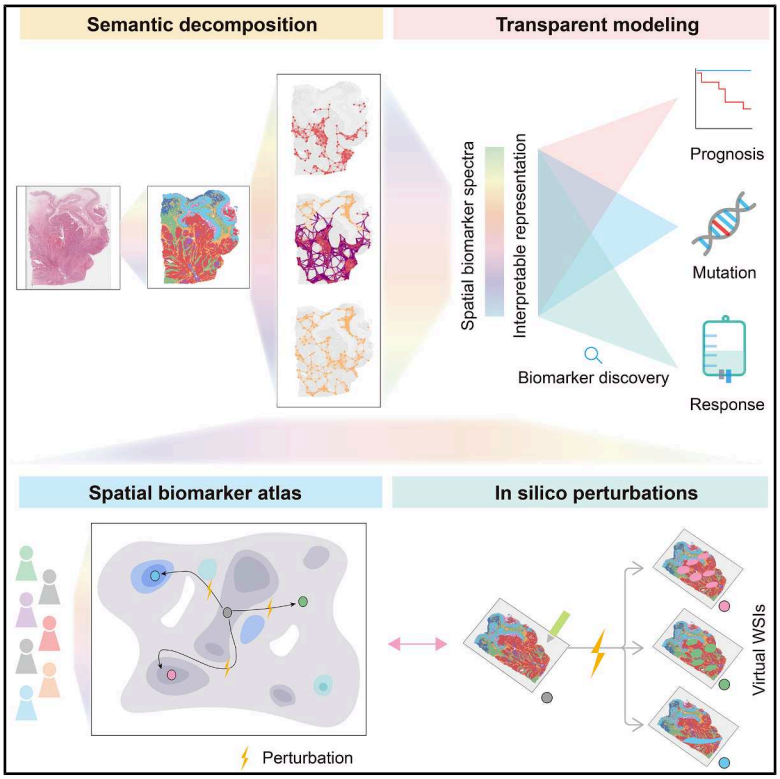
Cancer Cell

Article

Spatial biomarker discovery via interpretable semantic learning in histopathology

Graphical abstract



Highlights

- PathPrism converts WSIs into interpretable spatial biomarker spectra
- Spatial biomarkers enable transparent modeling of clinical outcomes
- Transparent modeling enables active, closed-loop biomarker discovery
- VirtualWSI enables the perturbation-driven exploration of spatial biology

Authors

Junhao Liang (梁君浩), Xiaofeng Jiang (江晓锋), Nic Gabriel Reitsam, ..., Magali Svrcek, Lingjie Kong (孔令杰), Jakob Nikolas Kather

Correspondence

konglj@tsinghua.edu.cn (L.K.), jakob_nikolas.kather@tu-dresden.de (J.N.K.)

In brief

Liang et al. introduce PathPrism, a framework that converts whole-slide images into interpretable spatial biomarker spectra, establishing a transparent representation of tissue architecture. This representation enables transparent linear modeling of prognosis, molecular alterations, and treatment response with high performance, while transforming histopathology into a platform for interpretable discovery and perturbation-driven exploration.

Liang et al., 2026, Cancer Cell *44*, 1605–1622
August 10, 2026 © 2026 The Authors. Published by Elsevier Inc.
https://doi.org/10.1016/j.ccell.2026.05.014

CellPress

Cancer Cell

CellPress
OPEN ACCESS

Article

Spatial biomarker discovery via interpretable semantic learning in histopathology

Junhao Liang (梁君浩),[1,2] Xiaofeng Jiang (江晓锋),[3] Nic Gabriel Reitsam,[1,4,5] Tim Lenz,[1] Li Zhang (张莉),[1] Marco Gustav,[1] Zunamys Itzell Carrero,[1] Hannah Sophie Muti,[1,31,32,33] Peter Neidlinger,[1] Jan Clusmann,[1,33,36] Laura Žigutytė,[1] Vidhya Sainath,[1] Fabian Wolf,[1] Lisa A. Boardman,[6] Amy J. French,[7] Ellen L. Goode,[8] Andrea Gsur,[9] Stefanie Brezina,[9] Marc J. Gunter,[10,11] Robert Steinfelder,[12] Tabitha Harrison,[12,13] Ulrike Peters,[12,13] Amanda I. Phipps,[12,13] Philip Quirke,[14] Nicholas P. West,[14] Michael Hoffmeister,[15,34] Hermann Brenner,[15,16,17] Durgesh Wankhede,[34] Jitendra Jonnagaddala,[18] Nicholas Hawkins,[19] Robyn L. Ward,[20] Elena Fountzilas,[21,37] Kyriaki Papadopoulou,[22] George Fountzilas,[22,35] Thierry André,[23] Julien Taieb,[24] Jean-François Emile,[25] Magali Svrcek,[26,27] Lingjie Kong (孔令杰),[2,28,*] and Jakob Nikolas Kather[1,14,29,30,38,*]

[1]Else Kroener Fresenius Center for Digital Health, Faculty of Medicine, TUD Dresden University of Technology, Dresden, Germany
[2]State Key Laboratory of Precision Measurement Technology and Instruments, Department of Precision Instrument, Tsinghua University, Beijing 100084, China
[3]Department of Thoracic Surgery, Sichuan Clinical Research Center for Cancer, Sichuan Cancer Hospital & Institute, Sichuan Cancer Center, University of Electronic Science and Technology of China (UESTC), Chengdu, China
[4]Pathology, Faculty of Medicine, University of Augsburg, Augsburg, Germany
[5]Bavarian Center for Cancer Research, BZKF, Augsburg, Germany
[6]Division of Gastroenterology and Hepatology, Mayo Clinic, Rochester, MN, USA
[7]Division of Laboratory Genetics, Department of Laboratory Medicine and Pathology, Mayo Clinic, Rochester, MN, USA
[8]Department of Quantitative Health Sciences, Division of Epidemiology, Mayo Clinic, Rochester, MN, USA
[9]Center for Cancer Research, Medical University of Vienna, Vienna, Austria
[10]Nutrition and Metabolism Branch, International Agency for Research on Cancer, World Health Organization, Lyon, France
[11]Cancer Epidemiology and Prevention Research Unit, School of Public Health, Imperial College London, London, UK
[12]Division of Public Health Sciences, Fred Hutchinson Cancer Center, Seattle, WA, USA
[13]Department of Epidemiology, University of Washington, Seattle, WA, USA
[14]Pathology & Data Analytics, Leeds Institute of Medical Research at St James's, University of Leeds, Leeds, UK
[15]Division of Clinical Epidemiology and Aging Research, German Cancer Research Center (DKFZ), Heidelberg, Germany
[16]Cancer Prevention Graduate School, German Cancer Research Center (DKFZ), Heidelberg, Germany
[17]German Cancer Consortium (DKTK), German Cancer Research Center (DKFZ), Heidelberg, Germany
[18]School of Population Health, Faculty of Medicine and Health, UNSW Sydney, Kensington, Sydney, NSW, Australia
[19]School of Medical Sciences, Faculty of Medicine and Health, UNSW Sydney, Sydney, NSW, Australia
[20]Faculty of Medicine and Health, The University of Sydney, Camperdown, Sydney, NSW, Australia
[21]Department of Medical Oncology, St Luke's Clinic, Thessaloniki, Greece
[22]Laboratory of Molecular Oncology, Hellenic Foundation for Cancer Research, Thessaloniki, Greece
[23]Medical Oncology, Institut Universitaire de Cancérologie AP-HP Sorbonne Université - Hôpital Saint-Antoine, Paris, France
[24]Department of Gastroenterology and Digestive Oncology, European Georges Pompidou Hospital, Assistance Publique Hôpitaux de Paris, University of Paris Cité, 75015 Paris, France
[25]Paris-Saclay University, Versailles SQY University, EA4340-BECCOH, Assistance Publique-Hôpitaux de Paris, Ambroise-Paré Hospital, Smart Imaging, Service de Pathologie, Boulogne, France
[26]Sorbonne Université, LIMICS, INSERM UMRS 1142, Paris, France
[27]Léon Bérard Comprehensive Cancer Center, Department of Pathology, Lyon, France
[28]IDG/McGovern Institute for Brain Research, Tsinghua University, Beijing, China
[29]Department of Medicine I, Faculty of Medicine, TUD Dresden University of Technology, Dresden, Germany
[30]Medical Oncology, National Center for Tumor Diseases (NCT), University Hospital Heidelberg, Heidelberg, Germany
[31]Department for Visceral, Thoracic and Vascular Surgery, University Hospital and Faculty of Medicine Carl Gustav Carus, Technische Universität Dresden, Dresden, Germany
[32]National Center for Tumor Diseases Dresden (NCT/UCC), A Partnership Between DKFZ, Faculty of Medicine and University Hospital Carl Gustav Carus, TU Dresden University of Technology, and Helmholtz-Zentrum Dresden - Rossendorf (HZDR), Dresden, Germany
[33]German Cancer Consortium (DKTK), Partner Site Dresden, German Cancer Research Center (DKFZ), Heidelberg, Germany
[34]Division of Clinical Epidemiology of Early Cancer Detection, German Cancer Research Center (DKFZ), Heidelberg, Germany
[35]Aristotle University of Thessaloniki, Thessaloniki, Greece
[36]Department of Internal Medicine III, Gastroenterology, Metabolic Diseases and Intensive Care, University Hospital RWTH Aachen, Aachen, Germany
[37]European University Cyprus, Engomi, Cyprus
[38]Lead contact
*Correspondence: konglj@tsinghua.edu.cn (L.K.), jakob_nikolas.kather@tu-dresden.de (J.N.K.)
https://doi.org/10.1016/j.ccell.2026.05.014

This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

SUMMARY

Spatial biomarkers are critical for precision oncology but remain challenging to systematically discover due to the complexity of whole-slide images. We present PathPrism, an interpretable AI framework for spatial biomarker discovery and virtual experimentation. Unlike black-box models, PathPrism encodes tissue architecture into pathologically informed spatial features, enabling transparent modeling of prognosis, molecular alterations, and therapy response. Applied to 7,000 patients with colorectal cancer across 11 cohorts, PathPrism uncovered hundreds of biomarkers predictive of survival, MSI, *BRAF*, and *TP53* mutations, and stratified chemotherapy benefit in stage II/III disease. Building on these interpretable findings, PathPrism uses large language models as auxiliary tools to generate hypotheses grounded in spatial semantics. We further introduce VirtualWSI, a platform for semantic perturbation within an interpretable spatial biomarker atlas. PathPrism provides a scalable and interpretable framework for spatial biomarker discovery.

INTRODUCTION

Histopathological biomarkers are central to cancer diagnosis, prognosis estimation, and treatment selection.[1–3] Given its accessibility and routine clinical use, histopathology offers substantial translational potential.[4–6] Yet biomarker discovery remains labor-intensive and reliant on expert interpretation, requiring meticulous review of large-scale whole-slide images (WSIs) and extensive validation.[3] Spatial biomarkers, defined by multiscale and multimodal patterns within the tumor microenvironment (TME), hold particular promise but are difficult to discover because tissue heterogeneity, complex spatial organization, and subtle visual patterns often exceed the capacity of human interpretation.[7–9] As a result, only a small number of spatial biomarkers have been clinically validated, leaving much of cancer's spatial complexity unexplored.[10] This mismatch between abundant histopathological data and limited human throughput motivates high-throughput, quantitative, and interpretable approaches to spatial biomarker discovery.

Artificial intelligence (AI) for scientific discovery,[11] centered on extracting patterns from complex data and translating them into testable hypotheses, has already driven advances in mathematics,[12] genomics,[13,14] and antibiotic discovery.[15,16] However, in computational pathology, most deep learning (DL) models remain black boxes due to the massive visual information embedded in WSIs and the high complexity of model parameters, offering limited insight into the features driving their predictions.[17,18] Recent studies have attempted to derive candidate biomarkers from DL models using post hoc explainability.[19–23] However, because these models rely on raw pixels or latent image representations, post hoc explanations can be diffuse and ambiguous, making it difficult to determine what the model has learned and to define, quantify, characterize, and validate the resulting signals as pathologically interpretable biomarkers.[17,18,24] This limits clinical translation and large language model (LLM)-assisted hypothesis generation,[25–28] and hinders the transition from prediction to active exploration through manipulable perturbation.[29–33] Here we present PathPrism, an AI-driven framework for interpretable spatial biomarker discovery, coupled with structure-preserving semantic perturbation to explore the spatial biomarker atlas.

RESULTS

Overview of PathPrism

The core principle of PathPrism is to maximize interpretability without compromising performance, achieved through biologically informed input representations and a transparent, simplified model design (Figure 1). To deconstruct the visual complexity of WSIs into semantically defined colorectal cancer (CRC) tissue types, PathPrism first acts like a prism, producing segmentation maps that are quantitatively defined and pathologically interpretable (Figure 1A). While expert annotation can provide such segmentation, it is labor-intensive, subjective, and infeasible at scale. To address this, we developed PrismNet, a DL model trained on 100,000 annotated CRC histopathology patches,[34,35] enabling high-throughput semantic segmentation of WSIs (Figure S1A, Methods). Specifically, PrismNet segments WSIs into eight key tissue types commonly observed in CRC[34]: adipose tissue (ADI), debris (DEB), lymphocytes (LYM), mucus (MUC), smooth muscle (MUS), normal colon mucosa (NORM), cancer-associated stroma (STR), and colorectal adenocarcinoma epithelium (TUM).

To determine whether these semantic segmentation maps preserve clinically relevant prognostic information, we constructed MacroNet, a DL model trained to associate macroscopic tissue architecture with patient survival (Figures 1B and S2A, Methods). Multi-angle interpretation of the trained MacroNet revealed predictive spatial patterns.[31,32] Guided by these insights, PathPrism performs a second-level decomposition of the semantic segmentation maps into quantified, human-interpretable spatial features. Analogous to how a prism further refracts light into a spectrum, this process represents tissue organization through pathology-grounded semantic features, including tissue spatial fractions, spatial entropy, and slide-level graph features (Figures 1C and S4A, Methods). Using this spectrum of interpretable spatial biomarkers, PathPrism applies minimal, transparent models to learn associations with prognosis, molecular alterations, and treatment response (Figure 1D). This design enables direct attribution of predictions to specific biomarkers, supporting pathologically grounded discovery. The resulting biomarkers further support LLM-assisted hypothesis generation (Figure 1E) and *in silico* semantic perturbation (Figure 1F).

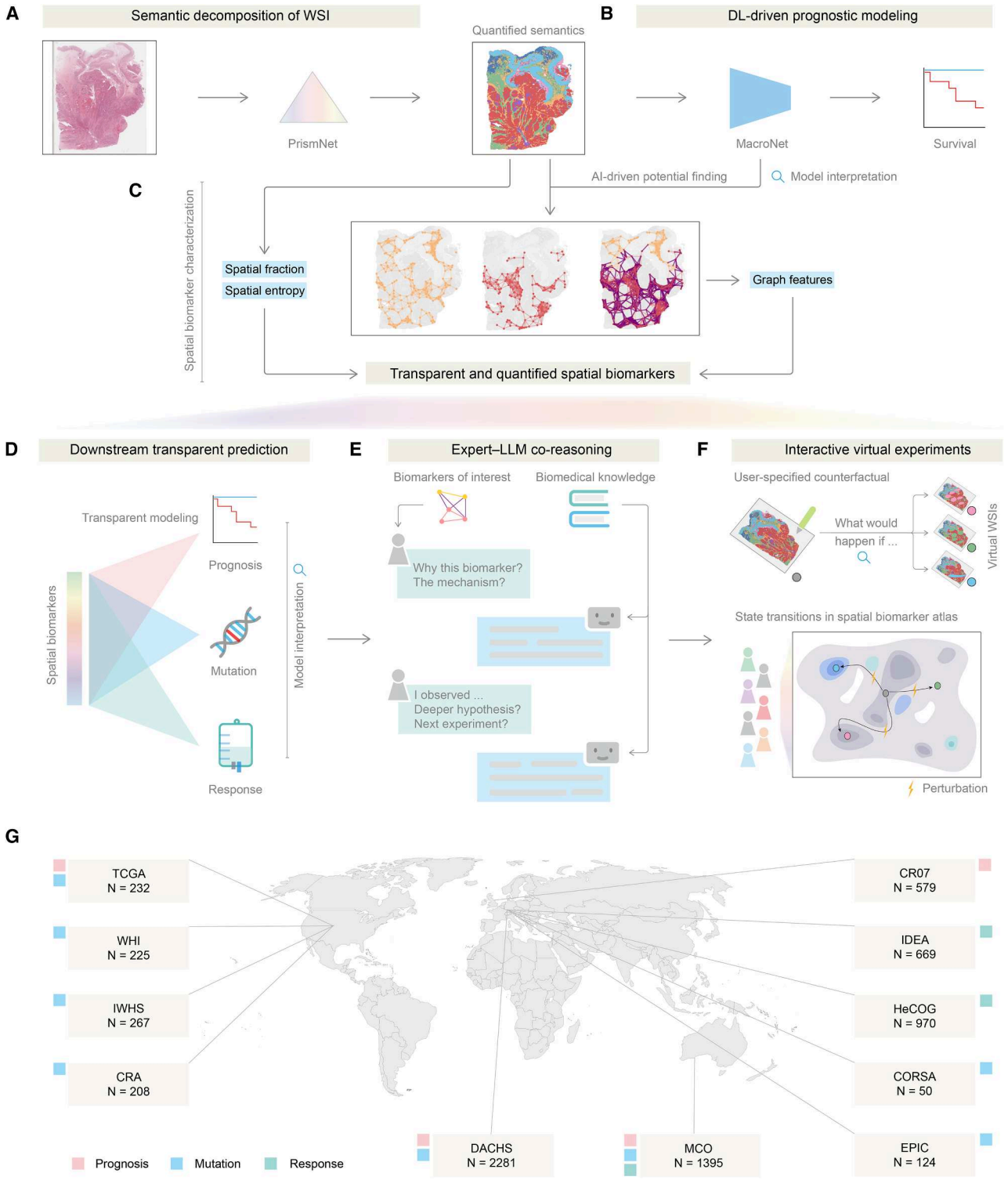


Figure 1. Overview of the PathPrism framework and study cohorts

(A) PrismNet decomposes CRC WSIs into semantic tissue maps from annotated patches.

(B) MacroNet predicts prognosis from semantic maps and enables interpretation of predictive spatial patterns.

(C) MacroNet-guided feature extraction generates interpretable spatial biomarkers, including tissue fractions, spatial entropy, and graph-based features.

(D) Transparent models use these biomarkers to predict prognosis, mutation status, and therapy response, with direct feature-level attribution.

(legend continued on next page)

PathPrism efficiently links semantics and prognosis

To validate the accuracy of histopathological semantic quantification and ensure robustness for downstream analyses, we evaluated PrismNet on the CRC100K dataset. Systematic benchmarking of multiple foundation models and dimensionality reduction strategies identified the combination of UNI features with PCA as the optimal configuration, effectively mitigating overfitting while maintaining high discriminative capacity (Figures S1B and S1C, Methods). Under this setting, PrismNet achieved a macro-average F1 score of 0.948 (95% CI: 0.942–0.954), a Matthews correlation coefficient (MCC) of 0.958 (95% CI: 0.952–0.964), and a macro-averaged AUROC of 0.988 (95% CI: 0.986–0.990) (Figure 2A), supporting its suitability for large-scale, high-fidelity tissue segmentation. Moreover, PrismNet remained robust when trained with partially annotated data, enabling efficient transfer to new cancer types with minimal additional labeling (Figure S1D).

To assess whether these semantic maps preserve clinically relevant information, we trained MacroNet to learn associations between macroscopic tissue architecture and patient disease-specific survival (DSS) directly from the segmentation outputs. Using 5-fold cross-validation on the DACHS cohort, we benchmarked MacroNet against state-of-the-art (SOTA) foundation models, including attMIL-UNI,[36,37] GigaPath,[38] CHIEF,[39] COBRA,[40] and PRISM[41] (Methods). MacroNet achieved a mean concordance index (C-index) of 0.716 ± 0.020, demonstrating performance comparable to other leading approaches (Figure 2B). Substituting PrismNet-derived semantic maps with baseline segmentations reduced prognostic performance, further supporting the integrated PrismNet-MacroNet framework (Figures S2B and S2C). We then evaluated MacroNet on three external cohorts. MacroNet ranked among the top two models in MCO and CR07 (Figure 2C). Kaplan-Meier analysis further showed risk stratification across cohorts, with hazard ratios (HRs) of 3.68 in MCO, 2.48 in CR07, and 3.27 in TCGA (Figure 2D). These results support that macroscopic spatial tissue organization encodes rich prognostic information that can be effectively leveraged through DL.

Spatial biomarker discovery via interpretation

Building on MacroNet's strong performance, we sought to explore what the model had learned. We first stratified patients into high- and low-risk groups based on MacroNet-predicted risk scores and analyzed tissue composition together with model attribution (Figure 3A, Methods). High-risk patients showed increased proportions of TUM, DEB, and STR, whereas low-risk patients were enriched for ADI, LYM, MUS, and NORM (Figure 3B). Saliency-based attribution analysis further showed greater contributions of TUM and MUC in the high-risk group despite comparable tissue proportions,[42] suggesting that MacroNet captured spatial organization beyond abundance alone. LYM and MUS showed both higher abundance and attribution in the low-risk group, consistent with their association with a favorable prognosis. We next analyzed pairwise tissue correlations and co-occurrence patterns using both composition and attribution data (Figure 3C). LYM and STR frequently co-occurred in low-risk cases, and their consistent interaction across tissue composition and attribution suggested a cooperative spatial pattern associated with favorable outcomes.

We then implemented a structure-preserving counterfactual explanation framework to further interpret MacroNet's decision-making[43] (Figure 3D, Methods). Figure 3E shows an example targeting STR, illustrating progressive modulation of STR regions while preserving structure (see also Figure S3). Across the cohort, amplifying LYM consistently reduced risk scores (Figure 3F), reinforcing its association with a favorable prognosis. Stratified analysis further showed that increasing NORM and LYM reduced predicted risk in high-risk patients, whereas increasing MUC, TUM, or STR elevated predicted risk in low-risk patients (Figure 3G). Notably, LYM consistently suppressed risk across all subgroups, aligning with its known protective role in CRC.[44–46]

To further interpret MacroNet's behavior, we integrated semantic maps with pathologist assessment. Visualizations revealed that low-risk samples displayed orderly architecture, with MUS separating TUM and ADI and abundant INF near MUS, while high-risk samples showed disorganized structures, particularly in STR and TUM, with TUM invading beyond MUS into ADI and often lacking INF (Figure 3H). Quantification showed significantly higher spatial entropy in high-risk samples (Figure 3I). Attribution maps revealed consistent focus on the invasive front in both groups (Figures 3J and 3K). In high-risk cases, attribution was fragmented across STR, TUM, and ADI, capturing heterogeneity and aligning with SARIFA-like infiltration patterns[47,48] (Figure 3J). In contrast, low-risk samples showed attribution concentrated along the TUM–MUS boundary and INF zones, where INF appeared to restrict deeper invasion (Figure 3K). Overall, multi-angle interpretation suggested that MacroNet learned prognostically relevant spatial patterns from semantic segmentation maps and identified candidate spatial biomarkers.

Potential biomarker characterization and verification

To complete the discovery-validation cycle, we translated insights from interpreting MacroNet into testable spatial biomarker hypotheses and systematically quantified them using the semantic segmentation maps (Figure 4A; Figure S4A, Methods). We characterized the candidates across spatial fraction, spatial entropy, and graph-based topology. To capture spatial topology, we constructed slide-level graphs that encode contiguous tissue regions and their local proximity or inter-tissue relationships, allowing intra-tissue and inter-tissue characterization of tissue architecture (Figure 4B, Methods; Tables S2 and S3). In total, we compiled 628 interpretable spatial features grounded in predefined semantic tissue maps.

(E) LLM-assisted analysis organizes selected biomarkers into structured hypotheses for expert review.
(F) VirtualWSI performs semantic perturbation for model interrogation within spatial biomarker space.
(G) Study cohorts and analytical tasks. Patient numbers indicate final samples after quality control. Lines indicate approximate regional association, not exact or exclusive location. See also Figures S1, S2 and S11; Table S1.

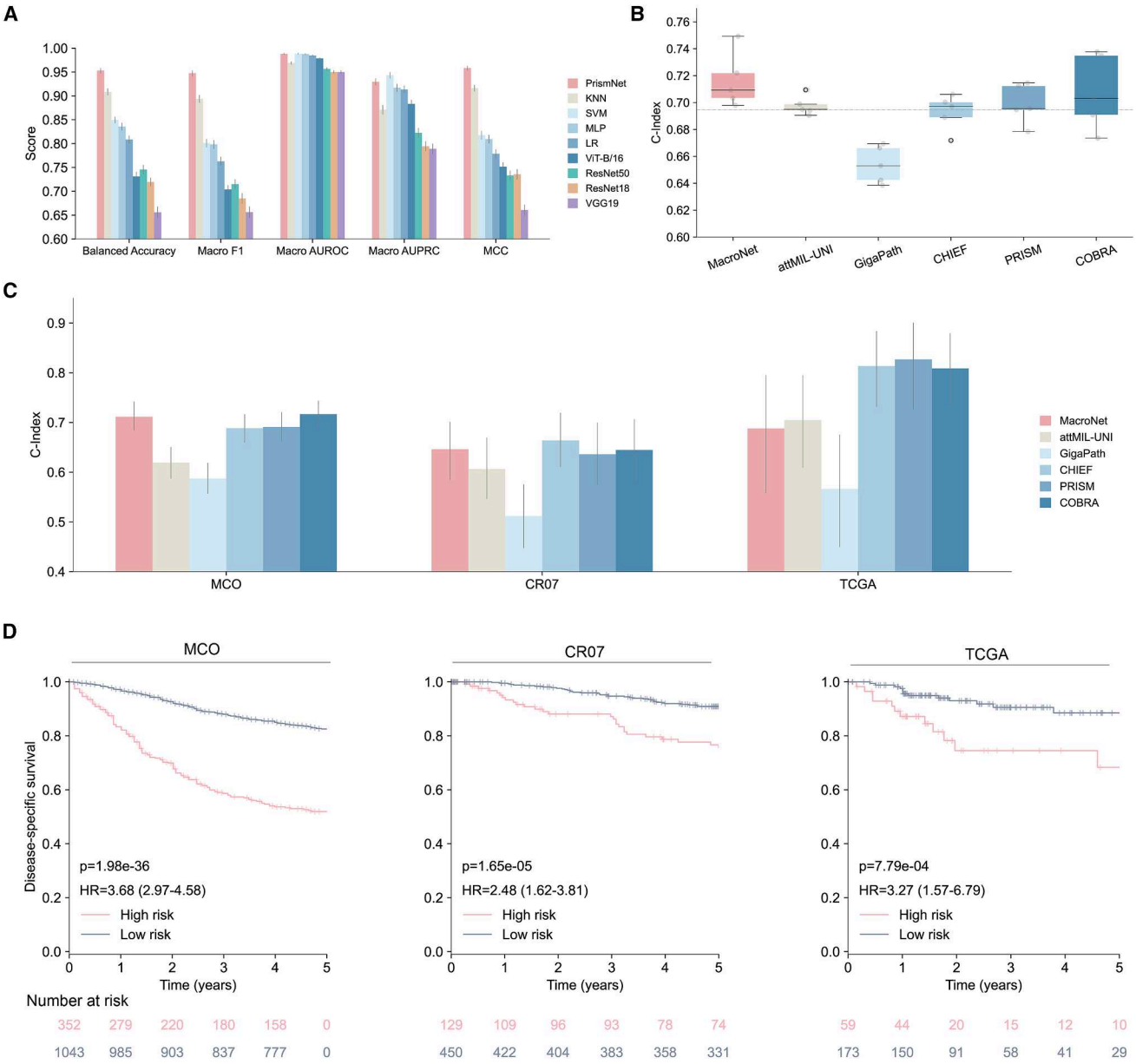


Figure 2. Benchmark evaluation of predictive performance

(A) Classification performance of PrismNet and baseline methods trained on NCT-CRC-HE-100K-NONORM and evaluated on CRC-VAL-HE-7K. PrismNet, KNN, SVM, MLP, and logistic regression (LR) used UNI-encoded features, whereas ViT-B/16, ResNet18, ResNet50, and VGG19 were trained on raw image patches. All models were trained without stain normalization. Data are means; error bars indicate 95% confidence intervals from 1,000 bootstrap samples.

(B) 5-fold cross-validation C-indices on DACHS for MacroNet, attMIL-UNI, GigaPath, CHIEF, PRISM, and COBRA. Boxplots show median, interquartile range, and whiskers extending to 1.5× the interquartile range. The dashed line indicates the mean C-index across models.

(C) Prognostic performance on external cohorts MCO, CR07, and TCGA. Data are means; error bars indicate 95% confidence intervals from 1,000 bootstrap samples.

(D) Kaplan-Meier analysis of MacroNet risk groups in MCO, CR07, and TCGA. Thresholds were defined in DACHS. *p* values were calculated by two-sided log rank tests. See also Figures S1 and S2.

We evaluated these features across four real-world CRC cohorts, including DACHS, MCO, CR07, and TCGA. Among single-tissue biomarkers, several STR- and TUM-related features showed prognostic performance with C-indices exceeding 0.6 and in some cases approaching 0.7 (Figure 4C). Cohort-specific patterns were also observed. DACHS and MCO shared effective STR, TUM, and MUS features, whereas CR07, which includes only rectal cancers with frequent preoperative radiotherapy and low MSI frequency,[49] showed stronger MUS-related signals, and TCGA additionally highlighted MUC and DEB. Feature performance further differed by tissue type. MUS was consistently captured by spatial fraction, whereas TUM required

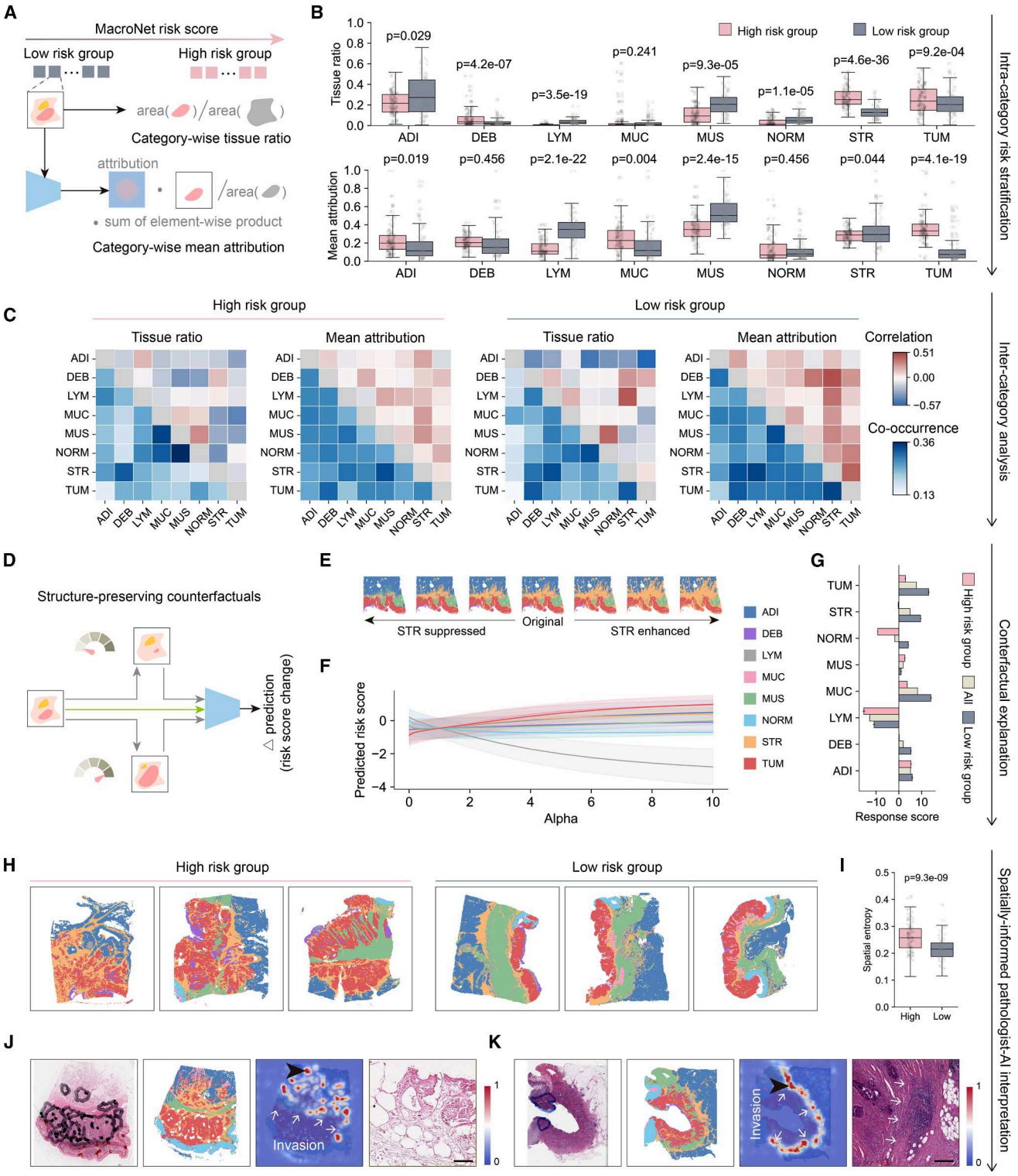


Figure 3. Model interpretation and potential biomarker discovery

(A) Calculation of tissue ratios and mean attributions.

(B) Tissue ratios and mean attributions in high- and low-risk groups. Boxplots show median, interquartile range, and whiskers extending to 1.5× the interquartile range. Tissue ratios were compared after centered log-ratio (CLR) transformation using two-sided Welch's t tests with Benjamini-Hochberg false discovery rate (FDR) correction across tissue categories. Mean attributions, which are not compositional, were compared using Welch's t tests with FDR correction.

(C) Pairwise tissue correlation and co-occurrence matrices. Upper right, correlation; lower left, co-occurrence.

(legend continued on next page)

graph-derived entropy for stronger prognostic value. Kaplan-Meier analyses confirmed this divergence, with TUM graph entropy achieving a C-index of 0.62 and HR of 1.97, and MUS spatial fraction achieving a C-index of 0.60 and HR of 0.57 (Figure 4D).

Pairwise interaction biomarkers derived from multi-tissue graphs also showed significant prognostic associations across cohorts (Figure 4E). Several individual features achieved strong predictive performance, including ADI-TUM-E in CR07, which reached a C-index of 0.69, and STR-TUM-EN1 and MUS-TUM-EN2 in TCGA, each reaching C-indices of 0.73. Interaction patterns varied across cohorts. Favorable biomarkers often involved LYM-MUS and LYM-STR in DACHS and LYM-NORM in TCGA, whereas poor prognosis was commonly associated with STR-TUM and ADI-TUM interactions. Spatial features also captured age-related architectural variation, including higher ADI-STR-AD1 and ADI-TUM-BT1 values and lower DEB-NNR values in patients with early-onset DACHS, with similar trends in MCO (Figures S5A and S5B).

Spatial biomarker spectrum enables transparent prediction

Motivated by the prognostic value of individual spatial biomarkers, we hypothesized that the full set of 628 interpretable features captures a multidimensional slide-level representation of tissue architecture. Functionally, this spectrum serves as an alternative to slide-level embeddings produced by foundation models.[38–41] However, unlike latent DL embeddings, each feature in the spatial biomarker spectrum is mathematically defined and linked to predefined semantic tissue maps, allowing predictions to be traced back to specific spatial patterns. We therefore asked whether this spectrum could support transparent prediction across prognosis, molecular alterations, and treatment response.

To evaluate this, we trained an elastic net-regularized Cox proportional hazards model (Coxnet) on all 628 spatial features to build a spatial biomarker prognostic model (SPM),[50,51] and refined it into a filtered SPM (FSPM) using the top 10 most contributive features (Figure 4F, Methods). In 5-fold cross-validation on the DACHS cohort, both models achieved C-indices exceeding 0.7, matching MacroNet and leading foundation models (Figure 4G). These results generalized to external cohorts (MCO, CR07, TCGA), where SPM and FSPM retained competitive prognostic accuracy (Figure 4H). Essentially, unlike deep models that require retrospective attribution, our interpretable models directly linked predictions to specific biomarkers, such as LYM–STR and ADI–TUM interactions, stromal area fraction, and tumor centrality (TUM–AC) (Figure 4I). This forward interpretability enabled direct visualization of prognostically salient features. Low-risk WSIs displayed orderly STR–TUM and ADI–TUM interactions and abundant lymphocyte involvement (e.g., LYM–STR and LYM–MUS), while high-risk samples showed disorganized architecture and diminished immune connectivity (Figure 4J). Notably, both SPM and FSPM outperformed models using fraction-based features alone (Figures S4B and S4C), further underscoring the importance of incorporating spatial information.

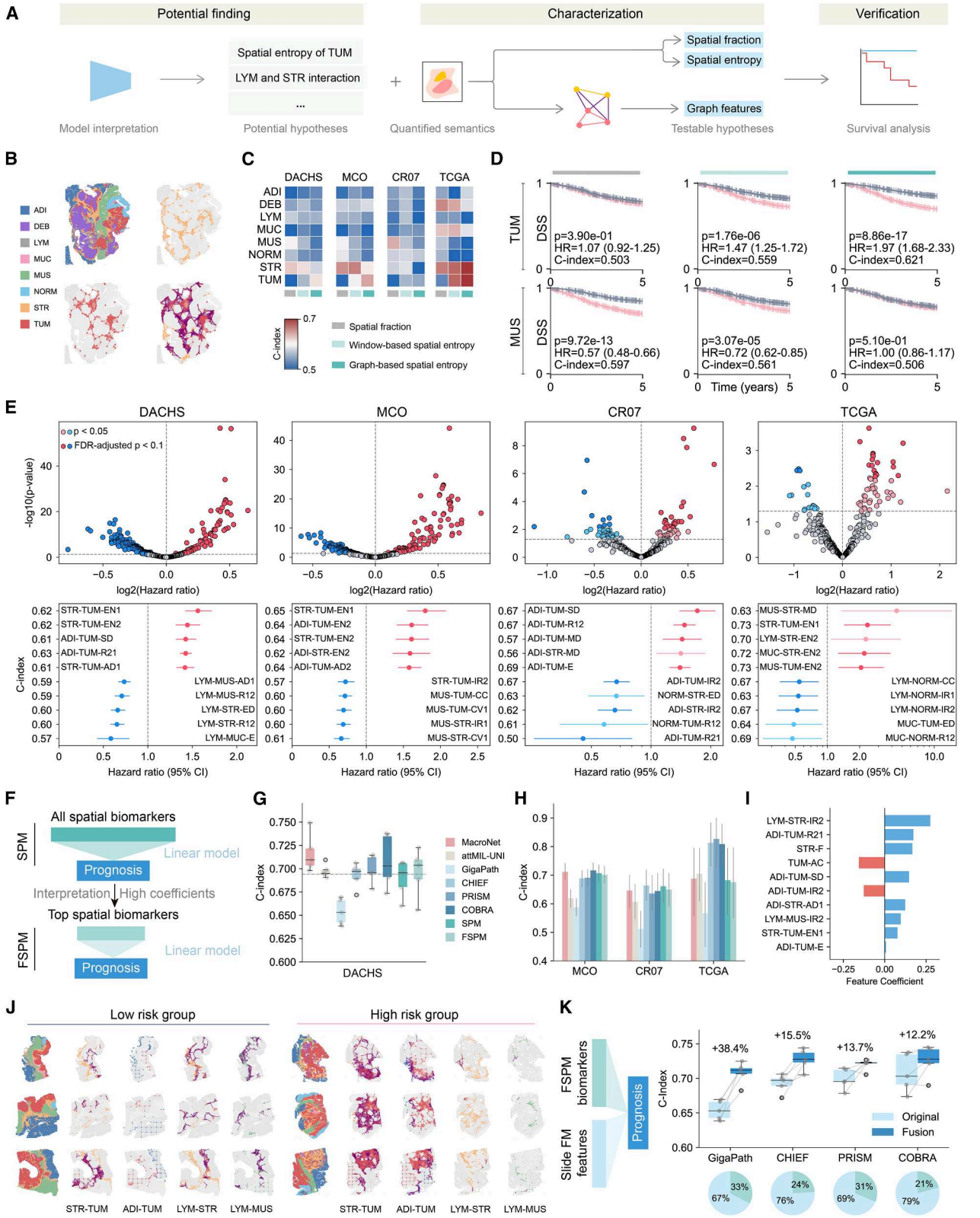
We further combined the 10 FSPM features with slide-level embeddings from foundation models trained on microscopic morphology (Methods). The hybrid model improved prognostic performance, with spatial biomarkers contributing nearly 25% of overall feature importance despite their smaller number (Figure 4K). These results suggest that interpretable spatial features provide complementary information to latent foundation-model representations.

We next evaluated whether the spatial biomarker spectrum could serve as a transferable slide-level representation for mutation and treatment response prediction beyond prognosis (Figure 5A, Methods). For MSI prediction, an L1-regularized logistic regression model achieved a median AUC of 0.85 in 5-fold cross-validation on DACHS and 0.78 across seven external cohorts (Figures 5B and 5C). Key predictive features included MUC-TUM spatial isolation (MUC-TUM-IR2) and debris entropy (DEB-ENT) (Figure 5D), suggesting that MSI is associated with macro-level architectural variation. To evaluate broader genomic associations, we trained models on CORSA, EPIC, WHI, IWHS, and CRA, and tested them on TCGA, where spatial features predicted BMPR2, BRAF, and TP53 mutations with AUCs of 0.78, 0.75, and 0.66, respectively (Figure 5E). Features such as DEB entropy and MUC-TUM interactions were among the top contributors across multiple mutation prediction tasks, suggesting shared spatial correlates of MSI and selected driver mutations (Figures 5F–5H). These findings demonstrate that macro-scale spatial organization, as captured by interpretable spatial biomarkers, encodes genetic information traditionally assumed to require cellular-level analysis.

Interpretable spatial biomarkers uncover divergent responses to adjuvant chemotherapy

To move beyond prediction and enable actionable discovery, we investigated whether specific spatial biomarkers could stratify prognosis among patients who received adjuvant chemotherapy (ACT). In stage II ACT-treated patients, several biomarkers showed prognostic stratification potential (Figure S6A), including LYM-TUM-AD2, which distinguished outcomes in MCO (HR = 5.19, p = 0.0005) and generalized to HECOG (HR = 2.00,

(D) Schematic of the structure-preserving perturbation-based counterfactual explanation approach.
(E) STR perturbation example shows suppression and enhancement.
(F) Risk score response curves after tissue-category perturbation. Colored lines denote the mean predicted risk score, with shaded areas indicating 95% confidence intervals. Alpha values greater than 1 indicate enhancement, and values less than 1 indicate suppression.
(G) Model response magnitudes across high-risk, low-risk, and all patients.
(H) Representative segmentation maps from high- and low-risk patients.
(I) Spatial entropy comparison between risk groups. Boxplots show median, interquartile range, and whiskers extending to 1.5× the interquartile range. p values were calculated using two-sided Mann-Whitney-Wilcoxon tests.
(J and K) Representative high-risk (J) and low-risk (K) cases showing WSI, segmentation, attribution, and zoomed morphology. White arrows indicate cancer invasion direction, and black arrows indicate zoomed-region locations. Scale bars, 200 μm. See also Figure S3.



(legend on next page)

$p = 0.007$) (Figure S6B). Similar patterns were observed in stage III ACT-treated patients (Figure S6C), where TUM-NN stratified risk in MCO (HR = 1.74, $p = 0.021$) and remained significant in HECOG, IDEA Arm A, and IDEA Arm B (Figure S6D).

We further investigated whether specific spatial biomarkers are associated with differential prognosis depending on whether patients received ACT. To this end, we developed a scalable Cox interaction-based framework for identifying biomarkers predictive of ACT benefit (Methods). In the MCO cohort, stage II patients receiving ACT showed worse outcomes overall (HR = 2.53, $p = 0.001$) (Figure 5I), likely reflecting treatment selection bias. Stratification by LYM-NORM-ED, a measure of LYM-normal tissue interaction density, identified subgroups with divergent associations. Patients with low LYM-NORM-ED (≤ 0.038) showed no significant adverse association with ACT (HR = 0.79, $p = 0.711$), whereas patients with high LYM-NORM-ED (>0.038; HR = 4.49, $p = 0.001$) or no LYM-NORM interaction (HR = 3.35, $p = 0.016$) had worse outcomes under ACT. These findings raise the hypothesis that aberrant LYM–NORM tissue interactions may underlie adverse ACT responses, whereas their absence may delineate a subgroup more likely to benefit from therapy.

Surprisingly, in stage III CRC, a setting where ACT is widely regarded as uniformly beneficial,[52,53] our analysis identified a spatial biomarker that stratifies patients into markedly divergent response groups. While Kaplan-Meier analysis across all stage III patients showed a trend favoring ACT (HR = 0.73, $p = 0.159$) (Figure 5J), stratification by LYM–MUC–CC, quantifying the fragmentation of LYM connectivity around mucinous regions, revealed striking heterogeneity. Among patients with high fragmentation (LYM–MUC–CC > 5), ACT was strongly associated with worse survival (HR = 9.08, $p = 0.008$). Conversely, patients with low fragmentation (LYM–MUC–CC ≤ 5) exhibited significant benefit from ACT (HR = 0.29, $p < 0.0001$). The interaction between LYM–MUC–CC status and ACT response was highly significant (interaction HR = 32.45, $p = 0.001$), while patients lacking any LYM–MUC structure showed no discernible treatment effect. These findings suggest that disorganized immune architecture in mucin-rich environments may reflect an aberrant tumor-immune interaction that compromises ACT efficacy,[54] transforming a generally beneficial treatment into a potential liability.[55]

Translating cohort-level spatial signatures into hypotheses through LLMs

Leveraging the interpretability of spatial biomarkers, we used LLMs as auxiliary tools to organize cohort-level spatial signatures into structured hypotheses for expert review. For LYM-MUC-CC, which stratified ACT-associated outcomes in stage III CRC (Figure 5J), prompts grounded in spatial feature definitions and H&E morphology suggested that fragmented LYM connectivity around mucinous regions may reflect immune exclusion and reduced ACT-associated benefit.[54,55] A second LLM reviewer assessed the reasoning and added complementary mechanisms, including chemokine sequestration, hypoxia-driven immune remodeling, and myeloid suppression (Figure 6B). We applied the same workflow to DEB entropy and MUC-TUM interaction, two top contributors across mutation prediction tasks. The LLM generated hypotheses related to tissue disruption, mucinous differentiation, and epithelial-stromal remodeling[56] (Figure S7A), while the reviewer LLM added complementary mechanisms (Figure S7B). Five CRC pathology and tumor biology experts evaluated the LLM-organized hypotheses (Figure S8A). Most were rated as biologically coherent and experimentally actionable (Figures S8B and S8C), supporting this workflow for hypothesis organization while not constituting experimental validation. We also compared coarse prompts with PathPrism-tailored prompts incorporating biomarker definitions, spatial interpretations, and survival statistics (Figure S9A). Across repeated runs, PathPrism-tailored prompts generated more consistent outputs than coarse prompts (Figure S9B).

Perturbation-based VirtualWSI enables exploration of the spatial biomarker atlas and refined stratification

To move from passive observation to active exploration, we developed VirtualWSI as an *in silico* semantic perturbation framework for model interrogation, enabling WSI-wide channel perturbation and interactive perturbation of selected regions while preserving global tissue structure[57] (Methods). To examine how perturbed samples move in the population-level space defined by spatial biomarkers, we constructed a spatial biomarker atlas of patients with CRC using the derived spatial biomarker spectrum (Figure 7A; Methods). MSI-H

Figure 4. Closed-loop spatial biomarker discovery and transparent prognosis

(A) Closed-loop workflow converts MacroNet-derived candidate patterns into quantitative spatial descriptors for cohort-level validation.

(B) Graph modeling strategies for single-tissue and multi-tissue spatial features. Node colors denote tissue categories, and purple edges indicate inter-tissue connections.

(C) Prognostic performance of single-tissue biomarkers across cohorts using spatial fraction, window-based entropy, and graph-based entropy.

(D) Prognostic impact of spatial characterizations for TUM and MUS. Color bars indicate feature classes as in (C). *p* values were calculated by two-sided log rank tests.

(E) Multi-tissue graph biomarkers. Volcano plot shows hazard ratios and *p* values for all biomarkers; forest plots show the top five highest- and lowest-HR biomarkers with C-indices. *p* values were obtained from univariable Cox proportional hazards models (Wald test).

(F) Construction of transparent prognostic models.

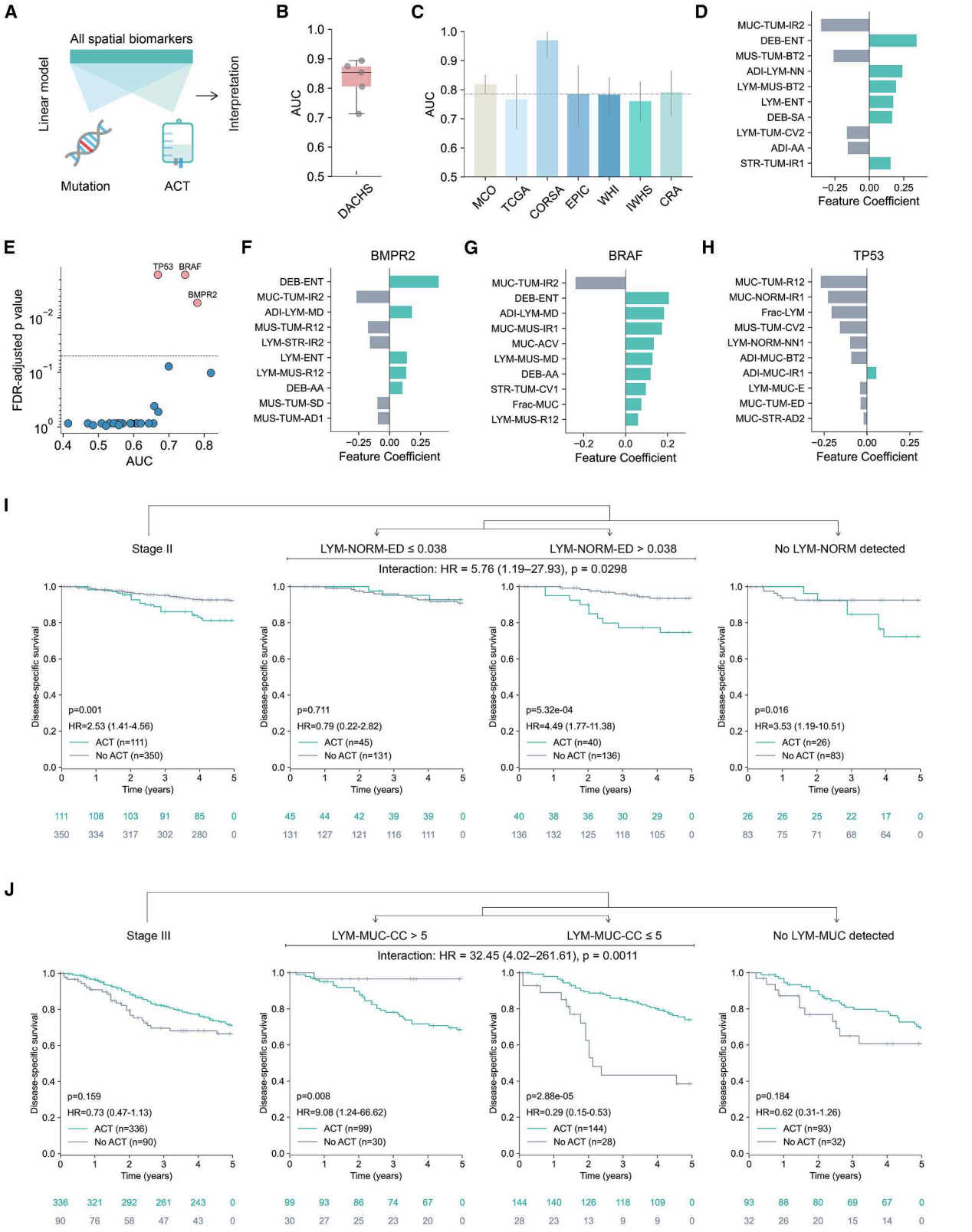
(G) 5-fold cross-validation C-indices. Boxplots show median, interquartile range, and whiskers extending to 1.5× the interquartile range. The dashed line indicates the mean C-index across models.

(H) Prognostic performance on external cohorts. Data are means; error bars indicate 95% confidence intervals from 1,000 bootstrap samples.

(I) FSPM biomarker contributions.

(J) Visualization of top-contributing spatial biomarkers in high- and low-risk patients.

(K) Prognostic modeling combines spatial biomarkers with foundation model-derived slide-level features. Boxplots show 5-fold cross-validation C-indices on DACHS (center line, median; box limits, interquartile range; whiskers, 1.5× the interquartile range); pie chart shows relative feature contributions. See also Figures S4 and S5; Tables S2 and S3.



(legend on next page)

tumors and BRAF-mutant tumors localized to overlapping regions and displayed a spatial pattern opposite to that of the biomarker MUC–TUM–IR2 (Figures 7B–7D), consistent with its negative contribution in the interpretable mutation prediction model (Figures 5D–5G). Among stage III patients in the MCO cohort, subgroups defined by the spatial biomarker LYM–MUC–CC formed structured clusters within the atlas (Figure 7E). Regions enriched for ACT-associated harm corresponded to areas with higher LYM–MUC–CC scores (Figure 7F).

We next used VirtualWSI to perturb the MUC and LYM channels and examine how spatial biomarkers respond to controlled tissue reconfiguration.[58] Increasing perturbation strength induced progressive changes in LYM–MUC–CC values and generated continuous trajectories across the atlas (Figures 7G and 7H). Across patients with stage III MCO, the mean LYM–MUC–CC score increased with perturbation strength (Figure 7I). Patient responses were categorized into three perturbation response archetypes, including positive, flat, and negative responses (Figure 7J). Most patients showed positive responses, whereas smaller subsets exhibited stable or inverse dynamics. During perturbation, a subset of patients transitioned between ACT-associated states defined by LYM–MUC–CC (Figure 7K). We therefore evaluated whether perturbation-induced state transitions could serve as a dynamic biomarker to refine ACT stratification (Figure 7L). When applied to patients without detectable LYM–MUC spatial structure in the original slides (Figure 5J), this perturbation-based biomarker further separated individuals into a subgroup with significant ACT benefit (HR = 0.27, p = 0.010) and a subgroup without a significant treatment association.

VirtualWSI also enabled region-level exploratory perturbation in individual samples. In a representative low-risk tumor, we perturbed selected regions to examine LYM-to-MUS replacement, STR expansion, and TUM extension into ADI (Figure S10A). FSPM risk scores changed minimally after LYM removal, whereas STR-driven boundary disruption and disorganized TUM-ADI invasion increased predicted risk. Projection into the atlas constructed from the ten FSPM features showed that TUM-ADI invasion moved samples toward higher-risk regions (Figure S10B). Visualization of individual biomarkers further indicated that elevated ADI–TUM interaction and STR enrichment characterized these regions, whereas cohesive tumor architecture marked by high TUM–AC corresponded to lower-risk areas (Figure S10C).

Transferability of the PathPrism framework to breast cancer

To evaluate the transferability of the PathPrism framework, we applied it to breast cancer (BRCA). A BRCA-specific dataset was assembled to train PrismNet for tissue segmentation (Figure S11A, Methods). PrismNet achieved strong classification performance on the test set (Figure S11B) and produced accurate whole-slide segmentation (Figure S11C). For prognosis, MacroNet, SPM, and FSPM performed comparably to leading foundation models in primary and external validation cohorts (Figures S11D–S11F), and FSPM-derived scores stratified risk in Kaplan-Meier analyses (Figures S11E–S11G). LYM-STR and NEC-STR were among the most influential biomarkers, with distinct spatial organization between risk groups (Figures S11H and S11I). We also evaluated spatial biomarkers in patients with BRCA receiving antibody-drug conjugates (ADCs). Among trastuzumab deruxtecan (T-DXd)-treated patients, STR-ACL was associated with survival (HR = 6.11, p = 0.018), with lower values reflecting more dispersed stromal topology in patients with favorable outcomes (Figure S11J). In TCGA-BRCA, STR-ACL was inversely correlated with ERBB2 and positively correlated with TOP1 expression (Figures S11K and S11L). Among sacituzumab govitecan (SG)-treated patients, higher ADI-NEC-BT2 was associated with improved outcomes (HR = 0.06, p = 0.002) (Figure S11M) and positively correlated with TACSTD2, but not TOP1, expression (Figures S11N and S11O). Although significant, these correlations were modest, suggesting that spatial biomarkers may capture tumor-context features beyond target gene expression alone.

DISCUSSION

A central challenge in deploying DL in medicine lies in its lack of interpretability, a limitation that undermines both AI-driven scientific discovery and clinician trust.[18,59,60] In computational pathology, this problem is particularly acute given that the complexity of WSIs amplifies the disconnect between prediction and interpretation.[17] Manually designed biomarkers offered high interpretability but limited scope, constrained by human limitations in recognizing and quantifying complex spatial patterns.[3] In contrast, DL and foundation models improve predictive performance by learning representations directly from data.[61] Yet these advances came at the cost of transparency.[24] PathPrism addresses this trade-off through a transparent, semantically structured framework that encodes WSIs into pathologically meaningful spatial representations based on tissue identities and spatial relationships. This design enables

Figure 5. Spatial biomarkers predict gene mutations and stratify benefit from adjuvant chemotherapy

(A) Linear models for mutation prediction and ACT treatment response stratification using spatial biomarkers.

(B) 5-fold cross-validation performance for MSI prediction in DACHS. Boxplots show median, interquartile range, and whiskers extending to 1.5× the interquartile range.

(C) MSI prediction performance in external cohorts. Data are means; error bars indicate 95% confidence intervals from 1,000 bootstrap samples. The dashed line indicates the median AUC across models.

(D) Top spatial biomarker contributions to MSI prediction ranked by mean absolute SHAP value.

(E) Mutation prediction performance for individual genes using spatial biomarkers, reported as AUROC. p values were calculated using two-sided t tests with FDR correction.

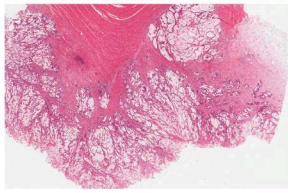
(F–H) Top spatial biomarker contributions for BMPR2 (F), BRAF (G), and TP53 (H) mutation prediction.

(I and J) Kaplan-Meier analysis of ACT treatment response stratified by spatial biomarkers in stage II (I) and stage III (J) CRC. Patients were grouped by biomarker presence and then by median biomarker value. ACT and no-ACT groups were compared within each risk group. Kaplan-Meier p values were calculated using two-sided log rank tests; interaction p values were calculated using Wald tests from Cox models. See also Figure S6.

A

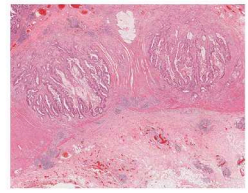
You are a biomedical reasoning agent with expertise in colorectal cancer (CRC), spatial pathology, and treatment response.
Context:
In stage III CRC, adjuvant chemotherapy (ACT) is generally considered beneficial. However, our analysis revealed a spatial biomarker—LYM–MUC–CC, quantifying the fragmentation of lymphocyte connectivity around mucinous regions—that strongly stratifies ACT efficacy:
• Patients with high LYM–MUC–CC (i.e., highly fragmented lymphocyte–mucin interaction) had worse survival with ACT (HR = 9.08).
• In contrast, those with low LYM–MUC–CC showed significant benefit from ACT (HR = 0.29).
• The interaction HR was 32.45 ($P < 0.05$), indicating a highly significant divergence in ACT response.
• Patients without LYM–MUC structures had no detectable ACT effect (HR ≈ 1.0).
This suggests that immune architectural disruption in mucin-rich environments may transform ACT from a beneficial to a harmful intervention.
Task:
I am curious with the finding and want to know why. Currently I have access only to H&E-stained whole-slide images (WSIs), which morphology is worth check?

Focus on the spatial organization of lymphocytes and mucin pools. Specifically, examine whether lymphocytes form cohesive clusters or appear diffusely scattered or fragmented around mucinous regions. Also assess whether mucin lakes physically separate lymphocyte aggregates or disrupt their access to tumor epithelium—this architectural barrier may underlie immune dysfunction and ACT resistance.



Low LYM-MUC-CC case

1.Mucinous adenocarcinoma (ca.90% mucinous differentiation)
2.Early pT2 with initial infiltration in muscularis propria
3.Low number of tumor-infiltrating lymphocytes.



High LYM-MUC-CC case

1.Conventional adenocarcinoma (NOS) with partial mucinous differentiation (ca. 20%)
2.pT2 with deep infiltration in muscularis propria
3.Lot of lymphoid aggregates (close to mucin pools but also in uninvolved adipose tissue).

This is the main observation from HE slides:
The high LYM-MUC cases showed immune-mucin dissociation and lymphoid aggregates in uninvolved tissue. The low LYM-MUC cases had sometimes very high mucinous proportions (over 70%) or no mucinous differentiation at all, with high stromal component and low to moderate number of tumor-infiltration lymphocytes.

This suggests that in high LYM–MUC cases, lymphocytes are spatially excluded from mucinous tumor regions, possibly trapped in distant lymphoid aggregates, indicating ineffective immune engagement. In contrast, low LYM–MUC cases retain structural integrity despite mucin or stromal abundance, allowing some degree of immune-tumor proximity, which may enable ACT synergy.

Based on these spatial patterns and ACT resistance, what is the most likely biological mechanism?

The most likely mechanism is immune exclusion driven by mucin-mediated physical and biochemical barriers that fragment lymphocyte infiltration and coordination. This spatial dissociation impairs cytotoxic function and may induce local immunosuppression, rendering ACT-triggered immunogenic cell death ineffective—or even harmful—by promoting tumor escape in a disorganized immune context.

Given this mucin-mediated immune sequestration hypothesis, what markers and spatial relationships should I test using mIHC or spatial transcriptomics to validate it?

Use mIHC or spatial transcriptomics to assess:
1.CD3/CD8 with MUC2/MUC5AC – test T cell proximity to mucin pools and quantify exclusion zones.
2.CXCL13, PNAd (HEV), and CD20 – identify ectopic lymphoid aggregates and their spatial disconnection from tumor nests.
3.PD-L1, IDO1, TGF-β – evaluate immunosuppressive signaling near mucin-rich, lymphocyte-depleted zones.

B

Review Summary
This is a strong and novel spatial biomarker-driven hypothesis with significant implications for personalized ACT use in stage III CRC. The reasoning is biologically plausible and builds on a well-integrated analysis of H&E morphology, spatial topology, and therapeutic outcomes.The identification of LYM–MUC–CC as a spatial biomarker that reverses ACT benefit is potentially paradigm-shifting. The high interaction HR (32.45, $P<0.05$) strongly supports that lymphocyte-mucin spatial disorganization can serve as a predictive biomarker—not just prognostic. If validated, this could stratify patients into ACT-beneficial or ACT-harmful groups, addressing a major unmet need in CRC treatment personalization

Suggested Alternative/Complementary Mechanistic Hypotheses
1.Mucin as a metabolic and immune sink:
Mucin pools may sequester DAMPs or chemokines (e.g., CXCL9/10), diluting immunogenic signals and misdirecting immune cell chemotaxis.
This could lead to ineffective immune-tumor alignment despite lymphocyte presence.
2.Myeloid cell-mediated suppression:
Mucinous regions may favor infiltration of M2-like macrophages or myeloid-derived suppressor cells (MDSCs) that actively exclude T cells and dampen ACT response.
Markers to check: CD163, CD206, ARG1.
3.Hypoxia-driven immune remodeling:
Large mucin pools can create localized hypoxic niches that suppress effector T cells and induce HIF1α/TGF-β signaling.
This might synergize with ACT-induced stress to favor immune escape.
4.Aberrant fibroblast activation:
In low LYM–MUC cases with high stroma but ACT benefit, activated fibroblasts (e.g., FAP+, αSMA+) might support lymphocyte positioning or antigen presentation.
In high LYM–MUC, mucin may inhibit such interactions.
5.Mismatch repair (MMR)/MSI interaction:
If high LYM–MUC correlates with MSI-low/MSS tumors, these might lack pre-existing immune pressure and depend more on ACT-triggered ICD—which fails when immune exclusion is present.

Figure 6. LLM-assisted hypothesis organization from interpretable spatial biomarkers

(A) Pathologist-LLM workflow using biomarker definitions, cohort-level associations, and morphology review to organize mechanistic hypotheses and potential follow-up experiments.

(B) A second LLM reviewed the workflow and proposed complementary hypotheses for the observed spatial phenotype. See also Figures S7, S8, and S9.

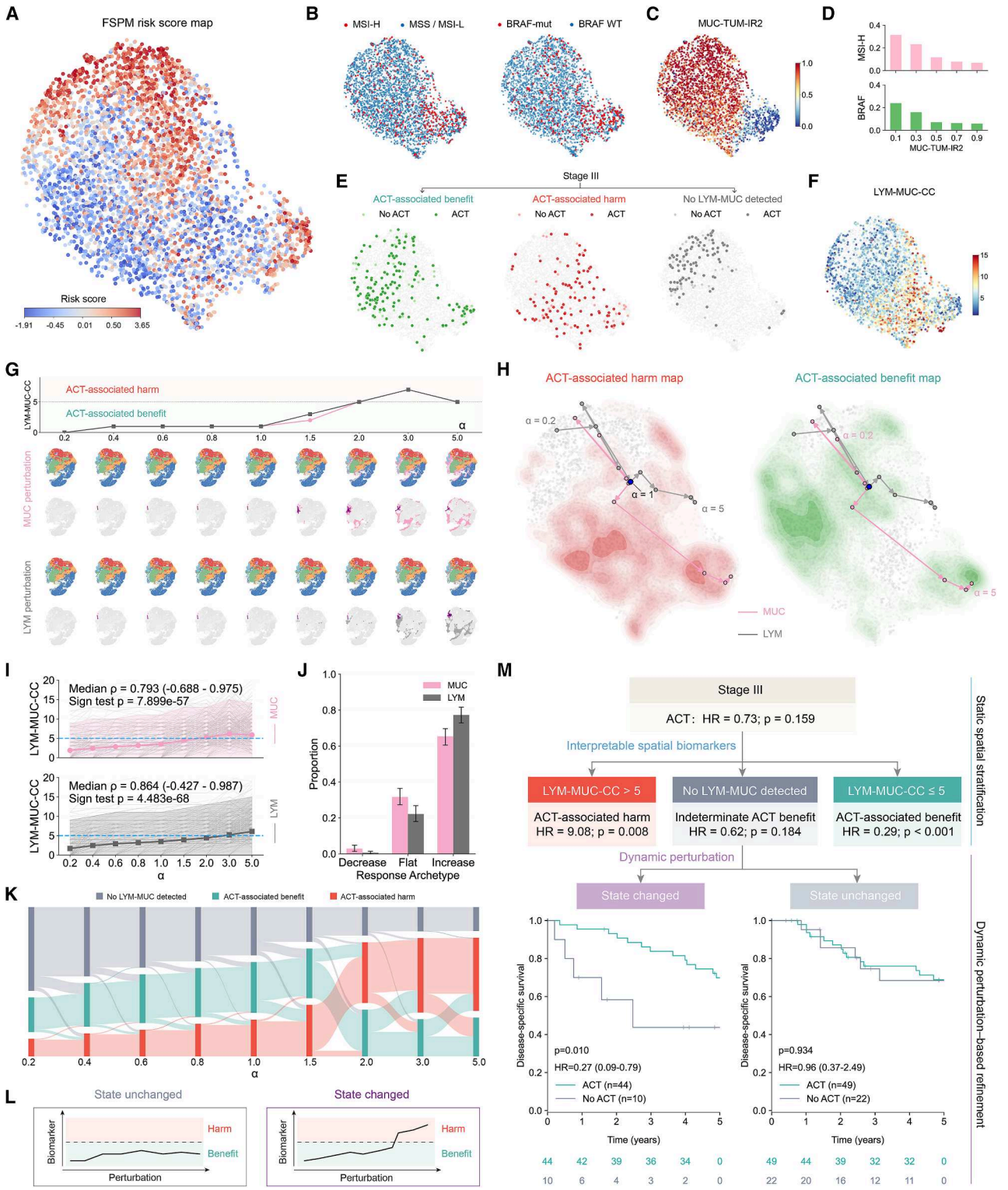


Figure 7. Spatial biomarker atlas and semantic perturbation with VirtualWSI

(A) UMAP spatial biomarker atlas of WSIs colored by FSPM-predicted risk.

(B) MSI status and BRAF mutation status projected onto the atlas.

(C) MUC-TUM-IR2 distribution on the atlas.

(D) Proportions of MSI-H and BRAF-mutant samples across MUC-TUM-IR2 intervals.

(legend continued on next page)

spatial patterns to be quantified, modeled, visualized, perturbed, and explored in a common language shared by models and human experts.

More broadly, PathPrism points toward a form of computational pathology in which AI systems are not only black-box predictors, but also tools for building interpretable representations. This interpretability supports detailed exploration and hypothesis-driven research rather than end-to-end prediction alone, helping relate tissue architecture to clinically relevant patterns. The framework enables cohort-scale analysis of patterns that are difficult to assess manually yet remain grounded in recognizable histology, connecting prediction, hypothesis generation, perturbation-based exploration, and spatial atlases of disease states. Its architecture is inherently compatible with emerging data modalities such as spatial multi-omics,[62–64] 3D pathology,[65,66] and multimodal integration.[67] Future applications may include therapeutic response modeling,[5] drug discovery,[68–70] and agentic AI workflows for interpretable autonomous exploration.[27,71] Overall, PathPrism provides a transparent and structured approach for spatial biomarker discovery through interpretable semantic learning in histopathology.

Limitations of the study

PathPrism has several limitations. The framework depends on predefined semantic tissue representations and is therefore constrained to spatial patterns captured by the selected segmentation categories. Although modular, application to new tumor types requires cancer-specific annotations, PrismNet training, and validation, as illustrated by the BRCA analysis. VirtualWSI uses semantic perturbations to probe model behavior, and these analyses should be interpreted as exploratory model interrogation rather than evidence of biological causality or physically realizable tissue states. In addition, although the prognostic models performed comparably to foundation-model approaches, their performance remains in a moderate range, so downstream biological interpretations, including ACT-associated findings, should be considered hypothesis-generating and require further validation. Finally, the LLM component helps organize interpretable spatial biomarkers into structured hypotheses for expert review, but does not generate validated biological insight without experimental confirmation.

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources should be directed to and will be fulfilled by the lead contact, Jakob Nikolas Kather (jakob_nikolas.kather@tu-dresden.de).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- The codebase and instructions are available at https://github.com/KatherLab/PathPrism.
- The molecular and clinical data for the TCGA and CPTAC cohorts, as well as the digitized WSIs for TCGA, are publicly accessible at the Genomic Data Commons Data Portal (https://portal.gdc.cancer.gov/) and cBioPortal for Cancer Genomics (https://www.cbioportal.org/). The datasets from DACHS, MCO, CR07, IDEA, HeCOG, and the GECCO consortium contain patient-sensitive information and are subject to institutional and consortium-specific data access policies. Access to these datasets may be obtained from the respective contributing institutions or consortia in accordance with their data access regulations and approval procedures.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

ACKNOWLEDGMENTS

We acknowledge Sanddhya Jayabalan, Silvia Barbosa, Asier Rabasco Meneghetti, and Michaela Unger for their help. LK is supported by the National Natural Science Foundation of China (NSFC) (T2425011). JNK is supported by the German Cancer Aid DKH (DECADE, 70115166), the German Federal Ministry of Research, Technology and Space BMFTR (PEARL, 01KD2104C; CAMINO, 01EO2101; TRANSFORM LIVER, 031L0312A; TANGERINE, 01KT2302 through ERA-NET Transcan; Come2Data, 16DKZ2044A; DEEP-HCC, 031L0315A; DECIPHER-M, 01KD2420A; NextBIG, 01ZU2402A; PROSURV, 01KD2509C), the German Research Foundation (DFG, Deutsche Forschungsgemeinschaft) as part of Germany's Excellence Strategy – EXC 2050/2 – Project ID 390696704 – Cluster of Excellence "Center for Tactile Internet with Human-in-the-Loop" (CeTI) of Technische Universität Dresden, as well as through DFG-funded collaborative research projects (TRR 412/1, 535081457; SFB 1709/1 2025, 533056198), 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 program (ODELIA, 101057091; GENIAL, 101096312), the European Research Council ERC (NADIR, 101114631), the Breast Cancer Research Foundation (BELLADONNA, BCRF-25-225) 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. JL acknowledges the Tsinghua Scholarship for Overseas Graduate Studies Fund. JC is supported by the Mildred-Scheel-Postdoktorandenprogramm of the German Cancer Aid (grant #70115730). HeCOG cohort analysis was partially supported by a Hellenic Society of Medical Oncology (HeSMO) grant. GECCO is funded by the National Cancer Institute, National Institutes of Health, and the US Department of Health and Human Services (U01 CA137088, R01 CA488857, P20 CA252733). Genotyping and sequencing

(E) LYM-MUC-CC-defined stage III MCO subgroups projected onto the atlas.
(F) LYM-MUC-CC distribution on the atlas.
(G) Structure-preserving MUC or LYM perturbation across increasing perturbation strength, alpha. Alpha values less than 1 attenuate, and those greater than 1 enhance, the target tissue component. Representative segmentations, biomarker visualizations, and LYM-MUC-CC scores are shown.
(H) ACT-associated harm and benefit density maps with trajectories of perturbed samples under MUC or LYM perturbation.
(I) Population-level LYM-MUC-CC response under perturbation. Gray lines show individual trajectories; colored lines and shading show mean responses and 95% confidence intervals. Directional consistency was assessed using Spearman correlation and a sign test.
(J) Response archetypes under MUC and LYM perturbation. Error bars indicate 95% confidence intervals from 1,000 bootstrap resamples.
(K) Sankey diagram of perturbation-induced state transitions in stage III MCO.
(L) Definition of unchanged and changed states based on perturbation-induced shifts in biomarker-defined ACT association.
(M) ACT benefit stratification refined by MUC perturbation-induced state changes in stage III patients. p values were calculated using two-sided log rank tests. See also Figure S10.

services were provided by the Center for Inherited Disease Research (contract number HHSN268201700006I). This research was funded in part through the National Institutes of Health (NIH)–National Cancer Institute Cancer Center (support grant P30 CA015704). Scientific Computing Infrastructure at Fred Hutch funded by the NIH's Office of Research Infrastructure Programs (grant S10OD028685). The CORSA study was funded by the Austrian Research Funding Agency (BRIDGE grant 829675, to AG) and the Herzfelder'sche Familienstiftung (grant to AG), and was supported by COST Action BM1206. The CRA study was supported by the National Institutes of Health (grant R01 CA068535). The coordination of the EPIC study is financially supported by the International Agency for Research on Cancer and by the Department of Epidemiology and Biostatistics, School of Public Health, Imperial College London, which has additional infrastructure support provided by the NIHR Imperial Biomedical Research Centre. The national cohorts are supported by Danish Cancer Society (Denmark); Ligue Contre le Cancer, Institut Gustave Roussy, Mutuelle Générale de l'Education Nationale, and Institut National de la Santé et de la Recherche Médicale (France); German Cancer Aid, German Cancer Research Center, German Institute of Human Nutrition Potsdam-Rehbruecke, and Federal Ministry of Education and Research (Germany); Associazione Italiana per la Ricerca sul Cancro-AIRC-Italy, Compagnia di San Paolo, and National Research Council (Italy); Dutch Ministry of Public Health, Welfare and Sports, Netherlands Cancer Registry, LK Research Funds, Dutch Prevention Funds, Dutch Zorg Onderzoek Nederland, World Cancer Research Fund, Statistics Netherlands (Netherlands); Health Research Fund–Instituto de Salud Carlos III, Regional Governments of Andalucía, Asturias, Basque Country, Murcia and Navarra, and the Catalan Institute of Oncology (Spain); Swedish Cancer Society, Swedish Research Council, and Region Skåne and Region Västerbotten (Sweden); Cancer Research UK (14136 to EPIC-Norfolk and C8221/A29017 to EPIC-Oxford) and the Medical Research Council (1000143 to EPIC-Norfolk and MR/M012190/1 to EPIC-Oxford; UK). The IWHS study was supported by the NIH (grants CA107333; HHSN261201000032C; and N01 contract awarded to the University of Iowa). The WHI study is funded by the National Heart, Lung, and Blood Institute (NIH, US Department of Health and Human Services) through contracts 75N92021D00001, 75N92021D00002, 75N92021D00003, 75N92021D00004, 75N92021D00005. We kindly thank all individuals who agreed to participate in the CORSA study. Furthermore, we thank all cooperating physicians, students, and the Biobank Graz of the Medical University of Graz. We also acknowledge the TCGA Research Network, which generated the data on which some of the results are based. Where authors are identified as personnel of the International Agency for Research on Cancer or WHO, the authors alone are responsible for the views expressed in this Article, and they do not necessarily represent the decisions, policies, or views of the International Agency for Research on Cancer or WHO. We thank the WHI investigators and staff for their dedication, and the study participants for making the program possible. A full list of WHI investigators can be found online (https://s3-us-west-2.amazonaws.com/www-whi-org/wp-content/uploads/WHI-Investigator-Long-List.pdf). CR07 was funded by the Medical Research Council and translational work supported by Cancer Research UK through the Leeds Radiotherapy Research Centre of Excellence (RRCOER-Jun24/100004). N.P.W. and P.Q. are supported by a Yorkshire Cancer Research program grant (L386). The DACHS study was supported by the German Research Council (BR 1704/6-1, BR 1704/6-3, BR 1704/6-4, CH 117/1-1, HO 5117/2-1, HO 5117/2-2, HE 5998/2-1, HE 5998/2-2, KL 2354/3-1, KL 2354 3-2, RO 2270/8-1, RO 2270/8-2, BR 1704/17-1, and BR 1704/17-2); the Interdisciplinary Research Program of the National Center for Tumor Diseases (NCT), Germany; and the German Federal Ministry of Education and Research (01KH0404, 01ER0814, 01ER0815, 01ER1505A, 01ER1505B, and 01KD2104A). The authors thank the NCT Tissue Bank, Heidelberg, for processing, staining, and scanning tumor tissue slides for this project.

AUTHOR CONTRIBUTIONS

J.L., X.J., L.K., and J.N.K. conceived the study and designed the experiments. J.L. performed the experiments and implemented the core analysis pipeline, with assistance from X.J., T.L., L.Z., M.G., P.N., J.C., L.Ž., V.S., and F.W. N.G.R. and H.S.M. provided expert pathological and clinical interpretation. L.A.B., A.J.F., E.L.G., A.G., S.B., M.J.G., R.S., T.H., U.P., A.I.P., P.Q., N.P.W., M.H., H.B., D.W., J.J., N.H., R.L.W., E.F., K.P., G.F., T.A., J.T., J.-F.E., and M.S. contributed patient cohorts and clinical datasets. J.L., Z.I.C., and J.N.K. prepared the manuscript with input from all authors. L.K. and J.N.K. supervised the research.

DECLARATION OF INTERESTS

JNK holds shares in StratifAI, Synagen, Spira Labs, Tremont AI, and Saterra AI; is Co-PI on institutional research grants from GSK and AstraZeneca, and declares honoraria or consulting fees from AstraZeneca, Bayer, Bioptimus, Daiichi Sankyo, Eisai, Janssen, Merck, MSD, Novartis, BMS, Roche, and Pfizer. MG has received honoraria for lectures sponsored by Techniker Krankenkasse, Sartorius, and AstraZeneca. NGR received travel support from Bruker Spatial Biology, unrelated to this study. GF has served on advisory boards for Pfizer and Novartis, received honoraria from AstraZeneca and Novartis, and holds shares in Genprex, Daiichi Sankyo, RFL Holdings, and FORMYCON. EF has served in an advisory role for Daiichi Sankyo and Amgen, received travel support from Pfizer, AstraZeneca, and Genesis, speaker fees from Lilly, GSK, and AstraZeneca, and holds stock in Genprex Inc. and Deciphera Pharmaceuticals, Inc.

DECLARATION OF GENERATIVE AI AND AI-ASSISTED TECHNOLOGIES IN THE WRITING PROCESS

During the preparation of this work, the authors used ChatGPT by OpenAI in order to refine the text. After using this tool or service, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS
  - Ethics statement
  - Patient cohorts
- METHOD DETAILS
  - PrismNet dataset for tissue segmentation
  - PrismNet
  - WSI semantic segmentation
  - MacroNet
  - Multi-angle interpretation
  - Spatial biomarker construction
  - Transparent prognostic model
  - Mutation prediction model
  - Adjuvant chemotherapy biomarker discovery
  - LLM co-reasoning
  - VirtualWSI
  - Construction of interpretable spatial biomarker atlas
- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.ccell.2026.05.014.

Received: September 20, 2025
Revised: January 29, 2026
Accepted: May 21, 2026
Published: June 11, 2026

REFERENCES

1. Swanson, K., Wu, E., Zhang, A., Alizadeh, A.A., and Zou, J. (2023). From patterns to patients: Advances in clinical machine learning for cancer diagnosis, prognosis, and treatment. Cell *186*, 1772–1791.

2. Shmatko, A., Ghaffari Laleh, N., Gerstung, M., and Kather, J.N. (2022). Artificial intelligence in histopathology: enhancing cancer research and clinical oncology. Nat. Cancer *3*, 1026–1038.

3. Song, A.H., Jaume, G., Williamson, D.F.K., Lu, M.Y., Vaidya, A., Miller, T.R., and Mahmood, F. (2023). Artificial intelligence for digital and computational pathology. Nat. Rev. Bioeng. *1*, 930–949.

4. Perez-Lopez, R., Ghaffari Laleh, N., Mahmood, F., and Kather, J.N. (2024). A guide to artificial intelligence for cancer researchers. Nat. Rev. Cancer *24*, 427–441.

5. Campanella, G., Kumar, N., Nanda, S., Singi, S., Fluder, E., Kwan, R., Muehlstedt, S., Pfarr, N., Schüffler, P.J., Häggström, I., et al. (2025). Real-world deployment of a fine-tuned pathology foundation model for lung cancer biomarker detection. Nat. Med. *31*, 3002–3010.

6. Wagner, S.J., Reisenbüchler, D., West, N.P., Niehues, J.M., Zhu, J., Foersch, S., Veldhuizen, G.P., Quirke, P., Grabsch, H.I., van den Brandt, P.A., et al. (2023). Transformer-based biomarker prediction from colorectal cancer histology: A large-scale multicentric study. Cancer Cell *41*, 1650–1661.e4.

7. Bressan, D., Battistoni, G., and Hannon, G.J. (2023). The dawn of spatial omics. Science *381*, eabq4964.

8. Gong, D., Arbesfeld-Qiu, J.M., Perrault, E., Bae, J.W., and Hwang, W.L. (2024). Spatial oncology: Translating contextual biology to the clinic. Cancer Cell *42*, 1653–1675.

9. Elhanani, O., Ben-Uri, R., and Keren, L. (2023). Spatial profiling technologies illuminate the tumor microenvironment. Cancer Cell *41*, 404–420.

10. Rozenblatt-Rosen, O., Regev, A., Oberdoerffer, P., Nawy, T., Hupalowska, A., Rood, J.E., Ashenberg, O., Cerami, E., Coffey, R.J., Demir, E., et al. (2020). The Human Tumor Atlas Network: Charting tumor transitions across space and time at single-cell resolution. Cell *181*, 236–249.

11. Wang, H., Fu, T., Du, Y., Gao, W., Huang, K., Liu, Z., Chandak, P., Liu, S., Van Katwyk, P., Deac, A., et al. (2023). Scientific discovery in the age of artificial intelligence. Nature *620*, 47–60.

12. Davies, A., Veličković, P., Buesing, L., Blackwell, S., Zheng, D., Tomašev, N., Tanburn, R., Battaglia, P., Blundell, C., Juhász, A., et al. (2021). Advancing mathematics by guiding human intuition with AI. Nature *600*, 70–74.

13. Novakovsky, G., Dexter, N., Libbrecht, M.W., Wasserman, W.W., and Mostafavi, S. (2023). Obtaining genetics insights from deep learning via explainable artificial intelligence. Nat. Rev. Genet. *24*, 125–137.

14. Elmarakeby, H.A., Hwang, J., Arafeh, R., Crowdis, J., Gang, S., Liu, D., AlDubayan, S.H., Salari, K., Kregel, S., Richter, C., et al. (2021). Biologically informed deep neural network for prostate cancer discovery. Nature *598*, 348–352.

15. Krishnan, A., Anahtar, M.N., Valeri, J.A., Jin, W., Donghia, N.M., Sieben, L., Luttens, A., Zhang, Y., Modaresi, S.M., Hennes, A., et al. (2025). A generative deep learning approach to de novo antibiotic design. Cell *188*, 5962–5979.e22. https://doi.org/10.1016/j.cell.2025.07.033.

16. Wong, F., Zheng, E.J., Valeri, J.A., Donghia, N.M., Anahtar, M.N., Omori, S., Li, A., Cubillos-Ruiz, A., Krishnan, A., Jin, W., et al. (2024). Discovery of a structural class of antibiotics with explainable deep learning. Nature *626*, 177–185.

17. Liang, J., Zhang, W., Yang, J., Wu, M., Dai, Q., Yin, H., Xiao, Y., and Kong, L. (2023). Deep learning supported discovery of biomarkers for clinical prognosis of liver cancer. Nat. Mach. Intell. *5*, 408–420.

18. Kim, C., Gadgil, S.U., and Lee, S.-I. (2025). Transparency of medical artificial intelligence systems. Nat. Rev. Bioeng. *4*, 11–29.

19. Song, A.H., Chen, R.J., Ding, T., Williamson, D.F.K., Jaume, G., and Mahmood, F. (2024). Morphological prototyping for unsupervised slide representation learning in computational pathology. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (IEEE), pp. 11566–11578.

20. Lee, Y., Park, J.H., Oh, S., Shin, K., Sun, J., Jung, M., Lee, C., Kim, H., Chung, J.-H., Moon, K.C., and Kwon, S. (2022). Derivation of prognostic contextual histopathological features from whole-slide images of tumours via graph deep learning. Nat. Biomed. Eng. *6*, 1452–1466.

21. L'Imperio, V., Wulczyn, E., Plass, M., Müller, H., Tamini, N., Gianotti, L., Zucchini, N., Reihs, R., Corrado, G.S., Webster, D.R., et al. (2023). Pathologist validation of a machine learning-derived feature for colon cancer risk stratification. JAMA Netw. Open *6*, e2254891.

22. Shi, J.-Y., Wang, X., Ding, G.-Y., Dong, Z., Han, J., Guan, Z., Ma, L.-J., Zheng, Y., Zhang, L., Yu, G.-Z., et al. (2021). Exploring prognostic indicators in the pathological images of hepatocellular carcinoma based on deep learning. Gut *70*, 951–961.

23. Mandair, D., Reis-Filho, J.S., and Ashworth, A. (2023). Biological insights and novel biomarker discovery through deep learning approaches in breast cancer histopathology. NPJ Breast Cancer *9*, 21.

24. Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat. Mach. Intell. *1*, 206–215.

25. Liu, F., Zhou, H., Gu, B., Zou, X., Huang, J., Wu, J., Li, Y., Chen, S.S., Hua, Y., Zhou, P., et al. (2025). Application of large language models in medicine. Nat. Rev. Bioeng. *3*, 445–464.

26. Clusmann, J., Kolbinger, F.R., Muti, H.S., Carrero, Z.I., Eckardt, J.-N., Laleh, N.G., Löffler, C.M.L., Schwarzkopf, S.-C., Unger, M., Veldhuizen, G.P., et al. (2023). The future landscape of large language models in medicine. Commun. Med. *3*, 141.

27. Gao, S., Fang, A., Huang, Y., Giunchiglia, V., Noori, A., Schwarz, J.R., Ektefaie, Y., Kondic, J., and Zitnik, M. (2024). Empowering biomedical discovery with AI agents. Cell *187*, 6125–6151.

28. Lu, M.Y., Chen, B., Williamson, D.F.K., Chen, R.J., Zhao, M., Chow, A.K., Ikemura, K., Kim, A., Pouli, D., Patel, A., et al. (2024). A multimodal generative AI copilot for human pathology. Nature *634*, 466–473.

29. Bunne, C., Roohani, Y., Rosen, Y., Gupta, A., Zhang, X., Roed, M., Alexandrov, T., AlQuraishi, M., Brennan, P., Burkhardt, D.B., et al. (2024). How to build the virtual cell with artificial intelligence: Priorities and opportunities. Cell *187*, 7045–7063.

30. Rood, J.E., Hupalowska, A., and Regev, A. (2024). Toward a foundation model of causal cell and tissue biology with a Perturbation Cell and Tissue Atlas. Cell *187*, 4520–4545.

31. Chen, V., Yang, M., Cui, W., Kim, J.S., Talwalkar, A., and Ma, J. (2024). Applying interpretable machine learning in computational biology-pitfalls, recommendations and opportunities for new developments. Nat. Methods *21*, 1454–1461.

32. Rotem, O., and Zaritsky, A. (2024). Visual interpretability of bioimaging deep learning models. Nat. Methods *21*, 1394–1397.

33. Kim, C., Gadgil, S.U., DeGrave, A.J., Omiye, J.A., Cai, Z.R., Daneshjou, R., and Lee, S.-I. (2024). Transparent medical image AI via an image-text foundation model grounded in medical literature. Nat. Med. *30*, 1154–1165.

34. Kather, J.N., Krisam, J., Charoentong, P., Luedde, T., Herpel, E., Weis, C.-A., Gaiser, T., Marx, A., Valous, N.A., Ferber, D., et al. (2019). Predicting survival from colorectal cancer histology slides using deep learning: A retrospective multicenter study. PLoS Med. *16*, e1002730.

35. El Nahhas, O.S.M., van Treeck, M., Wölflein, G., Unger, M., Ligero, M., Lenz, T., Wagner, S.J., Hewitt, K.J., Khader, F., Foersch, S., et al. (2025). From whole-slide image to biomarker prediction: end-to-end weakly supervised deep learning in computational pathology. Nat. Protoc. *20*, 293–316.

36. Jiang, X., Hoffmeister, M., Brenner, H., Muti, H.S., Yuan, T., Foersch, S., West, N.P., Brobeil, A., Jonnagaddala, J., Hawkins, N., et al. (2024). End-to-end prognostication in colorectal cancer by deep learning: a retrospective, multicentre study. Lancet Digit. Health *6*, e33–e43.

37. Chen, R.J., Ding, T., Lu, M.Y., Williamson, D.F.K., Jaume, G., Song, A.H., Chen, B., Zhang, A., Shao, D., Shaban, M., et al. (2024). Towards a general-purpose foundation model for computational pathology. Nat. Med. *30*, 850–862.

38. Xu, H., Usuyama, N., Bagga, J., Zhang, S., Rao, R., Naumann, T., Wong, C., Gero, Z., González, J., Gu, Y., et al. (2024). A whole-slide foundation model for digital pathology from real-world data. Nature *630*, 181–188.

39. Wang, X., Zhao, J., Marostica, E., Yuan, W., Jin, J., Zhang, J., Li, R., Tang, H., Wang, K., Li, Y., et al. (2024). A pathology foundation model for cancer diagnosis and prognosis prediction. Nature *634*, 970–978.

40. Lenz, T., Neidlinger, P., Ligero, M., Wölflein, G., van Treeck, M., and Kather, J.N. (2025). Unsupervised Foundation Model-Agnostic Slide-Level Representation Learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (IEEE), pp. 30807–30817.

41. Shaikovski, G., Casson, A., Severson, K., Zimmermann, E., Wang, Y.K., Kunz, J.D., Retamero, J.A., Oakley, G., Klimstra, D., Kanan, C., et al. (2024). PRISM: A multi-modal generative foundation model for slide-level histopathology. Preprint at arXiv. https://doi.org/10.48550/arXiv.2405.10254.

42. Simonyan, K., Vedaldi, A., and Zisserman, A. (2013). Deep inside Convolutional Networks: Visualising image classification models and saliency maps. Preprint at arXiv. https://doi.org/10.48550/arXiv.1312.6034.

43. DeGrave, A.J., Cai, Z.R., Janizek, J.D., Daneshjou, R., and Lee, S.-I. (2023). Auditing the inference processes of medical-image classifiers by leveraging generative AI and the expertise of physicians. Nat. Biomed. Eng. *9*, 294–306.

44. Pagès, F., Berger, A., Camus, M., Sanchez-Cabo, F., Costes, A., Molidor, R., Mlecnik, B., Kirilovsky, A., Nilsson, M., Damotte, D., et al. (2005). Effector memory T cells, early metastasis, and survival in colorectal cancer. N. Engl. J. Med. *353*, 2654–2666.

45. Pagès, F., Mlecnik, B., Marliot, F., Bindea, G., Ou, F.-S., Bifulco, C., Lugli, A., Zlobec, I., Rau, T.T., Berger, M.D., et al. (2018). International validation of the consensus Immunoscore for the classification of colon cancer: a prognostic and accuracy study. Lancet *391*, 2128–2139.

46. Westwood, A.C., Wilson, B.I., Laye, J., Grabsch, H.I., Mueller, W., Magee, D.R., Quirke, P., and West, N.P. (2025). Deep-learning enabled combined measurement of tumour cell density and tumour infiltrating lymphocyte density as a prognostic biomarker in colorectal cancer. BJC Rep. *3*, 12.

47. Reitsam, N.G., Grosser, B., Steiner, D.F., Grozdanov, V., Wulczyn, E., L'Imperio, V., Plass, M., Müller, H., Zatloukal, K., Muti, H.S., et al. (2024). Converging deep learning and human-observed tumor-adipocyte interaction as a biomarker in colorectal cancer. Commun. Med. (Lond.) *4*, 163.

48. Reitsam, N.G., Offermans, K., Simons, C.C.J.M., Grosser, B., Zimmermann, J., Grabsch, H.I., Märkl, B., and van den Brandt, P.A. (2025). Prognostic and predictive value of SARIFA-status within molecular subgroups of colorectal cancer: Insights from the Netherlands Cohort Study. Am. J. Surg. Pathol. *49*, 956–969.

49. Sebag-Montefiore, D., Stephens, R.J., Steele, R., Monson, J., Grieve, R., Khanna, S., Quirke, P., Couture, J., de Metz, C., Myint, A.S., et al. (2009). Preoperative radiotherapy versus selective postoperative chemoradiotherapy in patients with rectal cancer (MRC CR07 and NCIC-CTG C016): a multicentre, randomised trial. Lancet *373*, 811–820.

50. Simon, N., Friedman, J., Hastie, T., and Tibshirani, R. (2011). Regularization paths for cox's proportional hazards model via coordinate descent. J. Stat. Softw. *39*, 1–13.

51. Pölsterl, S. (2020). scikit-survival: A Library for Time-to-Event Analysis Built on Top of scikit-learn. J. Mach. Learn. Res. *21*, 1–212.

52. Wolpin, B.M., Meyerhardt, J.A., Mamon, H.J., and Mayer, R.J. (2007). Adjuvant treatment of colorectal cancer. CA Cancer J. Clin. *57*, 168–185.

53. Schmoll, H.J., Van Cutsem, E., Stein, A., Valentini, V., Glimelius, B., Haustermans, K., Nordlinger, B., van de Velde, C.J., Balmana, J., Regula, J., et al. (2012). ESMO Consensus Guidelines for management of patients with colon and rectal cancer. a personalized approach to clinical decision making. Ann. Oncol. *23*, 2479–2516.

54. Elomaa, H., Tarkiainen, V., Äijälä, V.K., Sirniö, P., Ahtiainen, M., Sirkiä, O., Karjalainen, H., Kastinen, M., Tapiainen, V.V., Rintala, J., et al. (2025). Associations of mucinous differentiation and mucin expression with immune cell infiltration and prognosis in colorectal adenocarcinoma. Br. J. Cancer *132*, 660–669.

55. Zhang, Y., Chen, Y., Huang, J., Wu, X., Tang, R., Huang, Q., Xu, Y., Peng, X., Fu, K., and Xiao, S. (2022). Mucinous histology is associated with poor prognosis in locally advanced colorectal adenocarcinoma treated with postoperative first-line adjuvant chemotherapy: A systematic review and meta-analysis. Eur. J. Surg. Oncol. *48*, 2075–2081.

56. Uderhardt, S., Martins, A.J., Tsang, J.S., Lämmermann, T., and Germain, R.N. (2019). Resident macrophages cloak tissue microlesions to prevent neutrophil-driven inflammatory damage. Cell *177*, 541–555.e17.

57. Wang, Z.J., Farooq, A.S., Chen, Y.-J., Bhargava, A., Xu, A.M., and Thomson, M.W. (2025). Identifying perturbations that boost T-cell infiltration into tumours via counterfactual learning of their spatial proteomic profiles. Nat. Biomed. Eng. *9*, 390–404.

58. Theodoris, C.V., Xiao, L., Chopra, A., Chaffin, M.D., Al Sayed, Z.R., Hill, M.C., Mantineo, H., Brydon, E.M., Zeng, Z., Liu, X.S., et al. (2023). Transfer learning enables predictions in network biology. Nature *618*, 616–624.

59. Rajpurkar, P., Chen, E., Banerjee, O., and Topol, E.J. (2022). AI in health and medicine. Nat. Med. *28*, 31–38.

60. Chen, R.J., Wang, J.J., Williamson, D.F.K., Chen, T.Y., Lipkova, J., Lu, M.Y., Sahai, S., and Mahmood, F. (2023). Algorithmic fairness in artificial intelligence for medicine and healthcare. Nat. Biomed. Eng. *7*, 719–742.

61. LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. Nature *521*, 436–444.

62. Vandereyken, K., Sifrim, A., Thienpont, B., and Voet, T. (2023). Methods and applications for single-cell and spatial multi-omics. Nat. Rev. Genet. *24*, 494–515.

63. Zhang, D., Schroeder, A., Yan, H., Yang, H., Hu, J., Lee, M.Y.Y., Cho, K.S., Susztak, K., Xu, G.X., Feldman, M.D., et al. (2024). Inferring super-resolution tissue architecture by integrating spatial transcriptomics with histology. Nat. Biotechnol. *42*, 1372–1377.

64. Coleman, K., Schroeder, A., and Li, M. (2024). Unlocking the power of spatial omics with AI. Nat. Methods *21*, 1378–1381.

65. Song, A.H., Williams, M., Williamson, D.F.K., Chow, S.S.L., Jaume, G., Gao, G., Zhang, A., Chen, B., Baras, A.S., Serafin, R., et al. (2024). Analysis of 3D pathology samples using weakly supervised AI. Cell *187*, 2502–2520.e17.

66. Liu, J.T.C., Glaser, A.K., Bera, K., True, L.D., Reder, N.P., Eliceiri, K.W., and Madabhushi, A. (2021). Harnessing non-destructive 3D pathology. Nat. Biomed. Eng. *5*, 203–218.

67. Almagro-Pérez, C., Song, A.H., Weishaupt, L., Kim, A., Jaume, G., Williamson, D.F.K., Hemker, K., Lu, M.Y., Singh, K., Chen, B., et al. (2025). AI-driven 3D Spatial Transcriptomics. Preprint at arXiv. https://doi.org/10.48550/arXiv.2502.17761.

68. Vamathevan, J., Clark, D., Czodrowski, P., Dunham, I., Ferran, E., Lee, G., Li, B., Madabhushi, A., Shah, P., Spitzer, M., et al. (2019). Applications of machine learning in drug discovery and development. Nat. Rev. Drug Discov. *18*, 463–477.

69. Jaume, G., de Brot, S., Song, A.H., Williamson, D.F.K., Oldenburg, L., Zhang, A., Chen, R.J., Asin, J., Blatter, S., Dettwiler, M., et al. (2024). Deep learning-based modeling for preclinical drug safety assessment. Preprint at bioRxiv. https://doi.org/10.1101/2024.07.20.604430.

70. Jaume, G., Peeters, T., Song, A.H., Pettit, R., Williamson, D.F.K., Oldenburg, L., Vaidya, A., de Brot, S., Chen, R.J., Thiran, J.-P., et al. (2024). AI-driven discovery of morphomolecular signatures in toxicology. Preprint at bioRxiv. https://doi.org/10.1101/2024.07.19.604355.

71. Ferber, D., El Nahhas, O.S.M., Wölflein, G., Wiest, I.C., Clusmann, J., Leßmann, M.-E., Foersch, S., Lammert, J., Tschochohei, M., Jäger, D., et al. (2025). Development and validation of an autonomous artificial intelligence agent for clinical decision-making in oncology. Nat. Cancer *6*, 1337–1349.

72. Brenner, H., Chang-Claude, J., Seiler, C.M., and Hoffmeister, M. (2011). Long-term risk of colorectal cancer after negative colonoscopy. J. Clin. Oncol. *29*, 3761–3767.

73. Hoffmeister, M., Jansen, L., Rudolph, A., Toth, C., Kloor, M., Roth, W., Bläker, H., Chang-Claude, J., and Brenner, H. (2015). Statin use and

survival after colorectal cancer: the importance of comprehensive confounder adjustment. J. Natl. Cancer Inst. *107*, djv045.

74. Jonnagaddala, J., Croucher, J.L., Jue, T.R., Meagher, N.S., Caruso, L., Ward, R., and Hawkins, N. (2016). Integration and analysis of heterogeneous colorectal cancer data for translational research. Nurs Informatics *225*, 387–391.
75. The Cancer Genome Atlas Network (2012). Comprehensive molecular characterization of human colon and rectal cancer. Nature *487*, 330–337.
76. Weinstein, J.N., Collisson, E.A., Mills, G.B., Shaw, K.R.M., Ozenberger, B.A., Ellrott, K., Shmulevich, I., Sander, C., Stuart, J.M., and Stuart, J.M. (2013). The Cancer Genome Atlas Pan-Cancer analysis project. Nat. Genet. *45*, 1113–1120.
77. dbGaP Study (2018). Common Variant GWAS, Genetics & Epidemiology of Colorectal Cancer Consortium. https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs001078.v1.p1.
78. André, T., Vernerey, D., Mineur, L., Bennouna, J., Desrame, J., Faroux, R., Fratte, S., Hug de Larauze, M., Paget-Bailly, S., Chibaudel, B., et al. (2018). Three versus 6 months of oxaliplatin-based Adjuvant Chemotherapy for patients with stage III colon cancer: Disease-free survival results from a randomized, open-label, international duration evaluation of adjuvant (IDEA) France, phase III trial. J. Clin. Oncol. *36*, 1469–1477.
79. Grothey, A., Sobrero, A.F., Shields, A.F., Yoshino, T., Paul, J., Taieb, J., Souglakos, J., Shi, Q., Kerr, R., Labianca, R., et al. (2018). Duration of adjuvant chemotherapy for stage III colon cancer. N. Engl. J. Med. *378*, 1177–1188.
80. Global Health Initiative (2024). Hellenic Cooperative Oncology Group (HeCOG) CLIMEDIN Study. https://www.snfghi.org/en/pillars/science-technology-and-innovation/hellenic-cooperative-oncology-group-hecog-climedin-study/.
81. Kather, J.N., Halama, N., and Marx, A. (2018). 100,000 histological images of human colorectal cancer and healthy tissue. Zenodo. https://doi.org/10.5281/zenodo.1214456.
82. Yamashita, R., Long, J., Longacre, T., Peng, L., Berry, G., Martin, B., Higgins, J., Rubin, D.L., and Shen, J. (2021). Deep learning model for the prediction of microsatellite instability in colorectal cancer: a diagnostic study. Lancet Oncol. *22*, 132–141.
83. Wang, X., Yang, S., Zhang, J., Wang, M., Zhang, J., Yang, W., Huang, J., and Han, X. (2022). Transformer-based unsupervised contrastive learning for histopathological image classification. Med. Image Anal. *81*, 102559.
84. Huang, Z., Bianchi, F., Yuksekgonul, M., Montine, T.J., and Zou, J. (2023). A visual-language foundation model for pathology image analysis using medical Twitter. Nat. Med. *29*, 2307–2316.
85. Xiang, J., Wang, X., Zhang, X., Xi, Y., Eweje, F., Chen, Y., Li, Y., Bergstrom, C., Gopaulchan, M., Kim, T., et al. (2025). A vision-language foundation model for precision oncology. Nature *638*, 769–778.
86. Xie, S., Girshick, R., Dollar, P., Tu, Z., and He, K. (2017). Aggregated Residual Transformations for Deep Neural Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (IEEE), pp. 1492–1500.
87. Verweij, P.J.M., and Van Houwelingen, H.C. (1994). Penalized likelihood in Cox regression. Stat. Med. *13*, 2427–2436.
88. Hurst, A., Lerer, A., Goucher, A.P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A.J., Welihinda, A., Hayes, A., Alec, R., et al. (2024). GPT-4o System Card. Preprint at arXiv. https://doi.org/10.48550/arXiv.2410.21276.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
The Cancer Genome Archive (TCGA)	https://portal.gdc.cancer.gov/	RRID:SCR_003193
NCT-CRC-HE-100K-NONORM	https://zenodo.org/records/1214456	N/A
NCT-CRC-HE-100K	https://zenodo.org/records/1214456	N/A
CRC-VAL-HE-7K	https://zenodo.org/records/1214456	N/A
Software and algorithms		
PathPrism	This paper; https://github.com/KatherLab/PathPrism	N/A
STAMP	https://github.com/KatherLab/STAMP	N/A
Python (3.9)	https://www.python.org/	RRID:SCR_008394
PyTorch (2.2.2)	https://pytorch.org	RRID:SCR_018536
NumPy (1.21.6)	http://www.numpy.org	RRID:SCR_008633
Pandas (1.3.5)	https://pandas.pydata.org	RRID:SCR_018214
SciPy (1.7.3)	http://www.scipy.org	RRID:SCR_008058
Matplotlib (3.4.3)	https://matplotlib.org	RRID:SCR_008624
scikit-learn (1.0.2)	https://scikit-learn.org	N/A
Lifelines (0.27.8)	https://lifelines.readthedocs.io/	N/A
Openslide (1.1.1)	https://openslide.org/	N/A
NetworkX (2.6)	https://networkx.org	N/A

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

Ethics statement

This retrospective study adhered to the Declaration of Helsinki and we obtained ethical approval from each contributing center. Data collection and analysis were done in an anonymised manner. Because of the retrospective nature of the study and the use of anonymised data, the ethics board at the Medical Faculty of TU Dresden waived the requirement for a formal ethics vote.

Patient cohorts

In this retrospective, multicentre study, we used anonymised haematoxylin and eosin–stained CRC slides from eleven cohorts in Australia, Europe (including the UK), and the USA for model training and validation. For data quality control, patients lacking essential clinical information (e.g., stage, treatment status, or survival outcome) were excluded. WSIs with low image quality, such as blurred regions or missing metadata (e.g., magnification level), were also removed. Given the focus on characterizing spatial architecture at the whole-slide level, samples with fragmented tissue or insufficient tissue coverage were excluded based on a quantitative threshold. Specifically, each WSI was divided into non-overlapping patches of 224 × 224 pixels at a resolution of 0.5 μm per pixel, and slides containing fewer than 10,000 tissue-containing patches, equivalent to a total tissue area below 1.25 cm^2, were excluded from the analysis. The final reported patient counts represent those who passed all quality control steps and were ultimately included in the analyses. The datasets used in this study include:

DACHS (n = 2281), the "Darmkrebs: Chancen der Verhütung durch Screening" study,[72,73] a large population-based case-control and patient cohort study on colorectal cancer conducted in southwestern Germany, designed to investigate genetic, lifestyle, and clinical risk factors for disease incidence and prognosis, with formalin-fixed paraffin-embedded (FFPE) tumor samples obtained from surgical resections and digitised into haematoxylin and eosin-stained WSIs, accompanied by harmonised long-term clinical and follow-up data.

MCO (n = 1395), the "Molecular and Cellular Oncology" study,[74] a prospective multicentre cohort from Australia enrolling patients who underwent curative-intent resection for colorectal cancer, with systematically collected clinical, pathological, and molecular data, and FFPE tumor specimens retrieved and digitised as haematoxylin and eosin-stained WSIs.

TCGA (n = 232), "The Cancer Genome Atlas",[75,76] a publicly funded US-based initiative encompassing CRC cases across all disease stages, designed to enable comprehensive molecular characterisation, with matched clinical annotations and digitised haematoxylin and eosin-stained WSIs obtained from FFPE surgical specimens via the Genomic Data Commons portal.

CR07 (n = 579), the "MRC CR07" trial,[49] a multicentre randomised clinical study conducted by the Medical Research Council (MRC) and the National Cancer Institute of Canada Clinical Trials Group (NCIC-CTG) evaluating preoperative radiotherapy versus selective postoperative chemoradiotherapy in operable rectal cancer, with FFPE tumor samples retrieved for a subset of patients and digitised as haematoxylin and eosin-stained WSIs.

GECCO (n = 874), the "Genetics and Epidemiology of Colorectal Cancer Consortium",[77] an international collaboration integrating data from multiple population-based and case-control studies on colorectal cancer, with FFPE tumor samples and harmonised clinical data available for five contributing cohorts and slides digitised as haematoxylin and eosin-stained WSIs: EPIC (n = 124), the "European Prospective Investigation into Cancer and Nutrition", a multicentre European cohort focusing on diet, lifestyle, and cancer risk. CORSA (n = 50), the "Colorectal Cancer Study of Austria", a population-based case-control study with biospecimen and clinical data. IWHS (n = 267), the "Iowa Women's Health Study", a US prospective cohort of postmenopausal women with data on dietary and hormonal exposures. CRA (n = 208), the "Cancer Risk Assessment" study, a US-based case-control study investigating genetic and environmental cancer risk factors. WHI (n = 225), the "Women's Health Initiative", a large US prospective cohort of postmenopausal women with lifestyle and biomarker data linked to colorectal cancer cases.

IDEA (n = 669), the "International Duration Evaluation of Adjuvant Chemotherapy" pooled study,[78,79] a multinational phase III trial comparing 3 versus 6 months of oxaliplatin-based adjuvant chemotherapy in stage III colon cancer, with tumor samples obtained from the French sub-cohort conducted by the "Groupe Coopérateur Multidisciplinaire en Oncologie" (GERCOR) and digitised as haematoxylin and eosin-stained WSIs with clinical data.

HeCOG (n = 970), the "Hellenic Cooperative Oncology Group",[80] a non-profit academic research network based in Greece and Cyprus conducting investigator-initiated clinical trials with integrated translational components, with FFPE tumor samples collected across multiple centers and digitised as haematoxylin and eosin-stained WSIs linked to harmonised clinical and outcome data.

Of these cohorts, DACHS, MCO, CR07, and TCGA were used for prognostic modeling; DACHS, MCO, TCGA, and all GECCO cohorts (EPIC, CORSA, IWHS, CRA, WHI) were used for mutation prediction; and MCO, IDEA, and HeCOG were used for adjuvant chemotherapy analysis.

BRCA cohorts were additionally included to evaluate the transferability of the PathPrism framework across tumor types. TCGA-BRCA (n = 642; Table S4), BRCA cohort from The Cancer Genome Atlas, comprises cases spanning multiple disease stages with matched molecular and clinical annotations.[75,76] FFPE tumor samples obtained from surgical resections were digitised into haematoxylin and eosin-stained WSIs and accessed through the Genomic Data Commons portal, together with matched mRNA expression profiles and curated clinical metadata. HeCOG-BRCA (n = 38; Table S5), the breast cancer cohort from the Hellenic Cooperative Oncology Group,[80] a non-profit academic research network based in Greece and Cyprus conducting investigator-initiated clinical trials with integrated translational components, includes patients receiving antibody–drug conjugate (ADC)-based therapies. FFPE tumor samples collected across multiple centers were digitised as haematoxylin and eosin-stained WSIs and linked to harmonised clinical annotations and treatment outcome data.

METHOD DETAILS

PrismNet dataset for tissue segmentation

PrismNet for CRC tissue segmentation was trained using the publicly available NCT-CRC-HE-100K-NONORM dataset and validated on CRC-VAL-HE-7K.[34,81] The NCT-CRC-HE-100K-NONORM dataset consists of 100,000 non-overlapping image patches derived from H&E-stained histological slides of human CRC and normal tissue. All images are 224x224 pixels at a resolution of 0.5 μm per pixel, and no stain normalization was applied. The patches were manually extracted from 86 H&E-stained whole-slide images obtained from FFPE tissue specimens from the NCT Biobank and the pathology archive of the University Medical Center Mannheim. These samples included primary CRC tumors and CRC liver metastases. To increase morphological variability in the normal tissue classes, additional non-tumorous regions from gastrectomy specimens were included.

Model validation was performed using the CRC-VAL-HE-7K dataset, which comprises 7,180 image patches extracted from H&E-stained slides of 50 patients with colorectal adenocarcinoma. The validation cohort contains no patient overlap with NCT-CRC-HE-100K-NONORM. Patches were extracted at the same resolution (224x224 pixels at 0.5 μm per pixel) and derived from FFPE tissue samples provided by the NCT tissue bank. Importantly, the tissue samples used to generate the NCT-CRC-HE-100K-NONORM and CRC-VAL-HE-7K patch datasets were independent of the patient cohorts used in this study for downstream prediction analyses.

For BRCA, a patch dataset (TCGA-BRCA-6K) was derived from TCGA-BRCA using a meta-annotation approach applied to WSIs that were excluded during slide-level quality control.[17] These slides did not overlap with the WSIs included in the prognostic analyses of this study. Image patches were extracted using the same size and resolution as in the CRC dataset (224 × 224 pixels at 0.5 μm per pixel). In total, 4,700 image patches extracted from 100 WSIs were used for training, and 1,149 patches extracted from an additional 20 WSIs were used as the test set.

PrismNet

PrismNet was developed to distill complex spatial information from histopathology into a set of interpretable, canonical categories describing tissue and microenvironmental components. Within the PathPrism framework, the number and identity of these components are flexible and can be adapted according to the available annotations or the specific clinical and biological questions under investigation. For example, when studying CRC, we defined eight commonly observed categories[34,82]: ADI (adipose tissue), DEB

(debris), LYM (lymphocytes), MUC (mucus), MUS (smooth muscle), NORM (normal colon mucosa), STR (cancer-associated stroma), and TUM (colorectal adenocarcinoma epithelium). When studying BRCA, we defined five commonly observed categories: ADI (adipose tissue), LYM (lymphocytes), NEC (necrosis), STR (cancer-associated stroma), and TUM (breast carcinoma epithelium). An additional class, BACK (background), was included to identify non-tissue regions.

To leverage the representational capacity of foundation models while reducing overfitting, PrismNet encodes image patches using a pretrained foundation model encoder, followed by dimensionality reduction of the resulting feature vectors and classification using a logistic regression module (Figure S1A). We benchmarked several foundation model encoders, including CTransPath,[83] PLIP,[84] CONCH,[84] MUSK,[85] and UNI,[37] together with two dimensionality reduction approaches, PCA and variational autoencoder (VAE). In all experiments, foundation model features were compressed to 32 dimensions (for example, UNI features from 1024 to 32). Dimensionality reduction substantially mitigated overfitting, and the UNI + PCA combination achieved the best overall performance (Figure S1C). Therefore, this configuration was used in subsequent experiments. Notably, CONCH, MUSK, and UNI showed highly similar performance, indicating that encoder choice had limited impact. In contrast, the PrismNet strategy of foundation model encoding followed by feature compression consistently outperformed direct linear probing without dimensionality reduction across all encoders.

For CRC, PrismNet was trained and validated on NCT-CRC-HE-100K-NONORM using raw, non–stain-normalized H&E images and evaluated on an independent test set (CRC-VAL-HE-7K). The NCT-CRC-HE-100K-NONORM dataset was split into training and validation subsets at a 4:1 ratio. During training, the UNI encoder parameters were frozen. PCA was fitted exclusively on the training data to avoid information leakage, and a logistic regression classifier was trained on the PCA-transformed features for multi-class classification. Early stopping based on validation performance was used to reduce overfitting. Final model evaluation was performed on the independent test set (Figure S1C). For BRCA, the same training pipeline was used, except that the NCT-CRC-HE-100K-NONORM training dataset and CRC-VAL-HE-7K test set were replaced with the training and test sets derived from TCGA-BRCA-6K.

WSI semantic segmentation

WSIs were first divided into non-overlapping tiles of size 224x224 pixels at a resolution of 0.5 μm per pixel (20x magnification), consistent with the patch parameters used in PrismNet. Each tile was then encoded using the same frozen UNI encoder as employed in PrismNet, producing a dense feature grid representation for the entire WSI (Figure S1A). This process was efficiently implemented using the open-source computational pathology pipeline STAMP.[35] Each feature vector in the grid was subsequently processed by the trained PrismNet classifier: initially projected into a lower-dimensional space via PCA, followed by classification using the logistic regression module. For each tile, a probability vector over tissue classes was obtained. By aligning these predictions with the spatial coordinates of the corresponding tiles, WSI-level tissue probability heatmaps and semantic segmentation maps were generated.

MacroNet

Model architecture

MacroNet was designed to automatically learn prognostically relevant spatial features from WSI-level tissue probability heatmaps (Figure S2A). The model architecture was based on ResNeXt50,[86] with the input channel number modified to 9 to accommodate the set of tissue classes represented in the probability maps. The modified convolutional layers served as an encoder to capture the macro-level spatial distribution and interaction features across tissue types, transforming them into a compact 2048-dimensional feature space. To enhance robustness and reduce feature dimensionality, the encoded feature vector was passed through a fully connected layer followed by batch normalization and ReLU activation, yielding a 32-dimensional representation. A final fully connected layer was used to generate a patient-level risk score. The model was trained using the negative Cox partial log likelihood loss,[87] enabling optimization for time-to-event survival outcomes. Network parameters were updated through backpropagation, allowing the model to automatically learn prognostic features from the spatial organization of tissues.

Training detail

MacroNet was trained end-to-end for 100 epochs with a mini-batch size of 128, using the Adam optimizer with a learning rate of 5×10^{-3}, $\beta_1 = 0.9$, $\beta_2 = 0.999$, and L_2 weight decay of 4×10^{-4}. A ReduceLROnPlateau scheduler was applied to reduce the learning rate by a factor of 0.2 when the validation loss failed to improve for 10 consecutive epochs. All training was performed using PyTorch and the tsai deep learning library on an NVIDIA Quadro RTX 8000 GPU with 48 GB of memory.

Benchmarking

To contextualize the performance of MacroNet, we benchmarked its prognostic ability against five representative baseline models derived from either patch-level or slide-level foundation models: attMIL-UNI,[36,37] GigaPath,[38] PRISM,[41] CHIEF,[39] and COBRA.[40] attMIL-UNI combines the frozen UNI encoder with an attention-based multiple instance learning (attMIL) framework. All tile-level features extracted by UNI from each WSI were aggregated into a patient-level representation using a lightweight attention mechanism. Specifically, each patch-level feature was first projected into a 256-dimensional embedding via a linear encoder, and attention weights were assigned to compute a weighted average of features across all tiles. The aggregated bag-level representation was then fed into a fully connected layer to produce a patient-level risk score. The model was optimized using the Adam optimizer with a learning rate of 5×10^{-3}. For slide-level foundation models, including GigaPath, PRISM, CHIEF, and COBRA, we directly used the publicly released pretrained models to extract WSI-level features. These models encode entire WSIs into a slide-level feature per patient, which was subsequently used as input to a Cox proportional hazards model implemented via

CoxnetSurvivalAnalysis with elastic net regularization (L1 ratio = 0.5).[50] The optimal regularization parameter (alpha) was selected through grid search over 25 logarithmically spaced values (0.002–0.1) using 5-fold cross-validation on the DACHS cohort.

Multi-angle interpretation

Tissue composition and attribution quantification

Patients were stratified into high- and low-risk groups based on MacroNet-predicted risk scores. Tissue composition was quantified from semantic segmentation maps by computing the proportion of each tissue type. Model attribution was assessed using gradient-based saliency maps, computed with the Captum Saliency method.[42] Attribution values were aggregated across tissue-specific masks and normalized by area to yield average class-wise attribution scores.

Inter-tissue co-occurrence and correlation analysis

Pairwise co-occurrence and correlation analyses were conducted to investigate relationships between tissue types. Co-occurrence matrices were derived from joint tissue presence within WSIs, while Spearman correlation coefficients were computed based on tissue ratios or attribution scores across patients. Analyses were stratified by the predicted risk group.

Structure-preserving counterfactual explanation

To assess the influence of individual tissue types on model predictions, a structure-preserving counterfactual explanation framework was implemented. For each sample, the probability map of a selected tissue channel was globally scaled by a factor α to simulate suppression ($\alpha < 1$) or enhancement ($\alpha > 1$), while preserving the spatial layout of other tissue classes (Figure S3). The modified probability maps were re-evaluated with trained MacroNet to obtain new risk scores. By analyzing changes in predicted risk across varying α values, the model's sensitivity to specific tissue alterations was quantified, enabling the identification of prognostically influential tissue components.

Visual pathology interpretation

Attribution maps were generated using gradient-based saliency,[42] implemented via the Saliency method from the Captum library. These maps were overlaid onto the corresponding semantic segmentation maps and aligned with the original WSIs for visual inspection. This alignment enabled examination of how model attention distributed across anatomical regions in representative high- and low-risk cases.

Spatial biomarker construction

To characterize and quantify spatial patterns from WSI-derived semantic segmentation maps, both basic and advanced spatial biomarkers were extracted (Figure S4A). Spatial fraction was defined as the proportion of each tissue type within the WSI, calculated by computing the pixel-wise ratio of each class (excluding background) from the segmentation map. Spatial entropy was computed by applying a sliding window over each tissue's binary mask and calculating local Shannon entropy. The average non-zero entropy values per tissue type were used as spatial heterogeneity features. To capture more complex spatial relationships, single-tissue and multi-tissue graphs were constructed from the segmentation maps, and topological features were extracted to describe higher-order spatial organization.

Single-tissue graph construction and feature extraction

To characterize the spatial organization within each tissue class, single-tissue graphs were constructed from WSI-derived semantic segmentation maps. For each class (excluding background), a binary mask was created and subjected to morphological erosion to remove boundary artifacts and reduce noise. Subsequently, connected component labeling was applied to identify spatially distinct tissue regions. To prevent excessively large or irregular regions from dominating graph structure, connected components exceeding an area threshold of 2,000 pixels ($\sim$0.25 cm^2) were subdivided into fixed-size blocks (30 $\times$ 30 pixels) using a sliding window approach. Each resulting spatial unit (either a small component or a subdivided block) was represented as a graph node, and its centroid coordinates were used as node positions. Edges were constructed between node pairs if their Euclidean distance was below a distance threshold of 20 pixels ($\sim$1.12 mm). This connectivity criterion was selected to reflect spatial proximity in histological context and implemented efficiently using KD-tree–based nearest neighbor search.

The resulting undirected graphs captured intra-class spatial connectivity. A set of topological and spatial features was extracted from each graph to serve as interpretable spatial biomarkers. These included node count, edge count, edge density, mean and standard deviation of inter-node distances, degree statistics (mean, max, min, coefficient of variation, entropy), clustering coefficient, number and ratio of connected components, average shortest path length, graph diameter, average area of tissue regions, spatial dispersion, nearest-neighbor ratio, centrality measures (betweenness, closeness), modularity, and degree assortativity (Table S2). All features were computed using the NetworkX package.

Multi-tissue graph construction and feature extraction

To model spatial interactions between distinct tissue types, multi-tissue graphs were constructed for each pairwise combination of tissue classes (excluding background). For each tissue pair (e.g., LYM–STR), the corresponding binary segmentation masks were extracted from the WSI-level semantic map. Morphological erosion was applied to each mask to reduce edge noise and ensure separation of closely adjacent structures. Connected component labeling was then used to identify discrete tissue regions within each class. To prevent large regions from dominating graph structure, connected components exceeding an area threshold of 2,000 pixels ($\sim$0.25 cm^2) were subdivided into non-overlapping 30 $\times$ 30-pixel blocks. All resulting regions, either small components or subdivided blocks, were treated as graph nodes, and each node was annotated with its tissue class (e.g., LYM or STR) based on origin. Edges were constructed exclusively between nodes from different tissue types, capturing inter-class spatial

proximity. Specifically, an undirected edge was added between a node from tissue A and a node from tissue B if their Euclidean distance was below a threshold of 20 pixels (~1.12 mm). KD-tree–accelerated nearest-neighbor search was used to efficiently identify such inter-tissue node pairs.

The resulting graphs were heterogeneous representations that captured the spatial architecture and interactions between specific tissue pairs. From each graph, a comprehensive set of interpretable features was extracted, including the number of nodes per tissue type, the number and density of inter-class edges, and the average and standard deviation of inter-node distances across inter-class connections. Additional features included the proportion of isolated nodes (i.e., regions with no inter-tissue connections), the average number of tissue-B neighbors per tissue-A node (and vice versa), and degree-based statistics such as mean degree, coefficient of variation, and entropy. Graph-level properties such as the number of connected components, modularity, assortativity based on tissue class, and mean betweenness centrality were also computed (Table S3). All graph construction and feature extraction procedures were implemented using the NetworkX library.

Transparent prognostic model

Spatial biomarker prognostic model

The spatial biomarker prognostic model (SPM) was constructed using Cox proportional hazards models with elastic net regularization,[50,51] trained on interpretable spatial features derived from WSI-based semantic segmentation maps. Model training was conducted using the CoxnetSurvivalAnalysis implementation (l1_ratio = 0.5, max_iter = 100,000), with 5-fold cross-validation. A grid search over 100 alpha values (log-uniform between 0.002 and 0.3) was performed, and the optimal alpha was chosen based on the highest mean validation C-index across folds. The models were trained and evaluated using disease-specific survival and event indicators, with performance measured by the C-index. All procedures were implemented in Python using scikit-survival, lifelines, and scikit-learn.

Filtered spatial biomarker prognostic model

The filtered spatial biomarker prognostic model (FSPM) was constructed using the same Cox proportional hazards framework as SPM, trained only on the top 10 spatial biomarkers with the largest absolute coefficients from the trained SPM. The FSPM was trained using the same procedure as SPM, including Z score normalization within each fold, 5-fold cross-validation, elastic net regularization (l1_ratio = 0.5, max_iter = 100,000), and grid search over 100 alpha values (log-uniform between 0.002 and 0.3). Performance was evaluated using DSS and the C-index. All procedures were implemented in Python using scikit-survival, lifelines, and scikit-learn.

Model interpretation

Model interpretation was performed by directly inspecting the learned coefficients from the trained Cox model. In this framework, each coefficient reflects the strength and direction of the association between a spatial biomarker and patient prognosis. Features with larger absolute coefficients were considered more prognostically informative. Positive coefficients indicate features associated with increased risk (hazard), while negative coefficients suggest a protective effect.

Fusion of spatial biomarkers and slide-level foundation model features

The 10 features from the FSPM were concatenated with slide-level feature vectors extracted from each foundation model (GigaPath, CHIEF, PRISM, COBRA). The combined feature sets were used as input to a Cox proportional hazards model with elastic net regularization.[50] All training procedures, including 5-fold cross-validation and log-uniform alpha selection, were identical to those used for the FSPM. To assess the relative importance of spatial biomarkers and foundation model features, we analyzed the learned coefficients from each fusion model. The contribution of each feature group was quantified as the sum of absolute coefficients, and expressed as a percentage of the total contribution.

Mutation prediction model

The mutation prediction model was constructed using L1-penalized logistic regression, trained on all interpretable spatial biomarkers derived from WSI-based semantic segmentation maps. For each mutation status (e.g., MSI, BRAF, TP53, BMPR2), a separate model was trained using the LogisticRegression implementation from scikit-learn (penalty = 'l1', solver = 'liblinear', C = 0.1, max_iter = 100). The model output was the predicted mutation probability, and coefficients were used to assess feature-level interpretability. All procedures were implemented in Python using scikit-learn and pandas.

Adjuvant chemotherapy biomarker discovery

To identify spatial biomarkers predictive of adjuvant chemotherapy (ACT) benefit in stage II/III colorectal cancer, both continuous and dichotomized interaction analyses were performed using Cox proportional hazards models. In the continuous setting, each spatial biomarker was standardized by mean-centering. ACT status was binarized, and a Cox model was fitted with three covariates: ACT (binary), the centered biomarker, and their interaction term (ACT × biomarker). Patients with high missingness (>80%), insufficient variation, or small subgroup sizes were excluded. Model fitting was conducted using the lifelines Python package, and p-values were computed via the Wald test from the model summary statistics. The prognostic relevance of the interaction was evaluated based on the interaction term's hazard ratio and corresponding p-value. For biomarkers with suggestive interaction signals ($p < 0.1$), a dichotomized analysis was conducted. Each biomarker was binarized using its cohort-specific median, and a new

Cox model was fitted including ACT status, the binary biomarker (risk group), and their interaction (ACT × risk). Wald test p-values were again used to quantify the significance of the interaction term, and effect sizes were reported as hazard ratios with 95% confidence intervals.

LLM co-reasoning

Mechanistic reasoning was performed using GPT-4o based on spatial biomarkers identified from cohort-level prognostic analyses.[88] Each session was initiated with a concise prompt detailing the biomarker's tissue interaction pattern, associated histological phenotype, and relevant clinical context. The model was prompted to interpret the biological relevance, propose mechanistic links to treatment response or genomic alterations, and suggest follow-up experimental strategies. Outputs were independently reviewed by a second GPT-4o agent acting as a scientific auditor to evaluate plausibility, internal coherence, and alternative hypotheses. To validate the reasoning, five domain experts in colorectal cancer pathology and translational oncology independently rated selected outputs for scientific soundness, experimental feasibility, and clinical relevance (Figure S8). All reasoning was conducted without fine-tuning or retrieval augmentation.

VirtualWSI

VirtualWSI enables structure-preserving in silico perturbation of tissue organization in WSIs based on the semantic segmentation maps generated by PathPrism. By modifying tissue probabilities while preserving spatial architecture, the framework allows systematic exploration of how tissue configurations influence spatial biomarkers and model-derived risk predictions. Two complementary perturbation strategies were implemented: global channel perturbation, which supports population-level perturbation analyses across cohorts, and interactive spatial perturbation, which enables hypothesis-driven exploration within individual samples.

Global channel perturbation

To investigate how perturbations of specific tissue categories influence spatial biomarkers and predicted risk, VirtualWSI performs WSI-wide channel perturbation on the probability maps produced by PrismNet. Each channel represents the probability of a specific tissue category, and the final segmentation map is obtained by assigning each pixel to the category with the highest probability across channels. Controlled perturbations are introduced by scaling the probability values of a selected tissue channel by a constant factor α (Figure S3A). Increasing the channel values ($\alpha > 1$) expands the corresponding tissue regions in the segmentation map, whereas decreasing them ($\alpha < 1$) leads to contraction of that tissue class (Figure S3B). Because final tissue labels are determined by selecting the maximum probability across channels, this operation preserves valid segmentation assignments while modulating the relative dominance of the perturbed tissue class. Operating directly on the probability maps allows the perturbation to preferentially affect regions with higher uncertainty, such as tissue boundaries, thereby preserving the overall tissue architecture of the slide. This procedure requires no additional model training or manual annotation and therefore enables efficient population-level perturbation analyses. The resulting perturbed segmentation maps can be used to recompute spatial biomarkers and evaluate the corresponding FSPM-predicted risk scores.

Interactive spatial perturbation

To support hypothesis-driven exploration, VirtualWSI also enables interactive perturbation of specific spatial regions within individual samples. Users can modify the semantic segmentation maps derived from PathPrism through a graphical interface by selecting regions of interest and overwriting the assigned tissue categories (for example, converting LYM to MUS or expanding TUM into adjacent tissues). Conceptually, this approach follows the same perturbation principle as global channel perturbation but restricts the perturbation to user-defined spatial regions. In practice, perturbations are applied through spatial masks so that only the selected regions are modified. This enables fine-grained, spatially localized perturbations while preserving the overall tissue architecture. Although this strategy enables precise exploration of spatial dependencies, it requires manual interaction and is therefore less suited for large-scale population-level analyses. As with global perturbations, the perturbed segmentation maps can be used to recompute spatial biomarkers and evaluate the corresponding FSPM-predicted risk scores.

Classification of dynamic response archetypes

Dynamic response archetypes were defined based on the dose–response trajectory of the LYM–MUC–CC biomarker under in silico perturbation. For each patient, Spearman rank correlation (ρ) between log-transformed perturbation strength (α) and LYM–MUC–CC values was calculated to capture the dominant trend direction. Patients were classified as monotonic increase ($\rho > 0.6$ and $p < 0.05$), monotonic decrease ($\rho < -0.6$ and $p < 0.05$), or flat (remaining cases). To avoid spurious classification driven by minimal variation, trajectories with small amplitude (Δ below the 25th percentile of the cohort-wide Δ distribution) were conservatively assigned to the flat category. Archetype classification was performed independently of survival and treatment outcome data.

Construction of interpretable spatial biomarker atlas

To characterize the organization of tissue phenotypes and visualize perturbation-induced state transitions, we constructed two spatial biomarker atlases from WSI-level spatial features derived from PathPrism.

Global spatial biomarker atlas

A global spatial biomarker atlas was constructed using the full set of spatial biomarker features describing tissue composition, spatial interactions, and structural organization (Figure 7A). For each WSI, the spatial biomarker spectrum was computed from the PrismNet segmentation maps. The resulting feature vectors were Z score normalized across samples and embedded into a two-dimensional manifold using UMAP. Each WSI was represented as a point in the resulting embedding space, where relative positioning reflects

similarity in spatial biomarker profiles. This atlas provides a population-level view of tissue phenotypes and was used to visualize cohort structure as well as perturbation-induced transitions in spatial biomarker space.

Prognostic spatial biomarker atlas

To examine perturbation effects in relation to prognostic signals, we constructed a second atlas using the ten interpretable spatial biomarkers selected by the FSPM prognostic model (Figure S10B). UMAP was applied to the *Z* score normalized values of these biomarkers to generate a two-dimensional embedding in which each WSI was positioned according to its prognostically relevant spatial configuration. Samples were visualized according to their predicted FSPM risk scores, enabling interpretation of how spatial tissue configurations relate to predicted risk. Individual biomarkers were further visualized across the embedding space to reveal regions enriched for specific spatial configurations, such as elevated ADI–TUM interaction, STR-dominant environments, or compact tumor architecture (Figure S10C).

Projection of perturbed samples

Perturbed WSIs generated by VirtualWSI were projected into both atlases by recomputing spatial biomarker features from the perturbed segmentation maps and applying the same UMAP transformations used for the original WSIs. This allowed direct visualization of perturbation-induced trajectories in spatial biomarker space and comparison between original and perturbed tissue states. All computations were implemented in Python using scikit-learn, matplotlib, and Plotly.

QUANTIFICATION AND STATISTICAL ANALYSIS

Model performance was evaluated using balanced accuracy, macro F1 score, macro AUROC, and Matthews correlation coefficient (MCC) for classification tasks, and concordance index (C-index) for survival prediction. Confidence intervals (95%) were estimated using nonparametric bootstrapping with 1,000 resamples. Survival analyses were conducted using Cox proportional hazards models. Kaplan–Meier curves were used for visualization, and group differences were assessed using log rank tests. Interaction effects were evaluated using Wald tests. For compositional data (tissue ratios), values were transformed using centered log-ratio (CLR) transformation prior to statistical testing. Group comparisons of tissue ratios and attribution values were performed using Welch's t-tests with Benjamini–Hochberg false discovery rate (FDR) correction. Spatial entropy was compared using the Mann–Whitney–Wilcoxon test. 5-fold cross-validation was used to assess model performance, and the resulting distributions of C-index are presented using boxplots, where the center line indicates the median, the box bounds represent the interquartile range (IQR), and whiskers extend to 1.5× IQR.