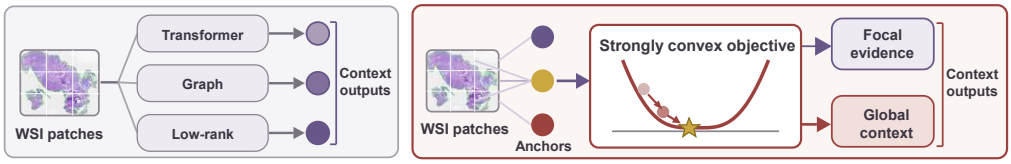
ENERGYMIL: MULTIPLE INSTANCE LEARNING AS STRONGLY CONVEX ENERGY MINIMIZATION FOR WHOLE-SLIDE IMAGE CLASSIFICATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Multiple instance learning (MIL) is the dominant framework for whole slide image (WSI) classification, where each slide becomes a bag of thousands of patch features supervised only by a slide label. Contextual MIL enriches these features through interactions among patches. Current designs specify the operations that compute context, while the target representation remains implicit. Making that target explicit gives contextual aggregation a principled endpoint. We design EnergyMIL to assign each WSI a unique contextual state defined over the entire slide. First, factorized anchor coupling connects tissue evidence across thousands of patches with storage linear in bag size. Then, a strongly convex contextual energy defines the state, and proximal refinement approaches it through a memory efficient learning path. Finally, complementary focal and global readouts produce the slide prediction. Across four datasets, EnergyMIL ranks first outright in eleven of the twelve dataset by metric comparisons and ties for first in the remaining one. Together, the formulation and results show how a well defined contextual state can provide a stable and scalable basis for integrating heterogeneous tissue evidence in WSI classification.



(a) Architecture-defined context. (b) Objective-defined context in EnergyMIL.

Figure 1: Architecture-defined and objective-defined contextual MIL.

1 INTRODUCTION

Whole-slide images (WSIs) are gigapixel records of heterogeneous tissue, yet supervision is commonly available only as a slide-level label (Campanella et al., 2019; Lu et al., 2021). Multiple instance learning (MIL) addresses this mismatch by representing each WSI as an unordered bag of patch features and learning from their shared label (Ilse et al., 2018; Lu et al., 2021). Pathology foundation models strengthen patch representations (Lu et al., 2024), but slide prediction still depends on how evidence from thousands of patches is combined. Attention MIL addresses this step by learning which patches support the label. Contextual MIL goes further by letting each patch representation depend on other evidence in the same slide (Li et al., 2021; Shao et al., 2021), making cross-patch interaction central to whole-slide classification.

Efforts to model these dependencies have followed several architectural routes. DSMIL relates instances to a critical instance, while TransMIL models morphological and spatial correlations with Transformer interactions (Li et al., 2021; Shao et al., 2021). Later methods use low-rank latent vectors for global interaction, re-embed features across regions, or construct task-aware subgraphs (Xiang & Zhang, 2023; Tang et al., 2024; Wang et al., 2025). Together, these advances enrich or

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

streamline slide-wide communication while leaving the contextual endpoint implicit. Making this endpoint explicit calls for a unique slide-wide target that remains tractable for thousands of patches. Figure 1 summarizes this distinction. Architectural contextual MIL identifies context with the output of a prescribed computation. EnergyMIL defines a contextual target first, then uses factorized global coupling and iterative refinement to approach it.

EnergyMIL first constructs anchor based global coupling, which gathers evidence across the WSI and returns slide wide context to every patch with storage linear in bag size. Second, a strongly convex nonnegative energy balances embedded morphology with anchor mediated interaction and defines a unique contextual state for the slide. Proximal state inference approaches this target, while one differentiable refinement carries the learning signal without retaining the full refinement trajectory. Finally, complementary focal and global readouts combine localized diagnostic evidence with the mean contextual state for slide prediction.

The analysis mirrors this construction. Normalized anchor coupling is symmetric, mass preserving, and nonexpansive. These properties retain the curvature imposed by the feature map, so the energy admits a unique nonnegative contextual state and proximal refinement converges linearly to it. The state is permutation equivariant and stable to changes in the injected patch evidence for a given coupling operator, while the slide prediction is permutation invariant. For learning, detached refinement approaches the state, followed by one differentiable update that provides a memory efficient approximation to implicit differentiation (Geng et al., 2021; Fung et al., 2022). The resulting formulation gives contextual aggregation a well defined target and keeps its training practical for large WSI bags.

Together, the formulation and analysis resolve a central modeling ambiguity in contextual MIL by supplying a representation level criterion for what interaction across a slide should recover and guarantees for how that state is reached. Across four primary datasets, EnergyMIL ranks first outright in eleven of the twelve comparisons spanning every dataset and metric and ties for first in the remaining one. Results on two supplementary datasets broaden the evaluation, while diagnostic analyses characterize the contribution of each design, sensitivity to model parameters, and computational behavior. More broadly, EnergyMIL establishes explicit contextual state design as a principled foundation for integrating heterogeneous tissue evidence in WSI classification.

2 Related Work

WSI Representation Learning. Gigapixel WSIs combine small diagnostic foci with tissue organization spanning the entire slide, while supervision is usually available only at slide level. Because native-resolution processing is computationally prohibitive, weakly supervised pipelines encode tiles and aggregate their features with MIL (Campanella et al., 2019; Lu et al., 2021). Pathology foundation models strengthen tile representations, and recent work further improves their task-specific adaptation (Lu et al., 2024; Shao et al., 2026; Xiong et al., 2026). These advances provide richer patch evidence, but slide prediction still depends on whether an aggregator can preserve local diagnostic cues while recovering their global context.

Contextual MIL for WSI. Pooling-based MIL forms a slide representation by selecting or weighting patch evidence (Ilse et al., 2018; Lu et al., 2021). This is effective for locating discriminative regions, but tissue interpretation also depends on how a patch relates to distant and neighboring morphology. Contextual MIL addresses this need through critical-instance relations, Transformer interactions, spatially constrained attention, or low-rank global communication (Li et al., 2021; Shao et al., 2021; Fourkioti et al., 2024; Xiang & Zhang, 2023). These architectures establish the value of cross-patch interaction, yet identify the contextual representation with the output of a prescribed finite computation. Consequently, the architecture determines not only how context is computed but also what contextual state is represented. Appendix A reviews complementary supervision, augmentation, and structural designs.

Optimization-Defined Representations. Optimization-defined networks replace a prescribed-depth representation with the solution of a parameterized problem. Deep equilibrium and monotone operator networks established fixed-point representations and uniqueness conditions (Bai et al., 2019; Winston & Kolter, 2020); recent studies extend this paradigm to memory-efficient score modeling and strongly convex energy-based graph states (Huang et al., 2024; Solodova et al., 2025).

Inspired by this perspective, EnergyMIL supplies a missing representation-level foundation for contextual MIL: a unique target that slide-wide interaction should recover. It realizes this criterion with a strongly convex, bag-dependent energy over variable-size patch sets and a factorized coupling that supports global interaction with storage linear in bag size.

3 ENERGYMIL

3.1 PROBLEM SETUP AND METHOD OVERVIEW

Let $\mathcal{D} = \{(\mathcal{B}_s, y_s)\}_{s=1}^{S}$ denote a WSI classification dataset, where $\mathcal{B}_s = \{x_{s,i}\}_{i=1}^{N_s}$ is a variable size bag of frozen patch features $x_{s,i} \in \mathbb{R}^{d_{\text{in}}}$ and only the slide label $y_s \in \{1, \dots, C\}$ is observed. For a single slide, omitting s, let $X \in \mathbb{R}^{N \times d_{\text{in}}}$ denote the matrix of preextracted patch features and $H = \phi(X) \in \mathbb{R}^{N \times d}$ its shared row wise embedding, which serves as the input to contextual modeling. Contextual MIL lets each embedded patch use evidence from the full bag, but a finite aggregation stack ties the resulting representation to the chosen computation. EnergyMIL instead defines $Z^\star(H) \in \mathbb{R}^{N \times d}$ as the unique minimizer of a bag dependent energy. As shown in Figure 2, learned anchors first establish factorized slide wide context. A strongly convex energy then defines a unique contextual state, which proximal refinement approaches with a memory efficient gradient approximation. Finally, complementary focal and global readouts map the state to slide logits.

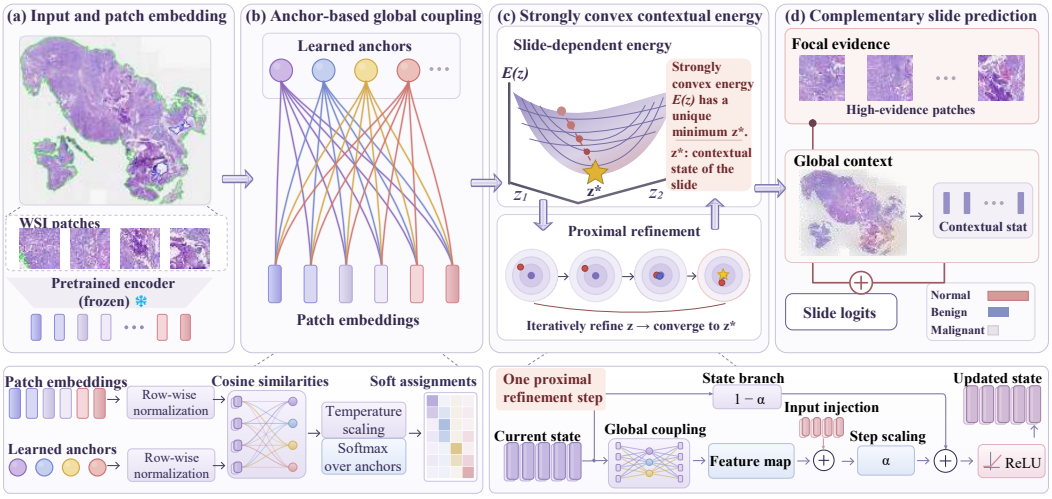


Figure 2: EnergyMIL in three stages. (a) Patch to anchor assignments induce normalized global coupling through an $N \times M$ factorization. (b) A symmetric feature map defines a strongly convex nonnegative energy, and proximal gradient refinement approaches its unique contextual state. One differentiable refinement follows detached state inference. (c) Gated attention and the mean contextual state provide focal and global summaries for slide prediction.

3.2 ANCHOR-BASED GLOBAL COUPLING

To share information across a variable-size WSI, we route patch interaction through M learned anchors. Let $E \in \mathbb{R}^{M \times d}$ be learned anchors with $M \ll N$. We define

$$P_{ij} = \frac{\exp(\widehat{h}_i^\top \widehat{e}_j / \tau)}{\sum_{k=1}^{M} \exp(\widehat{h}_i^\top \widehat{e}_k / \tau)}, \quad D = \operatorname{diag}(P^\top \mathbf{1}), \quad K = P D^{-1} P^\top, \tag{1}$$

where hats denote row-wise ℓ_2 normalization and $\tau > 0$ is the temperature. The action of K is evaluated as

$$KZ = P\big[D^{-1}(P^\top Z)\big], \tag{2}$$

which aggregates states to anchors and broadcasts them back. Assignments depend on the input but remain fixed during refinement.

Lemma 1 (Normalized coupling)**.** *For nonnegative row-stochastic P with positive column masses, K is symmetric, positive semidefinite, and row-stochastic. Its spectrum lies in* $[0, 1]$*,* $\|K\|_2 = 1$*, and* $\operatorname{rank}(K) \le M$.

The result follows from the Gram factorization $K = (PD^{-1/2})(PD^{-1/2})^\top$ and mass conservation. Appendix B gives the proof. Symmetry makes communication reciprocal, mass conservation preserves the slide average, and the spectral bound keeps the coupling compatible with the curvature of the state energy.

3.3 Strongly Convex Contextual Energy

For WSI classification, the contextual state should remain grounded in patch morphology, integrate evidence across the slide, and be uniquely defined. We therefore parameterize the feature map as

$$W = (1 - m)I_d - R^\top R, \qquad 0 < m < 1, \tag{3}$$

so $I_d - W = mI_d + R^\top R \succeq mI_d$. With $\mathcal{W}_H(Z) = KZW$ and an affine input injection $U = U(H)$, the contextual state is

$$Z^\star(H) = \arg\min_{Z \ge 0} f_H(Z), \qquad f_H(Z) = \tfrac{1}{2}\langle Z, (\mathcal{I} - \mathcal{W}_H)Z\rangle - \langle Z, U\rangle. \tag{4}$$

The linear term ties the state to embedded patch morphology, while the quadratic term incorporates slide-wide tissue context under a positive curvature margin.

Theorem 1 (Well-defined contextual state)**.** *Under Lemma 1 and Eq.* (3)*, the Hessian $Q_H = \mathcal{I} - \mathcal{W}_H$ satisfies $Q_H \succeq m\mathcal{I}$. Eq.* (4) *therefore has a unique minimizer for every finite U.*

If $\lambda_i \in [0, 1]$ is an eigenvalue of K and z_i is the corresponding transformed state row, then

$$\langle Z, Q_H Z\rangle = \sum_i \left(\|z_i\|^2 - \lambda_i z_i^\top W z_i\right) \ge \sum_i [1 - \lambda_i(1 - m)]\|z_i\|^2 \ge m\|Z\|_F^2. \tag{5}$$

Strong convexity and coercivity on the nonnegative orthant yield existence and uniqueness. Appendix B provides the full spectral and optimality arguments. Thus each WSI receives one contextual state even as its number of patches and anchor assignments vary.

3.4 Proximal State Inference and Learning

The unique minimizer is also a fixed point of a projected update. Because $\nabla f_H(Z) = Q_H Z - U$ and projection onto the nonnegative orthant is ReLU, proximal-gradient refinement (Parikh & Boyd, 2014) becomes

$$T_H(Z) = \operatorname{ReLU}((1 - \alpha)Z + \alpha(KZW + U)). \tag{6}$$

Its fixed points coincide with the constrained minimizer. Let $L = \|Q_H\|_2$. For $0 < \alpha < 2/L$, the map is a contraction and converges linearly from any initialization. We use

$$\alpha = \min\left\{\alpha_{\max}, \frac{2m}{(1 + \|W\|_2)^2}\right\}, \tag{7}$$

with an eight step power estimate of $\|W\|_2$. Eq. (7) is a conservative sufficient schedule, and less conservative admissible steps may exist for the symmetric energy. The stated guarantee holds when the true norm or a conservative upper bound is used. Appendix D characterizes the finite estimate used in the experiments.

We initialize $Z_0 = 0$ and stop when relative iterate change falls below a tolerance or the iteration cap is reached. State inference runs without recording the refinement history, so its memory use does not grow with the number of updates. One final update

$$\widetilde{Z} = T_H(\operatorname{stopgrad}(Z_T)) \tag{8}$$

provides a one step phantom gradient (Geng et al., 2021; Fung et al., 2022), a memory efficient leading term approximation to implicit differentiation. Appendix D gives the exact gradient relation.

The state guarantees concern the inner contextual problem for a given H and learned parameter setting. Parameter learning follows the standard nonconvex slide classification objective, with Eq. (8) providing its approximate gradient through state computation. Every refinement iterate approaches the same target state under that setting, while the stopping rule controls its numerical approximation.

3.5 SLIDE PREDICTION

Standard gated attention (Ilse et al., 2018) extracts a focal summary b_{att} from the contextual patch states. A second summary preserves the global state by weighting anchor averages by anchor mass:

$$b_{\mathrm{mass}} = \sum_j \frac{D_{jj}}{N} \left(D^{-1} P^\top \widetilde{Z} \right)_j = \frac{1}{N} \sum_i \widetilde{z}_i. \tag{9}$$

For the exact masses in Eq. (1), the second path is precisely the mean contextual state. Linear fusion, layer normalization, and a classifier map $[b_{\mathrm{att}}, b_{\mathrm{mass}}]$ to slide logits. The contextual state is permutation equivariant and both read-outs are invariant. Since $M < d$ in all reported settings, each refinement costs $O(Nd^2)$ under the reported evaluation order, with $O(NM + Nd + d^2)$ working storage. Algorithm 1 in Appendix D gives the complete forward pass, numerical details, and complexity accounting.

4 EXPERIMENTS

4.1 SETUP

Datasets. We evaluate four public WSI datasets. BML contains 111 H&E slides of resected brain metastases from primary lung cancer, comprising 12 small-cell lung cancers, 79 non-small-cell lung adenocarcinomas, and 20 non-adenocarcinoma non-small-cell lung cancers (Chadha et al., 2026). HER2 uses the 192-slide Yale dataset of invasive breast carcinomas from the HER2 Tumor ROIs collection, with 99 HER2-negative and 93 HER2-positive cases (Farahmand et al., 2022). AURORA contributes 225 H&E slides from 55 patients with metastatic breast cancer: 64 primary-tumor, 122 metastatic-tumor, and 39 adjacent-normal slides (Garcia-Recio et al., 2023). TCGA-KICH uses 109 diagnostic slides from the official TCGA Kidney Chromophobe project, grouped into stage I–II (84) and stage III–IV (25) (Davis et al., 2014). Every slide is represented by frozen 512 dimensional CONCH patch features (Lu et al., 2024).

Comparison methods. We compare ABMIL (Ilse et al., 2018), ACMIL (Zhang et al., 2024), AttriMIL (Cai et al., 2025), CLAM-SB (Lu et al., 2021), DSMIL (Li et al., 2021), DTFD-MIL (Zhang et al., 2022a), ILRA-MIL (Xiang & Zhang, 2023), MammothMIL (Shao et al., 2026), MR-Block (Xiong et al., 2026), RankMix (Chen & Lu, 2023), RRT-MIL (Tang et al., 2024), TAD-Graph (Wang et al., 2025), and TransMIL (Shao et al., 2021). Architecture specific hyperparameters follow the settings recommended in the original papers. For comparability, completed experiments use the same CONCH features and five fold splits and are trained for 100 epochs with unweighted cross entropy and Adam using learning rate 2×10^{-4} and weight decay 10^{-5}. Only task dependent input and output dimensions are changed.

EnergyMIL configuration. Appendix E.3 analyzes sensitivity to the number of anchors M, curvature margin m, and assignment temperature τ on BML and TCGA-KICH. Based on this analysis, the reported experiments use $M = 32$, $m = 0.3$, and $\tau = 0.1$. The remaining configuration uses embedding width 256, step cap $\alpha_{\max} = 0.5$, a cap of 400 updates, relative tolerance 10^{-4}, and dropout 0.25. Training uses one slide per batch, warmup over 10 epochs, and cosine annealing with minimum learning rate 10^{-4}.

Evaluation protocol. Following common practice for computationally intensive WSI experiments, we use one random seed (3407) rather than repeat the full protocol across multiple seeds and construct identical stratified five fold splits for every method. Every model is trained for a maximum of 100 epochs. We report the mean and standard deviation across held out folds for ACC, macro F1, and binary AUC for two class tasks or macro one versus rest AUC for multiclass tasks. TAD-Graph evaluation terminated unexpectedly for an unresolved reason after completing one AURORA fold and three TCGA-KICH folds. We report only these observed results and exclude the incomplete comparisons from cross dataset ranking. Appendix C details the datasets, training protocol, incomplete evaluation, and interpretation of uncertainty.

Table 1: Primary dataset comparison using CONCH features under five fold evaluation with random seed 3407. Each cell reports the fold mean on the first line and standard deviation on the second. Higher is better. Bold marks the highest displayed mean, including ties. † denotes the observed single fold TAD Graph result on AURORA.

Method	BML			HER2			AURORA			TCGA-KICH		
	ACC	AUC	F1	ACC	AUC	F1	ACC	AUC	F1	ACC	AUC	F1
ABMIL	0.911 ±0.052	0.878 ±0.066	0.856 ±0.085	0.813 ±0.086	0.814 ±0.083	0.811 ±0.088	0.818 ±0.043	0.844 ±0.035	0.795 ±0.042	0.780 ±0.022	0.620 ±0.180	0.487 ±0.116
ACMIL	0.902 ±0.065	0.869 ±0.089	0.850 ±0.111	0.844 ±0.102	0.847 ±0.098	0.841 ±0.106	0.822 ±0.035	0.828 ±0.046	0.780 ±0.063	0.789 ±0.042	0.619 ±0.197	0.555 ±0.122
AttriMIL	0.892 ±0.023	0.843 ±0.029	0.813 ±0.054	0.834 ±0.093	0.834 ±0.094	0.833 ±0.094	0.836 ±0.040	0.843 ±0.060	0.806 ±0.070	0.789 ±0.027	0.558 ±0.162	0.557 ±0.168
CLAM-SB	0.919 ±0.048	0.892 ±0.068	0.865 ±0.083	0.834 ±0.046	0.834 ±0.047	0.833 ±0.046	0.822 ±0.031	0.901 ±0.051	0.802 ±0.041	0.798 ±0.028	0.632 ±0.187	0.600 ±0.152
DSMIL	0.866 ±0.069	0.819 ±0.081	0.778 ±0.128	0.813 ±0.078	0.813 ±0.078	0.812 ±0.078	0.822 ±0.027	0.826 ±0.028	0.786 ±0.027	0.789 ±0.053	0.602 ±0.214	0.546 ±0.160
DTFD-MIL	0.883 ±0.050	0.832 ±0.065	0.795 ±0.109	0.844 ±0.091	0.844 ±0.090	0.843 ±0.092	0.818 ±0.040	0.832 ±0.043	0.781 ±0.051	0.789 ±0.027	0.631 ±0.201	0.539 ±0.142
ILRA-MIL	0.848 ±0.099	0.791 ±0.171	0.694 ±0.250	0.859 ±0.077	0.860 ±0.078	0.857 ±0.080	0.804 ±0.040	0.801 ±0.064	0.743 ±0.084	0.780 ±0.017	0.604 ±0.157	0.496 ±0.135
MammothMIL	0.901 ±0.037	0.907 ±0.061	0.842 ±0.040	0.844 ±0.084	0.876 ±0.066	0.841 ±0.088	**0.840 ±0.024**	0.913 ±0.054	0.807 ±0.064	0.771 ±0.005	0.499 ±0.271	0.435 ±0.002
MR-Block	0.875 ±0.064	0.889 ±0.075	0.795 ±0.113	0.818 ±0.076	0.835 ±0.073	0.817 ±0.077	0.813 ±0.025	0.865 ±0.074	0.770 ±0.072	0.798 ±0.043	0.480 ±0.227	0.558 ±0.172
RankMix	0.857 ±0.046	0.782 ±0.075	0.729 ±0.123	0.808 ±0.092	0.808 ±0.089	0.806 ±0.093	0.813 ±0.043	0.828 ±0.025	0.784 ±0.041	0.780 ±0.022	0.620 ±0.180	0.487 ±0.116
RRT-MIL	0.794 ±0.064	0.666 ±0.122	0.541 ±0.192	0.781 ±0.058	0.780 ±0.058	0.780 ±0.058	0.804 ±0.079	0.803 ±0.077	0.729 ±0.150	0.798 ±0.043	0.554 ±0.221	0.533 ±0.142
TAD-Graph	0.883 ±0.051	0.844 ±0.097	0.793 ±0.110	0.828 ±0.087	0.827 ±0.089	0.827 ±0.088	0.800 ±0.000†	0.790 ±0.000†	0.728 ±0.000†	0.803 ±0.026	0.624 ±0.277	0.607 ±0.168
TransMIL	0.848 ±0.099	0.766 ±0.172	0.688 ±0.257	0.807 ±0.071	0.806 ±0.071	0.806 ±0.071	0.809 ±0.020	0.894 ±0.048	0.754 ±0.033	0.789 ±0.053	0.541 ±0.146	0.555 ±0.175
EnergyMIL (ours)	**0.947 ±0.057**	**0.970 ±0.033**	**0.930 ±0.075**	**0.875 ±0.059**	**0.892 ±0.069**	**0.874 ±0.060**	**0.840 ±0.029**	**0.915 ±0.040**	**0.808 ±0.042**	**0.807 ±0.020**	**0.634 ±0.116**	**0.610 ±0.105**

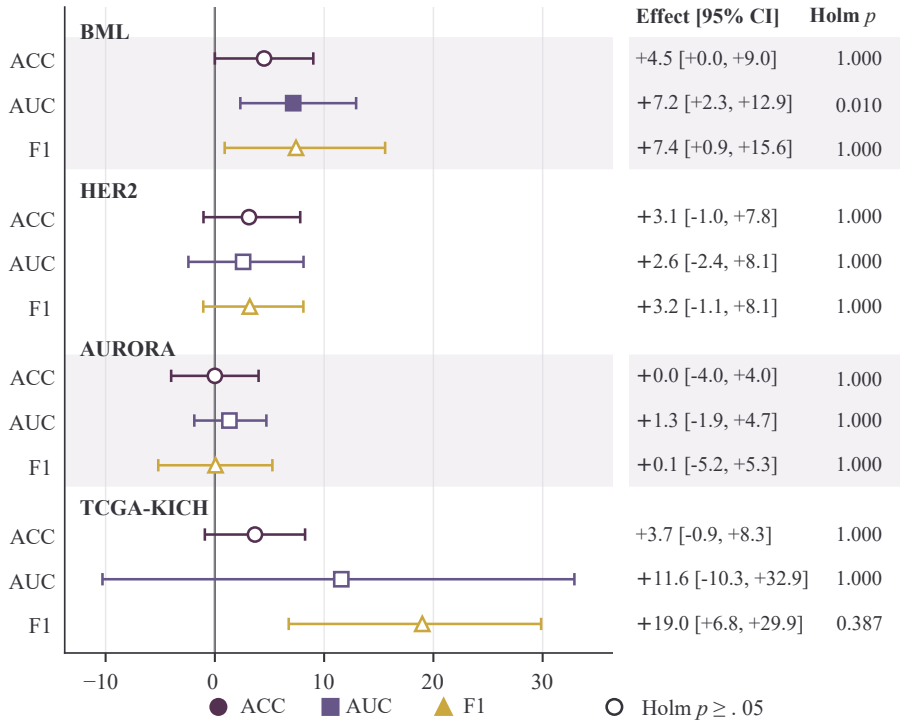


Figure 3: Paired slide-level effects relative to MammothMIL on matched out-of-fold predictions. Intervals are class-stratified paired-bootstrap 95% CIs. Filled markers denote Holm-corrected $p < 0.05$ from two-sided paired permutation tests.

4.2 Main Results

Table 1 tests whether the explicit contextual state delivers consistent gains across pathologies and label spaces. EnergyMIL ranks first outright in eleven of the twelve dataset by metric comparisons and ties for first in the remaining one. Relative to the strongest baseline with a complete five fold record in each comparison, it leads by 1.9 percentage points on average, counting the AURORA ACC tie as zero. The average margins are 1.3 points for ACC, 2.1 for AUC, and 2.3 for macro F1, indicating that the gain extends beyond raw accuracy to ranking quality and class balanced recognition. The strongest alternatives vary with the task. CLAM-SB and MammothMIL are closest on BML, ILRA-MIL and MammothMIL on HER2, and MammothMIL is especially competitive on AURORA, including an ACC tie. On TCGA-KICH, CLAM-SB leads the complete record baselines in AUC and F1, while CLAM-SB, MR-Block, and RRT-MIL share the highest baseline ACC. Despite these shifts, EnergyMIL remains at the top. This consistency supports energy defined context as a stable slide wide aggregator across multiclass and binary pathology tasks. Results on an anonymized hospital dataset and TCGA-KIRP are reported in Appendix E to assess breadth beyond the primary benchmark.

Aggregate means do not show whether the gains persist at the slide level. Figure 3 therefore compares EnergyMIL with MammothMIL, a recent method that performs strongly overall and matches EnergyMIL on AURORA ACC. Complete identifier-matched out-of-fold probabilities are also available for all four primary datasets, enabling the paired analysis. All twelve point estimates are non-negative. Paired-bootstrap intervals exclude zero for BML AUC and macro-F1 and for TCGA-KICH macro-F1. After Holm correction, BML AUC remains significant. Taken together, the paired analysis shows that the directional advantage of energy-defined context persists when the comparison is made on identical slides. At the same time, multiplicity correction separates broad empirical consistency from the narrower set of effects that can be claimed with statistical confidence. These intervals condition on the fitted fold models and quantify sample-level rather than retraining uncertainty.

4.3 Spatial Evidence

Figure 4 connects the slide prediction to tissue structure in a HER2 case selected by proximity to the median log patch count of its class without inspecting model correctness. Panel (a) gives the H&E overview. The green contours are dataset provided tumor ROIs rather than spatial HER2 status labels. Focal attention percentiles in panel (b) show where the readout concentrates, while panel (c) decomposes the HER2 negative versus HER2 positive logit contrast after contextualization, with red denoting support and blue opposition. Panel (d) shows the maximally supportive and opposing patches. Because contextual states already contain cross patch information, this decomposition is descriptive rather than a causal deletion map. EnergyMIL predicts HER2 negative with probability 0.846.

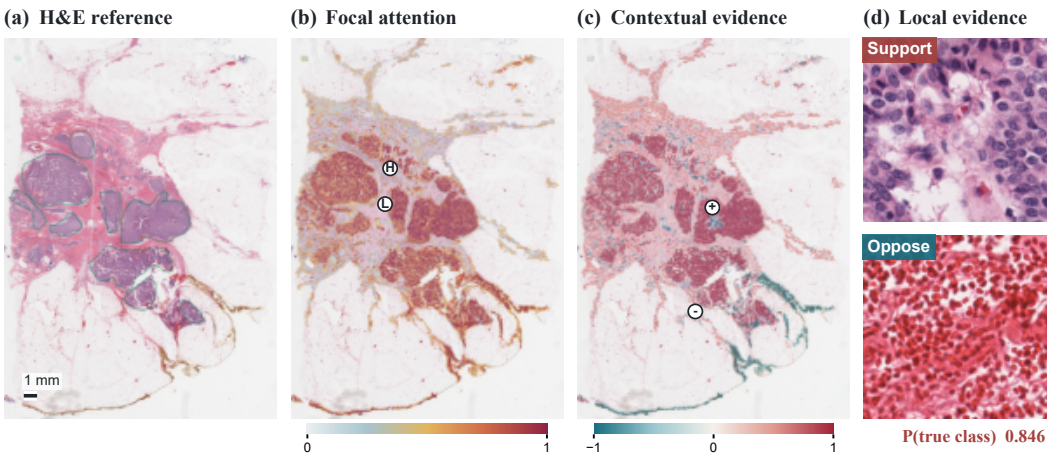


Figure 4: Spatial evidence for a HER2-negative slide. (a) H&E overview. (b) Focal-attention percentiles. (c) Signed contextual evidence. (d) Maximally supportive and opposing patches.

Figure 5 adds a between-method view on TCGA-KICH. Under the fixed selection rule in Appendix F, EnergyMIL correctly identifies the slide as early stage with probability 0.881, whereas ABMIL, CLAM-SB, and AttriMIL predict advanced stage from the same 4,079 patches. The spatial maps show that the methods differ not only in the final label but also in how tissue regions contribute to that decision. Additional fixed-baseline cases on BML, HER2, and AURORA appear in Appendix F.

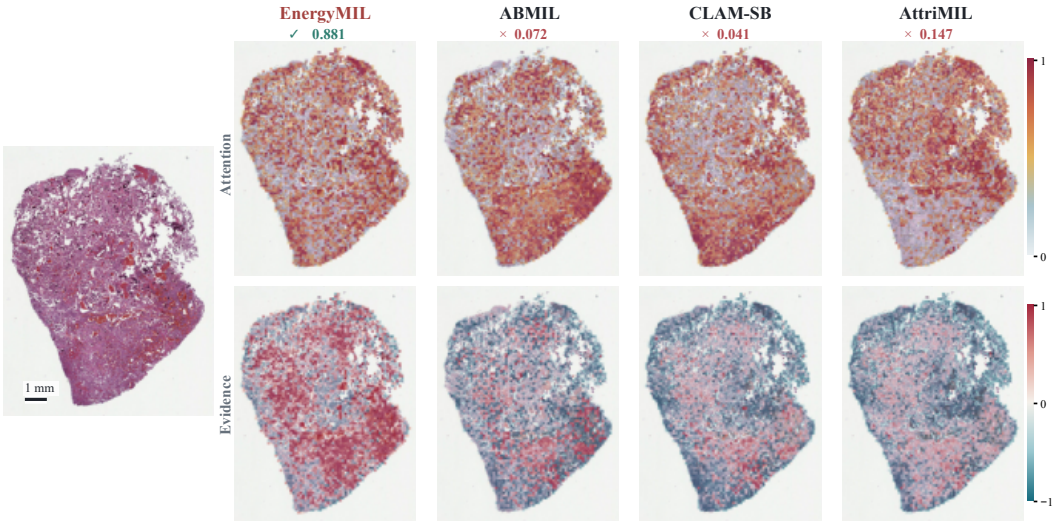


Figure 5: TCGA-KICH disagreement case. EnergyMIL correctly predicts early stage, while three fixed baselines predict advanced stage on the same slide.

4.4 Component Evidence

Table 2 reports the fold summaries, while Figure 6 exposes the direction and scale of all six dataset by metric effects. Anchor coupling has the largest mean contribution (4.0 points) and its removal degrades every comparison. Adaptive step sizing and the mean state readout are likewise positive in all six cells, with mean effects of 2.4 and 2.1 points. Detached state inference contributes positively in five cells. The sole exception is TCGA-KICH AUC, where the one step variant gains 3.1 points but has substantial fold level variation. Accordingly, the complete model leads five of the six means and provides the most consistent aggregate behavior across the two datasets and three metrics. Appendix D defines each variant.

Table 2: Five-fold component ablation on BML and TCGA-KICH (mean ± sample standard deviation). Bold marks the highest mean within each dataset and metric, including ties.

Variant	ACC	AUC	F1
	BML		
EnergyMIL (full)	**0.9466±0.0572**	**0.9698±0.0327**	**0.9297±0.0747**
One-step update	0.9379±0.0730	0.9393±0.0559	0.9217±0.0881
w/o anchor coupling	0.9107±0.0525	0.9501±0.0350	0.8724±0.0690
Fixed step size	0.9285±0.0384	0.9651±0.0267	0.9013±0.0595
w/o mean-state read-out	0.9289±0.0658	0.9392±0.0454	0.9042±0.0892
	TCGA-KICH		
EnergyMIL (full)	**0.8074±0.0197**	0.6337±0.1162	**0.6099±0.1050**
One-step update	0.7983±0.0236	**0.6647±0.1443**	0.5996±0.1505
w/o anchor coupling	0.7892±0.0227	0.5853±0.1638	0.5479±0.1541
Fixed step size	0.7983±0.0236	0.5750±0.1207	0.5835±0.1419
w/o mean-state read-out	0.7983±0.0236	0.6154±0.1217	0.5836±0.1423

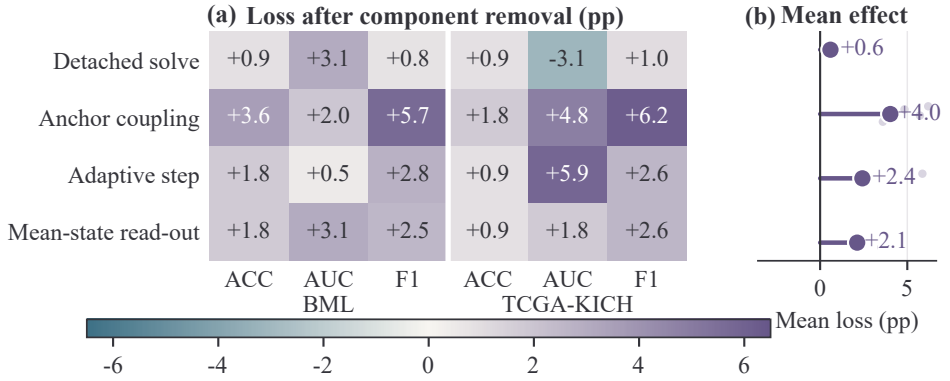


Figure 6: Signed component effects corresponding to Table 2. Each cell is full EnergyMIL minus the ablation in percentage points; positive values favor the complete model. The right panel summarizes the six effects for each ablation.

4.5 Model Complexity

Table 3 compares model size and profiled FLOPs under a shared input setting. EnergyMIL uses 0.535M parameters and 2.63G FLOPs, giving the fourth lowest FLOP count among the fourteen methods. Its parameter count is close to CLAM-SB and DTFD-MIL, while its profiled FLOPs are approximately half of theirs. Relative to RRT-MIL and TransMIL, EnergyMIL reduces profiled FLOPs by 81.6% and 89.5%, respectively. Together with the gains in Table 1, this profile shows that factorized anchor coupling provides global contextualization within a compact computational budget.

Table 3: Profiled model complexity under a shared protocol at $N = 10{,}000$ and $d = 512$ with two outputs.

Method	Params (M)	FLOPs (G)
EnergyMIL	0.535	2.63
ABMIL	0.330	3.28
ACMIL	0.263	1.97
AttriMIL	0.790	7.88
CLAM-SB	0.529	5.25
DSMIL	0.085	0.83
DTFD-MIL	0.527	5.25
ILRA-MIL	3.158	10.49
MammothMIL	0.613	1.35
MR-Block	0.566	4.26
RankMix	0.330	3.28
RRT-MIL	1.317	14.27
TAD-Graph	0.366	8.53
TransMIL	2.410	24.93

5 Conclusion

EnergyMIL formulates contextual WSI classification through the recovery of an optimization defined slide state. Its strongly convex energy assigns each slide a unique contextual target, while factorized anchor coupling, proximal inference, and complementary readouts make that target practical for variable size bags. Across four primary datasets, EnergyMIL ranks first outright in eleven of the twelve dataset by metric comparisons and ties for first in the remaining one. More broadly, contextual reasoning in MIL can be organized around the state to recover, with iterative depth controlling how accurately that state is reached.

AI USE STATEMENT

In this work, we used generative AI tools to aid and polish writing, assist literature retrieval and discovery, and support research ideation and execution. We have reviewed all AI assisted work and checked it against the original sources, method, and experimental evidence. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

The study uses retrospective histopathology datasets. The present results support methodological comparison. Establishing clinical utility would additionally require independent validation, patient level leakage checks, subgroup analysis, privacy review, and assessment by domain specialists.

REPRODUCIBILITY STATEMENT

Section 3 specifies the objective and forward pass. Appendix B provides proofs, while Appendices D and C document the algorithm, variants, datasets, metrics, and training conventions.

REFERENCES

Shaojie Bai, J. Zico Kolter, and Vladlen Koltun. Deep equilibrium models. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/01386bd6d8e091c2ab4c7c7de644d37b-Abstract.html.

Linghan Cai, Shenjin Huang, Ye Zhang, Jinpeng Lu, and Yongbing Zhang. AttriMIL: Revisiting attention-based multiple instance learning for whole-slide pathological image classification from a perspective of instance attributes. *Medical Image Analysis*, 103:103631, 2025. doi: 10.1016/j.media.2025.103631.

Gabriele Campanella, Matthew G. Hanna, Luke Geneslaw, Allen Miraflor, Vitor Werneck Krauss Silva, Klaus J. Busam, Edi Brogi, Victor E. Reuter, David S. Klimstra, and Thomas J. Fuchs. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nature Medicine*, 25(8):1301–1309, 2019. doi: 10.1038/s41591-019-0508-1.

Saahil Chadha, Durga V. Sritharan, Darin Dolezal, Sampada Chande, Thomas Hager, Khaled Bousabarah, Mariam S. Aboian, Veronica Chiang, MingDe Lin, Don X. Nguyen, and Sanjay Aneja. Matched MRI, segmentations, and histopathologic images of brain metastases from primary lung cancer. *Scientific Data*, 13(1):40, 2026. doi: 10.1038/s41597-025-06353-2. URL https://doi.org/10.1038/s41597-025-06353-2.

Yuan-Chih Chen and Chun-Shien Lu. RankMix: Data augmentation for weakly supervised learning of classifying whole slide images with diverse sizes and imbalanced categories. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23936–23945, 2023. doi: 10.1109/CVPR52729.2023.02292.

Caleb F. Davis, Christopher J. Ricketts, Min Wang, Lixing Yang, Andrew D. Cherniack, Hui Shen, Christian Buhay, Hyojin Kang, Sang Cheol Kim, Catherine C. Fahey, Kathryn E. Hacker, Gyan Bhanot, Dmitry A. Gordenin, Andy Chu, Preethi H. Gunaratne, Michael Biehl, Sahil Seth, Benny A. Kaipparettu, Christopher A. Bristow, Lawrence A. Donehower, Eric M. Wallen, Angela B. Smith, Satish K. Tickoo, Pheroze Tamboli, Victor Reuter, Laura S. Schmidt, James J. Hsieh, Toni K. Choueiri, A. Ari Hakimi, The Cancer Genome Atlas Research Network, Lynda Chin, Matthew Meyerson, Raju Kucherlapati, Woong-Yang Park, A. Gordon Robertson, Peter W. Laird, Elizabeth P. Henske, David J. Kwiatkowski, Peter J. Park, Margaret Morgan, Brian Shuch, Donna Muzny, David A. Wheeler, W. Marston Linehan, Richard A. Gibbs, W. Kimryn Rathmell, and Chad J. Creighton. The somatic genomic landscape of chromophobe renal cell carcinoma. *Cancer Cell*, 26(3):319–330, 2014. doi: 10.1016/j.ccr.2014.07.014.

Zhaolong Du, Shasha Mao, Xuequan Lu, Mengnan Qi, Yimeng Zhang, Jing Gu, and Licheng Jiao. Rethinking multiple-instance learning from feature space to probability space. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=torbeUlslS.

Saman Farahmand, Aileen I. Fernandez, Fahad Shabbir Ahmed, David L. Rimm, Jeffrey H. Chuang, Emily Reisenbichler, and Kourosh Zarringhalam. Deep learning trained on hematoxylin and eosin tumor region of interest predicts HER2 status and trastuzumab treatment response in HER2+ breast cancer. *Modern Pathology*, 35(1):44–51, 2022. doi: 10.1038/s41379-021-00911-w. URL https://doi.org/10.1038/s41379-021-00911-w.

Olga Fourkioti, Matt De Vries, and Chris Bakal. CAMIL: Context-aware multiple instance learning for cancer detection and subtyping in whole slide images. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=rzBskAEmoc.

Samy Wu Fung, Howard Heaton, Qiuwei Li, Daniel McKenzie, Stanley Osher, and Wotao Yin. JFB: Jacobian-free backpropagation for implicit networks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(6):6648–6656, 2022. doi: 10.1609/aaai.v36i6.20619.

Susana Garcia-Recio, Toshinori Hinoue, Gregory L. Wheeler, Benjamin J. Kelly, Ana C. Garrido-Castro, Tomas Pascual, Aguirre A. de Cubas, et al. Multiomics in primary and metastatic breast tumors from the AURORA US network finds microenvironment and epigenetic drivers of metastasis. *Nature Cancer*, 4(1):128–147, 2023. doi: 10.1038/s43018-022-00491-x. URL https://doi.org/10.1038/s43018-022-00491-x.

Zhengyang Geng, Xin-Yu Zhang, Shaojie Bai, Yisen Wang, and Zhouchen Lin. On training implicit models. In *Advances in Neural Information Processing Systems*, volume 34, 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/hash/cb8da6767461f2812ae4290eac7cbc42-Abstract.html.

Yuhao Huang, Qingsong Wang, Akwum Onwunta, and Bao Wang. Efficient score matching with deep equilibrium layers. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=J1djqLAa6N.

Maximilian Ilse, Jakub Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *International Conference on Machine Learning*, volume 80, pp. 2127–2136, 2018. URL https://proceedings.mlr.press/v80/ilse18a.html.

Bin Li, Yin Li, and Kevin W. Eliceiri. Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14318–14328, 2021. doi: 10.1109/CVPR46437.2021.01409.

Ming Y. Lu, Drew F. K. Williamson, Tiffany Y. Chen, Richard J. Chen, Matteo Barbieri, and Faisal Mahmood. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nature Biomedical Engineering*, 5(6):555–570, 2021. doi: 10.1038/s41551-020-00682-w.

Ming Y. Lu, Bowen Chen, Drew F. K. Williamson, Richard J. Chen, Ivy Liang, Tong Ding, Guillaume Jaume, Igor Odintsov, Long Phi Le, Georg Gerber, Anil V. Parwani, Andrew Zhang, and Faisal Mahmood. A visual-language foundation model for computational pathology. *Nature Medicine*, 30(3):863–874, 2024. doi: 10.1038/s41591-024-02856-4.

Mingrui Ma, Furong Luo, Binlin Ma, Shuxian Liu, Xiaoyi Lv, and Pan Huang. A multi-instance learning network with prototype-instance adversarial contrastive for cervix pathology grading. *Medical Image Analysis*, 109:103880, 2026. doi: 10.1016/j.media.2025.103880.

Neal Parikh and Stephen Boyd. Proximal algorithms. *Foundations and Trends in Optimization*, 1(3):123–231, 2014. doi: 10.1561/2400000003. URL https://web.stanford.edu/~boyd/papers/prox_algs.html.

Daniel Shao, Joel Runevic, Richard Chen, Drew Williamson, Ahrong Kim, Andrew Song, and Faisal Mahmood. Mixture of mini experts: Overcoming the linear layer bottleneck in multiple instance learning. In *International Conference on Learning Representations*,

2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/87425754bcc35f2bc62ef4a421a772d6-Abstract-Conference.html.

Zhuchen Shao, Hao Bian, Yang Chen, Yifeng Wang, Jian Zhang, Xiangyang Ji, and Yongbing Zhang. TransMIL: Transformer based correlated multiple instance learning for whole slide image classification. In *Advances in Neural Information Processing Systems*, volume 34, pp. 2136–2147, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/10c272d06794d3e5785d5e7c5356e9ff-Abstract.html.

Olga Solodova, Nick Richardson, Deniz Oktay, and Ryan P. Adams. Graph neural networks gone hogwild. In *International Conference on Learning Representations*, pp. 40323–40345, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/6462073c6bdf864ebfbbb11e80619f3e-Abstract-Conference.html.

Wenhao Tang, Fengtao Zhou, Sheng Huang, Xiang Zhu, Yi Zhang, and Bo Liu. Feature re-embedding: Towards foundation model-level performance in computational pathology. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11343–11352, 2024. doi: 10.1109/CVPR52733.2024.01078.

Fuying Wang, Jiayi Xin, Weiqin Zhao, Yuming Jiang, Maximus Yeung, Liansheng Wang, and Lequan Yu. TAD-Graph: Enhancing whole slide image analysis via task-aware subgraph disentanglement. *IEEE Transactions on Medical Imaging*, 44(6):2683–2695, 2025. doi: 10.1109/TMI.2025.3545680.

Ezra Winston and J. Zico Kolter. Monotone operator equilibrium networks. In *Advances in Neural Information Processing Systems*, volume 33, pp. 10718–10728, 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/hash/798d1c2813cbdf8bcdb388db0e32d496-Abstract.html.

Jinxi Xiang and Jun Zhang. Exploring low-rank property in multiple instance learning for whole slide image classification. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=01KmhBsEPFO.

Conghao Xiong, Zhengrui Guo, Zhe Xu, Yifei Zhang, Raymond Kai-yu Tong, Si Yong Yeo, Hao Chen, Joseph J. Y. Sung, and Irwin King. Exploiting low-dimensional manifold of features for few-shot whole slide image classification. In *International Conference on Learning Representations*, pp. 86203–86232, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/8ac50fd0a4eeeb1f077b17bb7c5353c3-Abstract-Conference.html.

Hongrun Zhang, Yanda Meng, Yitian Zhao, Yihong Qiao, Xiaoyun Yang, Sarah E. Coupland, and Yalin Zheng. DTFD-MIL: Double-tier feature distillation multiple instance learning for histopathology whole slide image classification. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18802–18812, 2022a. doi: 10.1109/CVPR52688.2022.01824.

Weijia Zhang, Xuanhui Zhang, Han-Wen Deng, and Min-Ling Zhang. Multi-instance causal representation learning for instance label prediction and out-of-distribution generalization. In *Advances in Neural Information Processing Systems*, volume 35, pp. 34940–34953, 2022b. doi: 10.52202/068431-2532. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/e261e92e1cfb820da930ad8c38d0aead-Abstract-Conference.html.

Yunlong Zhang, Honglin Li, Yunxuan Sun, Sunyi Zheng, Chenglu Zhu, and Lin Yang. Attention-challenging multiple instance learning for whole slide image classification. In *European Conference on Computer Vision*, pp. 125–143, 2024. doi: 10.1007/978-3-031-73668-1_8. URL https://doi.org/10.1007/978-3-031-73668-1_8.

Tingting Zheng, Kui Jiang, and Hongxun Yao. Dynamic policy-driven adaptive multi-instance learning for whole slide image classification. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8028–8037, 2024. doi: 10.1109/CVPR52733.2024.00767. URL https://doi.org/10.1109/CVPR52733.2024.00767.

Tingting Zheng, Kui Jiang, Yi Xiao, Sicheng Zhao, and Hongxun Yao. M3amba: Memory mamba is all you need for whole slide image classification. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15601–15610, 2025. doi: 10.1109/CVPR52734.2025.01454. URL https://doi.org/10.1109/CVPR52734.2025.01454.

APPENDIX ROADMAP

The supplementary material follows the argument of EnergyMIL from context to verification. Appendix A expands the methodological background, followed directly by the state guarantees in Appendix B. Appendix C documents the datasets and evaluation protocol. Appendix D presents the algorithm, numerical details, complexity, and controlled variants. Appendix E then presents the extended quantitative evidence, and Appendix F closes with spatial evidence across datasets.

A CONTEXTUAL MIL AND IMPLICIT REPRESENTATIONS

MIL aggregation. Attention-based MIL established a trainable mechanism for identifying informative instances from bag-level supervision (Ilse et al., 2018). CLAM adapted this principle to WSI classification with clustering constraints, while DSMIL connected instance-level and bag-level evidence through a critical instance (Lu et al., 2021; Li et al., 2021). These studies made weakly supervised slide prediction practical by turning slide labels into learnable evidence selection. Subsequent work broadened the role of the aggregator. TransMIL made cross-patch dependence explicit through Transformer interactions, CAMIL incorporated spatial neighborhoods, and ILRA-MIL showed that global communication can be made tractable through low-rank attention (Shao et al., 2021; Fourkioti et al., 2024; Xiang & Zhang, 2023). PSMIL further reconsidered the aggregation space by aligning and pooling instance predictions rather than features (Du et al., 2025). Together, this progression establishes contextual interaction as a central ingredient of WSI MIL. It also motivates a complementary question about the contextual endpoint that these interactions should recover.

Regularization and augmentation. A parallel line of work strengthens learning from slide-level supervision. DTFD-MIL distills information across pseudo-bags, RankMix augments ranked instances, ACMIL challenges concentrated attention, AttriMIL introduces attribute-level constraints, and PacMIL uses prototype–instance adversarial contrast (Zhang et al., 2022a; Chen & Lu, 2023; Zhang et al., 2024; Cai et al., 2025; Ma et al., 2026). CausalMIL uses bag supervision to identify causal instance representations for instance prediction and robustness under distribution changes (Zhang et al., 2022b). Together, these methods enrich the supervision and regularization available to the aggregator from coarse labels.

Feature and structure modeling. Other approaches reshape the representation on which aggregation operates. PAMIL adapts instance selection and aggregation through a learned policy, while M3amba carries information across instance groups with a memory driven state space model (Zheng et al., 2024; 2025). RRT-MIL re-embeds frozen features, TAD-Graph isolates task-relevant graph structure, MammothMIL expands aggregation capacity through mini experts, and MR-Block preserves local manifold structure (Tang et al., 2024; Wang et al., 2025; Shao et al., 2026; Xiong et al., 2026). Collectively, these advances address complementary bottlenecks in feature adaptation, structural organization, and model capacity. They refine how tissue evidence is represented and related, while opening a representation-level question about what state the resulting interaction is intended to produce.

Optimization-defined representations. Deep equilibrium models define representations as fixed points rather than outputs of networks with prescribed depth (Bai et al., 2019). Monotone operator equilibrium networks provide uniqueness through operator structure, while proximal algorithms supply convergent updates for constrained convex objectives (Winston & Kolter, 2020; Parikh & Boyd, 2014). Later studies extend this view to score modeling with reduced memory use and strongly convex graph energies (Huang et al., 2024; Solodova et al., 2025). Phantom gradients and backpropagation without a Jacobian enable training without retaining the full refinement trajectory (Geng et al., 2021; Fung et al., 2022). Adapting these ideas to WSI MIL requires an operator conditioned on each slide that supports variable bags, efficient global communication, and a unique state. EnergyMIL realizes these requirements through normalized factorized coupling, a strongly convex nonnegative energy, and complementary focal and global readouts. WSI MIL shows how tissue evidence can be selected and related, while models defined through optimization show how the resulting representation can be specified independently of depth. EnergyMIL unites these perspectives by making the contextual endpoint an explicit modeling object for WSI MIL.

B PROOFS AND STATE GUARANTEES

All statements in this section concern the inner state problem conditioned on an embedded bag H and the learned model parameters. Inner products and norms on matrix states are Frobenius unless marked otherwise.

Table 4: Notation. Bag size N varies across slides, while M and d are fixed model widths.

Symbol	Shape or domain	Meaning
X, H	$N \times d_{\text{in}}$, $N \times d$	Extracted features and embedded instances
E, P	$M \times d$, $N \times M$	Anchors and row stochastic assignments
D	$M \times M$ diagonal	Anchor masses
K	$N \times N$	Implicit instance coupling
R, W	$d \times d$	Learned factor and symmetric feature map
U, Z	$N \times d$	Input injection and variable state
Q_H	operator on $\mathbb{R}^{N \times d}$	State Hessian $\mathcal{I} - \mathcal{W}_H$
m, α, τ	positive scalars	Curvature margin, step, temperature

B.1 PROOF OF LEMMA 1

Let $s = P^\top \mathbf{1}_N$, with $s_j > 0$, and $D = \operatorname{diag}(s)$. The factorization

$$K = (PD^{-1/2})(PD^{-1/2})^\top \tag{10}$$

makes K symmetric and positive semidefinite and gives $\operatorname{rank}(K) \le \min(N, M)$. Its entries are nonnegative. Moreover,

$$K\mathbf{1}_N = PD^{-1}P^\top \mathbf{1}_N = PD^{-1}s = P\mathbf{1}_M = \mathbf{1}_N. \tag{11}$$

The induced infinity norm is one, so every eigenvalue has modulus at most one. Positive semidefiniteness makes every eigenvalue nonnegative. The vector of ones has eigenvalue one. Since K is symmetric, $\|K\|_2 = 1$.

B.2 CURVATURE, EXISTENCE, AND OPTIMALITY

The operator $\mathcal{W}_H$ is symmetric under the Frobenius inner product because

$$\langle Z_1, KZ_2W \rangle = \operatorname{tr}(Z_1^\top KZ_2W) = \langle KZ_1W, Z_2 \rangle. \tag{12}$$

Vectorization represents Q_H by $I - W \otimes K$. Thus, if λ_i and ω_j are eigenvalues of K and W, its eigenvalues are $1 - \lambda_i\omega_j$. Write $r_j \ge 0$ for an eigenvalue of $R^\top R$. Since $\omega_j = 1 - m - r_j$,

$$1 - \lambda_i\omega_j = (1 - \lambda_i) + \lambda_i(m + r_j) \ge m. \tag{13}$$

This establishes the curvature bound in Theorem 1. Since $\|\mathcal{W}_H\|_2 \le \|K\|_2\|W\|_2 = \|W\|_2$,

$$m \le \lambda_{\min}(Q_H), \qquad L = \|Q_H\|_2 \le 1 + \|W\|_2. \tag{14}$$

The objective satisfies $f_H(Z) \ge \frac{m}{2}\|Z\|_F^2 - \|U\|_F\|Z\|_F$. It is continuous and coercive. It therefore attains a minimum on the closed nonempty orthant, and strong convexity makes that minimizer unique. This completes the proof of Theorem 1.

Let $G = Q_HZ^\star - U$. The first order conditions are

$$Z^\star \ge 0, \qquad G \ge 0, \qquad Z^\star \odot G = 0. \tag{15}$$

They imply $Z^\star = \operatorname{ReLU}(Z^\star - \alpha G)$ for every $\alpha > 0$, and the converse follows coordinatewise. The fixed points of Eq. (6) therefore coincide with the constrained minimizer.

B.3 Convergence and residual bound

Projection onto a closed convex set is nonexpansive. For two states,

$$\|T_H(Z) - T_H(Y)\|_F \le \|(\mathcal{I} - \alpha Q_H)(Z - Y)\|_F \tag{16}$$

$$\le \max_{\lambda \in \mathrm{spec}(Q_H)} |1 - \alpha\lambda| \, \|Z - Y\|_F \tag{17}$$

$$\le q\|Z - Y\|_F, \tag{18}$$

where $q = \max\{|1 - \alpha m|, |1 - \alpha L|\}$. If $0 < \alpha < 2/L$, then $q < 1$. The contraction mapping theorem gives convergence from every initialization and

$$\|Z_t - Z^\star\|_F \le q^t \|Z_0 - Z^\star\|_F. \tag{19}$$

For $\alpha_{\max} > 0$ and $B = 1 + \|W\|_2$, Eq. (7) gives $0 < \alpha \le 2m/B^2 < 2/B \le 2/L$. The result also holds with any upper bound on $\|W\|_2$.

A fixed point residual also controls the error in the state:

$$\|Z - Z^\star\|_F \le \frac{\|T_H(Z) - Z\|_F}{1 - q}. \tag{20}$$

This follows by adding and subtracting $T_H(Z)$, applying the triangle inequality and contraction bound, and rearranging. Since $Z_{t+1} = T_H(Z_t)$, combining Eq. (20) with the stopping statistic in Appendix D gives

$$\|Z_t - Z^\star\|_F \le \frac{r_t \max(\|Z_t\|_F, 10^{-8})}{1 - q}. \tag{21}$$

Thus the tolerance on iterate change controls an absolute bound on state error. It is not itself a relative tolerance on state error.

B.4 Permutation equivariance and invariant prediction

Let Π be an $N \times N$ permutation matrix. Applying embedding and normalization to each row gives $P' = \Pi P$ and $U' = \Pi U$. Column masses are unchanged because $\Pi^\top \mathbf{1} = \mathbf{1}$, so $K' = \Pi K \Pi^\top$. For every feasible Z, $f_{\Pi H}(\Pi Z) = f_H(Z)$. The feasible orthant is preserved, and uniqueness gives $Z^\star(\Pi H) = \Pi Z^\star(H)$. Attention scores permute with instances while their weighted sum is unchanged, and mean pooling is invariant. The slide prediction at evaluation is therefore permutation invariant.

B.5 Stability for a fixed operator

Proposition 1 (Stability under input injection)**.** *For fixed K and W, let $Z_1^\star$ and $Z_2^\star$ minimize Eq.* (4) *with injections U_1 and U_2. Then*

$$\|Z_1^\star - Z_2^\star\|_F \le \frac{1}{m}\|U_1 - U_2\|_F. \tag{22}$$

The variational inequalities at the two optima, added together, give

$$m\|Z_1^\star - Z_2^\star\|_F^2 \le \langle U_1 - U_2, Z_1^\star - Z_2^\star \rangle. \tag{23}$$

The Cauchy Schwarz inequality completes the proof. The proposition characterizes stability for a fixed operator. A joint robustness analysis would additionally account for the dependence of P and K on the raw features.

B.6 Model interpretation

Interaction rank and state rank. The rank bound applies to the communication channel K, while the full state Z also receives the input injection. Without the nonnegativity constraint, $Z^\star = Q_H^{-1} U$. If $v^\top K = 0$, left multiplication of the stationarity equation gives $v^\top Z^\star = v^\top U$. Components of the input injection outside the interaction subspace therefore remain available. Under nonnegativity, the solution is piecewise linear in U for fixed operators and a fixed active set. Coordinate masking can also mix the subspaces. An interaction channel with rank at most M can therefore support a state whose rank exceeds M.

Symmetry. If a skew component $B - B^\top$ is added to W, its quadratic form vanishes because $Z^\top KZ$ is symmetric. Such an operator may still have a strongly monotone symmetric component, but the scalar energy in Eq. (4) represents only that symmetric component. More generally, the gradient of $\frac{1}{2}\langle Z, AZ\rangle$ is $\frac{1}{2}(A + A^*)Z$. The symmetric restriction therefore gives the precise energy interpretation used by EnergyMIL.

Mass weighted readout. Let $S = D^{-1}P^\top \widetilde{Z}$ contain the anchor averages and let $w_j = D_{jj}/N$. Then

$$\sum_j w_j S_j = \frac{1}{N}\mathbf{1}_M^\top P^\top \widetilde{Z} = \frac{1}{N}\mathbf{1}_N^\top \widetilde{Z}. \tag{24}$$

The corresponding ablation therefore removes the mean state path. It does not remove anchor interaction because the anchors still shape the contextual states used by both summaries.

C Datasets and Evaluation Protocol

C.1 Dataset composition

BML, HER2, AURORA, and TCGA-KICH follow their published dataset definitions (Chadha et al., 2026; Farahmand et al., 2022; Garcia-Recio et al., 2023; Davis et al., 2014). Each WSI is represented as a bag containing a variable number of frozen CONCH patch features (Lu et al., 2024). Every method uses the same tissue selection, slide labels, dataset membership, and cross validation partitions. This common input isolates differences in slide aggregation from differences in preprocessing or data selection.

Table 5: Sample counts and class distributions in the four primary datasets. Class counts follow the label order used in evaluation.

Dataset	Context	Samples	Class counts
BML	Brain metastases from lung	111	12 / 79 / 20
HER2	Yale HER2 breast dataset	192	99 / 93
AURORA	Metastatic breast dataset	225	39 / 122 / 64
TCGA-KICH	Kidney chromophobe stage groups	109	84 / 25

C.2 Training protocol

All methods use stratified five fold cross validation with identical partitions. A fresh model and optimizer are initialized for each fold. We use one fixed random seed of 3407, following common practice for computationally demanding WSI experiments. Training bags are shuffled within each fold. Every model is trained for a maximum of 100 epochs. The reported standard deviation describes variation across the five data partitions under this shared protocol.

The optimizer is Adam with learning rate 2×10^{-4} and weight decay 10^{-5}. The learning rate uses ten warmup epochs followed by cosine annealing with a period of ten epochs and a minimum of 10^{-4}. The loss is unweighted cross entropy. Experiments use Python 3.13, PyTorch 2.7 with CUDA 12.8, and an NVIDIA RTX 5090.

C.3 Metrics and uncertainty

ACC is the fraction of correct predictions. F1 is the unweighted mean of classwise F1. Binary AUC uses the probability of class 1, while multiclass AUC is the macro average over one versus rest comparisons. Main tables report the mean and standard deviation across folds. The deviations describe variation among the five data partitions. Confidence intervals and significance tests require a separate analysis. An unavailable result is left blank, and bold type is assigned at three decimal displayed precision. TAD-Graph completed one AURORA fold and three TCGA-KICH folds before its evaluation terminated unexpectedly for an unresolved reason. We report these entries only where explicitly indicated. We exclude them from aggregate ranks and do not extrapolate or impute the

missing folds. Sample level diagnostics include only methods and datasets with complete prediction scores.

D Algorithm and Complexity

D.1 Forward pass pseudocode

Algorithm 1 EnergyMIL forward pass for one slide

Require: Patch features X, learned parameters Θ, and inference settings m, τ, $\alpha_{\max}$, ε, and $T_{\max}$
Ensure: Slide logits ℓ and patch attention weights a

$H \leftarrow \phi(X)$
Compute P and D by Eq. (1) and set $U \leftarrow U(H)$
$W \leftarrow (1-m)I_d - R^\top R$ ▷ Eq. (3)
Estimate $\widehat{\sigma} \approx \|W\|_2$ with eight power iterations
$\alpha \leftarrow \min\{\alpha_{\max}, 2m/(1+\widehat{\sigma})^2\}$ ▷ Eq. (7)
$Z \leftarrow 0$ and $t \leftarrow 0$, then disable gradient history recording
repeat
  $Y \leftarrow P[D^{-1}(P^\top Z)]$ ▷ Factorized global coupling
  $Z^+ \leftarrow \mathrm{ReLU}((1-\alpha)Z + \alpha(YW + U))$ ▷ Eq. (6)
  $r \leftarrow \|Z^+ - Z\|_F / \max(\|Z\|_F, 10^{-8})$
  $Z \leftarrow Z^+$ and $t \leftarrow t+1$
until $r < \varepsilon$ or $t = T_{\max}$
$\widetilde{Z} \leftarrow T_H(\mathrm{stopgrad}(Z))$ with gradient recording enabled ▷ Eq. (8)
$(b_{\mathrm{att}}, a) \leftarrow \mathrm{GatedAttention}(\widetilde{Z})$
$b_{\mathrm{mass}} \leftarrow N^{-1}\sum_i \widetilde{z}_i$ ▷ Eq. (9)
$\ell \leftarrow \mathrm{Classifier}(\mathrm{LayerNorm}(\mathrm{Fuse}[b_{\mathrm{att}}, b_{\mathrm{mass}}]))$
return ℓ, a

The model follows the stages defined in the main text. It contains patch embedding, anchor based global coupling, strongly convex contextual energy, proximal state inference, gated attention, and the mean contextual state readout. The embedding is a linear layer followed by LeakyReLU, dropout, and LayerNorm. Anchors are orthogonally initialized. The attention hidden width is $d/2 = 128$. A linear fusion maps the concatenated $2d$ summary to width d. LayerNorm and a head with C classes then produce the slide logits.

D.2 Numerical details

We estimate $\|W\|_2$ using eight power iterations on $W^\top W$. The scalar estimate and step size are excluded from automatic differentiation. Column masses are floored at 10^{-6} before inversion. Writing $\bar{D}_{jj} = \max(D_{jj}, 10^{-6})$ gives $\bar{K} = P\bar{D}^{-1}P^\top$. The resulting operator remains symmetric and positive semidefinite and satisfies $\bar{K}\mathbf{1} \le \mathbf{1}$. Its spectrum therefore remains in $[0,1]$, so the curvature argument continues to apply when the floor is active.

At a differentiable fixed point, exact implicit differentiation gives $\mathrm{d}z^\star/\mathrm{d}\theta = (I - J_z)^{-1}J_\theta$, where $J_z = \partial T_\theta/\partial z$ and $J_\theta = \partial T_\theta/\partial\theta$. EnergyMIL instead differentiates $T_\theta(\mathrm{stopgrad}(z_T))$ and retains the leading term J_θ at the computed state. The readout and classifier are differentiated normally. Assignments remain differentiable functions of the patch features and anchors in the final update, which provides a training signal to the anchors. This yields a memory efficient approximation to implicit differentiation (Geng et al., 2021; Fung et al., 2022).

D.3 Complexity accounting

Assignments and state arrays use $O(NM + Nd)$ storage. Feature matrices and the input projection add $O(d^2 + d_{\mathrm{in}}d)$ parameter storage. Automatic differentiation retains the embedding, assignments, and final refinement while the T detached states are released, so memory use for the refinement history is independent of depth. Reverse mode differentiation traverses only the final update, making backward depth independent of T. Total activation storage grows with N, and the iteration count depends on conditioning and the stopping rule. Linear scaling in N therefore describes a given refinement budget.

Table 6: Leading arithmetic costs of the forward pass. T denotes detached refinement updates. Frozen feature extraction is excluded.

Operation	Cost	Frequency
Patch embedding	$O(Nd_{\text{in}}d)$	once
Assignments	$O(NMd)$	once
Feature map $R^\top R$	$O(d^3)$	once
Power iteration for $\|W\|_2$	$O(d^2)$	eight steps
Input injection	$O(Nd^2)$	once
Anchor aggregate and broadcast	$O(NMd)$	$T+1$ times
Dense feature mixing	$O(Nd^2)$	$T+1$ times
Gated attention projections	$O(Nd^2)$	once
Mass read-out	$O(NMd)$	once
Fusion and classification	$O(d^2)$	once

Computing KZ before multiplication by W costs $O(Nd^2)$ per update under the reported setting $M < d$. Iterative state inference therefore costs $O(TNd^2)$ for T updates. Associativity also permits $KZW = P[D^{-1}((P^\top Z)W)]$, whose cost reduces to $O(NMd)$ for WSI bags with $N > d$. Eq. (9) likewise permits direct mean pooling or an equivalent mass weighted average of anchor states. These equivalent forms clarify the available computational tradeoffs.

D.4 Variant definitions

Table 7: Component variants of EnergyMIL.

Variant	Change
EnergyMIL with one update	Skip detached iterations and apply one differentiable update from $Z_0 = 0$. Since $KZ_0 = 0$, this state update contains no K mediated cross patch coupling.
EnergyMIL without anchor coupling	Replace K by the identity so states evolve independently before pooling.
EnergyMIL with a fixed step	Use the configured constant step cap instead of the schedule based on the norm estimate. This variant is not covered by the guarantee associated with Eq. (7).
EnergyMIL without mean readout	Zero the second summary and retain gated attention.
EnergyMIL with an antisymmetric component	Add $B - B^\top$ to W. This diagnostic lies outside the symmetric energy model.

E Extended Experiments

This section follows the questions raised by the main results. We first assess breadth across datasets and metrics. We then quantify paired statistical reliability and examine sensitivity to model parameters. Every analysis states the available evidence, and incomplete results are excluded from aggregate claims.

E.1 Breadth across datasets and metrics

Figure 7 summarizes the twelve primary dataset by metric comparisons using only methods with complete five fold results. The faint points retain the underlying ranks, so the endpoint does not hide whether a method is consistently competitive or is driven by a few favorable tasks. EnergyMIL is first or tied first in every comparison and consequently has a mean dense rank of 1.04.

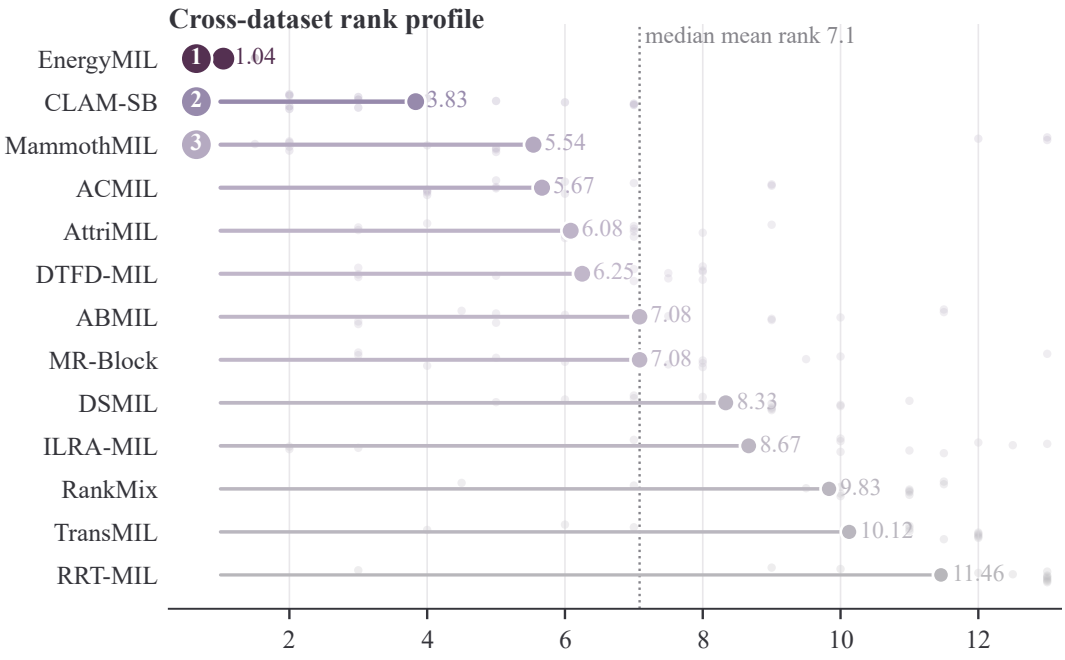


Figure 7: Rank profile over four datasets and three metrics. Each faint point is one dense rank for a dataset and metric. The large endpoint is their mean, and the line connects rank one to that mean. Lower is better. The profile includes only methods with complete results over five folds. Incomplete entries are neither imputed nor used for ranking.

Additional datasets. The two supplementary datasets examine performance beyond the primary benchmark. The first is an anonymized institutional dataset collected at a hospital. Its source and detailed composition will be disclosed after publication. We report both datasets separately because they were added after the primary comparison and therefore serve as breadth evidence rather than as a basis for selecting the method.

Table 8: Additional results on the hospital dataset and TCGA-KIRP. Entries are fold mean±standard deviation under the common frozen CONCH feature benchmark.

Method	Hospital dataset			TCGA-KIRP		
	ACC	AUC	F1	ACC	AUC	F1
ABMIL	0.769±0.054	0.731±0.080	0.733±0.078	0.843±0.060	0.808±0.063	0.785±0.067
ACMIL	0.739±0.088	0.699±0.143	0.664±0.188	0.865±0.056	0.815±0.080	0.825±0.060
AttriMIL	0.723±0.088	0.686±0.120	0.661±0.168	0.877±0.044	0.846±0.041	0.834±0.047
CLAM-SB	0.708±0.064	0.636±0.141	0.645±0.156	0.865±0.044	0.820±0.050	0.801±0.066
DSMIL	0.754±0.084	0.729±0.092	0.717±0.122	0.869±0.028	0.845±0.042	0.822±0.032
DTFD-MIL	0.708±0.064	0.606±0.162	0.627±0.159	0.869±0.047	0.826±0.065	0.820±0.051
ILRA-MIL	0.769±0.077	0.752±0.094	0.751±0.090	0.851±0.075	0.810±0.106	0.790±0.093
MammothMIL	0.677±0.064	0.686±0.147	0.598±0.140	0.873±0.043	0.865±0.067	0.833±0.044
MR-Block	0.723±0.069	0.634±0.222	0.645±0.170	0.869±0.055	0.823±0.067	0.820±0.065
RankMix	0.754±0.064	0.711±0.100	0.705±0.109	0.843±0.060	0.805±0.081	0.785±0.067
RRT-MIL	0.692±0.094	0.650±0.078	0.641±0.083	0.862±0.044	0.775±0.137	0.804±0.064
TAD-Graph	–	–	–	0.865±0.033	0.833±0.064	0.817±0.040
TransMIL	0.692±0.077	0.637±0.179	0.626±0.149	0.850±0.063	0.842±0.049	0.808±0.070
EnergyMIL (ours)	0.785±0.064	0.655±0.135	0.748±0.109	0.869±0.054	0.803±0.100	0.799±0.097

On the hospital dataset, EnergyMIL attains the highest ACC and is within 0.3 points of the best macro-F1, although its AUC is lower. On TCGA-KIRP, it remains competitive in ACC, while the

strongest baselines lead the displayed AUC and macro-F1. These results delimit the current scope of the evidence rather than extend the primary ranking claim.

E.2 Statistical reliability of the comparison

Aggregate means alone do not show whether the result is stable across data splits. Figure 8 therefore displays every EnergyMIL fold for each primary dataset and metric. The five points in each cell correspond to the five held out folds, while the interval is one fold standard deviation rather than a confidence interval. Variation is largest for F1 and AUC on TCGA-KICH, where the minority stage group is small, but no fold or dataset is suppressed by the aggregate summary.

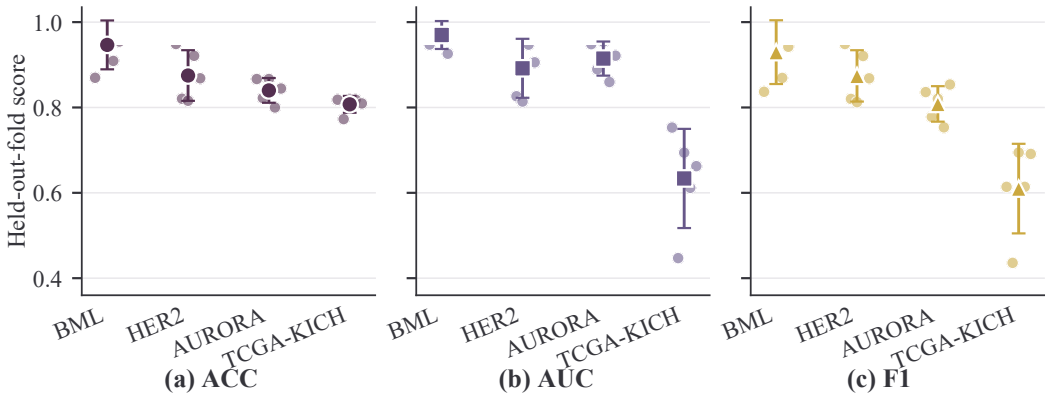


Figure 8: Fold-level stability of EnergyMIL on the four primary datasets. Small dots are the five held-out folds. Large markers and bars show the arithmetic mean and one sample standard deviation across folds. The intervals quantify variation across the fixed five-fold protocol and are not confidence intervals.

Fold summaries do not use the pairing available at the slide level. Figure 3 in Section 4.2 compares EnergyMIL with MammothMIL, the recent baseline for which complete, identifier-matched out-of-fold probabilities are available on all four datasets. Effects are recomputed from the same slides for both methods. The intervals use class-stratified paired bootstrap resampling, and the permutation tests exchange the two probability vectors within each slide before Holm correction across the twelve displayed comparisons.

This analysis adds uncertainty estimates and paired tests to the aggregate comparison.

E.3 Parameter Sensitivity

Figure 9 examines sensitivity to the number of anchors M, curvature margin m, and assignment temperature τ on BML and TCGA-KICH. We vary one factor at a time and summarize each candidate by its mean over the six dataset and metric cells. This analysis supports the reported values $M = 32$, $m = 0.3$, and $\tau = 0.1$. Their aggregate means exceed the strongest alternatives by 0.93, 0.36, and 2.15 percentage points in the respective sweeps. The individual metrics clarify the tradeoffs behind these averages. Setting $M = 128$ raises TCGA-KICH F1 by 6.26 points and lowers BML F1 by 4.40 points. Setting $m = 0.1$ raises TCGA-KICH AUC and F1 by 3.68 and 4.05 points and lowers their BML counterparts by 4.62 and 3.45 points. The reported values balance the observed cross dataset tradeoffs within the analyzed ranges and are used in Table 1.

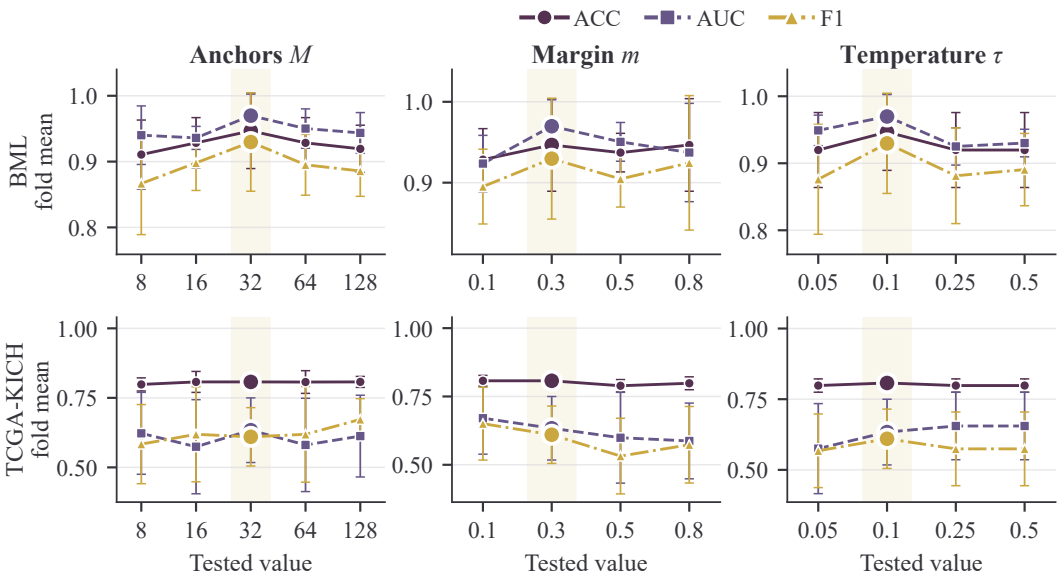


Figure 9: Parameter sensitivity for the number of anchors M, margin m, and assignment temperature τ. Each sweep varies one factor at a time. Curves are means over five folds and error bars are one fold standard deviation. Gold bands and white ring markers identify the values supported by the analysis and used in Table 1. Each tested value is a separately trained configuration.

F Spatial Evidence Across Datasets

Figure 5 in the main text provides the TCGA-KICH comparison. Figures 10, 11, and 12 extend the analysis to BML, HER2, and AURORA across distinct tasks and tissue appearances. Each case follows a fixed selection rule. We first require a correct EnergyMIL prediction and maximize the number of errors among ABMIL, CLAM-SB, and AttriMIL. We then maximize the gap between the probability that EnergyMIL assigns to the true class and the mean probability assigned by the three baselines. The slide identifier resolves any remaining tie. All methods receive the same patches and tissue extent. The upper row reports attention percentiles normalized within each slide, with AttriMIL using the branch for the labelled class. The lower row reports signed evidence percentiles for the contrast between the target and reference classes. Plus and cross symbols denote correct and incorrect predictions. The adjacent value is the probability assigned to the true class. Across these cases, the methods often attend to similar tissue regions, while their signed evidence differs substantially. Together, the four datasets show how the contextual state reshapes tissue contributions under shared spatial evidence.

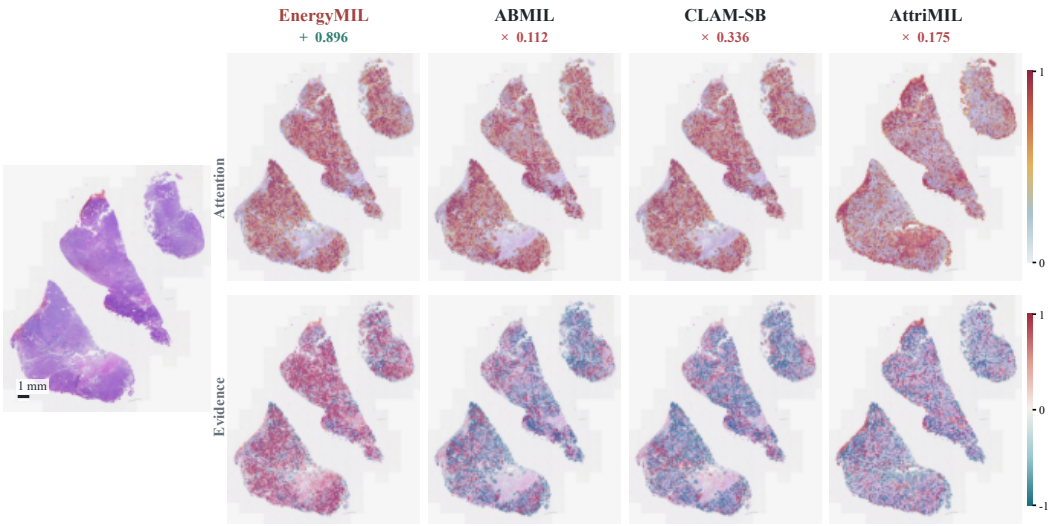


Figure 10: BML disagreement case over 7,280 shared patches. EnergyMIL correctly predicts class 2 with probability 0.896.

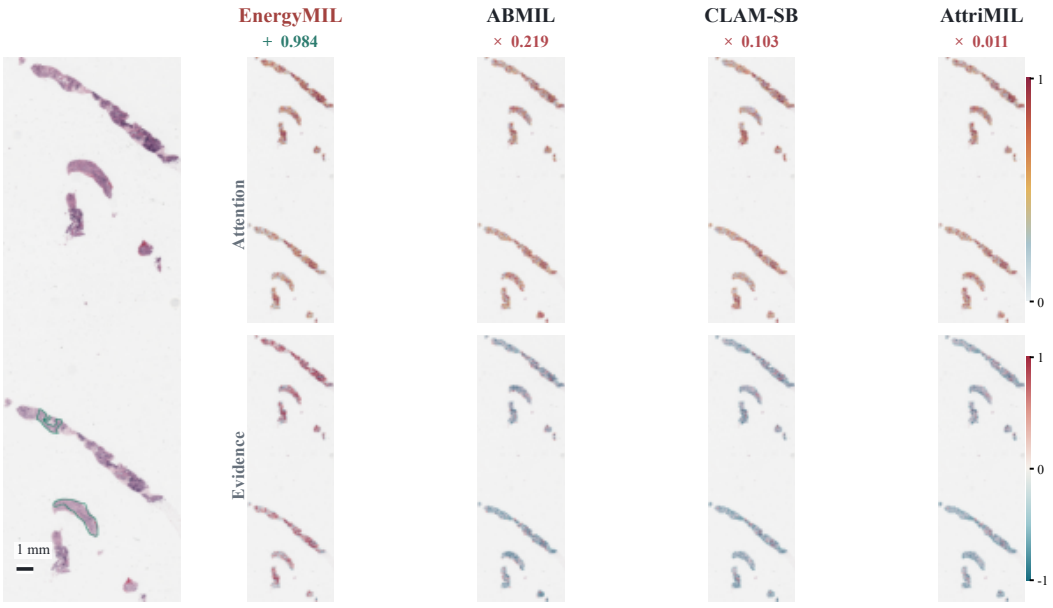


Figure 11: HER2 disagreement case over 1,604 shared patches. EnergyMIL correctly predicts HER2 positive status with probability 0.984.

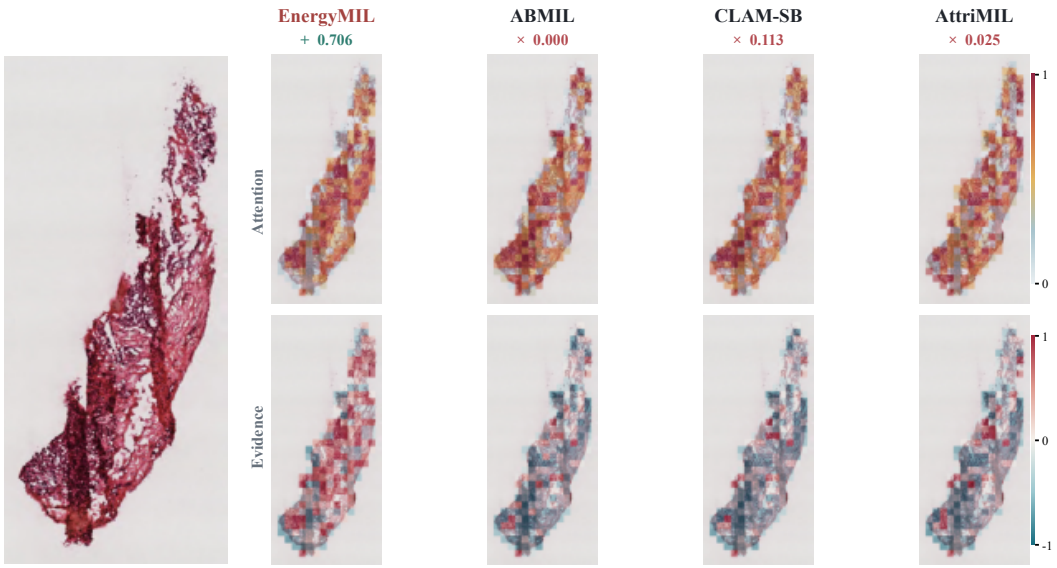


Figure 12: AURORA disagreement case over 215 shared patches. EnergyMIL correctly predicts class 1 with probability 0.706.