Counterfactual Diffusion Models for Interpretable Morphology-based Explanations of Artificial Intelligence Models in Pathology

Laura Žigutytė (1), Tim Lenz (1), Tianyu Han (2), Katherine J. Hewitt (1), Nic G. Reitsam (1,3,4), Sebastian Foersch (5), Zunamys I. Carrero (1), Michaela Unger (1), Alexander T. Pearson (6), Daniel Truhn (2), Jakob Nikolas Kather (1,7,8,9)

(1) Else Kroener Fresenius Center for Digital Health (EKFZ), Faculty of Medicine and University Hospital Carl Gustav Carus, TUD Dresden University of Technology, Dresden, Germany

(2) Department of Diagnostic and Interventional Radiology, University Hospital RWTH Aachen, Germany

(3) Pathology, Faculty of Medicine, University of Augsburg, Augsburg, Germany

(4) Bavarian Cancer Research Center (BZKF), Augsburg, Germany

(5) Institute of Pathology, University Medical Center of the Johannes Gutenberg University Mainz, Germany

(6) Section of Hematology/Oncology, Department of Medicine, University of Chicago, Chicago, IL, USA

(7) Department of Medicine I, Faculty of Medicine and University Hospital Carl Gustav Carus, TUD Dresden University of Technology, Dresden, Germany

(8) Medical Oncology, National Center for Tumor Diseases (NCT), University Hospital Heidelberg, Heidelberg, Germany

(9) Pathology & Data Analytics, Leeds Institute of Medical Research at St James's, University of Leeds, Leeds, United Kingdom

Abstract

Background: Deep learning (DL) can extract predictive and prognostic biomarkers from histopathology whole slide images, but its interpretability remains elusive.

Methods: We introduce MoPaDi (Morphing histoPathology Diffusion), a framework for generating counterfactual explanations for histopathology that reveal which morphological features drive deep learning classifier predictions. MoPaDi combines self-supervised diffusion autoencoders with task-specific classifiers to manipulate images and flip predictions by altering visual features. To address weakly supervised scenarios common in pathology, it integrates multiple instance learning. We validate MoPaDi across four tasks: tissue type, cancer subtypes, slide origin, and a biomarker (microsatellite instability) classification. Counterfactuals are evaluated through pathologist studies and quantitative analysis.

Results: MoPaDi achieves high image reconstruction quality (MS-SSIM 0.966–0.992) and good classification performance (AUCs 0.76–0.98). In a blinded observer study, two pathologists misclassified between 26.7% and 63.3% of synthetic images as real across all tasks, indicating that MoPaDi-generated images often exhibit high perceptual realism. Furthermore, counterfactual images revealed interpretable, pathology-consistent morphological changes recognizable by experts.

Conclusion: MoPaDi is a practical and extensible framework for counterfactual explanations generation in computational pathology. It enables interpretable, model-specific insight into what morphological changes drive classification outcomes, improving interpretability in clinical deep learning models.

Introduction

Deep learning (DL) is preferred for computer vision tasks in biomedical sciences requiring complex decision-making, prediction, or perception. In particular, in computational pathology, DL can predict diagnostic and prognostic biomarkers from hematoxylin and eosin (H&E) stained whole slide images (WSIs) derived from pathology slides[1–4]. DL performs well for many clinically relevant tasks, such as automating pathology workflows to make diagnoses or extract subtle visual information related to biomarkers[5]. However, the so-called 'black box' nature of DL can be problematic because it is often unclear which visual features the model bases its predictions on, limiting interpretability and trust in the classifier[6]. Established explainability methods, such as feature visualization[7] and pixel attribution (e.g., saliency maps, including attention heatmaps)[8], are limited and cannot guarantee that a model's decision is correct[6]. Although widely used in the biomedical field, interpreting these methods is challenging, often requiring domain expertise, and can result in overestimating a model's performance due to confirmation bias[6,9,10].

An alternative approach, counterfactual explanations[11,12], is available but has been underexplored in computational pathology. Counterfactual explanations answer the question: 'how should an image look to be classified as class X' (**Fig. 1a**). Here, we develop and validate MoPaDi (Morphing histoPathology Diffusion), a framework for the generation of counterfactual explanations for histopathology images. Our objective is to make the decision-making process of 'black-box' artificial intelligence (AI) models more understandable and trustworthy.

Previous studies in other image analysis domains have shown that counterfactual explanations can help uncover biases and confounding variables, as well as provide model prediction explanations, yet these efforts are limited to radiology data or outdated generative AI models[13–16]. Therefore, we propose utilizing state-of-the-art diffusion models to generate counterfactual

explanations. An alternative approach to diffusion models is generative adversarial networks (GANs), which have been explored in previous studies[17–23]. However, diffusion models are now considered the state-of-the-art technique, surpassing GANs in synthetic image quality and training stability[24]. Diffusion models work by gradually transforming images into noise and learning the reverse process, enabling the generation of novel images by sampling from the learned data distribution[25]. A key strength of diffusion models over GAN-based methods is their ability to modify real images while preserving details unrelated to the specific manipulation[26]. Furthermore, most computational pathology problems require a weakly supervised approach since labels are typically available only for WSIs consisting of many patches but not for every patch individually[27]. These tasks are usually addressed by multiple instance learning (MIL)[28]. Therefore, we extend the concept of counterfactual image generation to the weakly supervised scenario, in which we train an attention-based classifier on bags of tiles with a patient-level label, and then generate counterfactual explanations for the most predictive tiles. MoPaDi combines diffusion autoencoders[26] with MIL to generate high-quality counterfactual images of histopathology patches, preserving original histopathology patch details and enabling pathologists to comprehensively evaluate a classifier's learned morphologies. To the best of our knowledge, diffusion-based explainability methods have not been applied to histopathology data for counterfactual image generation combined with MIL.

The aim of MoPaDi is to reveal which morphological features in histopathology images drive the predictions of DL classifiers by creating counterfactual explanations, guided by task-specific classifiers and diffusion autoencoders trained in a self-supervised manner on unlabeled data. To validate our framework, we first demonstrate that meaningful features were learned by the model by generating tissue-type counterfactual images. Furthermore, we apply MoPaDi to explain DL model predictions for a broad range of clinically relevant tasks. The evaluation tasks include classifiers differentiating between WSI processing sites for batch effect analysis and microsatellite instability (MSI) status, a colorectal cancer biomarker predictable from H&E-stained WSIs[3,29]. Four datasets were utilized: two colorectal carcinomas (CRC) datasets, a breast cancer (BRCA) dataset, and a pan-cancer dataset comprising 32 cancer types. For the latter, we focused on hepatobiliary cancers (hepatocellular carcinoma [HCC], and cholangiocarcinoma [CCA]) and lung cancers (lung adenocarcinoma [LUAD], and lung squamous cell carcinoma [LUSC]) (**Fig. 1b**). Our framework, validated across diverse datasets, reveals key morphological features driving classifiers predictions through automated, targeted manipulation of histopathological images toward counterfactual images, enabling understanding of morphological patterns learned by the DL model across cancer types and a biomarker.

Results

Diffusion Autoencoders Efficiently Encode Histopathology Images

The first step in evaluating a generative model is to assess the fidelity of the image embeddings. Therefore, we evaluated the performance of our custom-trained diffusion autoencoders in reconstructing histopathology images by comparing 1,000 original and reconstructed images using mean squared error (MSE) and multi-scale structural similarity index measure (MS-SSIM) values. We found that the models achieved high performance across different datasets, as indicated by low MSE and high MS-SSIM values (**Table 1**). A qualitative evaluation confirmed a high rate of near-perfect reconstructions, although some minor artifacts were sometimes found to be accentuated by the compression-decompression process (**Fig. S1**). To further assess the encoder's ability to learn meaningful histopathological features, we trained additional denoising diffusion implicit models (DDIM) on feature vectors extracted by the feature extractor part of the diffusion autoencoder. This enabled the sampling of synthetic feature vectors that are subsequently used for novel image generation. These models produced high-quality synthetic images, indicated by low Fréchet Inception Distance (FID) between a held-out set of original images and 10,000 generated synthetic images (**Fig. S2**) for most datasets: 16.85 for TCGA CRC, 17.85 for TCGA BRCA, and 4.71 for TCGA Pan-cancer. In contrast, the CRC-VAL-HE-7K had a higher FID score of 36.95 (**Table 1**). Together, these data show that our custom-trained diffusion autoencoders effectively capture key image features and reconstruct histopathological patches across multiple datasets, demonstrating their potential as a suitable foundation for generating counterfactual explanations.

Tissue Type Counterfactuals Exhibit Meaningful Patterns

Next, we assessed MoPaDi's capabilities to generate counterfactual images for the tissue-type classifier, i.e., to flip the predicted class of a given image. We first trained a diffusion autoencoder alongside a feature extractor (**Fig. 1c**) on the NCT-CRC-HE-100k dataset, representing 9 tissue classes in colorectal cancer (**Fig. S3a**). Features extracted from the training dataset formed distinct clusters upon dimensionality reduction, suggesting the model learned meaningful class-specific representations (**Fig. 2a**). We then trained one logistic regression classifier on the extracted features to jointly predict all nine tissue classes. For the generation of counterfactual explanations, we focused on the clinically most relevant pair: normal colon mucosa (NORM) and dysplastic epithelium (TUM). However, counterfactual images can also be generated for other tissue types predicted by the classifier (representative examples for adipose, smooth muscle, mucus, debris, lymphocytes, and cancer-associated stroma counterfactuals shown in Supplementary **Fig. S4**). Evaluated on the independent CRC-VAL-HE-

7K dataset, the classifier reached an area under the receiver operating characteristic curve (AUROC) of 0.91 ± 0.01 for NORM and 0.98 ± 0.01 for TUM (bootstrapped 95% confidence intervals [CI]; **Fig. 2b**). AUROC values for the remaining classes ranged from 0.79 to 1.00 (**Fig. S3b**). We used the classifier's class-direction vector v to shift the latent features toward a selected target class X. The manipulated features then conditioned the diffusion-based decoder (**Fig. 1f**), enabling us to generate counterfactual images for patches from the CRC-VAL-HE-7K dataset at increasingly larger manipulation amplitudes (α). Manipulated features also shifted closer to the opposite class in the lower-dimensional space (**Fig. 2a** zoomed-in plot), particularly when features are near the boundary between classes (as seen in the colon mucosa tile bordering dysplastic and healthy tissue in **Fig. 2c**). To quantify the effectiveness of counterfactuals generation, we computed the percentage of manipulated tiles that were predicted as the opposite class across increasing amplitudes (**Fig. S5c**). We observed that NORM to TUM transitions generally required lower amplitudes to flip the classifier prediction, consistent with the lower AUC for NORM. Furthermore, to assess whether generated counterfactuals satisfy other axiomatic properties[30] besides effectiveness, we examined null and reverse transformations to evaluate composition and reversibility (Supplementary **Fig. S6**). In the null case (NORM to NORM), we observe minimal changes in appearance and high SSIM values, confirming that class-consistent manipulations preserve visual fidelity and do not introduce spurious alterations (Supplementary **Fig. S6a**). In contrast, reversing a transformation (e.g., TUM to NORM to TUM) does not yield the original image (Supplementary **Fig. S6b**), which is expected due to the design of MoPaDi's framework, which prioritizes classifier-specific interpretability over bijective transformations.

To further evaluate the quality of counterfactual images, we conducted a blinded evaluation study with two pathologists. Each observer was shown 30 transitions to a counterfactual image (15 for TUM and 15 for NORM) and asked to identify morphological changes. For example, counterfactual images showed that the tile was classified as NORM rather than TUM due to the presence of healthy, well-ordered goblet cells (**Fig. 2c**). Conversely, the reverse counterfactual pair (NORM to TUM) displayed loss of goblet cells and connective tissue, along with the appearance of hyperchromatic nuclei, revealing plausible morphological changes that would shift the classification toward TUM (**Fig. 2d**).

We further evaluated the quality of the generated images quantitatively by analyzing the distribution of six cell types in real and synthetic images (i.e., epithelial cells, connective tissue cells, lymphocytes, plasma cells, eosinophils, and neutrophils). To achieve this, we applied a pretrained DL-based nuclei segmentation and classification model. The results indicated that synthetic healthy colon mucosa images had cell type distributions similar to real images (**Fig.**

2e). Synthetic dysplastic epithelium tiles predominantly contained epithelial cells, with a few connective tissue cells and immune cells (**Fig. 2f**).

Collectively, these findings demonstrate that our model learned and could generate key histological features associated with normal colon mucosa and dysplastic epithelium, and that counterfactual images can serve as intuitive visual tools to reveal which morphological features influence the classifier's predictions.

Counterfactuals Enhance Interpretation of Linear Liver Cancer Predictions

We further validated our method by training a linear MoPaDi classifier (**Fig. 1d**) to distinguish two hepatobiliary cancer types, HCC and CCA, and generating related counterfactuals. A logistic regression model trained on features extracted from HCC and CCA cases in the TCGA Pan-cancer dataset achieved an AUROC of 0.90 ± 0.01 (95% CI) and an average precision (AP) of 0.57 ± 0.03 on the held-out test set (**Fig. 3a**). The lower AP is attributed to the imbalanced dataset, with only 10% of patients having CCA (**Table S1**). Averaging patch-level predictions to patient-level yielded an AUROC of 0.94 ± 0.10, indicating an effective distinction between hepatobiliary cancer types despite some patch-level errors. A qualitative analysis of the generated counterfactual images revealed that the model accurately learned representative morphological features for both classes. HCC patches manipulated to resemble CCA displayed a more glandular pattern, reduced eosinophilic cytoplasm, and lesser trabecular architecture (**Fig. 3d**). Meanwhile, CCA patches manipulated to resemble HCC displayed a more trabecular growth pattern with loss of glandular structure, eosinophilic cytoplasmic inclusions, possible Mallory bodies, and reduced stromal tissue (**Fig. 3e**). Increasing manipulation levels resulted in decreased MS-SSIM values and higher opposite-class predictions. These distinct morphological changes in counterfactual images demonstrate the model's ability to capture hepatobiliary cancer characteristics while enhancing the classifier's interpretability.

MIL Enables WSI-level Cancer-type Classification Counterfactuals

Given the complexity of weakly supervised WSI tasks, where histological phenotypes are visible only in select regions, we replace the linear classifier in MoPaDi with attention-based MIL to assess its potential for generating counterfactual explanations for tissue properties that are only defined by slide-level annotations, not patch-level annotations. For this, we trained a diffusion autoencoder, which contains a feature extractor, and combined it with an attention-based classifier of extracted features, enabling targeted generation of counterfactual images that reveal key morphologic features associated with different classes. We investigated the

efficacy of this approach in tumor classification problems: lung cancer (LUAD vs. LUSC), breast cancer (invasive lobular carcinoma [ILC] vs. invasive ductal carcinoma [IDC]), and the previously discussed HCC vs. CCA. Using HCC and CCA features from the feature extractor of our diffusion autoencoder, we trained an attention-based MIL classifier. Integrating MIL led to predictive performance similar to the patch-based linear approach for this task, reaching a mean AUROC of 0.96 ± 0.01 (**Fig. 3b**) and an AP of 0.82 ± 0.07 (**Fig. S7a**) in the 5-fold cross-validation. Top tiles from correctly classified patients displayed distinctive features: original HCC tiles showed cells arranged in sheets or trabecular patterns with lipid droplets, lacking glandular structures and fibrous stroma, while original CCA tiles showed glandular structures and more fibrous stroma (**Fig. S7b**). Counterfactual images guided by the MIL classifier resembled those generated using the linear approach with patch-level logistic regression (**Fig. S7c,d**). Cell type analysis showed similar distributions between counterfactual and real images (**Fig. S7e,f**). Increasing manipulation amplitude (α) resulted in higher opposite class predictions, with 100% of generated images being predicted as the opposite class at α = 0.08 (**Fig. 3c**)

We further validated counterfactual image generation guided by the MIL classifier by training a LUSC-LUAD classifier, achieving a mean AUROC of 0.91 ± 0.01 (with an improvement of 0.10 over the linear method) and an AP of 0.89 ± 0.02 (**Fig. S8a**). Pathologists analyzed LUSC to LUAD counterfactual transitions (**Fig. 4a**), with one noting increased glandular differentiation in 4 out of 8 patients and the other in 5 out of 8 (two top tiles per patient reviewed), while both observed lower pleomorphism (6/8). Additional features identified included increased vacuolization (4/8), clearer cells (4/8), and decreased tumor solidity (3/8), suggesting that the absence of these features contributes to classification as LUSC. However, some counterfactual tiles retained LUSC features atypical for LUAD, such as prominent dyskaryotic changes, nuclear molding, and residual squamous differentiation (**Fig. S8c**), indicating that the model may prioritize other features when making classification decisions. Additionally, LUAD to LUSC transitions showed decreased glandular architecture, increased solid growth patterns or tumor cell nests (4/8; 5/8), and increased cell density (2/8; 5/8). Other features included lower atypia (1/8; 2/8), intercellular bridges, and a keratin pearl (**Fig. S8d**). Counterfactual images maintained largely stable cell type compositions across most cell types, with minor differences observed, notably in immune cells (**Fig. S8e**). Counterfactual image generation effectiveness remained high across both classes, with a total of 98.8% of counterfactual images at amplitude α = 0.06 being predicted as the opposite class relative to the original (**Fig. S8f**). Taken together, this data demonstrates that MIL classifier-guided counterfactual image generation works effectively; thus, MoPaDi can be applied to WSI-level datasets.

We then investigated whether meaningful morphological features could be identified in counterfactuals for IDC and ILC classification in the TCGA BRCA cohort. The trained classifier reached a mean AUROC of 0.86 ± 0.02 and AP of 0.64 ± 0.04 (**Fig. S9a**). In ILC to IDC counterfactual transitions (**Fig. 4b top**), the most consistently observed changes were increased hyperchromasia (5/8) and atypia (4/8), showing that the tiles were classified as ILC due to the absence of these features. Other notable features included a shift towards less single-cell arrangement without clear glandular pattern formation (2-3/8), larger tumor cells (3/8), and tumor cell arrangement in larger nests (2/8). Some top tiles contained only muscle with single tumor cells, making feature assessment not feasible, but potentially representing the diffusely infiltrative nature of these neoplasms. Meanwhile, IDC to ILC counterfactual transitions (**Fig. 4b bottom**) lacked glandular architecture (7/8), showed less prominent nucleoli (8/8), and exhibited a single-cell arrangement of tumor cells (8/8), suggesting that the classifier distinguishes IDC based on the prominence of nucleoli and growth patterns. Although generated images appeared as realistic ILC representations with 97.4% of counterfactuals predicted as the flipped class (**Fig. S9e**), the top contributing tiles for which counterfactuals were generated often lacked characteristic IDC features. However, despite the classifier's moderate performance and inconsistent features in top tiles, we identified distinct ILC and IDC morphological characteristics, indicating that the classifier learned meaningful distinctions between these cancer types.

Counterfactual Transitions Explain Batch Effects

Considering the impact of batch effects on histopathological analyses, we next focused on whether batch effects could be identified by generating counterfactual explanations for classifiers differentiating the center of origin of WSIs. First, we demonstrated that features extracted with a pretrained diffusion autoencoder, when visualized in a reduced-dimensional space, revealed clustering of tiles from WSIs obtained at the same center, i.e., the institution from which the sample was derived, highlighting potential batch-related structure in the data (**Fig. 5a**).

We hypothesized that the originating institution of WSIs could be accurately predicted with a DL model. To test this, we trained a MIL classifier to distinguish WSIs in the BRCA cohort from Roswell Park Comprehensive Cancer Center in Buffalo, NY (identifier E2 in TCGA) versus those from other institutions. The 5-fold cross-validation models achieved a mean AUROC of 0.98 ± 0.00 and an AP of 0.87 ± 0.04 (**Fig. 5b**), confirming previous studies, which demonstrated that the original institution is encoded in patterns in histopathology images[31].

To assess whether these predictive signals reflect underlying confounders, we compared clinical and molecular variables across centers. As shown in Supplementary **Table S5**, center E2

differed from the others in breast cancer subtype frequencies, BRCA1/2 mutation status, and patient ethnicity. We then generated counterfactual explanations for histology patches from other (non-E2) centers (**Fig. 5c**). Quantitative analysis of the transitions revealed an increase in epithelial cell counts while connective tissue cell numbers remained relatively constant (**Fig. S10**). Some counterfactual images revealed significant changes ($p < 0.01$ or $p < 0.05$) in nuclear size and mean intensity. Qualitative analyses at the tissue level revealed a transition towards more intense eosin staining in the stroma, higher contrast between tissue components, the appearance of clusters of red blood cells, and the disappearance or shrinkage of adipocytes (**Fig. 5c**). Conversely, when E2 samples are manipulated to look less like E2 (**Fig. S11b**), epithelial and connective tissue cell counts had an inverse relation, i.e., epithelial cells decreased, while connective tissue cells increased or stayed in a similar range (**Fig. S12**). Counterfactual images also exhibited significant decreases ($p < 0.001$) in mean nuclear intensity for both cell types in all representative examples (**Fig. S12**).

Together, these data confirm pronounced batch effects between institutions in TCGA and show that MoPaDi can uncover the morphological patterns associated with them.

Counterfactual Images Help Identify Features of a Molecular Biomarker

Having developed and refined MoPaDi, we finally assessed its capability to predict molecular biomarkers from pathology images and to explain morphological patterns. We explored the link between genotype and phenotype by investigating counterfactual explanations for MSI in the TCGA CRC cohort. A 5-fold cross-validation was performed with MIL to predict MSI-high (MSIH) versus non-MSIH-high (nonMSIH), achieving a mean AUROC of 0.73 ± 0.08 and a mean AP of 0.40 ± 0.09, with lower AP reflecting high-class imbalance (**Fig. 6a**). We then visualized top predictive tiles, which showed high staining variability (**Fig. 6b**), and subsequently generated counterfactual transitions to explain the classifier's predictions. These transitions revealed center-specific batch effects through clearly visible color changes (**Fig. 6c,d**). Quantitative measurements supported these observations, showing significant ($p < 0.001$ or $p < 0.01$) changes in mean nuclear intensity for both epithelial and connective tissue cells (**Fig. S13, S14**).

Pathologists identified that top predictive tiles from MSIH patients typically contained inflammatory cells, atypical epithelium lacking normal bowel crypt structure, debris, high-grade nuclear pleomorphism, and in some cases, medullary or mucinous morphology, and signet-ring cell (**Fig. 6c and S15**). Counterfactual transitions for MSIH to nonMSIH typically displayed a more glandular pattern (7/8), decreased mucin (4/8), and a less solid or sheet-like growth pattern when the original tile displayed a medullary growth (1/8). Quantitative analysis showed

a marked increase in epithelial and connective tissue cell counts (**Fig. S13**). Regarding the nonMSIH to MSIH counterfactual transitions (**Fig. 6d**), pathologists identified the loss of glandular architecture as the most prominent feature (7/8). Other features included more solid growth patterns (5/8), increased mucinous appearance (3/8), and a more vacuolated appearance (4/8). Quantitative analysis revealed an increase in inflammatory cells, along with decreases in epithelial cell counts and nuclear eccentricity of both epithelial and connective tissue cells (**Fig. S14**).

To further quantify the realism of generated counterfactuals, we compared the cell-type composition between real and manipulated tiles, confirming that counterfactual images preserved epithelial and connective tissue composition (**Fig. S16a,b**). Additionally, we assessed the effectiveness of our counterfactual generation approach by evaluating classifier predictions across increasing manipulation amplitudes (α), finding that most counterfactuals reached the target class even at low α values (**Fig. S16c**). Furthermore, to further assess the perceptual quality of counterfactual images, we conducted a blinded evaluation with two pathologists. Each observer reviewed 300 tiles (150 real, 150 synthetic) and classified each tile as either real or synthetic (**Fig. S17a**). Pathologists misclassified between 26.7% and 63.3% of synthetic images as real across all tasks, indicating that MoPaDi-generated images often exhibit high perceptual realism (**Fig. S17b**).

These findings, from pathologists' observations of architectural changes in counterfactual transitions to quantitative analyses of cellular composition, reveal features the DL classifier uses to distinguish MSIH tumors from nonMSIH, enhancing the interpretability of the prediction process.

Discussion

Throughout this study, we developed a diffusion model – MoPaDi – capable of generating meaningful counterfactual images and enhanced it with MIL to handle WSIs. MoPaDi demonstrates how diffusion models can transform computational pathology from a passive descriptive tool into an instrument for virtual experimentation. Unlike traditional DL approaches that simply map images to predictions, MoPaDi enables researchers to ask and answer "what if" questions about biological specimens through counterfactual image generation.

The effectiveness of this approach was extensively validated across multiple experimental scenarios, spanning multiple tumor types and providing quantitative and qualitative evidence for the plausibility of generated images by using several orthogonal metrics. Our diffusion autoencoder achieved excellent image reconstruction quality (MS-SSIM) and generated highly realistic counterfactual images that were often indistinguishable from real histological samples

in blinded pathologist evaluations. Furthermore, we provide evidence that MoPaDi captures and recapitulates key biological processes. For example, the overall trend for cell type distributions in generated images of colorectal cancer aligns with established cancer and adjacent normal tissue microenvironment characteristics[32]. Tumor tiles showed high epithelial cell concentrations, especially in normal to tumor transitions, suggesting the model recognizes epithelial abundance as a key tumor characteristic, consistent with tumor biology[33]. Similarly, for distinguishing liver tumor types and lung tumor types, our model recapitulated key biological properties[34–36]. Finally, we extended our analysis to molecular biomarkers by investigating MSI status in colorectal cancer. While achieving moderate performance compared to state-of-the-art models[3], MoPaDi revealed key morphological patterns associated with MSI status through counterfactual explanations. Even with a limited set of top predictive tiles, pathologists identified characteristic features such as signet-ring cells, mucinous, medullary patterns, inflammatory infiltration, and poor differentiation that are known to be associated with MSI in colorectal cancer[37–41]. These findings are relevant because they demonstrate that DL models can not only recognize patterns in pathology images but can also manipulate them in biologically meaningful ways. Unlike standard attention heatmaps typically used in computational histopathology, which highlight broad regions of interest and can lead to confirmation bias[6], our approach identifies specific features the model uses for predictions.

As another application of MoPaDi, we demonstrate that it can pinpoint the underlying reason for batch effects. Batch effects are another important issue in computational pathology because DL models often learn variations in data collection techniques across different centers rather than focusing on target-specific morphological features[42]. We and others have shown that WSIs from the specific center can be predicted with high accuracy[31,43]. Our qualitative and quantitative analysis of generated counterfactual images indicated that the classifier learns not only staining variations but also histologic differences that may stem from biological variations in treated patient populations at different centers, such as differing cohort compositions. These findings are consistent with literature showing that TCGA WSIs contain a multifactorial site-specific signature[31]. The advantage of MoPaDi is that these center-level differences can be visualized and quantified for each image, deepening our understanding of batch effects at a granular level.

Finally, MoPaDi is the first generative model that works in a MIL framework to generate counterfactual explanations for biomarkers that are only defined on the level of slide annotations. This capability is important because many important molecular and clinical features in computational pathology are only available at the whole-slide level, not at the patch level. By combining diffusion models with MIL, we enable the interpretation of weakly-supervised learning problems that are common in clinical practice.

Our approach has several important limitations. Results are currently limited to patch-level analysis and require careful validation by domain experts to ensure biological plausibility. Each dataset requires training of a custom diffusion autoencoder model, which is computationally intensive and time-consuming, ranging from several days to weeks on consumer hardware, depending on dataset complexity. Although the model typically generates realistic images, it sometimes fails to capture fine biological nuances – for instance, in colorectal cancer transitions, we observed cases where generated images lacked proper basement membrane structure or showed insufficient nuclear hyperchromatism. Additionally, while our cell type distributions generally align with known tumor biology, the model shows consistent biases in generating certain cell types over others. Future work could address these limitations by integrating pathology foundation models with diffusion-based counterfactual generation to improve both predictive performance and biological accuracy while maintaining the advantages of high-quality image synthesis and selective feature manipulation.

Looking forward, this work opens new possibilities for computational pathology. By enabling virtual experimentation, tools like MoPaDi could help researchers pre-screen hypotheses before conducting expensive and time-consuming wet lab experiments. This could be particularly valuable given increasing restrictions on animal testing and the challenge of translating animal model results to humans. Additionally, these methods could provide new insights into morphological features that correlate with outcomes, potentially revealing previously unknown biological mechanisms. As these methods continue to evolve, they may become an integral part of the cancer research toolkit, complementing traditional experimental approaches and accelerating the pace of discovery in cancer biology.

Methods

Ethics Statement

This study was carried out in accordance with the Declaration of Helsinki. We retrospectively analyzed anonymized patient samples using publicly available data from “The Cancer Genome Atlas” (TCGA, https://portal.gdc.cancer.gov) and Zenodo. The Ethics commission of the Medical Faculty of the Technical University Dresden provided guidelines for study procedures.

Datasets

This study explored multiple datasets (**Fig. 1b**), starting with the publicly accessible NCT-CRC-HE-100K dataset[44]. This dataset comprises 100,000 H&E-stained histopathological images derived from 86 formalin-fixed paraffin-embedded (FFPE) samples. These samples represent both human CRC and normal tissues. The dataset encompasses nine distinct tissue classes: adipose, background, debris, lymphocytes, mucus, smooth muscle, normal colon mucosa, cancer-associated stroma, colorectal adenocarcinoma epithelium (**Fig. S3a**). All classes were utilized to train the diffusion autoencoder. The images, each with a resolution of 224 x 224 pixels at 0.5 microns per pixel (MPP), were color-normalized with Macenko's method[45]. The trained model was evaluated on 7,180 non-color-normalized image patches from the CRC-VAL-HE-7K dataset, published alongside the NCT-CRC-HE-100K, and serving as a non-overlapping dataset for testing purposes[44]. Further investigation extended to a second CRC dataset from The Cancer Genome Atlas Research Network (TCGA), comprising WSIs obtained from FFPE samples. The TCGA BRCA cohort was utilized as well to explore counterfactual explanations for this cancer type. Finally, we used another publicly available dataset, further referred to as the TCGA Pan-cancer dataset, where varying resolution patches (256 x 256 pixels at MPP ranging from 0.5 to 1) were extracted from tumor regions from WSIs of 32 cancer types[46].

Data Pre-processing

A 10–20% split of the data at the patient level, depending on the original cohort size, was performed before preprocessing to create the training and held-out test sets, with exact patient numbers given in **Table S1**. Testing data was used to evaluate the performance of the classifier and assess the quality of image reconstruction, generation, and manipulation. We ensured that many centers were only in the test set, except one bigger one for batch effects analysis in TCGA BRCA. Subsequently, for TCGA datasets, where raw WSIs were processed directly, the images were tessellated to obtain patches measuring 512 x 512 pixels at 0.5 MPP. WSIs that did not have the original MPP value in the metadata or had no available tumor annotations were discarded. This resulted in 1,115 WSIs from 1046 patients and 588 WSIs from 582 patients for the BRCA and CRC cohorts, respectively. Patches were derived from tissue regions determined by Canny edge detection. Then, for both cohorts, patches derived from non-tumorous regions were filtered out, retaining only those originating from tumor areas. This selection process was guided by tumor area annotations provided by pathology trainees supervised by an expert pathologist. Patches that overlapped by at least 60% with annotated tumor areas were included. During the diffusion model's training, the channels of each image were

scaled to achieve a mean of 0.5 and a standard deviation of 0.5, thereby standardizing the pixel value distribution throughout the dataset.

Diffusion Autoencoders

MoPaDi's architecture builds on the official diffusion autoencoders implementation by Preechakul et al.[26] It consists of a conditional DDIM image decoder (θ), conditioned on a feature vector obtained with a learnable convolutional neural network (CNN) encoder. We refer to this encoder as the feature extractor, which learns to extract meaningful high-level information (feature vectors $\mathbf{z}$) from input images (**Fig. 1c**). The encoder and decoder are trained jointly by optimizing a simplified noise prediction loss (**Eq. 1**):

$$\mathcal{L}_{simple} = \sum_{t=1}^{T} \quad \mathbb{E}_{x_0,\epsilon_t}\left[\left|\left|\epsilon_\theta(x_t, t, z) - \epsilon_t\right|\right|_2^2\right] \tag{1}$$

where ϵ_θ is the predicted noise at diffusion timestep t, x_t is the noisy image at that step, $\boldsymbol{z}$ is the encoded feature vector, and $\epsilon_t \sim \mathcal{N}(0, I)$ is the sampled Gaussian noise. This loss function encourages the encoder to produce meaningful feature vectors while enabling the decoder to reconstruct high-fidelity images from them.

We trained four separate autoencoding models for different datasets. For datasets with smaller images (224 x 224 or 256 x 256 px), we used 8 Nvidia A100 40 GB graphics processing units (GPUs), while for datasets with bigger images (512 x 512 px) – 4 Nvidia A100 80 GB GPUs. The models were trained until the convergence, i.e., until the approximate FID score, computed on 5,000 images, was below 20 (~10 for datasets with smaller patches). The training duration varied from several days to weeks due to the different complexities of datasets. A global batch size of 24 for 512 x 512 px models and 64 for smaller models, divided evenly among the GPUs, and a learning rate of 10^{-4} were used. To evaluate the quality of the latent space learned by the autoencoding model and generate novel synthetic images, we additionally train a latent DDIM on the distribution of feature vectors obtained from the diffusion-based encoder.

Classifiers and Counterfactual Image Generation

To enable classifier-guided counterfactual image generation, we first establish classification models for both linearly separable and non-separable tasks. For linearly separable tasks, such as tissue type classification, we train a multiclass logistic regression classifier on 512-dimensional features extracted with a pretrained MoPaDi feature extractor (**Fig. 1d**). Each tile is annotated with a one-hot target vector indicating its true class. Features are normalized to zero mean and unit variance before classification. Although the task is multiclass, we treat it

as a one-vs-rest problem by applying binary cross-entropy loss independently to each class. The training process involved 300,000 tile exposures.

For linearly non-separable classes, such as patient-level biomarker status, we adopt a MIL approach using transformer-based pooling. A transformer-based classifier processes bags of tile-level features, applying layer normalization, multi-head cross-attention, and a self-attention block, followed by a classifier head with two fully connected layers (with dropout and SiLU[47] activation), and a final linear output layer (**Fig. 1e**). The classifiers were trained using Adam optimizer, learning rate of 0.0001, early stopping with a patience of 20 epochs, and cross-entropy loss with class weighting to address class imbalance. Model's variability is assessed via 5-fold cross-validation. For counterfactual image generation, a final classifier is trained on the full training set, using 50–200 epochs based on the cross-validation results.

Both the linear and MIL classifiers guide counterfactual image generation by defining a class-specific direction of manipulation in feature space. In the linear case, the learned weight vector v of the logistic regression model defines the direction f in feature space that increases the logit for the target class. This direction is used to shift the tile's feature vector across the decision boundary. In the MIL setting, where the relationship between individual tiles and the label is non-linear, we compute the gradient of the loss ($\mathcal{L}$) for a chosen "counterfactual" class c with respect to the selected tile's feature vector z and use this gradient as the direction of manipulation:

$$\mathcal{L}(s(z), c) = s_c(z) \tag{2}$$

where $s_c(z)$ denotes the raw logit that the classifier assigns to a class c given a feature vector z.

$$f = -\nabla_z \mathcal{L}(s(z), c) = -\nabla_z s_c(z) \tag{3}$$

After modifying the original feature vector z by adding f (scaled by a manipulation amplitude α) (**Eq. 4**), the updated feature vector $\tilde{z}$ is passed to the conditional diffusion decoder (D), along with the original image encoded to noise (x_T), to generate the counterfactual image $\hat{x}$ (**Fig. 1f, Eq. 5)**.

$$\tilde{z} = z + \alpha \cdot f \tag{4}$$

$$\hat{x} = D(x_T, \tilde{z}) \tag{5}$$

We apply counterfactual manipulation by progressively increasing α along the target-class gradient direction in latent space. For each dataset, we use a fixed set of manipulation amplitudes (e.g., $\alpha \in \{0.2, 0.4, 0.6, 0.8, 1.0\}$ for the linear approach and $\alpha \in \{0.02, 0.04, 0.06, 0.08\}$ for the MIL approach). Even if some manipulated tiles do not cross the classifier's decision bound-

ary at the largest α, we apply the same α values uniformly across the cohort to ensure comparability. We increase α until the majority of re-encoded counterfactual images are classified as the target class.. The predictions are based on what the classifier would predict if it only had seen the particular tile. For the MIL approach, we generate counterfactual images for the top 5 most contributing tiles per WSI, with each tile being manipulated individually.

Evaluation Metrics

We used MSE between the pixels of the original image and the reconstructed one from the diffusion latent space and SM-SSIM[48] to assess the quality of the autoencoder's reconstruction capability for 1,000 images in each dataset. Structural similarity index[49] was used to visually represent similar and dissimilar regions between the two images converted to grayscale for more straightforward results interpretation. Latent space's ability to retain semantically meaningful information was also evaluated by sampling 10,000 synthetic images and computing the FID score to compare the distribution of generated images with the distribution of real images in the test set using the clean-fid library[50]. A lower FID score indicates higher similarity.

To evaluate the performance of classifiers for each experiment, we computed AUROC for the held-out test set. Bootstrapping was employed as a resampling technique to estimate robustness and confidence intervals. Resampling was done with replacement 1,000 times. We constructed ROC curves for each resampled set and interpolated between them to obtain a set of smoothed curves. 95% confidence intervals were extracted from smoothed curves. The latent space and the direction of manipulation were visualized using t-SNE, enabling a representation of extracted 512-dimensional feature vectors in a reduced two-dimensional format. For the TCGA BRCA dataset, we used UMAP instead of t-SNE due to the large number of tiles, which made t-SNE computationally inefficient.

Observer Studies

To evaluate the realism of generated counterfactual images, a pathologist and a pathology trainee supervised by a board-certified pathologist reviewed a total of 300 images (150 real and 150 synthetic, randomly selected) in a blinded setting (**Fig. S17a**). For each task, 30 real tiles (15 per class) were randomly selected from the test set, and corresponding counterfactuals were generated by transitioning each tile to the opposite class. The observers were presented with a mixed set of tiles and asked to classify each one as either real or synthetic. The proportion of correctly identified images served as a measure of the diffusion model's ability to generate realistic and visually convincing histopathology images.

To further assess whether the classifiers capture meaningful morphological features associated with specific classes, observers were also asked to identify features that change during the transition to a counterfactual image. In these experiments, class labels for both the original and the counterfactual images were provided, and pathologists were asked to identify histopathological features that change during the transition (**Fig. 1g**). They were given 16 examples for each classifier (8 examples per class), each containing two top predictive tiles from the same patient who was predicted correctly by the model with high confidence.

Quantitative Assessment of Morphological Changes

Pretrained nuclei segmentation and cell type classification models were applied to original and counterfactual images to quantify changes in certain cell types and morphological characteristics. *DeepCMorph* models were used for this purpose[51]. These models were trained to classify nuclei of six cell types: epithelial, connective tissue, lymphocytes, plasma cells, neutrophils, and eosinophils. Varying resolution 224 x 224 px patches from H&E WSIs coming from many different datasets were used to train *DeepCMorph*, including NCT-CRC-HE-100K and TCGA pan-cancer datasets. However, the models were primarily trained to classify cancer and tissue types, with the aim of enhancing performance by enabling the model to learn biological features such as cell types. Therefore, we were able to utilize the intermediate segmentation masks for our datasets.

To obtain instance segmentation masks only of nuclei of epithelial cells, we performed watershed-based splitting of touching objects in a semantic segmentation mask of these cells, an additional binary erosion operation to split remaining touching nuclei, and subsequently connected component labeling followed by a binary dilation to get back to the original nuclei size. We removed labels with an area lower than 30 pixels, as these typically were artifacts. As for the instance segmentation masks of connective tissue cell nuclei, we skipped these preprocessing steps because nuclei are more elongated and sparser. These steps enabled automatic counting of cells in the original and counterfactual images. Finally, quantitative shape and intensity-based measurements, i.e., area, mean intensity, and eccentricity, were extracted using the *scikit-image* library. We additionally quantify tissue architectural organization by computing the spatial entropy of cellular distributions. This involves creating a 2D histogram of cell centroids, normalizing it to a probability distribution, and calculating Shannon's entropy. This measurement was expected to capture tissue differentiation patterns, but its strong correlation with cell count suggests the need for more refined metrics to characterize architectural organization in histopathology patches. Lastly, we perform statistical hypothesis testing to compare measurements of original and counterfactual images. Since generated images are not independent and the measurements are not normally distributed, we perform a two-sided

permutation test with 10,000 resamples. For each cell type (epithelial or connective), the null hypothesis assumes no difference in the proportions of that cell type between original and counterfactual images. To evaluate this, the measurements for original and counterfactual images were shuffled randomly to create a null distribution of the test statistic under the null hypothesis. The p-value was then calculated as the proportion of permuted test statistics that are as extreme or more extreme than the observed test statistic. Cell type counts, measured as pixel counts in segmentation masks, were also compared between original and counterfactual images of the opposite class (i.e., manipulated to look like the compared images' class). For each cell type, Mann-Whitney U test was used to assess whether the distribution of percentages differed between the original and counterfactual groups. To account for multiple comparisons across cell types, Bonferroni correction was applied to adjust the p-values. Additionally, Cliff's delta was calculated to quantify the effect size and assess practical significance.

Data and Code Availability

MoPaDi code and instructions to download trained models are publicly available on GitHub at https://github.com/KatherLab/mopadi. The official implementation of diffusion autoencoders for natural images can be found on GitHub at https://github.com/phizaz/diffae. Digital histopathology WSIs from TCGA, together with the clinical data used in this study, are available at https://portal.gdc.cancer.gov and https://cbioportal.org. NCT-CRC-HE-100K/CRC-VAL-HE-7K[44] and TCGA Pan-cancer[46] datasets are publicly available at https://zenodo.org/records/1214456 and https://zenodo.org/records/5889558.

Tables

Table 1 | Evaluation of diffusion autoencoders trained on different datasets. The reconstruction quality of encoded images was evaluated using multi-scale structural similarity index measure (MS-SSIM) and mean square error (MSE), which perform pixel-by-pixel comparisons of 1,000 reconstructed images. The ability of the diffusion model's latent space to capture meaningful information was further assessed by generating 10,000 synthetic histopathology images and computing the Fréchet Inception Distance (FID) score between the generated and held-out test set images to evaluate their quality.

Model	**MS-SSIM ↑**	**MSE ↓**	**FID score ↓**
NCT100k	0.968 ± 0.036	$(0.4 \pm 0.5) * 10^{-3}$	36.95
TCGA Pan-cancer	0.992 ± 0.001	$(0.8 \pm 0.2) * 10^{-3}$	4.71
TCGA CRC	0.967 ± 0.060	$(8.1 \pm 14.0) * 10^{-3}$	16.85
TCGA BRCA	0.966 ± 0.062	$(4.5 \pm 11.3) * 10^{-3}$	17.85

Figures

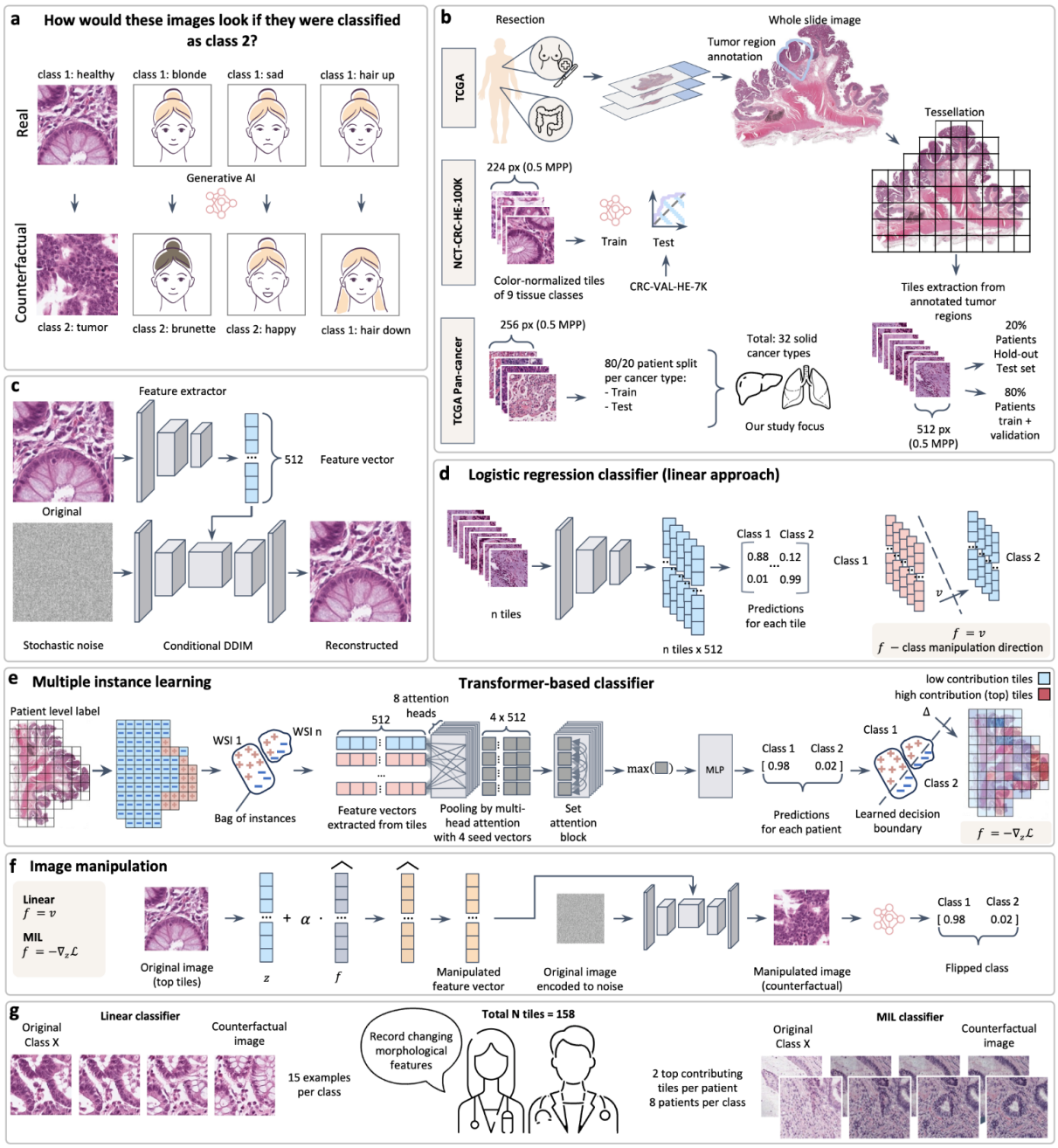

Fig. 1 | Overview of the experimental setup. a, Counterfactual explanations answer the question: 'What would this image need to look like to be classified as class X?'. **b,** Datasets and preprocessing steps. Colorectal and breast cancer datasets from TCGA, consisting of digitized whole slide images (WSIs), were resized to 512 x 512 px at 0.5 MPP, tessellated, and filtered to include only pathologist-annotated tumor regions. The NCT-CRC-HE-100k dataset includes 100,000 Macenko-normalized tiles from nine tissue classes. The TCGA Pan-cancer dataset contains tumor-region tiles from 32 solid cancer types. All datasets, except NCT-CRC-HE-100k, which had a separate test set, CRC-VAL-HE-7K, were divided into training and test sets based on patient-level splits. **c,** MoPaDi is trained simultaneously as a feature extractor and a denoising diffusion implicit model (DDIM), which is conditioned on feature vectors and the original image encoded into noise, enabling the reconstruction of original images. **d,** In

the linear approach, a logistic regression classifier is trained on tile-level feature vectors to define the decision boundary used for classifier-guided counterfactual generation. Binary classification is shown here for illustrative purposes; the method generalizes to multiclass settings using a one-vs-rest approach. **e,** In the multiple instance learning (MIL) approach, MoPaDi learns patient-level predictions from bags of tile features using a transformer-based classifier with attention-based pooling. **f,** For both classifier types, counterfactual images are generated by manipulating the feature vector of a selected tile with a manipulation amplitude α. In the linear case, f follows the classifier weight vector v; in the MIL case, f is computed as the gradient of the loss $\mathcal{L}$ with respect to the tile's feature vector $\mathbf{z}$. The denormalized manipulated feature vector is then used to condition DDIM, yielding a counterfactual image. **g**, Overview of the qualitative evaluation study. For linear classifiers (left), 15 class transitions were randomly selected by manipulating individual tiles from their original class to a counterfactual image of the target class. For MIL classifiers (right), the 2 most predictive tiles per patient were selected for 8 patients per class, resulting in 16 tiles per class transition, and 32 per classifier. Pathologists were shown the transitions from original to counterfactual images and asked to identify and describe key morphological changes. MLP, multilayer perceptron; MPP, micrometers per pixel.

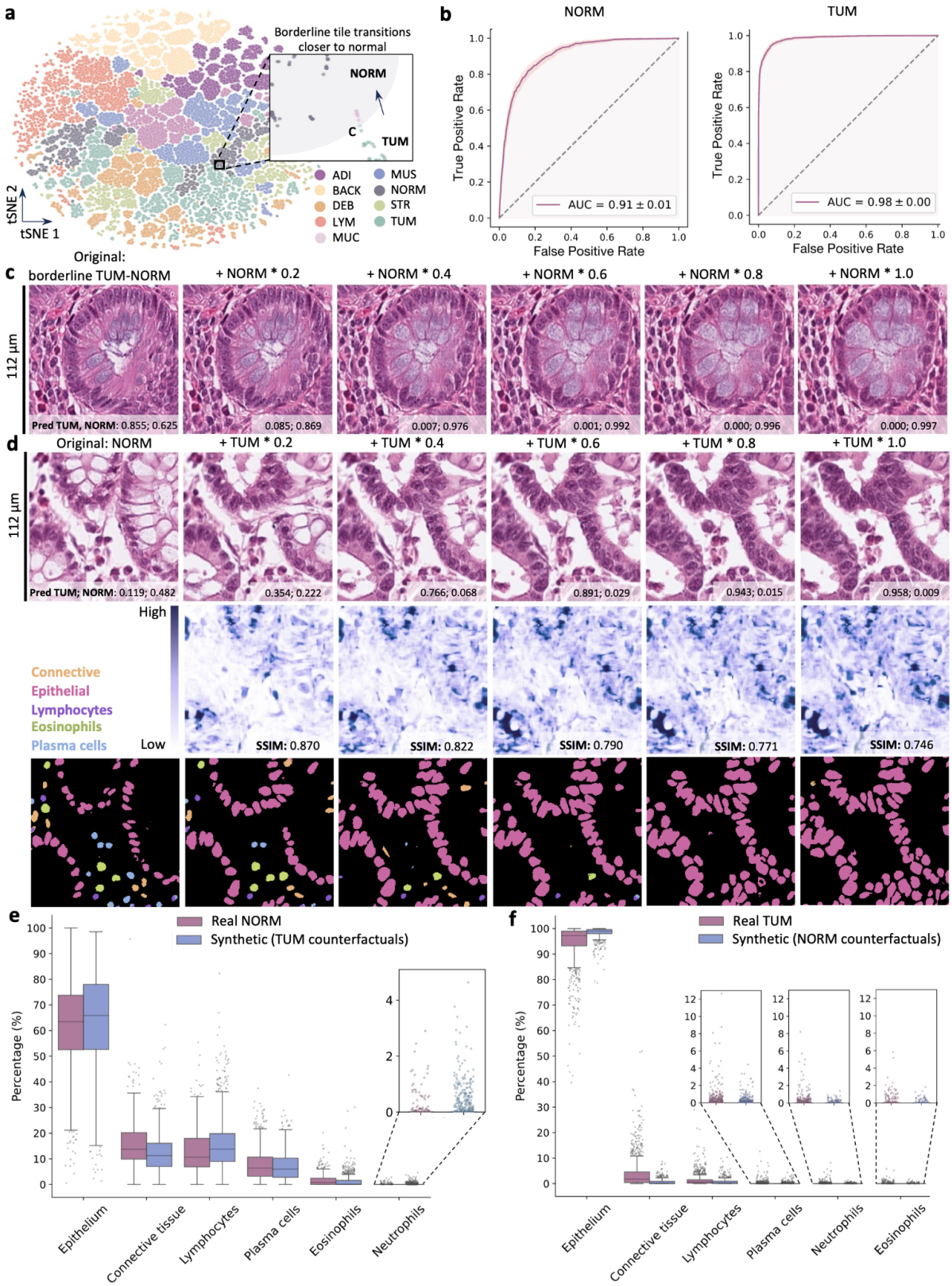

Fig. 2 | Results for counterfactual image generation for the CRC-VAL-HE-7K dataset. a, t-SNE projection of the feature extractor's latent space (training set), colored by class: adipose (ADI), background (BACK), debris (DEB), lymphocytes (LYM), mucus (MUC), smooth muscle (MUS), healthy colon mucosa (NORM), cancer-associated stroma (STR), colorectal adeno-carcinoma epithelium/dysplastic tissue (TUM). The features of dysplastic tile in **c**, initially near

the boundary between TUM (aqua) and NORM (gray), shift toward the healthy cluster (pink points in the zoomed-in plot; gray area – arbitrary healthy mucosa zone) as manipulation amplitude increases. **b,** Receiver operating characteristic (ROC) curves for the logistic regression classifier show areas under the curves (AUCs) greater than 0.9. ROC curves for all other classes are shown in extended data **Fig. S3**. **c,** Tissue-type classifier-guided counterfactuals of a dysplastic tile, generated by shifting its feature vector toward the predicted healthy colon mucosa class. Changes reflect what the model requires to flip the prediction, with increasing manipulation amplitude shown above each image. **d**, Counterfactuals of a healthy mucosa tile shifted toward dysplastic epithelium, rated highly realistic in a blinded observer study. Below: pixel-level difference maps (darker regions indicate greater changes), showing most change in gland regions, and Structural Similarity Index Measure (SSIM) values, which quantify the similarity between the original and manipulated tiles. Bottom row shows segmented and classified nuclei: pink (epithelial), orange (connective tissue), blue (plasma), deep purple (lymphocytes), aqua (neutrophils), green (eosinophils). **e,** Cell type distributions in real NORM tiles (n = 741) and synthetic (counterfactual images for TUM class, n = 1,233). All cell type percentages differed significantly ($p < 0.001$, Bonferroni corrected two-sided Mann-Whitney U test), except plasma cells, indicating the synthetic nature of generated images. **f,** Cell type distributions in real TUM tiles (n = 1,233) and synthetic images (counterfactual images for NORM class, n tiles = 741). All differences were significant ($p < 0.001$), except neutrophils and eosinophils. The model tended to generate more epithelial cells in counterfactual images, as it potentially learned this as a prominent feature of the malignant epithelium class. Additional statistical details for panels **e** and **f** are provided in **Tables S1 and S2**.

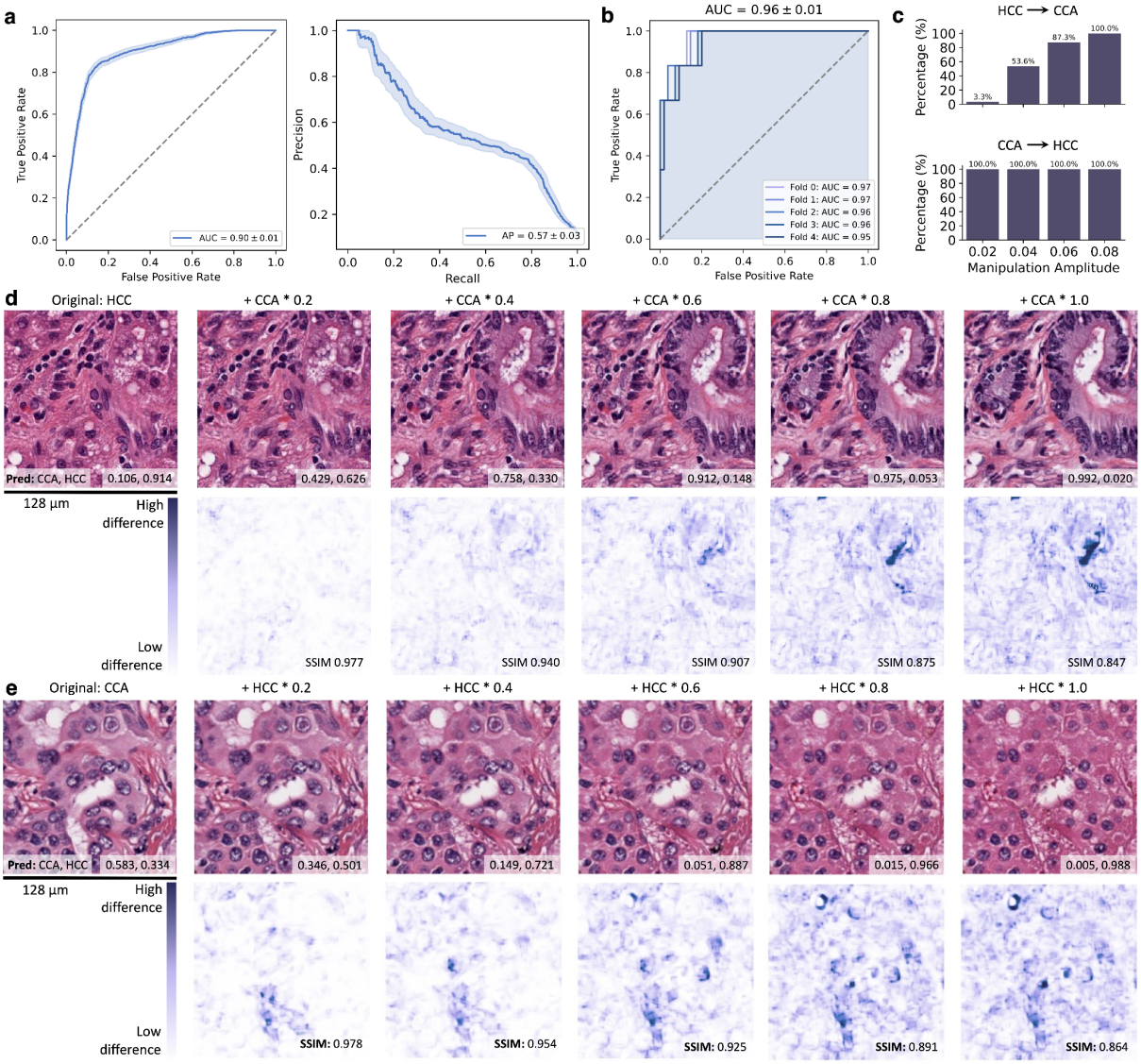

Fig. 3 | Results for the explainability of liver cancer types classifiers. a, Performance of the logistic regression classifier in distinguishing hepatobiliary cancers (hepatocellular carcinoma [HCC], and cholangiocarcinoma [CCA]) **b,** 5-fold cross-validation results of the MIL-based classifier for the same task. **c,** Counterfactual image generation effectiveness, measured as the percentage of generated images, across varying manipulation amplitudes, predicted as the opposite class. HCC to CCA required higher amplitudes to reach similar results, which likely reflects class imbalance in the dataset (**Table S1**). **d,** Representative example of HCC tile transitioning to its counterfactual HCC image generated with a linear approach. Comparison to the MIL approach can be found in the extended data (**Fig. S5**). Difference maps display pixel-wise differences between the original and synthetic tile. **e,** Representative example of CCA tile transitioning to its counterfactual HCC image. AUC, area under the curve; AP, average precision; SSIM, structural similarity index measure.

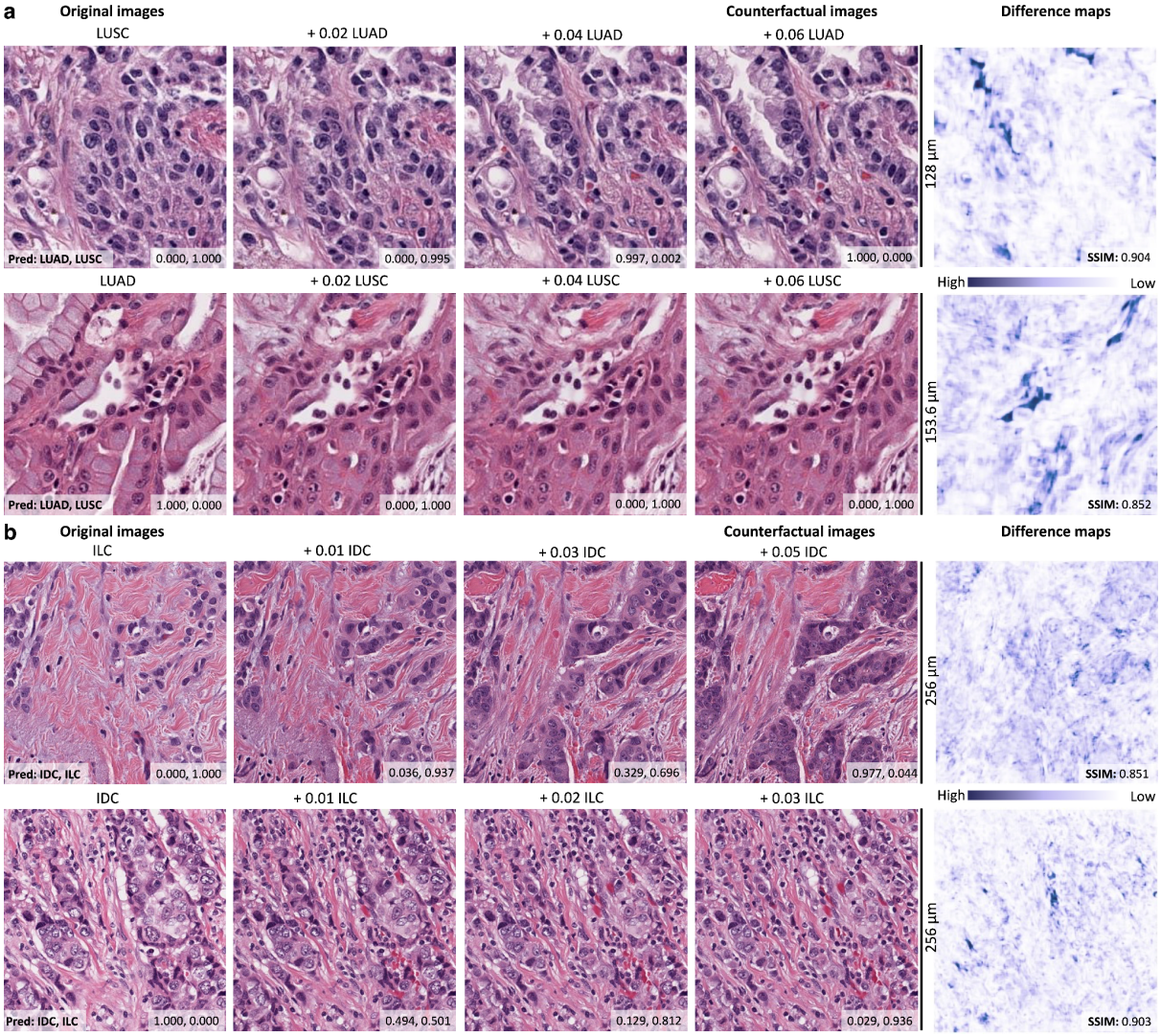

Fig. 4 | Results for explainability of breast and lung cancer type classifiers. Classifiers' performance metrics can be found in the extended data (Fig. S6). **a,** Representative examples of lung squamous cell carcinoma (LUSC) tile transitioning to its counterfactual lung adenocarcinoma (LUAD) image and vice versa. **b,** Representative examples of invasive lobular carcinoma (ILC) tile transitioning to its counterfactual invasive ductal carcinoma (IDC) image and vice versa. Difference maps display pixel-wise differences between the original and synthetic tile. SSIM, structural similarity index measure.

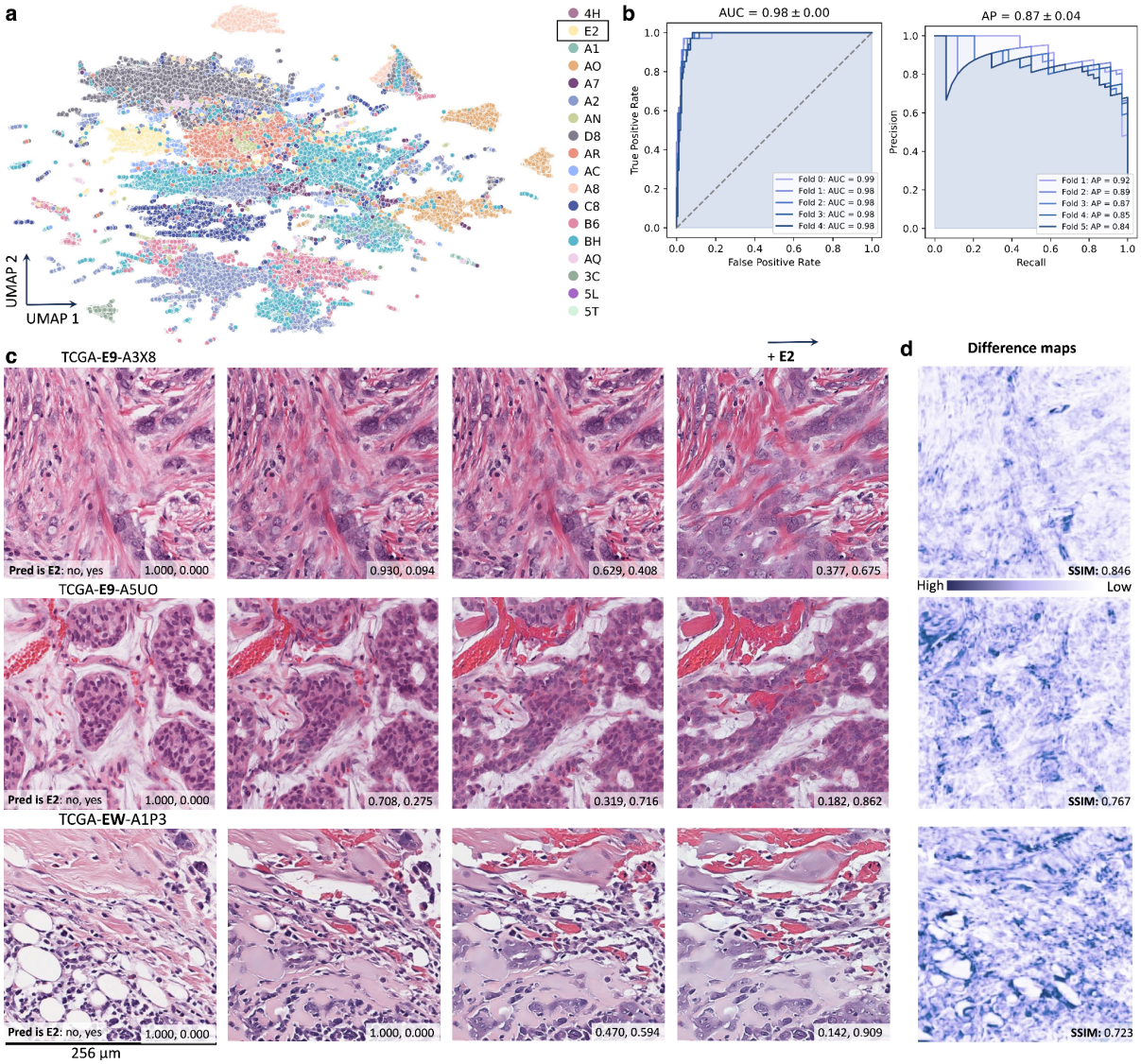

Fig. 5 | Results of the counterfactual image generation to explain batch effects. a, Dimensionality reduced space of TCGA BRCA training set's features according to the center of WSI origin (number of centers = 18). Twenty other centers that contributed less data were only in the test set, while E2 (Roswell Park) was in both for experimentation. **b**, 5-fold cross-validation results of the MIL-based classifier to differentiate between WSI coming from the E2 center vs. the rest. Even with a high class imbalance, the models reach a high AP value. **c,** Representative examples of counterfactual images for different centers. **d,** Pixel-wise difference images comparing the original image with the most manipulated one. The difference maps reveal that the areas of highest change predominantly occur in regions with dense cellular structures and at tissue interfaces. AUC, area under the curve; SSIM – structural similarity index measure.

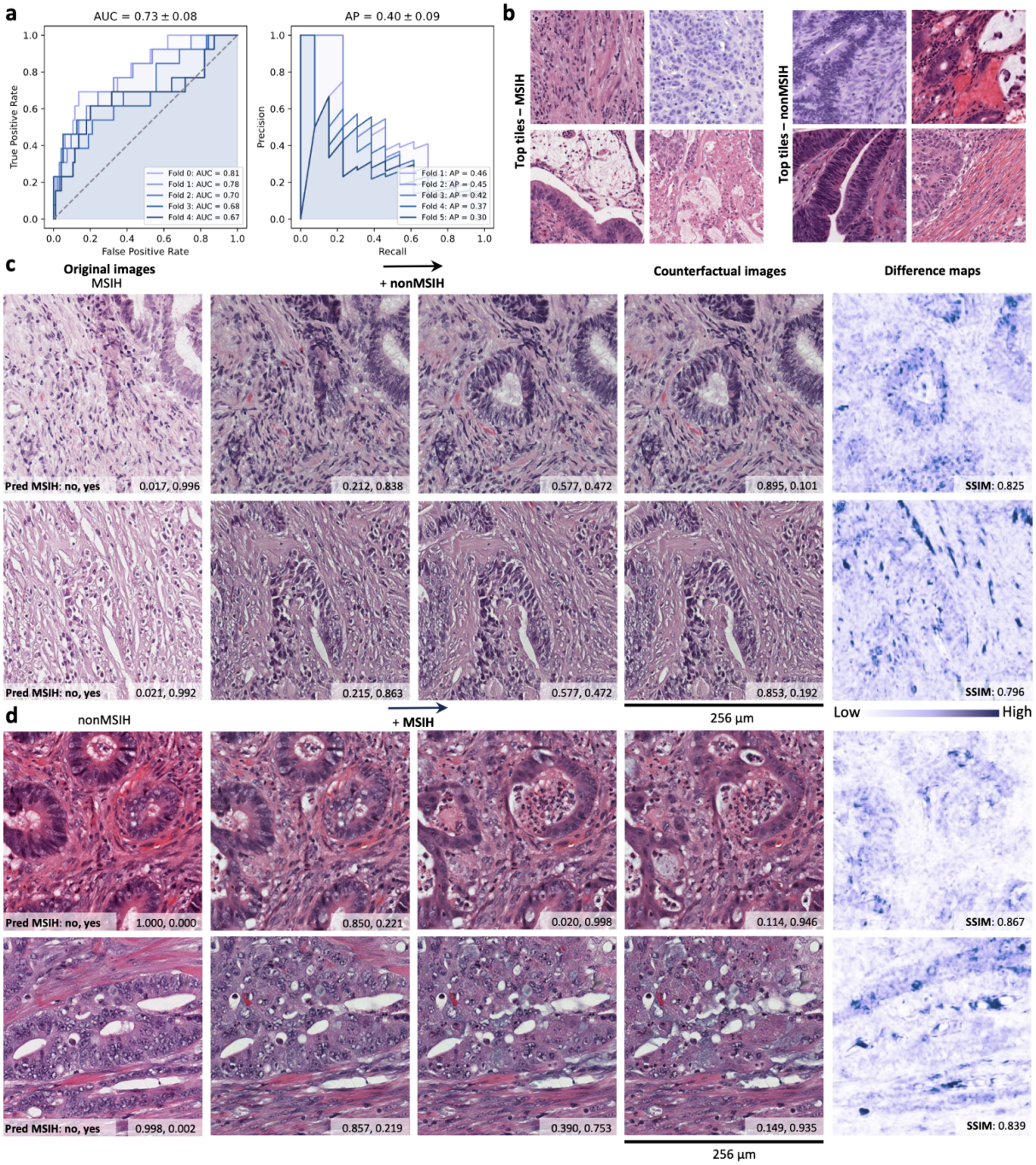

Fig. 6 | Results for the explainability of the classifier for the molecular biomarker in colorectal cancer. a, 5-fold cross-validation results of the MIL-based classifier to differentiate between high microsatellite instability (MSIH) and microsatellite stable (nonMSIH). **b,** The most predictive tiles from randomly selected patients for both classes. **c,** Representative examples of MSIH tiles transitioning to their counterfactual nonMSIH image. Pixel-wise difference images compare the original image with the most manipulated one. **d,** Representative examples of nonMSIH tiles transitioning to their counterfactual MSIH image. Pixel-based difference maps highlight changes in nuclei, stroma, cell density, and tissue architecture. AUC, area under the curve; SSIM – structural similarity index measure.

Abbreviations

AI – Artificial Intelligence

AP – Average precision

AUC – Area under the curve

AUROC – Area under the receiver operating characteristic

BRCA – Breast invasive carcinoma

CCA – Cholangiocarcinoma

CRC – Colorectal carcinoma

DDIM – Denoising diffusion implicit models

DL – Deep learning

FFPE – Formalin-fixed paraffin-embedded

FID – Fréchet inception distance

GAN – Generative adversarial network

H&E – Hematoxylin and eosin

GPU – Graphics processing unit

HCC – Hepatocellular carcinoma

IDC – Invasive ductal carcinoma

ILC – Invasive lobular carcinoma

HCC – Hepatocellular carcinoma

TCGA – The Cancer Genome Atlas

CCA – Cholangiocarcinoma

LUAD – Lung adenocarcinoma

LUSC – Lung squamous cell carcinoma

MIL – Multiple instance learning

MPP – Microns per pixel

MSE – Mean square error

MS-SSIM – Multi-scale structural similarity index measure

MSI – Microsatellite instability

MSIH – High microsatellite instability

NORM – Normal colon mucosa

TCGA – The Cancer Genome Atlas

TUM – Colorectal adenocarcinoma epithelium

WSI – Whole slide image

References

1. Kather, J. N. *et al.* Predicting survival from colorectal cancer histology slides using deep learning: A retrospective multicenter study. *PLoS Med.* **16**, e1002730 (2019).
2. Zeng, Q. *et al.* Artificial intelligence-based pathology as a biomarker of sensitivity to atezolizumab-bevacizumab in patients with hepatocellular carcinoma: a multicentre retrospective study. *Lancet Oncol.* **24**, 1411–1422 (2023).
3. Wagner, S. J. *et al.* Transformer-based biomarker prediction from colorectal cancer histology: A large-scale multicentric study. *Cancer Cell* **41**, 1650–1661.e4 (2023).
4. Calderaro, J., Žigutytė, L., Truhn, D., Jaffe, A. & Kather, J. N. Artificial intelligence in liver cancer - new tools for research and patient management. *Nat. Rev. Gastroenterol. Hepatol.* **21**, 585–599 (2024).
5. Unger, M. & Kather, J. N. Deep learning in cancer genomics and histopathology. *Genome Med.* **16**, 44 (2024).
6. Ghassemi, M., Oakden-Rayner, L. & Beam, A. L. The false hope of current approaches to explainable artificial intelligence in health care. *Lancet Digit. Health* **3**, e745–e750 (2021).
7. Huff, D. T., Weisman, A. J. & Jeraj, R. Interpretation and visualization techniques for deep learning models in medical imaging. *Phys. Med. Biol.* **66**, 04TR01 (2021).
8. Kim, B. *et al.* Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV). in *Proceedings of the 35th International Conference on Machine Learning* (eds. Dy, J. & Krause, A.) vol. 80 2668–2677 (PMLR, 10--15 Jul 2018).
9. Graziani, M., Lompech, T., Müller, H., & Andrearczyk, V. Evaluation and comparison of CNN visual explanations for histopathology. in *Proceedings of the AAAI Conference on Artificial Intelligence Workshops (XAI-AAAI-21), Virtual Event* 8–9.
10. Amorim, J. P. *et al.* Interpreting Deep Machine Learning Models: An Easy Guide for Oncologists. *IEEE Rev. Biomed. Eng.* **16**, 192–207 (2023).

11. Guidotti, R. Counterfactual explanations and how to find them: literature review and benchmarking. *Data Min. Knowl. Discov.* (2022) doi:10.1007/s10618-022-00831-6.
12. Goyal, Y. *et al.* Counterfactual Visual Explanations. in *Proceedings of Machine Learning Research* 2376–2384 (2019).
13. De Sousa Ribeiro, F., Xia, T., Monteiro, M., Pawlowski, N. & Glocker, B. High Fidelity Image Counterfactuals with Probabilistic Causal Models. *arXiv [cs.LG]* (2023).
14. Sanchez, P. *et al.* Causal machine learning for healthcare and precision medicine. *R Soc Open Sci* **9**, 220638 (2022).
15. Sanchez, P., Kascenas, A., Liu, X., O'Neil, A. Q. & Tsaftaris, S. A. What is Healthy? Generative Counterfactual Diffusion for Lesion Localization. in *Deep Generative Models* 34–44 (Springer Nature Switzerland, 2022). doi:10.1007/978-3-031-18576-2_4.
16. Han, T. *et al.* Reconstruction of patient-specific confounders in AI-based radiologic image interpretation using generative pretraining. *Cell Reports Medicine* **5**, (2024).
17. Dolezal, J. M. *et al.* Deep learning generates synthetic cancer histology for explainability and education. *NPJ Precis Oncol* **7**, 49 (2023).
18. Naglah, A., Khalifa, F., El-Baz, A. & Gondim, D. Conditional GANs based system for fibrosis detection and quantification in Hematoxylin and Eosin whole slide images. *Med. Image Anal.* **81**, 102537 (2022).
19. Wölflein, G., Um, I. H., Harrison, D. J. & Arandjelović, O. HoechstGAN: Virtual lymphocyte staining using generative adversarial networks. *arXiv [cs.CV]* 4997–5007 (2022).
20. DeGrave, A. J., Cai, Z. R., Janizek, J. D., Daneshjou, R. & Lee, S.-I. Auditing the inference processes of medical-image classifiers by leveraging generative AI and the expertise of physicians. *Nat Biomed Eng* (2023) doi:10.1038/s41551-023-01160-9.
21. Singla, S., Eslami, M., Pollack, B., Wallace, S. & Batmanghelich, K. Explaining the black-box smoothly-A counterfactual approach. *Med. Image Anal.* **84**, 102721 (2023).
22. Lang, O. *et al.* Using generative AI to investigate medical imagery models and datasets. *EBioMedicine* **102**, 105075 (2024).
23. Howard, F. M. *et al.* Generative adversarial networks accurately reconstruct pan-cancer

histology from pathologic, genomic, and radiographic latent features. *Sci. Adv.* **10**, eadq0856 (2024).

24. Dhariwal, P. & Nichol, A. Diffusion models beat gans on image synthesis. *Adv. Neural Inf. Process. Syst.* **34**, 8780–8794 (2021).
25. Song, J., Meng, C. & Ermon, S. Denoising Diffusion Implicit Models. *arXiv [cs.LG]* (2020).
26. Preechakul, K., Chatthee, N., Wizadwongsa, S. & Suwajanakorn, S. Diffusion autoencoders: Toward a meaningful and decodable representation. *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.* 10609–10619 (2021) doi:10.1109/CVPR52688.2022.01036.
27. Augustine, T. N. Weakly-supervised deep learning models in computational pathology. *EBioMedicine* **81**, 104117 (2022).
28. Gadermayr, M. & Tschuchnig, M. Multiple instance learning for digital pathology: A review of the state-of-the-art, limitations & future potential. *Comput. Med. Imaging Graph.* **112**, 102337 (2024).
29. Kather, J. N. *et al.* Deep learning can predict microsatellite instability directly from histology in gastrointestinal cancer. *Nat. Med.* **25**, 1054–1056 (2019).
30. Monteiro, M., Ribeiro, F. D. S., Pawlowski, N., Castro, D. C. & Glocker, B. Measuring axiomatic soundness of counterfactual image models. *arXiv [cs.CV]* (2023).
31. Howard, F. M. *et al.* The impact of site-specific digital histology signatures on deep learning model accuracy and bias. *Nat. Commun.* **12**, 4423 (2021).
32. Cai, Y. *et al.* An atlas of genetic effects on cellular composition of the tumor microenvironment. *Nat. Immunol.* **25**, 1959–1975 (2024).
33. Hamilton, S. R. Carcinoma of the colon and rectum. *Pathology and genetics of tumors of digestive system* (2000).
34. Travis, W. D. Pathology of lung cancer. *Clin. Chest Med.* **32**, 669–692 (2011).
35. Schlageter, M., Terracciano, L. M., D'Angelo, S. & Sorrentino, P. Histopathology of hepatocellular carcinoma. *World J. Gastroenterol.* **20**, 15955–15964 (2014).

36. Vijgen, S., Terris, B. & Rubbia-Brandt, L. Pathology of intrahepatic cholangiocarcinoma. *Hepatobiliary Surg. Nutr.* **6**, 22–34 (2017).

37. Jenkins, M. A. *et al.* Pathology features in Bethesda guidelines predict colorectal cancer microsatellite instability: a population-based study. *Gastroenterology* **133**, 48–56 (2007).

38. Shia, J. *et al.* Morphological characterization of colorectal cancers in The Cancer Genome Atlas reveals distinct morphology-molecular associations: clinical and biological implications. *Mod. Pathol.* **30**, 599–609 (2017).

39. Puccini, A. *et al.* Molecular profiling of signet-ring-cell carcinoma (SRCC) from the stomach and colon reveals potential new therapeutic targets. *Oncogene* **41**, 3455–3460 (2022).

40. Reitsam, N. G. *et al.* Concurrent loss of MLH1, PMS2 and MSH6 immunoexpression in digestive system cancers indicating a widespread dysregulation in DNA repair processes. *Front. Oncol.* **12**, 1019798 (2022).

41. Remo, A. *et al.* Morphology and molecular features of rare colorectal carcinoma histotypes. *Cancers (Basel)* **11**, 1036 (2019).

42. Kheiri, F., Rahnamayan, S., Makrehchi, M. & Asilian Bidgoli, A. Investigation on potential bias factors in histopathology datasets. *Sci. Rep.* **15**, 11349 (2025).

43. Dehkharghanian, T. *et al.* Biased data, biased AI: deep networks predict the acquisition site of TCGA images. *Diagn. Pathol.* **18**, 67 (2023).

44. Kather, J. N., Halama, N., Marx, A. 100,000 histological images of human colorectal cancer and healthy tissue. https://doi.org/10.5281/zenodo.1214456 (2018).

45. Macenko, M. *et al.* A method for normalizing histology slides for quantitative analysis. in *2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro* 1107–1110 (IEEE, 2009). doi:10.1109/ISBI.2009.5193250.

46. Komura, D. & Ishikawa, S. Histology images from uniform tumor regions in TCGA Whole Slide Images. *In Cell Reports (1.0, Vol. 38, Number 9, p. 110424)* https://doi.org/10.5281/zenodo.5889558 (2021).

47. Elfwing, S., Uchibe, E. & Doya, K. Sigmoid-weighted linear units for neural network

function approximation in reinforcement learning. *Neural Netw.* **107**, 3–11 (2018).

48. Wang, Z., Simoncelli, E. P. & Bovik, A. C. Multiscale structural similarity for image quality assessment. in *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003* vol. 2 1398–1402 Vol.2 (IEEE, 2003).
49. Wang, Z., Bovik, A. C., Sheikh, H. R. & Simoncelli, E. P. Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process.* **13**, 600–612 (2004).
50. Parmar, G., Zhang, R. & Zhu, J.-Y. On aliased resizing and surprising subtleties in GAN evaluation. *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.* 11400–11410 (2022) doi:10.1109/CVPR52688.2022.01112.
51. Ignatov, A., Yates, J. & Boeva, V. Histopathological image classification with cell morphology aware deep neural networks. (2024).

Acknowledgments

The authors gratefully acknowledge the GWK's support for funding this project by providing computing time through the Center for Information Services and HPC (ZIH) at TU Dresden. The authors also gratefully acknowledge the Gauss Centre for Supercomputing e.V. (www.gauss-centre.eu) for funding this project by providing computing time through the John von Neumann Institute for Computing (NIC) on the GCS Supercomputer JUWELS at Jülich Supercomputing Centre (JSC). Finally, the authors acknowledge the TCGA Research Network (https://www.cancer.gov/tcga), which generated the data on which the results shown in this manuscript are partly based.

Author Contributions

LZ, JNK, and DT conceptualized the study. LZ and TL performed the experiments and developed the software. LZ, KJH, NGR, and SF performed results analysis. LZ prepared the original draft and created the figures. JNK acquired funding and supervised the study. All authors provided scientific input, reviewed and edited the manuscript.

Competing Interests

NGR received compensation for travel expenses from nanoString (Bruker Company). **SF** has received honoraria from MSD and BMS. **DT** received honoraria for lectures by Bayer, GE, and Philips and holds shares in StratifAI GmbH, Germany, and in Synagen GmbH, Germany. **AP** has received grant/research funding from AbbVie, Kura Oncology, and Merck and is on the advisory board for AbbVie, Elevar, Ayala, Prelude, and ThermoFisher. **JNK** declares consulting services for Bioptimus, Panakeia, AstraZeneca, and MultiplexDx. Furthermore, he holds shares in StratifAI, Synagen, Tremont AI and Ignition Labs, and has received an institutional research grant by GSK, and has received honoraria by AstraZeneca, Bayer, Daiichi Sankyo, Eisai, Janssen, Merck, MSD, BMS, Roche, Pfizer, and Fresenius. The remaining authors have no competing interests to declare.

Funding

NGR received funding from the Manfred-Stolte-Stiftung and the Bavarian Cancer Research Center (BZKF). **SF** is supported by the German Federal Ministry of Education and Research (SWAG, 01KD2215C), the German Cancer Aid (DECADE, 70115166 and TargHet, 70115995) and the German Research Foundation (504101714). **DT** is supported by the German Federal Ministry of Education (TRANSFORM LIVER, 031L0312A; SWAG, 01KD2215B),

Deutsche Forschungsgemeinschaft (DFG) (TR 1700/7-1), and the European Union (Horizon Europe, ODELIA, GA 101057091). **JNK** is supported by the German Cancer Aid DKH (DEC-ADE, 70115166), the German Federal Ministry of Research, Technology and Space BMFTR (PEARL, 01KD2104C; CAMINO, 01EO2101; TRANSFORM LIVER, 031L0312A; TANGE-RINE, 01KT2302 through ERA-NET Transcan; Come2Data, 16DKZ2044A; DEEP-HCC, 031L0315A; DECIPHER-M, 01KD2420A; NextBIG, 01ZU2402A), the German Research Foundation DFG (CRC/TR 412, 535081457), the German Academic Exchange Service DAAD (SECAI, 57616814), the German Federal Joint Committee G-BA (TransplantKI, 01VSF21048), the European Union EU's Horizon Europe research and innovation programme (ODELIA, 101057091; GENIAL, 101096312), the European Research Council ERC (NADIR, 101114631), the National Institutes of Health NIH (EPICO, R01 CA263318) and the National Institute for Health and Care Research NIHR (Leeds Biomedical Research Centre, NIHR203331). The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health and Social Care. This work was funded by the European Union. Views and opinions expressed are, however, those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.