MOOZY: A Patient-First Foundation Model for Computational Pathology

Yousef Kotp[1,2], Vincent Quoc-Huy Trinh[3,4], Christopher Pal[2,5,6], and Mahdi S. Hosseini[1,2,7]

[1] CSSE, Concordia University, Montréal, Canada
[2] Mila – Québec AI Institute, Montréal, Canada
[3] CHUM, Université de Montréal, Montréal, Canada
[4] IRIC, Université de Montréal, Montréal, Canada
[5] Université de Montréal, Montréal, Canada
[6] Canada CIFAR Chair, Polytechnique Montréal, Montréal, Canada
[7] Dept. of Pathology, McGill University, Montréal, Canada

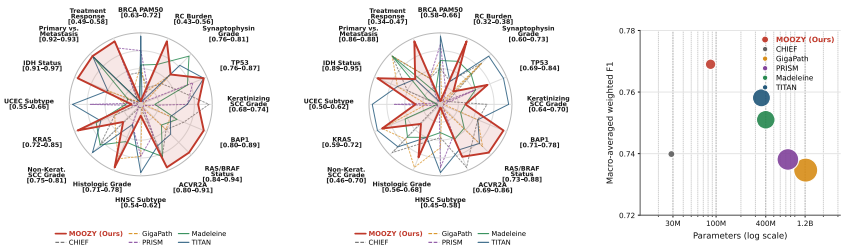
Code: `github.com/AtlasAnalyticsLab/MOOZY`

Abstract. Computational pathology needs whole-slide image (WSI) foundation models that transfer across diverse clinical tasks, yet current approaches remain largely slide-centric, often depend on private data and expensive paired-report supervision, and do not explicitly model relationships among multiple slides from the same patient. We present MOOZY, a patient-first pathology foundation model in which the patient case, not the individual slide, is the core unit of representation. MOOZY explicitly models dependencies across all slides from the same patient via a case transformer during pretraining, combining multi-stage self-supervision with scaled low-cost task supervision. In Stage 1, we pretrain a vision-only slide encoder on 77,134 public slide feature grids using masked self-distillation. In Stage 2, we align these representations with clinical semantics using a case transformer and multi-task supervision over 333 tasks from 56 public datasets, including 205 classification and 128 survival tasks across four endpoints. Across sixteen held-out tasks, MOOZY improves macro weighted F1, balanced accuracy, and macro weighted ROC-AUC relative to PRISM by +4.19%, +7.93%, and +6.95%, respectively. MOOZY is also parameter efficient with 85.77M parameters, 14× smaller than GigaPath. These results suggest that patient-level pretraining yields transferable embeddings, providing a path toward scalable patient-first histopathology foundation models.

Keywords: Computational Pathology · Whole-Slide Images · Foundation Models · Self-Supervised Learning · Slide Encoder

1 Introduction

The fundamental challenge in computational pathology is learning whole-slide image (WSI) representations that transfer across cancer types, clinical endpoints, and patient populations without task-specific retraining. Much of the field has



(a) **(b)** **(c)**

Fig. 1: **(a)** MLP-probe weighted F1 across sixteen held-out tasks. **(b)** Linear-probe weighted F1 across the same sixteen held-out tasks. In both (a) and (b), brackets report [min – max] weighted F1 per task, with the center corresponding to the minimum and the outer ring to the maximum observed value. **(c)** Macro-averaged weighted F1 (MLP probe) versus total parameter count (log scale).

historically advanced through task- and cohort-specific supervised pipelines [3, 9, 17, 45, 81] that must be rebuilt whenever the organ, scanner domain, or clinical objective changes, limiting scalability and reuse. A generalizable alternative requires models capable of encoding diagnostically relevant structure from unlabeled data at two levels of scale that are critical in clinical workflows: the intra-slide level, where diagnostic meaning emerges from long-range interactions across tissue regions rather than isolated patch morphology, and the patient level, where multiple slides from the same patient must be jointly interpreted to form predictions. Self-supervised learning in vision [12, 15, 16, 28, 31, 32, 67, 105] has demonstrated that scaling data and compute can yield general-purpose encoders with strong task-agnostic transferability [8,34,43,92], suggesting a similar paradigm shift is possible in pathology.

However, applying these methods to WSIs is non-trivial: WSIs are gigapixel-scale, and clinically relevant semantics arise from interactions between cellular detail and global tissue architecture. Consequently, early pathology foundation models emerged as *tile-level* encoders, with the field converging on the DINOv2 [67] pretraining recipe and scaling from ViT-Large [14, 26] to ViT-Giant backbones [60, 98] trained on up to millions of slides [89, 107]. In typical pipelines, tile features are extracted and a separate MIL aggregator [39, 56, 79] is trained for each downstream task, necessitating retraining whenever the clinical endpoint changes. More recently, the community has moved toward *slide-level* encoders that pretrain whole-slide representations, reducing reliance on task-specific MIL training. These can be broadly grouped into vision-only self-supervised methods [4, 13, 36, 48, 50, 99], multimodal approaches that align slides with text [20, 77, 96], genomic profiles [40, 87, 100], or cross-stain views [37, 41], and supervised methods that learn from task labels [63, 78, 91].

Despite this progress, structural limitations persist. Many top-performing models rely on proprietary data, and some do not release checkpoints [36, 87] or training recipes [20, 77, 87]. Current architectures concentrate capacity in heavy-

weight tile encoders [20,41,77,87] while using lightweight slide aggregators, even though the key challenge in WSIs is long-range context rather than per-tile morphology. Finally, while some methods accommodate multiple slides per patient, they typically rely on simple fusion heuristics: early fusion by concatenating or unioning patch features into one enlarged bag, or late fusion by averaging slide-level embeddings or predictions across slides [50,77,87,87,100]. These strategies treat a case as an unordered pool rather than explicitly modeling slide-to-slide relationships, discarding cross-slide interactions that carry diagnostic signal.

To this end, we introduce MOOZY, a *patient-first* model in the sense that the patient case, not the individual slide, serves as the fundamental unit of representation. Rather than encoding slides independently and merging their embeddings post-hoc, MOOZY explicitly models dependencies across all slides belonging to the same patient via a dedicated case transformer during pretraining. We adopt a lightweight patch encoder [42] that compresses gigapixel slides into feature grids, making slide-level SSL tractable. MOOZY decouples representation quality from task-specific adaptation where Stage 1 pretrains a slide encoder on unlabeled public whole-slide images via self-supervised learning, establishing general-purpose spatial representations without any label signal. Stage 2 then steers these representations toward clinical semantics through large-scale multi-task supervision, benefiting from the generalizable prior built in Stage 1.Stage 2 moves beyond per-slide encoding where a case-level aggregator explicitly models dependencies across all slides of the same patient, rather than collapsing multi-slide cases into a single bag or averaging independent predictions. To the best of our knowledge, this is the first method that jointly addresses slide-level representation learning and explicit patient-level inter-slide dependency modeling. Our contributions can be summarized as follows:

- We propose a two-stage framework that decouples vision-only slide SSL pretraining from patient-aware semantic alignment, where a case-level aggregator explicitly models dependencies across all slides of the same patient, making MOOZY the first attempt to move pathology foundational models beyond naive early/late multi-slide fusion.
- We construct a large-scale multi-task supervision regime spanning 333 tasks from 56 public datasets, covering classification and four survival endpoints across 23 anatomical sites, entirely from public data without private slides, paired reports, or expert annotations.
- We provide comprehensive quantitative and qualitative evaluation by benchmarking MOOZY on sixteen held-out tasks against both slide encoders and MIL baselines, complemented by attention map analysis and embedding visualization, showing competitive, transferable, and parameter-efficient representations.

2 Related Work

Pathology Patch Encoders. Pathology-specific SSL on public tiles outperforms ImageNet initialization [25,42], leveraging vision SSL advances [11,12,15,

16, 28, 31, 32, 67, 80, 102, 105]. The field has converged on the DINOv2 recipe [67] combining self-distillation, masked image modeling, KoLeo regularization [47, 74], and KDE-based objectives [90], with vision–language alignment [20, 55] and knowledge distillation [24] as complementary directions. Architectures have scaled from ViT-Large [14, 26, 101] to ViT-Huge [14, 89, 107] and ViT-Giant [6, 60, 75, 98] on proprietary [14, 60, 89] and public data [23, 25, 26, 29, 59]. Yet scaling laws for tile encoders remain unclear: models match much larger systems [44, 82], suggesting benchmark discriminability [94] and training-recipe effects dominate data volume. We hypothesize this saturation is fundamental as H&E tissue occupies a far more constrained visual space than natural images, with a narrow color palette and bounded set of morphological primitives (*e.g.*, cell types, glandular architectures, stromal patterns), so tile-level representations approach a performance ceiling well before general-vision thresholds. The true bottleneck therefore lies in *slide- and context-level modeling*.

Multi-Instance Learning. MIL treats a WSI as a bag of patch features with a single slide-level label, with approaches spanning permutation-invariant pooling [10], attention scoring [39, 56], transformer-based inter-patch modeling [79, 88], dual-stream objectives [51], pseudo-bag augmentation [104], and efficient variants via low-rank approximations [97], knowledge graphs [52], and regional re-embedding [84]. All aggregators are trained from scratch per task, motivating universal pretrained slide representations.

Slide Encoders. Slide-level pretraining operates on unordered sets of thousands of heterogeneous tile embeddings. Vision-only methods apply self-distillation [12, 13], contrastive tile sampling [15, 48], view transformations [36], dilated attention in masked autoencoders [19, 99], lightweight contextualizers [4], and state-space contrastive learning [30, 50]. Multimodal methods align slides with clinical text [20, 54, 77, 96], genomic or transcriptomic profiles [40, 87, 100], or cross-stain sections [37, 41]. Supervised approaches train on slide-level labels [63, 64, 78, 91]. Three gaps persist: many methods rely on expensive proprietary data [20, 77, 87], capacity concentrates in tile encoders over slide aggregators [20, 41, 77, 87], and multi-slide fusion remains naive [50, 77, 87, 100], treating cases as unordered pools. MOOZY directly addresses all three: a two-stage design decouples vision-only SSL pretraining from case-aware multi-task alignment, replacing naive multi-slide fusion with explicit inter-slide dependency modeling at the case level, while training with cheap available supervision.

3 Methodology

Stage 1: Self-Supervised Slide Encoder Pretraining. We cast WSI representation learning as self-supervised pretraining on precomputed patch features (Figure 2, up). Given a WSI $\mathcal{W}$, we partition tissue into non-overlapping 224-pixel patches and extract features with a frozen patch encoder f_{patch}:

$$\mathbf{z}_i = f_{\text{patch}}(p_i), \quad \mathbf{z}_i \in \mathbb{R}^{d_{\text{patch}}}. \tag{1}$$

We arrange patch features and coordinates into a 2-D grid $\mathbf{G} \in \mathbb{R}^{H \times W \times d_{\text{patch}}}$ with a binary validity mask for tissue positions (grid construction in Appendix C.1).

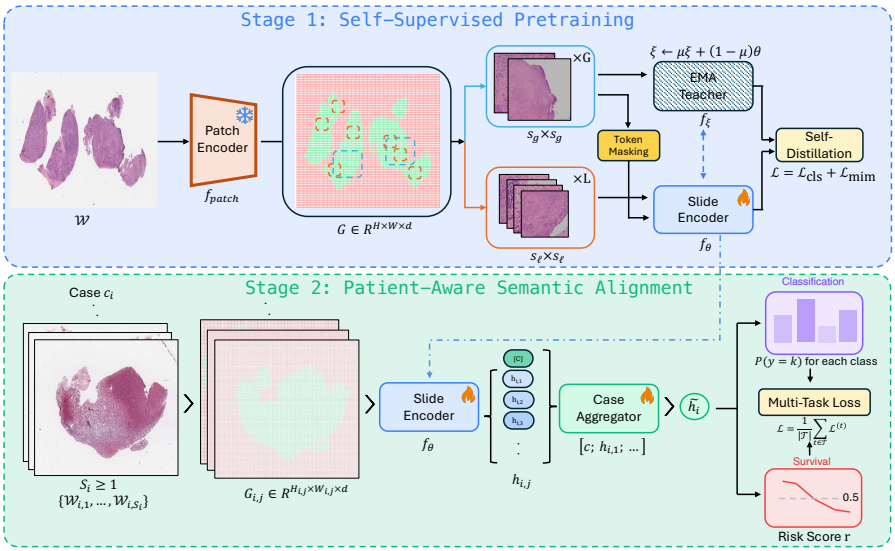


Fig. 2: Overview of the proposed two-stage framework. *Stage 1 (up):* A frozen patch encoder extracts per-patch features arranged into a spatial grid. Multi-scale crops are sampled with spatial augmentations and block-based masking. A student slide encoder and EMA teacher are jointly trained via CLS-level self-distillation ($\mathcal{L}_{\text{cls}}$) and masked patch prediction ($\mathcal{L}_{\text{mim}}$). *Stage 2 (down):* The pretrained slide encoder produces per-slide embeddings; a case transformer aggregates them into a unified case embedding $\tilde{\mathbf{h}}_i$, routed to task-specific classification and survival heads.

If a slide is available at multiple magnifications, each level-specific grid is treated as an independent training sample. To capture global context and local detail, we sample G global crops of size $S_g \times S_g$ and L local crops of size $S_l \times S_l$ ($S_g > S_l$) uniformly from valid grid locations, with each crop required to satisfy a minimum valid-token ratio $\rho_{\min}$. Unlike [20], which draws global and local views from the same fixed ROI, we sample crops independently over the full slide grid. This increases spatial diversity and lowers view mutual information, which benefits self-supervised WSI representation learning [36]. We apply DINOv3-style block masking [80] to global crops only. Because histopathology tissue is spatially continuous, contiguous masking encourages reasoning over broader morphology instead of reconstructing isolated tokens. In each batch, we select a fraction of global crops for masking, assign mask ratios uniformly over $[\gamma_{\min}, \gamma_{\max}]$, and shuffle them for uniform coverage of the masking range. We then iteratively place rectangular blocks with log-uniform aspect ratios until the target fraction of valid tokens is masked, restricting masking to tissue tokens. The full algorithm is in Appendix C.3.

Our slide encoder (Figure 3A) is a Vision Transformer [22] adapted to pre-computed feature grids. Patch features are projected to dimension d with a linear layer and GELU [33]. We prepend a learnable `[CLS]` token and R register

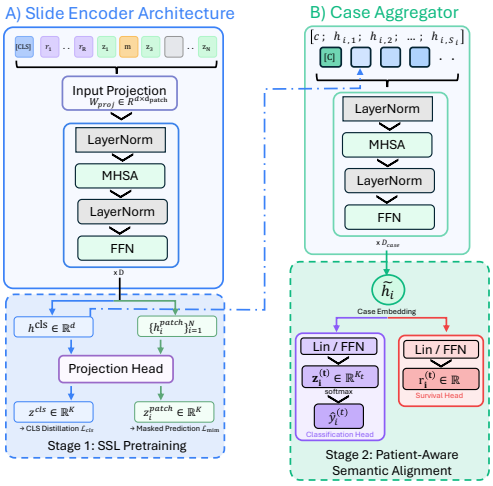


Fig. 3: Architecture of the slide encoder and case aggregator. **(A)** The slide encoder takes patch embeddings, a learnable `[CLS]` token, R register tokens, and mask tokens, processed through D transformer blocks. **(B)** The case aggregator prepends a learnable `[CASE]` token to per-slide embeddings and produces a case embedding $\tilde{\mathbf{h}}_i$, routed to heads for classification and survival prediction.

tokens [18], masked student positions are replaced by a learnable mask embedding. Each block uses pre-norm multi-head self-attention and an FFN with stochastic depth [38]. LayerScale [85] is disabled in our training runs. To encode spatial structure without learned positional embeddings, we use 2-D ALiBi [69] as adapted for WSIs in TITAN [20]. For each attention head h, we add:

$$b_{i,j}^{(h)} = -s_h \cdot \frac{\|\mathbf{p}_i - \mathbf{p}_j\|_2}{\Delta}, \tag{2}$$

where $\mathbf{p}_i$ are level-0 token coordinates, Δ is patch spacing, and $s_h > 0$ is a head-specific geometric slope. `[CLS]` and register tokens receive zero bias to remain spatially neutral. We also apply an additive attention mask that sets background-involving pairs to $-\infty$. The projection head maps encoder tokens to prototype logits with an MLP, an L2-normalized bottleneck, and a weight-normalized prototype layer [12, 105], shared for `[CLS]` and patch tokens (full formulation in Appendix C.6).

We use an EMA teacher for self-distillation, updating teacher parameters as $\xi \leftarrow \mu\xi + (1-\mu)\theta$ with cosine momentum schedule from μ_0 to μ_T. To avoid mode collapse, teacher outputs are centered with momentum-updated running averages [12]. The objective combines global CLS distillation and masked patch prediction. The teacher provides soft targets from global views, while the student

predicts from all views:

$$\mathcal{L}_{\text{cls}} = -\frac{1}{G(G+L-1)} \sum_{j=1}^{G} \sum_{\substack{i=1 \\ i \neq j}}^{G+L} \sum_{k=1}^{K} P_k^{(j)} \log Q_k^{(i)}, \tag{3}$$

where $P^{(j)}$ and $Q^{(i)}$ are teacher and student softmax distributions with temperatures τ_t and τ_s. For masked positions in global crops, the student additionally predicts teacher patch-level distributions:

$$\mathcal{L}_{\text{mim}} = -\frac{1}{|\mathcal{M}|} \sum_{(j,r,c) \in \mathcal{M}} \sum_{k=1}^{K} P_{r,c,k}^{(j)} \log Q_{r,c,k}^{(j)}, \tag{4}$$

where $\mathcal{M}$ is the set of masked valid positions, and patch distributions use temperature τ_t^{patch}. The total loss is $\mathcal{L} = \mathcal{L}_{\text{cls}} + \mathcal{L}_{\text{mim}}$.

Stage 2: Case-Aware Semantic Alignment. A central design principle of MOOZY is to decouple representation learning from semantic alignment: Stage 1 builds a general-purpose slide encoder on unlabeled data, and Stage 2 steers it toward clinical utility through multi-task supervision without re-learning spatial representations from task labels alone. This contrasts with task-specific MIL pipelines that learn both aggregation and task adaptation simultaneously from scratch, and with multimodal slide encoders that couple representation quality to the availability of paired text or genomic data. Concretely, we fine-tune the Stage 1 encoder with multi-task supervision across diverse clinical endpoints (Figure 2, down). Let $\mathcal{T} = \{\mathcal{T}_1, \ldots, \mathcal{T}_T\}$ be T supervised tasks. Each case c_i contains one or more WSIs $\{\mathcal{W}_{i,1}, \ldots, \mathcal{W}_{i,S_i}\}$, and each task provides either a class label (classification) or a time-to-event label with event indicator (survival). Stage 2 uses full-slide grids without crop sampling. To handle gigapixel inputs under GPU memory limits, we apply a hardware-adaptive token cap $K_{\max}(\cdot)$: if valid tokens exceed $K_{\max}$, we perform stratified random sampling to preserve whole-slide spatial coverage (algorithm in Appendix C.4). Retained tokens are compacted and passed to the slide encoder:

$$\mathbf{h}_{i,j} = f_\theta(\mathbf{X}_{i,j}^\star, \mathbf{P}_{i,j}^\star, \Delta_{i,j}) \in \mathbb{R}^d. \tag{5}$$

To form one case representation from slide embeddings $\mathbf{H}_i = \{\mathbf{h}_{i,1}, \ldots, \mathbf{h}_{i,S_i}\}$, we use a lightweight transformer aggregator (Figure 3B). A learnable `[CASE]` token is prepended and processed through D_{case} pre-norm transformer blocks with LayerScale, while DropPath is disabled, yielding:

$$\tilde{\mathbf{h}}_i = \text{LN}(\mathbf{z}_{i,D_{\text{case}}})[0] \in \mathbb{R}^d. \tag{6}$$

We apply this aggregator to all cases, including single-slide cases ($S_i = 1$), so that the learned embedding space is always case-centric and consistent regardless of slide count at inference. Each task $\mathcal{T}_t$ has a prediction head $g_t : \mathbb{R}^d \to \mathbb{R}^{o_t}$, either linear or MLP (formulations in Appendix C.7). For classification, we use weighted cross-entropy with label smoothing [83] coefficient ϵ:

$$\mathcal{L}_{\text{cls}}^{(t)} = \text{CE}_{w^{(t)},\epsilon}\left(\{\mathbf{z}_i^{(t)}, y_i^{(t)}\}_{i \in \mathcal{B}_t}\right), \tag{7}$$

where class weights use inverse frequency: $w_k^{(t)} = |\mathcal{D}_t|/(K_t \cdot |\{i \in \mathcal{D}_t : y_i^{(t)} = k\}|)$, and $\mathcal{B}_t$ is the valid labeled set. For survival prediction, we use a discrete-hazard objective: survival times are quantized into B_t bins with edges at training event-time quantiles, and B_t adapts to per-task event count. With predicted hazards $h_{i,k}^{(t)} = \sigma(a_{i,k}^{(t)})$, we minimize the negative log-likelihood over events and censored cases (full loss and bin-selection details in Appendix C.5). For ranking-based metrics, hazards are converted to scalar risk:

$$r_i^{(t)} = -\sum_{k=1}^{B_t} \log(1 - h_{i,k}^{(t)}). \tag{8}$$

Let $\mathcal{T}_{\text{active}} \subseteq \mathcal{T}$ be tasks with usable supervision in the current batch. We average losses over active tasks:

$$\mathcal{L} = \frac{1}{|\mathcal{T}_{\text{active}}|} \sum_{t \in \mathcal{T}_{\text{active}}} \mathcal{L}^{(t)}, \tag{9}$$

which naturally handles sparse multi-task labels by excluding unlabeled tasks for each case. At inference, the slide encoder and case transformer output $\tilde{\mathbf{h}}_i$ as the final case embedding, and task heads are discarded.

4 Experiment Setup

Dataset. We collect 56 different open-sourced datasets which include: REG Dataset [49], TCGA (all 32 cohorts) [59], CPTAC (all 10 cohorts) [23], BC-Therapy [76], BRACS [7], CAMELYON17 [2], DHMC Kidney [106], DHMC LUAD [93], EBRAINS [72, 73], IMP Colorectum [61, 62, 66], IMP Cervix [65], MBC [5, 27], MUT-HET-RCC [70], NADT Prostate [95], NAT-BRCA [68], and PANDA [9]. All collected slides are processed using AtlasPatch [1], which performs tissue segmentation using SAM2 [46,71] model. The resulting tissue masks define the valid regions from which non-overlapping 224×224 patches are extracted at both $20\times$ and $40\times$ magnification levels, reaching $\sim$1.6 billion extracted patches. We extract features for each patch using a pretrained lightweight patch encoder from [42], which has 21.67M parameters and the architecture of ViT-S [22]. Detailed patch counts at each magnification are reported in Appendix A.

For Stage 1 self-supervised pretraining, the preprocessing yields 77,134 slide feature grids: 53,286 at $20\times$ and 23,848 at $40\times$ magnification, sourced from $\sim$31.8 TB of raw WSI data. The two corresponding feature grids from those two magnification levels are treated as independent training samples and sampled uniformly. These slides span 23 distinct anatomical sites as shown in Figure 4.

For Stage 2 supervised fine-tuning, we construct 333 tasks (205 classification and 128 survival) across all 56 datasets, averaging 6 tasks per dataset. Survival supervision covers overall survival (OS), disease-specific survival (DSS), disease-free interval (DFI), and progression-free interval (PFI), depending on endpoint availability. Of these, 56 are slide-level and 277 are case-level (*i.e.*, predictions

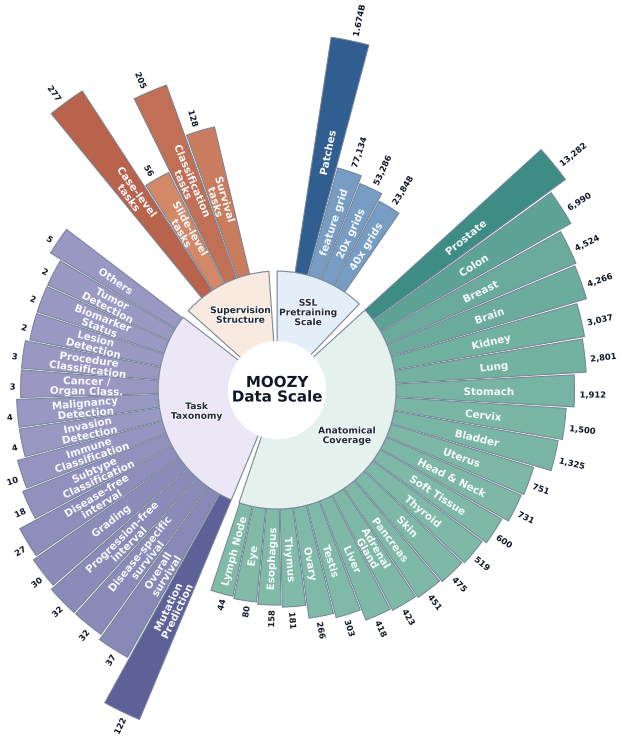


Fig. 4: Radial hierarchy of MOOZY data scale across four dimensions: pretraining scale, anatomical coverage, task taxonomy, and supervision structure.

aggregated over all slides of a case). The labeled Stage 2 subset comprises 41,089 supervised cases and 45,179 unique whole-slide images, a subset of the 53,286 Stage 1 slides, since unlabeled slides are excluded. The mean per-task slide-to-case ratio is 1.330[8]. Per-organ statistics, class counts, and task-construction details are provided in Appendices B and B.1 and Tab. 5. A small subset of tasks was drawn from prior work [87, 103]. For a broader landscape overview of computational pathology tasks, see [35]. A consolidated scale overview is shown in Figure 4.

SSL Pretraining. We train the slide encoder using the Stage 1 self-supervised framework (Sec. 3) on all 77,134 slide feature grids. Training uses 8 GPUs with an effective batch of 1,024 slides (micro batch 64, 2 accumulation steps) for 200 epochs (14,400 optimizer steps, ≈436 GPU-hours). Full hyperparameters are listed in Appendix E. The encoder is a 6-layer transformer (d=768, 12 heads, 4 register tokens [18]). Multi-crop sampling uses G=2 global crops of 20×20 tokens and L=4 local crops of 12×12 tokens, with block masking applied to global crops at mask ratio $\gamma \sim \mathcal{U}[0.1, 0.5]$. Optimization uses AdamW [53] with a cosine

[8] Computed over tasks as the mean of each task's unique-slide / unique-case ratio, not as the global union ratio.

learning rate schedule and an EMA teacher whose momentum follows a cosine schedule from μ_0=0.996 to μ_T=1.0.

Case-Aware Semantic Alignment. Following the Stage 2 framework (Sec. 3), training uses 8 GPUs with an effective batch of 1,024 cases (micro batch 1, 128 accumulation steps) for 20 epochs (1,000 optimizer steps, ≈512 GPU-hours). Complete hyperparameters are provided in Appendix F. For case-level aggregation, a case transformer (D_{case}=3 layers, 12 heads, learnable `[CASE]` token) pools all slides assigned to a supervised case into a single embedding. We jointly train on 205 classification and 128 survival tasks (333 total) spanning all 56 datasets. Classification tasks use cross-entropy with label smoothing (ϵ=0.03) and inverse-frequency class weighting. Survival tasks spanning OS, DSS, DFI, and PFI use a discrete-time hazard model with adaptive per-task binning (target 8 bins, minimum 2, maximum 16) and NLL loss. Optimization uses AdamW [53] with base learning rate 5×10^{-5}, cosine schedule, and gradient clipping at 0.3. A stratified 5% (task-wise) validation holdout is used to monitor overfitting. A summary of stage-specific augmentation strategies is provided in Appendix Table 10, and visual examples are provided in Appendix Figure 10.

5 Results

We evaluate on sixteen held-out tasks spanning diverse clinical settings. These cover Residual Cancer Burden (BC Therapy), TP53 mutation (CPTAC-BRCA), BAP1 mutation (CPTAC-CCRCC), ACVR2A mutation (CPTAC-COAD), Histologic Grade (CPTAC-LSCC), KRAS mutation (CPTAC-LUAD), IDH Status (EBRAINS), Treatment Response (MBC), molecular subtypes from TCGA-BRCA, TCGA-HNSC, and TCGA-UCEC, keratinizing and non-keratinizing squamous cell carcinoma grade and primary-versus-metastatic classification (HANCOCK) [21,103], and RAS/BRAF status and synaptophysin grade (VALENTINO) [86, 103]. These tasks are excluded from all training stages. Fifteen are case-level and IDH Status is slide-level[9]. Slide encoders use frozen-feature MLP probes, while MIL baselines train task-specific aggregators on frozen patch features. Both follow five-fold evaluation with case-level fold grouping. We report mean ± standard deviation for weighted F1, weighted ROC-AUC, and balanced accuracy. Full protocol details are in Appendix G. Linear probe results are in Appendix H and Tabs. 20 and 21.

Comparison with Slide Encoders. We compare MOOZY against five slide encoders. For case-level tasks, baseline encoders produce case representations by averaging per-slide embeddings before probing, while MOOZY uses its native case-level embedding from the case transformer. Results are in Table 1. Across all sixteen tasks, MOOZY achieves the strongest macro weighted F1 (0.769) and balanced accuracy (0.702), TITAN achieves the strongest weighted ROC-AUC (0.768 versus 0.763).

Parameter Efficiency. MOOZY totals 85.77M parameters (64.10M slide and case encoder + 21.67M patch encoder), making it 4–14× smaller than

[9] RC Burden is case-level task, but the slide to case ratio is 1.0

Table 1: Frozen-feature MLP probe comparison against slide encoder baselines on sixteen held-out tasks. **Bold**: best; underline: second best.

Task	Metric	CHIEF	Giga-Path	PRISM	Made-leine	TITAN	MOOZY (Ours)
Residual Cancer Burden	F1	$0.46_{\pm 0.03}$	$0.45_{\pm 0.05}$	$0.46_{\pm 0.07}$	$0.51_{\pm 0.03}$	$0.43_{\pm 0.07}$	$\mathbf{0.56}_{\pm 0.05}$
	AUC	$0.60_{\pm 0.05}$	$0.55_{\pm 0.08}$	$0.58_{\pm 0.06}$	$0.63_{\pm 0.03}$	$0.58_{\pm 0.07}$	$\mathbf{0.74}_{\pm 0.04}$
	Bal. Acc	$0.44_{\pm 0.05}$	$0.40_{\pm 0.05}$	$0.43_{\pm 0.04}$	$0.48_{\pm 0.03}$	$0.38_{\pm 0.11}$	$\mathbf{0.51}_{\pm 0.06}$
TP53 Mut.	F1	$0.82_{\pm 0.05}$	$0.76_{\pm 0.03}$	$0.85_{\pm 0.03}$	$0.84_{\pm 0.06}$	$\mathbf{0.87}_{\pm 0.04}$	$0.87_{\pm 0.04}$
	AUC	$0.81_{\pm 0.09}$	$0.76_{\pm 0.06}$	$0.85_{\pm 0.05}$	$0.85_{\pm 0.08}$	$\mathbf{0.91}_{\pm 0.04}$	$0.86_{\pm 0.06}$
	Bal. Acc	$0.83_{\pm 0.04}$	$0.76_{\pm 0.04}$	$0.84_{\pm 0.04}$	$0.84_{\pm 0.05}$	$\mathbf{0.88}_{\pm 0.04}$	$0.86_{\pm 0.05}$
BAP1 Mut.	F1	$0.86_{\pm 0.04}$	$0.84_{\pm 0.06}$	$0.80_{\pm 0.07}$	$0.85_{\pm 0.08}$	$0.84_{\pm 0.06}$	$\mathbf{0.89}_{\pm 0.06}$
	AUC	$0.75_{\pm 0.09}$	$0.63_{\pm 0.17}$	$0.71_{\pm 0.09}$	$0.78_{\pm 0.11}$	$\mathbf{0.82}_{\pm 0.06}$	$0.79_{\pm 0.12}$
	Bal. Acc	$0.75_{\pm 0.12}$	$0.66_{\pm 0.14}$	$0.66_{\pm 0.10}$	$0.75_{\pm 0.12}$	$0.75_{\pm 0.11}$	$\mathbf{0.78}_{\pm 0.11}$
ACVR2A Mut.	F1	$0.89_{\pm 0.07}$	$0.80_{\pm 0.10}$	$0.85_{\pm 0.03}$	$0.89_{\pm 0.09}$	$0.87_{\pm 0.05}$	$\mathbf{0.91}_{\pm 0.05}$
	AUC	$0.80_{\pm 0.13}$	$0.74_{\pm 0.11}$	$0.83_{\pm 0.08}$	$0.76_{\pm 0.19}$	$0.79_{\pm 0.10}$	$\mathbf{0.91}_{\pm 0.09}$
	Bal. Acc	$0.80_{\pm 0.12}$	$0.65_{\pm 0.10}$	$0.81_{\pm 0.08}$	$0.81_{\pm 0.16}$	$0.76_{\pm 0.15}$	$\mathbf{0.90}_{\pm 0.10}$
Histologic Grade	F1	$0.71_{\pm 0.07}$	$0.77_{\pm 0.03}$	$0.73_{\pm 0.11}$	$0.75_{\pm 0.06}$	$0.73_{\pm 0.06}$	$\mathbf{0.78}_{\pm 0.08}$
	AUC	$0.71_{\pm 0.09}$	$\mathbf{0.77}_{\pm 0.04}$	$0.67_{\pm 0.11}$	$0.74_{\pm 0.08}$	$0.71_{\pm 0.04}$	$0.75_{\pm 0.15}$
	Bal. Acc	$0.73_{\pm 0.06}$	$0.77_{\pm 0.03}$	$0.73_{\pm 0.12}$	$0.74_{\pm 0.06}$	$0.73_{\pm 0.06}$	$\mathbf{0.77}_{\pm 0.08}$
KRAS Mut.	F1	$0.77_{\pm 0.08}$	$0.77_{\pm 0.08}$	$0.72_{\pm 0.07}$	$0.81_{\pm 0.06}$	$0.80_{\pm 0.05}$	$\mathbf{0.85}_{\pm 0.04}$
	AUC	$0.76_{\pm 0.14}$	$0.72_{\pm 0.09}$	$0.61_{\pm 0.12}$	$0.70_{\pm 0.05}$	$0.80_{\pm 0.05}$	$\mathbf{0.80}_{\pm 0.06}$
	Bal. Acc	$0.74_{\pm 0.13}$	$0.76_{\pm 0.08}$	$0.63_{\pm 0.13}$	$0.77_{\pm 0.07}$	$\mathbf{0.81}_{\pm 0.05}$	$0.79_{\pm 0.10}$
IDH Status	F1	$0.92_{\pm 0.01}$	$0.94_{\pm 0.02}$	$0.91_{\pm 0.02}$	$0.92_{\pm 0.02}$	$0.94_{\pm 0.02}$	$\mathbf{0.97}_{\pm 0.02}$
	AUC	$0.96_{\pm 0.01}$	$0.97_{\pm 0.02}$	$0.95_{\pm 0.01}$	$0.96_{\pm 0.01}$	$0.97_{\pm 0.01}$	$\mathbf{0.99}_{\pm 0.01}$
	Bal. Acc	$0.92_{\pm 0.01}$	$0.94_{\pm 0.02}$	$0.91_{\pm 0.02}$	$0.91_{\pm 0.02}$	$0.94_{\pm 0.02}$	$\mathbf{0.97}_{\pm 0.02}$
Treatment Response	F1	$0.53_{\pm 0.07}$	$0.51_{\pm 0.07}$	$0.57_{\pm 0.08}$	$0.49_{\pm 0.05}$	$0.49_{\pm 0.02}$	$\mathbf{0.58}_{\pm 0.14}$
	AUC	$\mathbf{0.70}_{\pm 0.06}$	$0.68_{\pm 0.10}$	$0.69_{\pm 0.07}$	$0.59_{\pm 0.05}$	$0.60_{\pm 0.06}$	$0.68_{\pm 0.07}$
	Bal. Acc	$0.48_{\pm 0.10}$	$0.40_{\pm 0.08}$	$\mathbf{0.51}_{\pm 0.11}$	$0.35_{\pm 0.09}$	$0.37_{\pm 0.06}$	$0.48_{\pm 0.17}$
BRCA PAM50 Subtype	F1	$0.67_{\pm 0.02}$	$0.68_{\pm 0.05}$	$0.70_{\pm 0.04}$	$0.68_{\pm 0.05}$	$\mathbf{0.72}_{\pm 0.04}$	$0.63_{\pm 0.03}$
	AUC	$0.83_{\pm 0.03}$	$0.83_{\pm 0.05}$	$0.85_{\pm 0.04}$	$0.84_{\pm 0.04}$	$\mathbf{0.87}_{\pm 0.03}$	$0.80_{\pm 0.01}$
	Bal. Acc	$0.51_{\pm 0.02}$	$0.52_{\pm 0.06}$	$0.57_{\pm 0.07}$	$0.53_{\pm 0.06}$	$\mathbf{0.58}_{\pm 0.04}$	$0.50_{\pm 0.05}$
HNSC mRNA Subtype	F1	$0.54_{\pm 0.01}$	$0.60_{\pm 0.04}$	$0.60_{\pm 0.06}$	$0.59_{\pm 0.03}$	$\mathbf{0.62}_{\pm 0.07}$	$0.55_{\pm 0.03}$
	AUC	$0.75_{\pm 0.03}$	$0.77_{\pm 0.03}$	$0.78_{\pm 0.07}$	$0.77_{\pm 0.03}$	$\mathbf{0.79}_{\pm 0.02}$	$0.72_{\pm 0.01}$
	Bal. Acc	$0.55_{\pm 0.04}$	$0.61_{\pm 0.07}$	$0.59_{\pm 0.07}$	$0.58_{\pm 0.05}$	$\mathbf{0.61}_{\pm 0.08}$	$0.54_{\pm 0.04}$
UCEC Genomic Subtype	F1	$0.55_{\pm 0.04}$	$0.57_{\pm 0.04}$	$0.63_{\pm 0.03}$	$0.57_{\pm 0.04}$	$\mathbf{0.66}_{\pm 0.04}$	$0.56_{\pm 0.05}$
	AUC	$0.75_{\pm 0.03}$	$0.76_{\pm 0.04}$	$0.81_{\pm 0.04}$	$0.75_{\pm 0.02}$	$\mathbf{0.82}_{\pm 0.03}$	$0.74_{\pm 0.04}$
	Bal. Acc	$0.53_{\pm 0.04}$	$0.54_{\pm 0.04}$	$0.60_{\pm 0.04}$	$0.55_{\pm 0.03}$	$\mathbf{0.62}_{\pm 0.03}$	$0.52_{\pm 0.05}$
Synaptophysin Grade	F1	$0.78_{\pm 0.08}$	$0.79_{\pm 0.06}$	$0.76_{\pm 0.05}$	$\mathbf{0.81}_{\pm 0.01}$	$0.80_{\pm 0.06}$	$0.79_{\pm 0.08}$
	AUC	$0.61_{\pm 0.19}$	$\mathbf{0.74}_{\pm 0.08}$	$0.50_{\pm 0.09}$	$0.68_{\pm 0.08}$	$0.64_{\pm 0.11}$	$0.61_{\pm 0.15}$
	Bal. Acc	$0.64_{\pm 0.09}$	$\mathbf{0.69}_{\pm 0.08}$	$0.58_{\pm 0.03}$	$0.65_{\pm 0.08}$	$0.68_{\pm 0.04}$	$0.65_{\pm 0.08}$
RAS/BRAF Status	F1	$0.89_{\pm 0.07}$	$0.89_{\pm 0.07}$	$0.84_{\pm 0.11}$	$0.88_{\pm 0.08}$	$0.90_{\pm 0.07}$	$\mathbf{0.94}_{\pm 0.08}$
	AUC	$0.68_{\pm 0.14}$	$0.48_{\pm 0.28}$	$0.56_{\pm 0.16}$	$0.33_{\pm 0.20}$	$\mathbf{0.80}_{\pm 0.09}$	$0.75_{\pm 0.26}$
	Bal. Acc	$0.68_{\pm 0.20}$	$0.68_{\pm 0.18}$	$0.57_{\pm 0.22}$	$0.66_{\pm 0.19}$	$0.68_{\pm 0.19}$	$\mathbf{0.85}_{\pm 0.20}$
Keratinizing SCC Grade	F1	$\mathbf{0.74}_{\pm 0.02}$	$0.68_{\pm 0.06}$	$0.72_{\pm 0.04}$	$0.69_{\pm 0.02}$	$0.72_{\pm 0.03}$	$0.73_{\pm 0.03}$
	AUC	$\mathbf{0.75}_{\pm 0.04}$	$0.69_{\pm 0.07}$	$0.74_{\pm 0.02}$	$0.72_{\pm 0.04}$	$0.72_{\pm 0.05}$	$0.72_{\pm 0.06}$
	Bal. Acc	$\mathbf{0.73}_{\pm 0.02}$	$0.68_{\pm 0.06}$	$0.72_{\pm 0.03}$	$0.69_{\pm 0.02}$	$0.71_{\pm 0.03}$	$0.72_{\pm 0.03}$
Non-Keratinizing SCC Grade	F1	$0.80_{\pm 0.12}$	$0.78_{\pm 0.12}$	$0.75_{\pm 0.17}$	$0.80_{\pm 0.11}$	$\mathbf{0.81}_{\pm 0.12}$	$0.77_{\pm 0.14}$
	AUC	$0.66_{\pm 0.12}$	$0.61_{\pm 0.03}$	$0.52_{\pm 0.20}$	$0.66_{\pm 0.11}$	$\mathbf{0.71}_{\pm 0.11}$	$0.66_{\pm 0.07}$
	Bal. Acc	$\mathbf{0.79}_{\pm 0.15}$	$0.75_{\pm 0.13}$	$0.72_{\pm 0.17}$	$0.79_{\pm 0.12}$	$0.78_{\pm 0.14}$	$0.73_{\pm 0.15}$
Primary vs. Metastasis	F1	$0.92_{\pm 0.01}$	$0.92_{\pm 0.01}$	$0.92_{\pm 0.01}$	$0.93_{\pm 0.02}$	$0.93_{\pm 0.03}$	$\mathbf{0.93}_{\pm 0.02}$
	AUC	$0.62_{\pm 0.05}$	$0.60_{\pm 0.12}$	$0.66_{\pm 0.10}$	$0.75_{\pm 0.07}$	$\mathbf{0.77}_{\pm 0.03}$	$0.69_{\pm 0.13}$
	Bal. Acc	$0.57_{\pm 0.02}$	$0.58_{\pm 0.05}$	$0.63_{\pm 0.09}$	$0.63_{\pm 0.05}$	$0.64_{\pm 0.08}$	$\mathbf{0.65}_{\pm 0.10}$

Table 2: Macro-average MIL comparison across sixteen held-out tasks. Each patch-encoder entry averages over five MIL architectures (MeanMIL, ABMIL, CLAM, DSMIL, TransMIL).

Metric	Backbone	UNI v2	Phikon v2	CONCH v1.5	MUSK	MOOZY (Ours)
F1 (weighted)	0.723	0.722	0.714	<u>0.740</u>	0.720	**0.769**
ROC-AUC (weighted)	0.707	0.707	0.697	<u>0.720</u>	0.695	**0.763**
Balanced Acc	0.643	0.637	0.625	<u>0.661</u>	0.637	**0.702**

competing encoders (GigaPath 1.22B, PRISM 742M, Madeleine 400M, TITAN 355M). CHIEF is smaller in absolute size but trails MOOZY on all three macro metrics. This efficiency stems from concentrating capacity at the slide level and reusing a lightweight public patch encoder. A detailed breakdown is in Table 22.

Comparison with MILs. We compare MOOZY against five patch encoders in the MIL setting. Each encoder is paired with five MIL architectures (MeanMIL, ABMIL, CLAM, DSMIL, and TransMIL) following [78], and each entry is the arithmetic mean over the five architectures. Across all sixteen tasks, MOOZY leads the strongest MIL baseline, CONCH v1.5, by 0.029 weighted F1, 0.043 weighted ROC-AUC, and 0.041 balanced accuracy. Macro-average results are in Table 2. Task-wise and per-method breakdowns are in Appendix J.

Multi-stage Ablation. Table 3 reports macro-average results for five configurations. With the case aggregator held out, adding SSL pretraining improves all three metrics. Stage 2 only without the case aggregator reaches 0.743 weighted F1, 0.721 weighted ROC-AUC, and 0.664 balanced accuracy, which MOOZY without the case aggregator raises to 0.749, 0.737, and 0.682. The case aggregator further raises these values to 0.769, 0.763, and 0.702. When trained from scratch without SSL, adding the case aggregator reduces performance (Stage 2 only at 0.731/0.697/0.659). Task-wise breakdowns are in Appendix K and Tab. 29, Appendix L and Tab. 30, and Appendix M and Tab. 31.

Case Aggregator Ablation. We ablate the case aggregator by comparing MOOZY without the case aggregator (Stage 2 slide encoder with mean slide pooling)[10] against full MOOZY (Table 3). The aggregator raises macro weighted F1 from 0.749 to 0.769, AUC from 0.737 to 0.763, and balanced accuracy from 0.682 to 0.702. Task-wise, F1 improves on 14 of 16 tasks, AUC on 12, and balanced accuracy on 13. Full task-wise results are provided in Appendix N and Tab. 32.

Attention Map Analysis. To examine where each encoder concentrates on the slide, a board-certified pathologist reviewed attention maps for 20 representative WSIs sampled from eight held-out evaluation cohorts and five encoders. Two scores were assigned per image: the shift score (1–5, where 1 is cancer-focused, 3 is balanced, and 5 is non-cancer-focused) and the semantic-gap score (1–3, where

[10] The case aggregator is trained as part of MOOZY and discarded only at inference, where per-slide embeddings are mean-pooled instead.

Table 3: Unified macro-average ablation across sixteen held-out tasks. The table includes Stage 1 only, Stage 2 only with and without the case aggregator, MOOZY without the case aggregator (mean slide pooling), and full MOOZY.

Setting	Stage 1	Stage 2	Case Agg.	F1	AUC	Bal. Acc
Stage 1 only	✓	✗	✗	0.743	0.715	0.662
Stage 2 only w/o case agg.	✗	✓	✗	0.743	0.721	0.664
Stage 2 only	✗	✓	✓	0.731	0.697	0.659
MOOZY w/o case agg.	✓	✓	✗	0.749	0.737	0.682
MOOZY	✓	✓	✓	0.769	0.763	0.702

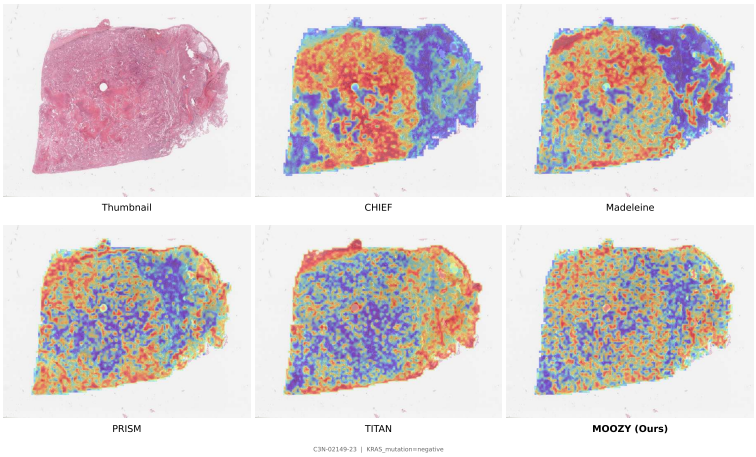


Fig. 5: Attention-map comparison on a lung adenocarcinoma slide. MOOZY and TITAN: balanced, comprehensive coverage (shift 3, gap 1). PRISM: balanced shift with moderate gaps (shift 3, gap 2). CHIEF and Madeleine: cancer-biased with frequent semantic gaps (shift 2, gap 3).

1 indicates rare and 3 indicates frequent gaps in diagnostically relevant tissue). MOOZY achieved the lowest mean gap score (1.00), followed by TITAN (1.38) and PRISM (1.75). On shift, MOOZY was near balanced (2.63) and TITAN closest to exact balance (3.13), while PRISM (2.38) were more cancer-biased. These two scores suggest that MOOZY attends broadly while maintaining balanced focus between tumor and surrounding tissue. A plausible explanation comes from the two-stage objective. In Stage 1, masked-region prediction and agreement across global and local views encourage the model to use distributed contextual evidence instead of relying on a few high-response hotspots. In Stage 2, supervision across diverse endpoints and cohorts favors features that stay informative in both malignant and non-malignant tissue contexts. A representative example is shown in Figure 5. Heatmap generation details and additional examples are provided in Appendix O.

Qualitative Analysis. To complement quantitative results, we visualize UMAP [58] embeddings for three representative tasks (CPTAC cancer type, pan-dataset anatomical site, and TCGA cancer type) across four slide encoders

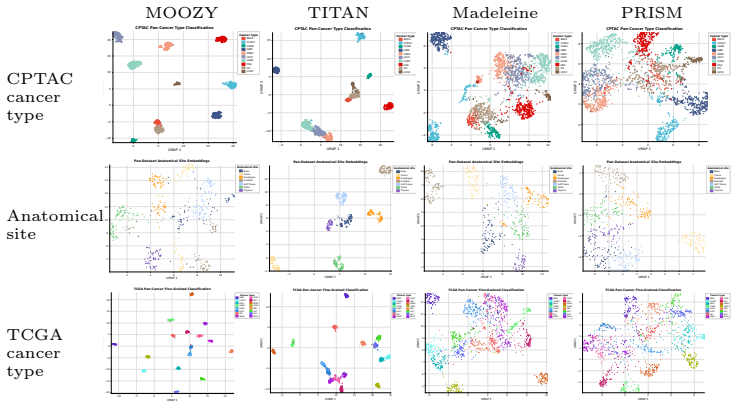


Fig. 6: UMAP qualitative comparison across four slide encoders (columns) and three tasks (rows), using matched class-balanced sampling and identical reduction settings.

(Figure 6). MOOZY shows the clearest separation on CPTAC and TCGA cancer type, TITAN is close behind, and Madeleine/PRISM show more overlap. On anatomical site, TITAN is strongest, with MOOZY and Madeleine comparable and PRISM weaker. t-SNE [57] and PCA stability diagnostics confirm the same pattern (Appendices P and Q and Figure 16).

6 Conclusion

We presented MOOZY, a two-stage patient-first framework that decouples vision-only SSL pretraining from case-aware multi-task alignment replacing the naive early/late multi-slide fusion with explicit case-level inter-slide dependency modeling, showing competitive representations without proprietary slides, paired clinical reports, or billion-parameter architectures. Across sixteen held-out tasks, MOOZY achieves the strongest macro weighted F1 and balanced accuracy among slide encoders and exceeds all trained MIL baselines across the three macro metrics, while using fewer parameters than the compared slide encoders. Like other slide encoders, MOOZY targets global case-level predictions, so dense or segmentation predictions are not supported. The benefit of explicit cross-slide aggregation also remains to be quantified on tasks specifically designed to require deeper multi-slide reasoning, alongside natural extensions toward case-level retrieval, slide-report co-training, and genomic fusion.

Acknowledgements

This work was supported by NSERC-DG RGPIN-2022-05378 [M.S.H], Amazon Research Award [M.S.H], and Gina Cody RIF [M.S.H], FRQNT scholarship [Y.K]. Computational resources were provided in part by Calcul Québec (`www.calculquebec.ca`) and the Digital Research Alliance of Canada (`www.alliancecan.ca`).

References

1. Alagha, A., Leclerc, C., Kotp, Y., Metwally, O., Moras, C., Rentopoulos, P., Rostami, G., Nguyen, B.N., Baig, J., Khellaf, A., Trinh, V.Q.H., Mizouni, R., Otrok, H., Bentahar, J., Hosseini, M.S.: Atlaspatch: An efficient and scalable tool for whole slide image preprocessing in computational pathology. arXiv preprint arXiv:2602.03998 (2026)
2. Bandi, P., Geessink, O., Manson, Q., Van Dijk, M., Balkenhol, M., Hermsen, M., Bejnordi, B.E., Lee, B., Paeng, K., Zhong, A., et al.: From detection of individual metastases to classification of lymph node status at the patient level: the CAMELYON17 challenge. IEEE Transactions on Medical Imaging **38**(2), 550–560 (2019). https://doi.org/10.1109/TMI.2018.2867350
3. Bejnordi, B.E., Veta, M., Van Diest, P.J., Van Ginneken, B., Karssemeijer, N., Litjens, G., Van Der Laak, J.A., Hermsen, M., Manson, Q.F., Balkenhol, M., et al.: Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. Jama **318**(22), 2199–2210 (2017)
4. Belagali, V., Kapse, S., Marza, P., Das, S., Li, Z., Boutaj, S., Pati, P., Yellapragada, S., Nandi, T.N., Madduri, R.K., Saltz, J., Prasanna, P., Christodoulidis, S., Vakalopoulou, M., Samaras, D.: Ticon: A slide-level tile contextualizer for histopathology representation learning. arXiv preprint arXiv:2512.21331 (2025)
5. Bergstrom, E.N., Abbasi, A., Díaz-Gay, M., Galland, L., Ladoire, S., Lippman, S.M., Alexandrov, L.B.: Deep learning artificial intelligence predicts homologous recombination deficiency and platinum response from histologic slides. Journal of Clinical Oncology **42**(30), 3550–3560 (oct 2024). https://doi.org/10.1200/JCO.23.02641, https://doi.org/10.1200/JCO.23.02641
6. Bioptimus: H-optimus-1 (2025), https://huggingface.co/bioptimus/H-optimus-1
7. Brancati, N., Anniciello, A.M., Pati, P., Riccio, D., Scognamiglio, G., Jaume, G., De Pietro, G., Di Bonito, M., Foncubierta, A., Botti, G., Gabrani, M., Feroce, F., Frucci, M.: Bracs: A dataset for breast carcinoma subtyping in h&e histology images. Database **2022** (jan 2022). https://doi.org/10.1093/database/baac093, https://doi.org/10.1093/database/baac093
8. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. Advances in neural information processing systems **33**, 1877–1901 (2020)
9. Bulten, W., Kartasalo, K., Chen, P.H.C., Ström, P., Pinckaers, H., Nagpal, K., Cai, Y., Steiner, D.F., Van Boven, H., Vink, R., et al.: Artificial intelligence for diagnosis and gleason grading of prostate cancer: the panda challenge. Nature medicine **28**(1), 154–163 (2022). https://doi.org/10.1038/s41591-021-01620-2
10. Campanella, G., Hanna, M.G., Geneslaw, L., Miraflor, A., Werneck Krauss Silva, V., Busam, K.J., Brogi, E., Reuter, V.E., Klimstra, D.S., Fuchs, T.J.: Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. Nature medicine **25**(8), 1301–1309 (2019)
11. Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., Joulin, A.: Unsupervised learning of visual features by contrasting cluster assignments. Advances in neural information processing systems **33**, 9912–9924 (2020)
12. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: Proceedings

of the IEEE/CVF international conference on computer vision. pp. 9650–9660 (2021)
13. Chen, R.J., Chen, C., Li, Y., Chen, T.Y., Trister, A.D., Krishnan, R.G., 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. pp. 16144–16155 (2022)
14. Chen, R.J., Ding, T., Lu, M.Y., Williamson, D.F., Jaume, G., Song, A.H., Chen, B., Zhang, A., Shao, D., Shaban, M., et al.: Towards a general-purpose foundation model for computational pathology. Nature medicine **30**(3), 850–862 (2024)
15. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International conference on machine learning. pp. 1597–1607. PmLR (2020)
16. Chen, X., He, K.: Exploring simple siamese representation learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 15750–15758 (2021)
17. Courtiol, P., Maussion, C., Moarii, M., Pronier, E., Pilcer, S., Sefta, M., Manceron, P., Toldo, S., Zaslavskiy, M., Le Stang, N., et al.: Deep learning-based classification of mesothelioma improves prediction of patient outcome. Nature medicine **25**(10), 1519–1525 (2019)
18. Darcet, T., Oquab, M., Mairal, J., Bojanowski, P.: Vision transformers need registers. In: The Twelfth International Conference on Learning Representations (2024), https://openreview.net/forum?id=2dnO3LLiJ1
19. Ding, J., Ma, S., Dong, L., Zhang, X., Huang, S., Wang, W., Zheng, N., Wei, F.: Longnet: Scaling transformers to 1,000,000,000 tokens. arXiv preprint arXiv:2307.02486 (2023)
20. Ding, T., Wagner, S.J., Song, A.H., Chen, R.J., Lu, M.Y., Zhang, A., Vaidya, A.J., Jaume, G., Shaban, M., Kim, A., et al.: Multimodal whole slide foundation model for pathology. arXiv preprint arXiv:2411.19666 (2024)
21. Dörrich, M., Balk, M., Heusinger, T., Beyer, S., Mirbagheri, H., Fischer, D.J., Kanso, H., Matek, C., Hartmann, A., Iro, H., Eckstein, M., Gostian, A.O., Kist, A.M.: A multimodal dataset for precision oncology in head and neck cancer. Nature Communications **16**, 7163 (2025). https://doi.org/10.1038/s41467-025-62386-6
22. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
23. Edwards, N.J., Oberti, M., Thangudu, R.R., Cai, S., McGarvey, P.B., Jacob, S., Madhavan, S., Ketchum, K.A.: The CPTAC data portal: a resource for cancer proteomics research. Journal of Proteome Research **14**(6), 2707–2713 (2015)
24. Filiot, A., Dop, N., Tchita, O., Riou, A., Dubois, R., Peeters, T., Valter, D., Scalbert, M., Saillard, C., Robin, G., Olivier, A.: Distilling foundation models for robust and efficient models in digital pathology. arXiv preprint arXiv:2501.16239 (2025)
25. Filiot, A., Ghermi, R., Olivier, A., Jacob, P., Fidon, L., Camara, A., Mac Kain, A., Saillard, C., Schiratti, J.B.: Scaling self-supervised learning for histopathology with masked image modeling. medRxiv (2023)
26. Filiot, A., Jacob, P., Mac Kain, A., Saillard, C.: Phikon-v2, a large and public feature extractor for biomarker prediction. arXiv preprint arXiv:2409.09173 (2024)

27. Galland, L., Ballot, E., Mananet, H., Boidot, R., Lecuelle, J., Albuisson, J., Arnould, L., Desmoulins, I., Mayeur, D., Kaderbhai, C., Ilie, S., Hennequin, A., Bergeron, A., Derangère, V., Ghiringhelli, F., Truntzer, C., Ladoire, S.: Efficacy of platinum-based chemotherapy in metastatic breast cancer and hrd biomarkers: utility of exome sequencing. npj Breast Cancer **8**(1), 28 (mar 2022). https://doi.org/10.1038/s41523-022-00395-0, https://doi.org/10.1038/s41523-022-00395-0
28. Grill, J.B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., et al.: Bootstrap your own latent-a new approach to self-supervised learning. Advances in neural information processing systems **33**, 21271–21284 (2020)
29. GTEx Consortium: The Genotype-Tissue Expression (GTEx) project. Nature Genetics **45**(6), 580–585 (2013)
30. Gu, A., Dao, T.: Mamba: Linear-time sequence modeling with selective state spaces. arXiv preprint arXiv:2312.00752 (2023)
31. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16000–16009 (2022)
32. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9729–9738 (2020)
33. Hendrycks, D., Gimpel, K.: Gaussian error linear units (gelus). arXiv preprint arXiv:1606.08415 (2016)
34. Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., Casas, D.d.L., Hendricks, L.A., Welbl, J., Clark, A., et al.: Training compute-optimal large language models. arXiv preprint arXiv:2203.15556 (2022)
35. Hosseini, M.S., Bejnordi, B.E., Trinh, V.Q.H., Chan, L., Hasan, D., Li, X., Yang, S., Kim, T., Zhang, H., Wu, T., et al.: Computational pathology: a survey review and the way forward. Journal of Pathology Informatics **15**, 100357 (2024)
36. Hou, X., Jiang, C., Kondepudi, A., Lyu, Y., Chowdury, A., Lee, H., Hollon, T.C.: A self-supervised framework for learning whole slide representations. arXiv preprint arXiv:2402.06188 (2024)
37. Hua, S., Yan, F., Shen, T., Ma, L., Zhang, X.: Pathoduet: Foundation models for pathological slide analysis of h&e and ihc stains. Medical Image Analysis **97**, 103289 (2024)
38. Huang, G., Sun, Y., Liu, Z., Sedra, D., Weinberger, K.Q.: Deep Networks with Stochastic Depth, pp. 646–661. Springer International Publishing (2016). https://doi.org/10.1007/978-3-319-46493-0_39, http://dx.doi.org/10.1007/978-3-319-46493-0_39
39. Ilse, M., Tomczak, J., Welling, M.: Attention-based deep multiple instance learning. In: International conference on machine learning. pp. 2127–2136. PMLR (2018)
40. Jaume, G., Oldenburg, L., Vaidya, A., Chen, R.J., Williamson, D.F., Peeters, T., Song, A.H., Mahmood, F.: Transcriptomics-guided slide representation learning in computational pathology. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2024)
41. Jaume, G., Vaidya, A., Zhang, A., H. Song, A., J. Chen, R., Sahai, S., Mo, D., Madrigal, E., Phi Le, L., Mahmood, F.: Multistain pretraining for slide representation learning in pathology. In: European Conference on Computer Vision. pp. 19–37. Springer (2024)

42. Kang, M., Song, H., Park, S., Yoo, D., Pereira, S.: Benchmarking self-supervised learning on diverse pathology datasets. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3344–3354 (2023)
43. Kaplan, J., McCandlish, S., Henighan, T., Brown, T.B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., Amodei, D.: Scaling laws for neural language models. arXiv preprint arXiv:2001.08361 (2020)
44. Karasikov, M., van Doorn, J., Känzig, N., Cesur, M.E., Horlings, H.M., Berke, R., Tang, F., Otálora, S.: Training state-of-the-art pathology foundation models with orders of magnitude less data (2025), https://arxiv.org/abs/2504.05186
45. Kather, J.N., Heij, L.R., Grabsch, H.I., Loeffler, C., Echle, A., Muti, H.S., Krause, J., Niehues, J.M., Sommer, K.A., Bankhead, P., et al.: Pan-cancer image-based detection of clinically actionable genetic alterations. Nature cancer **1**(8), 789–799 (2020)
46. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4015–4026 (2023)
47. Kozachenko, L.F., Leonenko, N.N.: Sample estimate of the entropy of a random vector. Problemy Peredachi Informatsii **23**(2), 9–16 (1987)
48. Lazard, T., Lerousseau, M., Decencière, E., Walter, T.: Giga-ssl: Self-supervised learning for gigapixel images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. pp. 4812–4822 (2023)
49. Lee, Y., Oh, H.R., Bychlov, A., Fukuoka, J., Kaushal, R.K., Sahay, A., Yadav, R., Sarioglu, S., Balci, S., Türkmen, İ., Tolkach, Y., Harder, C., Choi, J.H., Ahn, S.: Report generation of pathology using pan-asia giga-pixel wsis (2025) (2025). https://doi.org/10.5281/zenodo.15081614, https://doi.org/10.5281/zenodo.15081614
50. Lenz, T., Neidlinger, P., Ligero, M., Wölflein, G., van Treeck, M., Kather, J.N.: Unsupervised foundation model-agnostic slide-level representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2025)
51. Li, B., Li, Y., Eliceiri, K.W.: Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 14318–14328 (2021)
52. Li, J., Chen, Y., Chu, H., Sun, Q., Guan, T., Han, A., He, Y.: Dynamic graph representation with knowledge-aware attention for histopathology whole slide image analysis. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11323–11332 (2024)
53. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: International Conference on Learning Representations (2019)
54. Lu, M.Y., Chen, B., Williamson, D.F., Chen, R.J., Ikamura, K., Gerber, G., Liang, I., Le, L.P., Ding, T., Parwani, A.V., et al.: A foundational multimodal vision language ai assistant for human pathology. arXiv preprint arXiv:2312.07814 (2023)
55. Lu, M.Y., Chen, B., Williamson, D.F., Chen, R.J., Liang, I., Ding, T., Jaume, G., Odintsov, I., Le, L.P., Gerber, G., et al.: A visual-language foundation model for computational pathology. Nature medicine **30**(3), 863–874 (2024)
56. Lu, M.Y., Williamson, D.F., Chen, T.Y., Chen, R.J., Barbieri, M., Mahmood, F.: Data-efficient and weakly supervised computational pathology on whole-slide images. Nature biomedical engineering **5**(6), 555–570 (2021)

57. van der Maaten, L., Hinton, G.: Visualizing data using t-SNE. Journal of Machine Learning Research **9**(86), 2579–2605 (2008)
58. McInnes, L., Healy, J., Melville, J.: UMAP: Uniform manifold approximation and projection for dimension reduction. arXiv preprint arXiv:1802.03426 (2018). https://doi.org/10.48550/arXiv.1802.03426
59. National Cancer Institute: The cancer genome atlas program (tcga). https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga, accessed: 2025-10-19
60. Nechaev, D., Pchelnikov, A., Ivanova, E.: Hibou: A family of foundational vision transformers for pathology. arXiv preprint arXiv:2406.05074 (2024)
61. Neto, P.C., Montezuma, D., Oliveira, S.P., Oliveira, D., Fraga, J., Monteiro, A., Monteiro, J., Ribeiro, L., Gonçalves, S., Reinhard, S., et al.: An interpretable machine learning system for colorectal cancer diagnosis from pathology slides. npj Precision Oncology **8**(1), 56 (2024). https://doi.org/10.1038/s41698-024-00539-4, https://doi.org/10.1038/s41698-024-00539-4
62. Neto, P.C., Oliveira, S.P., Montezuma, D., Fraga, J., Monteiro, A., Ribeiro, L., Gonçalves, S., Pinto, I.M., Cardoso, J.S.: imil4path: A semi-supervised interpretable approach for colorectal whole-slide images. Cancers **14**(10), 2489 (2022). https://doi.org/10.3390/cancers14102489, https://doi.org/10.3390/cancers14102489
63. Nicke, T., Schacherer, D., Schäfer, J.R., Artysh, N., Prasse, A., Homeyer, A., Schenk, A., Höfener, H., Lotz, J.: Tissue concepts v2: A supervised foundation model for whole slide images. arXiv preprint arXiv:2507.05742 (2025)
64. Nicke, T., Schaefer, J.R., Hoefener, H., Feuerhake, F., Merhof, D., Kiessling, F., Lotz, J.: Tissue concepts: Supervised foundation models in computational pathology. Computers in Biology and Medicine **186**, 109621 (2025)
65. Oliveira, S.P., Montezuma, D., Moreira, A., Oliveira, D., Neto, P.C., Monteiro, A., Monteiro, J., Ribeiro, L., Gonçalves, S., Pinto, I.M., Cardoso, J.S.: A cad system for automatic dysplasia grading on h&e cervical whole-slide images. Scientific Reports **13**(1), 3970 (mar 2023). https://doi.org/10.1038/s41598-023-30497-z
66. Oliveira, S.P., Neto, P.C., Fraga, J., Montezuma, D., Monteiro, A., Monteiro, J., Ribeiro, L., Gonçalves, S., Pinto, I.M., Cardoso, J.S.: Cad systems for colorectal cancer from wsi are still not ready for clinical acceptance. Scientific Reports **11**(1), 1–15 (2021). https://doi.org/10.1038/s41598-021-93746-z, https://doi.org/10.1038/s41598-021-93746-z
67. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)
68. Peikari, M., Salama, S., Nofech-Mozes, S., Martel, A.L.: Automatic cellularity assessment from post-treated breast surgical specimens. Cytometry Part A **91**(11), 1078–1087 (oct 2017). https://doi.org/10.1002/cyto.a.23244, https://doi.org/10.1002/cyto.a.23244
69. Press, O., Smith, N.A., Lewis, M.: Train short, test long: Attention with linear biases enables input length extrapolation. In: International Conference on Learning Representations (2022)
70. Rajaram, S., Acosta, P., Panwar, V., Jarmale, V., Christie, A., Jasti, J., Margulis, V., Rakheja, D., Cheville, J., Leibovich, B.C., Parker, A., Brugarolas, J., Kapur, P.: Prediction of driver mutation heterogeneity in renal cancer from histopathology slides using deep learning (2022). https://doi.org/10.25452/figshare.plus.c.5983795.v1, https://doi.org/10.25452/figshare.plus.c.5983795.v1

71. Ravi, N., Gabeur, V., Hu, Y.T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., et al.: Sam 2: Segment anything in images and videos. arXiv preprint arXiv:2408.00714 (2024)
72. Roetzer-Pejrimovsky, T., Moser, A.C., Atli, B., Vogel, C.C., Mercea, P.A., Prihoda, R., Gelpi, E., Haberler, C., Höftberger, R., Hainfellner, J.A., Baumann, B., Langs, G., Woehrer, A.: The digital brain tumour atlas, an open histopathology resource. Scientific Data **9**(1), 55 (2022). https://doi.org/10.1038/s41597-022-01157-0, https://doi.org/10.1038/s41597-022-01157-0
73. Roetzer-Pejrimovsky, T., Moser, A.C., Atli, B., Vogel, C.C., Mercea, P.A., Prihoda, R., Gelpi, E., Haberler, C., Höftberger, R., Hainfellner, J.A., Baumann, B., Langs, G., Woehrer, A.: The digital brain tumour atlas, an open histopathology resource [data set] (2022). https://doi.org/10.25493/WQ48-ZGX, https://doi.org/10.25493/WQ48-ZGX
74. Sablayrolles, A., Douze, M., Schmid, C., Jégou, H.: Spreading vectors for similarity search. In: International Conference on Learning Representations (2019)
75. Saillard, C., Jenatton, R., Llinares-López, F., Mariet, Z., Cahané, D., Durand, E., Vert, J.P.: H-optimus-0 (2024), https://github.com/bioptimus/releases/tree/main/models/h-optimus/v0
76. Sammut, S.J.: Multi-omic machine learning predictor of breast cancer therapy response (2022). https://doi.org/10.5281/zenodo.6337925, https://doi.org/10.5281/zenodo.6337925
77. Shaikovski, G., Casson, A., Severson, K., Zimmermann, E., Wang, Y.K., Kunz, J.D., Retamero, J.A., Oakley, G., Klimstra, D., Kanan, C., et al.: Prism: A multimodal generative foundation model for slide-level histopathology. arXiv preprint arXiv:2405.10254 (2024)
78. Shao, D., Chen, R.J., Song, A.H., Runevic, J., Lu, M.Y., Ding, T., Mahmood, F.: Do multiple instance learning models transfer? In: Proceedings of the 42nd International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 267, pp. 54219–54238. PMLR (2025), https://proceedings.mlr.press/v267/shao25a.html
79. Shao, Z., Bian, H., Chen, Y., Wang, Y., Zhang, J., Ji, X., et al.: Transmil: Transformer based correlated multiple instance learning for whole slide image classification. Advances in neural information processing systems **34**, 2136–2147 (2021)
80. Siméoni, O., Vo, H.V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M., et al.: Dinov3. arXiv preprint arXiv:2508.10104 (2025)
81. Skrede, O.J., De Raedt, S., Kleppe, A., Hveem, T.S., Liestøl, K., Maddison, J., Askautrud, H.A., Pradhan, M., Nesheim, J.A., Albregtsen, F., et al.: Deep learning for prediction of colorectal cancer outcome: a discovery and validation study. The Lancet **395**(10221), 350–360 (2020)
82. Sophont AI, MedARC: How to train a state-of-the-art pathology foundation model with $1.6k. https://sophontai.com/blog/openmidnight (2025)
83. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2818–2826. IEEE (Jun 2016). https://doi.org/10.1109/cvpr.2016.308, http://dx.doi.org/10.1109/CVPR.2016.308
84. Tang, W., Zhou, F., Huang, S., Zhu, X., Zhang, Y., Liu, B.: Feature re-embedding: Towards foundation model-level performance in computational pathology. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11343–11352 (2024)

85. Touvron, H., Cord, M., Sablayrolles, A., Synnaeve, G., Jegou, H.: Going deeper with image transformers. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 32–42. IEEE (Oct 2021). https://doi.org/10.1109/iccv48922.2021.00010, http://dx.doi.org/10.1109/ICCV48922.2021.00010
86. Trahearn, N., Sakr, C., Banerjee, A., Lee, S.H., Baker, A.M., Kocher, H.M., Angerilli, V., Morano, F., Bergamo, F., Maddalena, G., Intini, R., Cremolini, C., Caravagna, G., Graham, T., Pietrantonio, F., Lonardi, S., Fassan, M., Sottoriva, A.: Computational pathology applied to clinical colorectal cancer cohorts identifies immune and endothelial cell spatial patterns predictive of outcome. The Journal of Pathology **265**(2), 198–210 (2025). https://doi.org/10.1002/path.6378
87. Vaidya, A., Zhang, A., Jaume, G., Song, A.H., Ding, T., Wagner, S.J., Lu, M.Y., Doucet, P., Robertson, H., Almagro-Perez, C., Chen, R.J., ElHarouni, D., Ayoub, G., Bossi, C., Ligon, K.L., Gerber, G., Le, L.P., Mahmood, F.: Molecular-driven foundation model for oncologic pathology. arXiv preprint arXiv:2501.16652 (2025)
88. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Advances in neural information processing systems. vol. 30 (2017)
89. Vorontsov, E., Bozkurt, A., Casson, A., Shaikovski, G., Zelechowski, M., Liu, S., Severson, K., Zimmermann, E., Hall, J., Tenenholtz, N., et al.: Virchow: A million-slide digital pathology foundation model. arXiv preprint arXiv:2309.07778 (2023)
90. Wang, T., Isola, P.: Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In: International Conference on Machine Learning. pp. 9929–9939. PMLR (2020)
91. Wang, X., Zhao, J., Marostica, E., Yuan, W., Jin, J., Zhang, J., Li, R., Tang, H., Wang, K., Li, Y., et al.: A pathology foundation model for cancer diagnosis and prognosis prediction. Nature **634**(8035), 970–978 (2024)
92. Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., et al.: Emergent abilities of large language models. arXiv preprint arXiv:2206.07682 (2022)
93. Wei, J.W., Tafe, L.J., Linnik, Y.A., Vaickus, L.J., Tomita, N., Hassanpour, S.: Pathologist-level classification of histologic patterns on resected lung adenocarcinoma slides with deep neural networks. Scientific reports **9**(1), 1–8 (2019)
94. Wei, J., Suriawinata, A., Ren, B., Liu, X., Lisovsky, M., Vaickus, L., Brown, C., Baker, M., Tomita, N., Torresani, L., Wei, J., Hassanpour, S.: A petri dish for histopathology image analysis. In: International Conference on Artificial Intelligence in Medicine. pp. 11–24. Springer (2021)
95. Wilkinson, S., Ye, H., Karzai, F., Harmon, S.A., Terrigino, N.T., VanderWeele, D.J., Bright, J.R., Atway, R., Trostel, S.Y., Carrabba, N.V., Whitlock, N.C., Walker, S.M., Lis, R.T., Sater, H.A., Capaldo, B.J., Madan, R.A., Gulley, J.L., Chun, G., Merino, M.J., Pinto, P.A., Salles, D.C., Kaur, H.B., Lotan, T.L., Venzon, D.J., Choyke, P.L., Turkbey, B., Dahut, W.L., Sowalsky, A.G.: Nascent prostate cancer heterogeneity drives evolution and resistance to intense hormonal therapy. medRxiv (2020). https://doi.org/10.1101/2020.09.29.20199711, https://www.medrxiv.org/content/10.1101/2020.09.29.20199711v1
96. Xiang, J., Wang, X., Zhang, X., Xi, Y., Eweje, F., Chen, Y., Li, Y., Bergstrom, C., Gopaulchan, M., Kim, T., Yu, K.H., Willens, S., Olguin, F.M., Nirschl, J.J., Neal, J., Diehn, M., Yang, S., Li, R.: A vision–language foundation model for precision oncology. Nature (2025)
97. Xiang, J., Zhang, J.: Exploring low-rank property in multiple instance learning for whole slide image classification. In: International Conference on Learning Representations (2023)

98. Xu, H., Usuyama, N., Bagga, J., Zhang, S., Rao, R., Naumann, T., Wong, C., Gero, Z., González, J., Gu, Y., et al.: A whole-slide foundation model for digital pathology from real-world data. Nature **630**(8015), 181–188 (2024)
99. Xu, H., Usuyama, N., Bagga, J., Zhang, S., Rao, R., Naumann, T., Wong, C., Gero, Z., González, J., Gu, Y., Xu, Y., Wei, M., Wang, W., Ma, S., Wei, F., Yang, J., Li, C., Gao, J., Rosemon, J., Bower, T., Lee, S., Weerasinghe, R., Wright, B.J., Robicsek, A., Piening, B., Bifulco, C., Wang, S., Poon, H.: A whole-slide foundation model for digital pathology from real-world data. Nature (2024)
100. Xu, Y., Wang, Y., Zhou, F., Ma, J., Jin, C., Yang, S., Li, J., Zhang, Z., Zhao, C., Zhou, H., Li, Z., Lin, H., Wang, X., Wang, J., Han, A., Chan, R.C.K., Liang, L., Zhang, X., Chen, H.: A multimodal knowledge-enhanced whole-slide pathology foundation model. Nature Communications **16**, 11406 (2025)
101. Yan, F., Wu, J., Li, J., Wang, W., Lu, J., Chen, W., Gao, Z., Li, J., Yan, H., Ma, J., et al.: Pathorchestra: A comprehensive foundation model for computational pathology with over 100 diverse clinical-grade tasks. arXiv preprint arXiv:2503.24345 (2025)
102. Zbontar, J., Jing, L., Misra, I., LeCun, Y., Deny, S.: Barlow twins: Self-supervised learning via redundancy reduction. In: International Conference on Machine Learning. pp. 12310–12320. PMLR (2021)
103. Zhang, A., Jaume, G., Vaidya, A., Ding, T., Mahmood, F.: Accelerating data processing and benchmarking of ai models for pathology. arXiv preprint arXiv:2502.06750 (2025)
104. Zhang, H., Meng, Y., Zhao, Y., Qiao, Y., Yang, X., Coupland, S.E., 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. pp. 18802–18812 (2022)
105. Zhou, J., Wei, C., Wang, H., Shen, W., Xie, C., Yuille, A., Kong, T.: ibot: Image bert pre-training with online tokenizer. arXiv preprint arXiv:2111.07832 (2021)
106. Zhu, M., Ren, B., Richards, R., Suriawinata, M., Tomita, N., Hassanpour, S.: Development and evaluation of a deep neural network for histologic classification of renal cell carcinoma on biopsy and surgical resection slides. Scientific reports **11**(1), 1–9 (2021)
107. Zimmermann, E., Vorontsov, E., Viret, J., Casson, A., Zelechowski, M., Shaikovski, G., Tenenholtz, N., Hall, J., Klimstra, D., Yousfi, R., et al.: Virchow2: Scaling self-supervised mixed magnification models in pathology. arXiv preprint arXiv:2408.00738 (2024)