GECKO: Gigapixel Vision-Concept Contrastive Pretraining in Histopathology

Supplementary Material

The supplementary presents the following materials:

- Generalizability Evaluation (Sec. 7)
- Additional ablations (Sec. 8)
- Additional implementation details (Sec. 9)
- Additional few-labels supervised evaluation (Sec. 10)
- Interpretability analysis (Sec. 11)

7. Generalizability Evaluation

Table 5 presents the generalization ability of GECKO compared to Intra and TANGLE pretraining. When only the WSI modality is available, both $\text{GECKO}_{\text{deep}}$ and $\text{GECKO}_{\text{ensemble}}$ significantly outperform at lower k values and maintain superior performance at higher k values. Additionally, when the gene modality is included, our performance matches that of TANGLE. Importantly, with gene modality in GECKO pretraining, the interpretable WSI-level concept embedding consistently outperforms Intra pretraining, even on out-of-domain datasets. This demonstrates the potential to develop powerful aggregators that leverage multiple modalities for pretraining, offering inherently interpretable predictions that can build trust in clinical settings.

	Methods	Embedding	$k=5$	$k=10$	$k=25$
WSI only	Intra [14]	deep	92.8 ± 2.1	95.0 ± 1.1	96.6 ± 0.7
	GECKO	deep	95.7 ± 1.2	96.2 ± 0.6	97.2 ± 0.7
		concept	93.0 ± 2.3	94.5 ± 1.8	95.7 ± 1.0
		ensemble	$\mathbf{96.3 \pm 1.0}$	$\mathbf{96.9 \pm 0.9}$	$\mathbf{97.3 \pm 0.8}$
WSI + Gene	TANGLE [14]	deep	97.0 ± 0.6	97.6 ± 0.6	$\mathbf{98.3 \pm 0.3}$
	GECKO	deep	$\mathbf{97.2 \pm 0.7}$	$\mathbf{97.7 \pm 0.6}$	$\mathbf{98.3 \pm 0.3}$
		concept	95.9 ± 1.2	95.7 ± 1.6	96.6 ± 0.7
		ensemble	97.0 ± 0.8	97.4 ± 0.6	97.9 ± 0.3

Table 5. Few Labels (out-of-domain) classification on binary CPTAC-Lung task. All AUCs are with linear probing. CONCH is used for extracting deep features.

8. Additional ablations

1. **WSI Concept-Encoding branch architecture:** By default, our dual-branch MIL uses the ABMIL [12] aggregator for the deep-encoding branch and a self-interpretable aggregator, inspired from SI-MIL [17], for the concept-encoding branch. For the ablation study, we replace the self-interpretable aggregator with an ABMIL in the concept-encoding branch that learns its own concept attention without reliance on the deep-encoding branch. As shown in Table 6, our default dual-branch MIL (referred as GECKO in the Table) consistently outperforms the variant using a ABMIL for both branches (referred to as Dual-ABMIL). Note that we removed the projector $H(\cdot)$ in the ABMIL for the concept prior to enforce linear aggregation and thus preserve interpretability.

Methods	Embedding	LUAD vs. LUSC $k=10$	LUAD vs. LUSC $k=25$	EBV+MSI vs. Others $k=10$	EBV+MSI vs. Others $k=25$
Dual-ABMIL [12]	concept	88.2 ± 0.7	90.3 ± 0.9	72.3 ± 5.9	73.8 ± 6.0
	ensemble	92.3 ± 0.7	94.6 ± 1.0	75.1 ± 6.0	78.0 ± 7.1
GECKO	concept	93.5 ± 1.3	94.6 ± 1.5	78.4 ± 3.8	80.3 ± 6.1
	ensemble	$\mathbf{95.3 \pm 0.9}$	$\mathbf{96.5 \pm 1.1}$	$\mathbf{79.8 \pm 4.8}$	$\mathbf{82.5 \pm 7.4}$

Table 6. WSI Concept-Encoding branch architecture. All AUC results reported with linear probing, and pretraining with WSI only. CONCH is used for extracting deep features.

2. **Effect of false negative elimination with keep ratio (r_{keep}):** In Figure 4, we demonstrate the effect of r_{keep} for false negative elimination [10] in contrastive pretraining across all five TCGA tasks. We report the performance of GECKO-Zero in an unsupervised 5-fold cross validation setting. We observed that the default contrastive pretraining with $r_{keep} = 1$ consistently results in poor performance. We attribute this to the fact that GECKO performs contrastive learning in C-dimensional embedding space, which is significantly smaller than a typical embedding size (256 or higher); thus, potentially contrasting WSIs with similar concept activations and introducing noise. Recall that, we project the WSI-level deep embedding to match the dimension of the WSI-level concept embedding before alignment. Empirically, $r_{keep} = 0.7$ consistently performed well across tasks, thus we fix r_{keep} as 0.7 for our experiments.

9. Implementation details

Pretraining setting. We pretrained our dual-branch MIL using GECKO for 50 epochs for all tasks with a learning rate of 1e-4. A warmup is applied for 5 epochs, increasing the learning rate from 1e-8 to 1e-4, followed by a cosine scheduler that decays the rate to 1e-8, consistent with TANGLE [14]. The same settings were used to train TANGLE and Intra for all comparisons, with a batch size of 64 for all pretraining methods.

Linear probing setting. For training the linear classifier across all methods, we use the same configuration

as above. Specifically, we train LogisticRegression classifier from `sklearn` with default parameters and set the number of iterations to 10,000.

Gene modality setting. For gene expression data, we adopt the curation method in [14, 34], resulting in 4,848 gene expressions per case across all datasets. To integrate the gene modality into GECKO, we employ the same MLP-based architecture as in TANGLE. We perform K-way contrastive alignment by aligning each pair of modalities. To contrast with the concept prior, we use a projection head prior to alignment on the deep- and the gene-encoding branches to match the output dimension C from the concept-branch. We directly align the outputs from the gene- and deep-encoding branches without additional projection, following TANGLE's design. Consequently, we optimize three losses in this multimodal setting of GECKO, which we average without any hyperparameter tuning.

10. Additional few-label setting evaluation

In Figure 5, we present the results of few-label supervised evaluation in the linear probing setting for the EBV vs. Others and BRCA datasets. In the unimodal setting with only WSI data (indicated by dashed lines), our $\text{GECKO}_{\text{ensemble}}$ significantly outperforms the Intra pretraining on the EBV vs. Others task, while performing on par for HER2 prediction. In the multimodal setting, where gene data is available (indicated by solid lines), $\text{GECKO}_{\text{ensemble}}$ pretrained with the gene modality alongside WSIs and our concept prior slightly outperforms TANGLE on the EBV vs. Others task, while achieving comparable performance on the HER2 prediction task.

11. Interpretability analysis

In Fig. 6 and 7, we illustrate the Top-K salient patches and the WSI-level concept activations produced by our GECKO-pretrained model for TCGA-Lung and TCGA-STAD in an unsupervised setting. In the WSI-level concept activation bar plots, we quantitatively demonstrate that for a WSI belonging to a particular class, our model not only identifies the important patches but also provides the WSI-level activation for each concept through its interpretable concept embedding M_{wsi}. Notably, the concepts with the highest activations align with those that are most relevant to the corresponding class, evaluated by a pathologist. In Table 8–12, we provide the concepts for each task along with their detailed descriptions, that were used as input to the text encoder of the CONCH model, in line with ConcepPath [42].

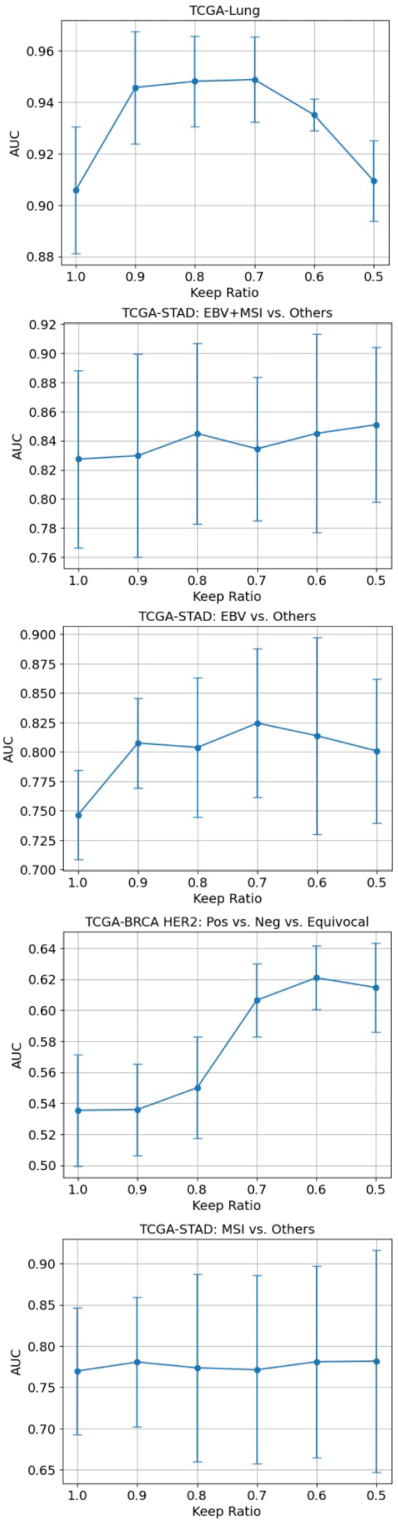

Figure 4. Effect of false negative elimination keep ratio (r_{keep}). All AUC values are reported in an unsupervised setting (5-fold cross validation) using our proposed heuristic. $r_{keep} = 0.7$ was found to work consistently well across all tasks.

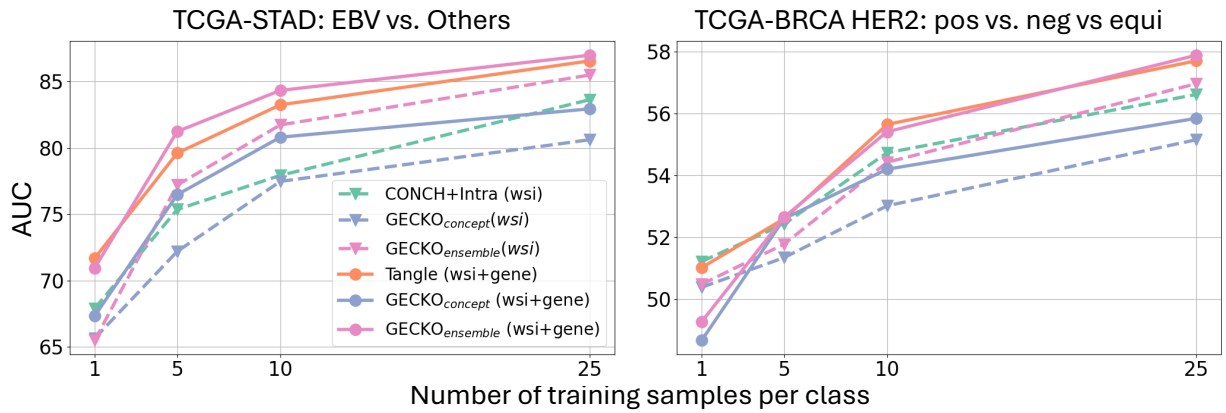

Figure 5. Few Labels (in domain) classification analysis. All AUC results are with linear probing. Dashed lines represent pretraining on WSI only, and solid lines represents multimodal pretraining with gene data. CONCH is used for extracting deep features.

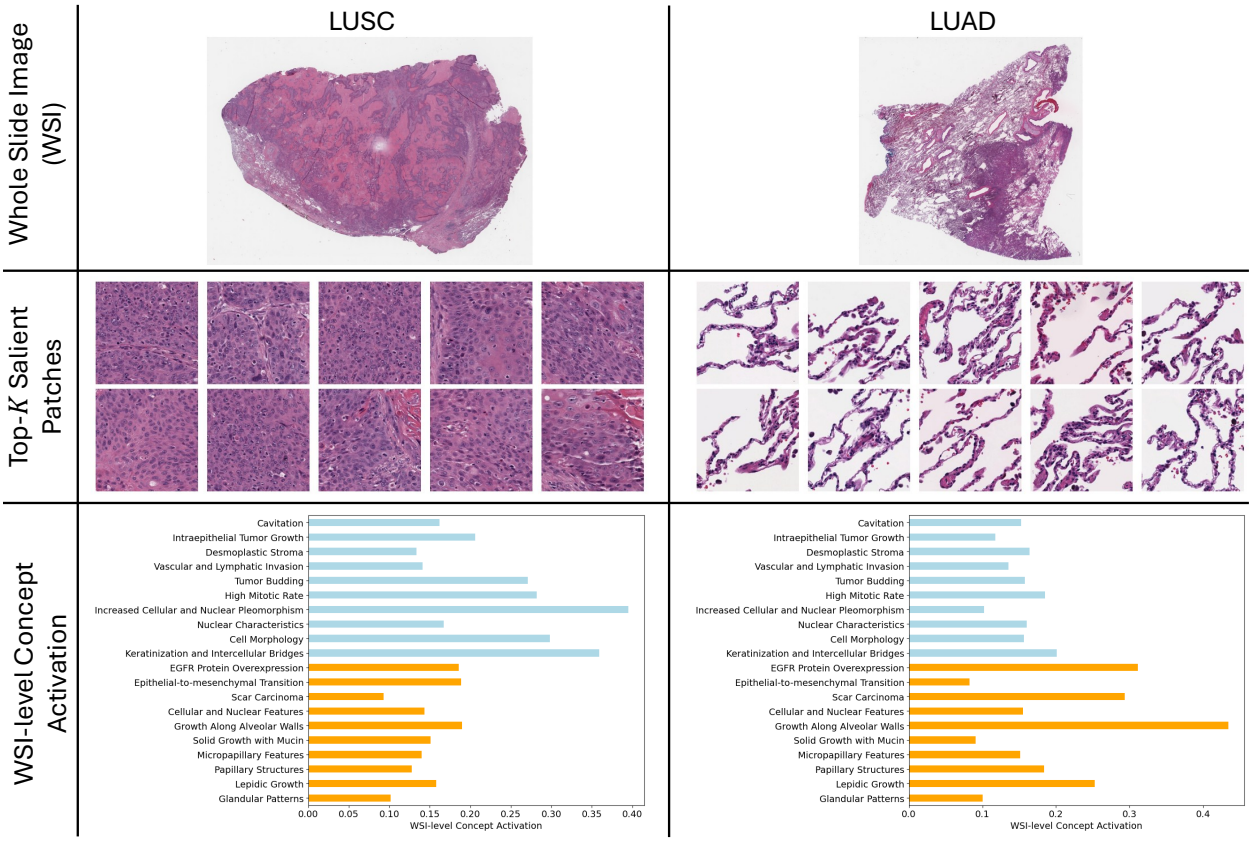

Figure 6. TCGA-Lung: LUSC vs. LUAD. Row 1 shows sample WSIs from LUSC and LUSC subtypes in TCGA-Lung. Row 2 shows the Top-K patches selected by our GECKO pretrained model. Row 3 illustrates the WSI-level aggregated concept activation (from interpretable concept embedding).

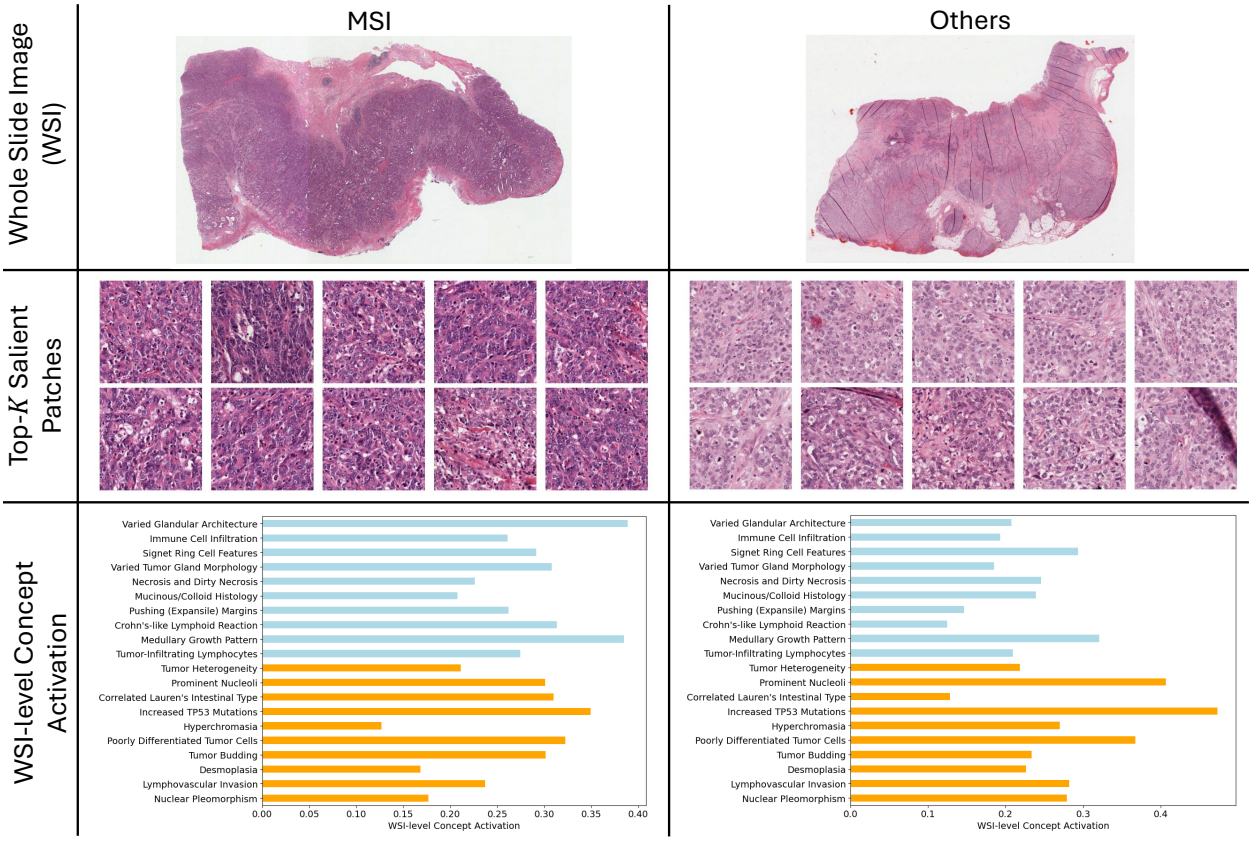

Figure 7. TCGA-STAD: MSI vs. Others. Row 1 shows sample WSIs from MSI and Others class in TCGA-STAD. Row 2 shows the Top-K patches selected by our GECKO pretrained model. Row 3 illustrates the WSI-level aggregated concept activation (from interpretable concept embedding).

Dataset	**#WSIs**	**Class name (#WSIs)**
TCGA-Lung	1042	LUAD: Lung adenocarcinoma (530) LUSC: Lung squamous cell carcinoma (512)
TCGA-BRCA	933	HER2-positive (164) Equivocal (186) HER2-Negative (583)
TCGA-STAD	268	EBV: Epstein-Barr virus (26) MSI: Microsatellite Instability (44) GS:Genomically Stable/CIN: Chromosomally Instable (199)
CPTAC-Lung	1091	LUAD: Lung adenocarcinoma (578) LUSC: Lung squamous cell carcinoma (513)

Table 7. Datasets (with class distribution) used for evaluation .

Type	Concept	Description
LUAD	Glandular Patterns	Gland-like structures; tubular; acinar; papillary formations; lined by atypical cells; mucin production;
	Lepidic Growth	Alveolar growth pattern; non-invasive; early adenocarcinomas; minimally invasive adenocarcinomas;
	Papillary Structures	Papillary architecture; fibrovascular cores; malignant cells lining; mucin content;
	Micropapillary Features	Micropapillary pattern; small cell clusters; no fibrovascular core; clear spaces from tissue processing;
	Solid Growth with Mucin	Solid growth pattern with mucin; mucicarmine; periodic acid-Schiff stains usage;
	Growth Along Alveolar Walls	Lepidic growth pattern; tumor cells along alveolar walls; non-invasive;
	Cellular and Nuclear Features	Cell morphology variable; cuboidal to columnar shape; hobnail appearance; pleomorphic nuclei; prominent nucleoli;
	Scar Carcinoma	Association with lung scarring or fibrosis; possible misdiagnosis on imaging; requires biopsy;
	Epithelial-to-mesenchymal Transition	E-cadherin staining decrease; mesenchymal markers increase; cytoplasmic/membranous staining; EMT at invasive front;
	EGFR Protein Overexpression	EGFR expression; membranous staining; possible cytoplasmic staining; cell membrane receptor;
LUSC	Keratinization and Intercellular Bridges	Squamous differentiation; keratin production; keratin pearls; intercellular bridges;
	Cell Morphology	Polygonal tumor cells; abundant eosinophilic cytoplasm; high keratin content;
	Nuclear Characteristics	Hyperchromatic nuclei; prominent nucleoli; variable pleomorphism;
	Increased Cellular and Nuclear Pleomorphism	Cellular and nuclear pleomorphism; increased variability; indicative of higher malignancy; IHC highlighted;
	High Mitotic Rate	High mitotic figure count; rapid cell proliferation; visualized by mitotic markers;
	Tumor Budding	Tumor budding presence; aggressive behavior indicator; cytokeratin stains highlight;
	Vascular and Lymphatic Invasion	Tumor cells in blood vessels or lymphatics; potential for metastasis; CD31 and podoplanin (D2-40) markers;
	Desmoplastic Stroma	Reactive stromal response; dense fibrous stroma surrounding tumor cells;
	Intraepithelial Tumor Growth	Intraepithelial growth; tumor spread within epithelial structures;
	Cavitation	Cavitation; central necrosis; more common in squamous cell carcinoma; visible on imaging;

Table 8. Pathology concepts for LUAD vs. LUSC

Type	Concept	Description
EBV+MSI	Lymphoepithelioma-like Histology	EBV-positive; lymphoepithelioma-like carcinoma; large sheets; syncytial clusters; undifferentiated cells; prominent lymphoid infiltration; no glandular formation; non-keratinizing; vesicular nuclei; prominent nucleoli; desmoplastic reaction;
	Syncytial trabecular pattern	Syncytial trabecular pattern; nested growth; cord-like structures; indistinct cell borders; interconnected net-like structure;
	Tumor Infiltrating Lymphocytes	Tumor-infiltrating lymphocytes; dispersed or clustered; infiltrating between cells or stromal; indicative of immune response;
	Intraepithelial Lymphocytosis	Intraepithelial lymphocytes; small, round; dense nuclei; disrupts architecture; associated with neoplastic epithelium; stromal lymphoplasmacytic infiltration;
	Stromal Lymphoplasmacytic Infiltration	Lymphocytes in stroma; plasma cells present; small cells with large nuclei; abundant basophilic cytoplasm; interspersed infiltration; reactive changes; possible fibrosis or edema;
	Medullary Growth Pattern	Carcinomas; colorectal; MSI-H status; high neoantigen load; poorly differentiated; syncytial growth; abundant intraepithelial lymphocytes; dMMR tumors; solid sheets of cells;
	Crohn's-like Lymphoid Reaction	Dense lymphoid aggregates; tumor margin; robust immune response; neoantigens; dMMR tumors; Crohn's-like reaction;
	Pushing (Expansile) Margins	Expansive growth pattern; pushing borders; high neoantigen levels; immune containment; dMMR tumors; non-infiltrative margin; microsatellite stable (MSS) tumors contrast;
	Pattern of Infiltration	Vigorous immune infiltrate; variable PD-L1 positive cell distribution; invasive tumor margins; tumor nests; 'brisk' infiltration pattern; T cell band at tumor margin; 'non-brisk' infiltration pattern; scattered T cells throughout tumor;
	Immune Cell Infiltration	Significant number; lymphocytes; tumor tissue presence;
Others	Nuclear Pleomorphism	Variation in nuclear size and shape; nuclei size disparity; irregular nuclear shapes; oval to highly irregular forms;
	Hyperchromasia	Nuclei appear darker; excess DNA content;
	Irregular Nuclear Contours	Uneven nuclear borders; indented nuclear contours;
	Prominent Nucleoli	Prominent nucleoli; increased number of nucleoli; sign of heightened protein synthesis; rapid cell division indicator;
	Chromatin Clumping	Irregular chromatin clumping; patchy nuclear appearance;
	Multipolar spindles	Multipolar spindles; asymmetric nuclear division; uneven genetic material distribution; cells with abnormal nuclear shapes and sizes;
	Lymphovascular Invasion	Tumor cells in lymphatic vessels; tumor cells in blood vessels; direct indication of metastasis;
	Tumor Budding	Small clusters of cancer cells at invasive front; individual cells at invasive front; sign of aggressive tumor phenotype; correlated with metastasis;
	Desmoplasia	Pronounced desmoplastic reaction; growth of fibrous tissue; connective tissue increase; association with aggressive tumors;
	Signet Ring Cells	Loss of E-cadherin function; CDH1 mutations; presence of signet ring cells; large vacuole in cells; nucleus at periphery; signet ring-like appearance; indicative of poor prognosis;

Table 9. Pathology concepts for EBV+MSI vs. Others

Type	Concept	Description
MSI	Tumor-Infiltrating Lymphocytes	High neoantigen load; immune cell infiltration; tumor tissue response; neoantigen presentation; dMMR tumors; prominent lymphocytic response; high TIL density; immune response to neoantigens;
	Medullary Growth Pattern	Carcinomas; colorectal; MSI-H status; high neoantigen load; poorly differentiated; syncytial growth; abundant intraepithelial lymphocytes; dMMR tumors; solid sheets of cells;
	Crohn's-like Lymphoid Reaction	Dense lymphoid aggregates; tumor margin; robust immune response; neoantigens; dMMR tumors; Crohn's-like reaction;
	Pushing (Expansile) Margins	Expansive growth pattern; pushing borders; high neoantigen levels; immune containment; dMMR tumors; non-infiltrative margin; microsatellite stable (MSS) tumors contrast;
	Mucinous/Colloid Histology	Abundance; extracellular mucin production; MSI-H tumors;
	Necrosis and Dirty Necrosis	High neoantigen loads; necrosis; cytotoxic immune response; tumor necrosis; 'dirty necrosis'; debris; nuclear dust; dMMR tumors commonality;
	Varied Tumor Gland Morphology	dMMR tumors; heterogeneous morphology; varied gland shapes; varied gland sizes; poor differentiation;
	Signet Ring Cell Features	Mucin-filled cells; peripheral nucleus; indicative of MSI-H; gastric cancer;
	Immune Cell Infiltration	Significant number; lymphocytes; tumor tissue presence;
	Varied Glandular Architecture	Disorganized structure; irregular gland formation; varied gland sizes; MSI-H tumors;
Others	Nuclear Pleomorphism	Variation in nuclear size and shape; nuclei size disparity; irregular nuclear shapes; oval to highly irregular forms;
	Lymphovascular Invasion	Tumor cells in lymphatic vessels; tumor cells in blood vessels; direct indication of metastasis;
	Desmoplasia	Pronounced desmoplastic reaction; growth of fibrous tissue; connective tissue increase; association with aggressive tumors;
	Tumor Budding	Small clusters of cancer cells at invasive front; individual cells at invasive front; sign of aggressive tumor phenotype; correlated with metastasis;
	Poorly Differentiated Tumor Cells	High-grade dedifferentiation; higher likelihood of metastasis;
	Hyperchromasia	Nuclei appear darker; excess DNA content;
	Increased TP53 Mutations	TP53 enrichment in high-CIN tumors; link to mitotic stress; TP53 malfunctions; increased mitotic figures in histology; atypical nuclear features; increased nuclear size; irregular nuclear contours; hyperchromasia; prominent nucleoli; genomic instability; altered cell cycle regulation; variety of cell types; abnormal tumor structures;
	Correlated Lauren's Intestinal Type	Well-formed glandular structures; intestinal epithelium resemblance; chronic gastritis initiation; progression to atrophy; intestinal metaplasia; dysplasia; carcinoma development; common in high-incidence regions; environmental factor association; diet-related; Helicobacter pylori infection;
	Prominent Nucleoli	Prominent nucleoli; increased number of nucleoli; sign of heightened protein synthesis; rapid cell division indicator;
	Tumor Heterogeneity	CIN-induced genetic heterogeneity; RAS-driven proliferation of diverse cells; increased tumor complexity; potential influence on drug resistance; enhancement of metastatic potential;

Table 10. Pathology concepts for MSI vs. Others.

Type	Concept	Description
EBV	Lymphoepithelioma-like Histology	EBV-positive; lymphoepithelioma-like carcinoma; large sheets; syncytial clusters; undifferentiated cells; prominent lymphoid infiltration; no glandular formation; non-keratinizing; vesicular nuclei; prominent nucleoli; desmoplastic reaction;
	Tumor Infiltrating Lymphocytes	Tumor-infiltrating lymphocytes; dispersed or clustered; infiltrating between cells or stromal; indicative of immune response;
	Intraepithelial Lymphocytosis	Intraepithelial lymphocytes; small, round; dense nuclei; disrupts architecture; associated with neoplastic epithelium; stromal lymphoplasmacytic infiltration;
	Stromal Lymphoplasmacytic Infiltration	Lymphocytes in stroma; plasma cells present; small cells with large nuclei; abundant basophilic cytoplasm; interspersed infiltration; reactive changes; possible fibrosis or edema;
	Syncytial trabecular pattern	Syncytial trabecular pattern; nested growth; cord-like structures; indistinct cell borders; interconnected net-like structure;
	Lace-like Pattern	Lace-like pattern; irregularly anastomosing tubules and cords; complex interconnected network; irregular net-like structure;
	Lymphoid Stroma	Lymphoid stroma infiltration; "lace-like" pattern; irregular tubules and cords; immune component in microenvironment; variable lymphoid infiltration;
	Invasion into the Submucosa	Invasion into submucosa; scattered cells to clusters; neoplastic cells breach muscularis mucosae; lymphocytic response around cancer cells;
	Poor Differentiation	Poorly differentiated adenocarcinomas; lacks specialized features; aggressive tumor; unformed glandular structures; infiltrating lymphoid stroma;
	Ulcered or saucer-like tumor	Central necrosis; ulceration with epithelial loss; robust inflammatory infiltrate; reactive cellular changes; marginal roll at ulcer edges; increased vascularity; surrounding fibrosis;
Others	Increased Mitotic Activity	Increased mitotic rate; atypical mitotic figures; abnormal mitoses; high cellular proliferation;
	Nuclear Pleomorphism	Variation in nuclear size and shape; nuclei size disparity; irregular nuclear shapes; oval to highly irregular forms;
	Hyperchromasia	Nuclei appear darker; excess DNA content;
	Irregular Nuclear Contours	Uneven nuclear borders; indented nuclear contours;
	Prominent Nucleoli	Prominent nucleoli; increased number of nucleoli; sign of heightened protein synthesis; rapid cell division indicator;
	Chromatin Clumping	Irregular chromatin clumping; patchy nuclear appearance;
	Multipolar spindles	Multipolar spindles; asymmetric nuclear division; uneven genetic material distribution; cells with abnormal nuclear shapes and sizes;
	Tumor Budding	Tumor budding presence; aggressive tumor phenotype; correlated with metastasis;
	Lymphovascular Invasion	Tumor cells in lymphatic vessels; tumor cells in blood vessels; direct indication of metastasis;
	Desmoplasia	Pronounced desmoplastic reaction; growth of fibrous tissue; connective tissue increase; association with aggressive tumors;

Table 11. Pathology concepts for EBV vs. Others

Type	Concept	Description
Positive	HER2 Overexpression	Strong; complete membrane staining; indicative of HER2 positivity;
	High Tumor Cellularity	Densely packed cells; high nuclear-to-cytoplasmic ratio; scant stroma; 'blue' appearance from dense nuclear staining;
	Mitotic Figures	Numerous in aggressive tumors; cells in division; high proliferation rate;
	Necrosis	Dead cell areas; cell debris; lost tissue architecture; outpaced blood supply;
	Pleomorphism	Variation in size and shape of cells and nuclei;
	High Tumor-infiltrating Lymphocytes Levels	Inferred from H&E sections; small, round, darkly stained nuclei; scant cytoplasm;
	Dense Clustering	Large, densely packed cellular areas on H&E; high nuclear to cytoplasmic ratio; minimal stroma;
	Loss of E-Cadherin	Negative staining pattern; distinguishes lobular from ductal carcinoma;
	GCDFP-15 Positive	Cytoplasmic granular staining; secreted protein indicating apocrine differentiation;
	Nuclear Markers	High density of nuclei; ER, PR, Ki-67, p53 staining; Ki-67 shows high proliferation index;
Negative	HER2 Protein Regular	No membrane staining or $\leq$10% staining; partial membrane staining in $\geq$10% of cells;
	Hormone Receptor Negative	No nuclear staining for ER/PR; consistent absence across cancer cells; uniform lack of staining;
	ER Negative	No nuclear staining; antibodies against ER don't bind;
	PR Negative	No nuclear staining; antibodies against PR don't bind;
	K67 Proteins	Nuclear staining; marks cell proliferation; absent in non-proliferative cells;
	DDR (DNA damage response) Effective	Lack/reduced expression; indicative of defective DNA repair; susceptibility to DDR inhibitors;
	Blood Vessel Density	CD31 or CD34 positive staining; lines blood vessels; increased density indicates active angiogenesis;
	Increased EMT (epithelial-mesenchymal transition)	Increased expression; suggestive of metastasis facilitation;
	Tumor Cell Invasion	Increased expression; indicates invasive potential;
	Vimentin Positive	Cytoplasmic staining; mesenchymal cell cytoskeletal component;
Equivocal	IHC Score 2+	No staining; faint staining; $\leq$ 10% tumor cells;
	HER2 Low Expression	Faint staining; barely perceptible; $\geq$10% tumor cells;
	HER2 Ultra-Low Expression	Weak to moderate staining; complete membrane; $\geq$ 10% tumor cells; Strong staining; complete membrane; $\leq$10% tumor cells;
	Heterogeneity	Variable HER2 expression; within the same tumor; challenging determination;
	Variable Staining Intensity	Variable intensity; across tumor areas; some regions stronger than others;
	Identified Invasive Tumor	Spread into tissues; potentially worse prognosis; beyond ducts/lobules;
	Moderate Tumor Proliferation	Lower than HER2-positive; higher than HER2-negative; complete but moderate membrane staining;
	Moderate Tumor Grading	Moderate uniformity; variable intensity and completeness; across tumor population;
	Metastatic Focus	Clusters of atypical cells; different from lymphoid cells; IHC markers highlight cancer cells;
	Moderate Residual Cancer Burden (RCB)	$\geq$10% tumor cells; weak/moderate intensity;

Table 12. Pathology concepts for Positive vs. Negative vs. Equivocal in BRCA HER2 prediction task