Patient-Level Representation Learning for Computational Pathology

Yousef Hassan

A Thesis
in
The Department
of
Computer Science and Software Engineering

Presented in Partial Fulfillment of the Requirements
For the Degree of
Master of Computer Science (Computer Science) at
Concordia University
Montréal, Québec, Canada

April 2026

CONCORDIA UNIVERSITY
School of Graduate Studies

This is to certify that the thesis prepared

By: **Yousef Hassan**

Entitled: **Patient-Level Representation Learning for Computational Pathology**

and submitted in partial fulfillment of the requirements for the degree of

Master of Computer Science (Computer Science)

complies with the regulations of this University and meets the accepted standards with respect to originality and quality.

Signed by the final examining committee:

_____ Chair
Dr. Yang Wang

_____ Examiner
Dr. Yang Wang

_____ Examiner
Dr. Eugene Belilovsky

_____ Thesis Supervisor
Dr. Mahdi Hosseini

_____ Co-supervisor
Dr. Chris Pal

Approved by _____
Dr. Denis Pankratov, Graduate Program Director

April 17, 2026 _____
Dr. Mourad Debbabi, Dean
Gina Cody School of Engineering and Computer Science

Abstract

Patient-Level Representation Learning for Computational Pathology

Yousef Hassan

Computational pathology requires whole-slide image (WSI) foundation models that transfer across diverse clinical tasks, yet current approaches remain largely slide-centric, often depend on proprietary data and expensive supervision from paired textual reports that are not publicly available, and do not explicitly model relationships among multiple slides from the same patient. This thesis presents MOOZY, a patient-first pathology foundation model in which the patient case, not the individual slide, serves as the fundamental unit of representation. Unlike existing methods that encode slides independently and merge their embeddings post-hoc, MOOZY explicitly models dependencies across all slides from the same patient via a dedicated case transformer during pretraining.

MOOZY follows a two-stage design that combines multi-stage open self-supervision with scaled low-cost task supervision. In Stage 1, a vision-only slide encoder is pretrained on 77,134 public slide feature grids using masked self-distillation to learn robust, context-aware representations. In Stage 2, these representations are aligned with clinical semantics through a case transformer and multi-task supervision over 333 tasks from 56 public datasets, spanning 205 classification tasks and 128 survival tasks that cover overall survival, disease-specific survival, disease-free interval, and progression-free interval across 23 anatomical sites.

Across eight held-out tasks evaluated with five-fold frozen-feature probing, MOOZY achieves best or tied-best performance on the majority of metrics, improving macro averages over TITAN by +7.37%, +5.50%, and +7.83% and over PRISM by +8.83%, +10.70%,

and +9.78% for weighted F1, weighted ROC-AUC, and balanced accuracy, respectively. MOOZY is also parameter-efficient, with 85.77M total parameters, 14.23 times smaller than GigaPath, while maintaining strong performance. These results demonstrate that open and reproducible patient-level pretraining is sufficient to produce transferable, generalizable, and parameter-efficient embeddings, providing a practical path toward scalable patient-first histopathology foundation models.

Acknowledgments

To my father, my first teacher and my greatest hero. He is an intensive care unit doctor who has spent his entire life in service of others, healing patients for as long as I can remember and continuing to do so to this day at the age of sixty-four. When I was a boy, I used to sneak into his room and sit quietly in the corner, watching him study. He would come home after long hours at the hospital, and instead of resting, he would open his books again, night after night, year after year, preparing to become a consultant doctor. That dream did not come easily, but he never stopped trying, and in watching him refuse to give up, he handed me something far greater than any title. He taught me that the pursuit of knowledge is its own reward, and that curiosity is a light you never let go out. "Ask questions," he would tell me. "Always ask questions." A man who has dedicated his life to healing others, and who, without knowing it, shaped the scientist I am still becoming. If there is courage in this thesis, it is borrowed from him.

To my mother, the gentlest voice across the widest ocean. Across two years and an ocean between us, I saw her only once, but somehow her love crossed every border and every time zone to find me on the days I needed it most. When I stumbled, she did not try to catch me. She simply reminded me who I am. "You will figure it out," she would say, her voice steady and warm through the phone, "it is just about time." She was never wrong. Whatever strength people see in me, she built it, quietly, from the other side of the world.

To my brothers, Marwan and Ziad, the two people who have known me longer than memory itself. There were nights in Montréal when the silence of being far from home

grew too heavy, and then, as if carried by some invisible thread that connects us, my phone would ring, and it would be one of them. They never needed me to explain. They just knew.

To my dearest friend, Yousef Walid, who has a way of seeing the world that makes everything clearer just by hearing him talk. He carried my doubts as if they were his own and handed them back to me lighter than he found them. If I ever sounded sure of myself during this degree, it is because somewhere, a few hours earlier, I had talked it through with him. To Seif, who turned a foreign city into a home, and with whom every corner of Montréal holds a story that belongs to both of us, and I would not trade a single one. To Nour, who sat beside me through countless lunches at Mila and long nights of research when the rest of the building had gone dark, making the solitude of science feel a little less solitary. She was always the first to share every release of my work on Linked In, and I will never forget that. To Mai, who sat next to my desk at Mila, filled its corridors with laughter, and turned working hours into long conversations that wandered into Jordanian words I had never heard before. Mila was always missing something on the days she was not there.

To my supervisor, Mahdi Hosseini for supervising the work during my Master's degree, and to my co-supervisor, Chris Pal, whose ideas and way of thinking through problems shaped how I approach science.

And finally, to my inner child, the boy who used to sit in the corner of his father's room, watching in silence, not yet understanding what he was looking at but already knowing that he wanted to spend his life looking. The same boy who, thanks to a computer his grandfather gave him long before most children ever touched one, sat for hours in front of a glowing screen, not playing but exploring, clicking through a world that had just opened its doors to him, teaching himself things no one around him had thought to teach. The early internet was a wilderness, and he wandered into it like a child wandering into a forest. That machine was his first laboratory, and he never really left it. He is still here. He never left.

I see him, fully, standing right in front of me, just as he was, small and wide-eyed and still convinced that the world is worth understanding. He does not know what a thesis is. He does not know what a deadline is. He only knows that there is more to see, more to ask, more to understand, and that has always been enough for him. The same boy who was too young and too short to reach the light switch, so he would grab his slipper, stand back, sprint across the room, and smash that button with everything he had, just to turn the light on. And he always did, at least that is what my mother says. I have always loved to call him "the child who runs with his slipper to the light switch.". This one is for you, buddy.

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`).

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

1.1.1 Clinical Context

Pathology is the branch of medicine concerned with the diagnosis and characterization of disease through the examination of tissue specimens [54]. When a patient undergoes a biopsy or surgical procedure, the excised tissue is processed into thin sections, stained with chemical dyes to reveal cellular structures, and mounted onto glass slides for microscopic examination. The most widely used staining protocol is hematoxylin and eosin (H&E) [30], where hematoxylin stains cell nuclei blue-purple and eosin stains cytoplasm and extracellular structures pink. A pathologist then examines these stained slides under a microscope to identify cellular abnormalities, determine the presence or absence of malignancy, classify tumor type and subtype, assign histologic grade, evaluate surgical margins, and assess other features that directly inform clinical decision-making. Histologic grade describes how abnormal and aggressive the tumor appears under the microscope.

This diagnostic process is fundamental to nearly every aspect of cancer care. Treatment selection, surgical planning, prognosis estimation, and eligibility for targeted therapies all

depend on the pathologist's interpretation of tissue morphology, meaning the visible structure and appearance of cells and tissue under the microscope [103, 1]. In many clinical workflows, a single patient may have multiple tissue slides prepared from different tissue blocks, meaning separate pieces of sampled tissue that are processed into slides, anatomical sites, or staining protocols, and the pathologist must integrate findings across all of these slides to form a coherent diagnostic conclusion at the patient level. For example, in a breast cancer surgery case, the pathologist may review slides from the main tumor, the surgical margins, and one or more lymph nodes, then combine these findings to determine whether tumor remains at the edge of resection, whether disease has spread beyond the breast, and what this means for staging and treatment.

The demand for pathology services continues to grow as the global incidence of cancer rises, with an estimated 19.3 million new cases diagnosed worldwide in 2020 alone [98], yet the supply of trained pathologists has not kept pace. Documented workforce shortages and declining per-capita pathologist counts have been reported across multiple countries [68], creating bottlenecks in diagnostic throughput. Moreover, histopathologic interpretation is inherently subjective, and studies have demonstrated significant inter-observer variability among pathologists even for common diagnostic tasks such as breast biopsy interpretation [25]. These factors collectively motivate the development of computational tools that can assist pathologists by providing consistent, reproducible, and scalable analysis of tissue specimens [40, 103].

1.1.2 The Clinical Diagnostic Workflow

The design choices in this thesis are motivated by the way a cancer diagnosis is actually produced in practice. A diagnosis does not arise from the interpretation of a single slide, but from the pathologist's integration of evidence across multiple slides acquired at different anatomic sites and, frequently, at different points in time [58, 54]. This subsection briefly

traces that workflow from specimen acquisition to the final synoptic report, because two observations about it, multi-slide sampling and case-level synthesis, directly motivate the patient-first framing adopted in the remainder of this thesis.

When a patient is worked up for a suspected malignancy, tissue is sampled through one or more procedures such as core biopsy, excisional biopsy, surgical resection, or lymph node dissection [58]. Each specimen is grossed into multiple tissue blocks that sample representative regions of the lesion, surgical margins, and adjacent anatomy, and each block is sectioned and stained, producing one or more slides per block. A single case therefore commonly comprises multiple slides spanning main-tumor regions, peritumoral stroma, resection margins, and regional lymph nodes. Where molecular or protein markers are clinically indicated, selected blocks are further stained with immunohistochemistry (for example, ER, PR, HER2, and Ki-67 in breast cancer, or p53, PD-L1, and mismatch-repair proteins in other organs), producing additional slides that complement, rather than replace, the morphologic H&E series.

Slides for a given patient are also not produced at a single instant. A typical oncologic trajectory begins with a small diagnostic biopsy used to confirm malignancy and guide initial treatment selection, followed by a larger resection specimen that supports definitive staging, and in many cases by subsequent specimens acquired during surveillance, recurrence workup, or metastasis sampling [3, 58]. Each encounter contributes additional slides to the same case, and the combined interpretation across these time points, rather than any single slide in isolation, is what informs decisions about staging, adjuvant therapy, and prognostic estimation. The patient case is therefore an inherently longitudinal, multi-slide object.

At the microscope, the pathologist reviews the entire set of slides belonging to a case and forms a single diagnostic conclusion by reasoning across them jointly. Different slides contribute different kinds of evidence. Main-tumor slides establish histologic type and

grade, margin slides determine whether tumor is present at the edge of resection, lymph node slides establish nodal involvement for staging, and immunohistochemistry slides provide biomarker status that informs targeted therapy eligibility. In settings such as multifocal disease, intratumoral heterogeneity, or synchronous lesions across anatomic sites, the diagnostic signal resides in the relationships between slides (for example, whether two foci share the same histologic subtype, or whether a nodal deposit is morphologically consistent with the primary tumor), and cannot be recovered by interpreting any slide alone.

The outcome of this cross-slide integration is a structured synoptic report, a standardized document that lists the diagnostic, prognostic, and predictive variables required for patient management, including histologic type and grade, tumor size, invasion patterns, margin status, nodal involvement, pathologic stage (pTNM), and relevant biomarker results [96, 84]. Widely adopted synoptic templates, such as the College of American Pathologists cancer protocols, are anchored to the AJCC staging system [3]. The synoptic report is defined at the level of the patient case, because every element it contains is derived from integrating findings across all slides in that case, and it is this case-level synthesis, not any single slide, that is ultimately communicated to the treating oncologist and used to select treatment.

Two design implications follow from this workflow. First, because the fundamental unit of clinical decision-making is the patient case, a model that encodes slides independently and merges them only post-hoc [90, 57, 102, 116] does not faithfully replicate the pathologist's diagnostic process. Explicitly modeling dependencies between slides of the same patient, as MOOZY does via the case transformer (Chapter 3), mirrors the cross-slide integration step that pathologists already perform at the microscope. Second, because the synoptic report spans diverse variable types, categorical diagnostic calls, ordinal grades, TNM categories, binary biomarker statuses, and time-to-event survival outcomes, a single task-specific objective cannot recover the breadth of information the clinical workflow is

designed to produce. The Stage 2 multi-task supervision used in this thesis, covering 333 tasks with 205 classification and 128 survival endpoints across 56 public datasets (Section 4.1), is a deliberate attempt to align the pretraining objective with the information content of a synoptic report rather than with any single clinical endpoint in isolation. The slide collection of each case is modeled as an unordered set aggregated by the case transformer. Explicit modeling of temporal ordering across specimens from different encounters is beyond the scope of this thesis and is flagged as a direction for future work (Chapter 6).

1.1.3 Whole-Slide Imaging and the Computational Challenge

The advent of high-resolution digital slide scanners has enabled the digitization of glass slides into whole-slide images (WSIs) [79, 26]. A WSI is a high-resolution digital scan of an entire glass slide, typically captured at 20× or 40× magnification. At these magnifications, a single WSI routinely contains tens of thousands to over one hundred thousand pixels along each spatial axis, resulting in images that range from hundreds of millions to several billion pixels in total [73]. This gigapixel scale is illustrated in Figure 1, which shows a representative TCGA breast carcinoma slide spanning over 86,000×66,000 pixels. This gigapixel scale distinguishes WSIs from the natural images commonly used in computer vision research, which typically range from a few hundred to a few thousand pixels per side.

The sheer size of WSIs poses a fundamental computational challenge. Standard deep learning architectures for image classification and recognition, such as convolutional neural networks [37] and vision transformers [23], are designed to operate on images of modest resolution. A single WSI exceeds this input size by several orders of magnitude, making direct end-to-end processing infeasible with current hardware. The dominant solution in the field is to partition each WSI into a grid of small, non-overlapping patches (also called tiles), each typically 224×224 pixels or 512×512 pixels, and to process these patches

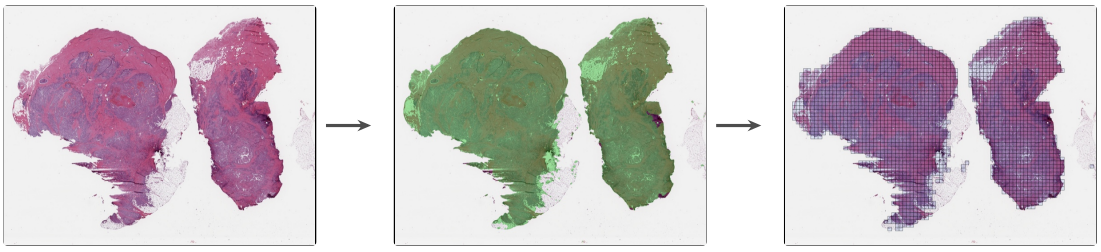


Figure 1: Standard WSI preprocessing pipeline. **(a)** A whole-slide image at full resolution (86,632 × 66,370 pixels). **(b)** Tissue detection segments foreground tissue from the background. **(c)** Patch coordinate extraction identifies 2,386 non-overlapping 512×512 tile locations at $20\times$ magnification within the detected tissue regions.

independently through a pretrained image encoder. A single WSI may yield thousands to tens of thousands of such patches, each capturing a small region of tissue at cellular resolution [52, 61].

However, processing patches independently discards the spatial relationships between tissue regions that carry critical diagnostic information. In clinical practice, a pathologist does not evaluate cells in isolation. The diagnostic interpretation depends on the spatial organization of tissue, including the arrangement of glandular structures, the interface between tumor and stroma, where stroma refers to the connective and supportive tissue around the tumor, the distribution of immune cells across the tissue landscape, and the overall architectural pattern of the specimen. Capturing these long-range interactions requires models that can reason over entire slides, not just individual patches [97, 66].

1.1.4 Levels of Representation in Computational Pathology

Representation learning in computational pathology [40, 1] operates at three distinct levels of granularity, each corresponding to a different unit of analysis, as illustrated in Figure 2. In this context, a representation, often called an embedding, is a numerical summary that encodes information from an image or a collection of images. Aggregation refers to the

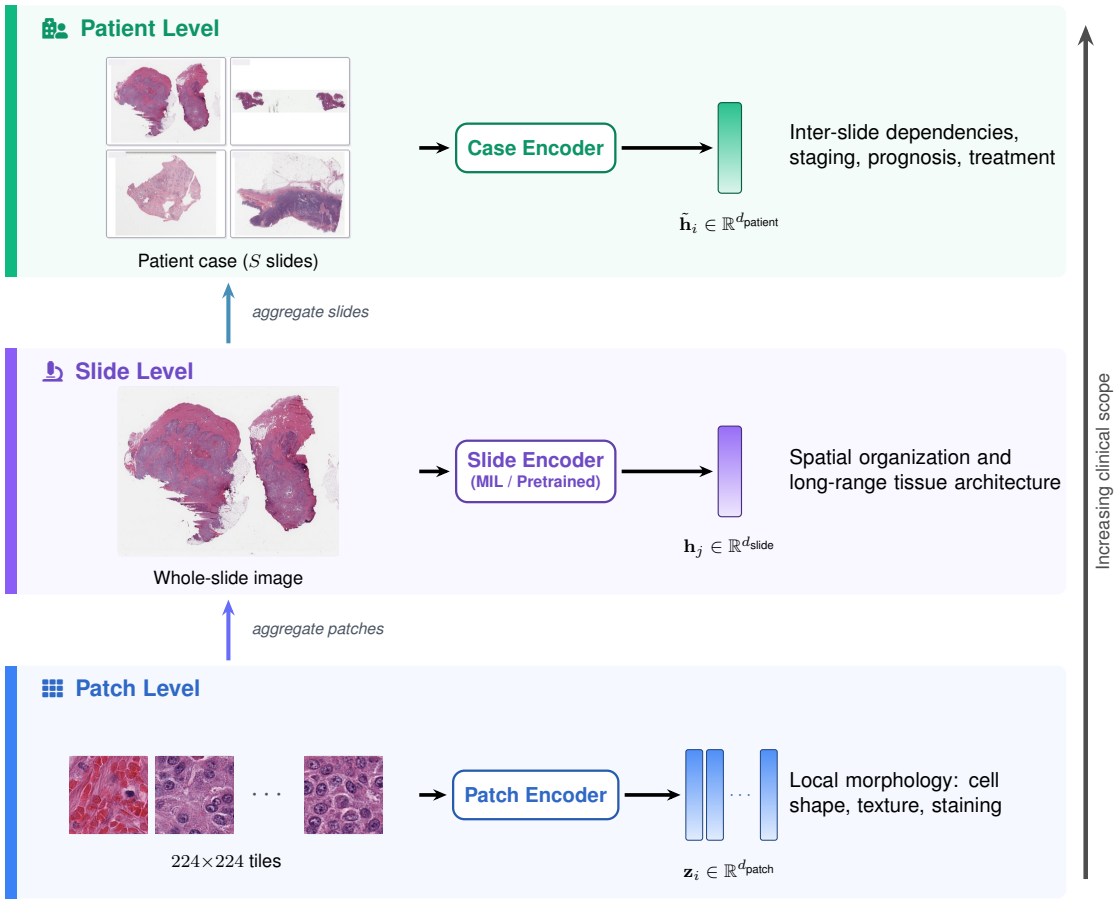


Figure 2: Levels of representation in computational pathology. At the patch level, individual tiles are mapped to feature vectors by a patch encoder. At the slide level, a slide encoder aggregates all patch embeddings from a whole-slide image into a single slide representation. At the patient level, a case encoder integrates slide embeddings from all slides belonging to a patient into a unified patient-level representation. Each successive level captures increasingly broader clinical context.

process of combining many lower-level representations into one higher-level summary.

Patch-level representations. At the finest level, a patch encoder maps each individual tile to a fixed-dimensional feature vector. These encoders are typically vision transformers [23] pretrained with self-supervised learning on large collections of pathology patches [47, 28]. The field has largely converged on the DINOv2 [78] pretraining recipe, combining self-distillation, masked image modeling, and uniformity regularization, and has scaled architectures from ViT-Large [16, 29] to ViT-Giant [114, 70] trained on up to millions of slides [105, 122]. The resulting patch features capture local morphological patterns such as cell shape, nuclear texture, staining intensity, and tissue microarchitecture within the 224×224 or 512×512 field of view. Patch-level encoders form the foundation of most computational pathology pipelines, as they provide the input features on which all higher-level representations are built.

Slide-level representations. At the intermediate level, a slide encoder or slide aggregator combines the thousands of patch features from a single WSI into a single slide-level representation [40]. The classical approach to this aggregation is multiple instance learning (MIL) [44, 65, 92], which treats the WSI as a "bag" of patch instances associated with a single slide-level label. MIL aggregators learn to pool patch features into a slide-level embedding through mechanisms such as attention-weighted summation [44, 65], transformer-based inter-patch modeling [92], permutation-invariant pooling [12], and dual-stream objectives [59]. Despite their architectural diversity, all MIL aggregators share a fundamental limitation: each is trained from scratch for each downstream task. The composition rules that the aggregator learns, including which patch interactions are diagnostically relevant and how spatial patterns should be weighted, are specific to the training objective and do not generalize across label spaces. An aggregator trained for breast cancer subtyping cannot be reused for lung mutation prediction without complete retraining, and every new

clinical endpoint requires its own labeled dataset and its own aggregator.

This task-specific regime motivated the development of pretrained slide encoders [15, 115, 22, 90], which learn general-purpose context modeling on large collections of unlabeled or weakly labeled slides before any downstream task is defined. The core idea mirrors the transition that previously occurred at the patch level: just as pretrained patch encoders eliminated the need to train a convolutional network from scratch for each tile-level task, pretrained slide encoders aim to eliminate the need to train an aggregator from scratch for each slide-level task. By applying self-supervised objectives such as self-distillation [15, 14], masked prediction [115], or multimodal alignment with clinical text [90, 22] directly to the set of patch features that constitute a slide, these models learn how patches relate to one another in tissue, capturing spatial organization and long-range dependencies as reusable structural knowledge rather than as a byproduct of a specific classification objective. The resulting slide embedding can then be applied to diverse downstream tasks with a simple linear or nonlinear probe, without retraining the encoder itself.

Patient-level (case-level) representations. At the coarsest and most clinically relevant level, a patient-level or case-level representation integrates information from all slides belonging to a single patient into a unified embedding [40]. In clinical pathology, a "case" refers to the complete set of tissue specimens and associated clinical information for a single patient encounter. A patient case may contain one slide or many slides, depending on the number of tissue blocks sampled, the number of anatomical sites biopsied, and the staining protocols applied. Clinical decisions such as cancer staging, treatment selection, and prognosis estimation are inherently patient-level tasks that require integrating evidence across all available slides. Despite this clinical reality, most existing computational pathology models operate at the slide level and handle multi-slide patients through simple post-hoc strategies, such as concatenating patch features from all slides into a single enlarged bag (early fusion) or averaging the independently computed slide embeddings

or predictions across slides (late fusion) [90, 57, 102, 116]. These fusion strategies treat the collection of slides as an unordered pool and do not model the relationships between slides, discarding inter-slide interactions that may carry diagnostic signal in settings such as multifocal staging, heterogeneity assessment, and prognosis.

1.1.5 From Task-Specific Pipelines to Foundation Models

Historically, computational pathology has advanced through task-specific and cohort-specific supervised pipelines [5, 11, 19, 50, 94]. In this paradigm, a model is trained from scratch for each combination of organ, cancer type, and clinical endpoint, using manually annotated data from a specific patient cohort. While these pipelines have achieved strong performance on their target tasks, they must be rebuilt whenever the organ, scanner domain, or clinical objective changes, limiting scalability and reuse across the diverse landscape of pathology applications.

A key enabler of this shift is self-supervised learning (SSL), a training paradigm in which a model learns representations from unlabeled data by solving pretext tasks derived from the data itself, rather than relying on manually provided labels. Common pretext tasks include predicting masked or corrupted portions of an input, enforcing consistency between different augmented views of the same sample, and matching representations across modalities. The broader machine learning community has demonstrated that scaling data and compute in self-supervised pretraining can yield general-purpose representations with strong task-agnostic transferability [10, 48, 39, 109]. In natural language processing, large language models pretrained on internet-scale text have shown emergent capabilities across diverse tasks without task-specific supervision. Analogous trends have emerged in computer vision, where self-supervised methods such as contrastive learning [17, 36, 32, 18], self-distillation [14, 78], and masked image modeling [35, 120] have produced encoders that transfer broadly across vision tasks.

These advances have motivated the development of foundation models for pathology, defined as large pretrained models that learn general-purpose representations from unlabeled or weakly labeled pathology data, with the goal of transferring to diverse clinical tasks with minimal adaptation. The field has progressed through successive levels of scale. Early pathology foundation models operated at the tile level, converging on the DINOv2 [78] pretraining recipe and scaling from ViT-Large [16, 29] to ViT-Giant backbones [114, 70] trained on up to millions of slides [105, 122]. In typical pipelines that use these tile encoders, patch features are extracted once and a separate MIL aggregator [44, 65, 92] is trained for each downstream task, necessitating retraining whenever the clinical endpoint changes.

More recently, the community has moved toward slide-level foundation models that pretrain whole-slide representations, reducing reliance on task-specific MIL training. These models can be broadly grouped into three families: vision-only self-supervised methods that learn from the structure of unlabeled slides [15, 55, 41, 115, 6, 57], multimodal approaches that align slides with paired text from clinical reports [90, 22, 112], genomic profiles [45, 116, 102], or cross-stain views [46, 42], and supervised methods that learn from task labels [74, 107].

1.1.6 Limitations of Current Approaches

Despite significant progress in slide-level foundation models, three structural limitations persist across the field.

Reproducibility. Many top-performing models are trained on proprietary datasets that are not publicly available. Some also rely on paired pathology reports or other rich clinical text supervision to train slide encoders, yet these reports are expensive to curate and are usually not publicly released with the corresponding slides [22, 90]. In addition, some

methods do not release pretrained checkpoints [41, 102] or complete training recipes [22, 102, 90]. This limits the ability of the research community to reproduce, validate, and build upon these methods, creating a barrier to scientific progress.

Capacity allocation. Current architectures concentrate the majority of their model capacity in heavyweight tile encoders [90, 22, 46, 102], often with hundreds of millions to over a billion parameters, while using comparatively lightweight slide aggregators. This allocation is misaligned with the nature of the problem. The key challenge in WSI understanding is capturing long-range spatial context and tissue-level organization, not encoding increasingly fine-grained per-tile morphology. Notably, recent work has shown that public-only tile encoders can match much larger systems trained on proprietary data [49, 95], and we hypothesize that this saturation is fundamental: H&E-stained tissue occupies a far more constrained visual space than natural images, with a narrow color palette and a bounded set of morphological primitives, so tile-level representations approach a performance ceiling well before general-vision thresholds.

Naive multi-slide fusion. While some models accommodate multiple slides per patient, they rely on simple fusion heuristics: early fusion by concatenating or pooling patch features from all slides into one enlarged bag, or late fusion by averaging slide-level embeddings or predictions across slides [90, 57, 102, 116]. These strategies treat a patient case as an unordered collection rather than explicitly modeling the dependencies between slides. This design discards cross-slide interactions that carry diagnostic signal. Such interactions matter in clinical settings involving multifocal disease, where distinct tumor foci may appear in separate samples, tumor heterogeneity, where different slides can reveal different morphologic patterns of the same disease, and multi-site biopsy interpretation.

These three gaps collectively point to the need for a patient-level foundation model that is open, reproducible, parameter-efficient, and capable of explicitly modeling inter-slide

dependencies during pretraining rather than through post-hoc fusion.

1.1.7 Problem Definition and Research Question

The problem addressed in this thesis is the following: given a patient case containing one or more whole-slide images, how can we learn a reusable representation that preserves diagnostically relevant structure within each slide, captures relationships across slides from the same patient, and transfers effectively across many downstream clinical tasks. This problem must be solved under practical constraints that matter to the field, including limited access to proprietary data, heterogeneous public supervision, and the need for reproducible training recipes.

Existing pathology foundation models address only part of this problem. Patch encoders learn local morphology but do not reason over full slides. Slide encoders model within-slide context but usually treat each slide independently. When multiple slides are available for a patient, they are commonly merged with simple early or late fusion rather than with an explicit patient-level model. The central research question of this thesis is therefore whether a pathology foundation model trained entirely on public data can learn a transferable patient-level representation that explicitly models relationships across a patient's slides and improves over slide-centric alternatives on diverse downstream tasks.

1.2 Thesis Statement

This thesis proposes MOOZY (Multi-stage Open self-supervised pretraining with lOw-cost supervision at siZe for patient-aware histopathologY), a *patient-first* foundation model for computational pathology in which 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 demonstrate that a two-stage framework, decoupling vision-only slide-level self-supervised learning from patient-aware multi-task semantic alignment, produces transferable, generalizable, and parameter-efficient embeddings that achieve competitive or superior performance across diverse clinical tasks, entirely from public data without proprietary slides, paired clinical reports, or billion-parameter architectures.

1.3 Objectives and Contributions

The contributions of this thesis can be summarized as follows:

- We propose a two-stage framework that decouples vision-only slide SSL pretraining (masked self-distillation on 77,134 unlabeled public slides) from patient-aware semantic alignment, where a case-level aggregator explicitly models dependencies across all slides of the same patient, making MOOZY the first open and reproducible attempt to move pathology foundation 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 (OS, DSS, DFI, PFI) across 23 anatomical sites, requiring harmonization of heterogeneous annotation formats, clinical records, and cohort conventions, entirely from public data without private slides, paired reports, or expert annotations.

- We provide comprehensive quantitative and qualitative evaluation by benchmarking MOOZY on eight held-out tasks against both slide encoders and MIL baselines, complemented by attention map analysis and embedding visualization, demonstrating that open patient-level pretraining yields competitive, transferable, and parameter-efficient representations.

1.4 Outline

The remainder of this thesis is organized as follows. Chapter 2 reviews the related literature, covering self-supervised learning, pathology patch encoders, multiple instance learning, and slide-level encoders. Chapter 3 presents the proposed MOOZY framework in detail, including the Stage 1 self-supervised slide encoder pretraining and Stage 2 patient-aware semantic alignment, along with all architectural and algorithmic details. Chapter 4 describes the experimental setup, including dataset construction, task preparation, hyperparameter configurations, and evaluation protocols. Chapter 5 presents the experimental results, including comparisons with slide encoders and MIL baselines, ablation studies, attention map analysis, and qualitative embedding visualizations. Chapter 6 concludes the thesis and discusses future research directions.

Chapter 2

Related Work

2.1 Self-Supervised Learning Pretraining

Representation learning has undergone a major shift: scaling data and compute has shown that a single foundation model can support diverse behaviors with minimal task-specific supervision, as demonstrated most prominently in large language models [10, 48, 39, 109]. Analogous scaling trends have transferred to vision, where large-scale self-supervised pretraining has been central to developing task-agnostic visual representations. Contrastive approaches such as SimCLR [17] and MoCo [36], non-contrastive methods like BYOL [32], Barlow Twins [118], and SimSiam [18], and online prototype learning (SwAV [13]) have advanced general-purpose encoders. More recently, masked image modeling and self-distillation via a teacher-student framework (MAE [35], iBOT [120], DINO [14], DINOv2 [78], DINOv3 [93]) have demonstrated strong transferability and zero-shot capacity across vision tasks.

2.2 Pathology Patch Encoders

Pathology-specific SSL on public tiles outperforms ImageNet initialization [47, 28], leveraging the vision SSL advances described above. The field has converged on the DINOv2 recipe [78], combining self-distillation, masked image modeling, and KoLeo regularization [53, 87], with modifications such as KDE-based uniformity objectives [106]. Vision–language alignment [64, 22] and knowledge distillation [27] serve as complementary directions. Architectures have scaled from ViT-Large [16, 29, 117] to ViT-Huge [105, 122, 16] and ViT-Giant [114, 88, 8, 70], trained on both proprietary [16, 105, 70] and public data [69, 24, 33, 28, 29].

Yet scaling laws for tile encoders remain unclear: public-only models match much larger systems [49, 95], suggesting that benchmark discriminability [108] and training-recipe effects dominate data volume beyond a modest threshold. We hypothesize this saturation is fundamental: 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*: aggregating heterogeneous tiles into whole-slide representations that capture spatial organization and long-range dependencies, motivating the slide-level pretraining framework proposed in this work.

2.3 Multiple Instance Learning

Multiple instance learning (MIL) treats a WSI as a bag of patch features with a single slide-level label. Approaches span permutation-invariant pooling [12], attention scoring [44, 65], transformer-based inter-patch modeling [92, 104], dual-stream objectives [59], pseudo-bag augmentation [119], and efficient variants via low-rank approximations [113], knowledge

graphs [60], and regional re-embedding [100]. Despite their architectural diversity, all these aggregators are trained from scratch for each downstream task, learning composition rules that do not generalize across label spaces. This task-specific regime motivates the shift toward pretrained slide encoders that learn universal whole-slide representations, decoupling context modeling from downstream supervision.

2.4 Slide Encoders

Slide-level pretraining operates on unordered sets of thousands of heterogeneous tile embeddings. Vision-only methods apply self-distillation [15, 14], contrastive tile sampling [55, 17], view transformations [41], dilated attention in masked autoencoders [115, 21], lightweight contextualizers [6], and state-space contrastive learning [57, 34]. Multimodal methods align slides with clinical text [90, 22, 112, 63], genomic or transcriptomic profiles [45, 116, 102], or cross-stain sections [46, 42]. Supervised approaches train on slide-level labels [74, 75, 107].

Three structural gaps persist across all families. First, reproducibility is limited by proprietary data and withheld checkpoints or training recipes [41, 102, 22, 90]. Second, capacity concentrates in heavyweight tile encoders rather than slide aggregators [90, 22, 46, 102], despite clinically relevant structure arising primarily from long-range spatial organization. Third, multi-slide fusion remains naive, relying on bag union or embedding averaging [90, 57, 102, 116], treating cases as unordered pools rather than explicitly modeling inter-slide relationships. Our framework addresses all three: a two-stage design decouples vision-only SSL pretraining from patient-aware multi-task alignment, replacing naive multi-slide fusion with explicit inter-slide dependency modeling at the case level, while training entirely on public data with a fully released recipe.

Chapter 3

Methodology

This chapter presents the proposed MOOZY framework in detail. MOOZY is a two-stage patient-first foundation model for computational pathology. 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 steers these representations toward clinical semantics through large-scale multi-task supervision, directly benefiting from the generalizable prior built in Stage 1. Critically, Stage 2 moves beyond per-slide encoding: 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.

3.1 Stage 1: Self-Supervised Slide Encoder Pretraining

3.1.1 Input Representation

We cast WSI representation learning as self-supervised pretraining on precomputed patch features (Figure 3, top). Given a WSI $\mathcal{W}$, we partition tissue into non-overlapping 224-pixel patches and extract features with a frozen patch encoder f_{patch}:

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

Spatial Grid Construction

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. For patches at level-0 coordinates (x_i, y_i), let $(x_{\min}, y_{\min})$ be the minimum coordinates and Δ the uniform patch spacing. The grid position (r, c) for patch i is:

$$r = \left\lfloor \frac{y_i - y_{\min}}{\Delta} + 0.5 \right\rfloor, \quad c = \left\lfloor \frac{x_i - x_{\min}}{\Delta} + 0.5 \right\rfloor. \tag{2}$$

Grid positions without tissue are filled with zero vectors, and a binary validity mask $\mathbf{V} \in \{0, 1\}^{H \times W}$ tracks tissue presence. If a slide is available at multiple magnifications, each level-specific grid is treated as an independent training sample.

3.1.2 Multi-Scale Crop Sampling

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 $\rho_{\min}$. For a crop $\mathbf{C}$ of size $S \times S$ with corresponding validity mask $\mathbf{V}_{\mathbf{C}}$, the minimum valid-token constraint is:

$$\frac{\sum_{r,c} V_{\mathbf{C},r,c}}{S^2} \geq \rho_{\min}. \tag{3}$$

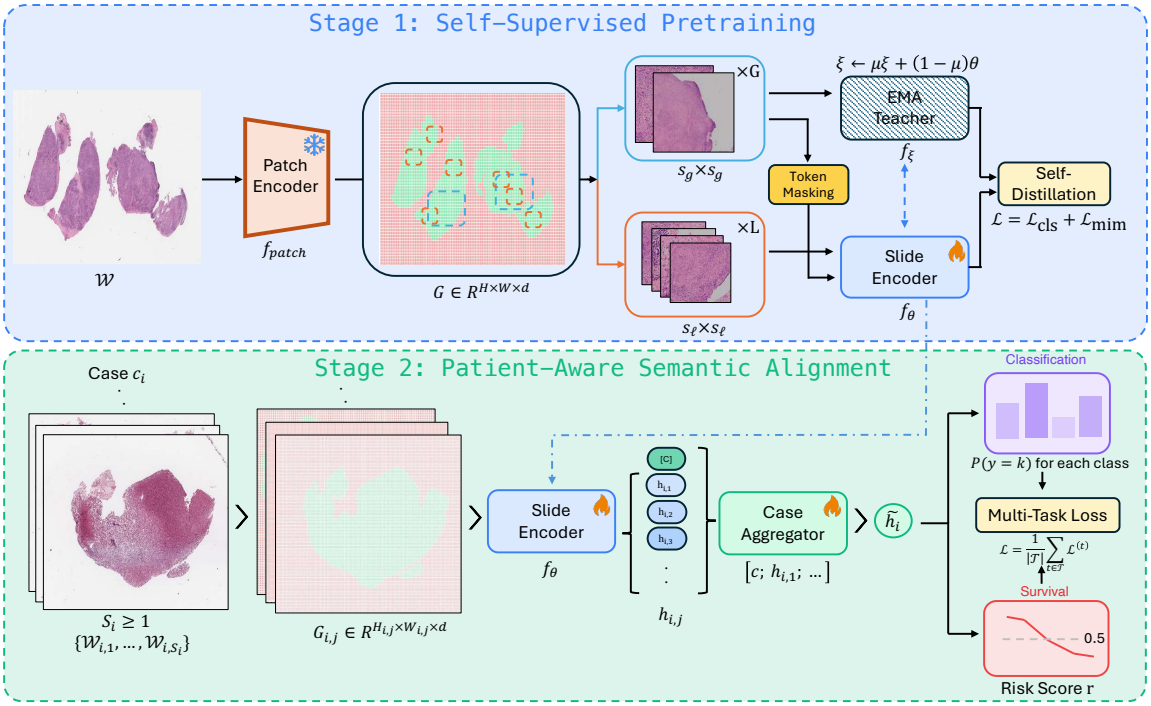


Figure 3: Overview of the proposed two-stage framework. *Stage 1 (top):* A frozen patch encoder extracts per-patch features arranged into a spatial grid. Multi-scale crops are sampled with spatial augmentations and block-based masking. A student slide encoder and EMA teacher are jointly trained via CLS-level self-distillation ($\mathcal{L}_{\text{cls}}$) and masked patch prediction ($\mathcal{L}_{\text{mim}}$). *Stage 2 (bottom):* 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.

Crops failing this criterion are resampled up to a fixed maximum number of attempts. Unlike [22], 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 [41].

3.1.3 Block-Based Masking

We apply DINOv3-style block masking [93] 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.

Within each batch, a fraction p_{mask} of global crops are selected for masking. Their mask ratios are distributed as $\gamma_1, \ldots, \gamma_n = \text{linspace}(\gamma_{\min}, \gamma_{\max}, n)$ and randomly shuffled across the selected crops, ensuring uniform coverage of the masking spectrum within each batch.

For each selected global crop $\mathbf{C}_j^g$ with assigned ratio γ_j, a binary mask $\mathbf{M}_j \in \{0,1\}^{S_g \times S_g}$ is constructed by iteratively placing rectangular blocks whose aspect ratios are drawn log-uniformly from $[\alpha_{\min}, \alpha_{\max}]$ at random positions, until the target count of masked valid tokens is reached:

$$|\{(r,c) : M_{j,r,c} = 1 \wedge V_{j,r,c} = 1\}| = \lfloor \gamma_j \cdot |\{(r,c) : V_{j,r,c} = 1\}| \rfloor. \tag{4}$$

Any remaining budget after block placement is filled by randomly selecting individual valid tokens.

3.1.4 Slide Encoder Architecture

Our slide encoder (Figure 4A) is a Vision Transformer [23] adapted to precomputed feature grids. Patch features are projected to dimension d with a linear layer and GELU [38]. We prepend a learnable `[CLS]` token and R register tokens [20], and masked student positions are replaced by a learnable mask embedding. Each block uses pre-norm multi-head self-attention and an FFN with LayerScale [101] and stochastic depth [43]. To encode spatial structure without learned positional embeddings, we use 2-D ALiBi [81] as adapted for WSIs in TITAN [22]. For each attention head h, we add:

$$b_{i,j}^{(h)} = -s_h \cdot \frac{\|\mathbf{p}_i - \mathbf{p}_j\|_2}{\Delta}, \tag{5}$$

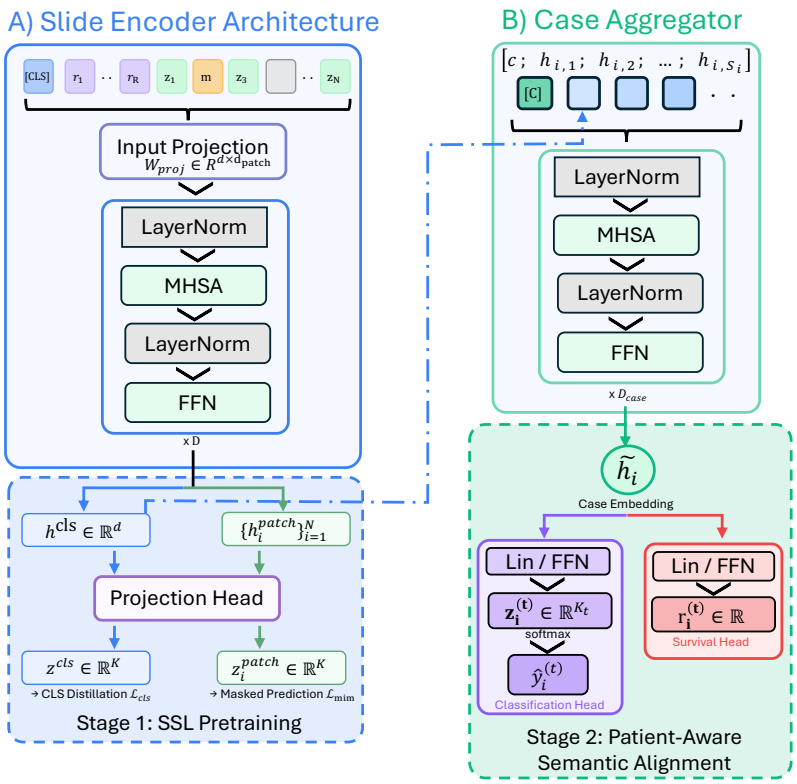


Figure 4: Architecture of the slide encoder and case aggregator. **(A)** The slide encoder takes patch embeddings, a learnable `[CLS]` token, R register tokens, and mask tokens, processed through D transformer blocks. **(B)** The case aggregator prepends a learnable `[CASE]` token to per-slide embeddings and produces a case embedding $\tilde{\mathbf{h}}_i$, routed to heads for classification and survival prediction.

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 $-\infty$.

3.1.5 Projection Head

The projection head maps encoder tokens to prototype logits with an MLP, an L2-normalized bottleneck, and a weight-normalized prototype layer [14, 120], shared for `[CLS]` and patch

tokens. For an input token embedding $\mathbf{h} \in \mathbb{R}^d$, the head computes:

$$\mathbf{u}_1 = \text{GELU}(\mathbf{W}_1^{\text{proj}}\mathbf{h} + \mathbf{b}_1^{\text{proj}}), \tag{6}$$

$$\mathbf{u}_2 = \text{GELU}(\mathbf{W}_2^{\text{proj}}\mathbf{u}_1 + \mathbf{b}_2^{\text{proj}}), \tag{7}$$

$$\mathbf{u}_3 = \frac{\mathbf{W}_3^{\text{proj}}\mathbf{u}_2 + \mathbf{b}_3^{\text{proj}}}{\|\mathbf{W}_3^{\text{proj}}\mathbf{u}_2 + \mathbf{b}_3^{\text{proj}}\|_2}, \tag{8}$$

$$\mathbf{z} = \mathbf{W}_4^{\text{proj}}\mathbf{u}_3. \tag{9}$$

Here $\mathbf{W}_1^{\text{proj}}, \mathbf{W}_2^{\text{proj}} \in \mathbb{R}^{d_h \times d}$ project to hidden dimension d_h, $\mathbf{W}_3^{\text{proj}} \in \mathbb{R}^{d_b \times d_h}$ maps to bottleneck dimension d_b before L2 normalization, and $\mathbf{W}_4^{\text{proj}} \in \mathbb{R}^{K \times d_b}$ is the weight-normalized prototype layer.

3.1.6 Self-Distillation Objective

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 [14]. 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:

$$\mathcal{L}_{\text{cls}} = -\frac{1}{G(G+L-1)} \sum_{j=1}^{G} \sum_{\substack{i=1 \\ i \neq j}}^{G+L} \sum_{k=1}^{K} P_k^{(j)} \log Q_k^{(i)}, \tag{10}$$

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:

$$\mathcal{L}_{\text{mim}} = -\frac{1}{|\mathcal{M}|} \sum_{(j,r,c) \in \mathcal{M}} \sum_{k=1}^{K} P_{r,c,k}^{(j)} \log Q_{r,c,k}^{(j)}, \tag{11}$$

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

3.2 Stage 2: Patient-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 multi-modal 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 3, bottom). 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).

3.2.1 Adaptive Token Capping

Stage 2 uses full-slide grids without crop sampling. To handle gigapixel inputs under GPU memory limits, we apply a hardware-adaptive token cap $K_{\text{max}}(\cdot)$: if valid tokens exceed K_{max}, we perform stratified random sampling to preserve whole-slide spatial coverage. Let $K = K_{\text{max}}$ and partition the valid-rank indices $\{0, \ldots, V_{i,j} - 1\}$ (in flattened raster order) into K equal-width bins, where s_b and e_b denote the start and end rank of bin b, respectively:

$$s_b = \left\lfloor \frac{b\, V_{i,j}}{K} \right\rfloor, \quad e_b = \left\lfloor \frac{(b+1) V_{i,j}}{K} \right\rfloor - 1, \quad b = 0, \ldots, K-1. \tag{12}$$

For each bin b, one offset is drawn uniformly:

$$u_b \sim \mathrm{Unif}\{0, \ldots, \max(1, e_b - s_b + 1) - 1\}, \tag{13}$$

and rank $r_b = s_b + u_b$ is retained. The final retained set $\{r_b\}_{b=0}^{K-1}$ is mapped back to original valid-token indices, yielding exactly one sampled token per bin. Retained tokens are compacted and passed to the slide encoder:

$$\mathbf{h}_{i,j} = f_\theta(\mathbf{X}^\star_{i,j}, \mathbf{P}^\star_{i,j}, \Delta_{i,j}) \in \mathbb{R}^d. \tag{14}$$

3.2.2 Case-Level Aggregation

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 4B). A learnable `[CASE]` token is prepended and processed through D_{case} pre-norm transformer blocks with LayerScale and DropPath, yielding:

$$\tilde{\mathbf{h}}_i = \mathrm{LN}(\mathbf{z}_{i,D_{\text{case}}})[0] \in \mathbb{R}^d. \tag{15}$$

We apply this aggregator to all cases, including single-slide cases ($S_i = 1$), so that the learned embedding space is always patient-centric and consistent regardless of slide count at inference.

3.2.3 Task Head Formulations

Each task $\mathcal{T}_t$ has a prediction head $g_t : \mathbb{R}^d \to \mathbb{R}^{o_t}$, either linear or MLP.

Linear head. The linear head applies feature dropout with rate p_{head} before projection:

$$g_t(\tilde{\mathbf{h}}) = \mathbf{W}_t \, \mathrm{Dropout}_{p_{\text{head}}}(\tilde{\mathbf{h}}) + \mathbf{b}_t. \tag{16}$$

MLP head. The MLP head uses LayerNorm, two hidden linear layers with GELU, and fixed internal dropout:

$$g_t(\tilde{\mathbf{h}}) = \mathbf{W}_3\phi\Big(\mathbf{W}_2\phi\Big(\mathbf{W}_1\mathrm{LN}(\tilde{\mathbf{h}})\Big)\Big), \quad \phi(\mathbf{u}) = \mathrm{Dropout}_{0.25}(\mathrm{GELU}(\mathbf{u})). \tag{17}$$

3.2.4 Classification Loss

For classification, we use weighted cross-entropy with label smoothing [99] coefficient ϵ:

$$\mathcal{L}_{\text{cls}}^{(t)} = \mathrm{CE}_{w^{(t)},\epsilon}\Big(\{\mathbf{z}_i^{(t)}, y_i^{(t)}\}_{i\in\mathcal{B}_t}\Big), \tag{18}$$

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.

3.2.5 Survival Loss

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. For each survival task t, let E_t denote the number of observed (uncensored) events in the training set. The provisional number of discrete time bins is chosen adaptively from E_t using the configured bounds $(B_{\min}, B_{\text{target}}, B_{\max})$:

$$\hat{B}_t = \begin{cases} \max(B_{\min},\ \max(1, E_t)), & E_t < B_{\text{target}}, \\ \min\Big(B_{\max},\ B_{\text{target}} + \Big\lfloor \dfrac{E_t - B_{\text{target}}}{3\,B_{\text{target}}} \Big\rfloor\Big), & E_t \geq B_{\text{target}}. \end{cases} \tag{19}$$

This rule limits the number of bins when few events are available and allows the discretization to grow gradually as the event count increases.

To construct the time discretization, we place $\hat{B}_t - 1$ equally-spaced quantile cut-points

of the observed event-time distribution in the training set, partitioning the time axis into $\hat{B}_t$ candidate intervals. When event times are tied, multiple cut-points may coincide and are merged, so the effective bin count $B_t \leq \hat{B}_t$ reflects the number of distinct intervals that remain.

For sample i, let $\tau_i^{(t)}$ denote the observed survival time and let $\delta_i^{(t)} \in \{0, 1\}$ indicate whether the event was observed ($\delta_i^{(t)} = 1$) or censored ($\delta_i^{(t)} = 0$). Let $j_i^{(t)} \in \{1, \ldots, B_t\}$ denote the index of the time bin containing $\tau_i^{(t)}$. The model outputs one logit $a_{i,k}^{(t)}$ per bin, which is converted to a discrete hazard:

$$h_{i,k}^{(t)} = \sigma(a_{i,k}^{(t)}), \qquad k = 1, \ldots, B_t. \tag{20}$$

Here, $h_{i,k}^{(t)}$ represents the conditional probability of experiencing the event in bin k, given survival through all preceding bins. The per-sample negative log-likelihood is:

$$\ell_i^{(t)} = \begin{cases} -\left(\displaystyle\sum_{k<j_i^{(t)}} \log(1 - h_{i,k}^{(t)}) + \log h_{i,j_i^{(t)}}^{(t)} \right), & \delta_i^{(t)} = 1, \\ -\displaystyle\sum_{k \leq j_i^{(t)}} \log(1 - h_{i,k}^{(t)}), & \delta_i^{(t)} = 0, \end{cases} \tag{21}$$

where the first case corresponds to an observed event in bin $j_i^{(t)}$, and the second corresponds to right censoring at bin $j_i^{(t)}$. For ranking-based metrics, hazards are converted to scalar risk:

$$r_i^{(t)} = -\sum_{k=1}^{B_t} \log(1 - h_{i,k}^{(t)}). \tag{22}$$

Table 1: Augmentation strategies used in each training stage.

Strategy	Stage 1 (SSL)	Stage 2 (Alignment)	Applied to
Spatial Augmentation (Flip, Rotation)	✓	✓	All crops / full grids
Block Masking	✓	✗	Global crops only
Token Dropout	✗	✓	Full slide grids

3.2.6 Multi-Task Loss Aggregation

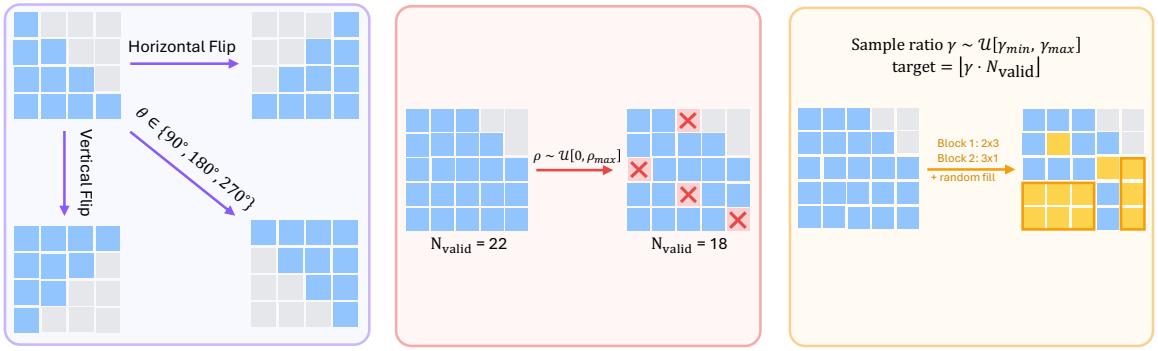
Let $\mathcal{T}_{\text{active}} \subseteq \mathcal{T}$ be tasks with usable supervision in the current batch. We average losses over active tasks:

$$\mathcal{L} = \frac{1}{|\mathcal{T}_{\text{active}}|} \sum_{t \in \mathcal{T}_{\text{active}}} \mathcal{L}^{(t)}, \tag{23}$$

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.

3.3 Augmentation Strategies

Table 1 summarizes the augmentation strategies used in each training stage. Spatial augmentation improves orientation robustness while preserving morphology, token dropout regularizes full-slide inputs during Stage 2, and block-based masking defines the masked prediction signal during Stage 1. Visual examples of these augmentation strategies are shown in Figure 5.



(a) Spatial Augmentation (b) Token Dropout (c) Block-based Masking

Figure 5: Visual examples of augmentation strategies across training stages.

Chapter 4

Experimental Setup

4.1 Dataset

We collect 56 open-sourced datasets: REG [56], TCGA (all 32 cohorts) [69], CPTAC (all 10 cohorts) [24], BC-Therapy [89], BRACS [9], CAMELYON17 [4], DHMC Kidney [121], DHMC LUAD [110], EBRAINS [86, 85], IMP Colorectum [77, 72, 71], IMP Cervix [76], MBC [7, 31], MUT-HET-RCC [82], NADT Prostate [111], NAT-BRCA [80], and PANDA [11]. All collected slides are processed using AtlasPatch [2], which performs tissue segmentation using SAM2 [51, 83] model fine-tuned on histopathology data. 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 approximately 1.6 billion extracted patches. We extract features for each patch using a pretrained lightweight patch encoder from [47], which has 21.67 million parameters and the architecture of ViT-S [23] trained using DINOv2 [78] on 40 million patches. The resulting per-patch feature vectors and their spatial coordinates are assembled into 2D feature grids, forming the shared input representation for both training stages.

Table 2 reports slide counts and extracted patch totals at $20\times$ and $40\times$ magnifications after segmenting tissues from the collected slides. Table 3 reports class-cardinality across

Table 2: Total number of patches extracted from the full dataset at each magnification level using non-overlapping 224×224 tiling.

Magnification Level	Number of Slides	Number of Patches
20X	53,286	449,943,195
40X	23,848	1,224,330,326
Total	77,134	1,674,273,521

Table 3: Distribution of the number of classes (i.e., labels) per classification task.

Number of Classes	2	3	4	5	6	7	8	9	10	12	30	46
Number of Tasks	148	36	5	5	1	2	1	2	2	1	1	1

classification tasks, where most tasks are low-cardinality and a smaller subset has higher-cardinality label spaces.

4.1.1 Stage 1 Data

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 approximately 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. Together, these slides span 23 distinct anatomical sites: adrenal gland, bladder, brain, breast, cervix, colon and rectum, esophagus, eye, head and neck, kidney, liver and bile ducts, lung, lymph node, ovary, pancreas, prostate, skin, soft tissue, stomach, testis, thymus, thyroid, and uterus.

4.1.2 Stage 2 Data

For Stage 2 supervised fine-tuning, we construct 333 tasks in total (205 classification and 128 survival) across all 56 datasets, averaging approximately 6 tasks per dataset. Survival

supervision includes overall survival (OS), disease-specific survival (DSS), disease-free interval (DFI), and progression-free interval (PFI), depending on cohort-level endpoint availability. Of these, 56 are slide-level and 277 are case-level (i.e., predictions aggregated over all slides of a patient). Not every slide processed in Stage 1 carries a label for at least one task. After merging all task-specific case lists and deduplicating, the labelled subset used for Stage 2 comprises $30{,}024$ unique patients and $45{,}179$ unique whole-slide images, a strict subset of the $53{,}286$ Stage 1 slides, as slides without any associated label are excluded from supervised training. A consolidated scale overview is shown in Figure 6.

4.1.3 Training Task Distribution by Anatomical Site

Table 4 reports the full distribution of training tasks across anatomical sites and associated task categories.

4.1.4 Sparse Supervision Structure

Stage 2 supervision is inherently sparse because each case is only labeled for the subset of tasks available in its source cohort. Let $M \in \{0,1\}^{N \times T}$ denote the case-task supervision matrix, where $M_{i,t} = 1$ if case c_i has a valid label for task $\mathcal{T}_t$ and $M_{i,t} = 0$ otherwise. For each task during training, loss is computed only on labeled cases, and per-batch optimization averages only over active tasks with usable labels. Figure 7 illustrates this sparse structure.

4.2 Task Preparation

4.2.1 TCGA Task Preparation

WSIs were collected from The Cancer Genome Atlas (TCGA) through the Genomic Data Commons (GDC) portal [69]. We first linked every TCGA slide to its case identifier, then harmonized case-level clinical and molecular attributes. We include 240 TCGA tasks (117 classification, 123 survival), corresponding to 72.1% of all Stage 2 tasks and 96.1% of all survival tasks. TCGA supervision spans 32 TCGA projects plus one pan-cancer pooled task, with 8 slide-level tasks and 232 case-level tasks. Across all TCGA tasks, the labeled union covers 9,732 unique cases and 11,857 unique slides. Survival endpoints include overall survival (OS), disease-specific survival (DSS), disease-free interval (DFI), and progression-free interval (PFI). Table 5 summarizes task families and cohort coverage.

Pan-cancer cancer-type classification (slide-level). We built a pooled TCGA benchmark with 11,185 slides and obtained 46 cancer subtypes labels from [22]. Labels are defined at slide level and retain cohort-specific subtype granularity.

Primary diagnosis classification (slide-level). We derived cohort-specific primary-diagnosis tasks from diagnosis records marked as primary disease, excluding missing and non-informative values. We required at least two classes with at least 25 cases per class. Seven cohorts satisfied these criteria and are detailed in Table 6.

Tumor grade classification (case-level). We extracted cohort-specific tumor-grade labels from TCGA diagnosis metadata after harmonizing grade strings into canonical G1 to G4 categories. A task was retained only when at least two classes remained after cleaning and each class had sufficient support (minimum 10 cases per class).

Survival prediction (case-level). We used four TCGA-CDR endpoints: overall survival (OS), disease-specific survival (DSS), disease-free interval (DFI), and progression-free interval (PFI). For each cohort-endpoint pair, event indicators and follow-up times

were filtered to valid entries, then modeled with a discrete-hazard objective using task-specific quantile time bins. The bin count was adaptive with target 8 bins and bounds of 2 to 16 bins. OS, DSS, and PFI were available in all 32 cohorts, while DFI was available in 27 cohorts (missing in TCGA-GBM, TCGA-MESO, TCGA-SKCM, TCGA-THYM, and TCGA-UVM).

Mutation status prediction (case-level). We constructed binary wildtype/mutant tasks for 25 driver genes (ALK, APC, ARID1A, ATRX, BAP1, BRAF, CDKN2A, CTNNB1, EGFR, ERBB2, FBXW7, IDH1, IDH2, KRAS, MET, NF1, PBRM1, PIK3CA, PTEN, RB1, SETD2, SMAD4, TERT, TP53, VHL). Cohort-gene tasks were retained only when both classes met minimum support (10 cases per class), yielding 99 mutation tasks across 20 cohorts.

For case-level tasks (tumor grade, survival, mutation), labels are defined at patient level and linked to all slides from that patient. Table 7 provides a detailed cohort-level coverage breakdown.

4.2.2 REG Task Preparation

REG dataset pairs whole-slide image identifiers with short, templated pathology report text. Each report follows a consistent schema of the form *"organ, procedure; histologic diagnosis[, grade]"*, which enables a deterministic decomposition into *organ*, *procedure*, and *diagnostic content*. The corpus contains 8,494 report–slide pairs spanning breast, prostate, stomach, lung, bladder, colorectal, and cervix specimens. After validation by checking if all of them follow the same structure, 8,493 records were retained; a single malformed entry lacking the required delimiter was excluded. This standardized structure permits rule-based label extraction, avoiding variability and potential bias introduced by LLM-based parsing, and yielding stable label definitions anchored to the original clinical phrasing.

Normalization and parsing. Before extracting labels, we unified organ names to a single vocabulary (e.g., *urinary bladder→bladder*, *uterine cervix→cervix*, *nipple→breast*). For tasks spanning multiple organs, colon and rectum were merged into a single *colorectal* category, as they share the same diagnostic criteria. Each report was then split into its three fields using the fixed delimiters, and all labels were derived solely from explicit wording in those fields, never inferred.

Labeling principle. Label assignment was deliberately conservative: a label was assigned only when the report contained clear, unambiguous textual evidence, and the sample was excluded from that task otherwise. For yes/no attributes, "Present" required an explicit positive statement, and "Absent" required either an explicit negation or a diagnosis that logically rules out the attribute. This conservative policy means each task has its own subset of samples. Slides whose reports lack sufficient evidence for a given task simply do not appear in that task's dataset. Figure 8 summarizes the resulting 36 tasks by class count, task type, and label structure. Below we describe the tasks constructed for each organ.

Breast (7 tasks). We extracted seven tasks from breast reports. *Histologic type* identifies the specific category of breast tissue abnormality (e.g., invasive carcinoma of no special type, invasive lobular carcinoma, ductal carcinoma in situ (DCIS), fibroadenoma, phyllodes tumor). When both an invasive component and an in-situ component are mentioned, the invasive component takes precedence. *DCIS presence* indicates whether a pre-invasive lesion confined to the milk ducts is mentioned. It is only marked *Absent* for cases where invasive carcinoma is confirmed with no DCIS mentioned. *Overall histologic grade*, *tubule formation*, *nuclear grade*, and *mitotic score* are the four components of the Nottingham grading system, a standardized scoring scheme that quantifies how abnormal the cancer cells look and how fast they are dividing. Each is labeled only when explicitly stated in the report. *Procedure type* records the biopsy method (core-needle, sono-guided core, mammotome, or biopsy NOS) from the report header.

Prostate (5 tasks). We extracted five tasks from prostate reports. *Gleason score* is the sum of the two most prevalent cancer growth patterns in the sample, each rated 1–5 by the pathologist, where higher scores indicate more aggressive cancer. *Primary* and *secondary Gleason pattern* are the individual pattern scores that make up the total. *ISUP grade group* is a simplified 1–5 scale derived from the Gleason score and used internationally to communicate prostate cancer severity. *Tumor presence* is marked *Present* when cancer terminology (carcinoma/adenocarcinoma) appears and *Absent* only when the report explicitly states no tumor was found. All four grading labels are assigned only when the report explicitly provides the values.

Stomach (5 tasks). We extracted five tasks from stomach reports. *Histologic type* classifies the tissue finding (e.g., adenocarcinoma, tubular adenoma, chronic gastritis, MALT lymphoma, gastrointestinal stromal tumor). *Differentiation grade* captures how closely cancer cells resemble normal stomach cells (well / moderately / poorly differentiated). It is assigned only for adenocarcinoma cases where the report explicitly states the grade. *Adenoma dysplasia grade* rates how abnormal the cells in a benign polyp look (low vs. high grade), assigned only for explicitly stated adenomas. *Intestinal metaplasia* flags a specific pre-cancerous change in the stomach lining, labeled *Present* only when explicitly mentioned and *Absent* only when gastritis is stated without it. *Malignancy status* uses a three-way scheme where *malignant* is assigned for carcinoma/lymphoma, *pre-malignant* for adenoma or high-grade dysplasia, and *benign* for gastritis or polyp, all based on explicit wording. Cases without clear cues are excluded.

Lung (3 tasks). We extracted three tasks from lung reports. *Histologic type* distinguishes the three most common lung cancer subtypes (adenocarcinoma, squamous cell carcinoma, small-cell carcinoma) when explicitly named. *Small-cell vs. non-small-cell* separates small-cell lung cancer, a fast-growing subtype requiring different treatment, from all other types, based on explicit terminology. *Malignancy status* is labeled only from explicit

malignant or benign cues.

Bladder (5 tasks). We extracted five tasks from bladder reports. *Tumor presence* is labeled *Present* when carcinoma terminology appears and *Absent* when the report explicitly states no tumor. *Invasiveness* distinguishes whether tumor cells are confined to the inner lining (non-invasive or in situ) versus having grown into deeper tissue layers (invasive). *Invasion depth* adds finer granularity: no invasion, invasion into the connective tissue just below the lining (subepithelial), or invasion into the muscle wall (muscularis propria), a critical distinction for treatment decisions. *Papillary grade* classifies non-invasive papillary tumors (finger-like growths from the bladder wall) as low or high grade based on explicit wording. *Consolidated tumor type* integrates all of the above into a single label, with an explicit "no tumor" statement taking precedence.

Colorectal (4 tasks). We extracted four tasks from colon and rectum reports. *Histologic type* identifies the tissue finding (e.g., adenocarcinoma, tubular adenoma, sessile serrated lesion, hyperplastic polyp, chronic colitis). *Differentiation grade* and *adenoma dysplasia grade* follow the same rules as stomach: assigned only for the relevant tissue type and only when explicitly stated. *Malignancy status* uses the same malignant/pre-malignant/benign scheme as stomach.

Cervix (4 tasks). We extracted four tasks from cervix reports. *CIN grade* (cervical intraepithelial neoplasia, stages 1–3) and *SIL grade* (squamous intraepithelial lesion, low or high) are two overlapping grading systems for pre-cancerous changes in the cervix. Both are assigned only in non-invasive contexts and are suppressed if invasive cancer is mentioned. *Invasive vs. pre-invasive* distinguishes cancer that has breached the cervical lining from changes still confined within it. *Procedure type* records the biopsy method (e.g., colposcopic or punch biopsy) from the report header.

Cross-organ tasks (3 tasks). Beyond the organ-specific tasks, we constructed three tasks that pool samples across all organs. *Organ classification* predicts which of the seven

tissue sites a slide comes from, using the normalized organ vocabulary (with colon and rectum merged into colorectal). *Procedure classification* predicts the biopsy method using a unified taxonomy spanning all organs (endoscopic, transurethral resection, colposcopic, colonoscopic, sono-guided, core, punch, mammotome, biopsy NOS). A label is assigned only when the procedure is explicitly named in the report header. *Global malignancy detection* applies the three-way malignant/pre-malignant/benign scheme across all organs. Figure 9 shows the class distributions for these three tasks.

4.3 Training Configuration

4.3.1 SSL Pretraining Configuration

We train the slide encoder using the Stage 1 self-supervised framework on all $77{,}134$ slide feature grids, treating $20\times$ and $40\times$ grids as independent samples. 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, approximately 436 GPU-hours).

The encoder is a 6-layer transformer (d=768, 12 heads, 4 register tokens [20]). 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 [62] with a cosine learning rate schedule and an EMA teacher whose momentum follows a cosine schedule from μ_0=0.996 to μ_T=1.0. Complete hyperparameters are provided in Tables 8–12.

4.3.2 Patient-Aware Semantic Alignment Configuration

We fine-tune the SSL-pretrained teacher encoder end-to-end alongside task-specific MLP heads, following the Stage 2 framework. Both $20\times$ and $40\times$ grids are used as independent inputs. 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, approximately 512 GPU-hours). Complete hyperparameters are provided in Tables 13–16.

For case-level aggregation, a case transformer (D_{case}=3 layers, 12 heads, learnable `[CASE]` token) pools all slides of a patient into a single embedding. We jointly train on 205 classification and 128 survival tasks (333 total) spanning all 56 datasets. Labels form a sparse matrix where each patient is labeled only for their active tasks, and loss is computed accordingly. 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 [62] 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.

4.4 Evaluation Protocol

We evaluate on eight held-out tasks that span diverse clinical settings: 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), and Treatment Response (MBC). All eight tasks are excluded from training. Seven are case-level and IDH Status is slide-level. All models are assessed under a unified frozen-feature MLP probe with five-fold evaluation and patient-level fold grouping. We report mean $\pm$ standard deviation for weighted F1, weighted ROC-AUC, and balanced accuracy.

4.4.1 MLP Probe Setup

Both the slide encoder and MIL comparisons share the same five-fold evaluation protocol. Folds are constructed with label stratification. Each fold applies an 80%/20% train-validation split. Model selection is based on best validation weighted F1 score. We report mean $\pm$ standard deviation across folds for weighted F1, weighted ROC-AUC, and balanced accuracy.

For the MLP probe setup in the slide encoder comparison, we use a three-layer MLP classifier on frozen slide representations:

$$\texttt{LayerNorm(D)} \rightarrow \texttt{Linear(D, h1)} \rightarrow \texttt{GELU} \rightarrow \texttt{Dropout} \rightarrow$$
$$\texttt{Linear(h1, h2)} \rightarrow \texttt{GELU} \rightarrow \texttt{Dropout} \rightarrow \texttt{Linear(h2, C)},$$

with adaptive hidden sizes $h_1 = \max(4, \mathrm{round}(0.66D))$ and $h_2 = \max(2, \mathrm{round}(0.5h_1))$. Optimization uses AdamW with learning rate 1×10^{-3} and weight decay 1×10^{-2}. Training runs for 200 epochs with batch size 64, cross-entropy loss, and dropout 0.25 in both hidden blocks. Class-balanced sampling is applied during training.

The MLP probe setup in the MIL comparison uses the same MLP head used in the slide encoder comparison, and follows the MIL-Lab implementation [91] for different MILs. Each patch encoder is paired with five MIL architectures (MeanMIL, ABMIL, CLAM, DSMIL, and TransMIL). Optimization uses AdamW with learning rate 1×10^{-3} and weight decay 1×10^{-2} for 100 epochs with one bag per iteration and class-balanced sampling. For multi-slide patients, patient-level predictions are obtained by averaging per-slide logits (late fusion).

4.4.2 Linear Probe Setup

For completeness, we additionally evaluate different slide encoders and MILs with a multinomial logistic-regression classifier. We use the same five-fold splits as the MLP comparison, with label stratification and case-level grouping and an 80% to 20% train-validation split in each fold. The classifier uses L2 regularization and selects regularization strength by minimizing validation loss over 45 logarithmically spaced values from 10^{-6} to 10^{5}. Optimization uses LBFGS with a maximum of 500 iterations and class-balanced weighting. Performance is reported as mean and standard deviation across folds for weighted F1, weighted ROC-AUC, and balanced accuracy.

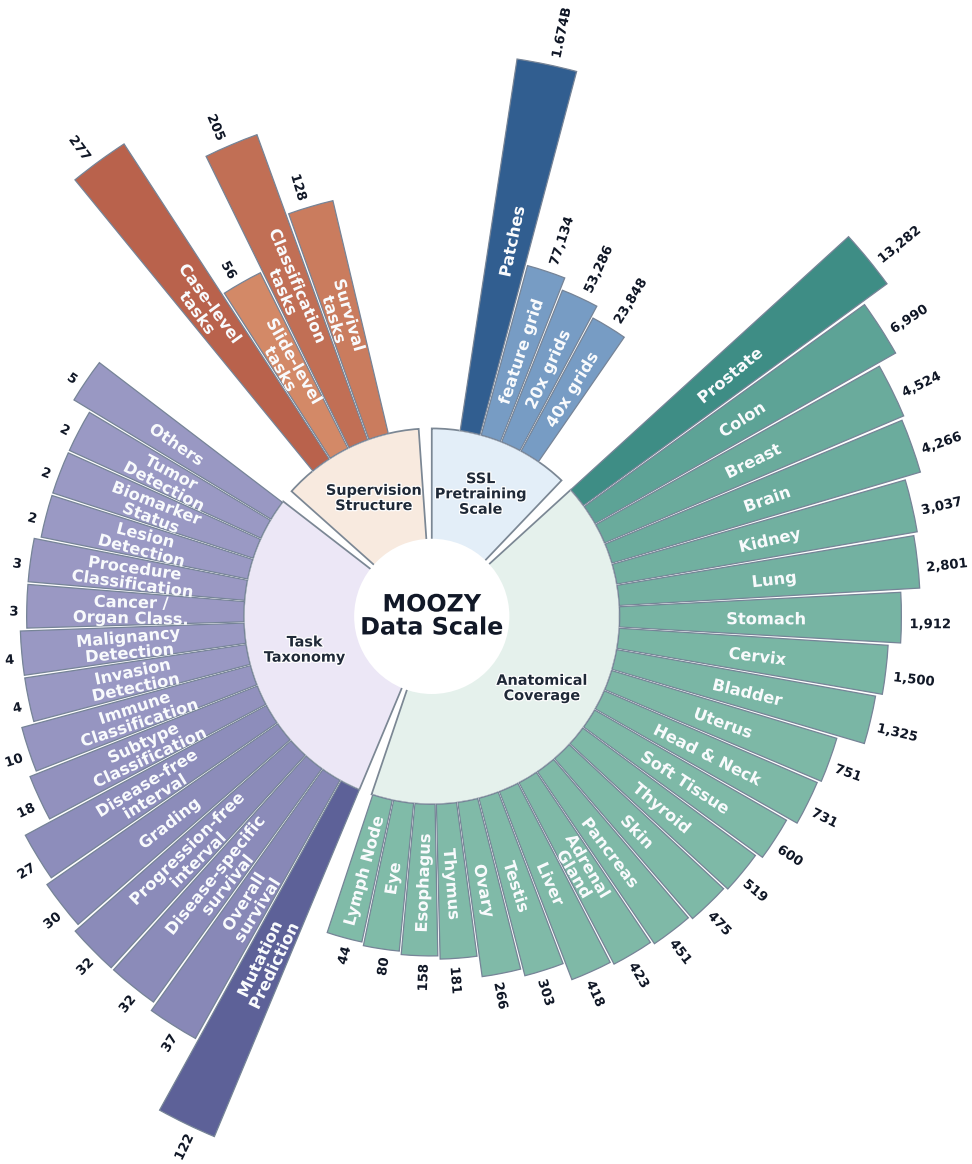


Figure 6: Radial hierarchy of MOOZY data scale across four dimensions: pretraining scale, anatomical coverage, task taxonomy, and supervision structure.

Table 4: Distribution of training tasks across anatomical sites and task categories.

Anatomical Site	Datasets	Tasks	Slides	Samples	Task Categories
Multi-organ	3	5	21,832	18,702	Cancer/Organ Classification, Malignancy Detection, Procedure Classification
Prostate	4	11	13,282	2,207	Grading, Survival Prediction, Treatment Response, Tumor Detection
Colon	5	35	6,990	6,977	Grading, Immune Classification, MSI Status, Malignancy Detection, Mutation Prediction, Subtype Classification, Survival Prediction
Breast	8	27	4,524	3,563	Biomarker Status, Grading, Immune Classification, Invasion Detection, Lesion Detection, Metastasis Detection, Mutation Prediction, Procedure Classification, Subtype Classification, Survival Prediction, Tumor Classification
Brain	4	19	4,266	3,126	Grading, Immune Classification, Mutation Prediction, Subtype Classification, Survival Prediction
Kidney	6	27	3,037	2,853	Grading, Immune Classification, Mutation Prediction, Subtype Classification, Survival Prediction
Lung	8	33	2,801	2,280	Histologic Pattern, Immune Classification, Malignancy Detection, Mutation Prediction, Subtype Classification, Survival Prediction
Stomach	2	20	1,912	1,886	Grading, Lesion Detection, Malignancy Detection, Mutation Prediction, Subtype Classification, Survival Prediction
Cervix	3	10	1,500	1,490	Grading, Invasion Detection, Procedure Classification, Survival Prediction
Bladder	2	17	1,325	1,254	Grading, Invasion Detection, Mutation Prediction, Subtype Classification, Survival Prediction, Tumor Detection
Uterus	3	35	751	656	Grading, Immune Classification, Mutation Prediction, Subtype Classification, Survival Prediction
Head & Neck	2	11	731	558	Grading, Immune Classification, Mutation Prediction, Survival Prediction
Soft Tissue	1	6	600	254	Mutation Prediction, Subtype Classification, Survival Prediction
Thyroid	1	5	519	506	Mutation Prediction, Survival Prediction
Skin	1	14	475	433	Mutation Prediction, Survival Prediction
Pancreas	2	10	451	288	Grading, Immune Classification, Mutation Prediction, Survival Prediction
Adrenal Gland	2	8	423	232	Survival Prediction
Liver	2	12	418	404	Grading, Mutation Prediction, Survival Prediction
Testis	1	5	303	225	Subtype Classification, Survival Prediction
Ovary	2	5	266	156	Immune Classification, Survival Prediction
Thymus	1	4	181	121	Subtype Classification, Survival Prediction
Esophagus	1	6	158	156	Grading, Subtype Classification, Survival Prediction
Eye	1	4	80	80	Mutation Prediction, Survival Prediction
Lymph Node	1	4	44	44	Survival Prediction

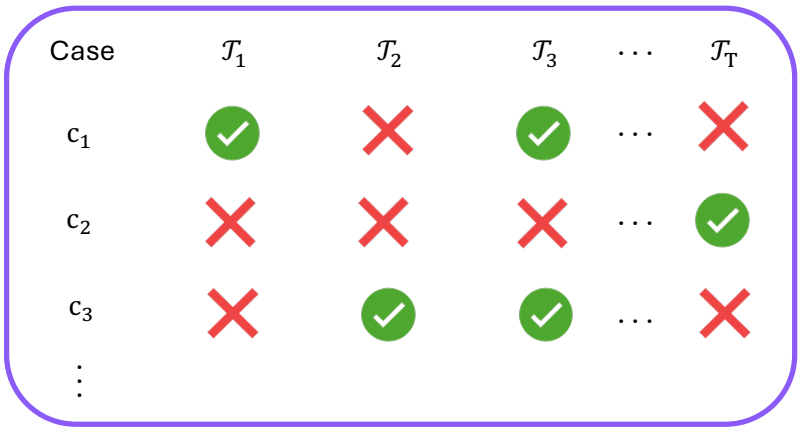


Figure 7: Schematic of sparse case-task supervision in Stage 2. Rows denote cases and columns denote tasks. A check mark indicates an available supervision target for that case-task pair; a cross indicates missing supervision.

Table 5: TCGA task families used in the supervised training run. Unique case/slide counts are union counts within each family and are not additive across rows.

Task Family	Tasks	Level	Cohorts	Label Space	Unique Cases	Unique Slides
Pan-cancer cancer type	1	Slide	Pooled TCGA	46 classes	9,148	11,185
Primary diagnosis	7	Slide	7 cohorts	2 to 3 classes	2,284	2,284
Tumor grade	10	Case	10 cohorts	G1 to G4 (2 to 4 observed classes)	3,274	3,790
Mutation status	99	Case	20 cohorts	Binary (wildtype vs mutant)	7,955	9,619
Survival endpoints	123	Case	32 cohorts (DFI in 27)	Adaptive discrete-hazard bins (2 to 16 bins)	9,578	11,675
TCGA union (all families)	240	Mixed	32 + pooled	Mixed	9,732	11,857

Table 6: Included TCGA cohorts and retained primary-diagnosis classes.

Cohort	Organ	Primary diagnosis classes (count)	Samples
TCGA-BRCA	Breast	Infiltrating duct carcinoma (710), Lobular carcinoma (182)	892
TCGA-COAD	Colorectal	Adenocarcinoma (378), Mucinous adenocarcinoma (59)	437
TCGA-ESCA	Esophagus	Squamous cell carcinoma (82), Adenocarcinoma (65)	147
TCGA-SARC	Soft tissue	Leiomyosarcoma (76), Dedifferentiated liposarcoma (41), Undifferentiated sarcoma (15)	132
TCGA-TGCT	Testis	Seminoma (130), Mixed germ cell tumor (25), Embryonal carcinoma (15)	170
TCGA-THYM	Thymus	Thymoma type AB (29), Thymoma type B2 (23)	52
TCGA-UCEC	Uterus	Endometrioid adenocarcinoma (350), Serous cystadenocarcinoma (104)	454

Table 7: Cohort-level coverage of TCGA tasks used in training. Columns report the number of tasks per family and the union of labeled cases/slides within each cohort across all included TCGA tasks.

Cohort	Surv.	Mut.	Grade	Prim. Dx	Total	Cases	Slides
TCGA-ACC	4	0	0	0	4	56	227
TCGA-BLCA	4	8	0	0	12	386	457
TCGA-BRCA	4	5	0	1	10	1,062	1,133
TCGA-CESC	4	0	1	0	5	269	279
TCGA-CHOL	4	0	1	0	5	39	39
TCGA-COAD	4	11	0	1	16	451	459
TCGA-DLBC	4	0	0	0	4	44	44
TCGA-ESCA	4	0	1	1	6	156	158
TCGA-GBM	3	1	0	0	4	389	860
TCGA-HNSC	4	2	1	0	7	450	472
TCGA-KICH	4	0	0	0	4	108	120
TCGA-KIRC	4	4	1	0	9	513	519
TCGA-KIRP	4	2	0	0	6	275	299
TCGA-LGG	4	5	1	0	10	491	844
TCGA-LIHC	4	2	1	0	7	365	379
TCGA-LUAD	4	6	0	0	10	478	541
TCGA-LUSC	4	2	0	0	6	478	512
TCGA-MESO	3	1	0	0	4	75	87
TCGA-OV	4	0	0	0	4	105	106
TCGA-PAAD	4	2	1	0	7	183	209
TCGA-PCPG	4	0	0	0	4	176	196
TCGA-PRAD	4	0	0	0	4	403	449
TCGA-READ	4	2	0	0	6	165	166
TCGA-SARC	4	1	0	1	6	254	600
TCGA-SKCM	3	11	0	0	14	433	475
TCGA-STAD	4	10	1	0	15	416	442
TCGA-TGCT	4	0	0	1	5	225	303
TCGA-THCA	4	1	0	0	5	506	519
TCGA-THYM	3	0	0	1	4	121	181
TCGA-UCEC	4	22	1	1	28	505	566
TCGA-UCS	4	0	0	0	4	57	91
TCGA-UVM	3	1	0	0	4	80	80

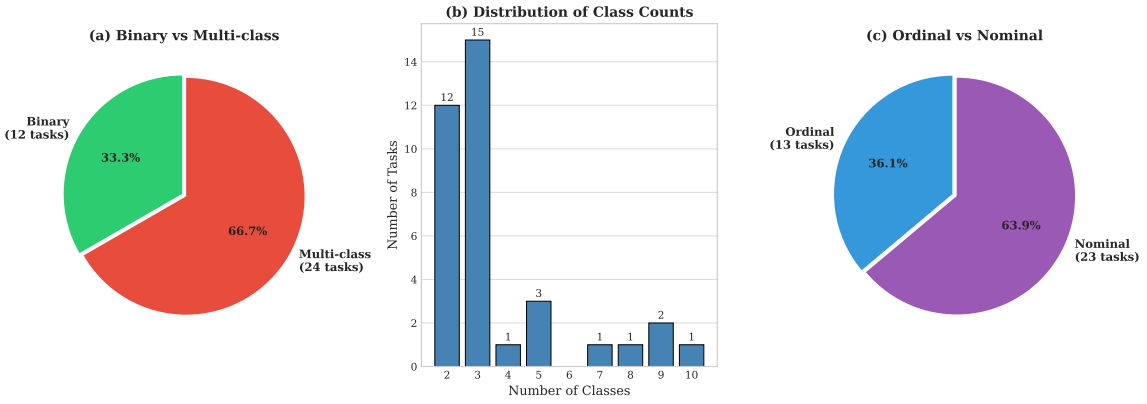


Figure 8: Characterization of REG classification tasks. **(a)** Proportion of binary (33.3%, 12 tasks) versus multi-class (66.7%, 24 tasks) tasks. **(b)** Distribution of class counts per task, with the majority having 2 or 3 classes. **(c)** Proportion of ordinal (36.1%, 13 tasks, e.g., grading) versus nominal tasks (63.9%, 23 tasks, e.g., subtype classification).

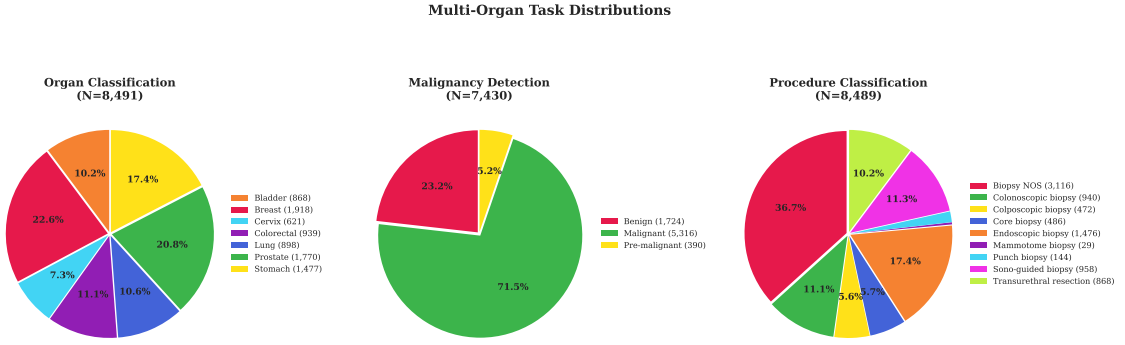


Figure 9: Class distributions for the three cross-organ REG tasks. *Organ Classification* (N=8,491) spans seven organs, with breast (22.6%) and prostate (20.8%) as the largest groups. *Malignancy Detection* (N=7,430) exhibits a strong class imbalance, with 71.5% malignant cases. *Procedure Classification* (N=8,489) covers nine biopsy types, with biopsy NOS (36.7%) being the most common.

Table 8: Encoder architecture hyperparameters.

Hyperparameter	**Value**
Input feature dimension (d_{patch})	384
Model dimension (d)	768
Number of attention heads (H)	12
Number of transformer layers (D)	6
Feed-forward dimension	3072
Number of register tokens (R)	4
MLP dropout rate	0.1
Attention dropout rate	0.0
Stochastic depth max rate	0.1
LayerScale initialization	Disabled
QK normalization	Disabled
Learnable ALiBi slopes	No (fixed)
Projection Head	
Hidden dimension	2048
Bottleneck dimension	256
Output dimension (K)	8192
Weight normalization	Enabled (frozen gain)

Table 9: Multi-crop sampling hyperparameters.

Hyperparameter	**Value**
Number of global crops (G)	2
Global crop size ($S_g \times S_g$)	20×20 tokens
Number of local crops (L)	4
Local crop size ($S_l \times S_l$)	12×12 tokens
Minimum valid token ratio ($\rho_{\min}$)	0.25
Maximum resampling attempts	3
Spatial Augmentations	
Horizontal flip probability (p_h)	0.5
Vertical flip probability (p_v)	0.5
Rotation probability (p_r)	0.5
Rotation angles	$\{90°, 180°, 270°\}$

Table 10: Block masking hyperparameters.

Hyperparameter	**Value**
Masking strategy	Block
Mask ratio minimum ($\gamma_{\min}$)	0.1
Mask ratio maximum ($\gamma_{\max}$)	0.5
Minimum patches per block	4
Maximum patches per block	Unlimited
Minimum aspect ratio ($\alpha_{\min}$)	0.3
Maximum aspect ratio ($\alpha_{\max}$)	$1/0.3 \approx 3.33$
Per-crop masking probability	0.5
Masking applied to	Global crops only

Table 11: Optimization hyperparameters for SSL pretraining.

Hyperparameter	**Value**
Optimizer	AdamW
Base learning rate	5×10^{-4}
Reference batch size for LR scaling	256
Minimum learning rate	2×10^{-6}
Learning rate schedule	Cosine decay
Learning rate warmup epochs	5
Weight decay (start)	0.04
Weight decay (end)	0.4
Weight decay schedule	Cosine
Gradient clipping (max norm)	0.3
Mixed precision	BFloat16
Training Scale	
Micro batch size per GPU	64
Number of GPUs	8
Gradient accumulation steps	2
Effective batch size	1024
Number of epochs	200
Total optimizer steps	14,400
Training time (GPU-hours)	$\approx$436

Table 12: Self-distillation hyperparameters.

Hyperparameter	**Value**
EMA Teacher	
Initial momentum (μ_0)	0.996
Final momentum (μ_T)	1.0
Momentum schedule	Cosine
Temperature	
Student temperature (τ_s)	0.1
Teacher CLS temperature (start)	0.04
Teacher CLS temperature (end, τ_t)	0.07
Teacher patch temperature (start)	0.04
Teacher patch temperature (end, τ_t^{patch})	0.07
Temperature warmup epochs	30
Centering	
Center momentum (λ)	0.9
Stabilization	
Freeze final projection layer (epochs)	3

Table 13: Architecture hyperparameters for semantic alignment. The slide encoder uses the same architecture as SSL pretraining (Table 8), initialized from the pretrained teacher weights.

Hyperparameter	**Value**
Case Transformer	
Number of layers (D_{case})	3
Number of attention heads	12
Feed-forward dimension	3072
Dropout rate	0.1
LayerScale initialization	10^{-5}
`[CASE]` token initialization std	0.02
Task Heads	
Head type	MLP
Head dropout rate	0.1

Table 14: Optimization hyperparameters for semantic alignment.

Hyperparameter	**Value**
Optimizer	AdamW
Base learning rate	5×10^{-5}
Minimum learning rate	2×10^{-7}
Learning rate schedule	Cosine annealing
Warmup steps	0
Weight decay	0.4
Gradient clipping (max norm)	0.3
Mixed precision	BFloat16
Training Scale	
Micro batch size per GPU	1 case
Number of GPUs	8
Gradient accumulation steps	128
Effective batch size	1024 cases
Number of epochs	20
Total optimizer steps	1,000
Training time (GPU-hours)	$\approx$512

Table 15: Data augmentation hyperparameters for semantic alignment.

Hyperparameter	**Value**
Spatial Augmentations	
Horizontal flip probability (p_h)	0.5
Vertical flip probability (p_v)	0.5
Rotation probability (p_r)	0.5
Rotation angles	$\{90°, 180°, 270°\}$
Token Dropout	
Maximum dropout ratio (ρ_{max})	0.1

Table 16: Loss function hyperparameters for semantic alignment.

Hyperparameter	**Value**
Classification Tasks	
Label smoothing (ϵ)	0.03
Class weighting	Inverse frequency
Survival Tasks	
Loss function	Discrete-time NLL (hazard)
Time bins (min / target / max)	2 / 8 / 16
Class weighting	None
Multi-Task Aggregation	
Task loss weighting	Equal (average)
Validation	
Validation split ratio	0.05 (5%)
Stratification	By task (class/event)

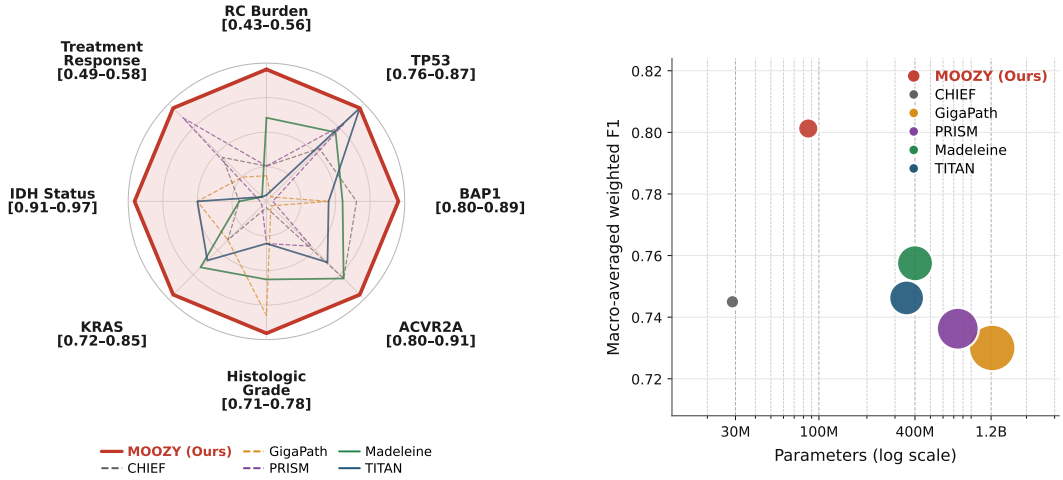
Chapter 5

Results and Discussion

5.1 Comparison with Slide Encoders

We compare MOOZY against five slide encoders: CHIEF, GigaPath, PRISM, Madeleine, and TITAN. For case-level tasks, baseline encoders produce patient representations by averaging per-slide embeddings before probing, while MOOZY uses its native case-level embedding from the case transformer, requiring no post-hoc fusion. For the slide-level IDH Status task, all methods are evaluated per slide. Results are presented in Table 17.

MOOZY achieves best or tied-best weighted F1 on all eight tasks and best or second-best AUC and balanced accuracy on seven of eight tasks. The largest margins are on Residual Cancer Burden (+0.05 F1, +0.11 AUC over Madeleine). On mutation tasks, MOOZY is strongest on ACVR2A across all three metrics and improves F1 on BAP1 and KRAS, while TITAN remains stronger on BAP1 AUC and KRAS balanced accuracy, suggesting complementary strengths between cross-slide aggregation and multimodal pretraining. TITAN is the closest overall competitor, also leading on TP53 AUC. Treatment Response is the main exception, where high variance (± 0.17) limits firm conclusions.



(a) **(b)**

Figure 10: **(a)** Weighted F1 across eight held-out tasks. Brackets report [min – max] weighted F1 for each task, with the center corresponding to the minimum and the outer ring to the maximum observed value. **(b)** Macro-averaged weighted F1 versus total parameter count (log scale), showing that MOOZY remains highly accurate while being highly parameter-efficient.

5.2 Parameter Efficiency

MOOZY totals 85.77M parameters (64.10M slide and patient encoder + 21.67M patch encoder), making it 4–14× smaller than competing encoders (GigaPath 1.22B, PRISM 742M, Madeleine 400M, TITAN 355M). CHIEF is smaller in absolute size but delivers substantially weaker performance. This efficiency stems from concentrating capacity at the slide level and reusing a lightweight public patch encoder. A detailed breakdown is provided in Table 18.

5.3 Comparison with MIL Baselines

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 [91], and each entry is the arithmetic mean over the five architectures. This

Table 17: Frozen-feature MLP probe comparison against slide encoder baselines on eight held-out tasks. **Bold**: best; underline: second best.

Task	Metric	CHIEF	Giga-Path	PRISM	Made-leine	TITAN	MOOZY (Ours)
Residual Cancer Burden	F1	$0.46_{\pm 0.03}$	$0.45_{\pm 0.05}$	$0.46_{\pm 0.07}$	$\underline{0.51}_{\pm 0.03}$	$0.43_{\pm 0.07}$	$\mathbf{0.56}_{\pm 0.05}$
	AUC	$0.60_{\pm 0.05}$	$0.55_{\pm 0.08}$	$0.58_{\pm 0.06}$	$\underline{0.63}_{\pm 0.03}$	$0.58_{\pm 0.07}$	$\mathbf{0.74}_{\pm 0.04}$
	Bal. Acc	$0.44_{\pm 0.05}$	$0.40_{\pm 0.05}$	$0.43_{\pm 0.04}$	$\underline{0.48}_{\pm 0.03}$	$0.38_{\pm 0.11}$	$\mathbf{0.51}_{\pm 0.06}$
TP53 Mut.	F1	$0.82_{\pm 0.05}$	$0.76_{\pm 0.03}$	$\underline{0.85}_{\pm 0.03}$	$0.84_{\pm 0.06}$	$\mathbf{0.87}_{\pm 0.04}$	$\underline{0.87}_{\pm 0.04}$
	AUC	$0.81_{\pm 0.09}$	$0.76_{\pm 0.06}$	$0.85_{\pm 0.05}$	$0.85_{\pm 0.08}$	$\mathbf{0.91}_{\pm 0.04}$	$\underline{0.86}_{\pm 0.06}$
	Bal. Acc	$0.83_{\pm 0.04}$	$0.76_{\pm 0.04}$	$0.84_{\pm 0.04}$	$0.84_{\pm 0.05}$	$\mathbf{0.88}_{\pm 0.04}$	$\underline{0.86}_{\pm 0.05}$
BAP1 Mut.	F1	$\underline{0.86}_{\pm 0.04}$	$0.84_{\pm 0.06}$	$0.80_{\pm 0.07}$	$0.85_{\pm 0.08}$	$0.84_{\pm 0.06}$	$\mathbf{0.89}_{\pm 0.06}$
	AUC	$0.75_{\pm 0.09}$	$0.63_{\pm 0.17}$	$0.71_{\pm 0.09}$	$0.78_{\pm 0.11}$	$\mathbf{0.82}_{\pm 0.06}$	$\underline{0.79}_{\pm 0.12}$
	Bal. Acc	$0.75_{\pm 0.12}$	$0.66_{\pm 0.14}$	$0.66_{\pm 0.10}$	$0.75_{\pm 0.12}$	$\underline{0.75}_{\pm 0.11}$	$\mathbf{0.78}_{\pm 0.11}$
ACVR2A Mut.	F1	$\underline{0.89}_{\pm 0.07}$	$0.80_{\pm 0.10}$	$0.85_{\pm 0.03}$	$0.89_{\pm 0.09}$	$0.87_{\pm 0.05}$	$\mathbf{0.91}_{\pm 0.05}$
	AUC	$0.80_{\pm 0.13}$	$0.74_{\pm 0.11}$	$\underline{0.83}_{\pm 0.08}$	$0.76_{\pm 0.19}$	$0.79_{\pm 0.10}$	$\mathbf{0.91}_{\pm 0.09}$
	Bal. Acc	$0.80_{\pm 0.12}$	$0.65_{\pm 0.10}$	$\underline{0.81}_{\pm 0.08}$	$0.81_{\pm 0.16}$	$0.76_{\pm 0.15}$	$\mathbf{0.90}_{\pm 0.10}$
Histologic Grade	F1	$0.71_{\pm 0.07}$	$\underline{0.77}_{\pm 0.03}$	$0.73_{\pm 0.11}$	$0.75_{\pm 0.06}$	$0.73_{\pm 0.06}$	$\mathbf{0.78}_{\pm 0.08}$
	AUC	$0.71_{\pm 0.09}$	$\mathbf{0.77}_{\pm 0.04}$	$0.67_{\pm 0.11}$	$0.74_{\pm 0.08}$	$0.71_{\pm 0.04}$	$\underline{0.75}_{\pm 0.15}$
	Bal. Acc	$0.73_{\pm 0.06}$	$\underline{0.77}_{\pm 0.03}$	$0.73_{\pm 0.12}$	$0.74_{\pm 0.06}$	$0.73_{\pm 0.06}$	$\mathbf{0.77}_{\pm 0.08}$
KRAS Mut.	F1	$0.77_{\pm 0.08}$	$0.77_{\pm 0.08}$	$0.72_{\pm 0.07}$	$\underline{0.81}_{\pm 0.06}$	$0.80_{\pm 0.05}$	$\mathbf{0.85}_{\pm 0.04}$
	AUC	$0.76_{\pm 0.14}$	$0.72_{\pm 0.09}$	$0.61_{\pm 0.12}$	$0.70_{\pm 0.05}$	$\underline{0.80}_{\pm 0.05}$	$\mathbf{0.80}_{\pm 0.06}$
	Bal. Acc	$0.74_{\pm 0.13}$	$0.76_{\pm 0.08}$	$0.63_{\pm 0.13}$	$0.77_{\pm 0.07}$	$\mathbf{0.81}_{\pm 0.05}$	$\underline{0.79}_{\pm 0.10}$
IDH Status	F1	$0.92_{\pm 0.01}$	$\underline{0.94}_{\pm 0.02}$	$0.91_{\pm 0.02}$	$0.92_{\pm 0.02}$	$0.94_{\pm 0.02}$	$\mathbf{0.97}_{\pm 0.02}$
	AUC	$0.96_{\pm 0.01}$	$0.97_{\pm 0.02}$	$0.95_{\pm 0.01}$	$0.96_{\pm 0.01}$	$\underline{0.97}_{\pm 0.01}$	$\mathbf{0.99}_{\pm 0.01}$
	Bal. Acc	$0.92_{\pm 0.01}$	$0.94_{\pm 0.02}$	$0.91_{\pm 0.02}$	$0.91_{\pm 0.02}$	$\underline{0.94}_{\pm 0.02}$	$\mathbf{0.97}_{\pm 0.02}$
Treatment Response	F1	$0.53_{\pm 0.07}$	$0.51_{\pm 0.07}$	$\underline{0.57}_{\pm 0.08}$	$0.49_{\pm 0.05}$	$0.49_{\pm 0.02}$	$\mathbf{0.58}_{\pm 0.14}$
	AUC	$\mathbf{0.70}_{\pm 0.06}$	$0.68_{\pm 0.10}$	$\underline{0.69}_{\pm 0.07}$	$0.59_{\pm 0.05}$	$0.60_{\pm 0.06}$	$0.68_{\pm 0.07}$
	Bal. Acc	$0.48_{\pm 0.10}$	$0.40_{\pm 0.08}$	$\mathbf{0.51}_{\pm 0.11}$	$0.35_{\pm 0.09}$	$0.37_{\pm 0.06}$	$\underline{0.48}_{\pm 0.17}$

comparison is deliberately asymmetric: MIL baselines train a task-specific aggregator from scratch on labeled data, whereas MOOZY is entirely frozen and evaluated with an MLP probe on its pretrained case embedding. Macro-average results are shown in Table 19 and the full per-task breakdown in Table 20.

At the macro level (Table 19), MOOZY leads all three metrics, improving over the strongest MIL baseline (CONCH v1.5) by +0.055 weighted F1, +0.064 weighted ROC-AUC, and +0.062 balanced accuracy. The remaining MIL baselines (Backbone, UNI v2, Phikon v2, MUSK) cluster within a narrow band, with no single encoder consistently second across all metrics.

Table 18: Parameter count comparison across slide encoders. MOOZY (85.77M total) is the most parameter-efficient slide-level encoder while achieving best or tied-best performance on most held-out tasks (Table 17).

Encoder	Slide Encoder Parameters	Patch Encoder Parameters	Total Parameters
CHIEF	1.19M	27.52M	28.71M
GigaPath	85.15M	1.13B	1.22B
PRISM	110.83M	631.23M	742.06M
Madeleine	5.00M	395.23M	400.23M
TITAN	48.54M	306.11M	354.65M
MOOZY (Ours)	42.8M (slide) + 21.3M (case)	21.67M	85.77M

The per-task results (Table 20) reveal that MOOZY achieves best weighted F1 and best balanced accuracy on all eight tasks, and best ROC-AUC on seven of eight. The largest margins appear on ACVR2A mutation, where MOOZY exceeds the next-best baseline by +0.07 F1, +0.17 AUC, and +0.17 balanced accuracy, and on Residual Cancer Burden (+0.09 F1, +0.13 AUC, +0.07 balanced accuracy over CONCH v1.5). On mutation tasks, MOOZY improves over the strongest MIL baseline by +0.07 F1 on TP53, +0.03 F1 on BAP1, and +0.05 F1 on KRAS. IDH Status shows the clearest separation, with MOOZY reaching 0.97 F1 and 0.99 AUC compared to 0.94 and 0.97 for CONCH v1.5. The sole metric where MOOZY does not lead is Histologic Grade AUC, where CONCH v1.5 edges ahead by 0.01 (0.76 vs. 0.75), though MOOZY retains the lead on F1 and balanced accuracy for this task. Treatment Response again exhibits high variance (± 0.14 F1), but MOOZY still leads on F1 and balanced accuracy and ties MUSK on AUC.

CONCH v1.5 is the strongest MIL baseline overall, ranking first or second among MIL methods on the majority of task-metric pairs. This is consistent with its vision-language pretraining, which produces richer patch features than vision-only encoders. MUSK follows as second-strongest on BAP1 and Treatment Response. Despite the advantage that task-specific aggregation gives MIL baselines, particularly the ability to attend to a few decisive patch regions, MOOZY's frozen patient-level representations consistently outperform, indicating that the structure captured during patient-aware pretraining subsumes

Table 19: Macro-average MIL comparison across eight held-out tasks. Each 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)	0.733	0.716	0.715	<u>0.746</u>	0.729	**0.801**
ROC-AUC (weighted)	0.735	0.719	0.724	<u>0.751</u>	0.725	**0.815**
Balanced Acc	0.686	0.660	0.654	<u>0.696</u>	0.679	**0.758**

much of what task-specific MIL training learns.

5.4 Multi-Stage Ablation

Table 21 reports macro-average results for four configurations. Stage 2 only (task supervision without SSL) slightly underperforms Stage 1 alone on weighted F1:

$$\text{F1}_{\text{Stage 2}} = 0.748 \quad \text{vs.} \quad \text{F1}_{\text{Stage 1}} = 0.760, \tag{24}$$

and AUC:

$$\text{AUC}_{\text{Stage 2}} = 0.725 \quad \text{vs.} \quad \text{AUC}_{\text{Stage 1}} = 0.753, \tag{25}$$

confirming that SSL provides a foundation that task supervision alone cannot recover. Combining both stages with the case aggregator yields the strongest results, validating the two-stage design: SSL provides a generalizable spatial prior, and patient-aware alignment provides clinical grounding that SSL alone cannot achieve.

5.4.1 Stage 1 Only vs. MOOZY

Table 22 provides the complete task-level breakdown. MOOZY improves most tasks, with the largest gains on ACVR2A mutation and BAP1 mutation. KRAS mutation remains close between stages, with near-parity in weighted F1 and small decreases in weighted ROC-AUC and balanced accuracy. Despite this task-specific variation, macro averages

Table 20: Comparison of MOOZY against patch encoder baselines with trained MIL aggregators on eight held-out tasks. MIL baselines train a task-specific aggregator from scratch on frozen patch features; MOOZY is entirely frozen and evaluated with an MLP probe. Each patch encoder entry is the arithmetic mean over five MIL architectures (MeanMIL, ABMIL, CLAM, DSMIL, and TransMIL). **Bold**: best; underline: second best.

Task	Metric	Backbone	UNI v2	Phikon v2	CONCH v1.5	MUSK	MOOZY (Ours)
Residual Cancer Burden	F1 (weighted)	$0.46_{\pm 0.05}$	$0.44_{\pm 0.05}$	$0.42_{\pm 0.06}$	$\underline{0.47}_{\pm 0.05}$	$0.44_{\pm 0.05}$	$\mathbf{0.56}_{\pm 0.05}$
	ROC-AUC (weighted)	$0.60_{\pm 0.06}$	$0.60_{\pm 0.06}$	$0.59_{\pm 0.06}$	$\underline{0.61}_{\pm 0.06}$	$0.59_{\pm 0.07}$	$\mathbf{0.74}_{\pm 0.04}$
	Balanced Acc	$\underline{0.44}_{\pm 0.06}$	$0.40_{\pm 0.06}$	$0.39_{\pm 0.06}$	$0.42_{\pm 0.06}$	$0.40_{\pm 0.06}$	$\mathbf{0.51}_{\pm 0.06}$
TP53 mutation	F1 (weighted)	$0.77_{\pm 0.06}$	$0.77_{\pm 0.07}$	$0.78_{\pm 0.07}$	$\underline{0.80}_{\pm 0.06}$	$0.79_{\pm 0.06}$	$\mathbf{0.87}_{\pm 0.04}$
	ROC-AUC (weighted)	$0.79_{\pm 0.07}$	$0.75_{\pm 0.09}$	$0.78_{\pm 0.08}$	$\underline{0.81}_{\pm 0.08}$	$0.80_{\pm 0.06}$	$\mathbf{0.86}_{\pm 0.06}$
	Balanced Acc	$0.76_{\pm 0.05}$	$0.77_{\pm 0.07}$	$0.77_{\pm 0.07}$	$\underline{0.79}_{\pm 0.07}$	$0.79_{\pm 0.04}$	$\mathbf{0.86}_{\pm 0.05}$
BAP1 mutation	F1 (weighted)	$0.84_{\pm 0.04}$	$0.82_{\pm 0.05}$	$0.83_{\pm 0.05}$	$0.85_{\pm 0.05}$	$\underline{0.86}_{\pm 0.06}$	$\mathbf{0.89}_{\pm 0.06}$
	ROC-AUC (weighted)	$0.66_{\pm 0.19}$	$0.65_{\pm 0.16}$	$0.67_{\pm 0.12}$	$\underline{0.75}_{\pm 0.13}$	$0.73_{\pm 0.13}$	$\mathbf{0.79}_{\pm 0.12}$
	Balanced Acc	$0.67_{\pm 0.10}$	$0.64_{\pm 0.10}$	$0.68_{\pm 0.10}$	$\underline{0.74}_{\pm 0.09}$	$0.73_{\pm 0.10}$	$\mathbf{0.78}_{\pm 0.11}$
ACVR2A mutation	F1 (weighted)	$0.83_{\pm 0.09}$	$\underline{0.84}_{\pm 0.07}$	$0.81_{\pm 0.11}$	$0.82_{\pm 0.07}$	$0.79_{\pm 0.09}$	$\mathbf{0.91}_{\pm 0.05}$
	ROC-AUC (weighted)	$\underline{0.74}_{\pm 0.10}$	$0.72_{\pm 0.16}$	$0.73_{\pm 0.13}$	$0.70_{\pm 0.14}$	$0.60_{\pm 0.19}$	$\mathbf{0.91}_{\pm 0.09}$
	Balanced Acc	$0.73_{\pm 0.09}$	$\underline{0.73}_{\pm 0.11}$	$0.68_{\pm 0.13}$	$0.70_{\pm 0.10}$	$0.65_{\pm 0.12}$	$\mathbf{0.90}_{\pm 0.10}$
Histologic Grade	F1 (weighted)	$0.74_{\pm 0.05}$	$0.74_{\pm 0.05}$	$0.72_{\pm 0.05}$	$\underline{0.76}_{\pm 0.05}$	$0.75_{\pm 0.08}$	$\mathbf{0.78}_{\pm 0.08}$
	ROC-AUC (weighted)	$0.73_{\pm 0.09}$	$0.73_{\pm 0.05}$	$0.72_{\pm 0.06}$	$\mathbf{0.76}_{\pm 0.08}$	$0.74_{\pm 0.08}$	$\underline{0.75}_{\pm 0.15}$
	Balanced Acc	$0.74_{\pm 0.05}$	$0.75_{\pm 0.06}$	$0.72_{\pm 0.05}$	$\underline{0.75}_{\pm 0.05}$	$0.75_{\pm 0.07}$	$\mathbf{0.77}_{\pm 0.08}$
KRAS mutation	F1 (weighted)	$0.78_{\pm 0.07}$	$0.74_{\pm 0.05}$	$0.74_{\pm 0.05}$	$\underline{0.80}_{\pm 0.08}$	$0.76_{\pm 0.07}$	$\mathbf{0.85}_{\pm 0.04}$
	ROC-AUC (weighted)	$\underline{0.74}_{\pm 0.10}$	$0.71_{\pm 0.09}$	$0.68_{\pm 0.07}$	$0.74_{\pm 0.12}$	$0.70_{\pm 0.10}$	$\mathbf{0.80}_{\pm 0.06}$
	Balanced Acc	$0.76_{\pm 0.09}$	$0.69_{\pm 0.08}$	$0.68_{\pm 0.06}$	$\underline{0.77}_{\pm 0.09}$	$0.72_{\pm 0.08}$	$\mathbf{0.79}_{\pm 0.10}$
IDH Status	F1 (weighted)	$0.93_{\pm 0.02}$	$0.93_{\pm 0.02}$	$0.93_{\pm 0.02}$	$\underline{0.94}_{\pm 0.02}$	$0.92_{\pm 0.02}$	$\mathbf{0.97}_{\pm 0.02}$
	ROC-AUC (weighted)	$0.96_{\pm 0.01}$	$0.97_{\pm 0.01}$	$\underline{0.97}_{\pm 0.01}$	$0.97_{\pm 0.01}$	$0.96_{\pm 0.01}$	$\mathbf{0.99}_{\pm 0.01}$
	Balanced Acc	$0.93_{\pm 0.02}$	$0.93_{\pm 0.02}$	$0.93_{\pm 0.02}$	$\underline{0.93}_{\pm 0.02}$	$0.92_{\pm 0.02}$	$\mathbf{0.97}_{\pm 0.02}$
Treatment Response	F1 (weighted)	$0.51_{\pm 0.08}$	$0.45_{\pm 0.04}$	$0.49_{\pm 0.08}$	$\underline{0.53}_{\pm 0.06}$	$0.52_{\pm 0.08}$	$\mathbf{0.58}_{\pm 0.14}$
	ROC-AUC (weighted)	$0.66_{\pm 0.10}$	$0.62_{\pm 0.04}$	$0.65_{\pm 0.08}$	$0.67_{\pm 0.08}$	$\underline{0.68}_{\pm 0.07}$	$\mathbf{0.68}_{\pm 0.07}$
	Balanced Acc	$0.46_{\pm 0.12}$	$0.37_{\pm 0.07}$	$0.38_{\pm 0.08}$	$\underline{0.47}_{\pm 0.11}$	$0.47_{\pm 0.09}$	$\mathbf{0.48}_{\pm 0.17}$

remain consistently positive across all three metrics.

5.4.2 Stage 2 Only vs. MOOZY

Table 23 provides the complete task-level breakdown for the Stage 2 only versus MOOZY comparison. Stage 2 only trains the slide encoder with multi-task supervision from scratch, without Stage 1 SSL pretraining, and serves as an ablation of the self-supervised initialization. MOOZY improves over Stage 2 only on every task-metric pair, confirming that

Table 21: Unified macro-average ablation across eight held-out tasks. The table includes Stage 1 only, Stage 2 only, 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	✓	✗	✗	0.760	0.753	0.701
Stage 2 only	✗	✓	✓	0.748	0.725	0.701
MOOZY w/o case agg.	✓	✓	✗	0.771	0.789	0.729
MOOZY	✓	✓	✓	0.801	0.815	0.758

Table 22: Task-wise comparison between Stage 1 and MOOZY (Ours) across the eight slide-encoder evaluation tasks. Values are mean $\pm$ standard deviation across five folds. Relative improvement is computed as $(\text{MOOZY} - \text{Stage 1})/\text{Stage 1} \times 100$ using fold means.

Task	F1 (S1)	F1 (MOOZY)	ΔF1 (%)	AUC (S1)	AUC (MOOZY)	ΔAUC (%)	Bal. Acc (S1)	Bal. Acc (MOOZY)	ΔBal. Acc (%)
Residual Cancer Burden	$0.51_{\pm 0.05}$	$0.56_{\pm 0.05}$	+9.80	$0.63_{\pm 0.06}$	$0.74_{\pm 0.04}$	+17.46	$0.48_{\pm 0.06}$	$0.51_{\pm 0.06}$	+6.25
TP53 mutation	$0.81_{\pm 0.07}$	$0.87_{\pm 0.04}$	+7.41	$0.79_{\pm 0.07}$	$0.86_{\pm 0.06}$	+8.86	$0.80_{\pm 0.06}$	$0.86_{\pm 0.05}$	+7.50
BAP1 mutation	$0.83_{\pm 0.07}$	$0.89_{\pm 0.06}$	+7.23	$0.66_{\pm 0.23}$	$0.79_{\pm 0.12}$	+19.70	$0.66_{\pm 0.12}$	$0.78_{\pm 0.11}$	+18.18
ACVR2A mutation	$0.84_{\pm 0.08}$	$0.91_{\pm 0.05}$	+8.33	$0.77_{\pm 0.12}$	$0.91_{\pm 0.09}$	+18.18	$0.69_{\pm 0.13}$	$0.90_{\pm 0.10}$	+30.43
Histologic Grade	$0.76_{\pm 0.05}$	$0.78_{\pm 0.08}$	+2.63	$0.73_{\pm 0.05}$	$0.75_{\pm 0.15}$	+2.74	$0.76_{\pm 0.06}$	$0.77_{\pm 0.08}$	+1.32
KRAS mutation	$0.85_{\pm 0.08}$	$0.85_{\pm 0.04}$	+0.00	$0.82_{\pm 0.10}$	$0.80_{\pm 0.06}$	-2.44	$0.83_{\pm 0.07}$	$0.79_{\pm 0.10}$	-4.82
IDH Status	$0.93_{\pm 0.01}$	$0.97_{\pm 0.02}$	+4.30	$0.96_{\pm 0.01}$	$0.99_{\pm 0.01}$	+3.13	$0.92_{\pm 0.01}$	$0.97_{\pm 0.02}$	+5.43
Treatment Response	$0.55_{\pm 0.09}$	$0.58_{\pm 0.14}$	+5.45	$0.66_{\pm 0.09}$	$0.68_{\pm 0.07}$	+3.03	$0.47_{\pm 0.14}$	$0.48_{\pm 0.17}$	+2.13
Macro average	0.760	0.801	+5.43	0.753	0.815	+8.31	0.701	0.758	+8.02

Stage 1 pretraining provides a representation foundation that facilitates downstream semantic alignment. The largest gains appear on KRAS mutation AUC (+29.03%), Treatment Response (+18.37% F1, +17.07% balanced accuracy), and Residual Cancer Burden (+12.00% F1). IDH Status and Histologic Grade show smaller but consistently positive improvements.

5.5 Case Aggregator Ablation

We ablate the patient-level case aggregator by comparing MOOZY without the case aggregator (Stage 2 slide encoder with mean slide pooling) against full MOOZY. Table 24 provides the complete task-level breakdown. The consistent gains on case-level tasks confirm that explicit inter-slide modeling captures information that per-slide embeddings do not encode individually. The near-parity on the slide-level IDH Status task is expected and confirms that the aggregator does not degrade transferability when case-level context is

Table 23: Task-wise comparison between Stage 2 only and MOOZY (Ours) across the eight slide-encoder evaluation tasks. Stage 2 only trains the slide encoder with multi-task supervision but without Stage 1 SSL pretraining. Values are mean $\pm$ standard deviation across five folds. Relative improvement is computed as $(\text{MOOZY} - \text{Stage 2 only})/\text{Stage 2 only} \times 100$ using fold means.

Task	F1 (S2)	F1 (MOOZY)	ΔF1 (%)	AUC (S2)	AUC (MOOZY)	ΔAUC (%)	Bal. Acc (S2)	Bal. Acc (MOOZY)	ΔBal. Acc (%)
Residual Cancer Burden	$0.50_{\pm 0.02}$	$0.56_{\pm 0.05}$	+12.00	$0.65_{\pm 0.02}$	$0.74_{\pm 0.04}$	+13.85	$0.47_{\pm 0.05}$	$0.51_{\pm 0.06}$	+8.51
TP53 mutation	$0.77_{\pm 0.07}$	$0.87_{\pm 0.04}$	+12.99	$0.75_{\pm 0.08}$	$0.86_{\pm 0.06}$	+14.67	$0.76_{\pm 0.07}$	$0.86_{\pm 0.05}$	+13.16
BAP1 mutation	$0.85_{\pm 0.04}$	$0.89_{\pm 0.06}$	+4.71	$0.72_{\pm 0.17}$	$0.79_{\pm 0.12}$	+9.72	$0.74_{\pm 0.13}$	$0.78_{\pm 0.11}$	+5.41
ACVR2A mutation	$0.89_{\pm 0.04}$	$0.91_{\pm 0.05}$	+2.25	$0.79_{\pm 0.12}$	$0.91_{\pm 0.09}$	+15.19	$0.81_{\pm 0.16}$	$0.90_{\pm 0.10}$	+11.11
Histologic Grade	$0.76_{\pm 0.05}$	$0.78_{\pm 0.08}$	+2.63	$0.71_{\pm 0.09}$	$0.75_{\pm 0.15}$	+5.63	$0.75_{\pm 0.05}$	$0.77_{\pm 0.08}$	+2.67
KRAS mutation	$0.77_{\pm 0.08}$	$0.85_{\pm 0.04}$	+10.39	$0.62_{\pm 0.07}$	$0.80_{\pm 0.06}$	+29.03	$0.72_{\pm 0.07}$	$0.79_{\pm 0.10}$	+9.72
IDH Status	$0.95_{\pm 0.01}$	$0.97_{\pm 0.02}$	+2.11	$0.97_{\pm 0.01}$	$0.99_{\pm 0.01}$	+2.06	$0.95_{\pm 0.01}$	$0.97_{\pm 0.02}$	+2.11
Treatment Response	$0.49_{\pm 0.07}$	$0.58_{\pm 0.14}$	+18.37	$0.59_{\pm 0.06}$	$0.68_{\pm 0.07}$	+15.25	$0.41_{\pm 0.11}$	$0.48_{\pm 0.17}$	+17.07
Macro average	0.748	0.801	+7.09	0.725	0.815	+12.41	0.701	0.758	+8.13

Table 24: Task-wise comparison between MOOZY w/o case aggregator (Stage 2 slide encoder alone with mean slide pooling) and MOOZY (Ours) across the eight slide-encoder evaluation tasks. Values are mean $\pm$ standard deviation across five folds. Relative improvement is computed as $(\text{MOOZY} - \text{MOOZY w/o case aggregator})/\text{MOOZY w/o case aggregator} \times 100$ using fold means.

Task	F1 (w/o Case Agg.)	F1 (MOOZY)	ΔF1 (%)	AUC (w/o Case Agg.)	AUC (MOOZY)	ΔAUC (%)	Bal. Acc (w/o Case Agg.)	Bal. Acc (MOOZY)	ΔBal. Acc (%)
Residual Cancer Burden	$0.49_{\pm 0.04}$	$0.56_{\pm 0.05}$	+14.29	$0.64_{\pm 0.08}$	$0.74_{\pm 0.04}$	+15.62	$0.45_{\pm 0.06}$	$0.51_{\pm 0.06}$	+13.33
TP53 mutation	$0.84_{\pm 0.04}$	$0.87_{\pm 0.04}$	+3.57	$0.84_{\pm 0.06}$	$0.86_{\pm 0.06}$	+2.38	$0.85_{\pm 0.05}$	$0.86_{\pm 0.05}$	+1.18
BAP1 mutation	$0.87_{\pm 0.05}$	$0.89_{\pm 0.06}$	+2.30	$0.76_{\pm 0.14}$	$0.79_{\pm 0.12}$	+3.95	$0.76_{\pm 0.12}$	$0.78_{\pm 0.11}$	+2.63
ACVR2A mutation	$0.88_{\pm 0.05}$	$0.91_{\pm 0.05}$	+3.41	$0.89_{\pm 0.10}$	$0.91_{\pm 0.09}$	+2.25	$0.80_{\pm 0.12}$	$0.90_{\pm 0.10}$	+12.50
Histologic Grade	$0.76_{\pm 0.06}$	$0.78_{\pm 0.08}$	+2.63	$0.74_{\pm 0.10}$	$0.75_{\pm 0.15}$	+1.35	$0.76_{\pm 0.06}$	$0.77_{\pm 0.08}$	+1.32
KRAS mutation	$0.84_{\pm 0.05}$	$0.85_{\pm 0.04}$	+1.19	$0.79_{\pm 0.08}$	$0.80_{\pm 0.06}$	+1.27	$0.81_{\pm 0.06}$	$0.79_{\pm 0.10}$	-2.47
IDH Status	$0.96_{\pm 0.01}$	$0.97_{\pm 0.02}$	+1.04	$0.99_{\pm 0.01}$	$0.99_{\pm 0.01}$	+0.00	$0.96_{\pm 0.01}$	$0.97_{\pm 0.02}$	+1.04
Treatment Response	$0.53_{\pm 0.06}$	$0.58_{\pm 0.14}$	+9.43	$0.66_{\pm 0.04}$	$0.68_{\pm 0.07}$	+3.03	$0.44_{\pm 0.05}$	$0.48_{\pm 0.17}$	+9.09
Macro average	0.771	0.801	+3.89	0.789	0.815	+3.33	0.729	0.758	+3.95

uninformative.

5.6 Linear Probe Results

We complement the MLP probe evaluation with linear probing to assess how much diagnostic information is directly accessible without a nonlinear readout. Table 25 compares frozen slide-encoder representations under a shared multinomial logistic-regression classifier, while Table 26 compares MOOZY against MIL baselines that still learn task-specific patch aggregation, with each entry averaged over MeanMIL, ABMIL, CLAM, DSMIL, and TransMIL.

Table 25: Linear-probe slide encoder comparison across tasks.

Task	Metric	CHIEF	Giga-Path	PRISM	Made-leine	TITAN	MOOZY (Ours)
Residual Cancer Burden	F1	$0.34_{\pm 0.06}$	$0.32_{\pm 0.05}$	$0.33_{\pm 0.09}$	$\underline{0.36}_{\pm 0.07}$	$0.34_{\pm 0.09}$	$\mathbf{0.38}_{\pm 0.11}$
	AUC	$0.58_{\pm 0.05}$	$0.57_{\pm 0.07}$	$0.61_{\pm 0.03}$	$\underline{0.62}_{\pm 0.03}$	$0.57_{\pm 0.06}$	$\mathbf{0.66}_{\pm 0.05}$
	Bal. Acc	$0.39_{\pm 0.04}$	$0.30_{\pm 0.06}$	$0.32_{\pm 0.02}$	$\underline{0.40}_{\pm 0.05}$	$0.38_{\pm 0.08}$	$\mathbf{0.44}_{\pm 0.07}$
TP53 Mut.	F1	$0.70_{\pm 0.08}$	$0.69_{\pm 0.09}$	$0.77_{\pm 0.07}$	$0.77_{\pm 0.06}$	$\mathbf{0.84}_{\pm 0.04}$	$\underline{0.80}_{\pm 0.06}$
	AUC	$0.80_{\pm 0.09}$	$0.75_{\pm 0.06}$	$0.85_{\pm 0.06}$	$0.85_{\pm 0.06}$	$\mathbf{0.88}_{\pm 0.04}$	$\underline{0.87}_{\pm 0.03}$
	Bal. Acc	$0.70_{\pm 0.07}$	$0.71_{\pm 0.08}$	$0.78_{\pm 0.08}$	$0.78_{\pm 0.05}$	$\mathbf{0.84}_{\pm 0.03}$	$\underline{0.81}_{\pm 0.05}$
BAP1 Mut.	F1	$0.75_{\pm 0.06}$	$0.75_{\pm 0.10}$	$0.71_{\pm 0.11}$	$0.73_{\pm 0.09}$	$\underline{0.75}_{\pm 0.10}$	$\mathbf{0.78}_{\pm 0.05}$
	AUC	$0.67_{\pm 0.14}$	$0.66_{\pm 0.15}$	$0.63_{\pm 0.16}$	$\underline{0.69}_{\pm 0.14}$	$0.68_{\pm 0.09}$	$\mathbf{0.74}_{\pm 0.11}$
	Bal. Acc	$\underline{0.60}_{\pm 0.19}$	$0.56_{\pm 0.15}$	$0.54_{\pm 0.18}$	$0.60_{\pm 0.19}$	$0.59_{\pm 0.16}$	$\mathbf{0.64}_{\pm 0.12}$
ACVR2A Mut.	F1	$\mathbf{0.86}_{\pm 0.03}$	$0.69_{\pm 0.13}$	$0.77_{\pm 0.06}$	$0.78_{\pm 0.09}$	$0.81_{\pm 0.06}$	$\underline{0.83}_{\pm 0.04}$
	AUC	$0.79_{\pm 0.09}$	$0.69_{\pm 0.15}$	$0.81_{\pm 0.11}$	$0.76_{\pm 0.16}$	$\underline{0.82}_{\pm 0.04}$	$\mathbf{0.89}_{\pm 0.07}$
	Bal. Acc	$\underline{0.76}_{\pm 0.07}$	$0.52_{\pm 0.14}$	$0.68_{\pm 0.15}$	$0.66_{\pm 0.14}$	$0.69_{\pm 0.13}$	$\mathbf{0.78}_{\pm 0.10}$
Histologic Grade	F1	$0.63_{\pm 0.03}$	$\mathbf{0.68}_{\pm 0.04}$	$0.56_{\pm 0.14}$	$0.62_{\pm 0.07}$	$0.62_{\pm 0.06}$	$\underline{0.66}_{\pm 0.08}$
	AUC	$0.65_{\pm 0.05}$	$\mathbf{0.77}_{\pm 0.06}$	$0.60_{\pm 0.18}$	$0.71_{\pm 0.12}$	$0.64_{\pm 0.06}$	$\underline{0.74}_{\pm 0.11}$
	Bal. Acc	$0.64_{\pm 0.04}$	$\mathbf{0.69}_{\pm 0.04}$	$0.56_{\pm 0.14}$	$0.62_{\pm 0.09}$	$0.63_{\pm 0.07}$	$\underline{0.68}_{\pm 0.08}$
KRAS Mut.	F1	$0.59_{\pm 0.09}$	$0.71_{\pm 0.07}$	$0.59_{\pm 0.07}$	$0.66_{\pm 0.15}$	$\mathbf{0.72}_{\pm 0.08}$	$\underline{0.71}_{\pm 0.08}$
	AUC	$0.63_{\pm 0.10}$	$0.69_{\pm 0.10}$	$0.59_{\pm 0.07}$	$0.62_{\pm 0.15}$	$\mathbf{0.77}_{\pm 0.06}$	$\underline{0.73}_{\pm 0.07}$
	Bal. Acc	$0.55_{\pm 0.10}$	$\mathbf{0.73}_{\pm 0.07}$	$0.58_{\pm 0.06}$	$0.63_{\pm 0.18}$	$\underline{0.73}_{\pm 0.09}$	$0.71_{\pm 0.09}$
IDH Status	F1	$0.92_{\pm 0.02}$	$\underline{0.93}_{\pm 0.01}$	$0.89_{\pm 0.02}$	$0.90_{\pm 0.01}$	$0.93_{\pm 0.02}$	$\mathbf{0.95}_{\pm 0.02}$
	AUC	$0.96_{\pm 0.01}$	$0.97_{\pm 0.01}$	$0.96_{\pm 0.01}$	$0.96_{\pm 0.01}$	$\underline{0.98}_{\pm 0.01}$	$\mathbf{0.99}_{\pm 0.01}$
	Bal. Acc	$0.91_{\pm 0.02}$	$0.93_{\pm 0.02}$	$0.89_{\pm 0.02}$	$0.90_{\pm 0.02}$	$\underline{0.93}_{\pm 0.02}$	$\mathbf{0.95}_{\pm 0.02}$
Treatment Response	F1	$0.39_{\pm 0.06}$	$\underline{0.39}_{\pm 0.10}$	$0.39_{\pm 0.12}$	$0.36_{\pm 0.11}$	$0.34_{\pm 0.04}$	$\mathbf{0.47}_{\pm 0.11}$
	AUC	$0.64_{\pm 0.06}$	$0.64_{\pm 0.08}$	$\mathbf{0.69}_{\pm 0.10}$	$0.61_{\pm 0.05}$	$\underline{0.64}_{\pm 0.04}$	$0.60_{\pm 0.08}$
	Bal. Acc	$\underline{0.37}_{\pm 0.09}$	$0.29_{\pm 0.10}$	$0.33_{\pm 0.10}$	$0.34_{\pm 0.16}$	$0.30_{\pm 0.02}$	$\mathbf{0.38}_{\pm 0.10}$

The MLP-to-linear drop is not specific to MOOZY. Across all six slide encoders in Tables 17 and 25, replacing the MLP probe with logistic regression reduces macro weighted F1 by 0.097 on average and balanced accuracy by 0.086, with F1 drops ranging from 0.078 (TITAN) to 0.110 (PRISM, Madeleine) and balanced-accuracy drops from 0.066 (TITAN) to 0.105 (PRISM). In contrast, weighted ROC-AUC decreases by only 0.027. This pattern is consistent with a well-established finding in self-supervised representation learning, where nonlinear probes consistently outperform linear ones when evaluating frozen features, as demonstrated across the DINO [14], iBOT [120], DINOv2 [78], and DINOv3 [93]

Table 26: Linear-probe MIL comparison across tasks, averaged over MeanMIL, ABMIL, CLAM, DSMIL, and TransMIL.

Task	Metric	Backbone	UNI v2	Phikon v2	CONCH v1.5	MUSK	MOOZY (Ours)
Residual Cancer Burden	F1 (weighted)	$\underline{0.46}_{\pm 0.05}$	$0.44_{\pm 0.05}$	$0.42_{\pm 0.06}$	$\mathbf{0.47}_{\pm 0.05}$	$0.44_{\pm 0.05}$	$0.38_{\pm 0.11}$
	ROC-AUC (weighted)	$0.60_{\pm 0.06}$	$0.60_{\pm 0.06}$	$0.59_{\pm 0.06}$	$\underline{0.61}_{\pm 0.06}$	$0.59_{\pm 0.07}$	$\mathbf{0.66}_{\pm 0.05}$
	Balanced Acc	$\mathbf{0.44}_{\pm 0.06}$	$0.40_{\pm 0.06}$	$0.39_{\pm 0.06}$	$0.42_{\pm 0.06}$	$0.40_{\pm 0.06}$	$\underline{0.44}_{\pm 0.07}$
TP53 mutation	F1 (weighted)	$0.77_{\pm 0.06}$	$0.77_{\pm 0.07}$	$0.78_{\pm 0.07}$	$\mathbf{0.80}_{\pm 0.06}$	$0.79_{\pm 0.06}$	$\underline{0.80}_{\pm 0.06}$
	ROC-AUC (weighted)	$0.79_{\pm 0.07}$	$0.75_{\pm 0.09}$	$0.78_{\pm 0.08}$	$\underline{0.81}_{\pm 0.08}$	$0.80_{\pm 0.06}$	$\mathbf{0.87}_{\pm 0.03}$
	Balanced Acc	$0.76_{\pm 0.05}$	$0.77_{\pm 0.07}$	$0.77_{\pm 0.07}$	$\underline{0.79}_{\pm 0.07}$	$0.79_{\pm 0.04}$	$\mathbf{0.81}_{\pm 0.05}$
BAP1 mutation	F1 (weighted)	$0.84_{\pm 0.04}$	$0.82_{\pm 0.05}$	$0.83_{\pm 0.05}$	$\underline{0.85}_{\pm 0.05}$	$\mathbf{0.86}_{\pm 0.06}$	$0.78_{\pm 0.05}$
	ROC-AUC (weighted)	$0.66_{\pm 0.19}$	$0.65_{\pm 0.16}$	$0.67_{\pm 0.12}$	$\mathbf{0.75}_{\pm 0.13}$	$0.73_{\pm 0.13}$	$\underline{0.74}_{\pm 0.11}$
	Balanced Acc	$0.67_{\pm 0.10}$	$0.64_{\pm 0.10}$	$0.68_{\pm 0.10}$	$\mathbf{0.74}_{\pm 0.09}$	$\underline{0.73}_{\pm 0.10}$	$0.64_{\pm 0.12}$
ACVR2A mutation	F1 (weighted)	$\underline{0.83}_{\pm 0.09}$	$\mathbf{0.84}_{\pm 0.07}$	$0.81_{\pm 0.11}$	$0.82_{\pm 0.07}$	$0.79_{\pm 0.09}$	$0.83_{\pm 0.04}$
	ROC-AUC (weighted)	$\underline{0.74}_{\pm 0.10}$	$0.72_{\pm 0.16}$	$0.73_{\pm 0.13}$	$0.70_{\pm 0.14}$	$0.60_{\pm 0.19}$	$\mathbf{0.89}_{\pm 0.07}$
	Balanced Acc	$0.73_{\pm 0.09}$	$\underline{0.73}_{\pm 0.11}$	$0.68_{\pm 0.13}$	$0.70_{\pm 0.10}$	$0.65_{\pm 0.12}$	$\mathbf{0.78}_{\pm 0.10}$
Histologic Grade	F1 (weighted)	$0.74_{\pm 0.05}$	$0.74_{\pm 0.05}$	$0.72_{\pm 0.05}$	$\mathbf{0.76}_{\pm 0.05}$	$\underline{0.75}_{\pm 0.08}$	$0.66_{\pm 0.08}$
	ROC-AUC (weighted)	$0.73_{\pm 0.09}$	$0.73_{\pm 0.05}$	$0.72_{\pm 0.06}$	$\mathbf{0.76}_{\pm 0.08}$	$0.74_{\pm 0.08}$	$\underline{0.74}_{\pm 0.11}$
	Balanced Acc	$0.74_{\pm 0.05}$	$0.75_{\pm 0.06}$	$0.72_{\pm 0.05}$	$\mathbf{0.75}_{\pm 0.05}$	$\underline{0.75}_{\pm 0.07}$	$0.68_{\pm 0.08}$
KRAS mutation	F1 (weighted)	$\underline{0.78}_{\pm 0.07}$	$0.74_{\pm 0.05}$	$0.74_{\pm 0.05}$	$\mathbf{0.80}_{\pm 0.08}$	$0.76_{\pm 0.07}$	$0.71_{\pm 0.08}$
	ROC-AUC (weighted)	$\mathbf{0.74}_{\pm 0.10}$	$0.71_{\pm 0.09}$	$0.68_{\pm 0.07}$	$\underline{0.74}_{\pm 0.12}$	$0.70_{\pm 0.10}$	$0.73_{\pm 0.07}$
	Balanced Acc	$\underline{0.76}_{\pm 0.09}$	$0.69_{\pm 0.08}$	$0.68_{\pm 0.06}$	$\mathbf{0.77}_{\pm 0.09}$	$0.72_{\pm 0.08}$	$0.71_{\pm 0.09}$
IDH Status	F1 (weighted)	$0.93_{\pm 0.02}$	$0.93_{\pm 0.02}$	$0.93_{\pm 0.02}$	$\underline{0.94}_{\pm 0.02}$	$0.92_{\pm 0.02}$	$\mathbf{0.95}_{\pm 0.02}$
	ROC-AUC (weighted)	$0.96_{\pm 0.01}$	$0.97_{\pm 0.01}$	$\underline{0.97}_{\pm 0.01}$	$0.97_{\pm 0.01}$	$0.96_{\pm 0.01}$	$\mathbf{0.99}_{\pm 0.01}$
	Balanced Acc	$0.93_{\pm 0.02}$	$0.93_{\pm 0.02}$	$0.93_{\pm 0.02}$	$\underline{0.93}_{\pm 0.02}$	$0.92_{\pm 0.02}$	$\mathbf{0.95}_{\pm 0.02}$
Treatment Response	F1 (weighted)	$0.51_{\pm 0.08}$	$0.45_{\pm 0.04}$	$0.49_{\pm 0.08}$	$\mathbf{0.53}_{\pm 0.06}$	$\underline{0.52}_{\pm 0.08}$	$0.47_{\pm 0.11}$
	ROC-AUC (weighted)	$0.66_{\pm 0.10}$	$0.62_{\pm 0.04}$	$0.65_{\pm 0.08}$	$\underline{0.67}_{\pm 0.08}$	$\mathbf{0.68}_{\pm 0.07}$	$0.60_{\pm 0.08}$
	Balanced Acc	$0.46_{\pm 0.12}$	$0.37_{\pm 0.07}$	$0.38_{\pm 0.08}$	$\mathbf{0.47}_{\pm 0.11}$	$\underline{0.47}_{\pm 0.09}$	$0.38_{\pm 0.10}$

families in the natural image domain. The asymmetry between the larger F1 and balanced-accuracy drops and the smaller ROC-AUC drop further suggests that pathology slide embeddings preserve linearly recoverable ranking information, yet their class boundaries are not linearly separable, likely because clinically relevant phenotypes depend on nonlinear mixtures of tumor architecture, stromal and immune context, and inter-slide relationships.

Despite this universal drop, MOOZY retains its macro-average lead. Macro scores drop from 0.801, 0.815, and 0.758 (F1, AUC, balanced accuracy) under the MLP probe to 0.698, 0.778, and 0.674, yet MOOZY remains the macro leader in the slide-encoder comparison (Table 25). The magnitude of the drop is task dependent. IDH Status is largely

preserved, while Residual Cancer Burden, BAP1, Histologic Grade, KRAS, and Treatment Response decline more substantially, indicating that some endpoints are near-linearly accessible while others require a nonlinear readout. Per-encoder strengths also shift under linear probing, with TITAN leading on TP53 and KRAS, GigaPath on Histologic Grade, and CHIEF on ACVR2A weighted F1. These structured differences suggest that each encoder exposes different clinical signals most cleanly to a linear classifier, reflecting differences in pretraining objectives and data.

Under the MIL linear probe comparison (Table 26), the balance shifts toward MIL baselines on several tasks. MOOZY still leads on IDH Status (0.95 F1, 0.99 AUC), ACVR2A mutation (0.89 AUC, +0.15 over the next-best baseline), and TP53 mutation (0.87 AUC, +0.06 over CONCH v1.5), but MIL baselines take the lead on BAP1 mutation (MUSK 0.86 vs. MOOZY 0.78 F1), Histologic Grade (CONCH v1.5 0.76 vs. MOOZY 0.66 F1), KRAS mutation (CONCH v1.5 0.80 vs. MOOZY 0.71 F1), and Treatment Response (CONCH v1.5 0.53 vs. MOOZY 0.47 F1). This shift is expected, since MIL aggregators learn task-specific attention over patches from labeled data, whereas MOOZY compresses the entire case into a single frozen vector. Under a nonlinear MLP readout, MOOZY can recover task-relevant structure from this compressed representation and leads across the board (Table 20), but a linear classifier cannot perform this recovery, giving MIL's task-specific aggregation a stronger advantage. Even on the tasks where MIL baselines lead F1 and balanced accuracy, MOOZY remains competitive on AUC (second-best on BAP1 and Histologic Grade), confirming that the patient-level representation preserves ranking information even when linear class separation is harder to achieve.

5.7 Attention Map Analysis

To examine where each encoder concentrates on the slide, a board-certified pathologist reviewed attention maps for 20 representative WSIs across all eight held-out datasets 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 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), PRISM (1.75), CHIEF (1.88), and Madeleine (2.50). On shift, MOOZY was near balanced (2.63) and TITAN closest to exact balance (3.13), while PRISM (2.38), CHIEF (2.13), and Madeleine (2.00) were more cancer-biased.

Together, 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.

5.7.1 Attention Map Generation

Heatmaps are produced from WSIs sampled across the held-out evaluation datasets. For each slide, we generate matched visualizations for all compared encoders using the same attribution objective, ensuring a common relevance scale. Let $X = \{x_i\}_{i=1}^{N}$ denote patch tokens for one slide, and let $z^{(m)}(X)$ be the corresponding slide embedding from encoder m, where:

$$m \in \{\text{CHIEF}, \text{Madeleine}, \text{PRISM}, \text{TITAN}, \text{MOOZY}\}. \tag{26}$$

As the attribution target we use the squared ℓ_2 norm of the slide embedding:

$$\phi^{(m)}(X) = \frac{1}{2} \left\| z^{(m)}(X) \right\|_2^2, \tag{27}$$

which is an encoder-intrinsic scalar that requires no downstream task head, making comparisons across models fair. We assign each patch the Grad×Input relevance:

$$s_i^{(m)} = \left\| x_i \odot \frac{\partial \phi^{(m)}}{\partial x_i} \right\|_1 . \tag{28}$$

Intuitively, $s_i^{(m)}$ measures how much a small perturbation of patch x_i shifts $\phi^{(m)}$, so high-score regions are those most influential to the encoder's slide-level representation. The relevance vector $\{s_i^{(m)}\}_{i=1}^{N}$ is then converted into a display map. We first apply rank normalization:

$$\tilde{s}_i^{(m)} = \frac{\mathrm{rank}(s_i^{(m)})}{N}, \tag{29}$$

then project normalized patch scores to level-0 patch boxes on the thumbnail and average them per pixel where boxes overlap.

5.7.2 Attention Map Examples

A representative example is shown in Figure 11. Additional cross-model attention-map comparisons are shown in Figures 12–16.

5.8 Qualitative Analysis: Embedding Visualization

To complement quantitative results, we visualize UMAP [67] embeddings for three representative tasks (CPTAC cancer type, pan-dataset anatomical site, and TCGA cancer type) across four slide encoders (Figure 17). 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.

For all qualitative plots, t-SNE uses perplexity 25 and 1000 optimization iterations,

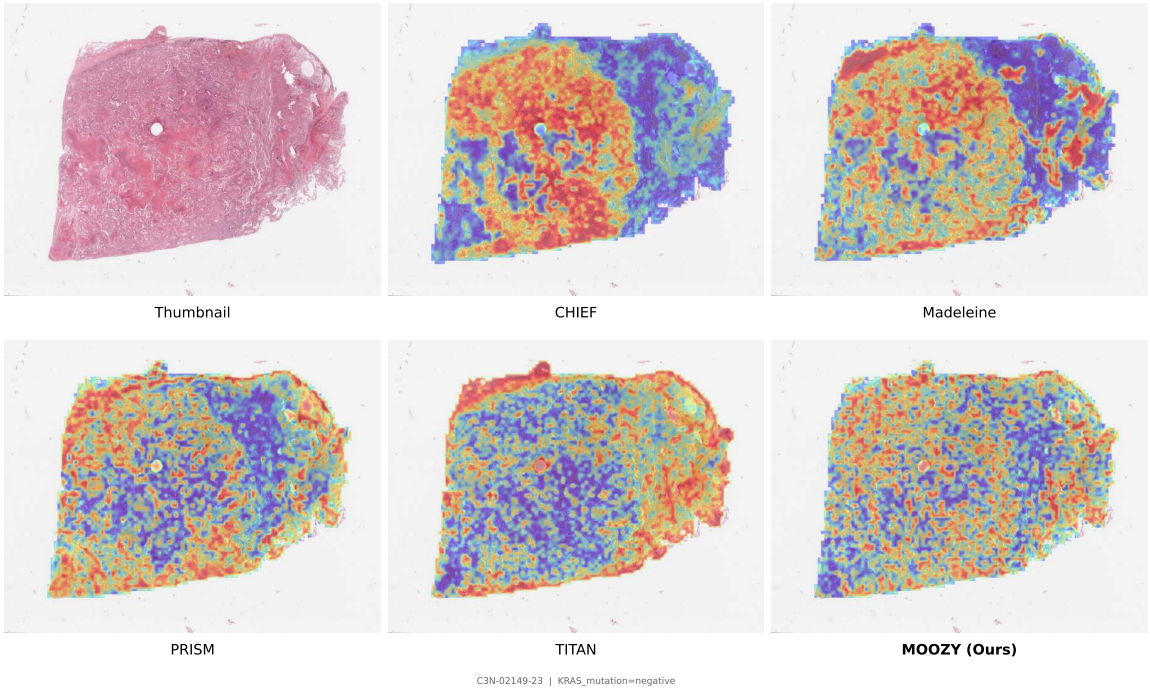


Figure 11: Attention-map comparison on a lung adenocarcinoma slide. MOOZY and TITAN: balanced, comprehensive coverage (shift 3, gap 1). PRISM: balanced shift with moderate gaps (shift 3, gap 2). CHIEF and Madeleine: cancer-biased with frequent semantic gaps (shift 2, gap 3).

while UMAP uses neighborhood size 120, minimum distance 0.3, and cosine metric. Effective sample counts are $N{=}2152$ for CPTAC cancer type, $N{=}1172$ for anatomical site, and $N{=}1280$ for TCGA cancer type.

The t-SNE layouts (Figure 18) exhibit patterns consistent with the UMAP results: MOOZY shows the clearest cluster separation on CPTAC and TCGA cancer type, TITAN is the strongest on anatomical site, and Madeleine and PRISM show comparatively weaker boundaries.

5.9 Unsupervised Embedding Geometry Analysis

We analyze embedding geometry on 3,300 unique slides using two unsupervised diagnostics shown in Figure 19: PCA compactness and bootstrap neighborhood stability. Let

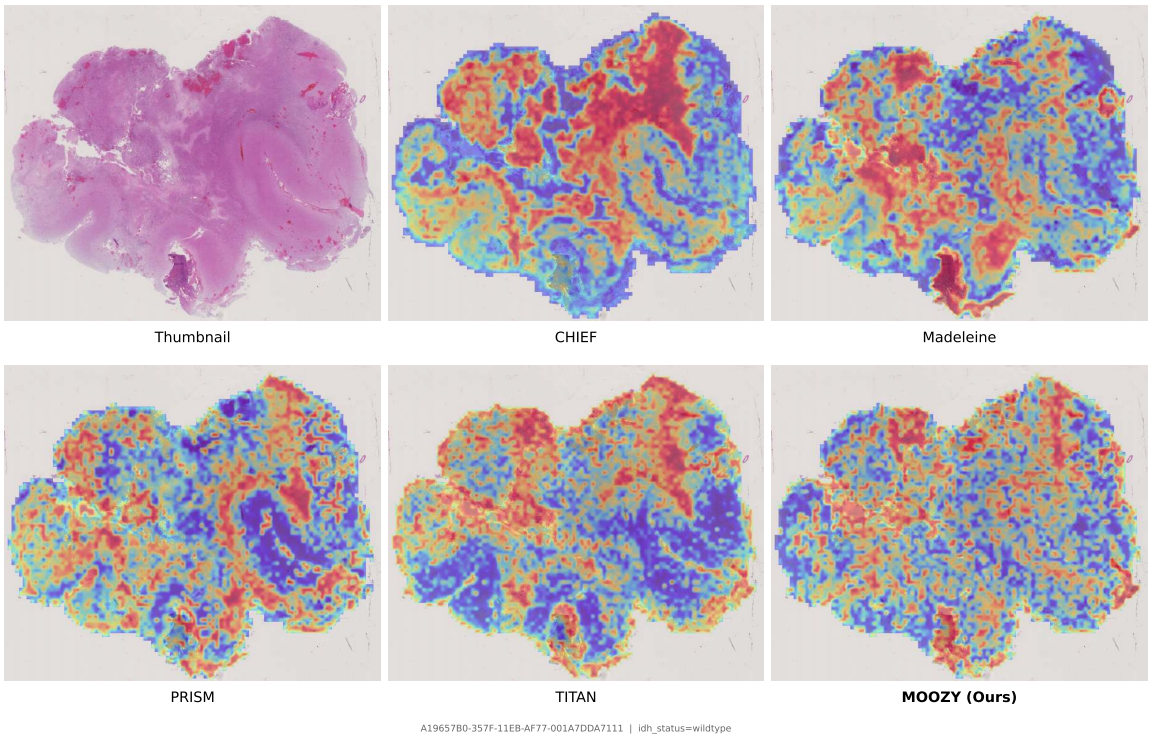


Figure 12: Additional attention map comparison across MOOZY and benchmarked models (example 2).

$X \in \mathbb{R}^{N \times D}$ denote the embedding matrix (one row per slide).

5.9.1 PCA Compactness

PCA compactness measures how many orthogonal directions are needed to explain embedding variance. We first center features:

$$\tilde{X} = X - \mathbf{1}\mu^\top, \quad \mu = \frac{1}{N}\sum_{i=1}^{N} x_i, \tag{30}$$

then compute covariance:

$$C = \frac{\tilde{X}^\top \tilde{X}}{N-1}. \tag{31}$$

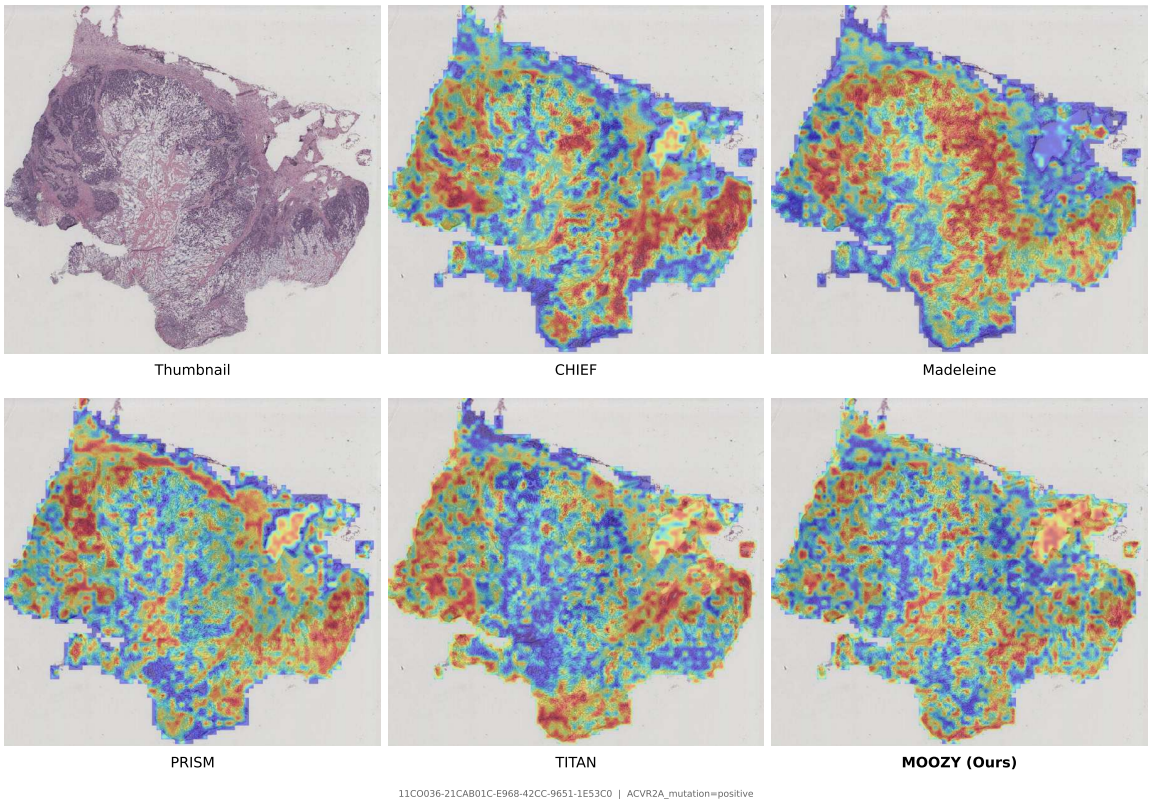


Figure 13: Additional attention map comparison across MOOZY and benchmarked models (example 3).

If $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_D$ are eigenvalues of C, we clamp:

$$\lambda_j \leftarrow \max(\lambda_j, 0) \tag{32}$$

to avoid numerical negatives, and compute cumulative explained variance:

$$V(r) = \frac{\sum_{j=1}^{r} \lambda_j}{\sum_{j=1}^{D} \lambda_j}. \tag{33}$$

For each threshold $\tau \in \{0.80, 0.90, 0.95\}$, compactness is the smallest rank r_τ such that $V(r_\tau) \geq \tau$. Lower r_τ indicates less redundancy and more information-efficient embeddings. MOOZY is the most compact encoder, requiring 9, 12, and 17 components at the 80%, 90%, and 95% thresholds, versus TITAN (17, 31, 56), CHIEF (19, 38, 67), Madeleine

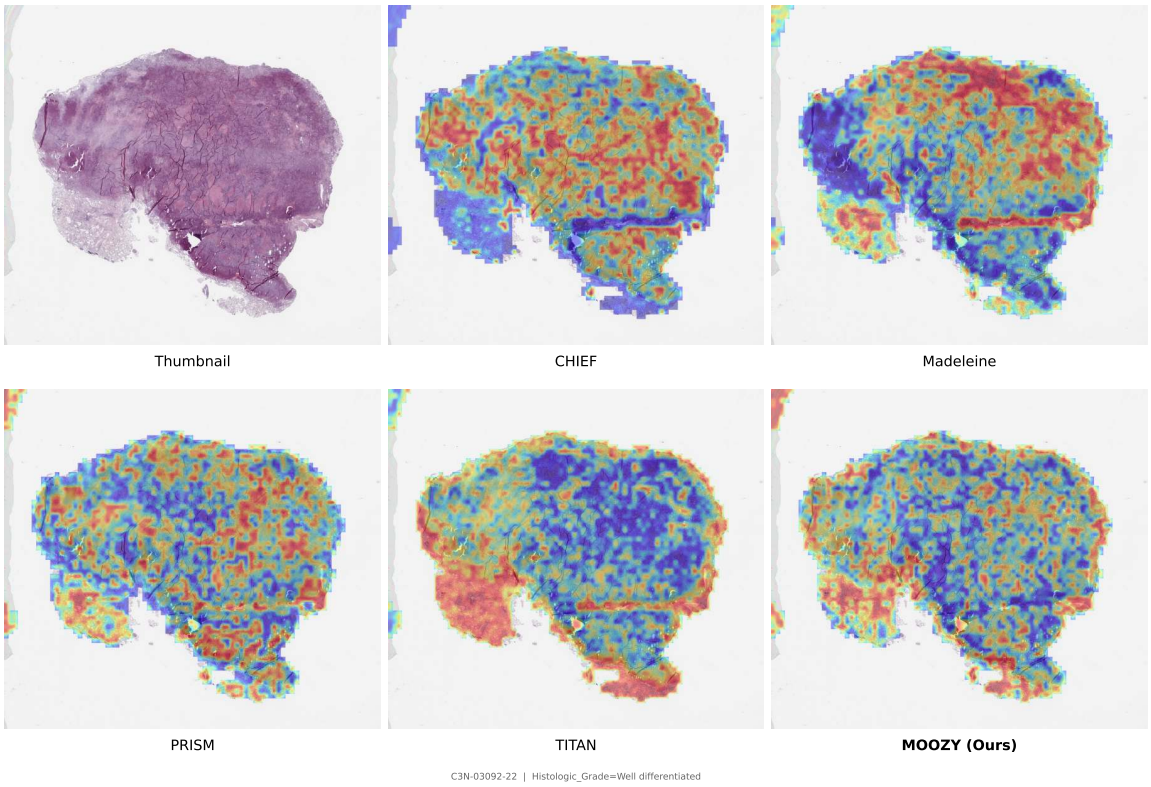


Figure 14: Additional attention map comparison across MOOZY and benchmarked models (example 4).

(17, 37, 72), PRISM (22, 45, 81), and GigaPath (58, 123, 205).

5.9.2 Bootstrap Neighborhood Stability

To quantify robustness of local neighborhood structure under sampling perturbations, we first L2-normalize each embedding:

$$\hat{x}_i = \frac{x_i}{\max(\|x_i\|_2, 10^{-12})}. \tag{34}$$

Using cosine distance, let $\mathcal{N}_k^{\text{full}}(i)$ be the k nearest neighbors of slide i on the full set (excluding self). For bootstrap repeat b, sample a subset $S_b \subset \{1, \ldots, N\}$ uniformly without replacement:

$$|S_b| = m = \text{round}(\rho N), \quad \rho = 0.8. \tag{35}$$

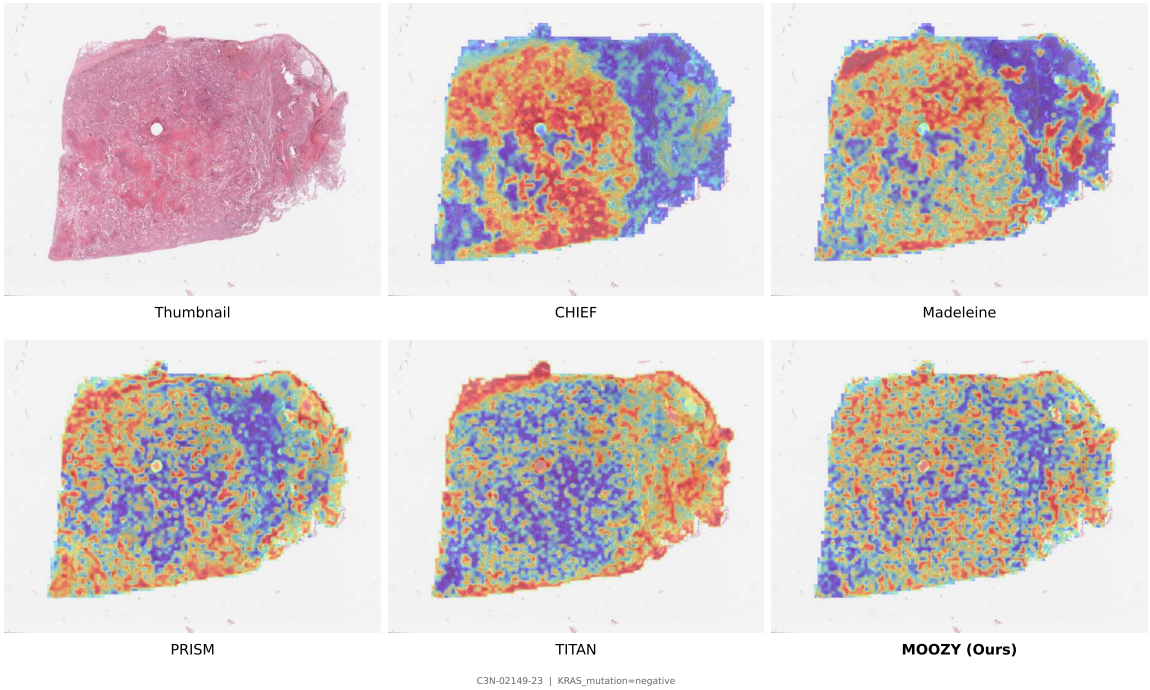


Figure 15: Additional attention map comparison across MOOZY and benchmarked models (example 5).

Recompute neighbors on the subset to obtain $\mathcal{N}_{k,b}^{\mathrm{sub}}(i)$ for $i \in S_b$, and define per-slide overlap:

$$o_{i,k,b} = \frac{\left|\mathcal{N}_k^{\mathrm{full}}(i) \cap \mathcal{N}_{k,b}^{\mathrm{sub}}(i)\right|}{k}, \tag{36}$$

repeat-level overlap:

$$\bar{o}_{k,b} = \frac{1}{|S_b|} \sum_{i \in S_b} o_{i,k,b}, \tag{37}$$

and the final stability curve over B repeats:

$$\mu_k = \frac{1}{B} \sum_{b=1}^{B} \bar{o}_{k,b}. \tag{38}$$

Higher μ_k indicates more stable local structure. In practice, encoders are near-tied on this metric: at k=30, scores span 0.7998 to 0.8032, and the mean spread across k is only 0.0029 (0.8002 to 0.8031). These unsupervised tests show that MOOZY has the strongest

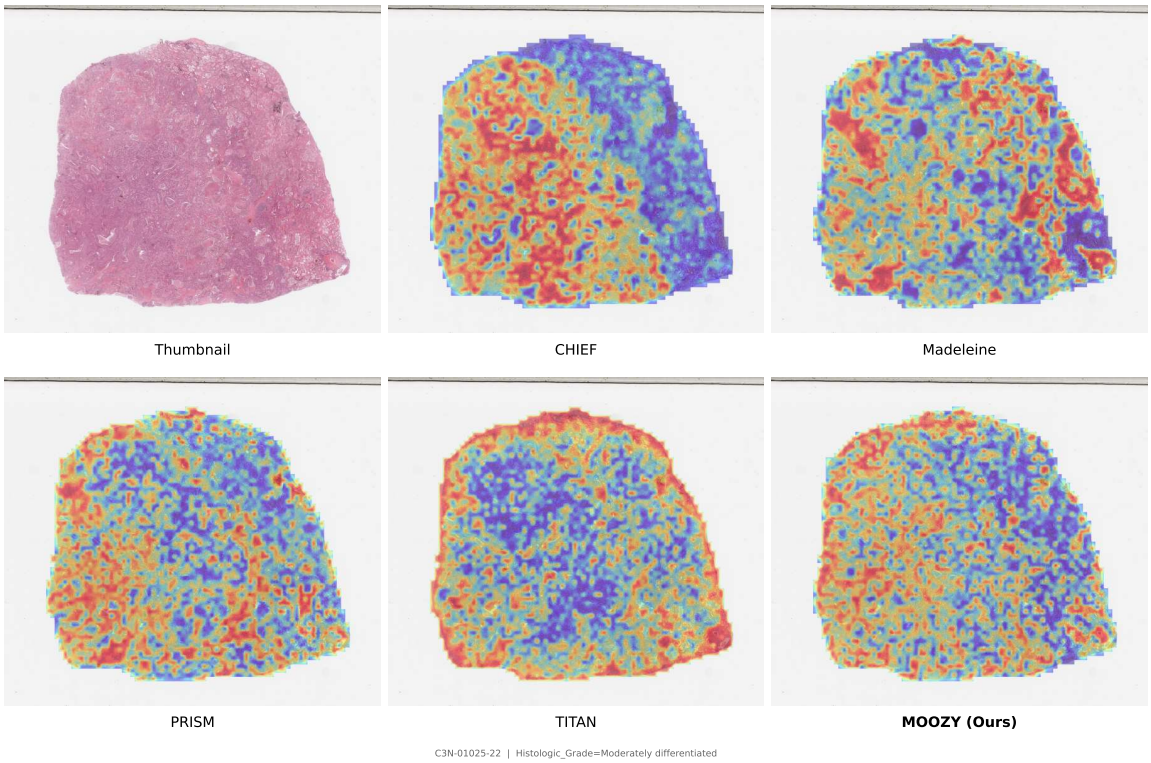


Figure 16: Additional attention map comparison across MOOZY and benchmarked models (example 6).

compactness while maintaining stability comparable to all baselines, indicating better representation efficiency without a meaningful robustness tradeoff.

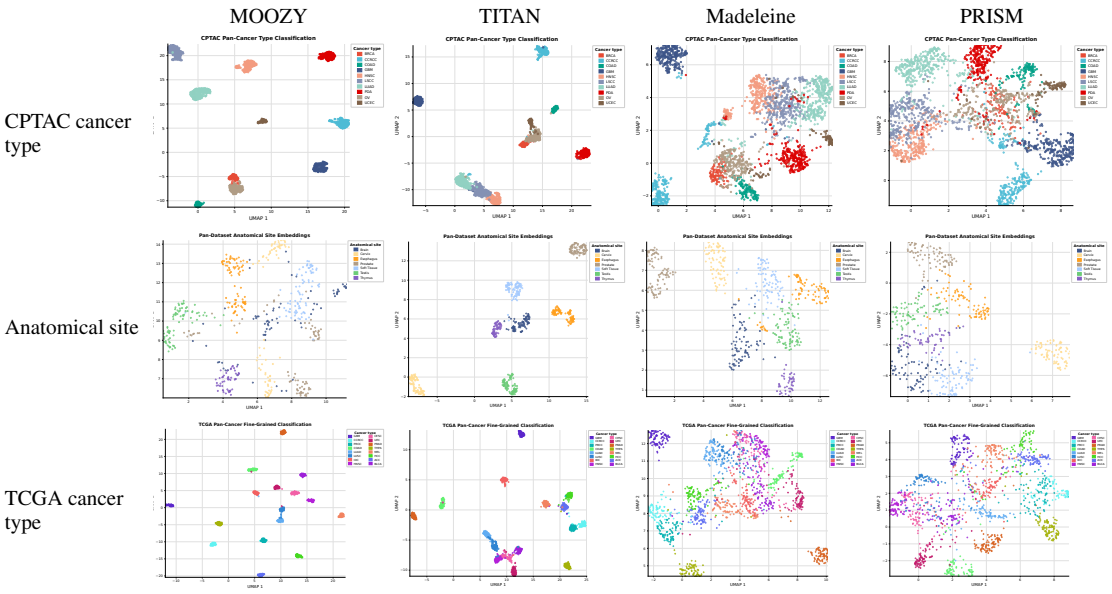


Figure 17: UMAP qualitative comparison across four slide encoders (columns) and three tasks (rows), using matched class-balanced sampling and identical reduction settings.

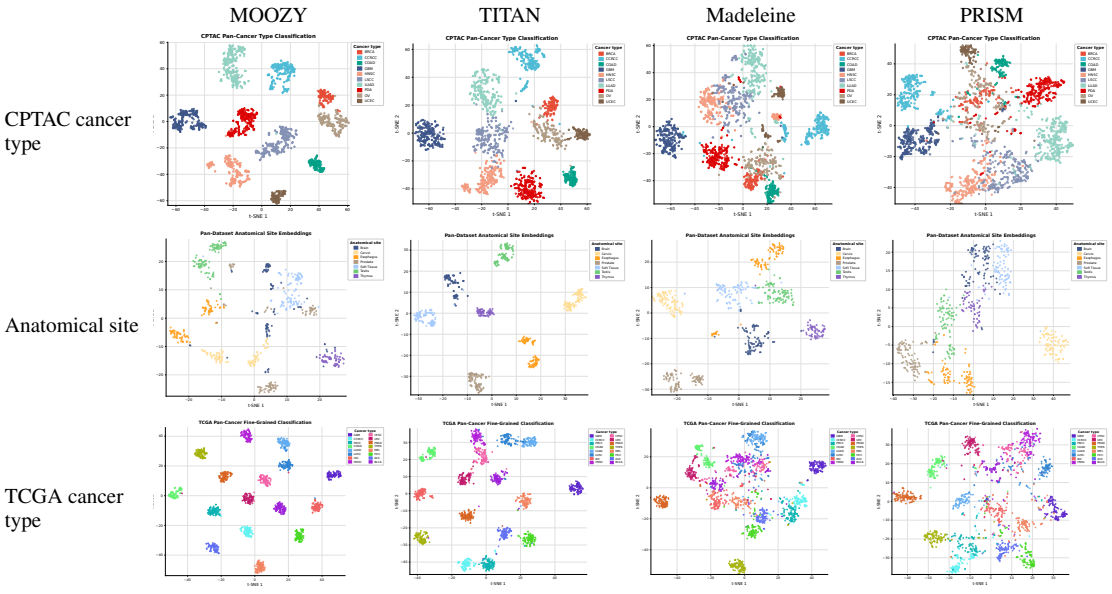


Figure 18: t-SNE qualitative comparison across four slide encoders (columns) and three tasks (rows).

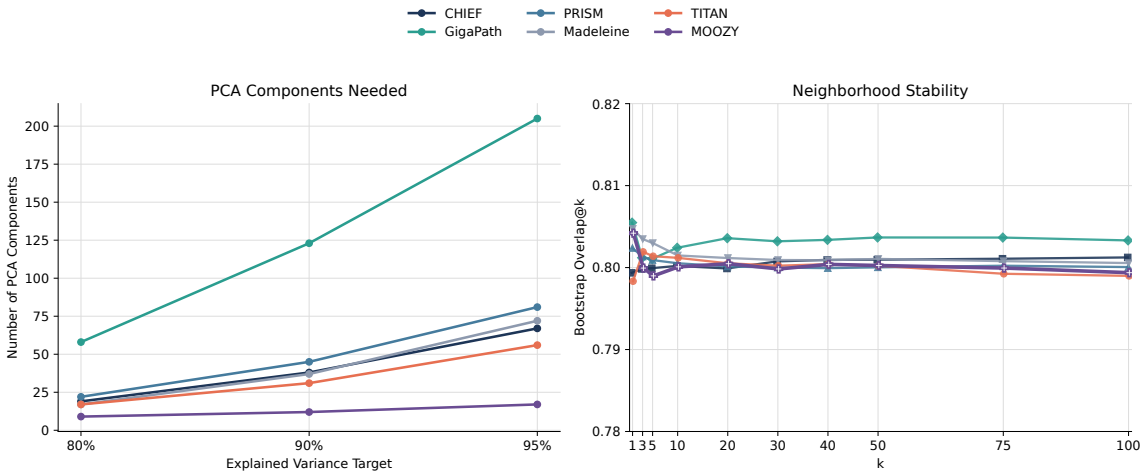


Figure 19: Encoder geometry comparison on 3,300 different slides. Left: PCA compactness, measured as the number of components needed to explain 80%, 90%, and 95% variance (lower is more compact). Right: bootstrap neighborhood stability measured by overlap@k, where each repeat randomly subsamples 80% of slides (higher is more stable).

Chapter 6

Conclusion and Future Work

We presented MOOZY, a two-stage patient-first framework that decouples vision-only slide-level self-supervised learning from patient-aware semantic alignment, and replaces the naive early/late multi-slide fusion of prior slide encoders with explicit case-level inter-slide dependency modeling. Across eight held-out tasks, MOOZY achieves best or tied-best performance on the majority of metrics against both slide encoders and MIL baselines while remaining substantially more parameter-efficient than most competing slide-level models. Together, these results show that open, public-data patient-level pretraining is sufficient to produce competitive and transferable pathology representations without proprietary slides, paired clinical reports, or billion-parameter architectures.

At the same time, an important limitation of the current formulation is its compression bottleneck. MOOZY ultimately reduces one or more gigapixel WSIs for a patient into a single case vector. That level of compression is well matched to global prediction tasks such as classification, mutation prediction, survival modeling, and retrieval, where one compact patient representation is desirable. However, it is likely too restrictive for tasks that require dense or compositional reasoning over many co-occurring findings. For example, report generation, grounded visual question answering, region-level localization, or any system that must describe multiple distinct pathologic processes within and across slides

may require preserving substantially more information than can be expressed in a single embedding. This limitation is not unique to MOOZY; it reflects a broader assumption in current slide-encoder research that whole-slide or whole-patient understanding can be faithfully summarized by one vector.

A natural future direction is therefore to move from single-vector patient representations to *multi-vector* patient representations. Rather than producing one `[CASE]` embedding, the model could output a learned set of patient latents, for example through a Perceiver-style latent array or a small bank of case tokens that attend jointly over all slides. Such a design would increase representational capacity while still retaining the efficiency benefits of learned compression. Different latent vectors could specialize to different tissue patterns, disease processes, anatomic regions, or slide subsets, enabling richer downstream reasoning than a single global summary permits. This could be especially valuable for report generation, where the model may need to preserve multiple simultaneously relevant findings rather than collapse them into one prediction-oriented representation.

Beyond representational capacity, several further directions follow naturally from this work. First, integrating paired pathology reports through slide-report co-training could improve semantic grounding while also providing a pathway toward clinically useful generation tasks. Second, combining patient-aware histomorphology with genomic, transcriptomic, or laboratory data could yield richer case representations that capture complementary sources of evidence unavailable from tissue appearance alone. Third, as more public datasets become available, the multi-task supervision regime can be expanded to additional anatomical sites, endpoints, and rare diseases, enabling a clearer study of scaling behavior at the patient level. Fourth, the patient-centric embeddings introduced here remain promising for case-level retrieval systems, where clinicians query by patient case rather than by individual slide, but future systems should ideally return not only similar cases but also the specific regions or slides that support the match. Finally, patient cases are inherently

longitudinal, with slides accumulating across distinct diagnostic, surgical, and surveillance encounters, yet MOOZY currently treats them as an unordered set. Incorporating explicit temporal ordering or encounter-level structure into the case aggregator is a promising extension for clinical tasks that hinge on disease progression over time.

Appendix A

Related Publications

The work presented in this thesis has been submitted to the European Conference on Computer Vision (ECCV 2026):

> **MOOZY: A Patient-First Foundation Model for Computational Pathology.**
> Yousef Hassan, Vincent Quoc-Huy Trinh, Christopher Pal, Mahdi S. Hosseini.
> Submitted to ECCV 2026.

To support reproducibility, all code, pretrained models, and the installable package are publicly available:

- Paper: `https://arxiv.org/abs/2603.27048`
- Project Page: `https://atlasanalyticslab.github.io/MOOZY/`
- Code: `https://github.com/AtlasAnalyticsLab/MOOZY`
- Model: `https://huggingface.co/AtlasAnalyticsLab/MOOZY`
- PyPI: `https://pypi.org/project/moozy/`

Bibliography

[1] E. Abels, L. Pantanowitz, F. Aeffner, M. D. Zarella, J. van der Laak, M. M. Bui, V. N. P. Vemuri, A. V. Parwani, J. Gibbs, E. Agosto-Arroyo, A. H. Beck, and C. Kozlowski. Computational pathology definitions, best practices, and recommendations for regulatory guidance: a white paper from the Digital Pathology Association. *The Journal of Pathology*, 249(3):286–294, 2019.

[2] A. Alagha, C. Leclerc, Y. Kotp, O. Metwally, C. Moras, P. Rentopoulos, G. Rostami, B. N. Nguyen, J. Baig, A. Khellaf, V. Q.-H. Trinh, R. Mizouni, H. Otrok, J. Bentahar, and M. S. Hosseini. AtlasPatch: Efficient tissue detection and high-throughput patch extraction for computational pathology at scale. *arXiv preprint arXiv:2602.03998*, 2026.

[3] M. B. Amin, S. B. Edge, F. L. Greene, D. R. Byrd, R. K. Brookland, M. K. Washington, J. E. Gershenwald, C. C. Compton, K. R. Hess, D. C. Sullivan, J. M. Jessup, J. D. Brierley, L. E. Gaspar, R. L. Schilsky, C. M. Balch, D. P. Winchester, E. A. Asare, M. Madera, D. M. Gress, and L. R. Meyer. *AJCC Cancer Staging Manual*. Springer, New York, NY, 8th edition, 2017.

[4] P. Bandi, O. Geessink, Q. Manson, M. Van Dijk, M. Balkenhol, M. Hermsen, B. E. Bejnordi, B. Lee, K. Paeng, A. Zhong, 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.

[5] B. E. Bejnordi, M. Veta, P. J. Van Diest, B. Van Ginneken, N. Karssemeijer, G. Litjens, J. A. Van Der Laak, M. Hermsen, Q. F. Manson, M. Balkenhol, 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.

[6] V. Belagali, S. Kapse, P. Marza, S. Das, Z. Li, S. Boutaj, P. Pati, S. Yellapragada, T. N. Nandi, R. K. Madduri, J. Saltz, P. Prasanna, S. Christodoulidis, M. Vakalopoulou, and D. Samaras. Ticon: A slide-level tile contextualizer for histopathology representation learning. *arXiv preprint arXiv:2512.21331*, 2025.

[7] E. N. Bergstrom, A. Abbasi, M. Díaz-Gay, L. Galland, S. Ladoire, S. M. Lippman, and L. B. Alexandrov. Deep learning artificial intelligence predicts homologous recombination deficiency and platinum response from histologic slides. *Journal of Clinical Oncology*, 42(30):3550–3560, oct 2024.

[8] Bioptimus. H-optimus-1. https://huggingface.co/bioptimus/H-optimus-1, 2025. Bioptimus model card. Accessed: 2026-04-19.

[9] N. Brancati, A. M. Anniciello, P. Pati, D. Riccio, G. Scognamiglio, G. Jaume, G. De Pietro, M. Di Bonito, A. Foncubierta, G. Botti, M. Gabrani, F. Feroce, and M. Frucci. Bracs: A dataset for breast carcinoma subtyping in h&e histology images. *Database*, 2022, jan 2022.

[10] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[11] W. Bulten, K. Kartasalo, P.-H. C. Chen, P. Ström, H. Pinckaers, K. Nagpal, Y. Cai, D. F. Steiner, H. Van Boven, R. Vink, et al. Artificial intelligence for diagnosis and gleason grading of prostate cancer: the panda challenge. *Nature medicine*, 28(1):154–163, 2022.

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

[13] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, and A. Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in neural information processing systems*, 33:9912–9924, 2020.

[14] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021.

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

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

[17] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PmLR, 2020.

[18] X. Chen and K. He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15750–15758, 2021.

[19] P. Courtiol, C. Maussion, M. Moarii, E. Pronier, S. Pilcer, M. Sefta, P. Manceron, S. Toldo, M. Zaslavskiy, N. Le Stang, et al. Deep learning-based classification of mesothelioma improves prediction of patient outcome. *Nature medicine*, 25(10):1519–1525, 2019.

[20] T. Darcet, M. Oquab, J. Mairal, and P. Bojanowski. Vision transformers need registers. In *The Twelfth International Conference on Learning Representations*, 2024.

[21] J. Ding, S. Ma, L. Dong, X. Zhang, S. Huang, W. Wang, N. Zheng, and F. Wei. Longnet: Scaling transformers to 1,000,000,000 tokens. *arXiv preprint arXiv:2307.02486*, 2023.

[22] T. Ding, S. J. Wagner, A. H. Song, R. J. Chen, M. Y. Lu, A. Zhang, A. J. Vaidya, G. Jaume, M. Shaban, A. Kim, et al. Multimodal whole slide foundation model for pathology. *arXiv preprint arXiv:2411.19666*, 2024.

[23] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[24] N. J. Edwards, M. Oberti, R. R. Thangudu, S. Cai, P. B. McGarvey, S. Jacob, S. Madhavan, and K. A. Ketchum. The CPTAC data portal: a resource for cancer proteomics research. *Journal of Proteome Research*, 14(6):2707–2713, 2015.

[25] J. G. Elmore, G. M. Longton, P. A. Carney, B. M. Geller, T. Onega, A. N. A. Tosteson, H. D. Nelson, M. S. Pepe, K. H. Allison, S. J. Schnitt, F. P. O'Malley, and D. L. Weaver. Diagnostic concordance among pathologists interpreting breast biopsy specimens. *JAMA*, 313(11):1122–1132, 2015.

[26] N. Farahani, A. V. Parwani, and L. Pantanowitz. Whole slide imaging in pathology: advantages, limitations, and emerging perspectives. *Pathology and Laboratory Medicine International*, 7:23–33, 2015.

[27] A. Filiot, N. Dop, O. Tchita, A. Riou, R. Dubois, T. Peeters, D. Valter, M. Scalbert, C. Saillard, G. Robin, and A. Olivier. Distilling foundation models for robust and efficient models in digital pathology. *arXiv preprint arXiv:2501.16239*, 2025.

[28] A. Filiot, R. Ghermi, A. Olivier, P. Jacob, L. Fidon, A. Camara, A. Mac Kain, C. Saillard, and J.-B. Schiratti. Scaling self-supervised learning for histopathology with masked image modeling. *medRxiv*, 2023.

[29] A. Filiot, P. Jacob, A. Mac Kain, and C. Saillard. Phikon-v2, a large and public feature extractor for biomarker prediction. *arXiv preprint arXiv:2409.09173*, 2024.

[30] A. H. Fischer, K. A. Jacobson, J. Rose, and R. Zeller. Hematoxylin and eosin staining of tissue and cell sections. *Cold Spring Harbor Protocols*, 2008(5):pdb.prot4986, 2008.

[31] L. Galland, E. Ballot, H. Mananet, R. Boidot, J. Lecuelle, J. Albuisson, L. Arnould, I. Desmoulins, D. Mayeur, C. Kaderbhai, S. Ilie, A. Hennequin, A. Bergeron, V. Derangère, F. Ghiringhelli, C. Truntzer, and S. Ladoire. Efficacy of platinum-based chemotherapy in metastatic breast cancer and hrd biomarkers: utility of exome sequencing. *npj Breast Cancer*, 8(1):28, mar 2022.

[32] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. Richemond, E. Buchatskaya, C. Doersch, B. Avila Pires, Z. Guo, M. Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020.

[33] GTEx Consortium. The Genotype-Tissue Expression (GTEx) project. *Nature Genetics*, 45(6):580–585, 2013.

[34] A. Gu and T. Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.

[35] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022.

[36] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020.

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

[38] D. Hendrycks and K. Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016.

[39] J. Hoffmann, S. Borgeaud, A. Mensch, E. Buchatskaya, T. Cai, E. Rutherford, D. d. L. Casas, L. A. Hendricks, J. Welbl, A. Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.

[40] M. S. Hosseini, B. E. Bejnordi, V. Q.-H. Trinh, L. Chan, D. Hasan, X. Li, S. Yang, T. Kim, H. Zhang, T. Wu, et al. Computational pathology: a survey review and the way forward. *Journal of Pathology Informatics*, 15:100357, 2024.

[41] X. Hou, C. Jiang, A. Kondepudi, Y. Lyu, A. Chowdury, H. Lee, and T. C. Hollon. A self-supervised framework for learning whole slide representations. *arXiv preprint arXiv:2402.06188*, 2024.

[42] S. Hua, F. Yan, T. Shen, L. Ma, and X. Zhang. Pathoduet: Foundation models for pathological slide analysis of h&e and ihc stains. *Medical Image Analysis*, 97:103289, 2024.

[43] G. Huang, Y. Sun, Z. Liu, D. Sedra, and K. Q. Weinberger. *Deep Networks with Stochastic Depth*, pages 646–661. Springer International Publishing, 2016.

[44] M. Ilse, J. Tomczak, and M. Welling. Attention-based deep multiple instance learning. In *International conference on machine learning*, pages 2127–2136. PMLR, 2018.

[45] G. Jaume, L. Oldenburg, A. Vaidya, R. J. Chen, D. F. Williamson, T. Peeters, A. H. Song, and F. Mahmood. Transcriptomics-guided slide representation learning in computational pathology. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

[46] G. Jaume, A. Vaidya, A. Zhang, A. H. Song, R. J. Chen, S. Sahai, D. Mo, E. Madrigal, L. Phi Le, and F. Mahmood. Multistain pretraining for slide representation learning in pathology. In *European Conference on Computer Vision*, pages 19–37. Springer, 2024.

[47] M. Kang, H. Song, S. Park, D. Yoo, and S. Pereira. Benchmarking self-supervised learning on diverse pathology datasets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3344–3354, 2023.

[48] J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

[49] M. Karasikov, J. van Doorn, N. Känzig, M. E. Cesur, H. M. Horlings, R. Berke, F. Tang, and S. Otálora. Training state-of-the-art pathology foundation models with orders of magnitude less data, 2025.

[50] J. N. Kather, L. R. Heij, H. I. Grabsch, C. Loeffler, A. Echle, H. S. Muti, J. Krause, J. M. Niehues, K. A. Sommer, P. Bankhead, et al. Pan-cancer image-based detection of clinically actionable genetic alterations. *Nature cancer*, 1(8):789–799, 2020.

[51] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026, 2023.

[52] D. Komura and S. Ishikawa. Machine learning methods for histopathological image analysis. *Computational and Structural Biotechnology Journal*, 16:34–42, 2018.

[53] L. F. Kozachenko and N. N. Leonenko. Sample estimate of the entropy of a random vector. *Problemy Peredachi Informatsii*, 23(2):9–16, 1987.

[54] V. Kumar, A. K. Abbas, and J. C. Aster. *Robbins and Cotran Pathologic Basis of Disease*. Elsevier, Philadelphia, PA, 10th edition, 2020.

[55] T. Lazard, M. Lerousseau, E. Decencière, and T. Walter. Giga-ssl: Self-supervised learning for gigapixel images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 4812–4822, 2023.

[56] Y. Lee, H. R. Oh, A. Bychlov, J. Fukuoka, R. K. Kaushal, A. Sahay, R. Yadav, S. Sarioglu, S. Balci, İ. Türkmen, Y. Tolkach, C. Harder, J.-H. Choi, and S. Ahn. REport Generation of pathology using Pan-Asia Giga-pixel WSIs (2025). `https://doi.org/10.5281/zenodo.15081614`, 2025. Zenodo record for the MICCAI 2025 challenge description.

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

[58] S. C. Lester. *Manual of Surgical Pathology*. Saunders/Elsevier, Philadelphia, PA, 3rd edition, 2010.

[59] B. Li, Y. Li, and K. W. Eliceiri. Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14318–14328, 2021.

[60] J. Li, Y. Chen, H. Chu, Q. Sun, T. Guan, A. Han, and Y. He. 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*, pages 11323–11332, 2024.

[61] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. W. M. van der Laak, B. van Ginneken, and C. I. Sánchez. A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42:60–88, 2017.

[62] I. Loshchilov and F. Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

[63] M. Y. Lu, B. Chen, D. F. Williamson, R. J. Chen, K. Ikamura, G. Gerber, I. Liang, L. P. Le, T. Ding, A. V. Parwani, et al. A foundational multimodal vision language ai assistant for human pathology. *arXiv preprint arXiv:2312.07814*, 2023.

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

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

[66] A. Madabhushi and G. Lee. Image analysis and machine learning in digital pathology: Challenges and opportunities. *Medical Image Analysis*, 33:170–175, 2016.

[67] L. McInnes, J. Healy, and J. Melville. UMAP: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

[68] D. M. Metter, T. J. Colgan, S. T. Leung, C. F. Timmons, and J. Y. Park. Trends in the US and Canadian pathologist workforces from 2007 to 2017. *JAMA Network Open*, 2(5):e194337, 2019.

[69] National Cancer Institute. The cancer genome atlas program (tcga). `https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga`. Accessed: 2026-04-19.

[70] D. Nechaev, A. Pchelnikov, and E. Ivanova. Hibou: A family of foundational vision transformers for pathology. *arXiv preprint arXiv:2406.05074*, 2024.

[71] P. C. Neto, D. Montezuma, S. P. Oliveira, D. Oliveira, J. Fraga, A. Monteiro, J. Monteiro, L. Ribeiro, S. Gonçalves, S. Reinhard, et al. An interpretable machine learning system for colorectal cancer diagnosis from pathology slides. *npj Precision Oncology*, 8(1):56, 2024.

[72] P. C. Neto, S. P. Oliveira, D. Montezuma, J. Fraga, A. Monteiro, L. Ribeiro, S. Gonçalves, I. M. Pinto, and J. S. Cardoso. imil4path: A semi-supervised interpretable approach for colorectal whole-slide images. *Cancers*, 14(10):2489, 2022.

[73] M. K. K. Niazi, A. V. Parwani, and M. N. Gurcan. Digital pathology and artificial intelligence. *The Lancet Oncology*, 20(5):e253–e261, 2019.

[74] T. Nicke, D. Schacherer, J. R. Schäfer, N. Artysh, A. Prasse, A. Homeyer, A. Schenk, H. Höfener, and J. Lotz. Tissue concepts v2: A supervised foundation model for whole slide images. *arXiv preprint arXiv:2507.05742*, 2025.

[75] T. Nicke, J. R. Schaefer, H. Hoefener, F. Feuerhake, D. Merhof, F. Kiessling, and J. Lotz. Tissue concepts: Supervised foundation models in computational pathology. *Computers in Biology and Medicine*, 186:109621, 2025.

[76] S. P. Oliveira, D. Montezuma, A. Moreira, D. Oliveira, P. C. Neto, A. Monteiro, J. Monteiro, L. Ribeiro, S. Gonçalves, I. M. Pinto, and J. S. Cardoso. A cad system for automatic dysplasia grading on h&e cervical whole-slide images. *Scientific Reports*, 13(1):3970, mar 2023.

[77] S. P. Oliveira, P. C. Neto, J. Fraga, D. Montezuma, A. Monteiro, J. Monteiro, L. Ribeiro, S. Gonçalves, I. M. Pinto, and J. S. Cardoso. Cad systems for colorectal cancer from wsi are still not ready for clinical acceptance. *Scientific Reports*, 11(1):1–15, 2021.

[78] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

[79] L. Pantanowitz, P. N. Valenstein, A. J. Evans, K. J. Kaplan, J. D. Pfeifer, D. C. Wilbur, L. C. Collins, and T. J. Colgan. Review of the current state of whole slide imaging in pathology. *Journal of Pathology Informatics*, 2:36, 2011.

[80] M. Peikari, S. Salama, S. Nofech-Mozes, and A. L. Martel. Automatic cellularity assessment from post-treated breast surgical specimens. *Cytometry Part A*, 91(11):1078–1087, oct 2017.

[81] O. Press, N. A. Smith, and M. Lewis. Train short, test long: Attention with linear biases enables input length extrapolation. In *International Conference on Learning Representations*, 2022.

[82] S. Rajaram, P. Acosta, V. Panwar, V. Jarmale, A. Christie, J. Jasti, V. Margulis, D. Rakheja, J. Cheville, B. C. Leibovich, A. Parker, J. Brugarolas, and P. Kapur. Prediction of driver mutation heterogeneity in renal cancer from histopathology slides using deep learning. `https://doi.org/10.25452/figshare.plus.c.5983795.v1`, 2022. Figshare+ dataset record.

[83] N. Ravi, V. Gabeur, Y.-T. Hu, R. Hu, C. Ryali, T. Ma, H. Khedr, R. Rädle, C. Rolland, L. Gustafson, et al. SAM 2: Segment Anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024.

[84] A. A. Renshaw, M. Mena-Allauca, E. W. Gould, and S. J. Sirintrapun. Synoptic reporting: Evidence-based review and future directions. *JCO Clinical Cancer Informatics*, 2:1–9, 2018.

[85] T. Roetzer-Pejrimovsky, A. C. Moser, B. Atli, C. C. Vogel, P. A. Mercea, R. Prihoda, E. Gelpi, C. Haberler, R. Höftberger, J. A. Hainfellner, B. Baumann, G. Langs, and A. Woehrer. The digital brain tumour atlas, an open histopathology resource. *Scientific Data*, 9(1):55, 2022.

[86] T. Roetzer-Pejrimovsky, A. C. Moser, B. Atli, C. C. Vogel, P. A. Mercea, R. Prihoda, E. Gelpi, C. Haberler, R. Höftberger, J. A. Hainfellner, B. Baumann, G. Langs, and A. Woehrer. The digital brain tumour atlas, an open histopathology resource [data set], 2022.

[87] A. Sablayrolles, M. Douze, C. Schmid, and H. Jégou. Spreading vectors for similarity search. In *International Conference on Learning Representations*, 2019.

[88] C. Saillard, R. Jenatton, F. Llinares-López, Z. Mariet, D. Cahané, E. Durand, and J.-P. Vert. H-optimus-0. `https://github.com/bioptimus/releases/tree/main/models/h-optimus/v0`, 2024. Bioptimus model release.

[89] S. J. Sammut. Multi-omic machine learning predictor of breast cancer therapy response, 2022.

[90] G. Shaikovski, A. Casson, K. Severson, E. Zimmermann, Y. K. Wang, J. D. Kunz, J. A. Retamero, G. Oakley, D. Klimstra, C. Kanan, et al. Prism: A multimodal generative foundation model for slide-level histopathology. *arXiv preprint arXiv:2405.10254*, 2024.

[91] D. Shao, R. J. Chen, A. H. Song, J. Runevic, M. Y. Lu, T. Ding, and F. Mahmood. Do multiple instance learning models transfer? In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 54219–54238. PMLR, 2025.

[92] Z. Shao, H. Bian, Y. Chen, Y. Wang, J. Zhang, X. Ji, et al. Transmil: Transformer based correlated multiple instance learning for whole slide image classification. *Advances in neural information processing systems*, 34:2136–2147, 2021.

[93] O. Siméoni, H. V. Vo, M. Seitzer, F. Baldassarre, M. Oquab, C. Jose, V. Khalidov, M. Szafraniec, S. Yi, M. Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

[94] O.-J. Skrede, S. De Raedt, A. Kleppe, T. S. Hveem, K. Liestøl, J. Maddison, H. A. Askautrud, M. Pradhan, J. A. Nesheim, F. Albregtsen, et al. Deep learning for prediction of colorectal cancer outcome: a discovery and validation study. *The Lancet*, 395(10221):350–360, 2020.

[95] Sophont AI. How to train a state-of-the-art pathology foundation model with $1.6k. `https://sophontai.com/blog/openmidnight.html`, 2025. OpenMidnight release note. Accessed: 2026-04-19.

[96] J. R. Srigley, T. McGowan, A. MacLean, M. Raby, J. Ross, S. Kramer, and C. Sawka. Standardized synoptic cancer pathology reporting: A population-based approach. *Journal of Surgical Oncology*, 99(8):517–524, 2009.

[97] C. L. Srinidhi, O. Ciga, and A. L. Martel. Deep neural network models for computational histopathology: A survey. *Medical Image Analysis*, 67:101813, 2021.

[98] H. Sung, J. Ferlay, R. L. Siegel, M. Laversanne, I. Soerjomataram, A. Jemal, and F. Bray. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*, 71(3):209–249, 2021.

[99] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2818–2826. IEEE, June 2016.

[100] W. Tang, F. Zhou, S. Huang, X. Zhu, Y. Zhang, and B. Liu. Feature re-embedding: Towards foundation model-level performance in computational pathology. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11343–11352, 2024.

[101] H. Touvron, M. Cord, A. Sablayrolles, G. Synnaeve, and H. Jegou. Going deeper with image transformers. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 32–42. IEEE, Oct. 2021.

[102] A. Vaidya, A. Zhang, G. Jaume, A. H. Song, T. Ding, S. J. Wagner, M. Y. Lu, P. Doucet, H. Robertson, C. Almagro-Perez, R. J. Chen, D. ElHarouni, G. Ayoub, C. Bossi, K. L. Ligon, G. Gerber, L. P. Le, and F. Mahmood. Molecular-driven foundation model for oncologic pathology. *arXiv preprint arXiv:2501.16652*, 2025.

[103] J. van der Laak, G. Litjens, and F. Ciompi. Deep learning in histopathology: the path to the clinic. *Nature Medicine*, 27(5):775–784, 2021.

[104] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, volume 30, 2017.

[105] E. Vorontsov, A. Bozkurt, A. Casson, G. Shaikovski, M. Zelechowski, S. Liu, K. Severson, E. Zimmermann, J. Hall, N. Tenenholtz, et al. Virchow: A million-slide digital pathology foundation model. *arXiv preprint arXiv:2309.07778*, 2023.

[106] T. Wang and P. Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International Conference on Machine Learning*, pages 9929–9939. PMLR, 2020.

[107] X. Wang, J. Zhao, E. Marostica, W. Yuan, J. Jin, J. Zhang, R. Li, H. Tang, K. Wang, Y. Li, et al. A pathology foundation model for cancer diagnosis and prognosis prediction. *Nature*, 634(8035):970–978, 2024.

[108] J. Wei, A. Suriawinata, B. Ren, X. Liu, M. Lisovsky, L. Vaickus, C. Brown, M. Baker, N. Tomita, L. Torresani, J. Wei, and S. Hassanpour. A petri dish for histopathology image analysis. In *International Conference on Artificial Intelligence in Medicine*, pages 11–24. Springer, 2021.

[109] J. Wei, Y. Tay, R. Bommasani, C. Raffel, B. Zoph, S. Borgeaud, D. Yogatama, M. Bosma, D. Zhou, D. Metzler, et al. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*, 2022.

[110] J. W. Wei, L. J. Tafe, Y. A. Linnik, L. J. Vaickus, N. Tomita, and S. Hassanpour. Pathologist-level classification of histologic patterns on resected lung adenocarcinoma slides with deep neural networks. *Scientific reports*, 9(1):1–8, 2019.

[111] S. Wilkinson, H. Ye, F. Karzai, S. A. Harmon, N. T. Terrigino, D. J. VanderWeele, J. R. Bright, R. Atway, S. Y. Trostel, N. V. Carrabba, N. C. Whitlock, S. M. Walker, R. T. Lis, H. A. Sater, B. J. Capaldo, R. A. Madan, J. L. Gulley, G. Chun, M. J. Merino, P. A. Pinto, D. C. Salles, H. B. Kaur, T. L. Lotan, D. J. Venzon, P. L. Choyke, B. Turkbey, W. L. Dahut, and A. G. Sowalsky. Nascent prostate cancer heterogeneity drives evolution and resistance to intense hormonal therapy. *medRxiv*, 2020.

[112] J. Xiang, X. Wang, X. Zhang, Y. Xi, F. Eweje, Y. Chen, Y. Li, C. Bergstrom, M. Gopaulchan, T. Kim, K.-H. Yu, S. Willens, F. M. Olguin, J. J. Nirschl, J. Neal, M. Diehn, S. Yang, and R. Li. A vision–language foundation model for precision oncology. *Nature*, 2025.

[113] J. Xiang and J. Zhang. Exploring low-rank property in multiple instance learning for whole slide image classification. In *International Conference on Learning Representations*, 2023.

[114] H. Xu, N. Usuyama, J. Bagga, S. Zhang, R. Rao, T. Naumann, C. Wong, Z. Gero, J. González, Y. Gu, et al. A whole-slide foundation model for digital pathology from real-world data. *Nature*, 630(8015):181–188, 2024.

[115] H. Xu, N. Usuyama, J. Bagga, S. Zhang, R. Rao, T. Naumann, C. Wong, Z. Gero, J. González, Y. Gu, Y. Xu, M. Wei, W. Wang, S. Ma, F. Wei, J. Yang, C. Li, J. Gao, J. Rosemon, T. Bower, S. Lee, R. Weerasinghe, B. J. Wright, A. Robicsek, B. Piening, C. Bifulco, S. Wang, and H. Poon. A whole-slide foundation model for digital pathology from real-world data. *Nature*, 2024.

[116] Y. Xu, Y. Wang, F. Zhou, J. Ma, C. Jin, S. Yang, J. Li, Z. Zhang, C. Zhao, H. Zhou, Z. Li, H. Lin, X. Wang, J. Wang, A. Han, R. C. K. Chan, L. Liang, X. Zhang, and H. Chen. A multimodal knowledge-enhanced whole-slide pathology foundation model. *Nature Communications*, 16:11406, 2025.

[117] F. Yan, J. Wu, J. Li, W. Wang, J. Lu, W. Chen, Z. Gao, J. Li, H. Yan, J. Ma, et al. Pathorchestra: A comprehensive foundation model for computational pathology with over 100 diverse clinical-grade tasks. *arXiv preprint arXiv:2503.24345*, 2025.

[118] J. Zbontar, L. Jing, I. Misra, Y. LeCun, and S. Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *International Conference on Machine Learning*, pages 12310–12320. PMLR, 2021.

[119] H. Zhang, Y. Meng, Y. Zhao, Y. Qiao, X. Yang, S. E. Coupland, and Y. Zheng. 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*, pages 18802–18812, 2022.

[120] J. Zhou, C. Wei, H. Wang, W. Shen, C. Xie, A. Yuille, and T. Kong. ibot: Image bert pre-training with online tokenizer. *arXiv preprint arXiv:2111.07832*, 2021.

[121] M. Zhu, B. Ren, R. Richards, M. Suriawinata, N. Tomita, and S. Hassanpour. 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.

[122] E. Zimmermann, E. Vorontsov, J. Viret, A. Casson, M. Zelechowski, G. Shaikovski, N. Tenenholtz, J. Hall, D. Klimstra, R. Yousfi, et al. Virchow2: Scaling self-supervised mixed magnification models in pathology. *arXiv preprint arXiv:2408.00738*, 2024.