Objective Design for Self-Supervised Slide-Level Aggregation over Frozen Pathology Foundation Features

Yancheng Liu [1] **Shaolei Lu** [2] **Randall Balestriero** [1]

Abstract

Pathology foundation models give strong patch features, but slide-level prediction still requires aggregating thousands of frozen embeddings into a single representation. We ask whether self-supervised pretraining of the aggregator improves over static pooling when the patch encoder is fixed, and which SSL objective is best matched to whole-slide images. Holding the patch encoder, aggregator, view policy, token budget, and training schedule constant, we vary only the SSL objective: VICReg, JEPA, LeJEPA-2C, and a WSI-adapted multi-crop variant LeJEPA-MC. Across morphology classification, molecular prediction, cohort transfer, ordinal grading, and a private clinical cohort, LeJEPA-MC emerges as the most robust initialization and the strongest representation for the vast majority of clinical endpoints. Crucially, we find that multi-crop self-supervision acts as a powerful regularizer against cohort shift for high-capacity aggregators and excels on ordinal grading tasks by forcing local focal features to align with global context. While static pooling remains a surprisingly high-floor baseline on saturated within-cohort tasks, it is clearly surpassed on harder, label-scarce, and out-of-distribution clinical tasks, suggesting that distributionally regularized global/local prediction is better matched to heterogeneous WSI aggregation than two-view invariance or latent context-target prediction alone.

1. Introduction

Whole-slide images (WSIs) are gigapixel histopathology scans, routinely tiled into 10^4–10^5 patches per slide before downstream analysis. Recent pathology foundation models—UNI (Chen et al., 2024), Prov-GigaPath (Xu et al., 2024), CONCH (Lu et al., 2024), Virchow (Vorontsov et al., 2024), and Phikon (Filiot et al., 2023)—produce strong, transferable patch-level features via self-supervised pretraining on hundreds of millions of patches. But clinical labels are assigned at the slide or patient level, so deployment requires *aggregating* thousands of frozen patch embeddings into a single slide representation. With modern foundation features, even simple static pooling—mean, max, or both—is a difficult baseline to beat.

This paper asks a focused question: *when the patch encoder is fixed, can self-supervised pretraining of the slide aggregator improve over static pooling, and which SSL objective is best suited to WSIs?* The question matters because patch-level pretraining has begun to saturate: scaling the encoder or its corpus yields diminishing returns on slide-level endpoints (Xu et al., 2024; Chen et al., 2024), while the aggregator on top is typically supervised and trained from scratch on each task. If aggregation is the next bottleneck, isolating which SSL objective is best matched to it is a prerequisite for any further work on slide-level pretraining.

We hold the patch encoder, pretraining corpus, aggregator, view policy, token budget, training schedule, checkpoint rule, and downstream protocol constant, and vary only the SSL objective. This isolates objective design from architectural and data-scale confounders that prior comparisons have not fully controlled.

We compare four objectives that test distinct mechanisms. **VICReg** (Bardes et al., 2022) probes whether two-view invariance with explicit variance/covariance regularization is sufficient. **JEPA** (Assran et al., 2023) tests latent context-to-target prediction with an EMA target, the family that also includes BYOL (Grill et al., 2020) and DINO (Caron et al., 2021). **LeJEPA-2C** (Balestriero & LeCun, 2025) replaces the EMA/predictor machinery with two-view centroid prediction and SIGReg distributional regularization. **LeJEPA-MC** extends this to a WSI-adapted global/local crop scheme inspired by DINO's multi-crop (Caron et al., 2021), where regional tissue views predict a stable global slide centroid.

[1]Department of Computer Science, Brown University, Providence, RI, USA [2]Department of Pathology and Laboratory Medicine, Warren Alpert Medical School of Brown University, Providence, RI, USA. Correspondence to: Yancheng Liu <yancheng_liu@brown.edu>.

Our main finding is that slide-level SSL helps when the objective and view structure match the data. Static pooling is strong on saturated within-cohort morphology, but the LeJEPA variants consistently produce the strongest slide-level representations, with LeJEPA-MC performing best on four of five full-label linear-probing endpoints and showing the clearest gains on cohort transfer, ordinal grading, and clinically heterogeneous tasks. Fine-tuning and low-label probing show the same qualitative pattern: multi-crop predictive pretraining improves both frozen representation quality and supervised initialization, although it does not win every saturated or ultra-low-label endpoint.

Contributions.

- A controlled head-to-head comparison of SSL objectives for slide aggregation over frozen pathology features, with all confounders fixed except the loss family.
- A WSI-adapted multi-crop LeJEPA objective that uses regional tissue crops as local views predicting a stable global centroid.
- An evaluation across five tasks spanning morphology, molecular phenotype, cohort transfer, ordinal grading, and a private clinical cohort, with linear probing, fine-tuning, and low-label regimes.

2. Related Work

Pathology foundation models. Self-supervised pretraining on large pathology corpora has produced patch-level encoders that transfer broadly: UNI (Chen et al., 2024) (DINOv2 on 100M patches), Prov-GigaPath (Xu et al., 2024) (DINOv2 with slide-level masked autoencoding on 1.3B patches), CONCH (Lu et al., 2024) (vision-language pretraining on H&E–text pairs), Virchow (Vorontsov et al., 2024) (DINOv2 on 1.5M slides), and Phikon (Filiot et al., 2023) (iBOT). These models share a deployment pattern: features are extracted once and reused frozen for slide-level tasks. We use UNI2 throughout for fairness and reproducibility; the question we ask is independent of which patch encoder is fixed.

Slide-level aggregation. Aggregating variable-length bags of patch features into slide-level predictions is most often handled by attention-based MIL (Ilse et al., 2018; Lu et al., 2021), Transformer-style MIL (Shao et al., 2021; Zhang et al., 2022), or—most recently—state-space MIL such as MambaMIL (Yang et al., 2024), all trained supervised end-to-end on each task. We use a memory-efficient MAMBA2MIL aggregator (Mamba-2 (Dao & Gu, 2024) + gated attention) and treat aggregator architecture as fixed; the focus is the objective.

Slide-level self-supervised learning. Several recent works pretrain slide-level encoders, with different supervisory signals and design choices that this paper builds on. HIPT (Chen et al., 2022) introduced hierarchical DINO over patch and region tokens. SPT (Hou et al., 2024) introduced the feature-space augmentation pipeline—splitting, cropping, masking—that we adopt and extend with shared D4 transforms and partial-overlap splitting. Madeleine (Jaume et al., 2024c) uses multi-stain consistency as the supervisory signal; for our purposes its key technical contribution is fixed-N batching of slide bags, which inspires our 2048-token view-finalization protocol. TANGLE (Jaume et al., 2024b) aligns slide embeddings with paired transcriptomics. The closest prior work is COBRA (Lenz et al., 2025), which also uses a Mamba-2 + gated-attention slide encoder over frozen patch features, but trains it with a contrastive objective using multiple foundation models as cross-encoder augmentations. We borrow the COBRA architectural recipe but differ in three ways: (i) we use a single foundation model and feature-space augmentations rather than cross-FM views, isolating the role of the SSL objective; (ii) our objective is non-contrastive (LeJEPA-MC's centroid prediction with SIGReg) rather than contrastive, avoiding negative-sample selection; and (iii) we add a global/local multi-crop scheme that COBRA does not use. Our work therefore complements COBRA: same backbone, different objective, isolating the objective contribution.

Self-supervised objectives. SSL methods differ mainly in how they define agreement and prevent collapse. Contrastive methods such as SimCLR (Chen et al., 2020) and MoCo (He et al., 2020) use InfoNCE with negative samples; asymmetric self-distillation methods such as BYOL (Grill et al., 2020) and DINO (Caron et al., 2021) use stop-gradient targets and EMA dynamics without negatives; VICReg (Bardes et al., 2022) uses explicit variance and covariance regularization. JEPA reframes SSL as latent prediction from a context to a target representation (Assran et al., 2023), and LeJEPA replaces EMA targets with explicit distributional regularization (SIGReg) (Balestriero & LeCun, 2025). We compare a representative subset (VICReg, JEPA, LeJEPA-2C, LeJEPA-MC) under one WSI aggregation setup; methods involving negatives or EMA centering schedules introduce additional confounders that this controlled comparison was designed to avoid.

3. Method

3.1. Setup

Each WSI is a bag of frozen patch features and 2D coordinates,

$$X = \{(x_j, c_j)\}_{j=1}^{N}, \quad x_j \in \mathbb{R}^{1536},\ c_j \in \mathbb{R}^2,$$

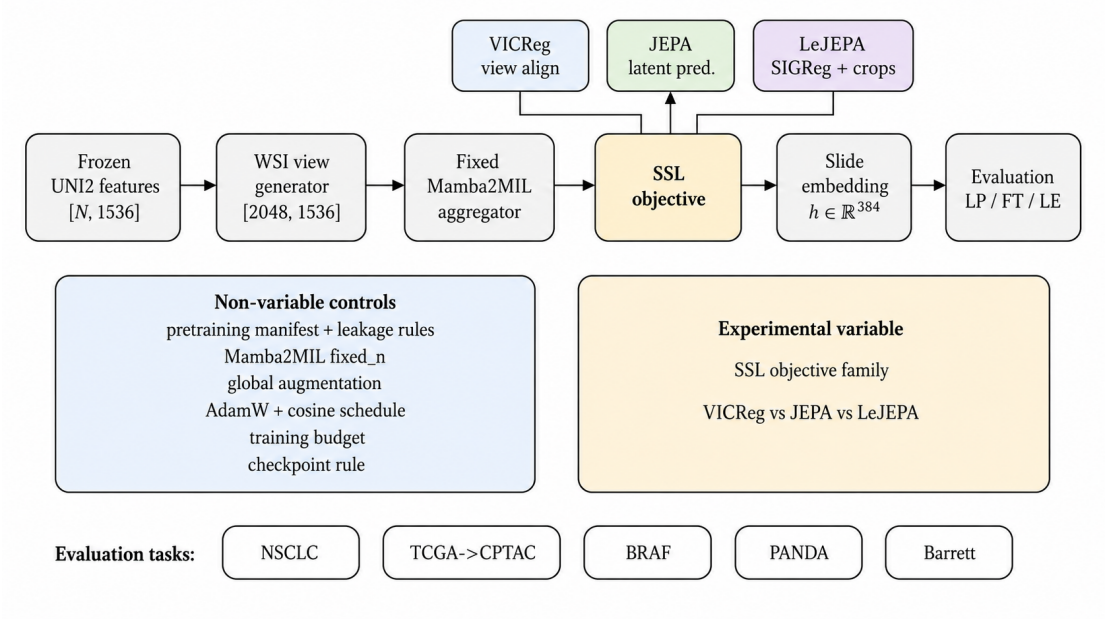

Figure 1. Controlled comparison. The patch encoder, aggregator, view generator, token budget, training schedule, and downstream protocol are fixed; only the SSL objective varies. Downstream evaluation uses the slide embedding h, not the projector output z.

and a slide encoder $f_\theta(X) = h \in \mathbb{R}^{384}$ produces a single slide embedding. SSL is performed in a higher-dimensional projector space $z = g(h)$ to give the loss room to spread features apart, but the projector g overfits to the SSL pretext and discards information that downstream tasks need; following standard practice in joint-embedding SSL, we therefore use h (not z) for all downstream evaluation, and discard g after pretraining.

3.2. Aggregator

All methods share the same MAMBA2MIL aggregator. Frozen UNI2 patch features are projected to 384 dimensions, ordered along a Hilbert curve over patch coordinates, processed by Mamba-2 (Dao & Gu, 2024) sequence blocks, and pooled with gated attention:

$$h = \mathrm{GatedAttn}\big(\mathrm{Mamba2}\big(\mathrm{HilbertSort}(W_{\mathrm{in}}\, x,\, c)\big)\big)\,.$$

Hilbert ordering is a space-filling curve that maps 2D coordinates to a 1D index while preserving local adjacency: patches close on the slide remain close in the sequence. This matters because Mamba-2 is a 1D selective state-space model with linear-time complexity in sequence length, and naive raster or random orderings would break the spatial structure that downstream tasks rely on. Hilbert ordering is deterministic given the coordinates, so the same patches produce the same sequence at training and inference time.

3.3. Feature-Space Views

Augmentations operate on patch embeddings and coordinates, not pixels, since the patch encoder is frozen and re-extraction would be prohibitive. For each slide we cap the bag at 12,000 patches by spatial-stratified subsampling—partition the slide into a coarse grid and sample evenly across cells—to bound memory while preserving slide coverage; uniform random subsampling can over-sample dense tumor regions and miss peripheral tissue. Two global views are then produced by (i) shared D4 transforms (one of 8 rotations/reflections of the patch coordinate frame, applied identically to both views so they remain registered), (ii) partial-overlap splitting that draws two index sets from the bag with a controlled overlap fraction (0.4), (iii) spatial cropping of a contiguous bounding box, and (iv) random token masking. Each view is finalized to exactly 2048 real patches: if the post-mask pool has $\geq$2048 indices, sample without replacement; otherwise backfill from the post-crop, post-split, and full-slide pools in turn (Appendix B).

Zero-padding to a fixed length would couple the SSL loss to slide size—a small slide would contribute many padding tokens with no information, while a large slide would not—so we enforce a 2048-patch minimum on the manifest and never pad. LeJEPA-MC additionally samples V_l=2 local views with the same pipeline but a smaller token budget (1024) and tighter crop area range; these correspond to regional tissue regions rather than whole-slide views.

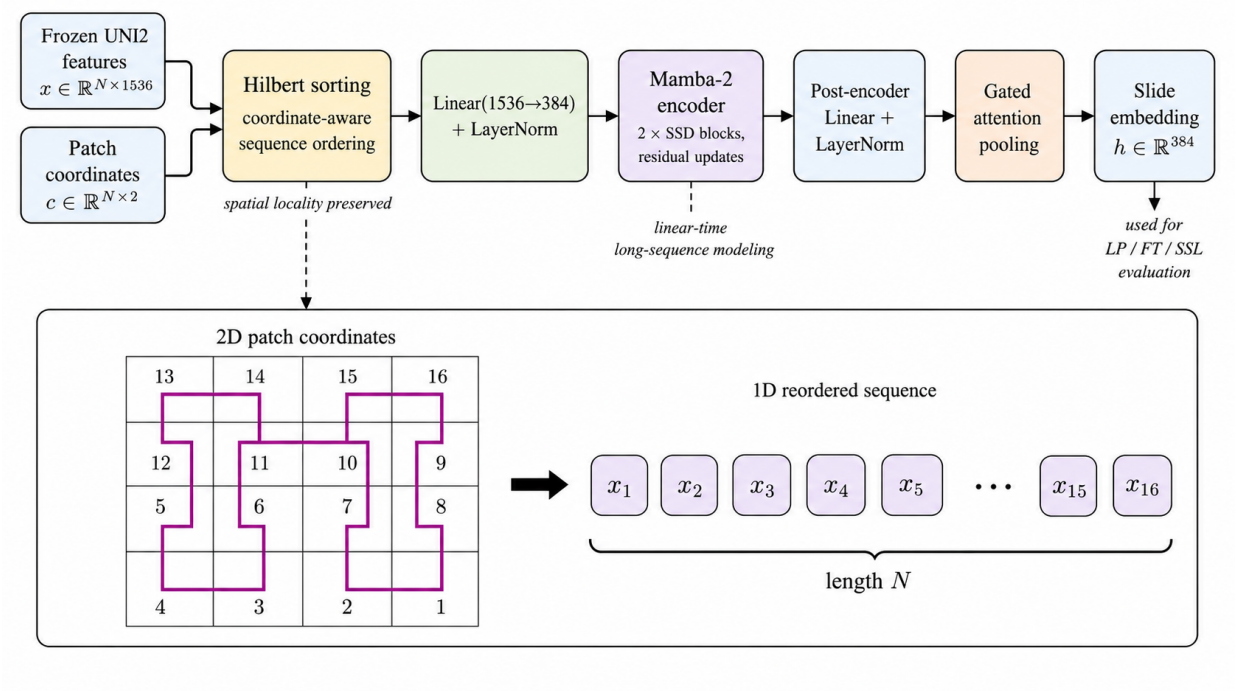

Figure 2. MAMBA2MIL aggregator. Frozen UNI2 patch features and coordinates are Hilbert-sorted into a locality-preserving sequence, projected, processed by Mamba-2 blocks, and pooled by gated attention. The pooled embedding h is the shared representation used for SSL pretraining and all downstream evaluation.

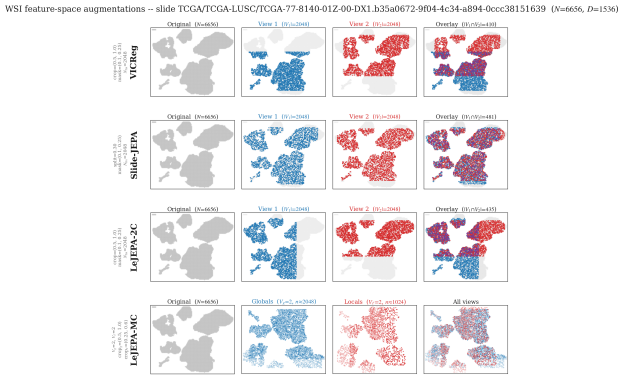

Figure 3. Feature-space WSI views. VICReg, JEPA, and LeJEPA-2C use two global views (2048 tokens). LeJEPA-MC adds two local views (1024 tokens, smaller crop area) that capture regional tissue structure.

3.4. Objectives

Let $z_r = g(f_\theta(v_r))$ denote a projected view embedding with projector g. Joint-embedding SSL methods all face the same hazard—representational *collapse*, where f_θ outputs the same vector regardless of input, trivially minimizing two-view agreement—and differ mainly in how they prevent it.

VICReg (Bardes et al., 2022) aligns two projected views with three explicit terms:

$$\mathcal{L}_{\text{VIC}} = \lambda_{\text{inv}}\mathcal{L}_{\text{inv}} + \lambda_{\text{var}}\mathcal{L}_{\text{var}} + \lambda_{\text{cov}}\mathcal{L}_{\text{cov}}.$$

$\mathcal{L}_{\text{inv}}$ is the mean-squared distance between matched views, $\mathcal{L}_{\text{var}}$ hinges each projector dimension's batch standard deviation above a threshold (preventing constant outputs), and $\mathcal{L}_{\text{cov}}$ penalizes off-diagonal projector covariance (preventing redundant dimensions). No negatives or EMA targets are used.

JEPA (Assran et al., 2023) predicts a target representation from a context representation, $p_c = q \circ g \circ f_\theta(v_c)$ vs. $z_t = g \circ f_{\bar\theta}(v_t)$, with SmoothL1 loss. The target encoder $f_{\bar\theta}$ shares architecture with f_θ but its weights are an exponential moving average (EMA) of f_θ, updated as $\bar\theta \leftarrow \tau\bar\theta + (1-\tau)\theta$ at each step. Gradients do not flow through $f_{\bar\theta}$ (stop-gradient on the target branch), which is what prevents collapse: the target is a slowly-moving anchor rather than a learnable copy of the context. We use a symmetric two-view variant in which each global view alternately plays the context and target role, doubling supervision per slide.

LeJEPA-2C (Balestriero & LeCun, 2025) dispenses with the EMA target and predictor altogether, predicting the projected centroid $\mu = \frac{1}{2}(z_1 + z_2)$ of two views regularized

by SIGReg:

$$\mathcal{L}_{\text{2C}} = (1-\lambda)\,\frac{1}{2}\sum_{r=1}^{2}\|z_r - \text{sg}(\mu)\|_2^2 + \lambda\,\frac{1}{2}\sum_{r=1}^{2}\text{SIGReg}(z_r),$$

where $\text{sg}(\cdot)$ denotes stop-gradient on the centroid target. SIGReg is a distributional regularizer that pushes the marginal distribution of z across the batch toward an isotropic Gaussian. This prevents collapse without needing negatives, EMA dynamics, or explicit variance/covariance terms: an isotropic Gaussian marginal automatically has full-rank covariance and non-zero variance in every direction.

LeJEPA-MC forms the centroid from *only* the V_g global views and trains *all* $V = V_g + V_l$ views (global and local) to predict it:

$$\mu_g = \frac{1}{V_g}\sum_{r=1}^{V_g} z_r,$$

$$\mathcal{L}_{\text{MC}} = (1-\lambda)\,\frac{1}{V}\sum_{r=1}^{V}\|z_r - \text{sg}(\mu_g)\|_2^2 + \lambda\,\frac{1}{V}\sum_{r=1}^{V}\text{SIGReg}(z_r).$$

The asymmetry is essential: keeping the prediction target at the global slide level while predicting from regional crops forces local features to remain consistent with whole-slide context, which is the structure WSI tasks tend to need (a small high-grade region must still be recognizable as belonging to a high-grade slide). If local views contributed to the target, the centroid would drift toward an average of regional content and the global signal would weaken.

4. Experimental Setup

4.1. Datasets and Patch Extraction

We use the publicly released UNI2-h slide-level patch features[1] from Chen et al. (2024), which provides 1536-dimensional patch embeddings for whole-slide images tiled into non-overlapping 256×256 patches at 20× magnification (∼0.5 microns/pixel). Foreground patches are detected with the HEST tissue segmentation model (Jaume et al., 2024a), a DeepLabV3 with a ResNet-50 backbone fine-tuned on annotated H&E and IHC tissue regions, which is more robust than Otsu thresholding to pen marks, fiducials, and stain artifacts. Using the public release directly fixes the patch encoder, tile selection, and feature extraction across all our experiments. The cached features are reused across all SSL methods and all downstream tasks, so any difference between methods reflects the slide-level model only.

Pretraining corpus. TCGA + CPTAC slides with ≥2048 patches, totaling 14,529 slides and 142.7M patches across 37 cancer types (Appendix A). The CPTAC NSCLC test split, the in-house GEJ/Barrett's cohort, and the PANDA challenge slides are all held out from pretraining.

[1] huggingface.co/datasets/MahmoodLab/UNI2-h-features

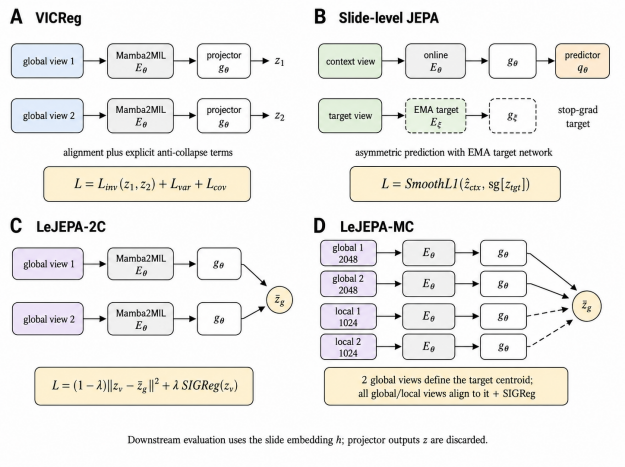

Figure 4. The four SSL objectives under the same MAMBA2MIL aggregator. VICReg aligns two projected views with variance/covariance regularization; JEPA predicts an EMA target; LeJEPA-2C predicts a two-view centroid with SIGReg; LeJEPA-MC extends the centroid target to global/local crops. The projector output z is discarded for downstream evaluation.

NSCLC subtyping. TCGA-LUAD vs TCGA-LUSC, binary classification at the slide level: 946 slides from 946 patients (468 LUAD, 478 LUSC), under patient-grouped 5-fold cross-validation.

NSCLC cohort transfer. Train on TCGA-LUAD/LUSC, evaluate AUROC on a held-out CPTAC-LUAD/LSCC test set of 453 slides from 453 patients (241 LUAD, 212 LSCC), excluded from SSL pretraining.

BRAF mutation prediction. TCGA-COAD/READ, slide-level binary classification of BRAF mutation status from H&E: 566 slides from 559 patients (423 COAD, 143 READ), with a 10.6% BRAF-mutant prevalence (60 positive, 506 negative). The label imbalance is reflected in higher AUROC variance for this task.

PANDA. The public PANDA prostate biopsy challenge (Bulten et al., 2022), used with the official Kaggle training-data split: 10,615 slides (5,455 from Karolinska, 5,160 from Radboud) with ordinal ISUP grade labels (0–5). Grade distribution: 2,891 / 2,666 / 1,343 / 1,242 / 1,249 / 1,224 across grades 0–5.

GEJ/Barrett's. Private cohort of 744 gastroesophageal junction biopsies from 305 cases, graded into 4 ordinal severity classes: BE-ND (305), BE-indefinite (111), BE-

LGD (113), and BE-HG/intramucosal carcinoma (215). The class distribution is moderately imbalanced rather than heavily skewed—the most common class is roughly $2.7\times$ the smallest. Collected under institutional review with patient consent.

4.2. SSL Pretraining

All four objectives share the same training recipe except for the loss-specific hyperparameters listed in Table 1. We use AdamW (Loshchilov & Hutter, 2019) with weight decay 0.05, gradient clipping at norm 1.0, a 10-epoch linear warmup followed by cosine decay over the remaining 90 epochs, bf16 mixed precision, and a global batch size of 256 slides. Pretraining runs on $2\times$ NVIDIA RTX 3090 GPUs, requiring approximately 10 GPU-hours per run.

Table 1. Objective-specific pretraining settings. Learning rates and loss-specific coefficients were chosen from pretraining-only pilot runs using SSL stability and representation diagnostics, not downstream labels.

Objective	LR	Key settings
VICReg	3×10^{-4}	inv/var/cov $= 25/25/4$
JEPA	2×10^{-4}	EMA target, SmoothL1
LeJEPA-2C	5×10^{-4}	SIGReg λ=0.05, V_g=2
LeJEPA-MC	5×10^{-4}	SIGReg λ=0.05, V_g=2, V_l=2

The SSL projector g is a 3-layer MLP with hidden dimension 1024 and output dimension 1024, followed by batchnorm; this is shared across all four objectives. JEPA additionally uses a 2-layer predictor q on the context branch with hidden dimension 512. The EMA decay for JEPA's target encoder is $\tau = 0.996$, increased to 1.0 over training following standard practice. Checkpoints are saved every epoch; downstream evaluation uses the final-epoch checkpoint without early stopping or validation-set selection.

Hyperparameters were selected using only pretraining-time information. Specifically, we ran a small pilot sweep over learning rates in $\{2, 3, 5\} \times 10^{-4}$ and selected, for each objective, the largest learning rate that produced stable SSL optimization without collapse or divergence. Stability was assessed on the held-out pretraining validation split using SSL loss curves and representation diagnostics, including embedding norm, per-dimension standard deviation, RankMe, and α-ReQ. Loss-specific coefficients were selected under the same pretraining-only criterion: for example, the VICReg covariance weight was fixed to 4 because it gave stable covariance control without overwhelming the invariance and variance terms. No downstream labels, downstream validation metrics, or task-specific checkpoints were used during SSL hyperparameter selection; after the pilot stage, all downstream experiments used the final-epoch checkpoint.

4.3. Downstream Tasks

Five tasks span the relevant axes: *NSCLC* subtyping (TCGA-LUAD/LUSC, within-cohort morphology); *TCGA→CPTAC* NSCLC transfer (cohort shift); *BRAF* mutation prediction in TCGA-COAD/READ (molecular); *PANDA* prostate ISUP grading (public ordinal); and *GEJ/Barrett's* severity grading (private clinical, ordinal). Binary tasks report AUROC; ordinal tasks report quadratic weighted kappa (QWK). Secondary metrics (AUPRC, macro-F1) are in the appendix.

4.4. Evaluation Protocols

Linear probing (LP) freezes f_θ, extracts slide embeddings h once, and trains an ℓ_2-regularized logistic head (binary tasks) or an ordinal cumulative-link model (PANDA, GEJ). Regularization strength is chosen by inner cross-validation within each outer fold. Folds are patient-grouped: all slides from a single patient are assigned to the same fold, preventing slide-to-slide leakage that would otherwise inflate scores when patients have multiple slides. We use 5-fold patient-grouped CV; reported standard deviations are across folds.

Fine-tuning (FT) updates the full aggregator and a task head with the same fold structure, using the SSL checkpoint as initialization. We fine-tune for 80 epochs with AdamW at learning rate 1×10^{-4} and a 2-epoch warmup; the same schedule is used for every initialization to keep the comparison fair. Each fold is repeated with 3 random seeds and we report the mean$\pm$std across folds and seeds.

Low-data probing (LD) matches the LP protocol but trains the linear head on a 1%, 10%, or 100% subset of labeled training slides within each fold, sampled uniformly at random with stratification on the label. We average over folds and 3 random seeds for each fraction.

Statistical comparison. Where we report $+\Delta$ between methods (e.g., LeJEPA-MC vs. Mean+Max), the difference is computed paired across folds. Because each task has only 5 outer folds, per-task paired tests have limited power and are not the primary basis for our claims. We therefore emphasize effect sizes and consistency of qualitative patterns across tasks and protocols: static pooling is strongest on saturated in-domain morphology, random MAMBA2MIL generally underperforms pooling, JEPA is weak, and the LeJEPA variants are the strongest overall, with LeJEPA-MC giving the clearest gains on shifted, ordinal, and clinically heterogeneous endpoints.

Baselines. We compare four SSL initializations (VICReg, JEPA, LeJEPA-2C, LeJEPA-MC) against three pooling baselines that operate directly on the frozen UNI2 patch features (mean, max, mean+max), an MAMBA2MIL aggregator with random initialization (no SSL), and—for fine-tuning only—

a supervised ABMIL (Ilse et al., 2018) baseline. The pooling baselines isolate how much of the gain comes from the patch encoder alone, while random MAMBA2MIL isolates how much comes from the aggregator architecture without learned weights.

5. Results

5.1. Linear Probing

Table 2 shows three patterns. (i) Static pooling is strong but not dominant: it is best on no endpoint, although Mean+Max remains highly competitive on saturated within-cohort morphology. On NSCLC, Mean+Max reaches 0.977 AUROC, only 0.006 below the best LP representation, LeJEPA-2C. Under cohort shift, however, Mean+Max reaches only 0.944 on TCGA→CPTAC, while LeJEPA-MC reaches 0.974. This separates the SSL story from a simple "frozen features are enough" story. (ii) The LeJEPA variants are the strongest SSL representations overall. JEPA is consistently weak, VICReg improves on JEPA on all five endpoints, and LeJEPA-2C improves on VICReg on all five endpoints. LeJEPA-MC improves over LeJEPA-2C on four of five endpoints, but not on saturated NSCLC, where LeJEPA-2C is slightly higher (0.983 vs. 0.981). (iii) The multi-crop extension is most useful away from the saturated in-domain task: relative to LeJEPA-2C, LeJEPA-MC changes NSCLC by -0.002, but improves BRAF by $+0.017$, TCGA→CPTAC by $+0.015$, PANDA by $+0.017$, and GEJ by $+0.020$. Thus the realistic claim is not that multi-crop universally dominates, but that it helps most on harder, shifted, ordinal, or clinically heterogeneous endpoints.

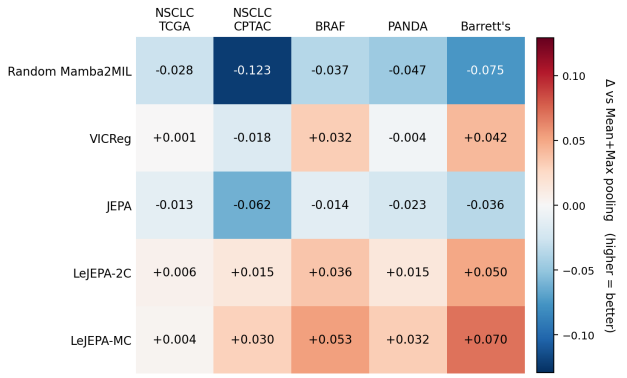

Figure 5. Improvement over Mean+Max pooling in full-label LP. Random MAMBA2MIL is uniformly negative, JEPA mostly negative, VICReg mixed; the LeJEPA variants are uniformly positive, with the largest gains on harder tasks.

Where the gains land. The size of the gain is task-dependent in a way that matches the claim that SSL helps when slide-level aggregation is informative. NSCLC subtyping is near saturation under UNI2 features: LeJEPA-MC improves over Mean+Max by only $+0.004$, and LeJEPA-2C is the best method by a small margin. Larger gains appear on BRAF mutation prediction ($+0.053$ over Mean+Max), where the H&E signal of a molecular phenotype is diffuse and weak; on TCGA→CPTAC transfer ($+0.030$); on PANDA grading ($+0.032$ QWK), where ISUP grade depends on the highest-grade region rather than on the average; and on GEJ ($+0.070$ QWK), the most clinically realistic and most label-scarce of the cohorts. The pattern is consistent: SSL aggregation helps most when the relevant signal is non-localized, distribution-shifted, ordinal, or label-scarce.

PANDA context. The PANDA result should be interpreted as an ordinal representation stress test rather than a prostate-specific leaderboard claim. LeJEPA-MC reaches 0.847 QWK using frozen slide embeddings and a lightweight downstream head, which is below prostate-specialized systems optimized directly for PANDA-style grading; for example, a recent hierarchical Vision Transformer reports 0.916 QWK on the PANDA test set (Grisi et al., 2025). However, our score is close to the externally evaluated PANDA algorithms reported by Bulten et al. (2022), which achieved QWKs of 0.862 and 0.868 on independent US and European validation sets, and falls within the range of public algorithm/pathologist agreement reported in later external evaluations (Faryna et al., 2024). We therefore use PANDA as an out-of-distribution ordinal benchmark for slide embeddings, not as a claim of state-of-the-art prostate grading.

Figure 5 summarizes where SSL helps over static pooling. Random MAMBA2MIL underperforms Mean+Max on every task, so architecture alone does not explain the gains. JEPA is mostly negative relative to Mean+Max, VICReg is mixed, and the LeJEPA variants are positive on the harder endpoints. NSCLC TCGA remains close to saturation; the larger gains are on BRAF, transfer, PANDA, and GEJ.

Random aggregation hurts. A non-trivial finding is that random-initialized MAMBA2MIL *underperforms* the best static-pooling baseline on all five tasks, by 0.069 on average. Relative to Mean+Max specifically, the average deficit is 0.062. The Mamba-2 + gated-attention architecture is therefore not by itself responsible for the SSL gains; without learned weights the aggregator destroys signal that simple pooling preserves. This is the strongest control against the alternative explanation that the LeJEPA-MC win is only an architecture effect.

5.2. Fine-Tuning

Fine-tuning broadly matches the LP pattern, but not as a strict total ordering. The LeJEPA variants remain the strongest initializations overall. LeJEPA-MC is best

Table 2. Full-label linear probing. Binary tasks: AUROC. Ordinal tasks: QWK. Best in **bold**, second-best underlined.

Representation	NSCLC AUROC	BRAF AUROC	TCGA→CPTAC AUROC	PANDA QWK	GEJ QWK
Mean Pool	0.971 ±.013	0.788 ±.061	0.940	0.838 ±.013	0.551 ±.073
Max Pool	0.965 ±.019	0.764 ±.094	0.927	0.789 ±.015	0.575 ±.085
Mean+Max Pool	0.977 ±.010	0.776 ±.075	0.944	0.815 ±.014	0.589 ±.079
Random MAMBA2MIL	0.949 ±.016	0.739 ±.087	0.821	0.768 ±.017	0.514 ±.063
VICReg	0.978 ±.011	0.808 ±.068	0.926	0.811 ±.012	0.631 ±.057
JEPA	0.964 ±.014	0.762 ±.091	0.882	0.792 ±.016	0.553 ±.071
LeJEPA-2C	**0.983 ±.008**	0.812 ±.058	0.959	0.830 ±.011	0.639 ±.054
LeJEPA-MC	0.981 ±.009	**0.829 ±.054**	**0.974**	**0.847 ±.010**	**0.659 ±.048**

Table 3. Fine-tuning. Primary: AUROC (binary) / QWK (GEJ). Secondary: AUPRC (binary) / macro-F1 (GEJ). NSCLC also reports external CPTAC AUROC after TCGA training.

Task	Initialization	Primary	Secondary	External AUROC
BRAF	ABMIL scratch	0.835 ±.057	0.461 ±.135	–
	MAMBA2MIL scratch	0.844 ±.083	0.415 ±.153	–
	JEPA	0.847 ±.069	0.421 ±.161	–
	VICReg	0.861 ±.055	0.495 ±.140	–
	LeJEPA-2C	0.873 ±.052	0.482 ±.147	–
	LeJEPA-MC	**0.888 ±.045**	**0.510 ±.128**	–
GEJ	ABMIL scratch	0.685 ±.058	0.575 ±.064	–
	MAMBA2MIL scratch	0.711 ±.071	0.571 ±.098	–
	JEPA	0.717 ±.086	0.605 ±.089	–
	VICReg	0.731 ±.078	**0.654 ±.054**	–
	LeJEPA-2C	0.741 ±.068	0.629 ±.059	–
	LeJEPA-MC	**0.755 ±.062**	0.651 ±.047	–
NSCLC	ABMIL scratch	0.977 ±.011	0.978 ±.013	0.965 ±.009
	MAMBA2MIL scratch	0.981 ±.014	0.982 ±.010	0.934 ±.026
	JEPA	0.980 ±.010	0.983 ±.009	0.951 ±.019
	VICReg	0.983 ±.009	0.984 ±.008	0.966 ±.013
	LeJEPA-2C	**0.985 ±.007**	0.985 ±.006	0.962 ±.015
	LeJEPA-MC	0.984 ±.008	**0.987 ±.005**	**0.970 ±.011**

on BRAF primary AUROC (0.888), GEJ primary QWK (0.755), NSCLC AUPRC (0.987), and external CPTAC transfer (0.970). LeJEPA-2C is slightly best on saturated NSCLC in-domain AUROC (0.985 vs. 0.984 for LeJEPA-MC), and VICReg is best on GEJ macro-F1 (0.654 vs. 0.651 for LeJEPA-MC). The realistic conclusion is therefore that LeJEPA-MC is the strongest overall initialization, especially for transfer and harder endpoints, but it does not win every secondary metric.

On BRAF, LeJEPA-MC improves over scratch MAMBA2MIL by +0.044 AUROC and +0.095 AUPRC. On GEJ, it improves over scratch MAMBA2MIL by +0.044 QWK and +0.080 macro-F1. On NSCLC, the in-domain gain over scratch MAMBA2MIL is small because the task is saturated (+0.003 AUROC, +0.005 AUPRC), but the external-transfer gain is much larger (+0.036 AUROC).

The cohort-transfer story. The most striking number is the external CPTAC AUROC under TCGA training. Scratch MAMBA2MIL reaches 0.981 on TCGA in-domain but drops to 0.934 on CPTAC, a 4.7-point gap that quantifies how much the supervised model has overfit to TCGA-specific tissue presentation. LeJEPA-MC initialization closes most of this gap: in-domain reaches 0.984, while external CPTAC reaches 0.970, leaving only a 1.4-point gap. SSL pretraining therefore does not primarily improve *in-domain* fine-tuning—which is saturated—but improves the *robustness* of the resulting fine-tuned model under cohort shift. A second pattern reinforces this: scratch ABMIL (0.965 external) beats scratch MAMBA2MIL (0.934 external) despite weaker in-domain performance, suggesting the simpler aggregator is harder to overfit. SSL initialization recovers the expressive aggregator's representational power without paying the same overfitting cost.

LP→FT consistency. The LP and FT results agree at the level that matters: JEPA is weak, VICReg is better, and the LeJEPA variants are strongest. The exact ordering is not perfectly preserved because NSCLC is near ceiling and small differences are within the fold-to-fold variation. Still, the same qualitative signal appears in both protocols: SSL pretraining improves the aggregator most on BRAF, GEJ, and external transfer, while in-domain NSCLC leaves little room for improvement.

5.3. Low-Data Probing

The low-data results are strongest for the LeJEPA family, but the winner is not always LeJEPA-MC. On BRAF at 1% labels, VICReg is best (0.658), with LeJEPA-MC second (0.652). On GEJ at 1% labels, LeJEPA-2C is best (0.247), with LeJEPA-MC second (0.243). LeJEPA-MC is best at 10% and 100% labels on both tasks, so the realistic conclusion is that LeJEPA-MC is the strongest low-data representation once a modest number of labels is available, while other SSL objectives can slightly edge it in the extreme 1% regime.

Relative to the best static baseline, LeJEPA-MC improves BRAF by $+0.061$ AUROC at 1% labels, $+0.065$ at 10%, and $+0.041$ at 100%. On GEJ, it improves by $+0.094$ QWK at 1%, $+0.085$ at 10%, and $+0.070$ at 100%.

Scaling pattern. The two tasks show different label-efficiency curves. On BRAF, the LeJEPA-MC gap to the best static baseline is largest at 10% labels ($+0.065$), similar at 1% ($+0.061$), and smaller at 100% ($+0.041$). This suggests that at the extreme 1% setting the bottleneck is partly label noise and class imbalance, not only representation quality. On GEJ, the gap widens as labels shrink ($+0.070 \rightarrow +0.085 \rightarrow +0.094$ from 100% to 10% to 1%), which is the practically valuable pattern: the harder and more label-scarce the clinical task is, the more SSL aggregation helps.

LeJEPA-MC at 10% labels does *not* fully match the best static baseline at 100% labels. On BRAF it reaches 0.747 vs. 0.788 for the best 100% static baseline; on GEJ it reaches 0.456 vs. 0.589. Therefore the correct claim is not an order-of-magnitude recovery of label efficiency. The stronger and safer claim is that SSL substantially narrows the low-label gap and gives the largest gains in the hardest label-scarce settings.

Comparison with random aggregation. Random MAMBA2MIL not only loses to static pooling at 100% labels but performs poorly at 1% on GEJ (0.103 QWK, near zero), while every SSL initialization remains positive. SSL pretraining is therefore not a marginal regularizer that only helps in plentiful-label regimes; it is what makes the aggregator usable when labels are scarce.

5.4. Spatial Attention

Figure 6 compares fine-tuned attention on a representative GEJ disagreement case. The slide was selected manually from the validation pool by a human reviewer, drawn from cases where the random-initialization model and the LeJEPA-MC model produced different predicted classes; we present it as a qualitative illustration rather than a systematic study. Inference is rerun deterministically (no random crop, mask, D4 transform, or token sampling) and patch attention is read directly from the gated-attention pooling layer. The random initialization spreads attention broadly, whereas LeJEPA-MC concentrates it on fewer regions and predicts the correct higher-grade class. This is consistent with the quantitative GEJ gain and suggests SSL initialization biases the supervised model toward more selective spatial aggregation, though attention is not a causal explanation, the case was not algorithmically sampled, and a single example cannot establish a mechanism.

6. Discussion

Three takeaways. First, static pooling over modern frozen patch features is a genuinely strong, high-floor baseline; it should be reported in any slide-level SSL study. On saturated tasks like NSCLC subtyping, it performs near the ceiling, and the gains from even the best SSL methods fall within the margin of error. Second, objective design dictates performance at the aggregation stage: under identical architecture, data, and training budgets, JEPA does not consistently improve over static pooling, VICReg only partially does, while LeJEPA-2C and LeJEPA-MC consistently dominate. The mechanism is the form of the collapse-prevention term—an EMA stop-gradient anchor (JEPA) and conditioning regularizers (VICReg) are weaker priors than a marginal-distribution constraint (SIGReg) when the input population is as heterogeneous as a 37-cancer-type WSI corpus. Third, multi-crop prediction is exceptionally useful for WSIs because the global/local asymmetry forces the model to reconcile local focal extremes with the global slide context. This aligns perfectly with WSI tasks like ordinal grading, where the clinical label is dictated by a small, localized region of high-grade dysplasia rather than the tissue average.

Where SSL helps and where it does not. The magnitude of the SSL gain tracks directly with clinical and spatial task difficulty: it is statistically marginal on saturated within-cohort morphology, but transformative on molecular phenotypes, ordinal grading, cohort-shift transfer, and label-scarce regimes. We interpret this fundamentally as a regularization and generalization story rather than purely a

Table 4. Low-data linear probing on BRAF (AUROC) and GEJ (QWK), averaged over 5 folds × 3 seeds. The 100% column matches Table 2.

	BRAF AUROC			GEJ QWK		
Representation	1%	10%	100%	1%	10%	100%
Mean Pool	0.591 ±.048	0.682 ±.041	0.788 ±.061	0.142 ±.041	0.337 ±.057	0.551 ±.073
Max Pool	0.551 ±.057	0.643 ±.034	0.764 ±.094	0.092 ±.061	0.346 ±.063	0.575 ±.085
Mean+Max	0.572 ±.049	0.673 ±.045	0.776 ±.075	0.149 ±.058	0.371 ±.068	0.589 ±.079
Random MAMBA2MIL	0.561 ±.079	0.656 ±.032	0.739 ±.087	0.103 ±.059	0.291 ±.052	0.514 ±.063
VICReg	**0.658 ±.045**	0.694 ±.027	0.808 ±.068	0.213 ±.045	0.368 ±.046	0.631 ±.057
JEPA	0.567 ±.061	0.663 ±.030	0.762 ±.091	0.118 ±.047	0.336 ±.055	0.553 ±.071
LeJEPA-2C	0.622 ±.043	0.722 ±.029	0.812 ±.058	**0.247 ±.040**	0.413 ±.044	0.639 ±.054
LeJEPA-MC	0.652 ±.039	**0.747 ±.026**	**0.829 ±.054**	0.243 ±.036	**0.456 ±.039**	**0.659 ±.048**

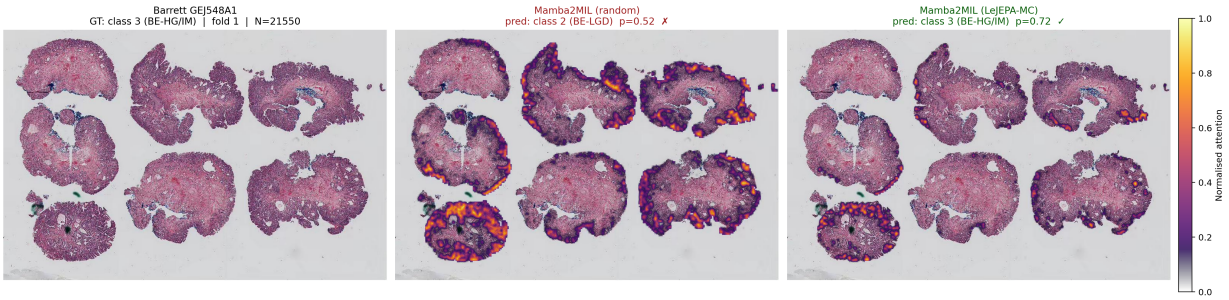

Figure 6. Gated-attention overlays on a GEJ validation slide (ground truth class 3, BE-HG/IM). The random-initialized MAMBA2MIL predicts class 2 (p=0.52) with diffuse attention across many fragments; LeJEPA-MC concentrates attention on a smaller subset and predicts class 3 correctly (p=0.72). Qualitative; not a causal explanation.

feature-extraction one. High-capacity aggregators trained from scratch suffer a severe overfitting penalty, heavily memorizing in-domain tissue presentation. SSL initialization acts as a powerful domain-robust regularizer, closing the generalization gap (as seen in the TCGA→CPTAC external transfer) and preventing model collapse. However, this complexity comes with a slight tradeoff: at the extreme 1% low-data limit, the complex optimization of multi-crop can be slightly edged out by simpler SSL objectives, suggesting that global/local alignment requires a minimum threshold of gradient updates to fully stabilize.

Limitations. The study uses one patch foundation model (UNI2), one aggregator (MAMBA2MIL), and one pretraining corpus (TCGA + CPTAC), and does not vary the augmentation policy across objectives. Although PANDA provides a public ordinal benchmark, our use of it is not directly comparable to prostate-specialized leaderboard systems that use task-specific supervised training, threshold tuning, ensembling, or test-time augmentation. We deliberately do not compare against contrastive methods (SimCLR, MoCo) or asymmetric self-distillation methods (BYOL, DINO), because their additional confounders—negative sets, stop-gradient predictors, prototype/centering schedules—would complicate the tightly controlled framing of this paper; comparison to those families is left to future work. Our standard deviations are across patient-grouped folds and three random seeds for low-data probing; with 5 folds, paired statistical tests have limited power per task, so we lean on consistency across 11 endpoints rather than per-endpoint significance. The spatial-attention analysis is qualitative and is not a causal explanation of the GEJ improvement; a quantitative attention-faithfulness study is a natural next step. Finally, an obvious ablation we did not run is a LeJEPA-2C with 4 global views to explicitly disentangle multi-crop diversity from the global/local asymmetry, nor did we explore tailored optimization schedules to stabilize multi-crop in the ultra-scarce 1% label regime.

7. Conclusion

Under a tightly controlled head-to-head evaluation, LeJEPA-MC proves to be the strongest SSL objective for slide-level aggregation over frozen pathology foundation features. It gives the strongest overall performance across linear probing, fine-tuning, and most low-label regimes, while consistently improving over static pooling and random aggregation on the harder, shifted, ordinal, and clinically heterogeneous endpoints. Rather than uniformly raising the ceiling on already-saturated tasks, the true value of distributionally

regularized global/local prediction lies in its ability to force local-to-global spatial alignment and heavily regularize high-capacity models against domain shift. Consequently, it delivers its largest and most critical gains on heterogeneous, out-of-distribution, and clinically complex tasks, offering a highly robust pathway for WSI representation learning without the prohibitive cost of retraining the underlying patch foundation model.

Impact Statement

This work studies self-supervised pretraining of slide-level aggregators over frozen pathology foundation features. The downstream tasks include cancer subtyping, molecular phenotype prediction, and clinical grading, all of which are decision-support tasks rather than autonomous diagnostic systems; any clinical use would require prospective validation and regulatory review. Pretraining uses only de-identified TCGA and CPTAC data and a private cohort collected under institutional approval. Because better label-efficient WSI representations could amplify both the benefits and the failure modes of automated histopathology, we believe the most consequential future considerations are dataset-shift robustness, calibration under cohort transfer, and equitable performance across patient subgroups—directions we view as essential follow-up work rather than addressed claims of this paper.

Code Availability

We release the clean OceanPath codebase for SSL pretraining, supervised MIL training, evaluation, statistical analysis, and model export at `https://github.com/Lyce24/OceanPath`. The repository includes Hydra configuration files, scripts for pretraining and downstream evaluation, SSL objectives and feature-space augmentations, MMap-backed data loading, model definitions, tests, and export utilities.

References

Assran, M., Duval, Q., Misra, I., Bojanowski, P., Vincent, P., Rabbat, M., LeCun, Y., and Ballas, N. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15619–15629, 2023.

Balestriero, R. and LeCun, Y. LeJEPA: Provable and scalable self-supervised learning without the heuristics. *arXiv preprint arXiv:2511.08544*, 2025.

Bardes, A., Ponce, J., and LeCun, Y. VICReg: Variance-invariance-covariance regularization for self-supervised learning. In *International Conference on Learning Representations (ICLR)*, 2022.

Bulten, W., Kartasalo, K., Chen, P.-H. C., Ström, P., Pinckaers, H., Nagpal, K., Cai, Y., Steiner, D. F., van Boven, H., Vink, R., Hulsbergen-van de Kaa, C. A., van der Laak, J., Amin, M. B., Evans, A. J., van der Kwast, T., Allan, R., Humphrey, P. A., Grönberg, H., Samaratunga, H., Delahunt, B., Srigley, J. R., et al. Artificial intelligence for diagnosis and Gleason grading of prostate cancer: the PANDA challenge. *Nature Medicine*, 28(1):154–163, 2022. doi: 10.1038/s41591-021-01620-2.

Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., and Joulin, A. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9650–9660, 2021.

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

Chen, R. J., Ding, T., Lu, M. Y., Williamson, D. F., Jaume, G., Chen, B., Zhang, A., Shao, D., Song, A. H., Shaban, M., Williams, M., Oldenburg, L., Weishaupt, L. L., Wang, J. J., Vaidya, A., Le, L. P., Gerber, G., Sahai, S., Williams, W., and Mahmood, F. Towards a general-purpose foundation model for computational pathology. *Nature Medicine*, 30(3):850–862, 2024. doi: 10.1038/s41591-024-02857-3.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, pp. 1597–1607. PMLR, 2020.

Dao, T. and Gu, A. Transformers are SSMs: Generalized models and efficient algorithms through structured state space duality. *arXiv preprint arXiv:2405.21060*, 2024.

Faryna, K., Tessier, L., Retamero, J., Bonthu, S., Samanta, P., Singhal, N., Kammerer-Jacquet, S.-F., Radulescu, C., Agosti, V., Collin, A., Farré, X., Fontugne, J., Grobholz, R., Hoogland, A. M., Leite, K. R. M., Oktay, M., Polonia, A., Roy, P., Salles, P. G., van der Kwast, T. H., van Ipenburg, J., van der Laak, J., and Litjens, G. Evaluation of artificial intelligence-based Gleason grading algorithms "in the wild". *Modern Pathology*, 37(11):100563, 2024. doi: 10.1016/j.modpat.2024.100563.

Filiot, A., Ghermi, R., Olivier, A., Jacob, P., Fidon, L., Mac Kain, A., Saillard, C., and Schiratti, J.-B. Scaling self-supervised learning for histopathology with masked image modeling. *medRxiv*, 2023. doi: 10.1101/2023.07.21.23292757.

Grill, J.-B., Strub, F., Altché, F., Tallec, C., Richemond, P. H., Buchatskaya, E., Doersch, C., Pires, B. A., Guo, Z. D., Azar, M. G., Piot, B., Kavukcuoglu, K., Munos, R., and Valko, M. Bootstrap your own latent: A new approach to self-supervised learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 21271–21284, 2020.

Grisi, C., Kartasalo, K., Eklund, M., Egevad, L., van der Laak, J., and Litjens, G. Hierarchical vision transformers for prostate biopsy grading: Towards bridging the generalization gap. *Medical Image Analysis*, 105:103663, 2025. doi: 10.1016/j.media.2025.103663.

He, K., Fan, H., Wu, Y., Xie, S., and Girshick, R. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9729–9738, 2020.

Hou, X., Jiang, C., Kondepudi, A., Lyu, Y., Chowdury, A., Lee, H., and Hollon, T. C. A self-supervised framework for learning whole slide representations. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

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

Jaume, G., Doucet, P., Song, A. H., Lu, M. Y., Almagro-Perez, C., Wagner, S. J., Vaidya, A. J., Chen, R. J., Williamson, D. F., Kim, A., and Mahmood, F. HEST-1k: A dataset for spatial transcriptomics and histology image analysis. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2024a.

Jaume, G., Vaidya, A., Chen, R. J., Williamson, D. F., Liang, P. P., and Mahmood, F. Modeling dense multimodal interactions between biological pathways and histology for survival prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024b.

Jaume, G., Vaidya, A., Zhang, A., Song, A. H., Chen, R. J., Sahai, S., Mo, D., Madrigal, E., Le, L. P., and Mahmood, F. Multistain pretraining for slide representation learning in pathology. In *European Conference on Computer Vision (ECCV)*, LNCS. Springer, 2024c.

Lenz, T., Neidlinger, P., Ligero, M., Wölflein, G., van Treeck, M., and Kather, J. N. Unsupervised foundation model-agnostic slide-level representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 30807–30817, 2025.

Loshchilov, I. and Hutter, F. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2019.

Lu, M. Y., Williamson, D. F., Chen, T. Y., Chen, R. J., Barbieri, M., and Mahmood, F. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nature Biomedical Engineering*, 5(6):555–570, 2021.

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

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

Vorontsov, E., Bozkurt, A., Casson, A., Shaikovski, G., Zelechowski, M., Severson, K., Zimmermann, E., Hall, J., Tenenholtz, N., Fusi, N., et al. A foundation model for clinical-grade computational pathology and rare cancers detection. *Nature Medicine*, 30(10):2924–2935, 2024.

Xu, H., Usuyama, N., Bagga, J., Zhang, S., Rao, R., Naumann, T., Wong, C., Gero, Z., Gonzalez, J., Gu, Y., Xu, Y., Wei, M., Wang, W., Ma, S., Wei, F., Yang, J., Li, C., Gao, J., Rosemon, J., Bower, D., Lee, S., Weerasinghe, M., Wright, B., Liu, A., Wang, X., Mermel, C., Chen, S., Hsu, C.-N., Xu, Z., et al. A whole-slide foundation model for digital pathology from real-world data. *Nature*, 630:181–188, 2024. doi: 10.1038/s41586-024-07441-w.

Yang, S., Wang, Y., and Chen, H. MambaMIL: Enhancing long sequence modeling with sequence reordering in computational pathology. In *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, 2024.

Zhang, H., Meng, Y., Zhao, Y., Qiao, Y., Yang, X., Coupland, S. E., and Zheng, Y. DTFD-MIL: Double-tier feature distillation multiple instance learning for histopathology whole slide image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18802–18812, 2022.

A. Pretraining Manifest

The pretraining manifest is built from UNI2 patch features extracted from TCGA and CPTAC slides with at least 2048 tissue patches.

Table 5. Manifest construction (left) and patch-count distribution after the 2048-patch minimum filter (right).

Field	Value
Cohorts	TCGA, CPTAC
H5 files scanned	28,000
Skipped (cohort)	10,615
Skipped (<2048)	2,856
Read errors	0
Kept slides	14,529

Statistic	Value
Total patches	142,733,112
Mean / median	9,824 / 8,340
Min / max	2,048 / 67,820
25th / 75th pct.	4,841 / 13,057

The 14,529 pretraining slides span 37 TCGA/CPTAC cancer types; the largest contributions are BRCA (1,453), GBM (953), LSCC (895), COAD (695), and LGG (646). PANDA, the in-house GEJ/Barrett's cohort, and CPTAC NSCLC test slides are excluded from SSL pretraining.

B. View Finalization

To avoid coupling the SSL loss to slide size via zero-padding, every view is finalized to exactly 2048 real tokens. Let M, C, S, F denote the post-mask, post-crop, post-split, and full-slide candidate index sets. If $|M| \geq 2048$ we sample 2048 indices from M without replacement. Otherwise we initialize $I \leftarrow M$ and add unique indices first from $C \setminus I$, then $S \setminus I$, then $F \setminus I$, until $|I| = 2048$. Because the manifest enforces $|F| \geq 2048$, the with-replacement fallback is unreachable. We log refill composition during training to monitor this.

C. Augmentation Configuration

Table 6. Augmentation parameters. Global: shared by VICReg, JEPA, LeJEPA-2C, and global views of LeJEPA-MC. Local: LeJEPA-MC only. Slide pre-cap: 12,000 patches, spatial-stratified.

Parameter	Global view	Local view
Tokens / view	2048	1024
D4 transform	shared	shared
Split overlap	0.4	0.4
Crop area range	$[0.50, 1.00]$	$[0.25, 0.60]$
Crop aspect range	$[0.75, 1.33]$	$[0.75, 1.33]$
Crop min keep frac	0.6	0.5
Mask ratio range	$[0.10, 0.30]$	$[0.05, 0.20]$
Mask min keep frac	0.65	0.65
Min tokens	256	128

D. Evaluation Details

Linear probing trains a regularized linear classifier on frozen slide embeddings h with patient-grouped cross-validation when patient identifiers are available. Fine-tuning updates the full MAMBA2MIL aggregator and a task head with the same fold structure. Low-data probing samples labeled subsets within each fold and averages over folds and seeds. SSL projector outputs are never used downstream.