Predicting breast cancer biomarkers from HE histopathology using weakly supervised whole-slide deep learning

Jennifer Ho
jenph@stanford.edu

Summer Royal
sroyal@stanford.edu

Luke Zhao
lukezhao@stanford.edu

Abstract

Treatment eligibility for cancer patients is determined by observing the biomarker status of the patient's tumor. The current standard is to identify the presence of various biomarkers via an assay but these assays are costly, time consuming, and are not always available in low-resource settings. Prior work suggests that staining whole slide images of tumor biopsies with Hematoxylin & Eosin (H&E) can elucidate these biomarkers. Our project investigates whether attention-based multiple-instance learning can predict the presence of three key biomarkers (ER, PR, and HER2) from whole-slide H&E-stained images. We compared three patch encoders of increasing pathology specificity: ImageNet-pretrained ResNet-50, UNI (ViT-L/16 pretrained on over 100 million pathology patches via DINOv2), and CONCH (ViT-B/16 trained via contrastive vision-language pretraining on histopathology image-caption pairs). We implemented three slide-level aggregation strategies: mean-pooling with logistic regression, attention-based CLAM-SB, and our proposed Tumor-Aware CLAM (a lightweight residual gating mechanism that is intended to extract tumor patches and tumor features before attention aggregation to reduce confounding from stromal and adipose tissue.)

Our patient-level 5-fold stratified cross-validation with bootstrap confidence intervals showed that all encoder-aggregator combinations performed near random chance (pooled AUROCs ranging 0.48 to 0.60 across the three prediction tasks.) Pathology foundation models (UNI, CONCH) did not consistently outperform ResNet-50, and attention-based MIL methods did not consistently outperform mean pooling. Although Tumor-Aware CLAM showed promising performance in a single train-test split, it failed to improve under cross-validation. Our attention heatmap analysis suggests the observed single-split gains were results of unevenly distributed data partitioning rather than a predictive signal. Overall, our findings suggest that limited dataset size, not encoder choice or aggregation strategy, is the primary constraint on WSI-based biomarker prediction performance in this setting.

1. Introduction

Diagnosis of cancer in the modern age depends not only on identifying what and where a tumor is located, but also on understanding the biological features of that tumor that determine how it should be treated. Three of the key "biomarkers" that pathologists use to identify a tumor's biological subtype are: the estrogen receptor (ER), the progesterone receptor (PR), the gene HER2[28]. The presence of these biomarkers determine whether or not a patient is eligible for endocrine therapy, HER2-targeted agents, and cytotoxic chemotherapy[10]. Physicians currently take a biopsy and use immunohistochemistry (IHC) or fluoresence in situ hybridization (FISH) to measure the presence of these biomarkers in the biopsy tissue, but ICC and FISH assays are costly, take time, can be misinterpreted by the pathologist, and are not always available in low-income or resource-limited areas[28][10].

Cancer biopsies are often stained with hematoxylin and eosin to see information about the tissue's architecture, nuclear morphology, and the tumor's environment[7]. If clinicians had a way of predicting ER, PR, and HER2 status from H&E-stained images alone, clinicians could provide treatment plans for patients sooner and at lower cost. A major challenge in solving this problem is the fact that these whole-slide images (WSIs) have extremely large file sizes; WSIs are typically $10^4 \times 10^5$ pixels at $20\times$ magnification[18][23]. Patch-level diagnostic labels are rarely available, which poses another problem. Thus, learning from WSIs requires researchers to aggregate information across thousands of patches only using slide-level labels. We wanted to investigate whether or not we could create an attention-based multiple-instance learning model to improve predictions ER, PR, and HER2 status from H&E-stained slides. The input to our algorithm is a H&E-stained image of a breast cancer biopsy. We then use a frozen patch encoder (ResNet-50, UNI, or CONCH) followed by a multiple-instance learning aggregator (mean-pooling with logistic regression, CLAM-SB, or our own Tumor-Aware CLAM) to output a predicted binary biomarker status (ER-positive/negative, PR-positive/negative, or HER2-positive/negative).

Our initial single-split ablation experiments suggested that our own Tumor-Aware CLAM model did not sig-

nificantly improve performance over CLAM-SB, but attention-based MIL could improve biomarker prediction from H&E whole-slide images. However, after running cross-validation, encoder comparisons, and attention heatmap analysis, we found that the observed gains in individual train-test splits did not generalize consistently across the cohort. These findings suggest that limitations in dataset size and data variability are likely a more significant barrier to performance than the choice of encoder or aggregation architecture. More broadly, our study highlights the importance of rigorous evaluation when developing weakly supervised models for computational pathology.

2. Related Work

A growing body of work leveraging different approaches has investigated whether routine H&E morphology contains enough information to infer molecular features of breast cancer. First, several studies ask whether breast cancer molecular phenotypes are visible in routine morphology. Couture et al. showed that deep learning features from H&E images could predict breast cancer grade, ER status, histologic subtype, PAM50 intrinsic subtype, and recurrence-risk categories, suggesting that molecular state leaves partially detectable morphologic correlates [7]. Naik et al. then framed ER prediction from H&E whole-slide images as a weakly supervised multiple-instance learning problem, demonstrating that slide-level receptor labels alone can supervise WSI-scale learning [28]. Gamble et al. extended this direction to ER, PR, and HER2 prediction and analyzed associated morphologic features, making it one of the closest predecessors to our task [10]. HER2-specific efforts such as the HEROHE challenge further explored whether HER2 status could be predicted from H&E without IHC or ISH, but also highlighted the difficulty and variability of this task across teams and datasets [6]. More recently, Shamai et al. evaluated ER, PR, and ERBB2 prediction in a large multi-institutional cohort and emphasized clinically realistic use cases such as high-confidence hormone-receptor screening, second-read quality assurance, and possible false-negative detection [44]. Overall, these studies suggest that ER is usually the most learnable receptor from H&E morphology, while PR and especially HER2 are more difficult, likely because their molecular signals are less consistently reflected in routine tissue architecture.

Second, there are also other studies that focus on technical methods for learning from gigapixel WSIs with only slide-level labels. Standard convolutional encoders such as ImageNet-pretrained ResNet provided a practical baseline for extracting patch-level features from WSIs [13]. Multiple-instance learning (MIL) becomes a natural framework for WSI classification because each slide can be represented as a bag of patches with a single slide-level label. Attention-based MIL introduced a learnable, permutation-invariant aggregation mechanism that assigns different weights to different instances, making it both effective and more interpretable than simple mean pooling [18]. Campanella et al. showed that weakly supervised learning could scale to clinical-grade WSI classification using only slide-level labels, supporting the broader feasibility of weak supervision in pathology [1]. CLAM adapted attention MIL specifically for computational pathology by adding clustering-constrained learning, improving data efficiency and producing attention maps that localize high-value regions [24]. Other approaches, such as DSMIL and TransMIL, further improved WSI modeling by incorporating dual-stream instance/bag learning, self-supervised features, or Transformer-based modeling of correlations among patches [20, 45]. These methods are potentially good, because they avoid the need for expensive patch-level annotation, but they are weak in that slide-level supervision can encourage shortcut learning from artifacts, tissue composition, or non-tumor regions rather than receptor-specific tumor morphology.

Third, there are studies that use pathology-pretrained encoders instead of generic natural-image backbones. Earlier pipelines often used ResNet-like encoders pretrained on ImageNet, which are robust and accessible but not optimized for histopathology texture, staining, and tissue organization [13]. Self-supervised pathology models such as CTransPath and HIPT showed that domain-specific pretraining can produce more transferable histology representations, including features that capture local tissue patterns and larger-scale WSI structure [49, 4]. Foundation models have pushed this further: UNI is a general-purpose pathology encoder pretrained on more than 100 million H&E image patches, while CONCH uses vision-language pretraining from histopathology image-caption pairs [5, 21]. These models represent the current state-of-the-art direction for feature extraction in computational pathology, although there is not yet a single universally accepted approach for ER/PR/HER2 prediction because datasets, label definitions, train/test splits, and clinical cohorts differ substantially across studies. Our work builds on these prior categories by comparing generic ImageNet-pretrained ResNet50 features with pathology-pretrained UNI and CONCH features, and by adding a lightweight tumor-aware gating mechanism before attention aggregation. This design is motivated by a key limitation of prior MIL approaches: receptor-specific signal is expected to be concentrated in invasive tumor regions, whereas benign tissue, stroma, necrosis, artifacts, and background may introduce confounding signal during slide-level learning. Our work differs from prior studies in two main ways:

1. Rather than evaluating a single feature extractor, we compare three generations of encoders (a ImageNet-pretrained ResNet-50 baseline, UNI, and CONCH) to isolate the contribution of pathology-specific pretraining to biomarker prediction performance.

2. We introduce our own lightweight Tumor-Aware

CLAM architecture that explicitly incorporates a learned tumor prior before attention aggregation. Prior MIL approaches allow attention to focus on any tissue region, including stroma and adipose tissue, but our method emphasizes patches that are more likely to contain invasive tumor tissue.

By evaluating these design choices under patient-level cross-validation, our work provides insight into whether recent advances in pathology foundation models and tumor-focused attention mechanisms translate into measurable gains for weakly supervised biomarker prediction.

3. Methods

A whole-slide image (WSI) is too large to feed directly into a CNN; a single slide can have tens of thousands of pixels on each side, on the order of 10^4 to 10^5 pixels per dimension. Thus, we treat each slide as a bag of smaller image patches.

After tissue filtering and tiling, each slide yields a variable-length bag of non-overlapping 256×256 RGB patches at a target magnification of $20\times$. Candidate tiles are generated on a regular grid and retained only if an HSV-saturation Otsu tissue mask marks at least 50% of pixels as tissue. Across the 574 successfully tiled slides in our remote preprocessing cohort, this produced 3,890,904 total tissue patches. The number of patches per slide ranged from 61 to 27,230, with median 5,884 and mean 6,779 patches per slide.

Because the slide and not the individual patches are labeled, we formulate biomarker prediction as a weakly supervised multiple-instance learning (MIL) problem. For slide i, let $X_i = \{x_{i1}, \ldots, x_{iK_i}\}$ denote the set of tissue patches extracted from the slide, where K_i varies by slide. Each slide has a binary biomarker label $y_i \in \{0, 1\}$, where $y_i = 1$ indicates biomarker-positive status for ER, PR, or HER2. A frozen patch encoder f_θ maps each patch to an embedding $h_{ik} = f_\theta(x_{ik}) \in \mathbb{R}^D$, producing a bag of patch features $H_i = \{h_{i1}, \ldots, h_{iK_i}\}$.

The goal of the MIL aggregator g_ϕ is to map this unordered bag of patch embeddings to slide-level class logits $s_i = g_\phi(H_i) \in \mathbb{R}^2$, where $s_{i,0}$ and $s_{i,1}$ are the logits for biomarker-negative and biomarker-positive status, respectively. The predicted probability of biomarker positivity is obtained using a softmax:

$$\hat{p}_i = P(y_i = 1 \mid H_i) = \frac{\exp(s_{i,1})}{\exp(s_{i,0}) + \exp(s_{i,1})}.$$

We train the slide-level classifier using two-class cross-entropy:

$$\mathcal{L}_{\text{slide}} = -\frac{1}{N} \sum_{i=1}^{N} \log \left(\frac{\exp(s_{i,y_i})}{\exp(s_{i,0}) + \exp(s_{i,1})} \right),$$

where $y_i \in \{0, 1\}$ is the ground-truth biomarker label.s We compared three aggregation strategies: (1) mean-pooling of patch features with logistic regression, (2) attention-based MIL using CLAM-SB, and (3) a learned patch-gated variant of CLAM. Briefly, mean-pooling with logistic regression averages all patches into one slide vector and trains a linear classifier. This method tests whether global slide-level morphology is predictive without learned attention. The other two methods will be discussed in more detail below.

3.1. Patch Encoding

First, each patch is mapped to a feature vector via a frozen pretrained encoder. Because we have a small dataset (574 slides were successfully tiled/preprocessed remotely; after target-specific label filtering, we used N=193 ER, N=189 PR, and N=144 HER2 cases), we do not backpropagate into the encoder. This is done to avoid overfitting and to train many MIL variants cheaply (only need to cache features once).

We evaluate three encoders of increasing pathology specificity, which are described below:

1. ResNet-50: uses ImageNet-1K pretrained weights (torchvision IMAGENET1K_V2), with the final classification head replaced by an identity layer, yielding 2,048-dimensional features per patch. Despite being trained on natural images, ResNet-50 provides a well-established baseline and allows us to isolate the effect of the MIL aggregator independently of encoder quality.

2. UNI: ViT-L/16 pretrained via DINOv2 self-supervised learning on over 100,000 whole-slide images spanning diverse tissue types and clinical institutions [5], producing 1,024-dimensional embeddings. Pathology-specific pretraining aligns the feature space with histological concepts such as nuclear morphology and stromal texture that are uninformative to ImageNet classification but are relevant for our biomarker prediction goal.

3. CONCH: ViT-B/16 trained via contrastive language-image pretraining on paired pathology images and diagnostic text reports [21], producing 512-dimensional features. We extracted image features with projection and contrastive normalization disabled so that the MIL aggregator receives visual appearance features rather than just retrieval-optimized embeddings.

3.2. CLAM: Attention-MIL

We implement Clustering-Constrained Attention MIL [23]. Our implementation builds on the publicly available CLAM codebase [22], from which we adapted the core attention-MIL architecture; the Tumor-Aware gating mechanism, training protocol, evaluation pipeline, and cross-validation infrastructure were written by us with the help of AI tools. Here, each patch is scored independently and the slide embedding is a single attention-weighted average

of the patch features. Note there is no patch-to-patch interaction.

We define single-branch variant CLAM-SB for binary targets ER, PR, and HER2. This forward pass consists of:

1. Projection of each patch embedding to a hidden dimension H=256: $z_k = \text{Dropout}(\text{ReLU}(W_p h_k)$.

2. Scoring each patch via gated attention and bag normalization:

$$a_k = \mathbf{w}^\top \left(\tanh(\mathbf{V} z_k) \odot \sigma(\mathbf{U} z_k)\right) \tag{1}$$

$$\alpha_k = \text{softmax}_k(\mathbf{a}) = \frac{\exp(a_k)}{\sum_{j=1}^{K} \exp(a_j)} \tag{2}$$

3. Aggregate into one slide embedding: $M = \sum_k \alpha_k z_k$

4. Classify: $s = W_c M \in \mathbb{R}^2$, with $\hat{p} = \text{softmax}(s)_1$ used as the predicted probability of biomarker positivity.

The original CLAM formulation can include an auxiliary instance-level clustering loss: within each slide, the top-k highest-attention patches are treated as pseudo-positive instances and the bottom-k patches as pseudo-negative instances for a small per-class instance classifier. This encourages the attention module to separate high- and low-evidence patches in feature space. Our implementation supports this optional loss with weight $\lambda_{\text{inst}} = 0.3$, following the CLAM convention [24]. However, in the experiments reported here, we trained CLAM-SB and Tumor-Aware CLAM using only the slide-level cross-entropy objective, so the reported results should be interpreted as attention-based MIL without the auxiliary CLAM instance-clustering term. Therefore, we keep: $\mathcal{L}_{\text{MIL}} = \mathcal{L}_{\text{slide}}$

3.3. Tumor-Aware Gating

CLAM's attention can attend anywhere (including stroma, fat, artifacts, etc). Tumor-Aware MIL inserts a lightweight, per-patch tumor prior before attention to bias the model towards likely invasive tumor regions. Specifically, a small scorer (Linear $D \to 128$, ReLU, Linear $128 \to 2$ outputs two logits per patch. Softmax turns this to a per patch tumor probability $p_k \in [0, 1]$. Next, these patch features are gated by p_k before being sent through CLAM. Importantly, because no patch-level tumor annotations are available, this gate is not directly a measure of tumor probability; we evaluate whether its learned weighting improves downstream biomarker prediction and inspect attention maps for plausibility.

Multiplicative vs. Residual Gate. First, we started with a naive multiplicative gate $h_k \cdot p_k$. This design only scaled down patches; for example, an uncertain patch where $p \approx 0.5$ could lose half its signal despite potentially being useful. Due to performance, we later replaced this with a residual gate $\tilde{h}_k = h_k \cdot (1 + \gamma \cdot p_k)$,. This design maintains the strength of non-tumor patches ($\times 1$) and amplifies tumor patches by up to $\times(1+\gamma)$.

Gate Regularization. Additionally, if the scorer outputs the same or similar p for all patches, CLAM's first linear layer would absorb the scale and Tumor-Aware MIL could collapse. To mediate this, we add a gate regularizer with two optional terms: entropy, which discourages 0.5 assignment and pushes p_k towards 0 or 1, and a budget $(\bar{p} - b)^2$ which maintains the average active fraction near some target b (e.g. 0.25). This is scaled by a weight λ and prevents the gate from flagging every patch. For gated CLAM variants with gate regularization, the objective becomes:

$$\mathcal{L}_{\text{gated}} = \mathcal{L}_{\text{slide}} + \lambda_{\text{inst}} \mathcal{L}_{\text{inst}} + \lambda_{\text{gate}} \mathcal{L}_{\text{gate}}.$$

Learnable Temperature. We also introduce a learnable temperature $\tau > 0$ in the scorer's softmax: $p_k = \text{softmax}(\text{MLP}(h_k)/\tau)_1$. Initializing $\tau = 1$ and allowing it to be trained jointly with the scorer lets the model control gate sharpness: a low τ pushes the gate toward hard binary decisions, while a high τ produces a near-uniform, smooth weighting.

3.4. Why This Approach?

All publicly available datasets only had biomarker labels at the slide level, as opposed to labeling the actual features within the slide. A single slide can have thousands of image patches, many of which will not be relevant for the purpose of training a model to identify the presence of biomarkers. MIL allows us to aggregate all this patch-level information without needing patch-level annotations. We specifically chose attention-based MIL so that we could prioritize the specific regions of the slide images where the tumor is present. Attention-based MIL is permutation invariant to patch ordering, which is very beneficial for our particular use case where we have a very small amount of data and need to develop a model that's generalizable to all types of tumor biopsy slide images.

We chose to use frozen pretrained encoders per the suggestion of TA Karan Singh during our milestone check-in because we had a limited amount of data. Training a large vision model with only 574 labeled data samples would likely result in overfitting, so we used pretrained models that had the benefit of learning visual features from much larger collections of images. We chose to test out a Tumor-aware gating mechanism because we hypothesized that the model would have improved performance if it focused on the tumor morphology in order to predict biomarker status as opposed to basing its prediction on all image patches (which would include irrelevant surrounding stromal and adipose tissue). We intentionally wanted to bias our models to guide attention to the regions that are more relevant for biomarker status prediction.

There were several other approaches that we could have tried, but due to compute limitations and the time constraint, we were limited in our options. Transformer-based

slide options that explicitly model the interactions between patches are very computationally expensive and generally require larger datasets. We had to avoid approaches that would likely result in overfitting due to our small dataset size. End-to-end fine-tuning of the image encoders with the MIL aggregator was another approach we could've taken, but this approach would have been very computationally expensive and likely would have resulted in overfitting. We were not able to try patch-level classification approaches since we did not have data with patch-level labels.

3.5. Training Protocol

MIL models are trained with the Adam optimizer ($\beta_1 = 0.9$, $\beta_2 = 0.999$, $\text{lr} = 2 \times 10^{-4}$) for 20 epochs per cross-validation fold using a batch size of one slide, the same parameters as the original CLAM paper [24]. We chose the Adam optimizer because it is well-suited for smaller datasets. At each training step, we process the full bag of tissue patches from a single slide. We retain the checkpoint achieving the best validation AUROC sfor test-set evaluation. For the mean-pool + logistic regression baseline, we average all patch features within a slide to a single vector and we fit an L2-regularized logistic regression classifier. We standardize the mean-pooled slide features using statistics from the training split and fit an L2-regularized logistic regression classifier with balanced class weights, $C = 1.0$. We did not tune the baseline regularization strength here. Neither did we apply patch subsampling here.

3.6. Evaluation

For each of our targets (ER, PR, HER2 status), we looked at four different metrics:

1. AUROC: area under the receiver operating curve, which measures rank discrimination between the positive and negative classes independent from the classification threshold.

2. AUPRC: area under the precision recall curve, which measures precision vs. recall trade-offs.

3. Balanced accuracy: mean per-class recall. This is to account for label imbalances in our dataset

4. Brier score: mean squared error of predicted probabilities against binary labels. Lower scores indicate better calibration quality: $BS = \frac{1}{N}\sum_{i=1}^{N}(\hat{p}_i - y_i)^2$ where $\hat{p}_i$ is the predicted probability of biomarker positivity and $y_i \in 0, 1$ is the ground-truth label.

4. Data

We used the TCGA-BRCA cohort dataset from the NCI Genomic Data Commons portal [15]. This full TCGA-BRCA dataset has 1,000 slide images of tumor biopsies that were stained with Hemotoxilin & Eosin, and we worked with a labeled subset of 574 images that had the shared features we desired. The images are whole-slide images, not cropped or distorted in any way. An example image from the dataset is shown below in Figure 1. 574 slides were successfully tiled/preprocessed remotely; after target-specific label filtering, we used N=193 ER, N=189 PR, and N=144 HER2 cases. The slides are labeled with ER status, PR status, and HER2 status that are coded as a binary positive or negative level based on the results of IHC and FISH assays.

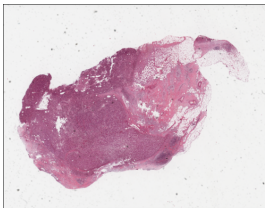

Figure 1. H&E-stained whole-slide image from the TCGA-BRCA cohort (case TCGA-C8-A1HM, ER-positive).

Our TCGA-BRCA cohort has an imbalance in the representation of different hormone receptors; 78% of patients are ER-positive, 70.9% are PR-positive, and 25% are HER2 positive. This imbalance might have affected our interpretation of the AUPRC curve, which tends to track the positive class prevalence - a naive majority-class classifier achieves AUPRC approximately equal to the positive rate. We therefore report both AUROC (threshold-free rank discrimination) and AUPRC (precision-recall summary), and treat AUROC as the primary metric.

We did the following preprocessing steps for each slide before feature extraction:

1. Read slides using OpenSlide at 20× effective magnification (selecting the closest native level and compensating for any residual downsample factor). We extracted non-overlapping 256×256-pixel tiles from a regular grid over the tissue region.

2. Converted tiles from RGB to HSV and thresholded the saturation channel using Otsu's method. We discarded any tiles that had less than 50% of the pixels exceeding Otsu threshold (because these correspond to empty space or fatty tissue). This was to ensure we are only keeping tiles that have a substantial amount of stained tumor tissue.

3. We used patient-level splits to prevent having label leakage from the same patient's slides appearing in both the training and the test set. We make 5-fold stratified cross-validation folds on the target label. We reserved a validation set for early stopping within each fold's training portion. Since each patient appears in the test set of exactly one fold, pooling out-of-fold predictions yields a single collection of held-out predictions over the full cohort. We could then compute the pooled AUROC and AUPRC on this cohort.

We computed bootstrap 95% confidence intervals on these pooled metrics by resampling patients with replacement. We had 1,000 resamples in total.

5. Experiments

We evaluate three slide-level aggregation strategies of increasing sophistication: a mean-pool and logistic regression baseline, CLAM-SB, and Tumor-Aware CLAM on three binary biomarker targets ER, PR, and HER2 status. We evaluate across ResNet-50, UNI, and CONCH encoders. For the Tumor-Aware gate, we also evaluate across gate variants.

5.1. Single-Split Ablation

We first evaluate on a single patient-level 70/15/15 train/validation/test split, training on the train set, selecting checkpoints by validation AUROC, and reporting held-out test metrics.

5.2. Tumor-Aware Gate Ablation

To test whether the tumor-aware gate's instability can be improved, we compare four gate configurations under the same 5-fold cross validation protocol: our naive multiplicative gate and three residual variants (residual gating alone, residual gating with regularization via entropy and budget terms, and residual with both regularization and learnable temperature evaluated across all three encoders and targets.

5.3. Cross-Validated Evaluation

Due to our small cohort, a single split potentially yields noisy results. In order to verify the single-split ablation's robustness, we re-evaluate all models under patient-level 5-fold stratified cross-validation. Splits are made at the patient level to prevent leakage between folds. Splits are also stratified on the target to preserve class balance. Because each patient is held out in exactly one fold, we pool the out-of-fold predictions into one set of held-out predictions over the full cohort and report the pooled AUROC and AUPRC, the mean $\pm$ standard deviation across folds, and bootstrap 95% confidence intervals (1,000 patient level resamples).

5.4. Attention Heatmap Analysis

In addition to the quantitative AUROC/AUPRC results above, we include a qualitative attention-heatmap analysis to inspect where the MIL model places high attention on representative slides. Because these visualizations are based on a small number of examples, we treat them as qualitative success and failure analysis rather than validation of tumor localization. To interpret what the attention mechanism attends to, we visualize CLAM-SB attention weights spatially over two ER-positive slides (chosen from a success and a failure case). This analysis examines whether attention localizes to diagnostically relevant tumor tissue or confounding non-tumor regions to motivate our tumor-aware mechanism.

5.5. Single-Split Ablation

We first ran a single patient-level split using ResNet-50 features to establish the pipeline and compare mean-pooled logistic regression against attention MIL and tumor-aware MIL. Figure 2 below summarizes our initial single-split ablation using ResNet-50 patch features. This experiment used one patient-level 70/15/15 train/validation/test split and was intended as an initial pipeline sanity check before larger cross-validated experiments. We compared a mean-pool + logistic regression baseline, CLAM-SB, and Tumor-Aware CLAM across ER, PR, and HER2 receptor status. Checkpoints for the MIL models were selected by validation AUROC.

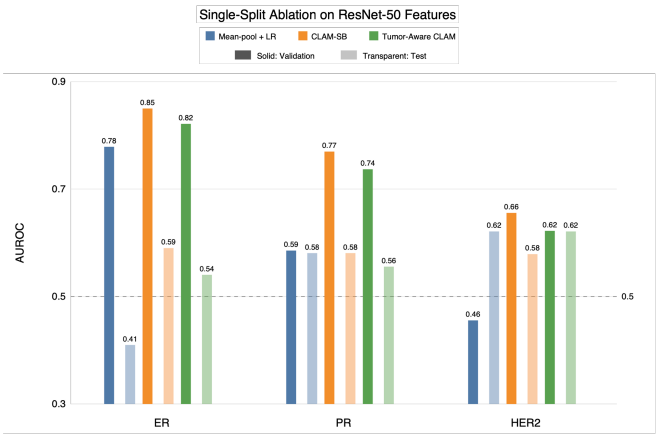

Figure 2. Initial Single-Split AUROC Summary Plot

On the validation split, attention-based models appeared to generally improve over the mean-pool baseline. CLAM-SB improved validation AUROC for all three targets: ER increased from 0.779 to 0.850, PR from 0.586 to 0.770, and HER2 from 0.456 to 0.656. Tumor-Aware CLAM also improved over the baseline on validation AUROC, reaching 0.821 for ER, 0.737 for PR, and 0.622 for HER2. These validation results suggested that attention-based multiple-instance learning could detect some receptor-associated morphology from H&E features.

However, these validation gains did not consistently transfer to the held-out test set. Generally, across all models and prediction tasks, there is a noticeable drop in performance between the validation and test evaluations. For ER, CLAM-SB achieved the highest test AUROC among the three models at 0.590, compared with 0.410 for mean-pool + logistic regression and 0.540 for Tumor-Aware CLAM. For PR, CLAM-SB and mean-pool + logistic regression were tied at a test AUROC of 0.581, while Tumor-Aware CLAM reached 0.556. For HER2, mean-pool + logistic regression and Tumor-Aware CLAM both reached a test AUROC of 0.621, while CLAM-SB reached 0.579.

Overall, the single-split ablation suggested possible morphology-based signal, especially for ER, but also revealed substantial validation-test instability. This instabil-

ity is expected given the small held-out test cohorts after target-specific missing labels are removed: 30 ER-labeled test patients, 29 PR-labeled test patients, and 24 HER2-labeled test patients. We therefore treat this experiment as exploratory rather than as the main estimate of generalization. The validation-test gap motivated the larger patient-level 5-fold cross-validation experiments across ResNet-50, UNI, and CONCH encoders.

5.6. Cross-Validated Results

Next, we performed patient-level 5-fold cross-validation experiments across multiple feature encoders: ResNet-50, UNI, and CONCH. We hypothesized that pathology-specific foundation model embeddings from UNI and CONCH would improve weakly supervised whole-slide biomarker prediction compared with ImageNet-pretrained ResNet-50 features. For each target, every patient was evaluated exactly once in an out-of-fold test fold. We report the mean and standard deviation of AUROC across folds, pooled out-of-fold AUROC and AUPRC, and bootstrap 95% confidence intervals computed over patients. This is reported in Figure 3 and Table 1.

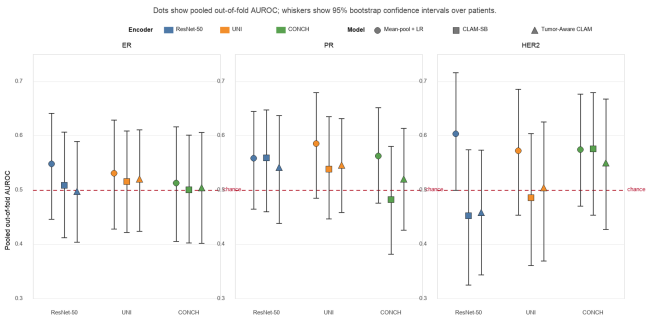

Figure 3. 5-Fold CV AUROC Summary Plot

Overall, cross-validated performance was modest across targets, encoders, and aggregation models. ER status was close to chance for nearly all settings. The best pooled ER AUROC was obtained by ResNet-50 mean-pool + logistic regression at 0.548, with a 95% CI of [0.447, 0.642]. UNI and CONCH did not improve ER prediction: their best pooled ER AUROCs were 0.531 and 0.513, respectively. The confidence intervals for ER broadly overlapped the chance line and each other, suggesting that none of the encoder-model combinations provided reliable ER discrimination under cross-validation.

PR status showed slightly stronger but still weak signal. The best pooled PR AUROC was achieved by UNI mean-pool + logistic regression at 0.586, with a 95% CI of [0.485, 0.679]. ResNet-50 CLAM-SB reached a similar pooled AUROC of 0.559, while CONCH mean-pool + logistic regression reached 0.563. However, the confidence intervals remained wide and overlapping, so these point-estimate differences should not be interpreted as statistically meaningful superiority of one encoder or model.

HER2 status showed the strongest point estimates, but uncertainty remained large. ResNet-50 mean-pool + logistic regression achieved the highest pooled HER2 AUROC at 0.603, with a 95% CI of [0.500, 0.717]. UNI mean-pool reached 0.572, while CONCH CLAM-SB reached 0.576. Although these results suggest possible HER2-associated morphology, the confidence intervals are broad, and the attention-based models did not consistently outperform the mean-pool baseline.

These cross-validated results substantially weaken the apparent gains observed in the single-split ablation. In particular, pathology foundation encoders did not clearly outperform ResNet-50, and neither CLAM-SB nor Tumor-Aware CLAM consistently improved over mean-pool logistic regression. The more rigorous cross-validated evaluation therefore suggests that receptor-status prediction from H&E morphology is challenging in the current cohort, and that the single-split improvements were not robust under patient-level cross-validation.

AUPRC values, shown in Table 1, should also be interpreted in the context of class prevalence. ER and PR have high positive-class prevalence in the labeled cohort, so their AUPRC values can appear relatively high even when AUROC is near chance. HER2 has lower positive-class prevalence, making its AUPRC values lower in absolute terms. The AUPRC numbers thus are more so a reflection of the distribution of our dataset as opposed to our model's ability to classify slides as ER or PR status.

Overall, while single-split evaluation had previously suggested attention-based MIL improved biomarker prediction, here, our cross-validated results reveal high variance and limited generalization on our dataset even. This motivated our next analysis, which evaluated whether alternative tumor-aware gate designs improve over the default tumor-aware CLAM design.

5.7. Tumor-Aware Gate Variant Ablation

Section 5.6 demonstrates that the original multiplicative tumor-aware gate was weak and unstable. We therefore evaluate three redesigned gate variants (see section 3.3) against the origional multiplicative gate under the same 5-fold cross-validation protocol across all three encoders and targets. Table 2 reports pooled out-of-fold AUROC with 95% confidence intervals.

When the best performing redesigned variant is chosen separately within each encoder-target combination, the residual-family designs outperform the multiplicative gat in 7 of the 9 combinations. This is a post-hoc best-of-variants comparison, thus this should not be interpreted as the behavior of a single pre-specified gate; no individual variant improves in more than 5 of 9 combinations. Residual and residual+regularization variants each improve in 5 of 9 combinations while residual+regularization+temperative improves in 3 of 9. The two exceptions are small: PR/ResNet-50 decreases from 0.541 to 0.525 and HER2/UNI decreases from 0.504 to

0.489.

Several improvements are substantial: with residual+regularization on ResNet-50 features, HER2 AUROC rises from 0.458 (multiplicative gate) to 0.626 while ER rises from 0.497 to 0.566. The best single MIL result in this experiment is HER2/ResNet-50 with residual+regularization (achieves pooled AUROC of 0.626 with 95% CI of [0.51, 0.73]). This is the only gated configuration to exceed its corresponding mean-pool baseline of 0.603 for HER2.

For PR/ResNet-50, the redesign also reduces the fold-to-fold variability that first motivated it. The multiplicative gate has an AUROC standard deviation of $\alpha = 0.120$ compared to $\alpha = 0.069$ for the residual gate and $\alpha = 0.074$ for residual+regularization. This indicates more consistent performance across folds in this setting. Adding a learnable termperature in addition to the residual+regularization generally did not improve pooled performance, having performed worse in 7 or 9 combinations, matching performance in 1, and improving by 0.001 in another. Thus, we conclude the added complexity did not provide any benefit on this cohort.

Overall, the residual designs partially mitigate the degradation and instability of the original multiplicative formulation and make tumor-aware MIL more competitive with the mean-pool baseline. However, it does not produce a decisive or consistent improvement over it. The best redesigned variant exceeds the corresponding mean-pool baseline in only 4 or 9 encoder-target combinations. For PR, mean-pool with logistic regression remains the strongest for UNI and CONCH and ties with CLAM-SB for ResNet-50 (pooled AUROS of 0.5587 and 0.5595, respectively). Moreover, the marginal confidence intervals for the gains overlap and thus do not independently establish any improvement. Combining these results with those in Section 8.2, we see a consistency in the limited cohort size being an important constraint.

5.8. Attention Heatmaps

Figures 5 and 6 show CLAM-SB attention heatmaps for two ER-positive slides. We chose these slides to compare a success and failure case. In Figure 5 (TCGA-A1-A0SF), attention is very concentrated in the middle tissue region. This is consistent with an invasive tumor nest that has a dense cellular mass in its center. This behavior aligns with the intended function of the attention mechanism.

Figure 6 (TCGA-C8-A1HM) illustrates a failure in the attention mechanism. Despite the slide containing a large continuous tissue mass, the model assigns its highest attention weights almost entirely to the boundary outline of the tissue rather than the center of the mass. The interior patches that are densely stained and likely contain invasive tumor cells are receiving little attention. This pattern suggests the model may be responding to edge-of-tissue texture at the tumor margin rather than nuclear or glandular features of the tumor parenchyma itself. This is exactly the confounding that the Tumor-Aware gating mechanism is designed to reduce by down-weighting patches with low estimated tumor probability to redirect the attention module away from non-tumor regions.

6. Conclusion

We investigated whether we could use weakly supervised whole-slide deep learning to predict the status of three breast cancer biomarkers (ER, PR, and HER2) from histopathology images stained with a Hematoxylin & Eosin stain. We compared three different patch encoders (ResNet-50, UNI, and CONCH) that had varying levels of pathology specificity to see whether pathology-specific encoding is important for solving this classification problem. We evaluated mean-pooling, CLAM-SB, and our proposed Tumor-Aware CLAM architecture as different slide-level aggregation strategies.

On individual train-test splits, all three slide-level aggregation strategies had a general improvement over the mean-pool baseline for ER status. CLAM-SB and Tumor-Aware CLAM had significant improvements over baseline for PR status, while mean-pool + LR had no increase in performance for PR. None of the aggregation strategies outperformed baseline for HER2 status.

Despite the promising results for ER and PR status on individual train-test splits, we did not see improvements under patient-level 5-fold cross-validation. For all three biomarker status prediction tasks, pooled AUROCs had values near random 0.50 chance prediciton, and the confidence intervals overlapped. We found that pathology foundation models did not consistently outperform conventional ImageNet-pretrained features, and attention-based MIL methods did not consistently improve upon a simple mean-pooling baseline. Attention heatmaps suggest that our results in in the train-test splits were a result of data partitions and were not related to biomarker-related morpholpgy.

Future work could look into sources for larger datasets to confirm whether or not our inconsistent results were a result of imbalanced data partitians and bias in the distribution of biomarker labels. With a larger cohort, future researchers could test out more sophisticated spatial modeling approaches such as transformer-based slide options that look at the interactions between image patches. Future work could also look into self-supervised adaption to breast cancer tissue specifically or auxilliary supervision adaptation. Overall, our results show examples of the limitations of working with small datasets and limited compute, and it would be interestinfg to see how future research could mitigate those limitations.

7. Appendix

Many software and code packages were used in our analysis [32, 38, 37, 36, 35, 14, 9, 12, 26, 48, 33, 42, 41, 43, 8, 30, 34, 11, 47, 40, 27, 16, 52, 51, 17, 50, 39, 19, 23, 5, 21, 25, 31, 46, **?**, 29, 3, 2].

Table 1. **Cross-Validated Receptor-Status Prediction Results**

Target	Encoder	Model	Fold AUROC mean +/- SD	Pooled AUROC [95% CI]	Pooled AUPRC [95% CI]
ER	ResNet-50	Mean-pool + LR	0.554 ± 0.065	**0.548 [0.447, 0.642]**	0.786 [0.707, 0.859]
ER	ResNet-50	CLAM-SB	0.512 ± 0.066	0.509 [0.412, 0.607]	0.762 [0.679, 0.848]
ER	ResNet-50	Tumor-Aware CLAM	0.499 ± 0.050	0.497 [0.404, 0.590]	0.759 [0.682, 0.844]
ER	UNI	Mean-pool + LR	0.532 ± 0.070	0.531 [0.428, 0.629]	0.768 [0.690, 0.848]
ER	UNI	CLAM-SB	0.523 ± 0.050	0.516 [0.422, 0.609]	0.773 [0.694, 0.851]
ER	UNI	Tumor-Aware CLAM	0.508 ± 0.065	0.519 [0.424, 0.611]	0.764 [0.685, 0.847]
ER	CONCH	Mean-pool + LR	0.507 ± 0.055	0.513 [0.406, 0.617]	0.733 [0.656, 0.829]
ER	CONCH	CLAM-SB	0.518 ± 0.059	0.501 [0.403, 0.601]	0.745 [0.663, 0.829]
ER	CONCH	Tumor-Aware CLAM	0.522 ± 0.070	0.504 [0.402, 0.606]	0.732 [0.649, 0.825]
PR	ResNet-50	Mean-pool + LR	0.565 ± 0.069	0.559 [0.465, 0.645]	0.745 [0.654, 0.830]
PR	ResNet-50	CLAM-SB	0.574 ± 0.099	0.559 [0.461, 0.648]	0.732 [0.653, 0.821]
PR	ResNet-50	Tumor-Aware CLAM	0.572 ± 0.120	0.541 [0.439, 0.638]	0.719 [0.635, 0.809]
PR	UNI	Mean-pool + LR	0.583 ± 0.079	**0.586 [0.485, 0.679]**	0.738 [0.654, 0.824]
PR	UNI	CLAM-SB	0.542 ± 0.041	0.539 [0.447, 0.636]	0.716 [0.636, 0.808]
PR	UNI	Tumor-Aware CLAM	0.562 ± 0.053	0.546 [0.459, 0.632]	0.731 [0.650, 0.824]
PR	CONCH	Mean-pool + LR	0.567 ± 0.062	0.563 [0.476, 0.652]	0.749 [0.669, 0.837]
PR	CONCH	CLAM-SB	0.502 ± 0.058	0.482 [0.382, 0.581]	0.673 [0.594, 0.774]
PR	CONCH	Tumor-Aware CLAM	0.523 ± 0.045	0.520 [0.427, 0.614]	0.701 [0.621, 0.793]
HER2	ResNet-50	Mean-pool + LR	0.623 ± 0.120	**0.603 [0.500, 0.717]**	0.316 [0.216, 0.489]
HER2	ResNet-50	CLAM-SB	0.473 ± 0.102	0.453 [0.325, 0.574]	0.219 [0.149, 0.338]
HER2	ResNet-50	Tumor-Aware CLAM	0.509 ± 0.088	0.458 [0.344, 0.574]	0.220 [0.150, 0.344]
HER2	UNI	Mean-pool + LR	0.560 ± 0.047	0.572 [0.454, 0.686]	0.304 [0.203, 0.477]
HER2	UNI	CLAM-SB	0.511 ± 0.108	0.486 [0.361, 0.604]	0.262 [0.169, 0.413]
HER2	UNI	Tumor-Aware CLAM	0.497 ± 0.155	0.504 [0.369, 0.626]	0.299 [0.185, 0.467]
HER2	CONCH	Mean-pool + LR	0.567 ± 0.057	0.575 [0.470, 0.677]	0.263 [0.185, 0.404]
HER2	CONCH	CLAM-SB	0.588 ± 0.167	0.576 [0.454, 0.680]	0.293 [0.189, 0.435]
HER2	CONCH	Tumor-Aware CLAM	0.575 ± 0.143	0.549 [0.427, 0.668]	0.324 [0.205, 0.482]

Patient-level 5-fold CV across ResNet-50, UNI, and CONCH.
Bold pooled AUROC marks the highest point estimate within each target;
overlapping CIs mean these should not be read as statistically distinct.

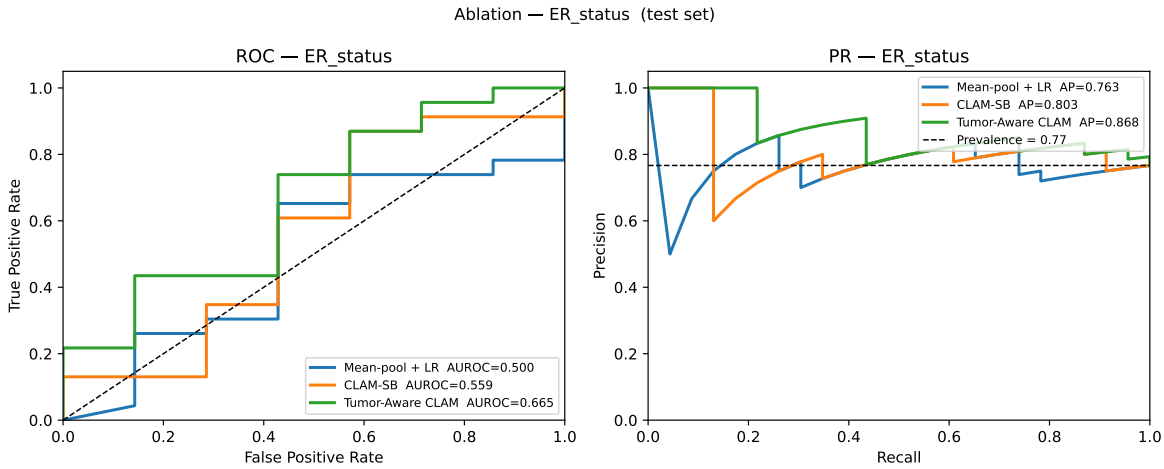

Figure 4. ROC + PR Curves, UNI Encoder, ER Status, Test Split

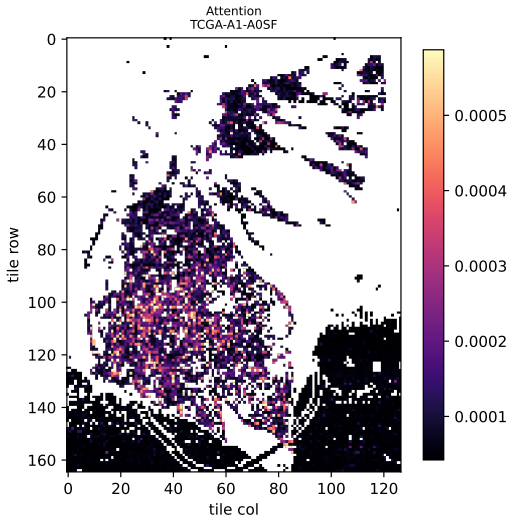

Figure 5. TCGA-A1-A0SF ER Status Attention Heatmap

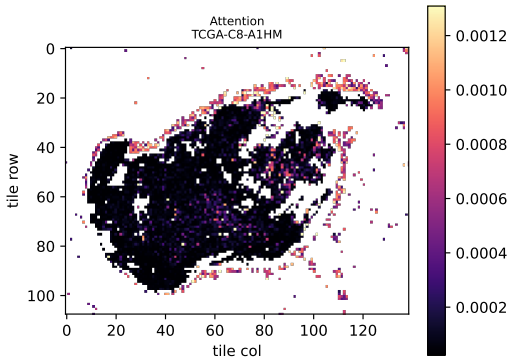

Figure 6. TCGA-C8-A1HM ER Status Attention Heatmap

Table 2. Tumor-Aware Gate Ablation under 5-fold Patient-level Cross-Validation. Each cell reports pooled out-of-fold AUROC with a 95% bootstrap confidence interval. Mean-pool + LR and CLAM-SB are non-gated references; the four gate rows are the ablation.

Target	Configuration	ResNet-50	UNI	CONCH
ER	Mean-pool + LR (ref.)	0.548 [0.45, 0.64]	0.531 [0.43, 0.63]	0.513 [0.41, 0.62]
ER	CLAM-SB (ref.)	0.509 [0.41, 0.61]	0.516 [0.42, 0.61]	0.501 [0.40, 0.60]
ER	Tumor-Aware (mult)	0.497 [0.40, 0.59]	0.519 [0.42, 0.61]	0.504 [0.40, 0.61]
ER	+ residual	0.554 [0.45, 0.65]	0.515 [0.42, 0.61]	0.518 [0.41, 0.61]
ER	+ residual+reg	**0.566** [0.47, 0.66]	**0.533** [0.44, 0.62]	**0.529** [0.43, 0.62]
ER	+ residual+reg+temp	0.529 [0.43, 0.62]	0.516 [0.41, 0.61]	0.529 [0.43, 0.62]
PR	Mean-pool + LR (ref.)	**0.559** [0.46, 0.64]	**0.586** [0.49, 0.68]	**0.563** [0.48, 0.65]
PR	CLAM-SB (ref.)	0.559 [0.46, 0.65]	0.539 [0.45, 0.64]	0.482 [0.38, 0.58]
PR	Tumor-Aware (mult)	0.541 [0.44, 0.64]	0.546 [0.46, 0.63]	0.520 [0.43, 0.61]
PR	+ residual	0.525 [0.44, 0.62]	0.575 [0.48, 0.66]	0.523 [0.42, 0.62]
PR	+ residual+reg	0.522 [0.43, 0.61]	0.522 [0.43, 0.62]	0.512 [0.42, 0.61]
PR	+ residual+reg+temp	0.523 [0.43, 0.62]	0.501 [0.41, 0.60]	0.496 [0.40, 0.60]
HER2	Mean-pool + LR (ref.)	0.603 [0.50, 0.72]	**0.572** [0.45, 0.69]	0.575 [0.47, 0.68]
HER2	CLAM-SB (ref.)	0.453 [0.33, 0.57]	0.486 [0.36, 0.60]	**0.576** [0.45, 0.68]
HER2	Tumor-Aware (mult)	0.458 [0.34, 0.57]	0.504 [0.37, 0.63]	0.549 [0.43, 0.67]
HER2	+ residual	0.530 [0.41, 0.64]	0.427 [0.31, 0.54]	0.541 [0.43, 0.65]
HER2	+ residual+reg	**0.626** [0.51, 0.73]	0.489 [0.36, 0.61]	0.570 [0.46, 0.69]
HER2	+ residual+reg+temp	0.581 [0.47, 0.69]	0.466 [0.35, 0.58]	0.483 [0.37, 0.59]

Table 3. Ablation Study Summary for Single-Split Test Across All Encoders

Encoder	Target	Split	Model	AUROC	AUPRC	Balanced Acc	Brier Score
conch	ER_status	test	CLAM-SB	0.5404	0.8156	0.5000	0.1838
resnet50	ER_status	test	CLAM-SB	0.5901	0.8274	0.6491	0.2029
uni	ER_status	test	CLAM-SB	0.5590	0.8113	0.5280	0.2286
conch	ER_status	test	Mean-pool + LR	0.3540	0.7301	0.4255	0.3695
resnet50	ER_status	test	Mean-pool + LR	0.4099	0.7532	0.4907	0.3593
uni	ER_status	test	Mean-pool + LR	0.5000	0.7649	0.5839	0.3124
conch	ER_status	test	Tumor-Aware CLAM	0.5466	0.8083	0.5000	0.1889
resnet50	ER_status	test	Tumor-Aware CLAM	0.5404	0.7993	0.4783	0.1881
uni	ER_status	test	Tumor-Aware CLAM	0.6646	0.8726	0.6273	0.2091
conch	ER_status	val	CLAM-SB	0.6643	0.8592	0.5000	0.1817
resnet50	ER_status	val	CLAM-SB	0.8500	0.9528	0.5929	0.1514
uni	ER_status	val	CLAM-SB	0.7929	0.9277	0.6429	0.1953
conch	ER_status	val	Mean-pool + LR	0.5357	0.7850	0.5893	0.2764
resnet50	ER_status	val	Mean-pool + LR	0.7786	0.9119	0.7571	0.1871
uni	ER_status	val	Mean-pool + LR	0.6500	0.8405	0.6143	0.2248
conch	ER_status	val	Tumor-Aware CLAM	0.6357	0.8471	0.5000	0.1830
resnet50	ER_status	val	Tumor-Aware CLAM	0.8214	0.9431	0.5000	0.1708
uni	ER_status	val	Tumor-Aware CLAM	0.7571	0.8990	0.6179	0.1661
conch	HER2_status	test	CLAM-SB	0.5684	0.3022	0.4421	0.2428
resnet50	HER2_status	test	CLAM-SB	0.5789	0.3064	0.4684	0.2829
uni	HER2_status	test	CLAM-SB	0.4368	0.2038	0.5211	0.2307
conch	HER2_status	test	Mean-pool + LR	0.6263	0.3162	0.5632	0.2458
resnet50	HER2_status	test	Mean-pool + LR	0.5684	0.2403	0.5737	0.2281
uni	HER2_status	test	Mean-pool + LR	0.6053	0.2769	0.5263	0.2312
conch	HER2_status	test	Tumor-Aware CLAM	0.5000	0.2317	0.5000	0.2104
resnet50	HER2_status	test	Tumor-Aware CLAM	0.5158	0.2494	0.5105	0.2091
uni	HER2_status	test	Tumor-Aware CLAM	0.5526	0.2520	0.5000	0.2088
conch	HER2_status	val	CLAM-SB	0.5526	0.2603	0.5000	0.2131
resnet50	HER2_status	val	CLAM-SB	0.5000	0.2335	0.5000	0.2163
uni	HER2_status	val	CLAM-SB	0.5658	0.2709	0.5132	0.2238
conch	HER2_status	val	Mean-pool + LR	0.4737	0.2185	0.4868	0.2483
resnet50	HER2_status	val	Mean-pool + LR	0.3421	0.1772	0.3750	0.2601
uni	HER2_status	val	Mean-pool + LR	0.4342	0.2078	0.4605	0.2307
conch	HER2_status	val	Tumor-Aware CLAM	0.5263	0.2435	0.5000	0.2092
resnet50	HER2_status	val	Tumor-Aware CLAM	0.5526	0.2483	0.5000	0.2061
uni	HER2_status	val	Tumor-Aware CLAM	0.5658	0.2741	0.5000	0.2107
conch	PR_status	test	CLAM-SB	0.5164	0.6384	0.5000	0.2427
resnet50	PR_status	test	CLAM-SB	0.5526	0.6483	0.5362	0.2464
uni	PR_status	test	CLAM-SB	0.5493	0.6558	0.5230	0.2606
conch	PR_status	test	Mean-pool + LR	0.5493	0.6698	0.5428	0.2678
resnet50	PR_status	test	Mean-pool + LR	0.5329	0.6366	0.5132	0.2642
uni	PR_status	test	Mean-pool + LR	0.5526	0.6401	0.5329	0.2520
conch	PR_status	test	Tumor-Aware CLAM	0.5395	0.6487	0.5000	0.2413
resnet50	PR_status	test	Tumor-Aware CLAM	0.5230	0.6125	0.5132	0.2439
uni	PR_status	test	Tumor-Aware CLAM	0.6151	0.7093	0.5757	0.2443
conch	PR_status	val	CLAM-SB	0.5789	0.7675	0.5000	0.2206
resnet50	PR_status	val	CLAM-SB	0.6447	0.8166	0.5000	0.2155
uni	PR_status	val	CLAM-SB	0.6776	0.8524	0.5000	0.2255
conch	PR_status	val	Mean-pool + LR	0.3816	0.7114	0.3783	0.3811
resnet50	PR_status	val	Mean-pool + LR	0.5855	0.7565	0.5559	0.3024
uni	PR_status	val	Mean-pool + LR	0.6842	0.8691	0.5197	0.2598
conch	PR_status	val	Tumor-Aware CLAM	0.5987	0.8080	0.5000	0.2038
resnet50	PR_status	val	Tumor-Aware CLAM	0.7368	0.8738	0.6711	0.1996
uni	PR_status	val	Tumor-Aware CLAM	0.6711	0.8542	0.5000	0.2075

8. Contributors

This project and paper is authored by Jennifer Ho, Summer Royal, and Luke Zhao. Fangrui Huang was our TA advisor. There was no involvement of any non-CS 231N contributors. Our project has not been submitted to a peer-reviewed conference or journal. We did not use this project for any other class.

8.1. Summer's Contributions

Report: Summer wrote the Abstract section, Introduction section, last two paragraphs of Related Work, parts of the Methods section, "Why This Approach?" subsection, Data section, parts of the Experiment/Results section, and the Conclusion section.
Code: built the entire WSI preprocessing pipeline from scratch, downloaded dataset and set up data split infastructure, made scripts for feature extraction, label preparation, modal pipeline, baseline testing, smoke tests, evaluation, figure generation, and plot ablation.

8.2. Luke's Contributions

Report: Wrote most of the Related Works section, parts of Methods, and parts of the Experiment section.
Code: Wrote the encoder loading code and ran the CONCH encoder experiment.

8.3. Jen's Contributions

9. AI Use

- We did all of the project ideation ourselves without using AI. We came up with the project idea and design decisions without consulting AI.
- We wrote this report ourselves without using AI.
- We used Cursor, Codex, and Claude Code to help with the process of writing code. A more detailed description of our AI usage can be found at this linked document.

References

[1] G. Campanella, M. G. Hanna, L. Geneslaw, A. Miraflor, V. W. K. Silva, K. J. Busam, E. Brogi, V. E. Reuter, D. S. Klimstra, and T. J. Fuchs. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nature Medicine*, 25(8):1301–1309, 2019.

[2] cBioPortal. cbioportal web api and clients, 2026. Accessed June 4, 2026.

[3] E. Cerami, J. Gao, U. Dogrusoz, B. E. Gross, S. O. Sumer, B. A. Aksoy, A. Jacobsen, C. J. Byrne, M. L. Heuer, E. Larsson, et al. The cbio cancer genomics portal: An open platform for exploring multidimensional cancer genomics data. *Cancer Discovery*, 2(5):401–404, 2012.

[4] R. J. Chen, C. Chen, Y. Li, T. Y. Chen, A. D. Trister, R. G. Krishnan, and F. Mahmood. Scaling vision transformers to gigapixel images via hierarchical self-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16144–16155, 2022.

[5] R. J. Chen, T. Ding, M. Y. Lu, D. F. K. Williamson, G. Jaume, A. H. Song, B. Chen, A. Zhang, D. Shao, M. Shaban, et al. Towards a general-purpose foundation model for computational pathology. *Nature Medicine*, 30(3):850–862, 2024.

[6] E. Conde-Sousa, J. Vale, M. Feng, K. Xu, Y. Wang, V. Della Mea, D. La Barbera, E. Montahaei, M. Soleymani Baghshah, A. Turzynski, J. Gildenblat, E. Klaiman, Y. Hong, G. Aresta, T. Ara'ujo, P. Aguiar, C. Eloy, and A. Pol'onia. HEROHE challenge: Predicting HER2 status in breast cancer from hematoxylin-eosin whole slide imaging. *Journal of Imaging*, 8(8):213, 2022.

[7] H. D. Couture, L. A. Williams, J. Geradts, S. J. Nyante, E. N. Butler, J. S. Marron, C. M. Perou, M. A. Troester, and M. Niethammer. Image analysis with deep learning to predict breast cancer grade, ER status, histologic subtype, and intrinsic subtype. *npj Breast Cancer*, 4:30, 2018.

[8] C. Davidson-Pilon. lifelines: Survival analysis in python. *Journal of Open Source Software*, 4(40):1317, 2019.

[9] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[10] P. Gamble, R. Jaroensri, H. Wang, F. Tan, M. Moran, T. Brown, I. Flament-Auvigne, E. A. Rakha, M. Toss, D. J. Dabbs, et al. Determining breast cancer biomarker status and associated morphological features using deep learning. *Communications Medicine*, 1:14, 2021.

[11] h5py Developers. h5py documentation, 2026. Accessed June 4, 2026.

[12] C. R. Harris, K. J. Millman, S. J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N. J. Smith, et al. Array programming with numpy. *Nature*, 585(7825):357–362, 2020.

[13] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[14] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. pages 770–778, 2016.

[15] A. P. Heath, V. Ferretti, S. Agrawal, M. An, J. C. Angelakos, R. Arya, R. Bajari, M. Bobohalm, D. Brooks, M. Brush, et al. The NCI genomic data commons. *Nature Genetics*, 53(3):257–262, 2021.

[16] Hugging Face. Hugging face hub documentation, 2026. Accessed June 4, 2026.

[17] J. D. Hunter. Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, 9(3):90–95, 2007.

[18] M. Ilse, J. M. Tomczak, and M. Welling. Attention-based deep multiple instance learning. *Proceedings of the 35th International Conference on Machine Learning*, 80:2127–2136, 2018.

[19] M. Ilse, J. M. Tomczak, and M. Welling. Attention-based deep multiple instance learning. In *Proceedings of the 35th International Conference on Machine Learning*, pages 2127–2136. PMLR, 2018.

[20] B. Li, Y. Li, and K. W. Eliceiri. Dual-stream multiple instance learning network for whole slide image classification

with self-supervised contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14318–14328, 2021.

[21] M. Y. Lu, B. Chen, D. F. K. Williamson, R. J. Chen, I. Liang, T. Ding, G. Jaume, I. Odintsov, L. P. Le, G. Gerber, et al. A visual-language foundation model for computational pathology. *Nature Medicine*, 30(3):863–874, 2024.

[22] M. Y. Lu, D. F. K. Williamson, T. Y. Chen, R. J. Chen, M. Barbieri, and F. Mahmood. CLAM: Data-efficient and weakly supervised computational pathology on whole slide images, 2021. GitHub repository.

[23] 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*, 5:555–570, 2021.

[24] 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*, 5:555–570, 2021.

[25] M. Macenko, M. Niethammer, J. S. Marron, D. Borland, J. T. Woosley, X. Guan, C. Schmitt, and N. E. Thomas. A method for normalizing histology slides for quantitative analysis. *Proceedings of the IEEE International Symposium on Biomedical Imaging*, pages 1107–1110, 2009.

[26] W. McKinney. Data structures for statistical computing in python. In *Proceedings of the 9th Python in Science Conference*, pages 56–61, 2010.

[27] Modal Labs. Modal documentation, 2026. Accessed June 4, 2026.

[28] N. Naik, A. Madani, A. Esteva, N. S. Keskar, M. F. Press, D. Ruderman, D. B. Agus, and R. Socher. Deep learning-enabled breast cancer hormonal receptor status determination from base-level he stains. *Nature Communications*, 11:5727, 2020.

[29] National Cancer Institute Genomic Data Commons. Genomic data commons: Access data, 2026. Accessed June 4, 2026.

[30] OpenSlide Project. Openslide python documentation, 2026. Accessed June 4, 2026.

[31] N. Otsu. A threshold selection method from gray-level histograms. *IEEE Transactions on Systems, Man, and Cybernetics*, 9(1):62–66, 1979.

[32] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*, 32, 2019.

[33] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[34] Pillow Contributors. Pillow documentation, 2026. Accessed June 4, 2026.

[35] PyTorch Foundation. torch.nn.crossentropyloss documentation, 2026. Accessed June 4, 2026.

[36] PyTorch Foundation. torch.optim.adam documentation, 2026. Accessed June 4, 2026.

[37] PyTorch Foundation. torch.utils.data dataloader documentation, 2026. Accessed June 4, 2026.

[38] PyTorch Foundation. Torchvision models: Resnet-50, 2026. Accessed June 4, 2026.

[39] ReportLab. Reportlab documentation, 2026. Accessed June 4, 2026.

[40] Requests Developers. Requests documentation, 2026. Accessed June 4, 2026.

[41] scikit-learn developers. scikit-learn metrics documentation, 2026. Accessed June 4, 2026.

[42] scikit-learn developers. sklearn.linear_model.logisticregression documentation, 2026. Accessed June 4, 2026.

[43] scikit-learn developers. sklearn.model_selection.train_test_split documentation, 2026. Accessed June 4, 2026.

[44] G. Shamai, R. Schley, A. Cretu, T. Neoran, E. Sabo, Y. Binenbaum, S. Cohen, T. Goldman, A. Pol'onia, K. Drumea, K. Stoliar, and R. Kimmel. Clinical utility of receptor status prediction in breast cancer and misdiagnosis identification using deep learning on hematoxylin and eosin-stained slides. *Communications Medicine*, 4:276, 2024.

[45] Z. Shao, H. Bian, Y. Chen, Y. Wang, J. Zhang, X. Ji, and Y. Zhang. TransMIL: Transformer based correlated multiple instance learning for whole slide image classification. In *Advances in Neural Information Processing Systems*, volume 34, pages 2136–2147, 2021.

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

[47] tqdm Developers. tqdm documentation, 2026. Accessed June 4, 2026.

[48] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, et al. Scipy 1.0: Fundamental algorithms for scientific computing in python. *Nature Methods*, 17(3):261–272, 2020.

[49] X. Wang, S. Yang, J. Zhang, M. Wang, J. Zhang, J. Huang, W. Yang, and X. Han. Transformer-based unsupervised contrastive learning for histopathological image classification. *Medical Image Analysis*, 81:102559, 2022.

[50] M. L. Waskom. Seaborn: Statistical data visualization. *Journal of Open Source Software*, 6(60):3021, 2021.

[51] Weights & Biases. Weights & biases experiment tracking documentation, 2026. Accessed June 4, 2026.

[52] R. Wightman. Pytorch image models, 2026. Accessed June 4, 2026.