Momentum Memory for Knowledge Distillation in Computational Pathology

Yongxin Guo[1]*, Hao Lu[1], Onur C. Koyun[1], Zhengjie Zhu[1], Muhammet F. Demir[1], Metin N. Gurcan[1]
[1]Wake Forest University School of Medicine, Winston-Salem, NC, USA
{Hao.Lu, Onur.Koyun, muhammet.demir}@advocatehealth.org
{Yongxin.Guo, Zhengjie.Zhu, Metin.Gurcan}@wfusm.edu

In this supplementary materials, we will include details of datasets and implementation, as well as more experimental results.

S1. Datasets

We utilized two distinct datasets for this study: The Cancer Genome Atlas Breast Invasive Carcinoma (TCGA-BRCA) and an internal cohort.

TCGA-BRCA Dataset.

Original Whole Slide Images (WSIs) in `.svs` format were downloaded from the official TCGA data portal. Our analysis was guided by the official clinical annotations for three tasks: Human Epidermal Growth Factor Receptor 2 (HER2) status, Progesterone Receptor (PR) status, and Oncotype DX (ODX) recurrence score. For the HER2 and PR tasks, we focused on the binary classification of positive versus negative status.

For the ODX risk stratification task, we used research-based ODX scores derived from normalized mRNA expression data, as calculated by [5]. The initial cohort of 1,133 WSIs underwent rigorous quality control, excluding cases with incomplete receptor status, missing ODX scores, gene profiles or processing failures. This resulted in a final analytical cohort of 997 WSIs. Within the hormone receptor-positive/HER2-negative (HR+/HER2-) subgroup ($n = 516$), patients were categorized into low-risk ($n = 443$) and high-risk ($n = 73$) groups. The normalized ODX scores for this dataset range from -2.009 to 2.744, with a risk threshold of 0.7169.

To ensure robust evaluation, patient-level stratification was employed to maintain a strict separation between training and testing cohorts, thereby preventing data leakage. To mitigate class imbalance issues, non-HR+/HER2- cases were also included in the training set.

*Corresponding author: Yongxin.Guo@wfusm.edu.
Code: https://github.com/CAIR-LAB-WFUSM/MoMKD

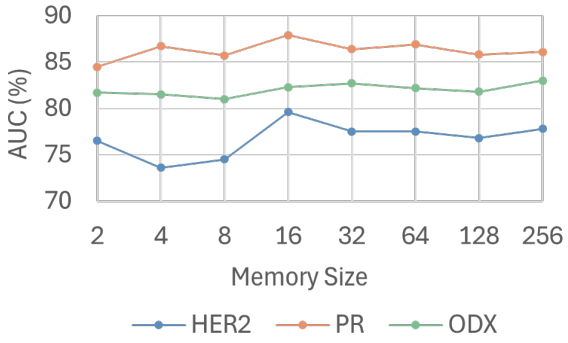

Figure S1. The variation of memory size within three tasks.

The In-House Dataset.

This independent institutional cohort comprises 1,123 H&E stained WSIs from HR+/HER2- breast cancer specimens. For this dataset, the ODX scores range from 0 to 100, with a clinical threshold of 25 used to differentiate low-risk ($n = 961$) from high-risk ($n = 162$) cases.

S2. Implementation Details

For WSI processing, we adopted the pipeline proposed by Trident [14]. Each WSI was initially segmented into non-overlapping tiles of 768×768 pixels. Features were then extracted from these tiles using the UNI v2 foundation model [2].

The original omics data presented a high-dimensional feature space with $D = 19085$. To address this, we employed an XGBoost model for feature selection, identifying the top-k genes, where k was set to 768. The resulting data was subsequently normalized using the z-score method.

Our model was implemented in Python with the PyTorch library and trained on a single NVIDIA H-100 GPU. We utilized the AdamW optimizer with a learning rate of 2×10^{-4} and a weight decay of 1×10^{-4}. The training was conducted for up to 80 epochs, incorporating an early stopping

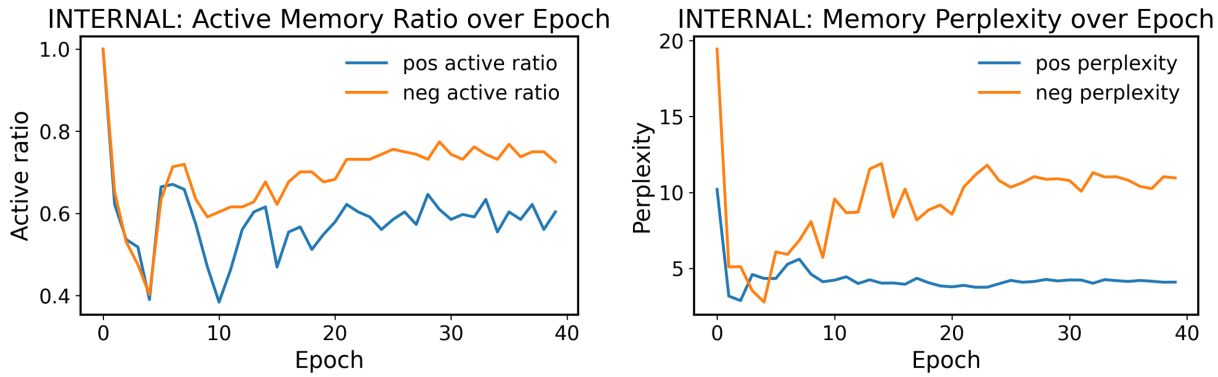

Figure S2. Memory dynamics on the in-house dataset. (Left) active memory ratio, and (right) perplexity evolution over training epochs.

mechanism to select the best-performing checkpoint based on validation set accuracy. The batch size used is 1. For all tasks, the memory size was set to $n = 16$. To stabilize training, model parameters were updated using a gradient accumulation strategy over 16 steps.

S3. Memory Dynamics on the Internal ODX Dataset

In this section, we introduce more details on the memory analysis as well as its dynamics on the in-house dataset which shown in Fig. S2.

Active Memory Ratio.

$$r_{\text{active}} = \frac{n_{\text{active}}}{n}, \tag{S1}$$

measures the proportion of memory mode entries updated within each epoch. Values near 1.0 indicate broad participation, whereas smaller ratios reflect selective activation. In our experiments, $r_{\text{active}} > 0.75$ throughout training, confirming the glocal activation.

Perplexity.

$$\text{Perplexity} = \exp(H(p)) \in [1, n], \tag{S2}$$

represents the effective number of active memory modes. Perplexity values in different tasks reflect the dynamics learning process on the proposed method.

Clinical understanding and analysis The feasibility of inferring molecular biomarkers directly from H&E stained whole slide image varies substantially across targets, reflecting the extent to which each biomarker manifests morphologically observable correlates. Among routinely assessed breast cancer biomarkers, HER2, progesterone receptor (PR), and Oncotype DX (ODX) recurrence scores represent three characteristic levels of morphological–molecular coupling.

HER2 (ERBB2 amplification and protein overexpression) is generally the most challenging target for image-based prediction [1, 8]. Its clinical definition relies on membranous protein overexpression and gene copy-number amplification assessed by immunohistochemistry (IHC) or in-situ hybridization (ISH), features that are not directly visible on standard H&E slides [6, 12]. Morphologic surrogates such as nuclear atypia, mitotic rate, or growth pattern provide only weak and indirect cues. Furthermore, even IHC-based HER2 scoring suffers from inter-observer variability , emphasizing its intrinsic diagnostic complexity.

In contrast, PR status tends to exhibit stronger alignment with histomorphological appearance. Hormone-receptor–positive tumors frequently present as low-grade, well-differentiated lesions with organized glandular architecture and lower mitotic activity that attributes readily captured in H&E morphology [3].

Finally, the Oncotype DX recurrence score, a multigene assay quantifying proliferation and differentiation-related transcripts which shows the closest association with H&E-derived phenotypes. Multiple clinical studies have demonstrated strong correlations between ODX and conventional histologic variables such as tumor grade, nuclear pleomorphism, and mitotic index [7, 11, 13]. Deep learning approaches leveraging WSIs and minimal clinical data have achieved concordance comparable to molecular testing (AUC $\approx$ 0.80–0.85 for high- vs low-risk classifica-

Algorithm 1 Momentum Memory Knowledge Distillation (MoMKD)

Require: WSI spatial graph G, Omics vector O, Ground truth label $Y \in \{0, 1\}$
Require: Momentum memory banks C^+ (positive) and C^- (negative)

// 1. Dual-Branch Modality Encoding
$F_{wsi} \leftarrow \text{WsiEncoder}(G)$ ▷ Extract patch-level WSI representations
$F_{omics} \leftarrow \text{OmicsEncoder}(O)$ ▷ Extract global omics representation

// 2. Memory-Guided Distillation
$Score \leftarrow \text{ComputeAttention}(F_{wsi}, \text{Detach}(C^+), \text{Detach}(C^-))$ ▷ Query memory for patch importance
$F_{slide} \leftarrow \text{Aggregate}(F_{wsi}, Score)$ ▷ Obtain slide-level WSI representation
$\hat{Y} \leftarrow \text{Classifier}(F_{slide})$ ▷ Generate diagnostic prediction

// 3. Indirect Cross-Modal Alignment
// Both modalities are aligned to the shared memory rather than to each other directly
$L_{align}^{wsi} \leftarrow \text{AlignmentLoss}(F_{wsi}, C^+, C^-, Y)$ ▷ Align WSI to class-specific memory
$L_{align}^{omics} \leftarrow \text{AlignmentLoss}(F_{omics}, C^+, C^-, Y)$ ▷ Align Omics to class-specific memory

// 4. Omics Semantic Anchoring
$L_{recon} \leftarrow \text{ReconstructionLoss}(\text{Decoder}(F_{omics}), O)$ ▷ Preserve biological structure

// 5. Gradient-Decoupled Optimization
$L_{task} \leftarrow \text{CrossEntropy}(\hat{Y}, Y)$

// Decouple gradients to prevent modality collapse:
Update WsiEncoder and Classifier using $\nabla(L_{task} + L_{align}^{wsi})$
Update OmicsEncoder using $\nabla(L_{recon} + L_{align}^{omics})$
Update C^+, C^- using $\nabla(L_{align}^{wsi} + L_{align}^{omics} + L_{mem})$ ▷ Shielded from L_{task}

tion [4, 9]). These findings suggest that ODX captures transcriptomic programs that are largely mirrored by morphological cues observable in H&E slides.

In summary, the variable predictability of HER2, PR, and ODX from H&E images arises from the differing degrees of morphological expressivity of their underlying biology. HER2 overexpression reflects membrane-localized molecular events poorly represented in tissue architecture; PR status influences global differentiation that leaves discernible patterns; and ODX aggregates proliferation-related signals that are morphologically pronounced. Acknowledging this gradient of morphologic–molecular coupling is essential when interpreting model performance.

Based on the statistics in Fig. S2 and the clinical characteristics of each biomarker, here we provide additional insight into the development of the memory used in our framework. The memory usage both in the TCGA-BRCA dataset (Figure 3) with the in-house dataset (Figure S2) indicate the memory component activation remains consistently broad, with over 75% of memory components active across epochs. This directly validates the efficacy of the gradient-decoupled momentum update, proving that the memory maintains rich semantic diversity and successfully avoids global collapse that often plaguing dynamic dictionaries.

Interestingly, task-dependent memory usage patterns emerge. The HER2 classification task activates a greater variety of memory usage, consistent with its higher histopathology complexity and known diagnostic ambiguity on H&E stains [8]. In contrast, PR and ODX tasks exhibit more concentrated usage, suggesting that fewer memory components suffice to capture discriminative features in these more visually distinguishable tasks [10]. This behavior aligns with the notion of functional sparsity: the model automatically compresses its memory usage when the task allows, while expanding representational breadth for histopathology heterogeneous conditions. These findings highlight that the momentum memory offering a dynamic balance between expressivity and compactness that is absent in static-memory or batch-local alignment methods.

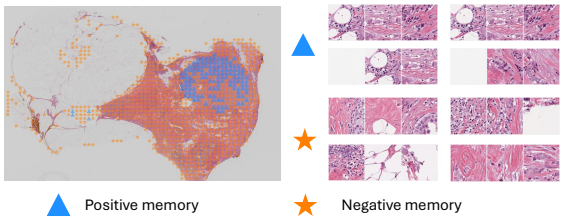

Figure S3. The visualization of memory usage on the misclassified case from the TCGA-BRCA dataset.

S4. Visualization on the misclassified case

To further investigate the misclassifications, we visualize the memory components of failure cases from the TCGA-BRCA dataset in the HER2 task (Fig. S3). As illustrated, the positive memory disproportionately attends to patches dominated by non-informative white background—redundant regions that theoretically should have been eliminated during initial preprocessing. This erroneous focus is similarly reflected in the negative memory representations. Consequently, these results highlight that rigorous background filtering and precise feature extraction remain crucial bottlenecks in standard WSI processing pipelines.

S5. Pseudo code for the proposed method

In the Algorithm 1, we provide the pseudo code for the MoMKD. The detailed implementation can be found in our GitHub repo.

References

[1] HER2 Testing in Breast Cancer - 2023 Guideline Update. https://www.cap.org/protocols-and-guidelines/cap-guidelines/current-cap-guidelines/recommendations-for-human-epidermal-growth-factor-2-testing-in-breast-cancer. 2

[2] Richard J. Chen, Tong Ding, Ming Y. Lu, Drew F. K. Williamson, Guillaume Jaume, Andrew H. Song, Bowen Chen, Andrew Zhang, Daniel Shao, Muhammad Shaban, Mane Williams, Lukas Oldenburg, Luca L. Weishaupt, Judy J. Wang, Anurag Vaidya, Long Phi Le, Georg Gerber, Sharifa Sahai, Walt Williams, and Faisal Mahmood. Towards a general-purpose foundation model for computational pathology. *Nature Medicine*, 30(3):850–862, 2024. 1

[3] Melina B. Flanagan, David J. Dabbs, Adam M. Brufsky, Sushil Beriwal, and Rohit Bhargava. Histopathologic variables predict Oncotype DX™ Recurrence Score. *Modern Pathology*, 21(10):1255–1261, 2008. 2

[4] Yongxin Guo, Ziyu Su, Onur C. Koyun, Hao Lu, Robert Wesolowski, Gary Tozbikian, M. Khalid Khan Niazi, and Metin N. Gurcan. BPMambaMIL: A bio-inspired prototype-guided multiple instance learning for oncotype DX risk assessment in histopathology. *Computer Methods and Programs in Biomedicine*, 272:109039, 2025. 3

[5] Frederick M. Howard, James Dolezal, Sara Kochanny, Galina Khramtsova, Jasmine Vickery, Andrew Srisuwananukorn, Anna Woodard, Nan Chen, Rita Nanda, Charles M. Perou, Olufunmilayo I. Olopade, Dezheng Huo, and Alexander T. Pearson. Integration of clinical features and deep learning on pathology for the prediction of breast cancer recurrence assays and risk of recurrence. *npj Breast Cancer*, 9(1):1–6, 2023. 1

[6] Mariia Ivanova, Francesca Maria Porta, Marianna D'Ercole, Carlo Pescia, Elham Sajjadi, Giulia Cursano, Elisa De Camilli, Oriana Pala, Giovanni Mazzarol, Konstantinos Venetis, Elena Guerini-Rocco, Giuseppe Curigliano, Giuseppe Viale, and Nicola Fusco. Standardized pathology report for HER2 testing in compliance with 2023 ASCO/CAP updates and 2023 ESMO consensus statements on HER2-low breast cancer. *Virchows Archiv*, 484(1):3–14, 2024. 2

[7] Soonmyung Paik, Steven Shak, Gong Tang, Chungyeul Kim, Joffre Baker, Maureen Cronin, Frederick L. Baehner, Michael G. Walker, Drew Watson, Taesung Park, William Hiller, Edwin R. Fisher, D. Lawrence Wickerham, John Bryant, and Norman Wolmark. A Multigene Assay to Predict Recurrence of Tamoxifen-Treated, Node-Negative Breast Cancer. *New England Journal of Medicine*, 351(27): 2817–2826, 2004. 2

[8] Gil Shamai, Ran Schley, Alexandra Cretu, Tal Neoran, Edmond Sabo, Yoav Binenbaum, Shachar Cohen, Tal Goldman, António Polónia, Keren Drumea, Karin Stoliar, and Ron 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(1):276, 2024. 2, 3

[9] Ziyu Su, Muhammad Khalid Khan Niazi, Thomas E. Tavolara, Shuo Niu, Gary H. Tozbikian, Robert Wesolowski, and Metin N. Gurcan. BCR-Net: A deep learning framework to predict breast cancer recurrence from histopathology images. *PloS One*, 18(4):e0283562, 2023. 3

[10] Ziyu Su, Yongxin Guo, Robert Wesolowski, Gary Tozbikian, Nathaniel S. O'Connell, M. Khalid Khan Niazi, and Metin N. Gurcan. Computational Pathology for Accurate Prediction of Breast Cancer Recurrence: Development and Validation of a Deep Learning-based Tool, 2024. 3

[11] Ziyu Su, Yongxin Guo, Robert Wesolowski, Gary Tozbikian, Nathaniel S. O'Connell, M. Khalid Khan Niazi, and Metin N. Gurcan. Computational Pathology for Accurate Prediction of Breast Cancer Recurrence: Development and Validation of a Deep Learning-based Tool, 2024. 2

[12] Renan Valieris, Luan Martins, Alexandre Defelicibus, Adriana Passos Bueno, Cynthia Aparecida Bueno de Toledo Osorio, Dirce Carraro, Emmanuel Dias-Neto, Rafael A. Rosales, Jose Marcio Barros de Figueiredo, and Israel Tojal da Silva. Weakly-supervised deep learning models enable HER2-low prediction from H &E stained slides. *Breast Cancer Research*, 26(1):124, 2024. 2

[13] Jon Whitney, German Corredor, Andrew Janowczyk, Shridar Ganesan, Scott Doyle, John Tomaszewski, Michael Feld-

man, Hannah Gilmore, and Anant Madabhushi. Quantitative nuclear histomorphometry predicts oncotype DX risk categories for early stage ER+ breast cancer. *BMC Cancer*, 18 (1):610, 2018. 2

[14] Andrew Zhang, Guillaume Jaume, Anurag Vaidya, Tong Ding, and Faisal Mahmood. Accelerating Data Processing and Benchmarking of AI Models for Pathology, 2025. 1