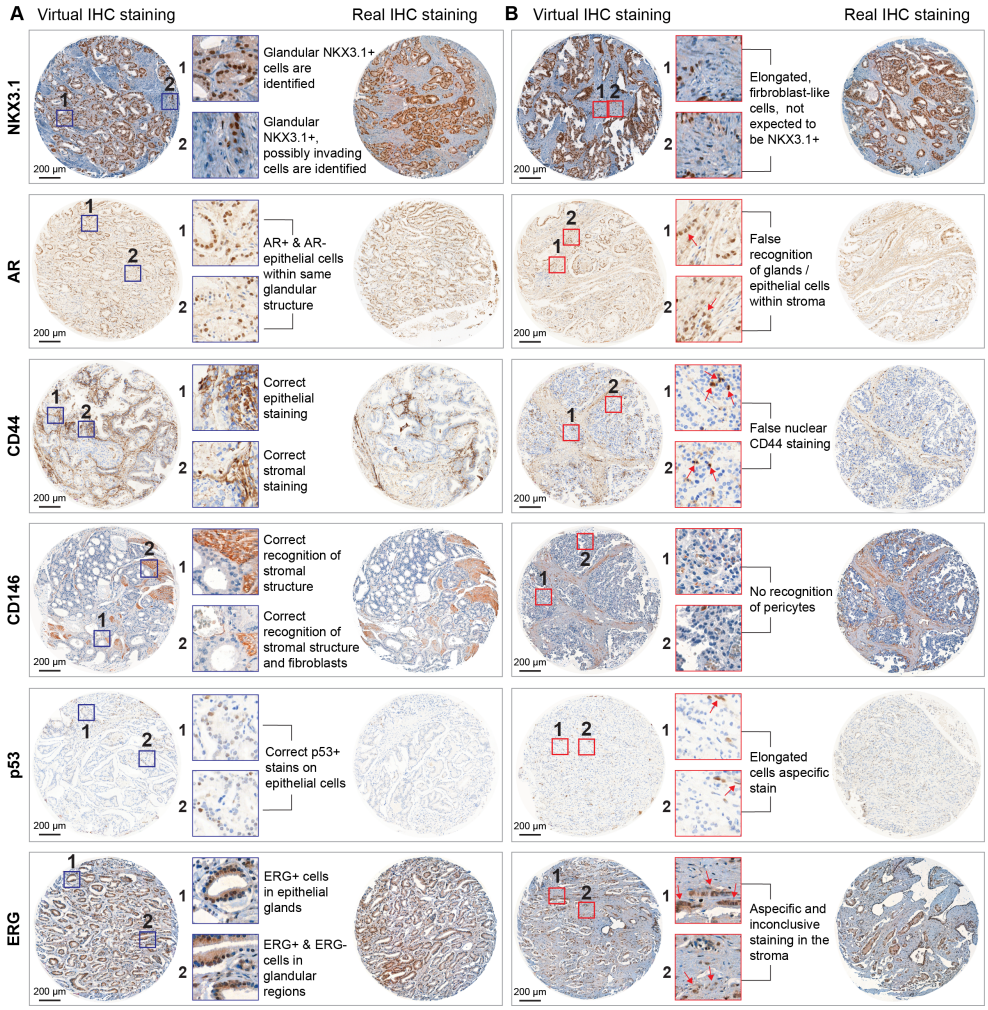

Fig. 4 Visual quality assessment of virtually stained IHC images of the EMPaCT prostate cancer TMA. (A) Example virtual TMA cores across all 6 markers (left column) and selected zoomed in regions (middle column) that highlight accurate staining patterns. Real reference stainings for each markers are given on the right column. **(B)** Same as (A) but highlighting regions with inaccurate or inconclusive staining.

but clearly leads to staining unreliability, *e.g.*, swapping of staining patterns between positive and negative cells. The introduction of the global consistency clearly mitigates this issue, and the further addition of the local consistency further optimizes the virtual staining at the cell-level.

Transfer learning 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 prostate tissue WSIs stained

bioRxiv preprint doi: https://doi.org/10.1101/2023.11.29.568996; this version posted December 1, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

with H&E, and generated virtually stained IHC images for the NKX3.1, AR, and CD146 markers, as before. We then stained for the same markers by IHC on the direct serial sections, thus generating a real stained WSI that can be directly used to visually validate the model predictions. For marker NKX3.1 (Figure 5), we observed that the virtually stained 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 Figure 5 and Appendix Figure A3) and accurate non-annotation of stromal or vascular structures (absence of signal) (example 3 in Figure 5 and Appendix Figure A3). 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 Figure 5 and Appendix Figure A3). For both CD146 and AR, we observed staining intensity discrepancies between virtual and real images, which were more striking in the case of CD146 where the overall signal intensity and background is higher in the virtual versus the real images (Figure 5 and Appendix Figure A3). These discrepancies can be reasoned to the fact that the VirtualMultiplexer has been trained on TMA images that have a different staining distribution than the WSIs. This might lead to false interpretation of the marker expression levels at a first visual inspection. However, when evaluating at higher magnification, the staining pattern in the matching regions of real and virtual stainings was effectively correct, *e.g.*, no glandular signal (example 5 in Figure 5) and appropriate stromal localization of CD146 (examples 6 and 7 in Figure 5) and nuclear localization of AR in luminal epithelial cells (example 1 in Appendix Figure A3). Lack of detection of vascular structures for CD146 was evident in both TMA cores and WSI (example 8 in Figure 5).

The VirtualMultiplexer leads to improved clinical predictions

We further substantiated the utility of the generated virtual IHC stainings in augmenting the performance of AI models when predicting clinically relevant endpoints. To this end, we benchmarked the classification performance of AI models that were trained using real H&E, real IHC, or virtual IHC images. Specifically, we encoded the images as tissue-graph representations and employed a Graph-Transformer (GT) [41] to map the representations to downstream class labels. First, we extracted patches from the images and used a pretrained ResNet-50 network [42] to encode patch features (Figure 6A and Methods). The patches and their features formed the nodes and node representations of the tissue-graph, and the edges were formed using the spatial distribution of the patches (Figure 6B and Methods). The graph representation underwent graph convolutions to contextualize the node features of the local tissue neighborhood. Afterwards, the node features were pooled and fed to a transformer layer, trained to predict clinical endpoints. Depending on how the patch features were combined, we trained the GT model under the following three settings (Figure 6C): (i) a unimodal setting, where independent GT models were trained for each H&E and IHC marker, (ii) a multimodal late fusion setting, where the outputs of independent GT models were fused at the last embedding stage, and (iii) a multimodal early fusion setting, where the patch features were combined early in the tissue-graph and fed into the GT model. While the unimodal setting resulted in a separate prediction per marker, in both multimodal settings the patch features were combined, resulting

bioRxiv preprint doi: https://doi.org/10.1101/2023.11.29.568996; this version posted December 1, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

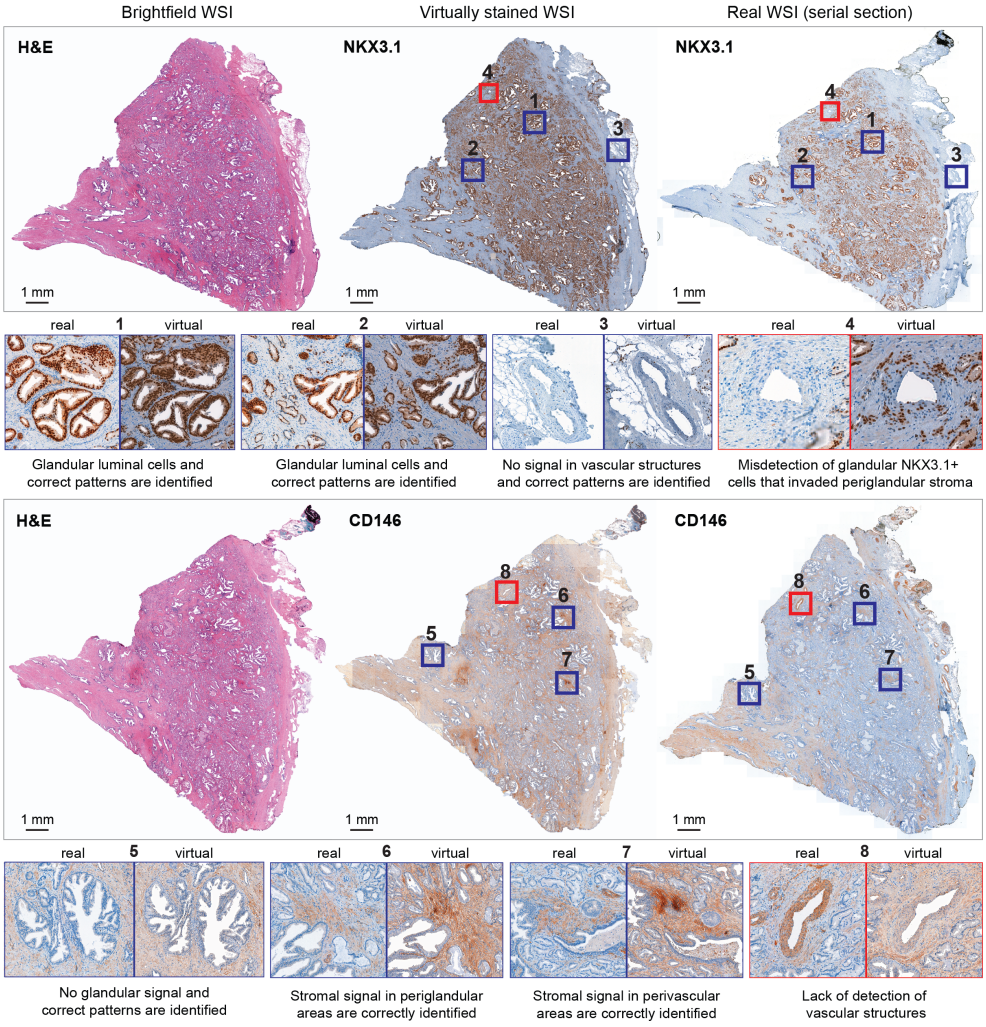

Fig. 5 Transfer learning from TMA to WSIs of prostate cancer tissue. Example of H&E (left image), virtual IHC (middle image), and real IHC (right image) staining for NKX3.1 (top) and CD146 (bottom) of prostate cancer tissue WSIs. Blue-framed zoomed-in regions display accurate staining pattern. Red-framed zoomed-in regions display examples of virtual staining mis-predictions.

in one prediction across all markers. However, in contrast to the late fusion multimodal setting that necessitated the training of several GT models, in the early fusion case only one model that learned from the joint spatial distribution across all markers was trained, mimicking a multiplexed imaging scenario. With the exception of the early fusion setting that was 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 (Figure 6D, legend).

We applied these settings to the EMPaCT dataset to predict two clinically relevant endpoints, namely the overall survival status and the disease progression of patients

bioRxiv preprint doi: https://doi.org/10.1101/2023.11.29.568996; this version posted December 1, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

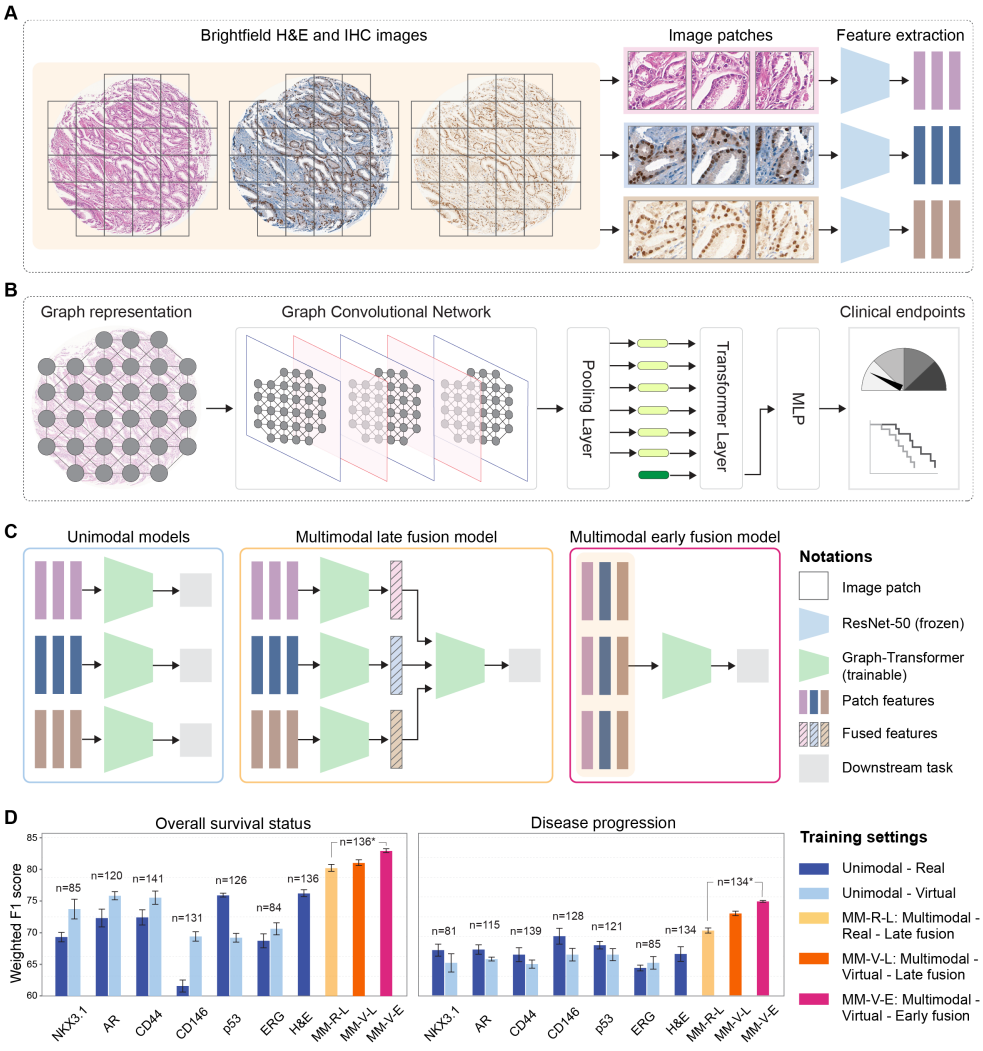

Fig. 6 Prediction of clinically relevant downstream tasks with virtually multiplexed data. **(A)** Patch extraction and computation of patch features with a frozen ResNet-50 model (blue trapezoid). **(B)** Overview of the Graph-Transformer model, implemented by first constructing a patch-level graph representation, followed by a transformer that processes the graph representation to predict clinically relevant endpoints. **(C)** Training of Graph-Transformer 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 colors indicate one of the five combinations of training setting and input data used (see legend). For each combination, barplot heights and errorbars indicate the mean and standard deviation of the weighted F1 score, as computed in the held-out test set from 3 independent runs with different initializations. The exact number of training samples used in each cases is given on the top of the barplots. * For all multimodal models, the reported number refers to the union across all markers.

(Figure 6D). We note that in the EMPaCT dataset few IHC marker stainings are missing, leading to small discrepancies in the number of real IHC images available across

bioRxiv preprint doi: https://doi.org/10.1101/2023.11.29.568996; this version posted December 1, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

all six markers. To ensure a fair comparison between real and virtual unimodal models, we matched the number of virtual IHC images to the number of available real IHC images, which implies that the dark and light blue barplots in Figure 6D are directly comparable. However, as H&E images were available for all patients, this issue did not affect the unimodal model trained on H&E images, which has a slight advantage over all other models in terms of number of samples used. Importantly, to compare all multimodal models, we again followed the same strategy of matching the number of virtual images used to the ones available in real data, and thus the last 3 bars in Figure 6D are also directly comparable. More details of the dataset distribution and the GT training are presented in Methods (Section 8). We observed that the unimodal models trained with virtual images are on par with unimodal ones trained with real images for both tasks, and the predictive performance of the unimodal models trained with virtual IHC images varied depending on the marker's predictive ability towards the downstream task. In the case of overall survival status prediction, two interesting exceptions concern CD146 and p53: for CD146 the unimodal models trained on virtual data outperformed the ones trained on real data, which is in accordance to the previous observation that virtual CD146 images achieved a higher quality assessment than real CD146 images (Figure 3F). The opposite is true for p53: virtual p53 images were of lower quality than real p53 images, and the corresponding unimodal - virtual prediction models achieved a lower performance than the unimodal - real ones. However, these observations were not replicated for disease progression prediction, which appeared to be an overall harder prediction task. In both prediction tasks, the multimodal settings outperformed the unimodal H&E results, indicating the utility of combining information from complementary markers over individual stainings. Furthermore, the multimodal early fusion model trained with virtual images achieved the best weighted F1 score of 82.9% and 74.8% for overall survival status and disease progression, respectively, establishing the potential of multiplexed analysis via virtual staining for augmenting the efficacy of AI models. Overall, we concluded that using virtual images generated by the VirtualMultiplexer can boost the performance of state-of-the-art AI models for clinically relevant endpoints.

Transferring the VirtualMultiplexer across patient cohorts and cancer types

We assessed the ability of the VirtualMultiplexer model to generalize to out-of-distribution data by employing two independent prostate cancer patient cohorts, namely SICAP [43] and PANDA [44], each one containing H&E stained needle biopsies with associated Gleason scores (details in Methods). We virtually stained the H&E images for four IHC markers, namely NKX3.1, CD146, AR and ERG, using the pre-trained VirtualMultiplexer on the EMPaCT dataset (example needle biopsy for SICAP in Figure 7A, additional examples for both SICAP and PANDA in Appendix Figure A4). The IHC markers were chosen due to their possible relevance towards Gleason score prediction. We observed that the virtual staining patterns of the IHC markers were overall correct and specific for each marker in terms of cell type and subcellular localization, with the only exception the occasional aspecific AR signal in the extracellular matrix areas. Other inconsistencies include the weak staining of interstitial tissue

bioRxiv preprint doi: https://doi.org/10.1101/2023.11.29.568996; this version posted December 1, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

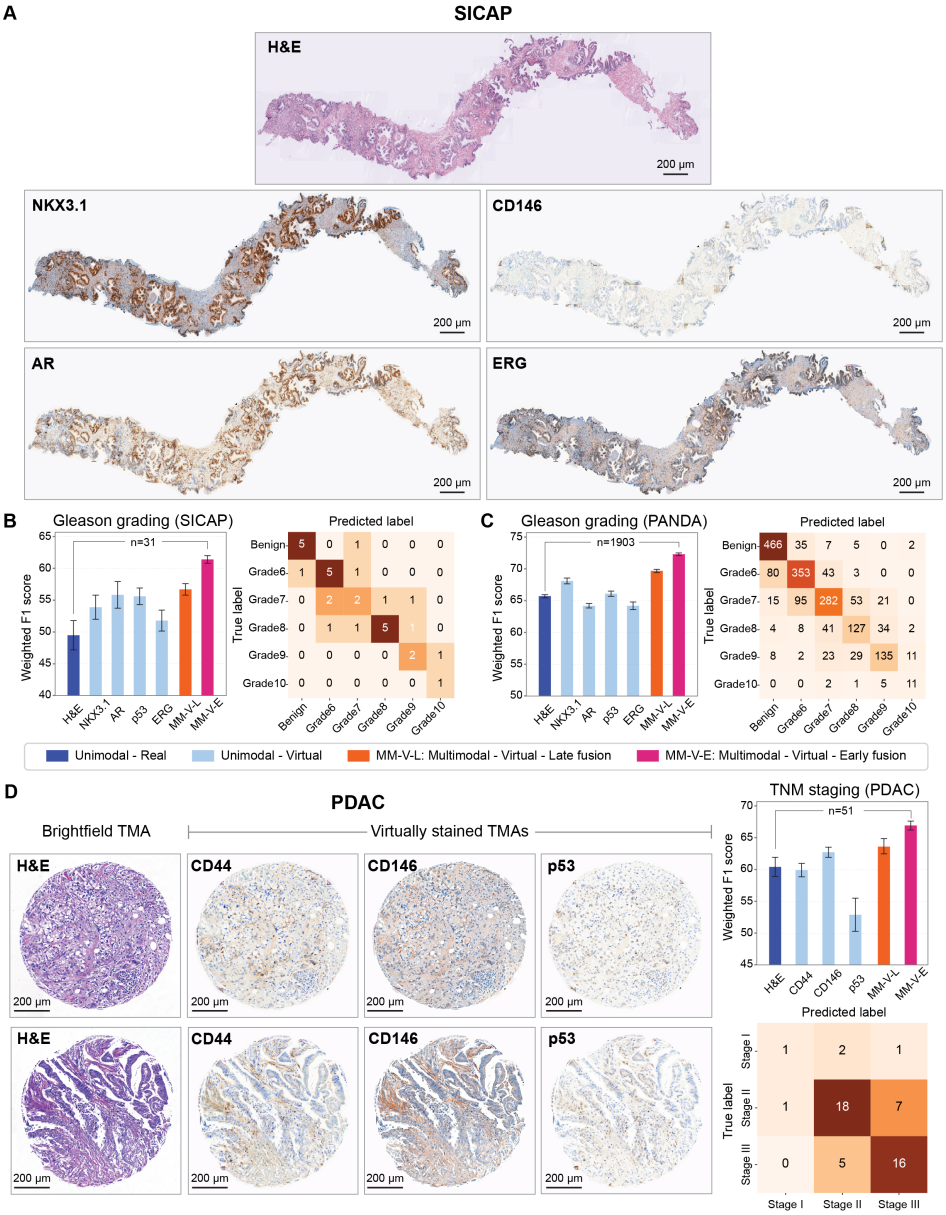

Fig. 7 Transfer learning across scales, cohorts, and cancer types. **(A)** Real H&E needle biopsy of the SICAP dataset (top) and matching virtual IHC stainings across 4 IHC markers (bottom), as generated from the EMPaCT-trained VirtualMultiplexer. **(B)** Prediction results of Gleason grading for the SICAP test set in terms of weighted F1 score and confusion matrix. Barplots and errorbars as in Figure 3, confusion matrices correspond to the Multimodal - Virtual Early fusion model. 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.

bioRxiv preprint doi: https://doi.org/10.1101/2023.11.29.568996; this version posted December 1, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

for CD146, and the heterogeneous staining of the glands for ERG. We also observed some recurring issues as in the quality assessment of the EMPaCT TMA (Figure 3), namely background (*e.g.*, stromal background occasionally present in NKX3.1 and ERG), border and tiling artifacts (*e.g.*, for CD146). Subsequently, we trained GT models under the previously described settings to predict the Gleason grade for both SICAP and PANDA datasets, shown in Figure 7B and C, respectively. We observe that the predictive performance of the unimodal models trained on virtual IHC stainings was close to or superior to the model using standalone H&E stainings for both the SICAP and PANDA datasets. Further improvement in Gleason score prediction was attained by the multimodal models built on the virtual IHC stainings. The early fusion model trained with the virtually multiplexed data achieved the best weighted F1 score of 61.4% and 72.3% for Gleason scoring on SICAP and PANDA, respectively, which significantly outperformed the H&E unimodal counterpart by 11.9% and 6.6% on SICAP and PANDA, respectively.

Finally, we evaluated the generalization ability of the VirtualMultiplexer to images of other cancer types. To this end, we first applied the pre-trained VirtualMultiplexer on the EMPaCT dataset to a PDAC TMA that was unseen by the model and used the available H&E images to generate virtual IHC stainings for CD44, CD146 and p53 (Figure 7D), 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 tumor, 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 pre-trained VirtualMultiplexer to generate virtual IHC stainings for CD44 and CD146 from colorectal [45] and breast cancer [46] H&E-stained WSIs from The Cancer Genome Atlas (TCGA) [47]. 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 (Appendix Figure A5).

The VirtualMultiplexer can greatly accelerate histopathology workflows

Lastly, we performed a runtime estimation of all components of the VirtualMultiplexer framework across imaging datasets of different scales, *i.e.*, TMAs, needle biopsies and WSIs (Appendix Figure A6). We calculated that applying the trained VirtualMultiplexer on a single EMPaCT TMA core (6000×6000 pixels at 20X magnification-0.24μm/pixel), an in-distribution sample, 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

bioRxiv preprint doi: https://doi.org/10.1101/2023.11.29.568996; this version posted December 1, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

normalization. However, the stain normalization step is crucial as it alleviates the appearance disparity between the training EMPaCT H&E TMAs and the out-of-distribution samples, 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 experiment 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 [48]. 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. We conclude that the VirtualMultiplexer leads to significant time gains when compared to a typical IHC staining, and can greatly accelerate histopathology workflows.

Discussion

Virtual staining has emerged as a promising direction in histopathology, with early attempts to apply GAN-based approaches on unpaired datasets across different stain-to-stain translation tasks. However, staining inconsistency issues limit its applications in translating marker-specific stains such as IHC. In this work, we proposed the VirtualMultiplexer, a generative toolkit that can effectively translate H&E to IHC images for several markers, and is able to preserve staining consistency across image scales without requiring access to consecutive tissue sections, image registration or extensive expert annotations. This was achieved by proposing a novel architecture that includes the joint optimization of multi-scale loss functions that encode different biological priors to ensure biological consistency on a cellular, neighborhood, and global, whole-image scale. Our results indicated that the VirtualMultiplexer consistently outperformed in image fidelity state-of-the-art S2S translation methods and generated IHC images that were indistinguishable from the real ones to the expert human eye. Detailed histopathological evaluation suggested that the staining quality of the generated images was on par or even exceeded that of the real images in terms of staining pattern and distribution, with staining artifacts (*e.g.*, high background, border effects) largely comparable in virtual versus real images for most markers. A thorough ablation study demonstrated that our novel local and global loss terms allowed us to mitigate staining unreliability and capture biologically consistent staining patterns, as opposed to solely using the adversarial and contrastive objectives employed in competing methods. When transferring the TMA-trained VirtualMultiplexer on prostate cancer WSIs and two unseen prostate cancer cohorts containing needle biopsies, we found that it generalized well to unseen prostate cancer images of different scales without any retraining or fine-tuning. Similar observations were made when generalizing

bioRxiv preprint doi: https://doi.org/10.1101/2023.11.29.568996; this version posted December 1, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

on tissues of different origins, namely a PDAC TMA and TCGA breast and colorectal cancer WSIs.

While our results demonstrate a clear potential for the use of virtual multiplexed staining in histopathology, several limitations remain and can be addressed in future extensions of the model. Firstly, in some cases we observed elevated background levels in the EMPaCT TMA. Although this issue was also present in the real data, it appeared more pronounced for markers that had an overall more faint staining pattern (*e.g.*, p53). More prominent background issues were present when we transferred the EMPaCT-trained model to the prostate cancer WSIs, which was expected considering that the EMPaCT IHC images and the prostate cancer WSIs were generated in two different institutions using different staining protocols and/or at different experiments and time points. Secondly, the patch-wise processing of the images induced tiling artifact in some cases, which is a well-known limitation of stain-to-stain translation approaches for large histopathology images [24, 39, 49]. In our algorithm, the tiling artifact was more pronounced at the border of a core, visible as patches of high staining intensity. One possible underlying cause of this effect is that, when the model receives as input an H&E patch at the border that contains very little tissue, the staining consistency losses "force" the model to stain the limited tissue with an overall higher intensity so as to match the staining distribution of the tissue-full patches that it has seen during training. Previous works [24, 39] have tried to address the tiling artifact, but it has been suggested that this caused less efficient translations [50]. As in our case the tiling artifact is observed in isolated, edge cases in the border, a straightforward solution would be to simply discard a narrow border surrounding the tissue, as empirically done in regular IHC when border artifacts are present. Thirdly, certain discrepancies in staining specificity were occasionally observed, such as failing to stain CD146+ vascular structures and glandular NKX3.1+ cells that invaded periglandular stroma. This can be partially attributed to the fact that these patterns were observed more rarely in the training images, and can be mitigated by ensuring the inclusion of adequate representative examples in the labeling of the IHC images of the training set.

Crucially, despite their limitations, the generated virtual images enabled the training of early fusion Graph-Transformer (GT) models, which consistently outperformed models trained on real data in the prediction on clinically relevant endpoints. This improvement was not only observed in the training TMA dataset across two prediction tasks, but also further confirmed on both independent prostate cancer cohorts, and the PDAC TMA cohort. In our experiments, we ensured that the multimodal early fusion GT models did not have an advantage in terms of number of samples used over the models trained with real data. At the same time, multimodal early fusion models had a much smaller parameter space in comparison to late fusion ones, as late fusion necessitated the training of as many GT models as the number of markers. This suggests that the higher model complexity of late fusion methods does not guarantee a performance improvement. A potential explanation of the observed performance improvement could be the quality of the generated images that are not affected by partially missing tumor areas or other artifacts occasionally found in real images. This explanation is corroborated by the fact that, for markers where virtual images were

bioRxiv preprint doi: https://doi.org/10.1101/2023.11.29.568996; this version posted December 1, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

of higher quality than real images, the corresponding unimodal - virtual models outperformed the unimodal - real ones, and vice versa. A more likely explanation could be that, as early fusion models were able to learn from the joint spatial distribution of several markers on the same tissue at once, they were able to pick up multimodal spatial relationships at the cellular level, mimicking data generated by advanced multiplexed imaging technologies. This explanation 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.

In conclusion, the current work establishes the potential of virtual multiplexed staining across images of different scales, patient cohorts and tissue types, with important implications towards AI-assisted histopathology. For example, the VirtualMultiplexer could be directly used for data inpainting, *i.e.*, filling out missing regions in a tissue image, or for sample imputation, *i.e.*, generating from scratch missing samples in large patient cohorts. As IHC marker panels are not always standardized across labs, filling out the gaps via virtual multiplexed staining could open the door towards harmonizing datasets within or across research labs, which is particularly important in cases of archival tissue samples with limited availability [51, 52]. As a result, the VirtualMultiplexer could enable the generation of comprehensive patient cohorts that could be used for clinically relevant predictions. An equally important application of our work is the use of virtual multiplexed staining for pre-histopathological experimental design: generating a large collection of IHC stains *in silico* and training AI models could support marker selection for actual experimentation, significantly reducing costs and preserving precious tissue. To reach its full potential, future work will be crucial to establish the use of the VirtualMultiplexer in real-world settings. From a technical standpoint, the generated virtual multiplexed stainings can enable the development of foundational models for IHC, as they have been successfully developed for brightfield H&E images [53–55]. However, developing such models requires high volumes of data, which is potentially challenging to acquire for IHC. Virtual stainings can be beneficial to this end and can pave the way for multimodal foundational tissue characterization. Interestingly, the virtual multiplexed stainings can also be exploited as biologically conditioned data augmentations to boost the development and in turn the predictive performance of foundational models in histopathology applications. From an applications standpoint, although here we presented a first proof-of-concept for transferring the model across tissue types, more thorough evaluations are needed to solidify our encouraging results. Finally, although here we focused on H&E-to-IHC translation, as our method is stain-agnostic, straight-forward adaptations of our work for S2S translation across cutting-edge multiplexed imaging technologies (*e.g.*, IMC, CODEX, MIBI) could significantly reduce their costs and find important applications in dataset harmonization or antibody panel selection and optimization. Our vision is that future extensions of our work could lead to the generation of an ever-growing and readily available dictionary of virtual stainers for IHC and beyond, that would be essentially limitless with respect to how many markers could be incorporated, surpassing in multiplexing ability even the most cutting-edge technologies and significantly accelerating spatial biology.

bioRxiv preprint doi: https://doi.org/10.1101/2023.11.29.568996; this version posted December 1, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Methods

Datasets

The VirtualMultiplexer was trained using the European Multicenter High Risk Prostate Cancer Clinical and Translational research group (EMPaCT) TMA dataset; an independent subset of EMPaCT was used for internal testing. The VirtualMultiplexer was further evaluated in a zero-shot fashion, *i.e.*, without any retraining or fine-tuning, on three external prostate cancer datasets, namely prostate cancer WSIs, SICAP [43] and PANDA [44] needle biopsies, on an independent PDAC dataset (PDAC TMAs) and on TCGA data from breast and colorectal cancer. In all cases, independent Graph-Transformers 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 the European Multicenter High Risk Prostate Cancer Clinical and Translational research group (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 NK3 Homeobox 1 (NKX3.1) and Androgen receptor (AR), tumor markers Tumor Protein p53 (p53) and Erythroblast transformation-specific related gene (ERG) and membrane markers Cell Surface Glycoprotein (CD44) and Melanoma Cell Adhesion molecule (CD146/MCAM). 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 minute room temperature incubation in 1% BSA in PBS–0.1%Tween 20. The following antibodies were used: anti-AR (Dako Agilent, M3562, 1:100 dilution), anti-NKX3.1 (Athena Enzyme Systems, 314, 1:200), anti-p53 (Dako Agilent, M7001, 1:800), anti-CD44 (Abcam, ab16728, 1:2000), anti-ERG (Abcam, ab133264, 1:500) and anti-CD146 (Abcam, ab75769 EPR3208, 1:500). Images were acquired using a 3D Histech Panoramic Flash II 250 scanner at 20$\times$ magnification (resolution 0.24μm/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 at 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 Table 1.

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 minutes in 3% H2O2, followed by 30 minute room temperature incubation in 1% BSA in PBS–0.1%Tween 20. The following

bioRxiv preprint doi: https://doi.org/10.1101/2023.11.29.568996; this version posted December 1, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Overall survival status prediction		
#cores 0: Alive/censored	#cores 1: Prostate-cancer related death	#total cores
472	205	677
Disease progression prediction		
#cores 0: No recurrence	#cores 1: Recurrence	#total cores
568	109	677

Table 1 Distribution of cores per clinical endpoint for the EMPaCT dataset.

primary antibodies were used: anti-CD146 (Abcam, ab75769 EPR3208, 1:500), anti-AR (Abcam, ab133273, EPR1535, 1:100) and anti-NKX3.1 (Cell Signaling, 83700T, 1:200). Secondary anti-rabbit antibody Envision HRP (DAKO, Agilent Technologies, Basel, Switzerland) for 30 minutes was used and signal detection was done using AEC substrate (DAKO, Agilent Technologies, Basel, Switzerland). 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 0.24μm/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 [43]. The WSIs were reconstructed by stitching the patches. The WSIs were scanned at 40× magnification by 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 [56], and 5,662 H&E stained needle biopsies from 1,222 patients at various hospitals in Stockholm, Sweden [57]. The slides from Radboud were scanned with a 3D Histech Panoramic Flash II 250 scanner at 20× magnification (resolution 0.24μm/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 0.45202μm and 0.5032μm, respectively. The Gleason grades of the biopsies were annotated by expert uro-pathologists and were released as part of the Prostate cANcer graDe Assessment (PANDA) challenge [44]. We removed the noisy and inconspicuously labeled 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 are shown in Table 2.

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 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

bioRxiv preprint doi: https://doi.org/10.1101/2023.11.29.568996; this version posted December 1, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

	Benign 0+0	Grade 6 3+3	Grade 7 3+4/4+3	Grade 8 3+5/4+4/5+3	Grade 9 4+5/5+4	Grade 10 5+5	Total
SICAP	36	31	29	36	16	7	155
	Benign 0+0	Grade 6 3+3	Grade 7 3+4/4+3	Grade 8 3+5/4+4/5+3	Grade 9 4+5/5+4	Grade 10 5+5	Total
PANDA	2603	2397	2327	1124	996	105	9552

Table 2 Distribution of WSIs across Gleason grades for both SICAP and PANDA datasets.

contained three spots from each case (tumor front, tumor center, tumor 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 Tumour, Node, and Metastasis (TNM) stage during a master thesis of student Jessica Lisa Rohrbach at ITMP, supervised by Martin Wartenberg. All cases were UICC tumor stage I, stage II or stage III cases on pathologic examination, according to the Union for International Cancer Control (UICC) TNM Classification of Malignant Tumours, 8th edition [58]; the TMA did not comprise of UICC tumor stage IV cases. In all our analysis including the TNM prediction (Figure 7D), we excluded the thirteen neoadjuvant cases and considered only the 104 cases that received adjuvant therapy. The distribution of cores across the three TNM stages is reported in Table 3.

Stage I	Stage II	Stage III	Total
24	121	115	260

Table 3 Distribution of cores per TNM stage for the PDAC dataset.

TCGA: The dataset includes example H&E WSIs from breast (BRCA) and colorectal cancer (CRC) from The Cancer Genome Atlas (TCGA), available at the GDC data portal (https://portal.gdc.cancer.gov).

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 [59]. 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

bioRxiv preprint doi: https://doi.org/10.1101/2023.11.29.568996; this version posted December 1, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

validation of the VirtualMultiplexer. The extracted H&E patches of all other unseen datasets (prostate cancer WSIs, SICAP, PANDA, PDAC, TCGA) were additionally first stain-normalized to mitigate the staining appearance variability with respect to the EMPaCT TMAs [60].

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 Figure 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 $X_p = \{x_p \in X_{\text{img}}\}$ and $Y_p = \{y_p \in Y_{\text{img}}\}$ 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 Contrastive Unpaired Translation (CUT) [33] 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, *i.e.*, $y'_p = G(x_p) = G_{\text{dec}}(G_{\text{enc}}(x_p))$. Finally, 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, namely local, neighborhood and global consistency (Figure 2B). The neighborhood consistency enforces effective staining translation at a patch-level, where a patch captures the neighborhood 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.

Neighborhood consistency: The neighborhood objective is a combination of an adversarial loss and a patch-wise multilayer contrastive loss, implemented as previously described in CUT [33] (Figure 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 patch-level [61]. The adversarial loss is a standard GAN min-max loss [34], 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 (Figure 2B, panel 1a). It is calculated as follows:

$$\mathcal{L}_{\text{adv}}(G, D, X_p, Y_p) = \mathbb{E}_{y_p \sim Y_p} \log D(y_p) + \mathbb{E}_{x_p \sim X_p} \log(1 - D(G(x_p))). \quad (1)$$

The patch-wise multilayer contrastive loss follows a noise contrastive estimation (NCE) concept as presented in [61, 62] and reused in [29, 33]. Specifically, it aims to maximize the resemblance between input H&E patch $x_p \in X_p$ and corresponding virtually synthesized IHC patch $y'_p \in Y'_p$ (Figure 2B, panel 1b). We first extract a *query* sub-patch y'_{sp} of size 64×64 from the target IHC domain patch y'_p (purple square

bioRxiv preprint doi: https://doi.org/10.1101/2023.11.29.568996; this version posted December 1, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

in panel 1b, Figure 2B) and match it to the corresponding sub-patch x_{sp}, *i.e.*, a sub-patch at the same spatial location as y'_{sp} but from the H&E source domain patch x_p (black square in panel 1b, Figure 2B). Since both sub-patches originate from the exact same tissue neighborhood, we expect that x_{sp} and y'_{sp} form a positive pair. We also sample N sub-patches $\{x^-_{sp}\}$ at different spatial locations from x_p (red squares in panel 1b, Figure 2B), 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 $\{x^-_{sp}\}$ to a d-dimensional embedding space $\mathbb{R}^d$ via G_{enc} and project them to a unit sphere, resulting in v, v^+ and $\{v^- \in \mathbb{R}^d$, 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:

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

Here, we use a variation of the loss in Eq. (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, 33]. 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 $\{z_l\}_L = \{H_l(G^l_{\text{enc}}(x_p))\}_L$ and $\{z'_l\}_L = \{H_l(G^l_{\text{enc}}(y'_p))\}_L = \{H_l(G^l_{\text{enc}}(G(x_p))\}_L$, $\forall l \in \{1, 2, ..., L\}$, 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 sub-patches, respectively (purple, black and red boxes in Figure 2B, panel 1b). The final patch-wise multilayer contrastive loss is computed as:

$$\mathcal{L}_{\text{contrastive}}(G, H, X_p) = \mathbb{E}_{x_p \sim X_p} \sum_{l=1}^{L} \sum_{s=1}^{S_l} \mathcal{L}(z'_{l,s}, z_{l,s}, z_{l,S_l \backslash s}) \tag{3}$$

We also employ contrastive loss $\mathcal{L}_{\text{contrastive}}(G, H, Y_p)$ on patches $y_p \in Y_p$, a domain-specific version of the identity loss [63, 64] that prevents the generator G from making unnecessary changes as proposed in [33]. Finally, the overall neighborhood consistency objective is computed as a weighted sum of the adversarial loss (1) and the multilayer contrastive loss (3) with regularization hyperparameter λ_{NCE}:

$$\mathcal{L}_{\text{neighborhood}} = \mathcal{L}_{\text{adv}}(G, D, X_p, Y_p) + \lambda_{\text{NCE}} \times \big(\mathcal{L}_{\text{contrastive}}(G, H, X_p) + \mathcal{L}_{\text{contrastive}}(G, H, Y_p)\big) \tag{4}$$

Global consistency: Inspired by seminal work in neural style transfer [65], 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 both in terms of tissue composition and staining pattern at the image (tile) level (Figure 2B, panel 2). Since 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

bioRxiv preprint doi: https://doi.org/10.1101/2023.11.29.568996; this version posted December 1, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

1024×1024 are extracted (boxes in Figure 2B, panel 2), denoted as $X_t = \{x_t \in X_{\text{img}}\}$ and $Y'_t = \{y'_t \in Y'_{\text{img}}\}$, respectively. Then, the tiles are encoded by a pretrained feature extractor F, specifically VGG16 [35] pretrained on ImageNet [66]. The tile-level content loss at layer l of F is calculated as:

$$\mathcal{L}^l_{\text{content}}(F, X_p, Y'_p) = \frac{\sum ||F^l(x_t) - F^l(y'_t)||^2}{h^l \cdot w^l \cdot c^l} \tag{5}$$

where, h, w and c are height, width, and channel dimensions of the feature map at l^{th} 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. Since 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 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:

$$\mathcal{L}^l_{\text{style}}(F, Y_p, Y'_p) = \frac{\sum ||\mathcal{G} \circ F^l(y_t) - \mathcal{G} \circ F^l(y'_t)||^2}{||\mathcal{G} \circ F^l(y_t)||^2 + ||\mathcal{G} \circ F^l(y'_t)||^2} \tag{6}$$

where $\mathcal{G}$ is the Gram matrix that measures the correlation between all the style in a feature map. The denominator is a normalization term which compensates for the under- or over-stylization of the tiles in a batch [67]. The overall global consistency loss is computed as:

$$\mathcal{L}_{\text{global}} = \lambda_{\text{content}} \times \sum_{l}^{L_{\text{content}}} \mathcal{L}^l_{\text{content}}(F, X_p, Y'_p) + \lambda_{\text{style}} \times \sum_{l}^{L_{\text{style}}} \mathcal{L}^l_{\text{style}}(F, Y_p, Y'_p) \tag{7}$$

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, namely a cell discriminator loss ($\mathcal{L}_{\text{cellDisc}}$) and a cell classification loss ($\mathcal{L}_{\text{cellClass}}$) (Figure 2B, panel 3). The cell discriminator loss is inspired by [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 Eq.(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_p} and $M_{Y'_p}$, respectively), which include bounding box demarcation around the cells (Figure 2B, panel 3). D_{cell} comprises a feature extractor followed by a RoIAlign layer [68], and a final discriminator. The goal of D_{cell} is to output, $D_{\text{cell}}(Y_p, M_{Y_p}) \rightarrow \mathbf{1}$ and $D_{\text{cell}}(Y'_p, M_{X_p}) \rightarrow \mathbf{0}$, where $\mathbf{1}$ and $\mathbf{0}$ indicate real and virtual cells (indicated in black and purple, respectively, in Figure 2B, panel 3). The cell discriminator loss is defined as:

bioRxiv preprint doi: https://doi.org/10.1101/2023.11.29.568996; this version posted December 1, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

$$\mathcal{L}_{\text{cellDisc}}(D_{\text{cell}}, Y_p, Y'_p, M_{Y_p}, M_{Y'_p}) = \frac{1}{2}\,\mathbb{E}_{y_p \in Y_p}\big(D_{\text{cell}}(y_p, M_{y_p}) - 1\big)^2 + \frac{1}{2}\,\mathbb{E}_{y'_p \in Y'_p}\big(D_{\text{cell}}(y'_p, M_{y'_p})\big)^2 \quad (8)$$

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 $C_{Y'_p}$ and C_{Y_p}, *i.e.*, binary variables depicting the positive or negative staining status of a cell (indicated as 1: yellow and 0: blue boxes in Figure 2B, panel 3). The computation of cell masks and labels is described in detail in section **Cell masking and labeling of IHC images**. The cell-level classification loss can be easily computed as cross-entropy loss, calculated as:

$$\mathcal{L}_{\text{cellClass}}(F_{\text{cell}}, Y_p, Y'_p, M_{Y_p}, M_{Y'_p}, C_{Y_p}, C_{Y'_p}) = \frac{-1}{|C_{y_p}|}\sum_{\substack{j=1 \\ i\in\{0,1\}}}^{|C_{y_p}|} \mathbb{1}_{C^j_{y_p}=i} \times \log\big(p(M^j_{y_p} \times y_p)\big) + \frac{-1}{|C_{y'_p}|}\sum_{\substack{j=1 \\ i\in\{0,1\}}}^{|C_{y'_p}|} \mathbb{1}_{C^j_{y'_p}=i} \times \log\big(p(M^j_{y'_p} \times y'_p)\big) \quad (9)$$

where, $|C_{y_p}|$ and $|C_{y'_p}|$ 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:

$$\mathcal{L}_{\text{local}} = \lambda_{\text{cellDisc}} \times \mathcal{L}_{\text{cellDisc}}(D_{\text{cell}}, Y_p, Y'_p, M_{Y_p}, M_{Y'_p}) + \lambda_{\text{cellClass}} \times \mathcal{L}_{\text{cellClass}}(F_{\text{cell}}, Y_p, Y'_p, M_{Y_p}, M_{Y'_p}, C_{Y_p}, C_{Y'_p}) \quad (10)$$

where $\lambda_{\text{cellDisc}}$ and $\lambda_{\text{cellClass}}$ are the regularization hyperparameteres 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 (*e.g.*, 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,

$$\mathcal{L}_{\text{VirtualMultiplexer}} = \mathcal{L}_{\text{neighborhood}} + \mathcal{L}_{\text{global}} + \mathcal{L}_{\text{local}} \quad (11)$$

Cell masking and labeling of IHC images: As already discussed, the local consistency loss of Eq.11 needs as input cell masks M_{X_p}, M_{Y_p} and cell labels C_{X_p}, C_{Y_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 (*e.g.*, HoVerNet [69]) 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, *i.e.*, its ability to synthesize

bioRxiv preprint doi: https://doi.org/10.1101/2023.11.29.568996; this version posted December 1, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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 $\rightarrow$ H&E translation. The VirtualMultiplexer$_{\text{IHC}\rightarrow\text{H\&E}}$ is trained using neighborhood 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 [69] 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. Since 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, *i.e.*, $D_{\text{cell}}^{\text{source}}$, and $D_{\text{cell}}^{\text{target}}$, respectively. Provided the annotations, $D_{\text{cell}}^{\text{target}}$ is trained as a CNN patch-classifier. The classifier predictions on Y_{img} combined with M_{Y_p} results in C_{Y_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 $D_{\text{cell}}^{\text{source}}$ as a CNN patch-classifier. The classifier predictions on X_{img} combined with M_{X_p} results in C_{X_p}.

Implementation and training details: The architectural choices of the VirtualMultiplexer were set as follows: G is a ResNet [42] with 9 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 2-layer MLP. We use Xavier weight initialization [70], instance normalization [71], and a batch size of 1 image. We use Least Square GAN loss [72] for $\mathcal{L}_{\text{adv}}$. The model hyperparameters for the loss terms of the VirtualMultiplexer are set as: 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 Adam optimizer [73] with momentum parameters $\beta_1 = 0.5$ and $\beta_2 = 0.999$. Different learning rates lr are employed for different consistency objectives, *i.e.*, for neighborhood consistency lr_G and lr_D is 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 8, 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 8.

Graph-Transformer Architecture

The Graph-Transformer (GT) architecture, proposed by [41], fuses a Graph Neural Network and a vision transformer (ViT) to process histopathology images. The GNN first operates on a graph-structured representation of a histopathology image, where the nodes and edges of the graph denote patches and inter-patch spatial connectivity,

bioRxiv preprint doi: https://doi.org/10.1101/2023.11.29.568996; this version posted December 1, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

and the nodes encode patch features extracted from a pre-trained deep learning model. Specifically, GT employs a graph convolution layer [74] to learn contextualized node embeddings through propagating and aggregating neighborhood node information. Subsequently, a ViT layer operates on the contextualized node features, leverages self-attention to weigh the significance 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. To note, 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 giga-pixel 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 [75], which reduces the number of nodes into a fixed number of tokens while preserving the local neighborhood information of the nodes.

Implementation and training details: The architecture of GT follows the official implementation on github[1]. 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 ¿10% were discarded. The patches were encoded using ResNet50 [42] model pretrained on ImageNet dataset [66]. A graph representation was constructed using the patches with a 8-node connectivity pattern. The GT network consisted of one graph convolutional layer, and set the ViT layer configurations as, number of ViT blocks = 3, MLP size = 128, embedding dimension of each patch = 32, and number of multi-head attention = 8. The model hyperparameters were set as: 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 8. The GT models were trained for 400 epochs with early stopping.

Method Evaluation

Patch-level evaluation: We use the FID score [76] to compare the distribution of the virtual IHC patches with the distribution of the real IHC patches, as shown in Figure 3. The computation begins with projecting the virtual and the real IHC patches to an embedding space using InceptionV3 [76] model, pretrained on ImageNet [66]. 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:

$$\text{FID} = ||\mu_r - \mu_v||^2 + Tr\big(\Sigma_r + \Sigma_v - 2(\Sigma_r\Sigma_v)^{\frac{1}{2}}\big) \tag{12}$$

where, μ_r and μ_v are the feature-wise mean of the real and virtual patches, Σ_r and Σ_v are co-variance 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, 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 errorbar in Figure 3). We used a 70%-30% ratio to split the data in 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 Table 4:

[1] A Graph-Transformer for Whole Slide Image Classification: https://github.com/vkola-lab/tmi2022

bioRxiv preprint doi: https://doi.org/10.1101/2023.11.29.568996; this version posted December 1, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

	NKX3.1	AR	CD44	CD146	p53	ERG
Train	248	359	426	390	389	251
Test	114	153	183	159	178	122
Total	362	512	609	549	567	373

Table 4 Distribution of cores per marker for training and testing the VirtualMultiplexer for the EMPaCT dataset.

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 8. Subsequently, Graph-Transformers [41] were trained under uni-modal and multi-modal settings using both real and virtually stained images, and evaluated on held-out independent test dataset. The final classification scores were reported using a weighted F1-metric, where a higher score depicts a better classification performance, thereby a 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 errorbars in Figures 6 and 7). In all cases, we used a 60%-20%-20% ratio to split the data in 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 Table 5.

Task: Overall survival status prediction								
	NKX3.1	AR	CD44	CD146	p53	ERG	H&E	Multimodal*
Train	232	333	397	343	377	230	407	407
Validation	74	110	133	123	120	82	134	134
Test	85	120	141	131	126	84	136	136
Task: Disease progression prediction								
	NKX3.1	AR	CD44	CD146	p53	ERG	H&E	Multimodal*
Train	223	341	405	348	384	225	408	408
Validation	87	107	127	121	121	86	135	135
Test	81	115	139	128	128	85	134	134

Table 5 Distribution of cores for training, validation, and testing of Graph-Transformers per downstream task on the EMPaCT dataset, for all unimodal (first 7 columns) and multimodal models. * The number of cores reported under multimodal setting refers to the union across all markers used.

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 Figure 7 and are reported in Table 6.

bioRxiv preprint doi: https://doi.org/10.1101/2023.11.29.568996; this version posted December 1, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

	SICAP	PANDA	PDAC
Training	95	5736	161
Validation	29	1913	48
Test	31	1903	51
Total	155	9552	260

Table 6 Number of available training, validation and test samples for the SICAP, PANDA and PDAC datasets.

Computational Hardware and Software

The image datasets were preprocessed on POWER9 CPUs (Central Processing Units) and one NVIDIA Tesla A100 GPU (Graphics Processing Unit) using the Histocartography library [59]. The deep learning models were trained on NVIDIA Tesla P100 GPUs using PyTorch (version 1.13.1) [77] and PyTorch Geometric (version 2.3.0) [78]. The entire pipeline was implemented in Python (version 3.9.1).

Data Availability

The main dataset used to support this study (EMPaCT) has been deposited in Zenodo, together with the prostate cancer WSIs (doi: 10.5281/zenodo.10066853). The SICAP dataset 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 (BRCA and CRC) are available 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 patient cohorts cannot be shared owing to patient-confidentiality obligations.

Author contributions

P.P. conceived and implemented the model. P.P., A.M. and M.A.R. designed and performed computational analyses. S.K., F.B. and M.R. performed the experiments. S.K., F.B., E.C., M.W. and M.K.d.J. assessed the virtual stainings. P.P., S.K., F.B., A.M. and M.A.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.A.R. wrote the paper with inputs from all authors. M.K.d.J. and M.A.R. were responsible for the overall planning and supervision of the project.

Code Availability

All source code of the VirtualMultiplexer is available under an open-source license in https://github.com/AI4SCR/VirtualMultiplexer.

bioRxiv preprint doi: https://doi.org/10.1101/2023.11.29.568996; this version posted December 1, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Acknowledgments

We would like to thank Guillaume Jaume, Jannis Born and Mara Graziani for constructive comments, discussions and suggestions. The results published here are in part based upon data generated by the TCGA Research Network: https://www.cancer.gov/tcga. This work was supported by the Swiss National Science Foundation (SNSF) Sinergia Grant (CRSII5_202297) to M.A.R. and M.K.dJ. The PDAC TMA construction has taken place at the Translational Research Unit (TRU) Platform of ITMP (https://www.ngtma.com/) in the setting of a grant by the Foundation for Clinical-Experimental Cancer Research Bern to M.W.

Competing interests

The authors declare no competing interests.

bioRxiv preprint doi: https://doi.org/10.1101/2023.11.29.568996; this version posted December 1, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

References

[1] Aditya Kashyap, Maria Anna Rapsomaniki, Vesna Barros, Anna Fomitcheva-Khartchenko, Adriano Luca Martinelli, Antonio Foncubierta Rodriguez, Maria Gabrani, Michal Rosen-Zvi, and Govind Kaigala. Quantification of tumor heterogeneity: from data acquisition to metric generation. *Trends in Biotechnology*, 40(6):647–676, 2022.

[2] John KC Chan. The wonderful colors of the hematoxylin–eosin stain in diagnostic surgical pathology. *International journal of surgical pathology*, 22(1):12–32, 2014.

[3] Leandro Luongo De Matos, Damila Cristina Trufelli, Maria Graciela Luongo De Matos, and Maria Aparecida da Silva Pinhal. Immunohistochemistry as an important tool in biomarkers detection and clinical practice. *Biomarker insights*, 5:BMI–S2185, 2010.

[4] Charlotte Giesen, Hao AO Wang, Denis Schapiro, Nevena Zivanovic, Andrea Jacobs, Bodo Hattendorf, Peter J Schüffler, Daniel Grolimund, Joachim M Buhmann, Simone Brandt, et al. Highly multiplexed imaging of tumor tissues with subcellular resolution by mass cytometry. *Nature methods*, 11(4):417–422, 2014.

[5] Yury Goltsev, Nikolay Samusik, Julia Kennedy-Darling, Salil Bhate, Matthew Hale, Gustavo Vazquez, Sarah Black, and Garry P Nolan. Deep profiling of mouse splenic architecture with codex multiplexed imaging. *Cell*, 174(4):968–981, 2018.

[6] Michael Angelo, Sean C Bendall, Rachel Finck, Matthew B Hale, Chuck Hitzman, Alexander D Borowsky, Richard M Levenson, John B Lowe, Scot D Liu, Shuchun Zhao, et al. Multiplexed ion beam imaging of human breast tumors. *Nature medicine*, 20(4):436–442, 2014.

[7] Sabrina M Lewis, Marie-Liesse Asselin-Labat, Quan Nguyen, Jean Berthelet, Xiao Tan, Verena C Wimmer, Delphine Merino, Kelly L Rogers, and Shalin H Naik. Spatial omics and multiplexed imaging to explore cancer biology. *Nature methods*, 18(9):997–1012, 2021.

[8] Nir Pillar and Aydogan Ozcan. Virtual tissue staining in pathology using machine learning. *Expert Review of Molecular Diagnostics*, 22(11):987–989, 2022.

[9] Bijie Bai, Xilin Yang, Yuzhu Li, Yijie Zhang, Nir Pillar, and Aydogan Ozcan. Deep learning-enabled virtual histological staining of biological samples. *Light: Science & Applications*, 12(1):57, 2023.

[10] Maximilian E Tschuchnig, Gertie J Oostingh, and Michael Gadermayr. Generative adversarial networks in digital pathology: a survey on trends and future potential. *Patterns*, 1(6):100089, 2020.

bioRxiv preprint doi: https://doi.org/10.1101/2023.11.29.568996; this version posted December 1, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

[11] Laya Jose, Sidong Liu, Carlo Russo, Annemarie Nadort, and Antonio Di Ieva. Generative Adversarial Networks in Digital Pathology and Histopathological Image Processing: A Review. *Journal of Pathology Informatics*, 12(1):43, 2021.

[12] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.

[13] Jingxi Li, Jason Garfinkel, Xiaoran Zhang, Di Wu, Yijie Zhang, Kevin De Haan, Hongda Wang, Tairan Liu, Bijie Bai, Yair Rivenson, et al. Biopsy-free in vivo virtual histology of skin using deep learning. *Light: Science & Applications*, 10(1):233, 2021.

[14] Yair Rivenson, Tairan Liu, Zhensong Wei, Yibo Zhang, Kevin de Haan, and Aydogan Ozcan. Phasestain: the digital staining of label-free quantitative phase microscopy images using deep learning. *Light: Science & Applications*, 8(1):23, 2019.

[15] Yair Rivenson, Hongda Wang, Zhensong Wei, Kevin de Haan, Yibo Zhang, Yichen Wu, Harun Günaydın, Jonathan E Zuckerman, Thomas Chong, Anthony E Sisk, et al. Virtual histological staining of unlabelled tissue-autofluorescence images via deep learning. *Nature biomedical engineering*, 3(6):466–477, 2019.

[16] Aman Rana, Alarice Lowe, Marie Lithgow, Katharine Horback, Tyler Janovitz, Annacarolina Da Silva, Harrison Tsai, Vignesh Shanmugam, Akram Bayat, and Pratik Shah. Use of deep learning to develop and analyze computational hematoxylin and eosin staining of prostate core biopsy images for tumor diagnosis. *JAMA network open*, 3(5):e205111–e205111, 2020.

[17] Kevin de Haan, Yijie Zhang, Jonathan E Zuckerman, Tairan Liu, Anthony E Sisk, Miguel FP Diaz, Kuang-Yu Jen, Alexander Nobori, Sofia Liou, Sarah Zhang, et al. Deep learning-based transformation of h&e stained tissues into special stains. *Nature communications*, 12(1):4884, 2021.

[18] Yijie Zhang, Kevin de Haan, Yair Rivenson, Jingxi Li, Apostolos Delis, and Aydogan Ozcan. Digital synthesis of histological stains using micro-structured and multiplexed virtual staining of label-free tissue. *Light: Science & Applications*, 9(1):78, 2020.

[19] Shengjie Liu, Chuang Zhu, Feng Xu, Xinyu Jia, Zhongyue Shi, and Mulan Jin. Bci: Breast cancer immunohistochemical image generation through pyramid pix2pix. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1815–1824, 2022.

[20] Weisi Xie, Nicholas P Reder, Can Koyuncu, Patrick Leo, Sarah Hawley, Hongyi Huang, Chenyi Mao, Nadia Postupna, Soyoung Kang, Robert Serafin, et al. Prostate cancer risk stratification via non-destructive 3d pathology with deep

bioRxiv preprint doi: https://doi.org/10.1101/2023.11.29.568996; this version posted December 1, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

learning-assisted gland analysis. *Cancer research*, 82(2):334, 2022.

[21] Parmida Ghahremani, Yanyun Li, Arie Kaufman, Rami Vanguri, Noah Greenwald, Michael Angelo, Travis J Hollmann, and Saad Nadeem. Deep learning-inferred multiplex immunofluorescence for immunohistochemical image quantification. *Nature machine intelligence*, 4(4):401–412, 2022.

[22] Ranran Zhang, Yankun Cao, Yujun Li, Zhi Liu, Jianye Wang, Jiahuan He, Chenyang Zhang, Xiaoyu Sui, Pengfei Zhang, Lizhen Cui, et al. Mvfstain: Multiple virtual functional stain histopathology images generation based on specific domain mapping. *Medical Image Analysis*, 80:102520, 2022.

[23] Caner Mercan, GCAM Mooij, David Tellez, Johannes Lotz, Nick Weiss, Marcel van Gerven, and Francesco Ciompi. Virtual staining for mitosis detection in breast histopathology. In *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, pages 1770–1774. IEEE, 2020.

[24] Amal Lahiani, Irina Klaman, Nassir Navab, Shadi Albarqouni, and Eldad Klaiman. Seamless virtual whole slide image synthesis and validation using perceptual embedding consistency. *IEEE Journal of Biomedical and Health Informatics*, 25(2):403–411, 2020.

[25] Shuting Liu, Baochang Zhang, Yiqing Liu, Anjia Han, Huijuan Shi, Tian Guan, and Yonghong He. Unpaired stain transfer using pathology-consistent constrained generative adversarial networks. *IEEE transactions on medical imaging*, 40(8):1977–1989, 2021.

[26] Joseph Boyd, Irène Villa, Marie-Christine Mathieu, Eric Deutsch, Nikos Paragios, Maria Vakalopoulou, and Stergios Christodoulidis. Region-guided cyclegans for stain transfer in whole slide images. In *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pages 356–365. Springer, 2022.

[27] Yiyang Lin, Bowei Zeng, Yifeng Wang, Yang Chen, Zijie Fang, Jian Zhang, Xiangyang Ji, Haoqian Wang, and Yongbing Zhang. Unpaired multi-domain stain transfer for kidney histopathological images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 1630–1637, 2022.

[28] Nassim Bouteldja, Barbara M Klinkhammer, Tarek Schlaich, Peter Boor, and Dorit Merhof. Improving unsupervised stain-to-stain translation using self-supervision and meta-learning. *Journal of Pathology Informatics*, 13:100107, 2022.

[29] Kutsev Bengisu Ozyoruk, Sermet Can, Berkan Darbaz, Kayhan Başak, Derya Demir, Guliz Irem Gokceler, Gurdeniz Serin, Uguray Payam Hacisalihoglu, Emirhan Kurtuluş, Ming Y Lu, et al. A deep-learning model for transforming the style of tissue images from cryosectioned to formalin-fixed and paraffin-embedded. *Nature Biomedical Engineering*, pages 1–13, 2022.

bioRxiv preprint doi: https://doi.org/10.1101/2023.11.29.568996; this version posted December 1, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

[30] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.

[31] Bowei Zeng, Yiyang Lin, Yifeng Wang, Yang Chen, Jiuyang Dong, Xi Li, and Yongbing Zhang. Semi-supervised pr virtual staining for breast histopathological images. In *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pages 232–241, 2022.

[32] Ali Borji. Pros and cons of gan evaluation measures: New developments. *Computer Vision and Image Understanding*, 215:103329, 2022.

[33] Taesung Park, Alexei A Efros, Richard Zhang, and Jun-Yan Zhu. Contrastive learning for unpaired image-to-image translation. In *ECCV*, pages 319–345, 2020.

[34] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[35] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations (ICLR*, 2014.

[36] Alberto Briganti, Steven Joniau, Paolo Gontero, Firas Abdollah, Niccolò M Passoni, Bertrand Tombal, Giansilvio Marchioro, Burkhard Kneitz, Jochen Walz, Detlef Frohneberg, et al. Identifying the best candidate for radical prostatectomy among patients with high-risk prostate cancer. *European urology*, 61(3):584–592, 2012.

[37] Burkhard Kneitz, Markus Krebs, Charis Kalogirou, Maria Schubert, Steven Joniau, Hein van Poppel, Evelyne Lerut, Susanne Kneitz, Claus Jürgen Scholz, Philipp Ströbel, 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 research*, 74(9):2591–2603, 2014.

[38] Lorenzo Tosco, Annouschka Laenen, Alberto Briganti, Paolo Gontero, R Jeffrey Karnes, Patrick J Bastian, Piotr Chlosta, Frank Claessens, Felix K Chun, Wouter Everaerts, et al. The empact classifier: a validated tool to predict postoperative prostate cancer-related death using competing-risk analysis. *European urology focus*, 4(3):369–375, 2018.

[39] Ming-Yang Ho, Min-Sheng Wu, and Che-Ming Wu. Ultra-high-resolution unpaired stain transformation via kernelized instance normalization. In *European Conference on Computer Vision*, pages 490–505. Springer, 2022.

bioRxiv preprint doi: https://doi.org/10.1101/2023.11.29.568996; this version posted December 1, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

[40] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

[41] Yi Zheng, Rushin H Gindra, Emily J Green, Eric J Burks, Margrit Betke, Jennifer E Beane, and Vijaya B Kolachalama. A graph-transformer for whole slide image classification. *IEEE transactions on medical imaging*, 41(11):3003–3015, 2022.

[42] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[43] Julio Silva-Rodríguez, Adrián Colomer, María Sales, Rafael Molina, and Valery Naranjo. Going deeper through the gleason scoring scale: An automatic end-to-end system for histology prostate grading and cribriform pattern detection. *Computer Methods and Programs in Biomedicine*, 195:105637, 2020.

[44] Wouter Bulten, Kimmo Kartasalo, Po-Hsuan Cameron Chen, Peter Ström, Hans Pinckaers, Kunal Nagpal, Yuannan Cai, David F Steiner, Hester van Boven, Robert Vink, et al. Artificial intelligence for diagnosis and gleason grading of prostate cancer: the panda challenge. *Nature medicine*, 28(1):154–163, 2022.

[45] Cancer Genome Atlas Network et al. Comprehensive molecular characterization of human colon and rectal cancer. *Nature*, 487(7407):330, 2012.

[46] Comprehensive molecular portraits of human breast tumours. *Nature*, 490(7418):61–70, 2012.

[47] The cancer genome atlas (tcga) research network. https://www.cancer.gov/tcga.

[48] Annika Blank, Heather Dawson, Caroline Hammer, Aurel Perren, and Alessandro Lugli. Lean management in the pathology laboratory. *Der Pathologe*, 38:540–544, 2017.

[49] Thomas de Bel, Meyke Hermsen, Jesper Kers, Jeroen van der Laak, and Geert Litjens. Stain-transforming cycle-consistent generative adversarial networks for improved segmentation of renal histopathology. In *Medical Imaging with Deep Learning*, pages 151–163, 2018.

[50] Mario Siller, Lea Maria Stangassinger, Christina Kreutzer, Peter Boor, Roman D Bulow, Theo JF Kraus, Saskia Von Stillfried, Soraya Wolfl, Sebastien Couillard-Despres, Gertie Janneke Oostingh, et al. On the acceptance of "fake" histopathology: A study on frozen sections optimized with deep learning. *Journal of Pathology Informatics*, 13:100168, 2022.

bioRxiv preprint doi: https://doi.org/10.1101/2023.11.29.568996; this version posted December 1, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

[51] Yang Nan, Javier Del Ser, Simon Walsh, Carola Schönlieb, Michael Roberts, Ian Selby, Kit Howard, John Owen, Jon Neville, Julien Guiot, et al. Data harmonisation for information fusion in digital healthcare: A state-of-the-art systematic review, meta-analysis and future research directions. *Information Fusion*, 82:99–122, 2022.

[52] Jean-Philippe Vert. How will generative ai disrupt data science in drug discovery? *Nature Biotechnology*, pages 1–2, 2023.

[53] Alexandre Filiot, Ridouane Ghermi, Antoine Olivier, Paul Jacob, Lucas Fidon, Alice Mac Kain, Charlie Saillard, and Jean-Baptiste Schiratti. Scaling self-supervised learning for histopathology with masked image modeling. *medRxiv*, pages 2023–07, 2023.

[54] Richard J Chen, Tong Ding, Ming Y Lu, Drew FK Williamson, Guillaume Jaume, Bowen Chen, Andrew Zhang, Daniel Shao, Andrew H Song, Muhammad Shaban, et al. A general-purpose self-supervised model for computational pathology. *arXiv:2308.15474*, 2023.

[55] Eugene Vorontsov, Alican Bozkurt, Adam Casson, George Shaikovski, Michal Zelechowski, Siqi Liu, Philippe Mathieu, Alexander van Eck, Donghun Lee, Julian Viret, et al. Virchow: A million-slide digital pathology foundation model. *arXiv:2309.07778*, 2023.

[56] Wouter Bulten, Hans Pinckaers, Hester van Boven, Robert Vink, Thomas de Bel, Bram van Ginneken, Jeroen van der Laak, Christina Hulsbergen-van de Kaa, and Geert Litjens. Automated deep-learning system for gleason grading of prostate cancer using biopsies: a diagnostic study. *The Lancet Oncology*, 21(2):233–241, 2020.

[57] Peter Ström, Kimmo Kartasalo, Henrik Olsson, Leslie Solorzano, Brett Delahunt, Daniel Berney, David Bostwick, Andrew Evans, David Grignon, Peter Humphrey, Kenneth Iczkowski, James Kench, Glen Kristiansen, Theodorus Van der Kwast, Katia Leite, Jesse McKenney, Jon Oxley, Chin-Chen Pan, Hemamali Samaratunga, and Martin Eklund. Pathologist-level grading of prostate biopsies with artificial intelligence. *Bioinformatics*, 2019.

[58] James D Brierley, Mary K Gospodarowicz, and Christian Wittekind. *TNM classification of malignant tumours*. John Wiley & Sons, 2017.

[59] Guillaume Jaume, Pushpak Pati, Valentin Anklin, Antonio Foncubierta, and Maria Gabrani. Histocartography: A toolkit for graph analytics in digital pathology. In *MICCAI Workshop on Computational Pathology (MICCAI-W)*, pages 117–128, 2021.

bioRxiv preprint doi: https://doi.org/10.1101/2023.11.29.568996; this version posted December 1, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

[60] Abhishek Vahadane, Tingying Peng, Shadi Albarqouni, Maximilian Baust, Katja Steiger, Anna Schlitter, Amit Sethi, Irene Esposito, and Nassir Navab. Structure-preserving color normalization and sparse stain separation for histological images. *IEEE Transactions on Medical Imaging*, 35(8):1962–1971, 2016.

[61] Michael Gutmann and Aapo Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Artificial Intelligence and Statistics*, pages 297–304, 2010.

[62] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[63] Yaniv Taigman, Adam Polyak, and Lior Wolf. Unsupervised cross-domain image generation. In *International Conference on Learning Representations (ICLR)*, 2016.

[64] Lin Zhang, Lei Zhang, Xuanqin Mou, and David Zhang. Fsim: A feature similarity index for image quality assessment. *IEEE transactions on Image Processing*, 20(8):2378–2386, 2011.

[65] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2414–2423, 2016.

[66] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255, 2009.

[67] Jiaxin Cheng, Ayush Jaiswal, Yue Wu, Pradeep Natarajan, and Prem Natarajan. Style-aware normalized loss for improving arbitrary style transfer. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 134–143, 2021.

[68] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *IEEE International Conference on Computer Vision (ICCV)*, pages 2961–2969, 2017.

[69] Simon Graham, Quoc Dang Vu, Shan E Ahmed Raza, Ayesha Azam, Yee Wah Tsang, Jin Tae Kwak, and Nasir Rajpoot. Hover-net: Simultaneous segmentation and classification of nuclei in multi-tissue histology images. *Medical image analysis*, 58:101563, 2019.

[70] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 249–256, 2010.

bioRxiv preprint doi: https://doi.org/10.1101/2023.11.29.568996; this version posted December 1, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

[71] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Instance normalization: The missing ingredient for fast stylization. *arXiv:1607.08022*, 2016.

[72] Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. In *IEEE International Conference on Computer Vision (ICCV)*, pages 2794–2802, 2017.

[73] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv:1412.6980*, 2014.

[74] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations (ICLR*, 2017.

[75] Filippo Maria Bianchi, Daniele Grattarola, and Cesare Alippi. Spectral clustering with graph neural networks for graph pooling. In *International Conference on Machine Learning (ICML)*, pages 874–883, 2020.

[76] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2818–2826, 2016.

[77] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Neural Information Processing Systems (NeurIPS)*, pages 8024–8035, 2019.

[78] Matthias Fey and Jan Eric Lenssen. Fast graph representation learning with pytorch geometric. arxiv 2019. In *ICLR Workshop on Representation Learning on Graphs and Manifolds*, 2019.