Delft University of Technology

Faculty of Electrical Engineering, Mathematics and Computer Science

MASTER'S THESIS

AUGUR: Interpretable Breast Cancer Molecular Subtyping from H&E Whole-Slide Images via Attention-Based Multiple-Instance Learning with Auxiliary Mutational-Signature Supervision

by

Ho Thi Ngoc Phuong

in partial fulfillment of the requirements for the degree of

Master of Science in Data Science and Artificial Intelligence Technology

at Delft University of Technology,
to be defended publicly on Monday, 13 July 2026.

Supervisor:	Marcel J.T. Reinders
Thesis committee:	Marcel J.T. Reinders, Jie Yang

AUGUR: Interpretable Breast Cancer Molecular Subtyping from H&E Whole-Slide Images via Attention-Based Multiple-Instance Learning with Auxiliary Mutational-Signature Supervision

Ho Thi Ngoc Phuong and Marcel J.T. Reinders

***Abstract*—**

Motivation. PAM50 molecular subtype guides breast-cancer treatment but requires sequencing or multi-gene assays that are costly, slow, and often unavailable. Routine H&E histology, by contrast, is universal, and because subtype-driving alterations leave morphological traces, the image itself may carry usable signal. We ask whether H&E whole-slide images can predict molecular subtype as a low-cost triage that prioritizes which patients to send for confirmatory molecular testing, provided each prediction carries interpretable visual evidence rather than a black-box score.

Approach. We present AUGUR, which predicts PAM50 subtype from a whole-slide image and localizes the tissue behind each call. Because slides are gigapixel with only slide-level labels, AUGUR adopts the standard weakly-supervised decomposition into two stages: a tile encoder first learns histology features through multi-task pretraining, then an attention-based multiple-instance-learning (MIL) aggregator pools tiles into a slide-level prediction. Our aggregator, DualCLAM, builds on CLAM [1], an existing model that scores tiles with per-class attention to form a slide representation and adds an instance-clustering loss that refines tile features from slide-level labels alone. DualCLAM augments it with auxiliary mutational-signature regression heads that contribute a molecular training signal and per-signature spatial explanations.

Results. On the TCGA-BRCA cohort under 5-fold stratified cross-validation, the encoder's pretext design dominates accuracy: it adds $+0.077$ macro AUROC over a segmentation-only baseline, roughly $5\times$ the effect of any MIL-side change. The best configuration, the four-pretext encoder paired with a CNV signature-regression head, reaches macro AUROC 0.7038 ± 0.0074. Never trained on annotations, the predicted-subtype attention maps align with pathologist-delineated tumor outlines on held-out slides. The auxiliary signature regression itself, however, is weak: the COSMIC exposure vectors are sparse, high-dimensional, and unevenly covered, ill-suited to the regression objective we used and likely better cast as multilabel classification. Even so, its attention heads still surface a small, stable subset of signatures (CN17, ID14, SBS33, SBS39).

Conclusions. H&E whole-slide images can yield interpretable, spatially-grounded subtype predictions at a level useful for triage. Accuracy hinges on the tile encoder's representation, so encoder pretraining warrants the most effort, but the aggregator should be redesigned, not abandoned: its weak signature branch points to two concrete fixes, predicting only a curated subset of signatures to shrink the decoder and recasting its objective from regression to multilabel classification. By pairing each prediction with visual and molecular evidence, AUGUR offers a low-cost, auditable pre-screen that prioritizes confirmatory molecular testing.

***Index Terms*—**Attention-based multiple instance learning, breast cancer subtyping, computational pathology, interpretable deep learning, mutational signatures, whole-slide imaging.

I. INTRODUCTION

Precision oncology depends on identifying tumor-specific molecular alterations (e.g., point mutations, copy-number alterations, and gene fusions) that can be used for diagnosis, prognosis, treatment selection, and trial matching. These alterations are not merely associative markers but the molecular mechanisms driving tumor growth and progression, making them direct targets for therapy [2]–[4].

In practice, these alterations are diagnosed using molecular pathology assays that range from targeted single-gene tests to broad next-generation sequencing (NGS) panels, whole-exome sequencing, or whole-genome sequencing. Although the cost per sequencing run has decreased over time, routine deployment of genome-scale testing still requires substantial investment in bioinformatics infrastructure (compute, storage, pipelines) and specialized personnel to run and maintain these systems. In their tertiary-hospital implementation study, Kim et al. [5] explicitly highlight both the financial investment required and the bioinformatics specialists and server engineers needed to support NGS in routine care. More broadly, the rapid clinical uptake of NGS has created persistent downstream bottlenecks in data processing, quality control, storage, and,

Manuscript prepared June 2026. *(Corresponding author: Ho Thi Ngoc Phuong.)*

Ho Thi Ngoc Phuong is the MSc candidate; Marcel J.T. Reinders is the thesis supervisor.

The authors are with the Faculty of Electrical Engineering, Mathematics and Computer Science, Delft University of Technology, Delft, Netherlands (e-mail: HoThiNgocPhuong@student.tudelft.nl; M.J.T.Reinders@tudelft.nl).

most importantly, clinical interpretation of complex variant findings [6]. These deployment barriers (cost and specialized expertise) explain the growing interest in lower-cost triage signals that prioritize which patients warrant confirmatory molecular testing, particularly in settings where universal sequencing is not feasible.

One promising candidate signal is tissue morphology. Across several well-studied cancers, driver mutations leave reproducible morphological imprints visible on hematoxylin-and-eosin (H&E)-stained tissue, though the strength, mechanism, and discoverability of these imprints vary by gene, tissue, and cellular context. At one end of the spectrum, the association is mechanistically obligate and visually unmistakable, so a pathologist can often infer the driver from morphology alone. Loss of CDH1 (E-cadherin, a cell-adhesion protein) yields the discohesive, single-file cells of lobular carcinoma [7], [8]; VHL inactivation produces the clear-cell phenotype of renal cell carcinoma [9]–[12]; the PML-RARα fusion gives the hypergranular promyelocytes of acute promyelocytic leukemia [13]–[16]; and MYC activation drives the "starry-sky" pattern of Burkitt lymphoma [17]–[19]. For most other recurrent driver mutations (KRAS, EGFR, PIK3CA, TP53, ATM, etc.), the morphological correlates are more subtle and confounded by the cell-of-origin and the subtype context in which the mutation occurs. These genotype-to-morphology relationships open a complementary triage axis: routine H&E morphology can act as a low-cost triage layer that prioritizes which patients warrant confirmatory molecular testing, provided the imprints, including the subtle and confounded ones, can be extracted reliably and explained in pathologist-interpretable terms.

The last decade has seen a rapid expansion of digital pathology, driven by improvements in whole-slide imaging (WSI) scanners, storage/compute infrastructure, and clinical validation efforts [20]–[22]. These developments have made computational pathology a natural domain for deep neural networks (DNNs), because DNNs can learn predictive features directly from pixel data at multiple spatial scales [23]. Early work established strong proof-of-concept that clinically relevant molecular alterations are, to some extent, predictable from routine H&E morphology using convolutional neural networks (CNNs) trained on tiled WSIs. Coudray et al. [24] train a CNN on The Cancer Genome Atlas (TCGA) WSIs and report that multiple recurrent lung adenocarcinoma gene mutations (including EGFR, KRAS, STK11, and TP53) can be predicted from H&E with meaningful discrimination performance. Kather et al. [25] extend this approach to 14 cancer types in over 5,000 patients, with mean cross-validated AUROCs for the top driver mutations clustering in the 0.6–0.8 range across lung, colorectal, breast, and gastric cancers. Fu et al. [26] show that whole-genome doubling has a universal morphological signature across cancer types.

Translation to routine clinical workflows, however, has been limited, both because external-cohort generalization remains uneven [27] and because established biomarkers are already covered by mature IHC and ISH assays. A separate concern is that these models offer little mechanistic insight into the genotype-morphology relationships they exploit. Subsequent work using multiple instance learning (MIL) and attention-based aggregation [1], [28] improves discrimination performance and produces post-hoc attention maps that intend to localize the morphological evidence behind each prediction, although the fidelity of these maps to underlying biology is contested.

A recurring observation across these efforts is that molecular subtypes are substantially more predictable from H&E than the individual driver mutations they contain. Strong subtype-from-H&E results have now been reproduced for breast intrinsic subtypes [29], [30], endometrial ProMisE classes [31], and bladder muscle-invasive subtypes [32]. Kather et al. [25] make the same point in their own data, finding that integrated features such as MSI, homologous recombination deficiency, and expression-based subtypes were markedly more predictable than the individual mutations comprising them. This asymmetry has a straightforward interpretation: each subtype integrates many co-occurring molecular features that push morphology in correlated directions, producing a stronger and more reproducible histologic signal than any single driver. A methodological consequence is that some apparent single-mutation predictions are partly subtype prediction in disguise: a model nominally calling TP53 or PIK3CA in breast cancer recovers, in part, the basal-like or luminal-A context in which those mutations are enriched.

A second, more recent shift is the rise of pathology foundation models: large self-supervised vision encoders pre-trained on millions of WSIs across institutions and tissues. UNI [33] (100M tiles, 100K slides, 20 tissue types), Prov-GigaPath [34] (1.3B tiles from 171K real-world non-TCGA slides), Virchow [35] (a 632M-parameter ViT trained on 1.5M slides routed through Memorial Sloan Kettering), and the open-weights Phikon [36] have collectively reset the floor for what counts as a strong off-the-shelf representation in computational pathology. The vision-language variant CONCH [37] reports zero-shot subtype classification accuracies above 90% on NSCLC, RCC, and breast cancer. These models substantially mitigate the representation-learning and external-cohort generalization problems that limited the previous CNN/MIL generation. They do not, however, resolve the explainability problem, since hierarchical attention, joint vision-language pre-training, and patch-level pooling all complicate post-hoc attribution to specific morphological structures, and large-scale pre-training does not by itself guarantee that the features driving a prediction correspond to the biology a pathologist would recognize. This gap is not specific to foundation models: any classifier that routes a slide through a learned tile embedding can offer only post-hoc attention as evidence, and attention is not guaranteed to be faithful. We therefore make no claim that our method is intrinsically more interpretable; instead, we treat faithful, pathologist-checkable localization as an explicit design goal and validate it directly rather than assuming attention constitutes an explanation.

The contributions of this work are threefold.

- We introduce AUGUR, a two-stage pipeline pairing a multi-task tile encoder with DualCLAM, an attention-based MIL aggregator with auxiliary mutational-signature heads whose per-class attention serves as candidate visual evidence.

- We show that the tile encoder's pretext design, not the MIL aggregator, is the dominant driver of subtyping accuracy on TCGA-BRCA, where AUGUR reaches a macro AUROC of 0.7038.
- We show that the predicted-subtype attention maps align with pathologist-delineated tumor outlines without ever being supervised on them, and that the signature heads surface a small, stable set of COSMIC signatures.

II. Related work

A. Genotype, subtype, and morphology

The literature supports a hierarchical view of the genotype-morphology relationship, from near-obligate, mechanistically-traceable lesions at the top to point mutations with no recognizable imprint at the bottom.

At the top, a small set of genes produce near-pathognomonic morphology in their tissue context. CDH1 loss disrupts adherens junctions, producing single-cell discohesion characteristic of invasive lobular carcinoma [7], [8]. VHL loss stabilizes HIF, which induces PLIN2 to package lipid droplets [9] and represses CPT1A to block fatty acid oxidation [10]; together with HIF-driven glycogen accumulation [11], [12], this produces the clear cytoplasm of clear cell renal cell carcinoma. PML-RARα arrests myeloid maturation at the promyelocyte stage through enhanced corepressor/HDAC recruitment [13], [14], producing the hypergranular promyelocytes with faggot cells characteristic of acute promyelocytic leukemia [15], [16]. MYC deregulation couples proliferation with apoptosis [17], [18]; combined with cooperating mutations, it produces Burkitt lymphoma [38], where high cell turnover generates the characteristic starry-sky pattern of tingible-body macrophages clearing apoptotic cells [15], [19]. The lesion-morphology association recurs across many tumor types: the DNAJB1-PRKACA fusion is present in essentially all fibrolamellar hepatocellular carcinomas [39], which are defined by their lamellar-fibrosis architecture [40]; germline FH mutation followed by somatic loss of the second allele [41] drives fumarate accumulation [42], [43] and is associated with the inclusion-like nucleolar morphology recognized as the hallmark of HLRCC [44]; the ETV6-NTRK3 fusion produces the microcystic, secretion-rich morphology shared by secretory breast carcinoma [45] and its salivary-gland counterpart, mammary analog secretory carcinoma (MASC) [46]. In each case a single defining lesion, which either disrupts a tissue-specific differentiation program or imposes an obligatory metabolic or architectural phenotype, is associated with morphology that a pathologist can recognize at low magnification.

Beneath this layer, the link weakens from obligate to probabilistic: a subtype enriches for a recognizable morphologic cluster, but the same morphology can arise outside it and a member of that subtype can present without it. In breast cancer, the PAM50 intrinsic subtypes [47] partition tumors along an axis of luminal differentiation: luminal tumors retain ESR1, GATA3, and FOXA1 expression, while basal-like tumors lose this program and present as high-grade carcinomas [48] with medullary-like features [49]. Deep learning can partially recover this signal from H&E at $\approx$77% accuracy for the binary basal-vs-non-basal call [29] and $\approx$66% for the four-class call [30]. In colorectal cancer, the four consensus molecular subtypes, including MSI immune (CMS1), WNT/MYC canonical (CMS2), KRAS metabolic (CMS3), and TGF-β mesenchymal (CMS4) [50], leave H&E imprints (Crohn-like lymphoid infiltrates in CMS1 [51]; desmoplastic stroma in CMS4 [52]) that the imCMS classifier [53] predicts at macro-AUROC of 0.84 on held-out cohorts. In endometrial cancer, the four TCGA molecular classes cluster along axes of mutational load and chromosomal stability: POLE-ultramutated and MMRd hypermutated tumors carry high neoantigen loads and dense TILs [54], while p53-abnormal copy-number-high tumors display marked nuclear atypia and serous-like architecture [31], [55]. Pathologist discordance motivated the molecular reclassification [56], so all four classes can now be predicted from H&E [31] as a component of the HECTOR recurrence prognosticator [57]. The pattern recurs in thyroid cancer, where BRAF V600E drives strong MAPK activation and the classical-type papillary architecture while RAS-family mutations produce weaker MAPK signaling and a follicular-pattern morphology [58], and in small cell lung cancer, where the ASCL1-, NEUROD1-, POU2F3-, and YAP1-defined transcriptional subtypes are defined transcriptionally rather than morphologically [59]. In each case the morphology emerges from the aggregate of co-occurring alterations and the transcriptional program they engage; *no single mutation suffices to produce it*. This is why subtype prediction from H&E achieves consistently higher within-cohort AUROCs than individual-mutation prediction, even when the same model architecture is used for both targets [25], [60].

At the bottom of the hierarchy, most somatic point mutations have no human-recognizable morphologic signature and are predictable from H&E only via cell-of-origin, subtype enrichment, or coincidental association with a more morphologically penetrant alteration. Pan-cancer integrative analyses confirm that cell-of-origin, not individual driver mutations, dominates the molecular taxonomy of human tumors [61], and the same asymmetry is observed when the predictor is H&E: pan-cancer studies that train one model per recurrent driver mutation report AUROCs clustered in the 0.6–0.8 range, with mean single-nucleotide-variant performance significantly lower than subtype prediction using the same architectures [24]–[26], [60]. Where mutation prediction does work, the signal commonly reflects a confounder rather than mutation-specific morphology: TP53 status in breast cancer tracks the basal-like subtype it is enriched in, so apparent "TP53 prediction" is largely subtype prediction [48]; in colorectal cancer, the same pipeline that generalizes well for MSI and BRAF fails for KRAS, NRAS, and PIK3CA on external validation [27]; and gene-level predictors can collapse onto site- or stain-specific shortcuts not visible to a pathologist [62]. The apparent exceptions are precisely the cases where the mutation drives a broader subtype state with its own morphology, such as MSI in colorectal cancer (functionally CMS1) [50], [63], BRAF V600E in thyroid (driving classical papillary architecture) [58], so the "mutation predictor" is operating at the subtype level.

A consequence of this hierarchy is that any single somatic point mutation is unlikely to leave a reproducible H&E

imprint on its own. This motivates predicting morphology from summary representations of the somatic genome, such as subtype assignments and mutational signature exposures (Section II-C), rather than from the presence or absence of individual point mutations. Subtype prediction from H&E is already a mature direction [29]–[31], [53], [57]; signature-exposure prediction is younger but the most clinically actionable summary, homologous-recombination deficiency, can be predicted from H&E with strong, bias-controlled discrimination (AUROC 0.83–0.86 after correcting for technical batch and molecular-subtype confounders) in luminal breast cancer [64].

B. Deep learning on whole slide images

We have traced the broader trajectory of deep learning on H&E images, from the CNN-MIL era to large self-supervised foundation encoders, in Section I. This subsection focuses on the methodological lineage most directly relevant to AUGUR: weakly-supervised aggregation of tile features into slide-level predictions, multi-task and pretext supervision for tile feature extraction, and attention-based explainability.

1) Weakly-supervised aggregation: Whole-slide images are gigapixel objects, routinely $100{,}000 \times 100{,}000$ pixels or more. They cannot be processed by a CNN in a single forward pass, while slide-level diagnoses are abundant in pathology reports and pixel-level annotations are not. The standard response partitions each slide into thousands of tiles, encodes them independently into feature vectors, and aggregates those features into a single slide-level prediction. The formal substrate is multiple-instance learning (MIL), which treats a slide as a bag of instances carrying a single bag-level label and parameterizes the Bernoulli bag-label distribution end-to-end via a tile encoder, a permutation-invariant pooling operator, and a bag classifier [28]. Campanella et al. [65] trained max-pool MIL models on 44,732 H&E WSIs from 15,187 patients on three binary tumor-detection tasks, namely prostate cancer in core biopsies, basal cell carcinoma in skin biopsies, and breast cancer metastasis in axillary lymph nodes, using only report-derived slide-level labels, reaching test AUROCs of 0.97–0.99 across the three tasks and outperforming pixel-annotated fully-supervised baselines on cross-cohort transfer.

The choice of pooling operator is consequential. Fixed max- or mean-pooling collapses the bag to a single score without indicating which tiles contributed, sacrificing both flexibility and interpretability. Attention-based MIL replaces fixed pooling with a learned attention distribution over tiles, achieving both higher accuracy in the small-sample regime and per-tile saliency maps that align with ground-truth annotations under bag-only supervision [28]. CLAM [1] extends this with an instance-clustering auxiliary loss that pushes high-attention tiles to be class-discriminative and low-attention tiles to be uninformative; its CLAM-MB variant additionally allocates one attention head per output class for class-conditional saliency, and CLAM reaches 10-fold macro-AUROCs of 0.991, 0.956, and 0.953 on renal cell carcinoma subtyping, NSCLC subtyping, and lymph-node metastasis detection, while retaining AUROC $>$0.9 with as little as 25–50% of the training data. TransMIL [66] replaces the i.i.d.-tile assumption with correlated MIL realized through self-attention with multi-scale positional encoding, achieving AUROCs of 0.93, 0.96, and 0.99 on CAMELYON16, TCGA-NSCLC, and TCGA-RCC, roughly 5 AUROC points above CLAM on CAMELYON16, with the trade-off that the tile-to-tile attention matrix is harder to read as per-tile saliency than the scalar attention weights of gated MIL.

2) Multi-task and pretext supervision: Early supervised pathology pipelines trained one model per task (tile classification, nuclear segmentation, mutation prediction, etc.), each requiring dense human annotation that scaled poorly. Pan-cancer mutation predictors require matched H&E and sequencing data for every recurrent driver and reach only modest AUROCs (0.73–0.86 across NSCLC drivers; only 6 of 10 tested genes reached significance) [24]; nuclear-instance models require exhaustive per-instance labels in the thousands.

Multi-task supervision reduces annotation cost across tasks that share a feature substrate. Within a single image, HoVer-Net [67] allocates three decoder branches to nuclear pixel segmentation, horizontal-vertical distance regression, and per-nucleus classification, producing simultaneous state-of-the-art performance on each (CoNSeP PQ 0.547, classification F1 0.42–0.64 across nuclear types) on a dataset of 24,319 manually annotated nuclei. Its Cerberus [68] extension scales the same principle to nuclei, glands, lumina, and tissue regions, exceeding HoVer-Net on every shared task (internal PQ 0.612 vs 0.583, external mPQ 0.332 vs 0.285) and showing that the shared multi-task encoder transfers to downstream tasks not seen during training. At the cross-dataset level, Mormont et al. [69] pretrain a single backbone jointly on 22 histopathology classification tasks ($\sim$900k images), beating ImageNet on 15 of 20 downstream evaluations under frozen feature extraction, although the advantage disappears under full fine-tuning, which lets ImageNet's generic features catch up.

Self-supervised learning takes the annotation burden off entirely by deriving supervision from the input itself. In Self-Path framework [70], domain-specific pretexts tailored to pathology (magnification prediction, magnification-jigsaw task (JigMag), and hematoxylin-channel reconstruction) outperform supervised baselines by 6–13 AUC points with as little as 1% of WSI labels. Generic contrastive learning (SimCLR) [71] applied to a 57-dataset histopathology corpus boosts downstream linear-probe macro-F1 by $\sim$28 points over ImageNet pretraining, with diminishing returns past $\sim$50k unlabeled images. Transformer-based SSL with semantically-mined positives (CTransPath/SRCL) [72] pretrains a hybrid CNN-Swin backbone on $\sim$15.6M TCGA and PAIP patches and, paired with CLAM, reaches WSI-level AUROCs of 0.922 on CAMELYON16, 0.912 on TCGA-NSCLC, and 0.967 on TCGA-RCC. Masked image modeling at TCGA scale (Phikon, iBOT ViT-B on the PanCancer40M corpus of 43M tiles) [36] extends this further, beating CTransPath by +1.4% and HIPT by +4.0% on average across 14 weakly-supervised WSI tasks and reaching 98.2 AUROC on PAIP-CRC MSI external validation.

3) Explainability: Early CNN-based pathology pipelines explained predictions post-hoc with generic computer-vision tools, such as tile-level probability heatmaps tiled back onto the WSI [24] or class activation maps [73] generated via gradient-based localization [74] that highlight regions the CNN attended

to. Both are coarse, applied after training without architectural support, and offer no class-conditioning or excitatory/inhibitory polarity.

Attention-based MIL repositions saliency from a post-hoc artifact to an intrinsic property: the same bag-normalized attention weights a_k that drive aggregation also serve as a per-tile saliency map [28], and CLAM-MB's per-class branches yield class-conditional saliency under bag-only supervision [1]. However, the fidelity of this saliency to tumor biology is contested. Deep learning detects TCGA submitting-site signatures at AUROCs of 0.96–0.99 across six cancer cohorts, biases the prediction of mutations, stage, and even ethnicity, and 91% of features predictable under standard cross-validation lose AUROC under preserved-site cross-validation [62].

Recent work pushes toward intrinsic, structured interpretability. Additive MIL [75] swaps the function composition at the predictor head so each instance contributes additively to the bag prediction, yielding exact per-patch, per-class contributions whose signs encode excitation/inhibition and are proportional to true Shapley values; on CAMELYON16 the additive variant of Attention-MIL reaches AUC 0.846 (vs 0.750 for ABMIL) while improving pathologist-segmentation AUPRC from 0.36 to 0.42. Prototype-based MIL [76], [77] replaces opaque attention scores with case-based reasoning against learned visual prototypes, matching or exceeding CLAM and TransMIL on standard benchmarks while supplying human-inspectable class descriptors and supporting multi-label tumor heterogeneity (PAMIL achieves AUC 0.8985 on TCGA-NSCLC, 0.9758 on TCGA-RCC, and 0.8475 on TCGA-STAD multi-label).

C. Genome-wide mutational signatures

Mutational signatures decompose a tumor's somatic mutation catalog into latent components, each carrying a characteristic context-conditional substitution profile and a per-tumor activity (*exposures*). The framework was introduced via non-negative matrix factorization of the 96-class trinucleotide-context single-base-substitution (SBS) spectrum across pan-cancer cohorts [78], and the COSMIC v3 reference catalog derived from 84.7M somatic mutations in 23,829 PCAWG / TCGA samples now lists 49 SBS, 11 DBS, and 17 indel (ID) signatures, each ascribed where possible to a known endogenous or exogenous mutational process, such as 5-methylcytosine deamination, ultraviolet light, tobacco carcinogens, and defective homologous-recombination repair [79], [80]. Two complementary layers extend the framework beyond point mutations: copy-number signatures decompose CN-segment summary statistics into 21 pan-cancer CN processes across 9,873 cancers from 33 types [81], while rearrangement (SV) signatures decompose a 32-class structural-variant taxonomy and identify 6 rearrangement processes in 560 breast-cancer whole genomes, several of which mark BRCA1/BRCA2-deficient tumors [82]. The output of a signature analysis for a single tumor is a vector of exposure values (one per signature) whose support typically includes a small number of co-active processes. This exposure vector summarizes the tumor's mutational history in a low-dimensional space.

A subset of signatures has well-established tissue-level morphologic correlates, providing biological grounding for the idea that signature exposures might be partially recoverable from H&E. MMR-deficiency signatures (SBS6/15/20/26 + SBS44) mark MSI-H tumors that, in colorectal cancer, present with dense lymphocytic infiltrates, Crohn-like peritumoral reaction, and mucinous differentiation [51]; the same dense-TIL morphology accompanies POLE-driven (SBS10a/b) and MMRd hypermutated endometrial cancers via shared high neoantigen loads [54]. HRD signatures, principally SBS3 and ID6, alongside BRCA-linked tandem-duplication and deletion rearrangement processes [82], mark BRCA1/BRCA2-deficient breast and ovarian cancers, with laminated fibrosis and clear-cytoplasm tumor cell foci as morphologic correlates [64]. APOBEC signatures (SBS2 + SBS13) are markedly enriched in HER2-amplified breast cancers [83], supplying an indirect morphologic correlate via HER2 subtype. UV-light signatures (SBS7a–d) dominate melanoma and cutaneous squamous-cell carcinoma, and tobacco signatures (SBS4) dominate lung squamous-cell and small-cell carcinomas. In both cases the morphology is tissue-of-origin rather than signature-specific, but the correlation between a dominant exogenous mutagen and a recognizable cancer type makes the link near-deterministic. In each case the morphology emerges from the subtype the signature defines (HRD, MSI), from cell-of-origin enrichment (UV-skin, smoking-lung), or from a co-occurring driver alteration (APOBEC-HER2), rather than from any direct visual manifestation of the mutational process itself.

Deep learning has been applied to histology-to-genomics prediction along two converging tracks. The first targets individual driver mutations: Coudray et al. [24] predict 10 NSCLC drivers from H&E (per-slide AUROCs 0.73–0.86 for the 6 of 10 reaching significance), and pan-cancer extensions across 14–32 cancer types report mutation AUROCs clustered in 0.6–0.8, with subtypes consistently more predictable than the mutations they enrich for [25], [26]. The second track targets aggregate genomic-instability summaries: microsatellite instability has been pushed to clinical-grade discrimination on routine H&E (AUROC 0.95–0.96 on external resection cohorts) [63], and HRD has been predicted from H&E with bias-controlled AUROCs of 0.83–0.86 in luminal breast cancer [64]. The literature is, however, almost exclusively focused on a single binarized signature target at a time, most commonly HRD, treated as the deployment goal. Predicting the full mutational signature exposure vector jointly with intrinsic subtype, with attention disentangled per process so that each process's morphologic correlates can be inspected separately, is underexplored. Our proposed method is positioned at exactly this gap: mutational signature exposure prediction as a process-resolved pretext on top of the clinically established subtype-classification task.

III. Methodology

AUGUR classifies molecular subtypes of breast tumors from H&E whole-slide images (WSIs). Because WSIs reach gigapixel resolution and only carry a single slide-level label, we decompose the task into a tile encoder and a multiple-instance-learning aggregator.

The pipeline (Figure 1) proceeds in three stages: WSI preprocessing into tissue tiles (Section III-A); a shared tile

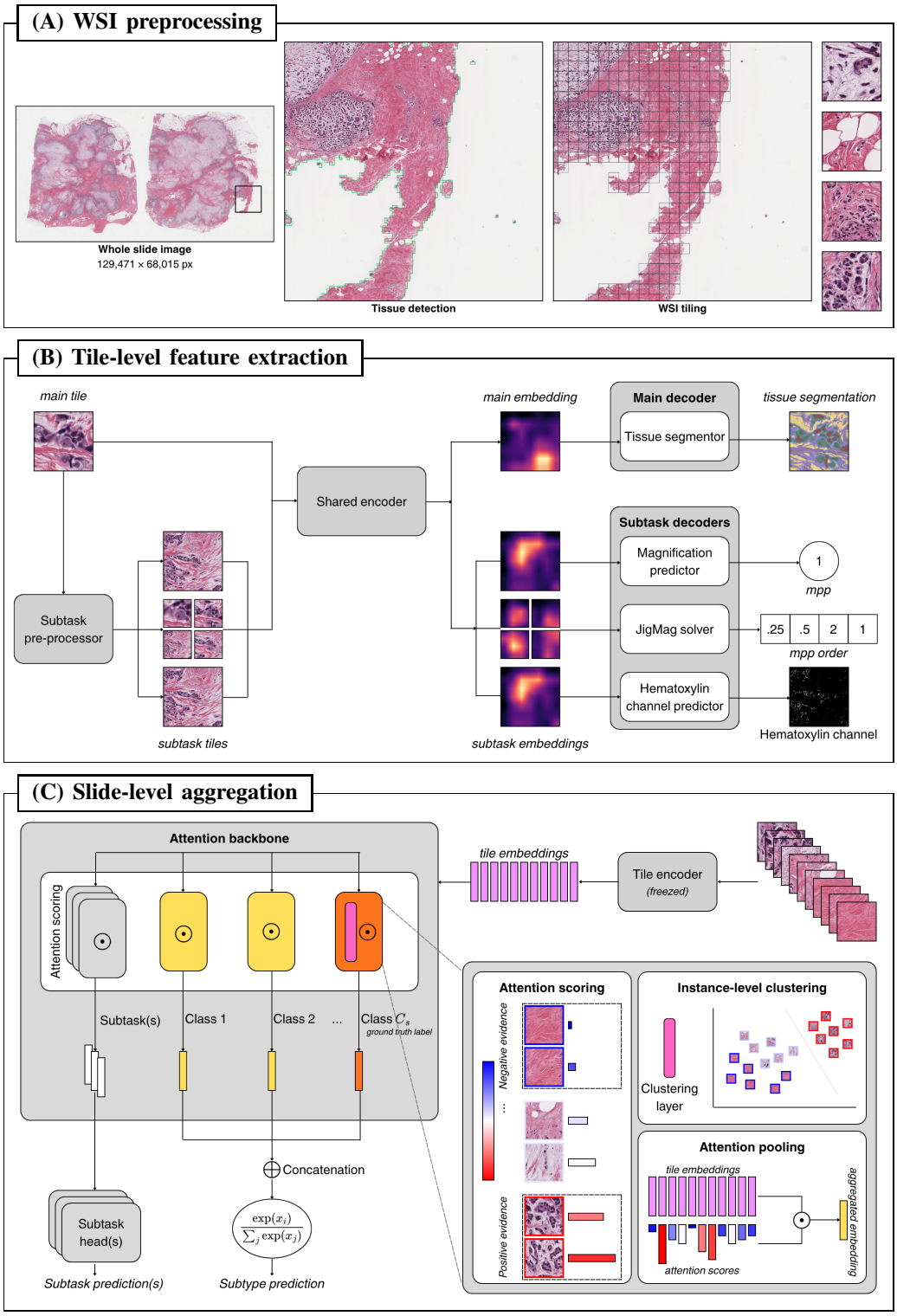


Figure 1: **Overview of the AUGUR pipeline.** ***(A) WSI preprocessing:*** a thumbnail tissue mask drives the enumeration of valid tile centers, each read at a target MPP and resized to a fixed encoder input resolution. ***(B) Tile-level feature extraction:*** a shared encoder is trained on four pretext tasks (tissue segmentation, magnification prediction, magnification jigsaw, and hematoxylin channel prediction), each routed through its own subtask preprocessor and decoder head. ***(C) Slide-level aggregation:*** the frozen tile encoder produces a precomputed bag of embeddings per slide, which the DualCLAM aggregator pools via attention to predict the slide subtype, one or more mutational-signature exposure vectors, and a per-class attention map.

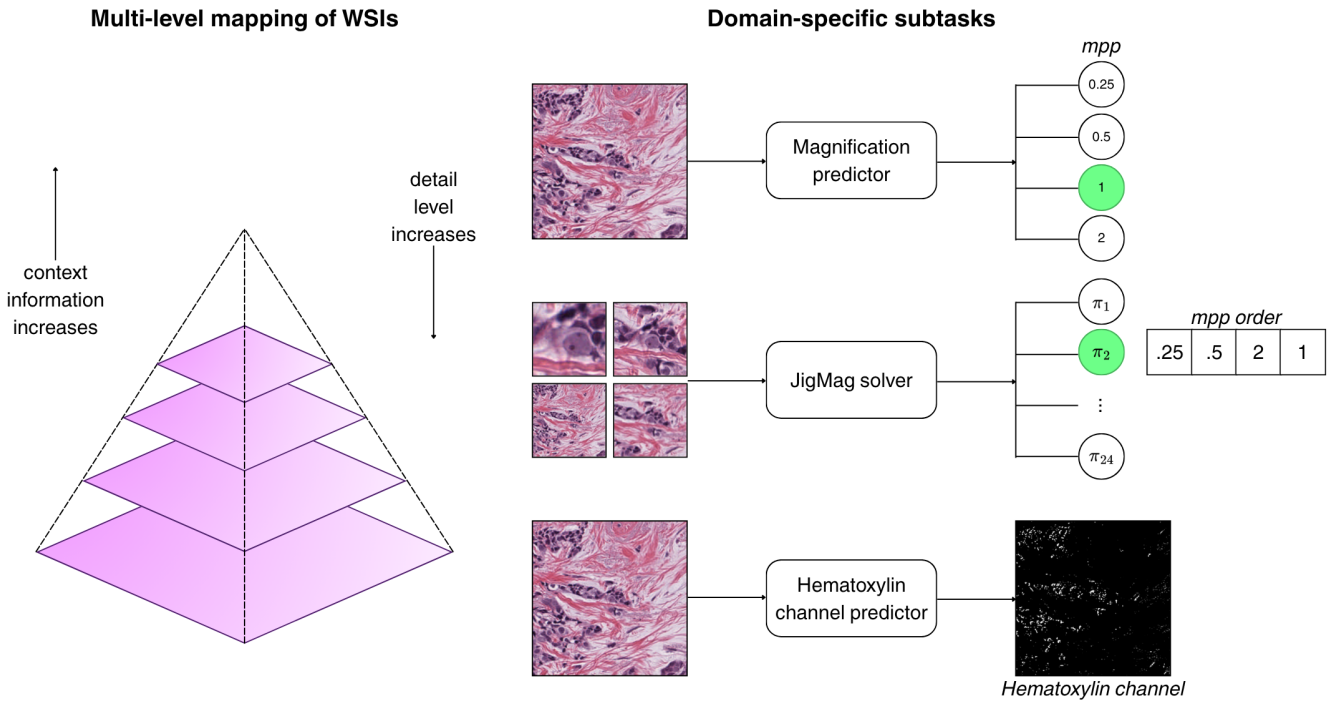


Figure 2: **The three tile-level self-supervised pretext subtasks used to train the shared encoder.** ***Magnification prediction:*** the tile is re-read at a randomly chosen MPP from a fixed candidate set, and the encoder must identify which MPP was used. ***Magnification jigsaw:*** four sub-tiles read at different MPP are arranged on a 2×2 grid according to a random permutation, and the encoder must identify the permutation. ***Hematoxylin channel prediction:*** the base tile is fed to a dense-prediction decoder that reconstructs the hematoxylin channel obtained from the input via Ruifrok-Johnston color deconvolution.

encoder trained with auxiliary supervised and self-supervised pretext tasks that shape its representations around interpretable histological concepts (Section III-B); and an attention-based slide-level aggregator that produces subtype prediction, one or more mutational-signature contributions, and a spatial attention map identifying the tiles driving each class decision (Section III-C).

A. WSI preprocessing

For each slide, OpenSlide [84] produces a low-resolution thumbnail from which we compute an approximate tissue mask: a pixel is flagged as tissue when its mean RGB intensity is below the brightness threshold θ_w and the spread between its largest and smallest color components is above the chroma threshold θ_g. The brightness condition filters white scanner background; the chroma condition rejects near-grayscale artifacts such as pen marks and scanner stripes.

We then slide a $W \times W$ window over the mask at stride S and retain windows whose mean tissue fraction exceeds the tissue threshold θ_t. Each retained window's center is projected back to level-0 slide coordinates, yielding a deterministic list of valid tile centers per slide. When a downstream task reads the same center at multiple resolutions (Section III-B), W is interpreted at the MPP with the widest physical extent so that finer-resolution reads at the same center also fall on tissue.

Tiles are parameterized by microns-per-pixel (MPP) rather than by an OpenSlide pyramid level. For a target MPP μ and side length H, we select the pyramid level closest to μ, read the corresponding square region via OpenSlide, and resize to $H \times H$ pixels. The output is a tile of resolution μ regardless of the level's native MPP, which decouples the pipeline from each slide's scanner-specific magnification metadata and lets the same center be re-read at several MPPs, a property that the multi-scale pretext tasks (Section III-B) rely on.

B. Tile-level feature extraction

We train a shared encoder on a set of tile-level pretext tasks $\mathcal{P} = \{t, m, j, h\}$: tissue segmentation ($t$), magnification prediction (m), magnification jigsaw (j), and hematoxylin channel prediction (h).

Let $\mathbf{X} \in \mathbb{R}^{3\times H\times H}$ denote a base tile produced by the preprocessing stage (Section III-A). For each task $p \in \mathcal{P}$, a task-specific preprocessor produces an input view $\mathbf{X}_p = \mathrm{Sub}_p(\mathbf{X}) \in \mathbb{R}^{3\times H\times H}$ from the base tile.

All views are encoded by the same encoder $\mathrm{Enc}(\cdot)$, yielding per-task embeddings:

$$\mathbf{z}_p = \mathrm{Enc}(\mathbf{X}_p) \in \mathbb{R}^D. \tag{1}$$

A task-specific decoder $\mathrm{Dec}_p(\cdot)$ maps each $\mathbf{z}_p$ to the target space of p:

$$\hat{\mathbf{Y}}_p = \mathrm{Dec}_p(\mathbf{z}_p).$$

After training, only the encoder is retained: the base-view embedding $\mathbf{z} = \mathrm{Enc}(\mathbf{X})$ is what the slide-level aggregator consumes (Section III-C).

Each task $p \in \mathcal{P}$ is paired with a positive scalar weight w_p and a loss $\mathcal{L}_p(\hat{\mathbf{Y}}_p, \mathbf{Y}_p)$ that compares its decoder output $\hat{\mathbf{Y}}_p$ to a task-specific target $\mathbf{Y}_p$. The shared encoder and all per-task decoders are trained jointly by minimizing the weighted sum

$$\mathcal{L} = \sum_{p \in \mathcal{P}} w_p \, \mathcal{L}_p, \tag{2}$$

with weights normalized to $\sum_{p \in \mathcal{P}} w_p = 1$ so each w_p can be read as the fraction of the total loss contributed by task p.

1) Tissue segmentation: The tissue segmentation pretext asks the encoder to support dense, per-pixel prediction of tissue type. It is the only supervised task in $\mathcal{P}$, in which the target is an external annotation rather than a transformation of the tile itself.

The subtask preprocessor is the identity, $\mathbf{X}_t = \mathbf{X}$. The target $\mathbf{Y}_t \in \{0, 1, \ldots, C_t\}^{H \times H}$ is a per-pixel class-index map assembled by walking every annotated region-of-interest (ROI) that overlaps the tile's footprint and pasting its class labels into the tile-aligned target. In total the map spans $C_t + 1$ classes: the C_t biologically meaningful tissue types (indices $1, \ldots, C_t$) plus one unknown class (index 0) for pixels outside any annotated region. Class 0 is reserved for "Outside ROI" and covers pixels not annotated by any overlapping ROI, including the entire map for tiles drawn from slides without annotations.

The decoder Dec_t produces per-pixel logits $\hat{\mathbf{Y}}_t \in \mathbb{R}^{(C_t+1) \times H \times H}$, one channel per class including the unknown. We train with pixel-wise cross-entropy that ignores class 0:

$$\mathcal{L}_t = \frac{1}{|\Omega_t|} \sum_{(i,j) \in \Omega_t} \mathrm{CE}\big(\hat{\mathbf{Y}}_{t,ij},\, \mathbf{Y}_{t,ij}\big), \tag{3}$$

where $\Omega_t = \{(i,j) : \mathbf{Y}_{t,ij} \neq 0\}$ and $\mathrm{CE}(\cdot,\cdot)$ is the standard categorical cross-entropy. Tiles drawn from slides without annotations therefore contribute zero tissue-typing gradient (their Ω_t is empty), but still drive the shared encoder through the other three pretext tasks.

2) Magnification prediction: The magnification prediction pretext is self-supervised, in which the encoder is asked to identify the resolution at which a tile was read, forcing it to encode scale-dependent histological features such as nuclear size relative to the field of view and the granularity of tissue texture.

The subtask preprocessor Sub_m samples a target MPP μ_k uniformly at random from a fixed candidate set $\mathcal{M} = \{\mu_1, \mu_2, \ldots, \mu_{C_m}\}$ and re-reads the tile at resolution μ_k from the same slide location as $\mathbf{X}$, using the MPP-anchored reading procedure of Section III-A. The resulting input view is $\mathbf{X}_m \in \mathbb{R}^{3 \times H \times H}$, and the target $y_m = k \in \{1, \ldots, C_m\}$ is the index of the chosen MPP.

The decoder Dec_m produces C_m logits $\hat{\mathbf{y}}_m \in \mathbb{R}^{C_m}$, and the loss is the standard categorical cross-entropy:

$$\mathcal{L}_m = \mathrm{CE}\big(\hat{\mathbf{y}}_m, y_m\big). \tag{4}$$

3) Magnification jigsaw: The magnification jigsaw pretext, also self-supervised, combines multi-scale awareness with spatial reasoning. The encoder is asked to identify how four sub-tiles, each read at a potentially different resolution, were arranged on a 2×2 grid.

The subtask preprocessor reads four sub-tiles of side length $H/2$, all centered on the same slide location as $\mathbf{X}$, at four configured MPPs $\mu_1, \mu_2, \mu_3, \mu_4$. It then samples a permutation π uniformly at random from the set Π of all $C_j = 4! = 24$ permutations of $(1, 2, 3, 4)$, and tiles the four sub-tiles into a 2×2 grid according to π. The resulting input view is $\mathbf{X}_j \in \mathbb{R}^{3 \times H \times H}$, and the target $y_j = k \in \{1, \ldots, C_j\}$ is the index of π in Π.

The decoder Dec_j produces C_j logits $\hat{\mathbf{y}}_j \in \mathbb{R}^{C_j}$, and the loss is the standard categorical cross-entropy:

$$\mathcal{L}_j = \mathrm{CE}\big(\hat{\mathbf{y}}_j, y_j\big). \tag{5}$$

4) Hematoxylin channel prediction: The hematoxylin channel prediction pretext is self-supervised: the encoder is asked to reconstruct the hematoxylin component of the H&E stain from the raw RGB tile. Because hematoxylin preferentially stains cell nuclei, succeeding at this task requires the encoder to localize nuclear structures and distinguish them from other tissue components.

The subtask preprocessor is the identity, $\mathbf{X}_h = \mathbf{X}$. The target $\mathbf{Y}_h \in [0, 1]^{H \times H}$ is the per-pixel hematoxylin intensity, extracted from the base tile via the Ruifrok-Johnston color deconvolution method [85]. Let $\mathbf{X}_{ij} \in (0, 1]^3$ denote the normalized RGB value at pixel (i, j) of $\mathbf{X}$. The per-pixel concentrations of the three reference stains, hematoxylin (c^H_{ij}), eosin (c^E_{ij}), and DAB (c^D_{ij}), are recovered by transforming RGB to optical density via the Beer-Lambert relation and applying the inverse of the reference stain matrix $\mathbf{M}$:

$$\big(c^H_{ij},\, c^E_{ij},\, c^D_{ij}\big)^\top = -\,\mathbf{M}^{-1} \log \mathbf{X}_{ij}. \tag{6}$$

The hematoxylin channel is retained, clipped at zero, and rescaled into $[0, 1]$ using the 1st and 99th percentiles (p_1, p_{99}) of its per-tile distribution, providing stable supervision across tiles with different overall staining intensity:

$$\mathbf{Y}_{h,ij} = \mathrm{clip}\left(\frac{\max(c^H_{ij}, 0) - p_1}{p_{99} - p_1},\, 0,\, 1 \right). \tag{7}$$

The decoder Dec_h produces a per-pixel intensity prediction $\hat{\mathbf{Y}}_h \in \mathbb{R}^{H \times H}$. We train with pixel-wise mean squared error:

$$\mathcal{L}_h = \frac{1}{H^2} \big\| \hat{\mathbf{Y}}_h - \mathbf{Y}_h \big\|^2_F. \tag{8}$$

C. Slide-level aggregation

The slide-level aggregator predicts a tumor subtype for each whole-slide image and, optionally, one or more auxiliary mutational-signature exposure vectors. Because slide-level labels are weak, only one label per slide of millions of pixels, we adopt a MIL framing: each slide is represented as a bag of tiles, the bag carries the label, and the aggregator must learn to combine bag instances into a slide-level prediction while also exposing which instances most strongly support each candidate class.

Let $\mathcal{B} = \big\{\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, \ldots, \mathbf{X}^{(K_\mathcal{B})}\big\}$ denote one bag, where each instance $\mathbf{X}^{(i)} \in \mathbb{R}^{3 \times H \times H}$ is a tile produced by the preprocessing stage (Section III-A). Parenthesized superscripts index instances within a bag, and the bag size $K_\mathcal{B}$ varies across slides because each slide contributes a different number of tissue tiles.

Each bag carries a subtype class index $y_s \in \{1, \ldots, C_s\}$ (the main supervised task) and, for each subtask $e \in \mathcal{E}$, a mutational-signature exposure vector $\mathbf{y}_e \in [0,1]^{C_e}$ with $\sum_i y_{e,i} = 1$, where $\mathcal{E}$ is the set of configured signature subtasks and C_e is the size of subtask e's reference catalog, i.e., the number of mutational signatures it contains.

1) Feature precomputation: End-to-end training of the aggregator and the tile encoder Enc is prohibitively expensive, since even after the per-bag sub-sampling introduced below, every forward–backward pass would have to push gradients through hundreds to thousands of tiles per bag, across every bag in every epoch. We therefore hold Enc frozen for the entire slide-level stage, reusing the weights trained in Section III-B. With Enc fixed, each tile's embedding $\mathbf{z}^{(i)} = \mathrm{Enc}(\mathbf{X}^{(i)})$ never changes; re-encoding the same tiles on every forward pass would recompute an output already known. We instead precompute the encoder output for every valid tissue tile of every slide once, before aggregator training begins, and cache the result on disk.

For each slide we enumerate all valid tissue tile centers using the procedure of Section III-A, read the corresponding tiles, and pass them through the frozen encoder. The slide's bag of embeddings is

$$\hat{\mathcal{B}} = \big\{\mathbf{z}^{(1)}, \mathbf{z}^{(2)}, \ldots, \mathbf{z}^{(K_{\mathcal{B}})}\big\}, \quad \mathbf{z}^{(i)} \in \mathbb{R}^D, \tag{9}$$

where $K_{\mathcal{B}}$ is the total number of tissue tiles enumerated for that slide. The aggregator consumes $\hat{\mathcal{B}}$ directly, replacing the encoder forward pass with a disk load.

A full $\hat{\mathcal{B}}$ may contain thousands of embeddings for large slides, which is too many to feed to the aggregator in one forward pass. Each training-time bag is therefore randomly sub-sampled at rate $\rho \in (0,1]$, retaining $\max\big(1, \lfloor \rho\, K_{\mathcal{B}} \rfloor\big)$ instances. The sub-sample is redrawn at every epoch, so the aggregator sees a different cross-section of each slide on each pass.

2) DualCLAM: Subtyping with auxiliary signature regressions: The slide-level aggregator builds on CLAM [1], a clustering-constrained attention multiple-instance learning model. CLAM pools bag instances via the attention mechanism of Ilse et al. [28], available in a plain or gated form, and supports two backbone layouts: a single-branch (SB) variant with one attention distribution shared across tasks, and a multi-branch (MB) variant that allocates one attention distribution per subtype class and thereby exposes per-class spatial evidence at inference time. An auxiliary instance clustering loss further encourages the attention map to localize tiles diagnostic of each class.

We propose DualCLAM, which extends CLAM with a set of slide-level regression heads, one per configured subtask $e \in \mathcal{E}$, that predict the corresponding mutational-signature exposure vectors $\{\mathbf{y}_e\}_{e \in \mathcal{E}}$ from the same bag of embeddings. Each subtask shapes the aggregator alongside the main subtype classification.

Given a bag of precomputed embeddings $\hat{\mathcal{B}}$ from Section III-C1, DualCLAM proceeds in three stages. First, each per-tile embedding $\mathbf{z}^{(i)} \in \mathbb{R}^D$ is mapped to a lower-dimensional representation $\mathbf{h}^{(i)} \in \mathbb{R}^L$ via a small MLP shared across all tiles. Second, the projected bag is fed to a multi-head attention net [28], in either its plain or gated form, with M heads. Each head $m \in \{1, \ldots, M\}$ computes a softmax-normalized attention distribution $\boldsymbol{\alpha}_m \in [0,1]^{K_{\mathcal{B}}}$ over the bag and returns a pooled embedding

$$\mathbf{z}_m = \sum_{i=1}^{K_{\mathcal{B}}} \boldsymbol{\alpha}_{m,i}\, \mathbf{h}^{(i)} \in \mathbb{R}^L, \tag{10}$$

where $M = 1$ for SB and $M = C_s + |\mathcal{E}|$ for MB, which assigns one attention branch to each of the C_s subtype classes—the per-class branches that yield class-specific spatial maps—plus one branch to each of the $|\mathcal{E}|$ configured subtasks. Third, the pooled embeddings $\{\mathbf{z}_m\}_{m=1}^{M}$ are routed to task-specific heads via a branch layout that allocates a contiguous slice of branches to each task.

a) Subtype classification: The subtype head Φ_s is a single linear layer that takes the pooled embeddings of the subtype branches (one shared branch in SB, or C_s per-class branches concatenated in MB) and produces C_s logits $\hat{\mathbf{y}}_s \in \mathbb{R}^{C_s}$. The loss is the standard categorical cross-entropy:

$$\mathcal{L}_s = \mathrm{CE}\big(\hat{\mathbf{y}}_s, y_s\big). \tag{11}$$

Slides without an annotated subtype are dropped from the slide-level cohort entirely, so y_s always indexes a real PAM50 class.

b) Mutational-signature exposure regression: Mutational signatures decompose a tumor's somatic mutation profile into a vector of exposures against a fixed reference catalog, each entry quantifying the contribution of one mutagenic process to the observed mutations (Section II-C). For each subtask $e \in \mathcal{E}$ and each tumor, the open-source SigProfiler suite [86]–[88] produces a non-negative exposure vector against subtask e's reference catalog of size C_e, which is normalized to sum to one so that $\mathbf{y}_e \in [0,1]^{C_e}$ with $\sum_i y_{e,i} = 1$ defines a discrete probability distribution over the reference signatures. The pipeline supports four signature classes, one per COSMIC mutation-type catalog: single-base substitutions (SBS), doublet-base substitutions (DBS), small insertions and deletions (ID), and copy-number alterations (CN); each contributes one element to $\mathcal{E}$ when enabled, with its own catalog size C_e, and which subset is active in any given run is a configuration choice. We use these vectors as targets of slide-level subtasks, on the hypothesis that some mutagenic processes leave morphological traces visible at H&E magnification and therefore force the aggregator to attend to tiles whose morphology carries biologically informative signal beyond what a subtype label alone provides.

The exposure vectors are computed offline, independently of the H&E pipeline. For each subtask $e \in \mathcal{E}$, the subtask head Φ_e is a single linear layer followed by a softmax that takes the subtask's allocated branch pool (the shared $\mathbf{z}_1$ in SB, or a dedicated per-subtask branch in MB) and produces a predicted exposure distribution $\hat{\mathbf{y}}_e \in [0,1]^{C_e}$ with $\sum_i \hat{y}_{e,i} = 1$. Because both $\mathbf{y}_e$ and $\hat{\mathbf{y}}_e$ are probability distributions over the reference catalog rather than arbitrary vectors in $\mathbb{R}^{C_e}$, we adopt the Kullback-Leibler divergence of the prediction from the target as the loss:

$$\mathcal{L}_e = \mathrm{KL}\big(\mathbf{y}_e \,\big\|\, \hat{\mathbf{y}}_e\big). \tag{12}$$

This matches the simplex geometry of the target, penalizes underprediction of a small but present signature by its log-

ratio rather than its small absolute deviation, and places $\mathcal{L}_e$ in the same nat-scale regime as the subtype cross-entropy $\mathcal{L}_s$, which keeps the task weights of (15) interpretable as loss-share fractions.

c) Instance clustering: The auxiliary instance-clustering loss of CLAM [1] provides a self-distilled supervisory signal that pushes each subtype attention branch to localize on tiles that look like its class. For each subtype class $c \in \{1, \dots, C_s\}$, we attach a binary instance classifier $f_c : \mathbb{R}^L \to \mathbb{R}^2$.

For a slide with true label c, we read class c's attention distribution $\boldsymbol{\alpha}_c$ from the multi-head attention (branch c in MB, the shared branch in SB) and identify the K most-attended tile indices $\mathcal{T}_c$ and the K least-attended tile indices $\mathcal{N}_c$ within the bag, where K is a small fixed positive integer. The in-class clustering loss treats the top-K tiles as positive (label 1) and the bottom-K as negative (label 0):

$$\mathcal{L}^{\text{in}}_{\text{inst},c} = \frac{1}{2K} \sum_{i \in \mathcal{T}_c \cup \mathcal{N}_c} \text{CE}\big(f_c(\mathbf{h}^{(i)}), \mathbf{1}[i \in \mathcal{T}_c]\big). \tag{13}$$

An optional out-of-class term passes the same top-K tiles through every other class's instance classifier with an all-negative target, formalizing the assumption that tiles diagnostic of class c should look unlike every other class:

$$\mathcal{L}^{\text{out}}_{\text{inst},c} = \frac{1}{K} \sum_{c' \neq c} \sum_{i \in \mathcal{T}_c} \text{CE}\big(f_{c'}(\mathbf{h}^{(i)}), 0\big). \tag{14}$$

The combined per-slide instance loss is $\mathcal{L}^{\text{in}}_{\text{inst},c} + \mathcal{L}^{\text{out}}_{\text{inst},c}$ (with the out-of-class term omitted when not used), averaged across slides in the batch.

3) Joint training objective: The aggregator is trained end-to-end (with the encoder held frozen, Section III-C1) by minimizing a weighted sum of the subtype classification loss $\mathcal{L}_s$ ((11)), the mutational-signature exposure regression loss $\mathcal{L}_e$ ((12)) for each subtask $e \in \mathcal{E}$, and the instance clustering loss $\mathcal{L}_{\text{inst}}$ ((13) and (14)):

$$\mathcal{L} = w_s \, \mathcal{L}_s + \sum_{e \in \mathcal{E}} w_e \, \mathcal{L}_e + w_{\text{inst}} \, \mathcal{L}_{\text{inst}}. \tag{15}$$

The task weights are normalized to $w_s + \sum_{e \in \mathcal{E}} w_e = 1$ so each can be read as the fraction of the total task loss contributed by that task. The instance-clustering weight $w_{\text{inst}} \geq 0$ is a separate non-negative coefficient that controls the contribution of CLAM's auxiliary supervision, with $w_{\text{inst}} = 0$ disabling it entirely.

4) Visual explainability via per-class attention: In the MB variant, the multi-head attention assigns a distinct distribution $\boldsymbol{\alpha}_c \in [0,1]^{K_{\mathcal{B}}}$ over the bag's $K_{\mathcal{B}}$ tiles to each subtype class $c \in \{1, \dots, C_s\}$, with $\boldsymbol{\alpha}_{c,i}$ measuring the relative weight tile i receives in class c's pooled representation (the subtask branches each carry their own attention distributions as well, but we focus on the subtype branches for class-level explainability). Because the slide's tile centers $\{(x_i, y_i)\}_{i=1}^{K_{\mathcal{B}}}$ are recorded during feature precomputation (Section III-C1), the attention values are placed back at their source coordinates to produce a per-class spatial attention map over the slide. The SB variant produces a single shared map and is therefore less informative for class-level interpretation.

At inference, the model emits both the subtype prediction $\hat{\mathbf{y}}_s$ and the full set of attention distributions $\{\boldsymbol{\alpha}_c\}_{c=1}^{C_s}$. The map for the predicted class $\hat{c} = \arg\max \hat{\mathbf{y}}_s$ identifies the tiles that most strongly support that prediction, while the maps for the other classes show the spatial evidence weighed against each alternative. The instance-clustering loss ((13), (14)) explicitly trains each branch's attention to localize on tiles diagnostic of its corresponding class, making the resulting maps an interpretable spatial complement to the subtype prediction rather than an artifact of the model's internal weighting.

IV. Experiments

A. Datasets

1) TCGA-BRCA: The Cancer Genome Atlas breast invasive carcinoma (TCGA-BRCA) [48] cohort is the primary data source for AUGUR, supplying 1,093 breast-cancer cases with paired diagnostic H&E whole-slide images (WSIs), somatic mutation calls in mutation annotation format (MAFs), allele-specific copy-number variation (CNV) segment calls, and PAM50 intrinsic-subtype labels. All modalities are downloaded from the GDC portal with the GDC Data Transfer Tool against a single submitter-level manifest, so the four data sources align on the same set of submitter identifiers. At the slide level, the aggregator is evaluated by 5-fold patient-grouped, subtype-stratified cross-validation: each fold serves once as the held-out test set, and the remaining four folds are split into train and validation at an $80/20$ ratio with seed 42. The tile level uses a separate, single train/validation/test split that allocates BCSS-annotated submitters by a budget triplet, described under BCSS (Section IV-A2).

a) Tiling and resize: The WSIs are formalin-fixed paraffin-embedded diagnostic slides scanned at $\sim$0.25 µm per pixel and stored as multi-resolution pyramids. They are read through OpenSlide and tiled by the MPP-anchored procedure of Section III-A with window size $W = 512$, stride $S = 512$, target MPP $\mu = 0.25\,\mu\text{m/px}$, resize side length $H = 224$, brightness threshold $\theta_w = 0.85$, chroma threshold $\theta_g = 0.01$, and tissue-fraction threshold $\theta_t = 0.5$, computed over slide thumbnails capped at 1024 pixels per side. At the tile level, 64 centers per slide are randomly sampled per epoch from the candidate set; at the slide level, the full enumeration of valid centers is fed to the feature-precompute stage of Section III-C1 and then sub-sampled at rate $\rho = 0.25$ per training-time bag.

b) Labels: PAM50 [47] is a 50-gene expression-based classifier that partitions breast tumors into four intrinsic molecular subtypes, namely luminal A, luminal B, HER2-enriched, and basal-like, plus a normal-like class. We adopt these five published labels as the classes of the main subtype-classification task of Section III-C2, giving $C_s = 5$ subtype classes; submitters with a missing or ambiguous PAM50 call are dropped from the slide-level cohort entirely. The MAFs and CNV segment files feed the four mutational-signature subtasks of Section III-C2: MAFs are passed to SigProfiler-MatrixGenerator [86], [87] to produce the SBS, DBS, and ID mutation matrices, and the CNV segments are passed to its CNV mode to produce the CN48 matrix. SigProfilerAssignment [88] then decomposes each tumor's mutation spectrum into a non-negative exposure vector against the corresponding COSMIC

	Active sigs. / tumor	Catalog size C_e	Patients labeled
sbs_regression	4.0 ± 1.1	97	968
dbs_regression	1.0 ± 0.2	22	178
id_regression	1.5 ± 0.6	25	873
cnv_regression	2.8 ± 1.0	26	1,084

Table I: **Per-subtask COSMIC catalog size C_e, label coverage, and exposure-vector sparsity.** For each signature class we report the catalog size, the number of TCGA-BRCA submitters with a non-zero exposure vector (out of 1,093), and the mean $\pm$ SD number of signatures with non-zero exposure per tumor. Exposure vectors are sparse: only a few of the C_e signatures are active in any tumor. DBS coverage is roughly $5\times$ below the other three, so dbs_regression is excluded from aggregator training.

reference catalog. Each exposure vector is normalized to sum to 1 per tumor and used as the target of the corresponding regression subtask: sbs_regression, dbs_regression, id_regression, and cnv_regression. Of the four, dbs_regression is dropped from training because DBS exposure vectors are available for only 178 TCGA-BRCA submitters (Table I), roughly $5\times$ fewer than for the other three classes; the aggregator is therefore optimized with three signature-regression auxiliary targets: sbs_regression, id_regression, and cnv_regression.

2) BCSS: The tissue-segmentation pretext task of Section III-B is supervised by the Breast Cancer Semantic Segmentation (BCSS) dataset [89], a publicly released collection of pixel-level tissue annotations on a subset of the TCGA-BRCA cohort. Amgad et al. recruited 25 participants tiered by experience (two senior pathologists, three junior pathology residents, and twenty medical students) to delineate tissue regions in 151 triple-negative breast cancer slides through the Digital Slide Archive web platform, with task assignment by experience (the twenty-one most challenging slides reserved for senior reviewers, the remaining one hundred thirty distributed evenly across all participants) and two phases of senior-pathologist review and correction. The procedure produced 20 340 polygonal regions over an average representative region of interest (ROI) of $1.18\,\mathrm{mm}^2$ per slide; the rest of each labeled slide carries no tissue-typing ground truth.

a) Tissue classes: The released ground truth labels comprise $C_t = 21$ biologically meaningful tissue classes spanning predominant types (tumor, stroma, lymphocytic infiltrate, necrosis/debris), non-predominant types (adipose tissue, blood vessels, blood, glandular secretions, mucoid material, artifacts), and challenging types (plasma cells, other immune infiltrate, normal ducts/acini, metaplasia, lymphatics, skin adnexa, angioinvasion, nerves, DCIS), plus a reserved outside_roi code at index 0 that marks pixels falling outside any annotated region, for $C_t + 1 = 22$ classes in total. By the index-0 ignore convention of (3), tiles whose entire mask is outside_roi contribute zero tissue-segmentation gradient while still driving the shared encoder through the magnification, JigMag, and hematoxylin pretexts.

b) Split budget: Of the 1,093 TCGA-BRCA submitters, only 151 carry BCSS annotations, and within each annotated slide only the representative ROI is labeled, so BCSS supervision is scarce along both axes. A naive $70/15/15$ patient split would land only ~ 23 labeled submitters in the held-out test set, leaving the tile-level tissue-segmentation evaluation under-powered. We therefore allocate the labeled submitters by an asymmetric budget triplet of $0.4/0.2/0.4$ across train/-validation/test, then draw from the unlabeled remainder to fill each split until the overall slide-count membership matches the $70/15/15$ ratio. The cohort-level split fractions are preserved while approximately 40% of the BCSS-labeled slides land in the held-out test set, roughly $2.6\times$ the test-set labeled budget that uniform allocation would yield, at the cost of a slightly smaller labeled training subset.

3) Coarse tumor annotations: The per-class slide-level attention maps produced by DualCLAM (Section III-C4) are interpreted against the coarse tumor-area dataset of Banerjee et al. [90], a complementary TCGA-BRCA annotation resource that supplies tumor outlines at whole-slide scope, the granularity BCSS does not provide. The dataset contains coarse polygonal tumor annotations and the derived binary masks for 1,039 TCGA-BRCA WSIs (one slide per case), released as a paired inclusion set (incl_vec) and exclusion set (excl_vec) per slide; the latter cuts out non-tumor inclusions from within the tumor outline. The annotations are deliberately coarse: the authors argue that fine-grained delineation is labor-intensive and unnecessary for downstream tasks such as region selection for mitotic-count assessment, tumor microenvironment analysis, and tumor growth-pattern analysis.

We use these polygons strictly for qualitative attention-map interpretation, not as a training target or a quantitative segmentation benchmark. For each test slide, the predicted-class attention map of Section III-C4 is overlaid on the slide thumbnail and on a level-0 ROI crop with the Banerjee tumor outline traced on top, so the figure shows whether the subtype-driving attention concentrates inside the pathologist-delineated tumor area.

B. Model and training configuration

The pipeline runs in three sequential stages: tile-encoder training, one-shot frozen-encoder feature precomputation per slide, and slide-level aggregator training under 5-fold cross-validation. Both training stages share the same Lightning recipe: AdamW with learning rate 10^{-4} and weight decay 10^{-2}, a cosine-annealing learning-rate schedule with $T_{\max} = 10$ epochs and $\eta_{\min} = 10^{-6}$, 16-bit mixed precision, sync_batchnorm across devices, and seed 42 throughout. Training runs to completion in both stages with no early stopping, and the final checkpoint for each run is the one with minimum validation loss across all epochs, exported to a stable path so the next stage can load it without re-discovering the run.

a) Tile encoder: The tile encoder is trained for 50 epochs on the multi-task tile dataloader of Section IV-A1, with a per-GPU batch size of 128 tiles and 2 gradient-accumulation steps (effective batch 256). The five encoder configurations evaluated in this report differ only in which subset of the three self-supervised pretext heads is active (none, magnification, jigmag, hematoxylin, full); the BCSS-supervised tissue-segmentation head is always on, so each run sees a multi-task loss in the sense of (2). After fitting, the best-validation checkpoint is automatically re-tested on the held-out tile-level test split, and its weights are saved for the next two stages to load.

b) Tile feature precomputation: Before the aggregator is trained, the frozen tile encoder is run once over every valid tissue tile of every slide carrying a PAM50 label. The pass is single-device, runs in fp16, processes 128 tiles per encoder forward call, and skips slides whose per-slide output file already exists so the cache can be built incrementally. Each slide produces a `<slide_id>.pt` containing the $(K_{\mathcal{B}}, D)$ feature tensor and the corresponding tile-center coordinates. No encoder forward pass is performed at aggregator-training time; the aggregator consumes these cached bags directly (Section III-C1).

c) Slide aggregator: The aggregator is trained on the cached features for 100 epochs per cross-validation fold, with a per-GPU batch size of 64 bags and 2 gradient-accumulation steps (effective batch 128). After each fold finishes fitting, the held-out test partition is scored using the best-validation checkpoint to produce one per-slide prediction table per fold; the five tables are pooled before the metrics of Section IV-C are computed.

C. Evaluation metrics

Both the subtyping head and the signature regression heads sit on the same aggregator and share the 5-fold split of the slide-level cohort defined in Section IV-A1, so the two protocols below operate on the same pooled set of out-of-fold test slides (subject to per-slide label availability for each signature class).

a) Subtyping: Slide-level subtyping is evaluated by pooling the five out-of-fold test predictions and computing per-class one-vs-rest AUROC, subtyping macro AUROC (unweighted mean over classes with both positive and negative support), and support-weighted AUROC of the $\arg\max$ prediction. Standard errors and 95% percentile confidence intervals are obtained by row-resampling bootstrap on the pooled prediction frame.

b) Signature regression: The four signature subtasks (`sbs_regression`, `dbs_regression`, `id_regression`, `cnv_regression`) are scored with a rank- and distribution-based protocol rather than an L2 fit, because the regression head trains under the KL loss of (12) and emits a probability distribution over the COSMIC reference catalog. For each slide, we identify the top-m signatures by true exposure (T) and the top-m by predicted exposure (P), with $m = \min(k, n_{\text{active}})$ where n_{active} is the slide's number of strictly positive true exposures; capping m by the activity count prevents a signal-free predictor from gaining spurious matches against zero-exposure tie-fillers, since COSMIC profiles are sparse and a fixed top-k would pad the true set with zeros on most slides. We report top-k IoU $|T \cap P|/|T \cup P|$ as the headline rank-agreement metric and top-k overlap $|T \cap P|/m$ (equal to both precision@k and recall@k at this cardinality) at $k \in \{3, 5\}$. Two k-independent distribution metrics complement the top-k summary: the mean total variation $\frac{1}{2}\|\hat{\mathbf{y}}_e - \mathbf{y}_e\|_1$ between the predicted and true exposure distributions, and the signature-active macro AUROC of "signature is active" ($\mathbf{y}_{e,i} > 0$) scored by predicted exposure. Two baselines anchor the chance floor: a prevalence predictor that emits the cohort-mean exposure profile for every slide (the marginal bar a useful model must beat), and a random predictor whose exact expected IoU and overlap@k follow from a hypergeometric draw with $\mathbb{E}[\text{overlap@}k \mid m] = m/C_e$; because m varies per slide with n_{active}, the reported random-baseline overlap is the across-slide mean of m/C_e rather than a single ratio. Both baselines score signature-active macro AUROC 0.5 by construction. Prevalence is a meaningful threshold because PAM50 subtype is correlated with mutational-signature exposure (basal-like tumors carry substantially different signature profiles from luminal tumors): a model that only learns to identify subtype from morphology already beats the random baseline through the subtype prior, but it cannot beat prevalence unless it picks up signature-specific signal beyond that prior. As with subtyping, the five folds' slides are pooled and a row-resampling bootstrap on the pooled mean yields standard errors and 95% percentile confidence intervals.

	DB index	**Avg. cos distance**
No pretext tasks	4.806	0.684
Magnification	4.344	0.648
JigMag	4.243	0.623
Hematoxylin	4.603	0.665
All pretext tasks	**4.157**	**0.591**

Table II: **Tile-level feature-space quality on the held-out BCSS-annotated test tiles, by tile-encoder pretext configuration.** Each row corresponds to one tile-encoder run: *No pretext tasks* uses tissue segmentation alone, the three single-pretext rows add the indicated auxiliary head on top of tissue segmentation, and *All pretext tasks* activates all four. For each configuration, encoder embeddings are extracted on the held-out tile-level test split, every tile is assigned its dominant ground-truth tissue class (BCSS taxonomy), and we report the Davies-Bouldin index (DB; lower = better-separated tissue clusters) and the macro-averaged within-class mean cosine distance (lower = tighter intra-class compactness). Best value per column is in bold.

D. Results

Table II reports tile-encoder feature-space quality on the held-out BCSS-annotated test tiles across the five pretext configurations. All three single-pretext runs improve on the tissue-segmentation-only baseline on both metrics, though the improvement is uneven: JigMag accounts for most of the available gain (11.7% DB and 8.9% cosine improvement over baseline), magnification is moderate (9.6% DB, 5.3% cosine), and hematoxylin alone is closer to baseline than to the other

two (4.2% DB, 2.8% cosine). Activating all three pretexts simultaneously yields the best feature space on both metrics (4.157 DB, 0.591 cosine), but the marginal gain over JigMag alone is small ($\sim 2\%$ DB, $\sim 5\%$ cosine); the pretexts compose without interference, but most of the benefit is already captured by the multi-scale jigsaw signal alone.

Figure 3 supports the quantitative ranking qualitatively. On the no-pretext panel, the projected tiles form a single diffuse cloud where tumor, stroma, and lymphocytic-infiltrate tiles overlap, with the "Agnostic" (gray) points, tiles whose dominant predicted class falls below $p = 0.5$, filling the center of the embedding. The all-pretext panel resolves this into recognizable clusters: stroma (orange) and necrosis or debris (red) become spatially distinct, fat (purple) detaches as its own region at the top, and the agnostic mass shrinks. JigMag and magnification each reproduce most of this separation in the single-pretext panels, while hematoxylin alone leaves a structure closer to the baseline, matching the per-pretext ranking in Table II. Adding any auxiliary pretext also reshapes the top-5 class composition: the no-pretext encoder is the only one to surface "Exclude" as a top-5 class, while all four pretext-augmented encoders displace it with "Fat", suggesting that even a single auxiliary head shifts the encoder from routing uncertain tiles into the junk class toward resolving them as a distinct biological tissue type.

Table III reports the pooled out-of-fold macro AUROC across the 20 trained slide-aggregator configurations. The dominant axis of variation is the tile-encoder pretext setting: holding the aggregator fixed, switching the encoder from tissue-segmentation-only to the four-pretext encoder lifts the macro AUROC by roughly $+0.077$ on average (e.g. for plain CLAM-MB, from 0.6092 to 0.6880), consistent with the tile-level feature-space analysis above and indicating that most of the slide-level signal is bottlenecked by the encoder rather than the aggregator. The aggregator-side changes are smaller. Switching the attention layout from single-branch (CLAM-SB) to multi-branch (CLAM-MB), holding the encoder pretext and the auxiliary subtask fixed, contributes a mean gain of only $+0.005$ macro AUROC, order of magnitude smaller compared to the pretext effect. Adding a signature-regression auxiliary subtask (CLAM $\rightarrow$ DualCLAM with one of `sbs_regression`, `id_regression`, `cnv_regression`) also improves subtyping, but only in some cases: the best DualCLAM-MB-full configuration (`cnv`, macro 0.7038) is above plain CLAM-MB-full (0.6880), however, on average the auxiliary head costs about -0.004 macro AUROC. The signature-regression heads therefore, in most cases, deliver the per-signature spatial explanations of Figure 4 at a small cost to the main subtype objective rather than at a benefit to it.

Figure 5 shows the predicted-class attention maps of the DualCLAM-MB aggregator on three TCGA-BRCA test slides, overlaid with the Banerjee et al. coarse tumor outline. On all three slides the warm-attention tiles concentrate within the inclusion polygons of the tumor outline and the cool-attention tiles concentrate outside, even though the aggregator is supervised only by the slide-level PAM50 label and the optional signature-regression head, and never sees the Banerjee mask during training. This supports the interpretive claim of Section III-C4 that the per-class attention branches localize on subtype-diagnostic morphology, and (as detailed in the caption) the model further discounts known slide-preparation artifacts such as the blue pen-marker region on TCGA-A1-A0SK.

Table IV reports the signature-regression metrics for the three trained subtasks (SBS, ID, CNV) against the prevalence and random baselines. Across all three subtasks the model only marginally beats prevalence on the IoU and overlap@k metrics, and on CNV the prevalence baseline is actually the best column (top-3 IoU 0.2677 vs. 0.2610 for the best model row). The signature-active macro AUROC tells the same story more sharply: prevalence and random both score 0.5 by construction, while the model sits near chance on CNV (0.4959–0.5217) and below chance on ID (0.41–0.44) and SBS (0.38), which means on those two subtasks the model is, on average, worse than random at ranking patients by whether a given signature is active. This is consistent with the expectation of Section IV-C that not every mutational signature leaves an H&E-visible morphological imprint; when a signature's mutational process has no morphological correlate, the regression head can do no better than the prevalence profile, and below-chance ID and SBS macro AUROCs further suggest that the aggregator has learned a subtype-aligned signature profile that predicts the wrong direction for any signature whose activity diverges from subtype.

The per-signature AUROCs in the rightmost column surface the subset of signatures whose morphology the aggregator most reproducibly attempts to localize. A small set ranks in the top-3 consistently across configurations (CN17 for CNV, ID14 for ID, and SBS33 and SBS39 for SBS) present whether the encoder is full-pretext or none, and whether the attention is single-branch or multi-branch. These are the candidate H&E-recognizable signatures in the TCGA-BRCA cohort; the rest of the catalog is too morphologically silent for this aggregator to discriminate.

Figure 4 examines what each per-signature attention head actually localizes on. The top row shows the CNV head's highest- and lowest-attention tiles on a basal-like slide whose ground-truth-active CN signatures (CN11, CN12, CN17, CN20) are dominated by WGD-associated (CN11, CN12) and HRD (CN17) processes [81]. The high-attention tiles localize on tumor regions with enlarged-appearing cells and prominent nuclei, a morphology compatible with the polyploid cytology WGD produces; the lowest-attention tiles fall on uniform stromal regions where this signal is absent.

The bottom row shows the SBS head on a different basal-like slide whose ground-truth-active signatures (SBS1, SBS2, SBS101) include the APOBEC cytidine-deaminase signature SBS2 (paired with SBS13 in the broader APOBEC family), a process consistently associated with high-grade tumor biology [83]. The high-attention tiles concentrate on densely cellular tumor regions with marked nuclear pleomorphism, the morphological hallmark of high-grade breast cancer; the lowest-attention tiles fall on acellular, hyalinised stroma.

These observations are anecdotal (two slides, one head each) but they suggest the per-signature branches are not collapsing to a shared subtype prior: the CNV and SBS heads on two basal-like slides attend to morphologically distinct cues, each

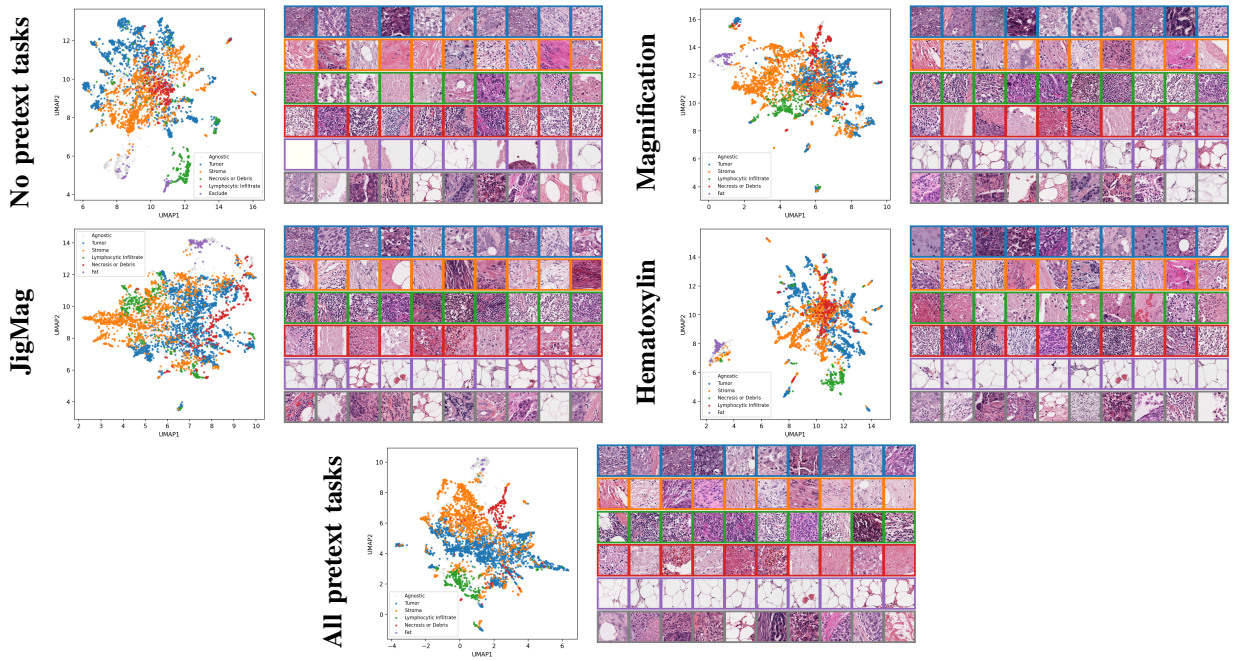


Figure 3: **Visualization of the tile-level feature space after Stage 2.** For each pretext-task configuration, we randomly sampled 2% of tiles from each slide in the independent test cohort and projected their 512-dimensional feature representations to two dimensions using UMAP (the scatter on the left of each panel). Each tile is shaded by its dominant tissue class, defined as the class whose proportion within the tile, as predicted by the tissue-segmentation tile encoder, is at least $p = 0.5$; tiles in which no class meets this threshold are labeled "Agnostic" (grey). For clarity we display only the five tissue classes most frequent among the sampled tiles; representative tiles for each class are shown to the right of each scatter, outlined in the matching cluster color. Colors are assigned by per-configuration frequency rank, so the same color can denote different classes across panels. Across all five configurations the two most common classes are tumor (blue) and stroma (orange). The third (green) and fourth (red) are lymphocytic infiltrate and necrosis/debris, but their order varies: necrosis/debris is third under the no-pretext and hematoxylin configurations, whereas lymphocytic infiltrate is third under magnification, JigMag, and the full configuration. The fifth class (purple) is fat in every configuration except no-pretext, where it is instead `Exclude` (tiles the encoder predicts as the unknown/exclude category).

Architecture	**Tile pretext** / **Subtask**	**Macro AUROC**		**Weighted AUROC**	
		full	none	full	none
CLAM-SB	-	0.6829 ± 0.0074	0.6215 ± 0.0082	0.7060 ± 0.0080	0.6373 ± 0.0087
CLAM-MB	-	0.6880 ± 0.0073	0.6092 ± 0.0081	0.7109 ± 0.0080	0.6193 ± 0.0085
DualCLAM-SB	cnv	0.6985 ± 0.0075	0.6252 ± 0.0082	0.7202 ± 0.0080	$\mathbf{0.6421 \pm 0.0086}$
	id	0.6766 ± 0.0080	0.5888 ± 0.0090	0.7012 ± 0.0092	0.6050 ± 0.0095
	sbs	0.6799 ± 0.0075	0.5866 ± 0.0083	0.7082 ± 0.0083	0.6009 ± 0.0088
DualCLAM-MB	cnv	$\mathbf{0.7038 \pm 0.0074}$	$\mathbf{0.6264 \pm 0.0081}$	$\mathbf{0.7257 \pm 0.0079}$	0.6419 ± 0.0086
	id	0.6757 ± 0.0079	0.6067 ± 0.0090	0.7020 ± 0.0090	0.6223 ± 0.0092
	sbs	0.6823 ± 0.0076	0.6082 ± 0.0090	0.7098 ± 0.0084	0.6202 ± 0.0092

Table III: **Slide-level PAM50 subtyping performance on the held-out TCGA-BRCA test slides.** Each row reports one aggregator configuration; columns list the subtyping macro AUROC and the support-weighted AUROC of the $\arg\max$ prediction. Predictions from the five cross-validation folds are pooled before scoring (Section IV-C); reported values are the pooled mean and $\pm$ the bootstrap standard error of that mean over 1,000 row-resampled draws. Best in each column is in bold.

Subtask	Model	Tile pretext	IoU		Overlap@k		Total variation	Macro AUROC	Top-3 signatures by AUROC
			k=3	k=5	k=3	k=5			
CNV	Random	-	0.0635	0.0695	0.0978	0.1084	0.8915	0.5000	-
	Prevalence	-	**0.2677**	**0.3010**	**0.3696**	**0.4118**	0.7639	0.5000	-
	DualCLAM-SB	full	0.2590	0.2872	0.3591	0.3962	0.7615	**0.5217**	CN18 (0.6501) CN17 (0.6227) CN1 (0.6193)
		none	0.2621	0.2921	0.3627	0.4024	0.7682	0.5057	CN15 (0.6571) CN6 (0.5877) CN11 (0.5847)
	DualCLAM-MB	full	0.2610	0.2887	0.3623	0.3988	**0.7611**	0.5057	CN17 (0.6152) CN18 (0.6102) CN1 (0.6090)
		none	0.2588	0.2890	0.3587	0.3987	0.7679	0.4959	CN6 (0.5626) CN17 (0.5585) CN26 (0.5410)
ID	Random	-	0.0483	0.0483	0.0614	0.0614	0.9386	0.5000	-
	Prevalence	-	0.1614	0.1614	0.1913	0.1913	0.8924	**0.5000**	-
	DualCLAM-SB	full	0.1731	0.1730	0.2046	0.2045	0.8855	0.4434	ID1 (0.6472) ID14 (0.5112) ID23 (0.4989)
		none	0.1651	0.1650	0.1967	0.1965	0.8851	0.4294	ID24 (0.5159) ID11 (0.5151) ID14 (0.5095)
	DualCLAM-MB	full	**0.1737**	**0.1736**	**0.2046**	**0.2046**	**0.8835**	0.4343	ID1 (0.5332) ID14 (0.5146) ID23 (0.4911)
		none	0.1691	0.1690	0.2007	0.2006	0.8853	0.4154	ID14 (0.5260) ID11 (0.5243) ID23 (0.4800)
SBS	Random	-	0.0185	0.0236	0.0303	0.0405	0.9586	0.5000	-
	Prevalence	-	0.3299	0.4365	0.4609	0.5680	0.6864	**0.5000**	-
	DualCLAM-SB	full	**0.3359**	0.4411	0.4676	0.5716	0.6835	0.3823	SBS39 (0.7118) SBS33 (0.6244) SBS89 (0.6027)
		none	0.3348	0.4390	0.4663	0.5694	0.6876	0.3835	SBS39 (0.5947) SBS33 (0.5624) SBS91 (0.5549)
	DualCLAM-MB	full	0.3366	0.4401	0.4684	0.5706	**0.6809**	0.3819	SBS39 (0.7050) SBS33 (0.6286) SBS87 (0.5548)
		none	0.3370	**0.4433**	**0.4690**	**0.5737**	0.6908	0.3768	SBS88 (0.5517) SBS7a (0.5316) SBS2 (0.5268)

Table IV: **Slide-level signature-agreement metrics on the held-out TCGA-BRCA test slides.** For each trained signature subtask (`sbs_regression`, `id_regression`, `cnv_regression`; `dbs_regression` excluded per Table I) we report top-k IoU $|T \cap P|/|T \cup P|$ and top-k overlap $|T \cap P|/m$ at $k \in \{3, 5\}$, the mean total variation $\frac{1}{2}\|\hat{\mathbf{y}}_e - \mathbf{y}_e\|_1$ between predicted and true exposure distributions, and the signature-active macro AUROC (Section IV-C). Each row is shown alongside the prevalence baseline (cohort-mean exposure profile) and the random baseline (uniform distribution, hypergeometric IoU/overlap expectation) so the model row is interpretable against both the marginal bar a useful model must beat and the no-information floor. Predictions from the five cross-validation folds are pooled before scoring; reported values are the pooled mean over $1{,}000$ row-resampled draws.

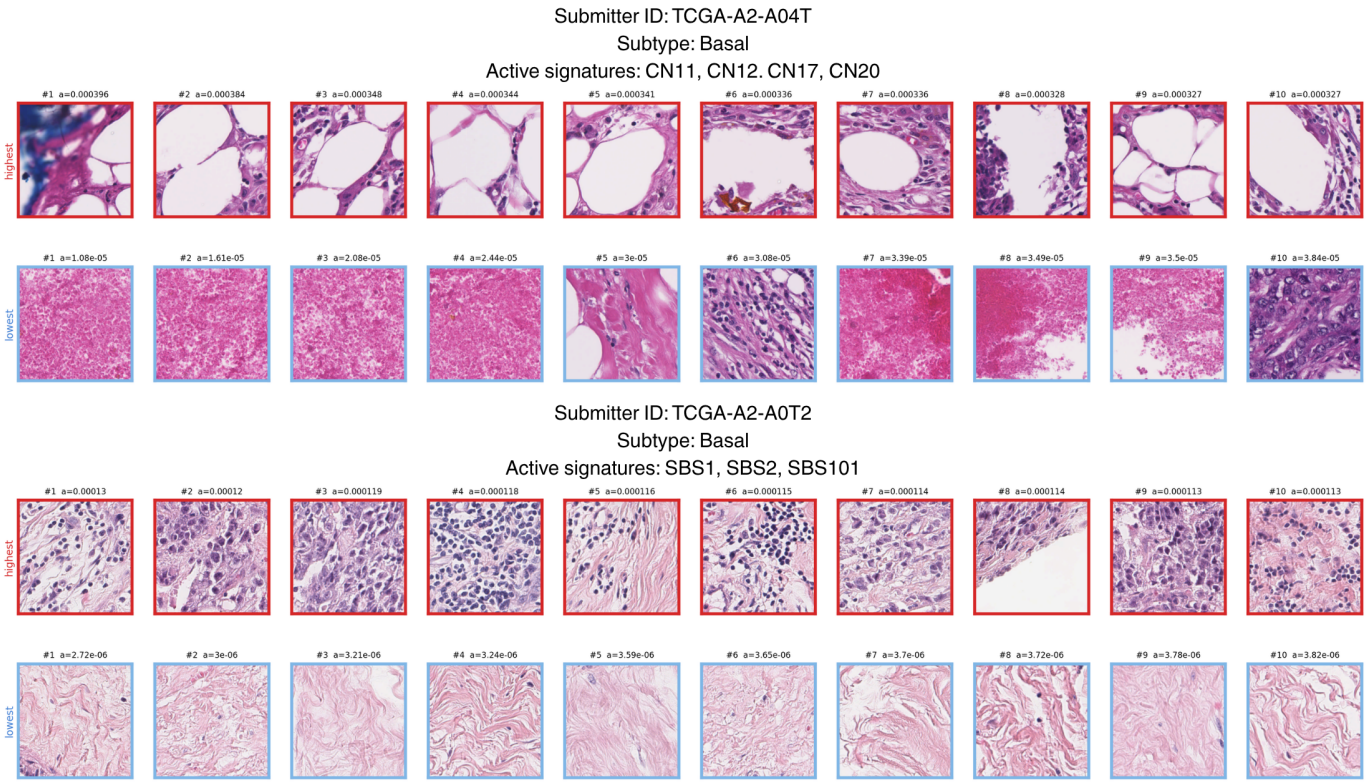


Figure 4: **Per-signature-subtask extreme-attention tiles from the multi-branch DualCLAM aggregator on two TCGA-BRCA test slides.** Each row corresponds to one (slide, signature class) pair: the top row visualizes the `cnv_regression` attention head on slide TCGA-A2-A04T (Basal subtype, ground-truth-active CN signatures $11, 12, 17, 20$), and the bottom row visualizes the `sbs_regression` attention head on slide TCGA-A2-A0T2 (Basal subtype, ground-truth-active SBS signatures $1, 2, 101$). Each row displays the top-10 highest-attention tiles for that subtask framed in red and the top-10 lowest-attention tiles framed in blue, with the raw attention weight printed above each tile. Both slides carry the same PAM50 subtype yet the highlighted tiles look morphologically distinct between the two signature heads, suggesting that the per-subtask attention branches of Section III-C2 respond to signature-specific features rather than collapsing to a shared subtype prior.

consistent with the dominant biology of its target signature class. The strongest evidence for head-specific (rather than slide-specific) localization would come from processing the same slide through both heads.

V. Discussion

The empirical headline of this report is that the tile-encoder pretext design dominates the slide-level subtyping signal. Switching the encoder from tissue-segmentation-only to the four-pretext configuration (tissue segmentation, magnification, JigMag, hematoxylin) lifts the pooled out-of-fold macro AUROC by approximately $+0.077$, which is roughly $5\times$ the gain of any MIL-side architectural change we tested. The qualitative UMAP analysis of Figure 3 corroborates the lift: the all-pretext encoder produces tissue clusters that are visibly more compact and better-separated, the agnostic-tile mass shrinks, and "Fat" replaces "Exclude" in the displayed top-5 classes, suggesting the encoder routes fewer uncertain tiles into a catch-all junk class.

The subtype attention maps emerge as a useful interpretability tool without explicit supervision. Although the aggregator sees only the slide-level PAM50 label, the predicted-class attention heads of the multi-branch CLAM and DualCLAM aggregators consistently concentrate on regions inside the Banerjee coarse tumor outline and avoid clear non-tumor regions, including slide-preparation artifacts such as the blue pen-marker region on TCGA-A1-A0SK (Figure 5). This qualitative tumor-mask agreement is the strongest empirical support for the per-class attention-map interpretation of Section III-C4, and it emerges without the model showing the tumor outline at training time.

The best-performing slide-level configuration is DualCLAM-MB with the CNV signature-regression head on top of the four-pretext encoder, reaching macro AUROC 0.7038 ± 0.0074 and weighted AUROC 0.7257 ± 0.0079 (Table III). The gain over plain CLAM-MB-full is small ($+0.016$ macro AUROC), and the SBS and ID auxiliary heads do not improve subtyping, so the DualCLAM contribution to the main classification objective is modest and class-dependent rather than a uniform architectural advance. We interpret this as an interpretability gain at small subtype cost: the per-signature attention heads add a spatial explanation that plain CLAM does not provide (Figure 4), while the main subtype metric is only marginally affected.

The signature-regression heads underperform across all three trained classes: top-k IoU and Overlap@k sit only marginally

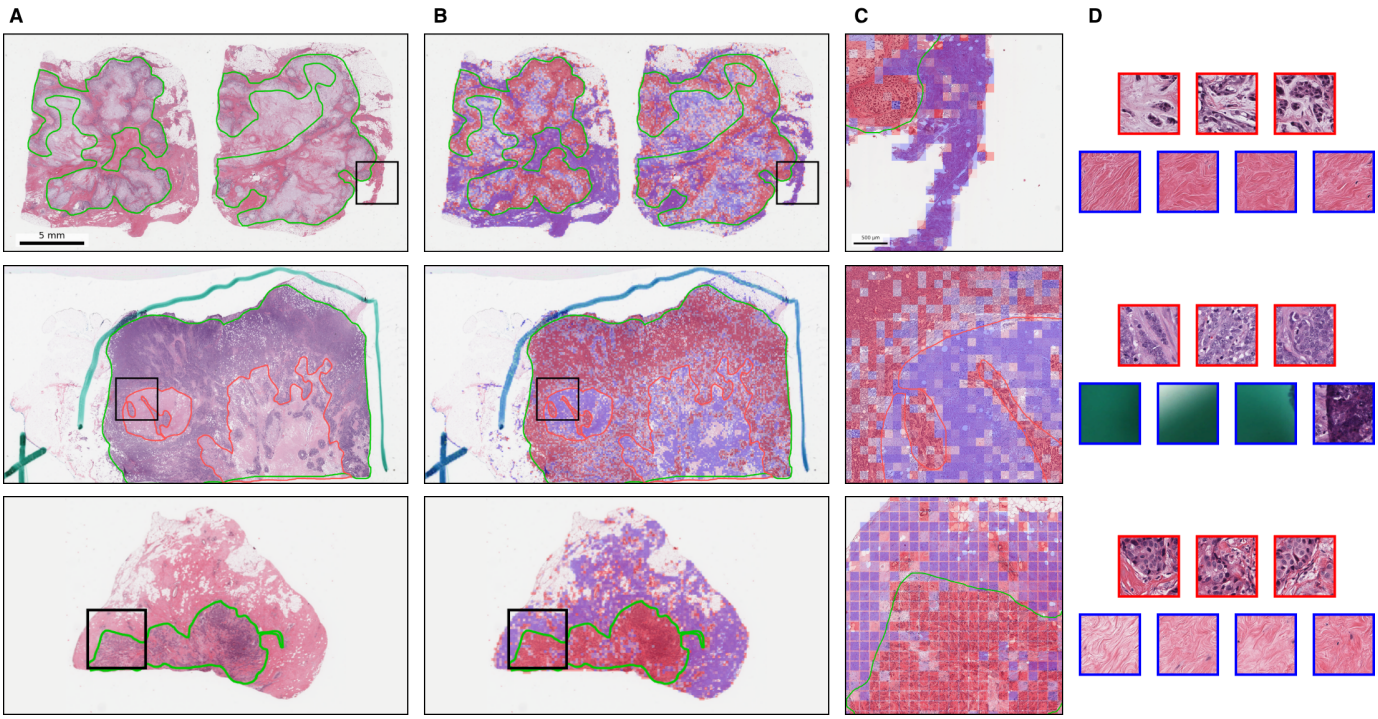


Figure 5: **Per-slide attention maps from the multi-branch DualCLAM aggregator on three TCGA-BRCA test slides.** Each row shows one slide laid out across four columns (A–D). ***A. Slide thumbnail:*** Slide thumbnail with the coarse tumor outline from Banerjee et al. [90] traced on top: tumor area in green and non-tumor exclusions in red. ***B. Attention heatmap:*** Predicted-class attention painted over every tile center sampled by the aggregator, colored from blue (low attention) to red (high attention), with a black rectangle marking the ROI shown enlarged in C. ***C. Boundary close-up:*** A level-0 crop on the tumor boundary (the rectangle in B), chosen so the boundary behavior of the attention is visible. ***D. Extreme-attention tiles:*** The three highest-attention tiles framed in red on top, and the four lowest-attention tiles framed in blue on bottom. In the middle row (TCGA-A1-A0SK), three of the four lowest-attention tiles capture the green pen-marker artifact at the lower-left of the slide, consistent with the aggregator learning to discount slide-preparation artifacts rather than attending to them as morphologically informative.

above the prevalence baseline, and the signature-active macro AUROC hovers near the 0.5 chance line (Table IV). Three properties of the labels make this largely expected. First, the exposure targets are extremely sparse: only 1.5–4 of the 25–97 catalog signatures are active in any tumor (Table I). On near-one-hot targets like these, the cohort-averaged KL objective is minimized by predicting the cohort-mean exposure, so a head that cannot extract tumor-specific signal collapses to exactly the prevalence-level behavior we observe; continuous-exposure regression is a poor match for what is closer to a sparse, near-binary activity pattern. Second, only a minority of mutational signatures are known to leave a morphologically recognizable trace on H&E (Section II-C); for the rest, no image-recoverable signal exists for any objective to learn. Third, label coverage is thin relative to target width: each head must fit a 25- to 97-dimensional exposure vector from at most $\sim$1,000 labeled slides (Table I).

Although the signature-regression metrics sit close to the prevalence baseline overall, the per-signature AUROCs surface a small but stable subset of signatures that the aggregator can rank: CN17 for CNV, ID14 for ID, and SBS33 and SBS39 for SBS (Table IV). These signatures consistently rank in the top-3 of their respective classes across the four DualCLAM configurations, and on the two illustrative slides of Figure 4 the per-signature attention heads attend to morphologically plausible cues consistent with the dominant biology of each slide's active signatures: polyploid cytology on the WGD-dominated CNV slide, and dense high-grade-tumor morphology on the APOBEC-positive SBS slide. A same-slide CNV-vs-SBS attention comparison would strengthen the head-specificity claim.

Whether this stable signature subset—or any of the results above—generalizes beyond our setup is the open question, and the most basic limit on the answer is data. Every result in this report comes from a single cohort: TCGA-BRCA supplies the training, validation, and test data for both stages. This leaves the reported performance open to inflation from per-site batch effects [62] and cohort-specific selection bias, which a single-cohort design cannot disentangle from genuine morphological signal. External validation on an independent BRCA cohort with comparable staining and scanning conventions is the natural next step, and would test whether the encoder-pretext advantage and the unsupervised tumor-mask alignment hold beyond TCGA.

Data is not the only axis the design holds fixed; the encoder is another. This report does not benchmark against pathology

foundation models, and the ResNet50 tile encoder is a deliberate scoping choice rather than an oversight. The question we set out to answer is whether a small set of self-supervised pretext tasks can improve a tile encoder under limited human-label supervision; beginning from an encoder already pre-trained on millions of whole-slide images would confound the very effect we wanted to isolate. Demonstrating a pretext gain on top of a foundation backbone would also demand far more than the $\sim$1,000 TCGA-BRCA cases available, and the size of these models made fine-tuning impractical here, as loading the encoder and the per-slide tile bags together repeatedly exhausted GPU memory. We therefore fixed the encoder to ResNet50 for this study. Given the encoder-bottleneck finding, however, a foundation-model backbone is the most likely route to substantially higher subtyping performance, and is the first item we would benchmark in a follow-up.

The encoder was the easiest component to hold fixed; the signature head is a deeper compromise. Its objective is the one already discussed—regressing a sparse, high-dimensional exposure vector rewards predicting the cohort mean—and the design is equally coarse in space. The multi-branch aggregator gives the subtype task one attention branch per class ($C_s = 5$), so every subtype gets its own spatial map, but each signature subtask receives a single branch that must account for its entire C_e-dimensional exposure vector at once. Genuine per-signature attention would require $97 + 25 + 26 = 148$ branches across the three trained classes, about $50\times$ the present three, and is infeasible at the current label coverage, so the head cannot localize any individual signature within its class. Both problems share one fix: shrink the output. Restricting each head to a small, biologically-defined panel (e.g. APOBEC, HRD, WGD, CIN, MSI, MMRd) would make the per-target multi-branch design tractable and turn the task into a near-binary multilabel classification rather than a high-dimensional regression, at the cost of hypothesis generation, since a fixed known panel cannot surface unanticipated morphology-signature links. A data-driven alternative keeps only the top-k most active signatures in the cohort, which would at least test whether the most prevalent processes leave recoverable morphology.

Architecture aside, the more fundamental constraint is simply the amount of labeled data. Each signature head is trained on around a thousand submitters (Table I) and the tile-level segmentation is evaluated on only $\sim$60 labeled slides across the $C_t = 21$ tissue classes. The existing labels can at least be used more sharply: training the signature heads on subtype-balanced batches would test whether their below-chance ID and SBS AUROCs reflect a subtype confound rather than a hard limit on H&E-recoverable signal. But the obvious remedy is more data, and two routes differ in ambition. The narrower one is more breast cancer: additional TCGA-BRCA slides, or a second BRCA cohort, would tighten every estimate above. The broader one is to recast the main task from breast-cancer subtyping to a pan-cancer objective, where a far larger and more diverse pool of labeled slides would give the auxiliary signature heads orders of magnitude more training signal than a single tumor type can supply.

One last source of uncertainty is internal to the training itself: every reported metric is a point estimate from a single seed per configuration, and the bootstrap intervals capture sampling variation over patients rather than over training initializations. Across-seed variance is therefore uncharacterized. Repeating each configuration under several seeds would quantify it, though we treat this as low priority, since the headline encoder-versus-aggregator gap is large enough that seed noise is unlikely to overturn it.

VI. Conclusion

This report introduced AUGUR, a two-stage pipeline that classifies PAM50 breast cancer subtypes from H&E whole-slide images while exposing per-class spatial attention maps as a visual explanation of each subtype prediction. The tile encoder is trained with one supervised pretext (BCSS tissue segmentation) and three self-supervised pretexts (magnification, JigMag, hematoxylin); the slide-level DualCLAM aggregator extends CLAM with auxiliary mutational-signature regression heads that supply additional per-signature attention maps. On the held-out TCGA-BRCA test cohort under 5-fold cross-validation, the best configuration, DualCLAM-MB on the four-pretext encoder with a CNV signature-regression head reaches macro AUROC 0.7038 ± 0.0074 and weighted AUROC 0.7257 ± 0.0079, and the pretext design choice on its own accounts for roughly $5\times$ the gain of any MIL-side architectural change. The subtype attention maps qualitatively concentrate within pathologist-delineated tumor regions despite never being supervised on them, and a small subset of COSMIC signatures (CN17, ID14, SBS33, SBS39) is consistently ranked across architectural variants.

VII. Contributions

We introduced AUGUR, a two-stage pipeline that predicts PAM50 breast-cancer subtype from H&E whole-slide images while exposing per-class spatial attention as a visual explanation of each prediction. Its contributions are:

- **An interpretable two-stage pipeline**: a multi-task tile encoder (BCSS tissue segmentation plus the self-supervised magnification, JigMag, and hematoxylin pretexts) feeding a DualCLAM aggregator that extends CLAM with auxiliary mutational-signature regression heads, with both stages emitting per-class attention maps.
- **An encoder-dominated subtyping result**: the tile-encoder pretext design is the dominant axis of variation, adding $+0.077$ macro AUROC over a segmentation-only baseline ($\sim 5\times$ any MIL-side change), with the best configuration reaching macro AUROC 0.7038 ± 0.0074.
- **Validated visual evidence**: the predicted-subtype attention maps concentrate within pathologist-delineated tumor regions without ever being supervised on them, and the per-signature heads consistently rank a small, stable COSMIC subset (CN17, ID14, SBS33, SBS39).

References

[1] M. Y. Lu, D. F. K. Williamson, T. Y. Chen, R. J. Chen, M. Barbieri, and F. Mahmood, "Data-efficient and weakly supervised computational pathology on whole-slide images," *Nature Biomedical Engineering*, vol. 5, no. 6, pp. 555–570, Jun. 2021.

[2] F. Sanchez-Vega *et al.*, "Oncogenic Signaling Pathways in The Cancer Genome Atlas," *Cell*, vol. 173, no. 2, pp. 321–337.e10, Apr. 2018.

[3] A. Drilon *et al.*, "Efficacy of Larotrectinib in TRK Fusion-Positive Cancers in Adults and Children," *The New England Journal of Medicine*, vol. 378, no. 8, pp. 731–739, Feb. 2018.

[4] J. Mateo *et al.*, "Delivering precision oncology to patients with cancer," *Nature Medicine*, vol. 28, no. 4, pp. 658–665, Apr. 2022.

[5] J. W. Kim *et al.*, "Clinical implementation of next-generation sequencing testing and genomically-matched therapy: A real-world data in a tertiary hospital," *Scientific Reports*, vol. 15, p. 2171, Jan. 2025.

[6] H. K. Karakoyun, C. Sayar, and K. Yararbaş, "Challenges in clinical interpretation of next-generation sequencing data: Advantages and Pitfalls," *Results in Engineering*, vol. 20, p. 101421, Dec. 2023.

[7] U. H. Frixen *et al.*, "E-cadherin-mediated cell-cell adhesion prevents invasiveness of human carcinoma cells." *Journal of Cell Biology*, vol. 113, no. 1, pp. 173–185, Apr. 1991.

[8] P. W. B. Derksen *et al.*, "Somatic inactivation of E-cadherin and p53 in mice leads to metastatic lobular mammary carcinoma through induction of anoikis resistance and angiogenesis," *Cancer Cell*, vol. 10, no. 5, pp. 437–449, Nov. 2006.

[9] B. Qiu *et al.*, "HIF2α-Dependent Lipid Storage Promotes Endoplasmic Reticulum Homeostasis in Clear-Cell Renal Cell Carcinoma," *Cancer Discovery*, vol. 5, no. 6, pp. 652–667, Jun. 2015.

[10] W. Du *et al.*, "HIF drives lipid deposition and cancer in ccRCC via repression of fatty acid metabolism," *Nature Communications*, vol. 8, no. 1, p. 1769, Nov. 2017.

[11] S.-l. Chen *et al.*, "GYS1 induces glycogen accumulation and promotes tumor progression via the NF-κB pathway in Clear Cell Renal Carcinoma," *Theranostics*, vol. 10, no. 20, pp. 9186–9199, Jul. 2020.

[12] N. Pescador *et al.*, "Hypoxia Promotes Glycogen Accumulation through Hypoxia Inducible Factor (HIF)-Mediated Induction of Glycogen Synthase 1," *PLoS ONE*, vol. 5, no. 3, p. e9644, Mar. 2010.

[13] F. Grignani *et al.*, "Fusion proteins of the retinoic acid receptor-α recruit histone deacetylase in promyelocytic leukaemia," *Nature*, vol. 391, no. 6669, pp. 815–818, Feb. 1998.

[14] R. J. Lin, L. Nagy, S. Inoue, W. Shao, W. H. Miller, and R. M. Evans, "Role of the histone deacetylase complex in acute promyelocytic leukaemia," *Nature*, vol. 391, no. 6669, pp. 811–814, Feb. 1998.

[15] S. H. Swerdlow *et al.*, *WHO Classification of Tumours of Haematopoietic and Lymphoid Tissues*, 4th ed. WHO Press, 2008.

[16] T. Sehgal and P. Sharma, "Auer rods and faggot cells: A review of the history, significance and mimics of two morphological curiosities of enduring relevance," *European Journal of Haematology*, vol. 110, no. 1, pp. 14–23, 2023.

[17] G. I. Evan *et al.*, "Induction of apoptosis in fibroblasts by c-myc protein," *Cell*, vol. 69, no. 1, pp. 119–128, Apr. 1992.

[18] C. M. Eischen, J. D. Weber, M. F. Roussel, C. J. Sherr, and J. L. Cleveland, "Disruption of the ARF–Mdm2–p53 tumor suppressor pathway in Myc-induced lymphomagenesis," *Genes & Development*, vol. 13, no. 20, pp. 2658–2669, Oct. 1999.

[19] J. L. Dy-Ledesma, J. D. Khoury, R. L. M. C. Agbay, M. Garcia, R. N. Miranda, and L. J. Medeiros, "Starry Sky Pattern in Hematopoietic Neoplasms: A Review of Pathophysiology and Differential Diagnosis," *Advances in Anatomic Pathology*, vol. 23, no. 6, pp. 343–355, Nov. 2016.

[20] J. van der Laak, G. Litjens, and F. Ciompi, "Deep learning in histopathology: The path to the clinic," *Nature Medicine*, vol. 27, no. 5, pp. 775–784, May 2021.

[21] A. J. Evans *et al.*, "US Food and Drug Administration Approval of Whole Slide Imaging for Primary Diagnosis: A Key Milestone Is Reached and New Questions Are Raised," *Archives of Pathology & Laboratory Medicine*, vol. 142, no. 11, pp. 1383–1387, Apr. 2018.

[22] S. Mukhopadhyay *et al.*, "Whole Slide Imaging Versus Microscopy for Primary Diagnosis in Surgical Pathology," *The American Journal of Surgical Pathology*, vol. 42, no. 1, pp. 39–52, Jan. 2018.

[23] N. Dimitriou, O. Arandjelović, and P. D. Caie, "Deep Learning for Whole Slide Image Analysis: An Overview," *Frontiers in Medicine*, vol. 6, Nov. 2019.

[24] N. Coudray *et al.*, "Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning," *Nature Medicine*, vol. 24, no. 10, pp. 1559–1567, Oct. 2018.

[25] J. N. Kather *et al.*, "Pan-cancer image-based detection of clinically actionable genetic alterations," *Nature Cancer*, vol. 1, no. 8, pp. 789–799, Aug. 2020.

[26] Y. Fu *et al.*, "Pan-cancer computational histopathology reveals mutations, tumor composition and prognosis," *Nature Cancer*, vol. 1, no. 8, pp. 800–810, Aug. 2020.

[27] J. M. Niehues *et al.*, "Generalizable biomarker prediction from cancer pathology slides with self-supervised deep learning: A retrospective multi-centric study," *Cell Reports Medicine*, vol. 4, no. 4, p. 100980, Apr. 2023.

[28] M. Ilse, J. M. Tomczak, and M. Welling, "Attention-based Deep Multiple Instance Learning," Jun. 2018.

[29] H. D. Couture *et al.*, "Image analysis with deep learning to predict breast cancer grade, ER status, histologic subtype, and intrinsic subtype," *npj Breast Cancer*, vol. 4, no. 1, p. 30, Sep. 2018.

[30] M. I. Jaber *et al.*, "A deep learning image-based intrinsic molecular subtype classifier of breast tumors reveals tumor heterogeneity that may affect survival," *Breast Cancer Research*, vol. 22, no. 1, p. 12, Jan. 2020.

[31] S. Fremond *et al.*, "Interpretable deep learning model to predict the molecular classification of endometrial cancer from haematoxylin and eosin-stained whole-slide images: A combined analysis of the PORTEC randomised trials and clinical cohorts," *The Lancet Digital Health*, vol. 5, no. 2, pp. e71–e82, Feb. 2023.

[32] A.-C. Woerl *et al.*, "Deep Learning Predicts Molecular Subtype of Muscle-invasive Bladder Cancer from Conventional Histopathological Slides," *European Urology*, vol. 78, no. 2, pp. 256–264, Aug. 2020.

[33] R. J. Chen *et al.*, "Towards a general-purpose foundation model for computational pathology," *Nature Medicine*, vol. 30, no. 3, pp. 850–862, Mar. 2024.

[34] H. Xu *et al.*, "A whole-slide foundation model for digital pathology from real-world data," *Nature*, vol. 630, no. 8015, pp. 181–188, Jun. 2024.

[35] E. Vorontsov *et al.*, "Virchow: A Million-Slide Digital Pathology Foundation Model," Jan. 2024.

[36] A. Filiot *et al.*, "Scaling Self-Supervised Learning for Histopathology with Masked Image Modeling," p. 2023.07.21.23292757, Dec. 2024.

[37] M. Y. Lu *et al.*, "A visual-language foundation model for computational pathology," *Nature Medicine*, vol. 30, no. 3, pp. 863–874, Mar. 2024.

[38] R. Schmitz *et al.*, "Burkitt lymphoma pathogenesis and therapeutic targets from structural and functional genomics," *Nature*, vol. 490, no. 7418, pp. 116–120, Oct. 2012.

[39] J. N. Honeyman *et al.*, "Detection of a Recurrent DNAJB1-PRKACA Chimeric Transcript in Fibrolamellar Hepatocellular Carcinoma," *Science (New York, N.Y.)*, vol. 343, no. 6174, pp. 1010–1014, Feb. 2014.

[40] J. R. Craig, R. L. Peters, H. A. Edmondson, and M. Omata, "Fibrolamellar carcinoma of the liver: A tumor of adolescents and young adults with distinctive clinico-pathologic features," *Cancer*, vol. 46, no. 2, pp. 372–379, Jul. 1980.

[41] I. P. Tomlinson *et al.*, "Germline mutations in FH predispose to dominantly inherited uterine fibroids, skin leiomyomata and papillary renal cell cancer," *Nature Genetics*, vol. 30, no. 4, pp. 406–410, Apr. 2002.

[42] P. Pollard *et al.*, "Accumulation of Krebs cycle intermediates and over-expression of HIF1α in tumours which result from germline FH and SDH mutations," *Human Molecular Genetics*, vol. 14, no. 15, pp. 2231–2239, Aug. 2005.

[43] J. Adam *et al.*, "Renal Cyst Formation in Fh1-Deficient Mice Is Independent of the Hif/Phd Pathway: Roles for Fumarate in KEAP1 Succination and Nrf2 Signaling," *Cancer Cell*, vol. 20, no. 4, pp. 524–537, Oct. 2011.

[44] M. J. Merino, C. Torres-Cabala, P. Pinto, and W. Marston Linehan, "The morphologic spectrum of kidney tumors in hereditary leiomyomatosis and renal cell carcinoma (HLRCC) syndrome," *American Journal of Surgical Pathology*, vol. 31, no. 10, pp. 1578–1585, Oct. 2007.

[45] C. Tognon *et al.*, "Expression of the ETV6-NTRK3 gene fusion as a primary event in human secretory breast carcinoma," *Cancer Cell*, vol. 2, no. 5, pp. 367–376, Nov. 2002.

[46] A. Skálová *et al.*, "Mammary Analogue Secretory Carcinoma of Salivary Glands, Containing the ETV6-NTRK3 Fusion Gene: A Hitherto Undescribed Salivary Gland Tumor Entity," *The American Journal of Surgical Pathology*, vol. 34, no. 5, p. 599, May 2010.

[47] J. S. Parker *et al.*, "Supervised Risk Predictor of Breast Cancer Based on Intrinsic Subtypes," *Journal of Clinical Oncology*, vol. 27, no. 8, pp. 1160–1167, Mar. 2009.

[48] D. C. Koboldt *et al.*, "Comprehensive molecular portraits of human breast tumours," *Nature*, vol. 490, no. 7418, pp. 61–70, Oct. 2012.

[49] C. A. Livasy *et al.*, "Phenotypic evaluation of the basal-like subtype of invasive breast carcinoma," *Modern Pathology*, vol. 19, no. 2, pp. 264–271, Feb. 2006.

[50] J. Guinney *et al.*, "The consensus molecular subtypes of colorectal cancer," *Nature Medicine*, vol. 21, no. 11, pp. 1350–1356, Nov. 2015.

[51] J. K. Greenson *et al.*, "Pathologic Predictors of Microsatellite Instability in Colorectal Cancer," *The American Journal of Surgical Pathology*, vol. 33, no. 1, p. 126, Jan. 2009.

[52] C. Isella *et al.*, "Stromal contribution to the colorectal cancer transcriptome," *Nature Genetics*, vol. 47, no. 4, pp. 312–319, Apr. 2015.
[53] K. Sirinukunwattana *et al.*, "Image-based consensus molecular subtype (imCMS) classification of colorectal cancer using deep learning," *Gut*, vol. 70, no. 3, pp. 544–554, Mar. 2021.
[54] B. E. Howitt *et al.*, "Association of Polymerase e–Mutated and Microsatellite-Instable Endometrial Cancers With Neoantigen Load, Number of Tumor-Infiltrating Lymphocytes, and Expression of PD-1 and PD-L1," *JAMA Oncology*, vol. 1, no. 9, pp. 1319–1323, Dec. 2015.
[55] D. A. Levine *et al.*, "Integrated genomic characterization of endometrial carcinoma," *Nature*, vol. 497, no. 7447, pp. 67–73, May 2013.
[56] C. B. Gilks, E. Oliva, and R. A. Soslow, "Poor interobserver reproducibility in the diagnosis of high-grade endometrial carcinoma," *The American Journal of Surgical Pathology*, vol. 37, no. 6, pp. 874–881, Jun. 2013.
[57] S. Volinsky-Fremond *et al.*, "Prediction of recurrence risk in endometrial cancer with multimodal deep learning," *Nature Medicine*, vol. 30, no. 7, pp. 1962–1973, Jul. 2024.
[58] N. Agrawal *et al.*, "Integrated Genomic Characterization of Papillary Thyroid Carcinoma," *Cell*, vol. 159, no. 3, pp. 676–690, Oct. 2014.
[59] C. M. Rudin *et al.*, "Molecular subtypes of small cell lung cancer: A synthesis of human and mouse model data," *Nature Reviews Cancer*, vol. 19, no. 5, pp. 289–297, May 2019.
[60] S. Arslan *et al.*, "A systematic pan-cancer study on deep learning-based prediction of multi-omic biomarkers from routine pathology images," *Communications Medicine*, vol. 4, no. 1, p. 48, Mar. 2024.
[61] K. A. Hoadley *et al.*, "Cell-of-Origin Patterns Dominate the Molecular Classification of 10,000 Tumors from 33 Types of Cancer," *Cell*, vol. 173, no. 2, pp. 291–304.e6, Apr. 2018.
[62] F. M. Howard *et al.*, "The impact of site-specific digital histology signatures on deep learning model accuracy and bias," *Nature Communications*, vol. 12, no. 1, p. 4423, Jul. 2021.
[63] A. Echle *et al.*, "Clinical-Grade Detection of Microsatellite Instability in Colorectal Tumors by Deep Learning," *Gastroenterology*, vol. 159, no. 4, pp. 1406–1416.e11, Oct. 2020.
[64] T. Lazard *et al.*, "Deep learning identifies morphological patterns of homologous recombination deficiency in luminal breast cancers from whole slide images," *Cell Reports Medicine*, vol. 3, no. 12, p. 100872, Dec. 2022.
[65] G. Campanella *et al.*, "Clinical-grade computational pathology using weakly supervised deep learning on whole slide images," *Nature medicine*, vol. 25, no. 8, pp. 1301–1309, Aug. 2019.
[66] Z. Shao *et al.*, "TransMIL: Transformer based correlated multiple instance learning for whole slide image classification," in *Proceedings of the 35th International Conference on Neural Information Processing Systems*, ser. NIPS '21. Red Hook, NY, USA: Curran Associates Inc., Dec. 2021, pp. 2136–2147.
[67] S. Graham *et al.*, "Hover-Net: Simultaneous segmentation and classification of nuclei in multi-tissue histology images," *Medical Image Analysis*, vol. 58, p. 101563, Dec. 2019.
[68] S. Graham *et al.*, "One model is all you need: Multi-task learning enables simultaneous histology image segmentation and classification," *Medical Image Analysis*, vol. 83, p. 102685, Jan. 2023.
[69] R. Mormont, P. Geurts, and R. Marée, "Multi-Task Pre-Training of Deep Neural Networks for Digital Pathology," *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 2, pp. 412–421, Feb. 2021.
[70] N. A. Koohbanani, B. Unnikrishnan, S. A. Khurram, P. Krishnaswamy, and N. Rajpoot, "Self-Path: Self-supervision for Classification of Pathology Images with Limited Annotations," Aug. 2020.
[71] O. Ciga, T. Xu, and A. L. Martel, "Self supervised contrastive learning for digital histopathology," *Machine Learning with Applications*, vol. 7, p. 100198, Mar. 2022.
[72] X. Wang *et al.*, "Transformer-based unsupervised contrastive learning for histopathological image classification," *Medical Image Analysis*, vol. 81, p. 102559, Oct. 2022.
[73] B. Korbar *et al.*, "Looking Under the Hood: Deep Neural Network Visualization to Interpret Whole-Slide Image Analysis Outcomes for Colorectal Polyps," in *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Jul. 2017, pp. 821–827.
[74] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," in *2017 IEEE International Conference on Computer Vision (ICCV)*, Oct. 2017, pp. 618–626.
[75] S. A. Javed, D. Juyal, H. Padigela, A. Taylor-Weiner, L. Yu, and A. Prakash, "Additive MIL: Intrinsically Interpretable Multiple Instance Learning for Pathology," *Advances in Neural Information Processing Systems*, vol. 35, pp. 20 689–20 702, Dec. 2022.
[76] D. Rymarczyk, A. Pardyl, J. Kraus, A. Kaczyńska, M. Skomorowski, and B. Zieliński, "ProtoMIL: Multiple Instance Learning with Prototypical Parts for Whole-Slide Image Classification," Sep. 2022.
[77] J. Liu *et al.*, "PAMIL: Prototype Attention-Based Multiple Instance Learning for Whole Slide Image Classification," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, M. G. Linguraru *et al.*, Eds. Cham: Springer Nature Switzerland, 2024, pp. 362–372.
[78] L. B. Alexandrov *et al.*, "Signatures of mutational processes in human cancer," *Nature*, vol. 500, no. 7463, pp. 415–421, Aug. 2013.
[79] L. B. Alexandrov *et al.*, "The repertoire of mutational signatures in human cancer," *Nature*, vol. 578, no. 7793, pp. 94–101, Feb. 2020.
[80] A. Degasperi *et al.*, "Substitution mutational signatures in whole-genome–sequenced cancers in the UK population," *Science*, vol. 376, no. 6591, p. abl9283, Apr. 2022.
[81] C. D. Steele *et al.*, "Signatures of copy number alterations in human cancer," *Nature*, vol. 606, no. 7916, pp. 984–991, Jun. 2022.
[82] S. Nik-Zainal *et al.*, "Landscape of somatic mutations in 560 breast cancer whole-genome sequences," *Nature*, vol. 534, no. 7605, pp. 47–54, Jun. 2016.
[83] S. A. Roberts *et al.*, "An APOBEC cytidine deaminase mutagenesis pattern is widespread in human cancers," *Nature Genetics*, vol. 45, no. 9, pp. 970–976, Sep. 2013.
[84] A. Goode, B. Gilbert, J. Harkes, D. Jukic, and M. Satyanarayanan, "OpenSlide: A vendor-neutral software foundation for digital pathology," *Journal of Pathology Informatics*, vol. 4, p. 27, Sep. 2013.
[85] A. Ruifrok and D. Johnston, "Quantification of histochemical staining by color deconvolution," *Analytical and quantitative cytology and histology*, vol. 23, no. 4, Aug. 2001.
[86] E. N. Bergstrom *et al.*, "SigProfilerMatrixGenerator: A tool for visualizing and exploring patterns of small mutational events," *BMC Genomics*, vol. 20, no. 1, p. 685, Aug. 2019.
[87] A. Khandekar *et al.*, "Visualizing and exploring patterns of large mutational events with SigProfilerMatrixGenerator," *BMC Genomics*, vol. 24, no. 1, p. 469, Aug. 2023.
[88] M. Díaz-Gay *et al.*, "Assigning mutational signatures to individual samples and individual somatic mutations with SigProfilerAssignment," *Bioinformatics*, vol. 39, no. 12, p. btad756, Dec. 2023.
[89] M. Amgad *et al.*, "Structured crowdsourcing enables convolutional segmentation of histology images," *Bioinformatics*, vol. 35, no. 18, pp. 3461–3467, Sep. 2019.
[90] S. Banerjee *et al.*, "Comprehensive Dataset of Coarse Tumor Annotations for The Cancer Genome Atlas Breast Invasive Carcinoma," in *Bildverarbeitung für die Medizin 2025*, C. Palm *et al.*, Eds. Wiesbaden: Springer Fachmedien, 2025, pp. 260–265.

APPENDIX