Molecular Aspects of Medicine 112 (2026) 101504

Contents lists available at ScienceDirect

Molecular Aspects of Medicine

journal homepage: www.elsevier.com/locate/mam

Role of foundation models in data-driven tissue diagnostics☆

Mohsin Bilal [a,*], Aadam [b], Manahil Raza [c], Anas Alsuhaibani [a], Youssef N. Altherwy [a], Abdulrahman Alabduljabbar [a], Fahdah Almarshad [a], Paul Golding [d], Nasir Rajpoot [c,e,*]

[a] Information Systems Department, College of Computer Engineering and Sciences, Prince Sattam bin Abdulaziz University, Al-Kharj, Saudi Arabia
[b] Luddy School of Informatics, Computing, and Engineering, Indiana University Indianapolis, United States of America
[c] Tissue Image Analytics Centre, Department of Computer Science, University of Warwick, Coventry, United Kingdom
[d] Frontier AI, San Francisco Bay Area, United States of America
[e] Histofy Ltd, The Venture Centre, University of Warwick Science Park, Coventry, United Kingdom

ARTICLE INFO

Keywords:
Tissue diagnostics
Computational pathology
Artificial intelligence
Foundation models
Downstream analysis
Reporting recommendations

ABSTRACT

Artificial intelligence (AI)-based tissue diagnostics is entering a new phase driven by pathology foundation models: large-scale encoders and multimodal systems pretrained with self-supervised and vision–language objectives. Yet the evidence base has expanded faster than the methods used to evaluate and interpret claimed capabilities. The central clinical question is no longer whether these models can achieve strong retrospective performance, but what *new capabilities* they provide, under what conditions those capabilities are demonstrated, and whether they can translate into reliable diagnostics in clinical settings. This review makes three contributions. First, we define histopathology-centric foundation models and distill the technical factors that shape their behavior: pretraining data regime, learning objective, and downstream adaptation to clinical endpoints. Second, we introduce a practical capability framing that distinguishes *general-purpose* capabilities (coverage across tissues, scales, stains, scanners, and institutions) from *functional breadth* (task primitives, adaptability, and analysis level), while accounting for modality scope spanning vision, language, and genomics. Third, we synthesize reported results as an *evidence map* rather than a leaderboard, clarifying where capability is supported by reported evidence, where reproducibility is constrained by access or reporting limits, and where further validation is needed. We then analyze current benchmarking practice and identify common confounders, including heterogeneous adaptation protocols and pretraining-evaluation overlap, and propose deployment-aware recommendations built around broad-and-deep benchmarks, standardized adaptation "budgets," overlap auditing, and domain-agnostic evaluation. Finally, we review emerging clinical-utility evidence and argue that foundation models are most compelling when tied to workflow-defined endpoints, calibrated operating points, and measurable operational benefit. We conclude with actionable recommendations for converting capability demonstrations into clinically reliable data-driven diagnostic systems.

1. Introduction

Computational pathology (Perez-Lopez et al., 2024; Song et al., 2023) applies machine learning to digitized histology, typically whole-slide images (WSIs), to support diagnosis, grading, subtyping, prognostic stratification, and biomarker discovery (Bilal et al., 2023a,b; Lu et al., 2024a; Xu et al., 2024). Deep learning, a subset of machine learning, has further accelerated this progress, advancing data-driven diagnostics for cancer detection, triaging and prescreening, tissue-type identification, molecular and genetic biomarker prediction, and the assessment of outcomes and treatment response (Atallah et al., 2023; Bilal et al., 2021c, 2023; Campanella et al., 2019; Graham et al., 2023; Hoang et al., 2024; Javed et al., 2020; Kather et al., 2016, 2019, 2020; Lu et al., 2024a; Saldanha et al., 2023; Wahab et al., 2023). Yet, most of these advances rely on task-specific supervised learning pipeline, where progress depends on access to annotated datasets. This remains a major bottleneck, because high-quality annotations are expensive to generate, often heterogeneous, and only weakly standardized across sites. Consequently, computational pathology has largely advanced through isolated task-dataset development, a paradigm in which clinical variability across scanners, stains, institutions, and patient populations alongside the gigapixel scale, pre-analytic heterogeneity of WSIs, and persistent inter- and intra-observer variability can substantially limit

☆ This article is part of a Special issue entitled: 'Data-Driven Diagnostics' published in Molecular Aspects of Medicine.

* Corresponding authors.
E-mail addresses: m.farooq@psau.edu.sa (M. Bilal), N.M.Rajpoot@warwick.ac.uk (N. Rajpoot).

https://doi.org/10.1016/j.mam.2026.101504
Received 8 April 2026; Received in revised form 23 June 2026; Accepted 31 August 2026
Available online 18 September 2026
0098-2997/© 2026 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

out-of-distribution generalization (the ability of AI algorithm to do well on unseen new data).

Foundation models (Bommasani et al., 2022) are reshaping computational pathology by shifting innovation from training task-specific models (Bilal et al., 2023a; Campanella et al., 2019; Graham et al., 2023; Saillard et al., 2023) toward reusing large pretrained representations that can be efficiently adapted to new endpoints (R.J. Chen et al., 2024; Vorontsov et al., 2024; Zimmermann et al., 2024). Foundation models learn "generic" representation first and solve "specialized" tasks next. Rather than training for each specific task from scratch, a foundation model is pretrained at scale on broad data, often with self-supervised objectives, so it learns reusable visual (and potentially multimodal) structure (Bommasani et al., 2022). This pretraining stage yields a general-purpose representation that can be adapted to multiple specialized downstream tasks with substantially fewer annotations than conventional supervised pipelines. In practice, the strengths are label (annotation) efficiency, transfer across endpoints, and an engineering shift toward reusable backbones with standardized adaptation interfaces. This shift in paradigm has enabled further advance in data-driven diagnostics through cross-modality retrieval (Wang et al., 2023), multimodal question answering (Ahmed et al., 2024; Tran et al., 2025), and report generation (Guo et al., 2024; Tan et al., 2024).

Despite rapid progress, the evidence base remains difficult to interpret for clinical adoption. Reported gains are hard to compare because studies differ in cohort composition, benchmark choice, reporting practice, adaptation strategy, and evaluation framework. As a result, apparent improvements may reflect differences in data regime, adaptation protocol, or experimental design rather than intrinsic superiority of a pretrained representation.

Our central thesis is that pathology foundation models should be judged by clinical utility, robustness and workflow integration, rather than by best-case performance on a single benchmark. Accordingly, this review adopts a decision framework organized around four questions: (1) what capability is demonstrated, (2) under what evaluation regime and operating point, (3) what adaptation is required, and (4) what additional evidence is needed to justify clinical use. To support this goal, we (i) define foundation models and the technical factors that stimulate transferability in pathology AI, (ii) curate and characterize representative pathology foundation models, using capability criteria that later structure our evidence map, and (iii) synthesize benchmarking gaps, clinical utility evidence, and a forward-looking direction for building clinically reliable data-driven diagnostic systems.

A key implication is that our evidence map presents a structured gap-analysis rather than a ranking exercise. Controlled benchmarks answer which model performs best under a fixed protocol; an evidence map answers where the field has generated evidence, where evaluation is sparse, and where claims remain limited by heterogeneous adaptation, incomplete reporting, or pretraining-evaluation overlap.

Scope of the review: We focus on histopathology-centric foundation models, primarily pretrained on WSI-derived tiles, with Hematoxylin and Eosin (H&E) staining as the dominant modality. This includes both vision-only encoders and vision–language aligned models used for representation learning and downstream adaptation. We discuss multimodal extensions (e.g., text, genomics) only where they materially change WSI interpretation or enable new clinical workflows. We do not attempt to produce a universal leaderboard; instead, we synthesize reported evidence and its reproducibility constraints, identify methodological limitations that impede comparability, and identify what additional validation is required for safe deployment.

Prior reviews have provided useful overviews but have not fully addressed key translation questions. Ochi et al. (2025) discussed five vision-only and five image–language aligned models emphasizing potential applications. Chanda et al. (2024) surveyed tools, datasets, and recent advances in large self-supervised models, with limited focus on critical evaluation practices, benchmarking rigor, and clinical utility. We address this gap by centering the review on how pathology foundation models are *evaluated and operationalized*: the dependence of reported gains on evaluation design and adaptation choices; foundational capability (robustness) versus functional breadth; representativeness and skew in pretraining corpora; benchmark design and leakage or overlap risks; and the practical constraints that shape adoption in data-driven diagnostic workflows.

Literature Search: To provide an evidence-grounded and tractable view, we conducted a structured search across PubMed and Google Scholar (for last five years), focused specifically on pathology foundation model reports, screening recent literature and identifying approximately 50 eligible pathology foundation models within scope. We conducted a structured literature review to identify histopathology-centric pathology foundation models relevant to whole-slide image (WSI)-based computational pathology and to synthesize evidence on capability, adaptation, and deployment relevance. The review focused on models with reusable pretrained representations (e.g., self-supervised, vision–language, multimodal, or generative pretraining) and pathology-relevant downstream evaluation.

Search strategy and sources: We searched PubMed (n = 1000 explored records) and Google Scholar (n = 200 explored records) and supplemented this with backward/forward citation chaining of key model, benchmark, and review papers. Search terms were iteratively refined to capture terminology variation across: foundation/pretraining paradigms (e.g., foundation model, pretrained model, self-supervised learning, contrastive learning, masked image modeling, representation learning, ViT); vision–language/multimodal paradigms (e.g., VLM, multimodal, image–text, copilot, generative); and pathology terms (e.g., digital pathology, computational pathology, histopathology, WSI, H&E).

Selection and eligibility: Records were deduplicated and screened at title/abstract and full-text stages using predefined criteria aligned with the review scope. The unit of inclusion was the study/report used for evidence synthesis. A study was included in the pathology foundation model set if it: (i) focused on histopathology/WSI analysis, (ii) described a reusable pretrained model/model family for transfer across tasks/settings, (iii) used a transferable pretraining/representation-learning strategy, (iv) reported pathology-relevant downstream evaluation across multiple tasks and task families, and (v) was materially relevant to histopathology-centric workflows. Records were excluded if outside scope, not foundation model relevant, lacking primary empirical/model content, insufficiently relevant to WSI pathology interpretation, or duplicates after consolidation. For supporting benchmarking/evaluation evidence used in interpretation, only peer-reviewed studies were retained.

Organization: Section 2 defines foundation models and the technical choices that shape their behavior in pathology. Sections 3–5 summarize the evidence map, benchmarking gaps, evaluation practices, and clinical utility considerations, culminating in forward-looking recommendations.

2. Foundation models: Definition and technical foundations

This section establishes a working definition of foundation models and the technical foundations needed to interpret their reported performance in data-driven diagnostics. Foundation-model era can be viewed as a third phase in machine learning: after manual feature engineering (classical ML) and representation learning from labeled data (deep learning), foundation models learn universal representations of unlabeled data through large-scale pretraining, often with self-supervision (Fig. 1a). In Fig. 1a, we introduce the core levers: data regime, learning objective, and adaptation interface that determine whether a pretrained representation reliably transfers to the downstream diagnostics endpoints.

The key shift is that a single pretrained backbone can support many endpoints, reducing the need to retrain entire model from scratch

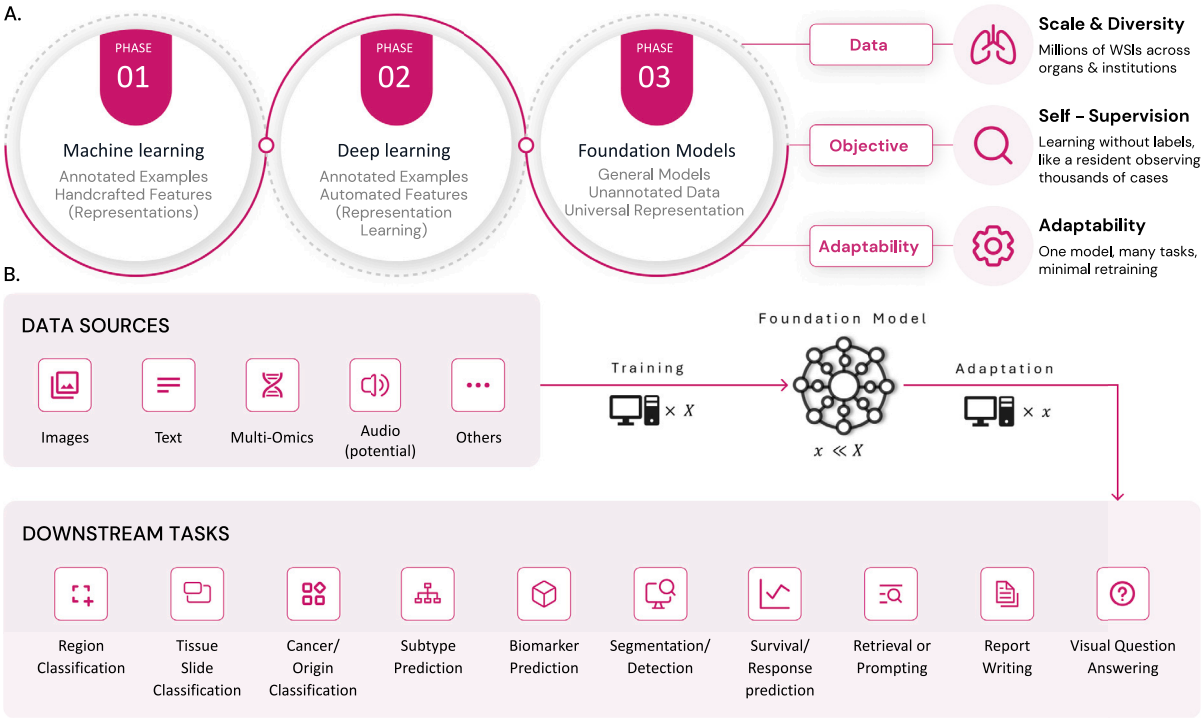


Fig. 1. A. Evolution from machine learning and deep learning to foundation models and their core technical levers. B. Pathology foundation model pipeline linking data, large-scale pretraining, and efficient downstream adaptation for tissue diagnostics. Current pathology-relevant modalities are dominated by images, reports/text, and molecular/omics data; audio modality shown in the schematic should be interpreted as potential inputs rather than routine pathology data sources.

for each task (Bommasani et al., 2022). In pathology, this shift is particularly consequential because labels are costly and heterogeneous, WSIs are gigapixel-scale, and clinical translation depends on robustness to acquisition and population shift. To support broad readership and to avoid conflating pretraining with downstream performance, Fig. 1b presents a pipeline-level schematic of pathology foundation models, connecting data sources and curation, compute-intensive pretraining objectives and compute, and efficient downstream adaptation and evaluation across data-driven diagnostic and prognostic tasks.

In this review, a foundation model refers to a large pretrained model that (i) is trained at scale on broad data, often self-supervised, (ii) provides a general-purpose representation or generative interface, and (iii) can be efficiently adapted to diverse downstream tasks (e.g., via linear probing, fine-tuning, or parameter-efficient adaptation) (Bommasani et al., 2022). The promise stems from learned invariances and semantic structure that reduce label requirements and enable transfer across endpoints.

In computational pathology, foundation models are best understood through a small set of technical factors. Data regime governs what variability the model can absorb (e.g., single-site vs multi-institution; limited vs diverse tissues; single vs multi-scanner; single-scale tiles vs multi-scale context). Learning objective shapes the representation geometry (e.g., contrastive learning/self-distillation Chen et al., 2020; Zhang et al., 2019, masked modeling Xie et al., 2022 and vision–language alignment Radford et al., 2021). Using conventional machine/deep learning approaches, adaptation interface and weak supervision (Bilal et al., 2023a; S. Chen et al., 2024), determine how capabilities are expressed downstream (e.g., frozen encoder plus lightweight head vs end-to-end fine-tuning), and it influences reproducibility, compute requirements, and susceptibility to overfitting under distribution shift.

3. Pathology foundation models

We review pathology foundation models through the lens of deployment-relevant decisions: what capability is demonstrated, what adaptation is required, and what evidence supports robustness under shift. The landscape spans (i) vision-only encoders for WSI representation, (ii) vision–language aligned models that enable retrieval and text-conditioned inference, and (iii) multimodal and copilot-style systems that extend beyond representation learning to interaction and generation. Across these categories, the distinctive strength of foundation models is not merely higher accuracy, its scope: the potential for reusable representations, label-efficient adaptation, and broader functional coverage, provided claims are supported by rigorous evaluation. In this sense, foundation models are increasingly becoming platforms rather than point solutions.

From the identified model set, we curated a representative subset of 40 models for visualization in Fig. 2, while using the broader evidence base throughout the review to construct the evidence map and interpret capability claims. In descriptive summaries, we emphasize the latest representative version of a model family where possible; older versions are retained when they are the actual models evaluated in the source paper, benchmark, or evidence table. Fig. 2 defines the capability taxonomy used throughout the manuscript by organizing models along: (i) modality scope (vision, text, genomics and their combinations), (ii) breadth of the pretraining data regime (multi-institution, multi-tissue, multi-scale/resolution, and multi-scanner), and (iii) downstream evaluation targets across data-driven pathology tasks, structured by task type (classification, segmentation, detection) and analysis level (WSI-level vs cellular-level). These dimensions operationalize what "generality" means in pathology in practice: transfer across acquisition settings, tissues, and clinically relevant task families.

Building on this context, we distinguish two distinctive capability notions. *General-purpose (foundational) capability* reflects breadth and robustness across domains and acquisition variability (e.g., tissue diversity, labs, scanning, staining, and scale variation). *Multipurpose (emerging specialty) capability* reflects functional breadth across task families, from the same representation model can support different task types (classification, segmentation, detection etc.) and analytic

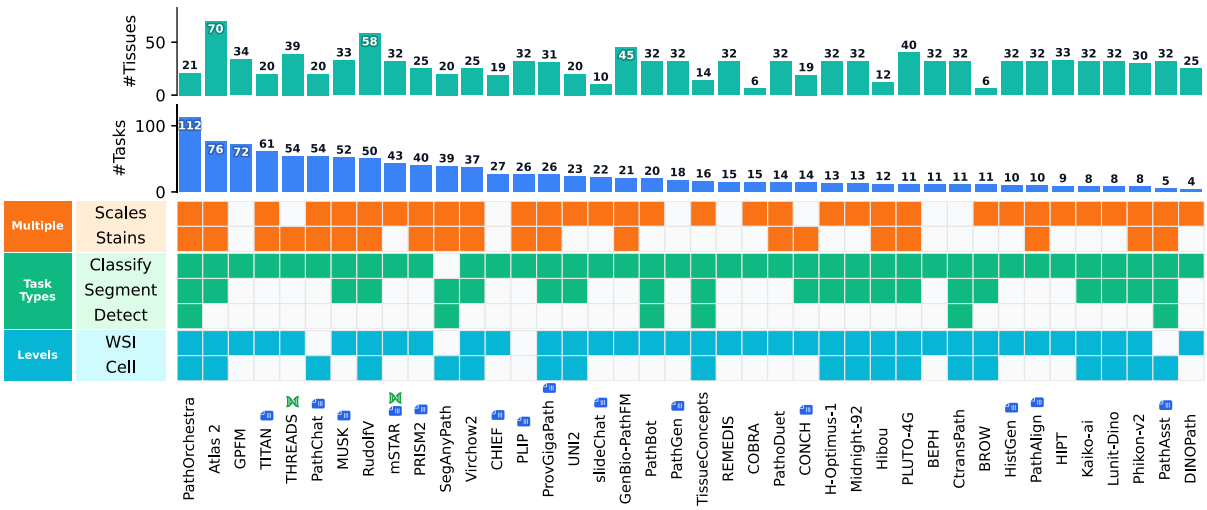


Fig. 2. Capability profile of representative pathology foundation models across modality scope and evidence breadth. All models include vision pretraining; document and DNA icons next to model names indicate models with language- or genomics-associated pretraining, respectively. General-purpose capability summarizes domain breadth across tissues, scales, scanners, and stains, noting that multi-scanner data are common across pretraining cohorts. Functional breadth summarizes task primitives (classification, segmentation, detection), downstream task coverage, and analysis level (cell/patch/WSI).

granularity (e.g., WSI-level and cellular-level tasks). The key idea is this: foundation models separate *learning the world* from *solving the task*. General understanding comes first. Specialization emerges later through adaptation.

3.1. Data scale and diversity

Computational pathology has rapidly scaled pretraining datasets to millions of multi-gigapixel WSIs, alongside a rise in model size to billions of trainable parameters. **Virchow2/2G** (Zimmermann et al., 2024) and **Atlas 2** (Alber et al., 2026) illustrate this shift: Virchow2 was pretrained on 3.1 million WSIs with 1.85 billion parameters, while Atlas 2 scales even further with a 5.5 million WSI corpus and 2 billion parameters. Other efforts reflect related scaling strategies: **PathBot** (Lu et al., 2025) explores billion-parameter ViT backbones with hybrid training recipes, whereas **Kaiko-Midnight** (Karasikov et al., 2025) examines how far strong performance can be achieved with smaller mixed public and proprietary datasets. **GenBio-PathFM** (Kapse et al., 2026) focuses on building a public-data pathology foundation model through automated data curation and a two-stage learning strategy, leading to strong benchmark performance. **PathOrchestra** (Yan et al., 2025) emphasizes broad clinical-task coverage across WSI- and ROI-level pathology applications. Taken together, these studies show that progress in pathology foundation models is being driven not only by scale itself, but by how model size interacts with pretraining data, composition, and learning objective.

Importantly, parameter count is an imperfect proxy for clinical readiness. Pathology models typically target a narrow visual domain (2 billion parameters of Atlas 2 (Alber et al., 2026)) and can therefore achieve strong in-domain results with far fewer parameters than general-purpose multimodal model systems such as GPT-4V (OpenAI, 2023) (estimated > 1 trillion parameters) or PaLM-E (Driess et al., 2023) (> 550 billion parameters) and Med-PaLM 2 (Singhal et al., 2025) (> 500 billion parameters), or more recent large medical multimodal models such as MedGemma 1.5 (Sellergren et al., 2025) (4 and 27 billion parameters). However, parameter efficiency or scale alone does not guarantee robustness under stain, scanner, site, or population shift. Scale should therefore not be conflated with clinical readiness; deployment relevance is established through domain-agnostic evaluation, calibrated operating points, and workflow-aligned decision-threshold reporting.

Fig. 3 highlights the diversity of pretraining corpora from which claimed general-purpose behavior emerges, but also an often-overlooked constraint: these corpora are typically imbalanced across organs and tissue types, suggesting that bias in computational pathology is often data-driven before it is algorithm-driven. No model achieves uniform pan-organ coverage, yet many models converge toward similar representation patterns, likely reflecting shared data availability and acquisition pipelines. Such skew can translate into uneven representation quality, particularly for rare entities or underrepresented tissues, and it limits the validity of broad "generalist" claims without rigorous evaluation. This observation motivates our evidence-map logic in Section 4: we treat breadth of evaluation (tasks and tissues), external validation, and explicit domain-shift testing as first-order signals of demonstrated capability.

Two further trends shape how evidence should be interpreted. First, vision–language and multimodal alignment extend pathology foundation models beyond image encoding to enable report-conditioned retrieval, text conditioning, and prompting; yet clinical utility depends on robust operating points and workflow-defined endpoints rather than generic "multimodal" capability. Second, many published comparisons remain confounded by non-uniform benchmarks, dataset overlap, and heterogeneous adaptation protocols. Accordingly, throughout this review we interpret claims only when the capability, adaptation setting, and evaluation regime are explicitly specified, and we use the evidence map to summarize where such evidence is available.

3.2. Training objectives

Unlike single-task pipelines, pathology foundation models are pretrained on gigapixel WSIs decomposed into tiles/patches, often across multiple magnifications and sometimes across stains, so that pretraining learns reusable morphological representations that can later be adapted to many downstream data-driven diagnostic and prognostic tasks. Most current pathology foundation models employ transformer-centric backbones (ViT/Swin-style variants Dosovitskiy et al., 2021) trained with large-scale self-supervision, with Fig. 4 summarizing representative vision-only models and their pretraining specifications (architecture, scale, and objective). This pretraining stage concentrates compute into a one-time cost and enables label-efficient adaptation; however, it also shifts the central question from "which model scores highest?" to "which pretraining regime yields robust transfer under clinical variability?".

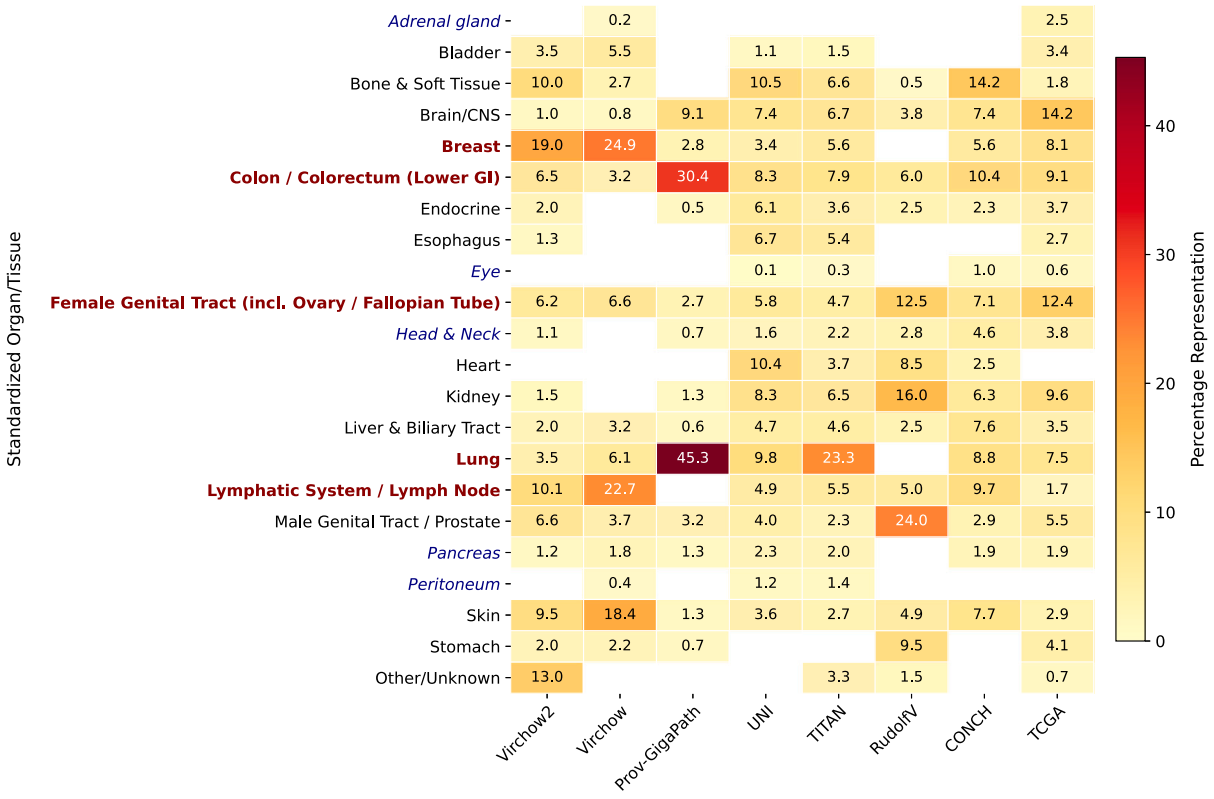


Fig. 3. Tissue/organ coverage highlights representational diversity as the origin of generalization behavior and its skewness as the organ-specific bias in pretraining corpora of lead foundation models.

Across literature, several training strategies recur. For vision-only encoders, three objective families dominate. (1) Contrastive or self-distillation SSL, including DINO/DINOv2 (Oquab et al., 2024) (e.g., **Virchow** Vorontsov et al., 2024/ **Virchow2/2G** Zimmermann et al., 2024, **UNI** R.J. Chen et al., 2024, **Hibou** Nechaev et al., 2024, and **Kaiko-ai** kaiko.ai et al., 2024), MoCo v3 (**CtransPath** Wang et al., 2022), SimCLR-style (Xie et al., 2022) variants (**REMEDIS** Azizi et al., 2023 and PathSSL Lai et al., 2023), and domain-specific intra-/inter-WSI contrast formulations (e.g., **CHIEF** Wang et al., 2024, **mSTAR** Y. Xu et al., 2025), learns invariances by matching different views of the same tissue content and underpins many strong encoders. (2) Masked prediction objectives (e.g., iBOT/BEiT/MAE variants; **Phikon** Filiot et al., 2023; **BEPH** (Yang et al. 2025)) learn to reconstruct missing content and can emphasize global context, but may remain sensitive to stain- and texture-specific statistics. (3) Hybrid recipes mix multiple signals, multi-task supervision (**TissueConcepts** Nicke et al., 2025), supervision injection (e.g., unified knowledge distillation in **GPFM** (Ma et al. 2025)), pathology-oriented segmentation systems (e.g.,**SegAnyPath** Wang et al., 2024), or pixel-/region-level signals when available to incorporate clinically relevant priors.

For vision–language alignment, CLIP-style objectives (Radford et al., 2021) bind patches to captions/reports and enable retrieval and promptable adaptation (e.g., **PLIP** Huang et al., 2023, **CONCH** Lu et al., 2024a and **PathAlign** Ahmed et al., 2024). Because these objective families induce different invariances, comparisons are interpretable only when the objective, data regime, and adaptation protocol are clearly reported.

Training pipelines also explicitly target sources of variability that undermine deployment. For example, multi-stain pretraining expands the stain distribution encountered during representation learning and can improve robustness to stain shift (e.g., **PathoDuet** Hua et al., 2024; **Madeleine** Jaume et al., 2025). When available, multi-institution and multi-scanner pretraining broadens acquisition variability, while multi-scale sampling exposes the representation to both histologic detail and tissue context.

Stain and scanner variation are also addressed through aggressive color augmentation, stain normalization/augmentation (e.g., **BROW** Wu et al., 2023; **Hibou's** RandStainNA Nechaev et al., 2024), and multi-site data (e.g., **Atlas/Atlas 2** Alber et al., 2026, 2025 and **RudolfV** Dippel et al., 2024). Tissue and cohort generalization is pursued through curated multi-organ collections (pathologist-led, e.g., RudolfV) and/or sheer corpus scale (e.g., **PRISM** Shaikovski et al., 2024/**PRISM2** Vorontsov et al., 2025, **Virchow2/2G** Zimmermann et al., 2024, **H-Optimus-1** Bioptimus, 2025, **Phikon-v2** Filiot et al., 2024).

Scale invariance is addressed via mixed-magnification training (e.g., **Virchow2/2G** Zimmermann et al., 2024), dynamic patch extraction (e.g., **Kaiko-ai** kaiko.ai et al., 2024), variable patching (e.g., **PLUTO** Juyal et al., 2024 and **PLUTO-4G** Padigela et al., 2025), hierarchical/multi-scale encoders, and cross-scale alignment (e.g., **HIPT** Chen et al., 2022, **COBRA** Lenz et al., 2025). These design choices align with Fig. 2 capability dimensions (modality, breadth of data regime, and task/level targets) and guide evidence-map interpretation: a claimed "generalist" model is more credible when its pretraining regime plausibly covers the variability it is expected to withstand at deployment.

Patch-level & Slide-level Foundation Models: A practical distinction is whether the foundation model primarily learns representations from local tissue regions or from whole-slide context. Patch-level foundation models encode small tiles or regions extracted from WSIs and learn local morphology features, including nuclei, glands, stroma, necrosis, immune infiltrates, tumor architecture, and stain-specific patterns. Models such as **UNI** (R.J. Chen et al., 2024), **Virchow/Virchow2** (Vorontsov et al., 2024; Zimmermann et al., 2024), **H-Optimus** (Saillard et al., 2024; Bioptimus, 2025), **Phikon** (Filiot et al., 2023), and **CTransPath** (Wang et al., 2022) are commonly used in this way. Their produced tile embeddings can be pooled, probed, fine-tuned, or passed to task-specific heads and WSI-level aggregation modules.

Fig. 4. Representative vision-only pathology foundation models by architecture, parameter scale, pretraining data scale, and training objective. In the radial encoding, the outer ring indicates training paradigm, named ring lists model names, the outer bar track represents parameter count, and the inner bar track represents WSI scale. Bar colors denote model architecture.

Slide-level foundation models, by contrast, aim to produce global WSI representations or decision interfaces for whole-slide analysis. Most current slide-level models still begin from tile or patch features, but add an additional layer of modeling to connect local morphology into a global representation. This may involve multiple-instance learning (MIL), transformer or graph aggregation, hierarchical context modeling, report alignment, molecular alignment, or weakly supervised training from slide-level labels. Examples include **TITAN** (Ding et al., 2025), **PRISM/PRISM2** (Shaikovski et al., 2024; Vorontsov et al., 2025), **Prov-GigaPath** (Xu et al., 2024), and other WSI encoders. This distinction matters because a strong patch encoder is not, by itself, a complete slide-level diagnostic system. Slide-level performance also depends on tissue sampling, aggregation design, adaptation budget, calibration, and external validation.

Fig. 5 summarizes the representative vision and vision–language aligned pipelines through which patch-level representation learning is extended into slide-level pathology foundation models. First, tile-level pretraining yields patch encoders that are subsequently adapted via lightweight heads or WSI aggregation methods for slide-level endpoints, as in MIL-style systems such as **OmniScreen** (Y.K. Wang et al., 2025). Second, vision–language alignment learns paired image–text embeddings to support retrieval, zero-shot transfer, and promptable adaptation (e.g., **PLIP** Huang et al., 2023 and **CONCH** Lu et al., 2024a). Third, some systems incorporate WSI aggregation during training to link patch representations to WSI-level decision making (e.g., **Prov-GigaPath** Xu et al., 2024). More recent slide-level or multimodal foundation models further align WSI representations with pathology reports, clinical text or molecular profiles, as in **TITAN** (Ding et al., 2025), **PRISM2** (Vorontsov et al., 2025), **mSTAR** (Y. Xu et al., 2025) and **THREADS** (Vaidya et al., 2025). Finally, when the intended use includes question answering, report generation, or interactive copilot-style support, contrastive alignment is often combined with generative objectives, as in **PathChat** (Lu et al., 2024b), **slideChat** (Chen et al., 2025), **PRISM2** (Vorontsov et al., 2025), and **PathAsst** (Sun et al., 2024b).

Two practical constraints shape this landscape. First, the most diverse pretraining corpora are often proprietary; when public data are used, TCGA is frequently a dominant source, and it is also commonly reused directly or via overlapping derivatives in downstream evaluations, including proprietary adaptation cohorts. In our review of the

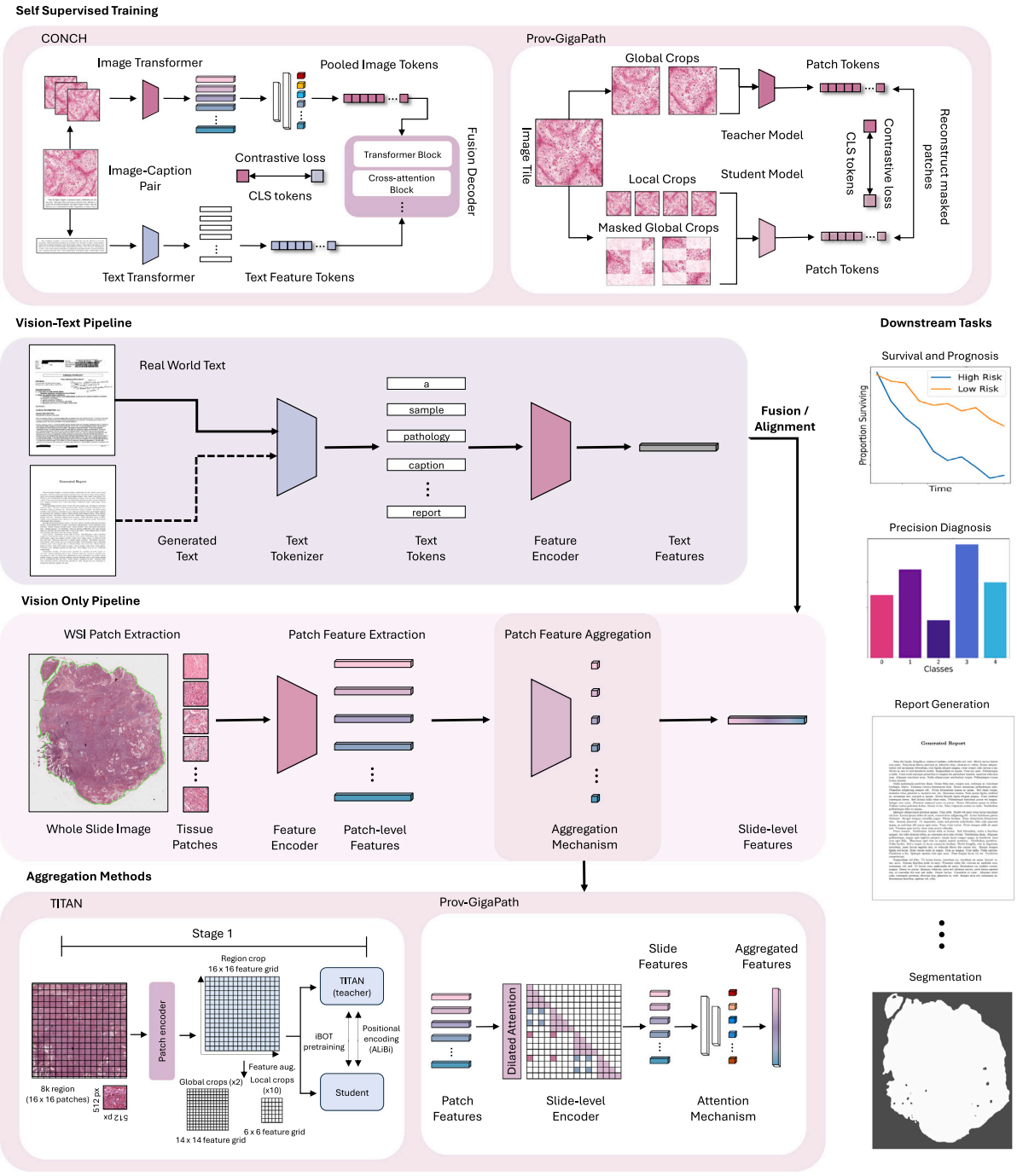


Fig. 5. Pathology Foundation Model Pretraining: Vision–text (CONCH), vision-only (Prov-GigaPath), and patch into WSI aggregation pipelines (TITAN vs Prov-GigaPath).

49-model curated set, 26 models (53%) report TCGA cohorts in pretraining. Across 394 extracted downstream evaluation entries in Supplementary Table S1, 104 (26.4%) are confirmed as TCGA-derived or TCGA-containing, 57 (14.5%) are CPTAC-derived or CPTAC-containing, and 233 (59.1%) use other non-TCGA/non-CPTAC evaluation sources. These counts are therefore conservative, especially when pretraining corpora are proprietary or incompletely disclosed, but they make the leakage risk concrete. This practice can inflate apparent generalization, introduce data leakage or implicit cohort familiarity, and makes it difficult to disentangle the effects of objective choice versus data regime in ablation studies. Second, downstream performance is strongly mediated by adaptation choices (aggregation design, fine-tuning protocol, and operating point), so results reported without explicit adaptation and evaluation specifications are prone to confounding. Overall, pathology foundation model training should be interpreted as a coupled design space: data regime, objective family, and adaptation interface, rather than a single "pretraining recipe".

3.3. Adaptation

A foundation model produces reusable tissue embeddings, but adaptation turns those embeddings into a task-effective pathology system. This step is central for downstream application, because most users will not pretrain a new foundation model; they will start from an off-the-shelf encoder and adapt it to a local clinical or research endpoint.

The simplest route is frozen feature extraction: tiles or regions are embedded once, and a lightweight classifier, regressor, retrieval index, or segmentation head is trained on top. Linear probing is useful for measuring representation quality with limited tuning, but it may underestimate performance when the endpoint requires slide-level context. For WSI-level diagnosis, prognosis, and biomarker prediction, patch embeddings are usually connected to slide labels through MIL pooling, attention pooling, transformer aggregation, graph aggregation, or hierarchical models. Recent work shows that the adaptation interface itself can be redesigned: contextualizers can enrich patch embeddings with slide-level context, mixture-of-aggregator or mixture-of-expert modules can route instances through different pooling strategies or feature-processing branches, and self-configuring MIL pipelines can standardize sampling, optimization, and uncertainty estimation (Belagali et al., 2025; Ozlugedik et al., 2026; Shao et al., 2026; Luo et al., 2025). These mechanisms are the bridge between patch-level encoders and slide-level endpoints, and they can be as important as the pretrained backbone itself.

When more labels or compute are available, adaptation can involve parameter-efficient fine-tuning, partial unfreezing, or full end-to-end fine-tuning. These approaches may improve task performance, but they also increase the risk of overfitting, site-specific calibration, and leakage when pretraining, adaptation, and evaluation cohorts overlap. For vision–language models, adaptation may also involve prompt engineering, text-template design, contrastive retrieval, or instruction tuning, which adds another source of variability. In clinical use cases, adaptation is not complete until the system is calibrated to a prespecified operating point, such as specificity at fixed high sensitivity for prescreening.

For these reasons, adaptation should be reported as a controlled "budget" rather than an implementation detail: number and type of labels, patient-level splits, aggregation design, tuning trials, compute, model-selection criteria, calibration method, operating-point selection, external validation, and uncertainty. Without this information, it is difficult to know whether a reported gain comes from the pretrained representation, the downstream head, the tuning budget, or hidden data overlap (Alowais et al., 2023; Bommasani et al., 2022; Firoozi et al., 2025; J. Liu et al., 2025).

4. Foundation models evaluation

Foundation models are evaluated under a different regime than single-task pipelines: performance is an interaction between (i) the pretrained representation, (ii) the adaptation interface and hyperparameters, and (iii) the evaluation design (cohort, split, operating point, and shift tests). Accordingly, evaluation can be viewed as *intrinsic* (measuring representation quality independent of any specific task) versus *extrinsic* (task performance after adaptation), but in both cases conclusions are strongly mediated by evaluation design and dataset construction, including split integrity and overlaps (Bommasani et al., 2022; Tamkin et al., 2021).

Importantly, adaptation is not a minor implementation detail. The same pretrained encoder can appear state-of-the-art or mediocre depending on tuning budget, aggregation choice, and thresholding policy. Therefore, evaluation should treat the adapted system as the unit of analysis, while explicitly separating what is attributable to the pretrained representation versus the downstream adaptation protocol.

A rigorous evaluation should therefore report not only outcome metrics, but also the conditions that determine comparability and deployment relevance: patient-level splits and overlap checks, independent, domain-shift agnostic validation (stain/scanner/site), calibration and operating-point reporting (e.g., optimal cutoff point), adaptation cost (compute, tuning, supervision requirements), and, where appropriate, decision-analytic/clinical-utility evaluation (Van Calster et al., 2025). Reporting model access constraints and method-selection cost is also essential because they affect reproducibility and independent validation. Where possible, uncertainty estimates (confidence intervals and variability across seeds) should be reported, as ranking differences among strong models are often within statistical noise.

Because published studies vary widely in cohorts, tuning/adaptation budgets, and evaluation design, headline metrics are not directly comparable without additional context. The reporting checklist is provided in Fig. 6 to support reproducible adaptation, credible generalization assessment, and workflow-relevant interpretation. To illustrate its use, Table 1 applies the checklist to **TITAN** (Ding et al., 2025) and **CTransPath** (Wang et al., 2022). The example uses an explicit reporting-evidence rubric: 2 = clearly reported or demonstrated, 1 = partially reported or demonstrated only for a subset of settings, and 0 = not reported or not demonstrated in the public source study or official model resource. The resulting totals (18/32 for TITAN and 11/32 for CTransPath) do not rank representation quality or clinical performance; they indicate how completely the available public evidence supports interpretation of leakage control, robustness, calibration, reproducibility, and deployment assumptions.

The worked example shows why the checklist is useful even when models are scientifically strong. TITAN demonstrates good practice through explicit benchmark exclusions, site-preserved splitting, external testing, detailed tiling/preprocessing reporting, and calibration assessment, but the checklist still exposes deployment gaps in fixed operating-point policy, threshold transport, drift monitoring, and prospective validation. CTransPath demonstrates good practice through transparent SSL training, pretraining source reporting, patch-generation details, computational requirements, and public code/weights, but the same checklist highlights the unresolved TCGA pretraining-evaluation overlap risk and missing calibration, operating-threshold, and workflow-utility evidence.

4.1. Evaluating pathology foundation models for data-driven diagnostics

Most studies evaluate pathology foundation models using extrinsic benchmarks, adapting the representation to downstream tasks and reporting standard metrics. For decision-making, we recommend an evaluation ladder: (1) intrinsic checks (retrieval or prompting clustering, and zero-shot evaluation), (2) extrinsic in-distribution and out-distribution (with respect to pretraining corpora) task performance with a standardized adaptation protocol, (3) external (with respect to adaptation cohort) validation on disjoint sites/cohorts, (4) explicit stress tests for stain/scanner/site/population shift and subgroup performance, and (5) diagnostic/prognostic workflow-defined endpoints and operating points. In pathology, intrinsic checks should be concretely specified (e.g., cross-site retrieval consistency, stain-robust nearest neighbors, and tissue-type clustering without labels)

Vision-only encoders are usually assessed by adapting patch embeddings to WSI-level endpoints (diagnosis/subtyping, biomarkers, survival) and pixel/object-level endpoints (segmentation/detection) using task-specific heads and aggregators. While families such as DINO/DINOv2 often improve in-distribution performance, comparability is routinely undermined by cohort reuse between pretraining and evaluation, heterogeneous adaptation choices (aggregation, tuning depth, hyperparameter budgets), inconsistent task definitions/reporting, and best-case metrics reported without explicit operating points or decision policies.

Vision–language models (e.g., **PLIP** Huang et al., 2023**, CONCH** Lu et al., 2024a) extend evaluation to retrieval, zero-shot transfer, and text-conditioned inference. Because their value depends on image–text alignment, evaluation should pair task metrics with alignment diagnostics (retrieval accuracy, prompt sensitivity, and failure modes) and must rigorously control for overlap, especially TCGA-derived images/captions, using patient-disjoint splits and external cohorts.

Multimodal systems integrate WSIs with reports, transcriptomics, or genomics to model morphology-molecular correspondence. Their evaluations are more confounded by multi-stage fine-tuning and modality-specific preprocessing; studies should therefore document data sources

A. Data sourcing and leakage control
- Pretraining cohort description: institutions, tissues, scanners, stains, magnifications, and time window.
- Overlap + leakage control: disclose any overlap with downstream benchmarks and how leakage is prevented.
- Split integrity: patient-level separation; preserve slide linkage; verify no overlap across pretraining, tuning, and test sets (patient/slide IDs + site/scanner metadata).

B. Training and preprocessing
- Objective + transforms: state pretraining objective and key augmentation/normalization choices (including stain strategy).
- Tiling + QC: report tissue detection, artifact filtering, tiling strategy (size/stride), and magnification sampling.

C. Adaptation or finetuning protocol
- Adaptation budget: frozen vs PEFT vs fine-tuned; MIL/aggregation design (if used); label counts; hyperparameters; compute budget.
- Model selection transparency: validation design, search space, number of trials, early stopping, random seeds; report method-selection cost.

D. Evaluation regime and robustness
- Generalization regime: distinguish in-distribution vs external cohorts with respect to pretraining as well as adaptation/finetuning; report site/scanner diversity; run explicit domain-shift tests.
- Subgroups + failures: performance by tissue/site/scanner and clinically relevant strata; document common error modes.
- Use-case metrics: metrics at operational thresholds (e.g., optimal cutoff point or specificity at fixed high sensitivity) with uncertainty and confidence intervals.

E. Calibration and deployment assumptions
- Calibration + thresholding: calibration method and threshold selection procedure (and calibration metrics where applicable).
- Operating-point policy: state whether thresholds are transferred across sites/scanners or refit per site; declare drift-monitoring assumptions.

F. Reproducibility and access
- Reproducibility package: code/configs, seeds, and sufficient details to rerun adaptation and evaluation.
- Access constraints: weights availability, licensing, compute requirements; provide embeddings and/or an evaluation harness when possible.

G. Workflow endpoint alignment
- Clinical intent: define clinical objective, intended users, and the workflow decision point.
- Utility evidence: prospective or silent-trial evaluation when claiming utility (or a strong retrospective workflow simulation tied to the same operating point).

Fig. 6. Recommended reporting checklist summarizes seven decision-critical reporting components (A–G) for pathology foundation model studies. The accompanying checklist blocks specify minimum items required to reduce confounding (including pretraining-benchmark overlap), enable comparability under standardized adaptation budgets, and support deployment-relevant interpretation (shift robustness, operating points, and calibration).

Table 1
Worked example applying the reporting checklist in Fig. 6. Scores assess reporting completeness and deployment-evidence support, not intrinsic model quality or task performance.

Checklist domain	TITAN reporting evidence	CTransPath reporting evidence
Data sourcing and training objectives (6)	**4/6**: reports multimodal WSI/report/caption pretraining and explicit exclusion of TCGA, PAIP, CPTAC, and PANDA from pretraining; full corpus composition and reuse constraints remain only partially reproducible.	**4/6**: reports self-supervised objective, pretraining source datasets, patch generation, and compute; inclusion of TCGA in pretraining creates downstream overlap concerns.
Split integrity and overlap auditing (6)	**4/6**: reports benchmark exclusions, site-preserved TCGA splitting, and external CPTAC/DHMC testing; residual uncertainty remains for complete corpus-level overlap auditing.	**1/6**: downstream WSI evaluations include TCGA cohorts despite TCGA-containing pretraining, without documented case-level exclusion or overlap audit.
Adaptation or finetuning protocol (4)	**3/4**: tiling, preprocessing, and downstream evaluation setup are reported, but threshold-selection and full tuning-cost transportability are incomplete.	**3/4**: reports patch-generation procedure, downstream aggregation approach, and computational requirements, but adaptation budgets are not standardized across endpoints.
Validation and robustness evidence (6)	**3/6**: includes multi-task evaluation with external CPTAC/DHMC testing; scanner/site stress testing and prospective or silent-trial validation remain limited.	**1/6**: reports downstream evaluations, but external validation, subgroup/shift testing, and robustness stress tests are limited.
Calibration and deployment assumptions (4)	**2/4**: reports expected calibration error, but no fixed clinical operating-point policy or threshold-transport strategy across sites/scanners.	**0/4**: calibration, prespecified operating thresholds, and clinical decision policies are not reported.
Reproducibility and access (2)	**1/2**: public resources support reuse, but full independent reproduction remains constrained by data and pipeline access.	**2/2**: code and weights are publicly released, supporting independent reuse.
Workflow endpoint alignment (4)	**1/4**: clinically relevant tasks are demonstrated, but drift monitoring, failure escalation, and prospective workflow utility are not established.	**0/4**: prospective workflow utility, drift monitoring, and deployment governance are not reported.
Total	**18/32**	**11/32**

and split integrity, enforce patient-level disjointness across modalities, and provide ablations that separate foundation model contribution from fusion design and downstream supervision.

Copilot-style systems ((PathChat Lu et al., 2024b, SlideChat Chen et al., 2025, PathAsst Sun et al., 2024b, PathGen Sun et al., 2025)) require evaluation beyond static accuracy: instruction-following, grounded multi-turn interaction, abstention/refusal behavior, and faithfulness (avoiding hallucinations). Existing benchmarks (PathQABench, SlideBench Meseguer et al., 2025, PathVQA He et al., 2021/PathMMU Sun et al., 2024a variants) are a starting point, but automated text metrics are poor proxies for clinical correctness (Ray, 2023) and risk; retrieval-grounded protocols that force explicit evidence citation from the WSI are more informative, ideally paired with curated QA sets and, when feasible, prospective or silent-trial evaluations aligned with workflow endpoints (Zhong et al., 2024).

Selected task	Metric	Virchow2	H-Optimus-1	UNI2-h	Prov-GigaPath	TITAN	PRISM2	Atlas-2	RudolfV	Hibou	PLUTO-4	Primary source(s)
CRC-100k	bAcc	96.7	95.6	**96.9**	95.9			**97.2**		95.8	96.8	Atlas-2; Pluto-4
PCAM	bAcc	93.9		95.1	94.6			**96.1**		94.3	**95.2**	Atlas-2; Pluto-4
BreakHIS	bAcc	82.1		**86.3**	84.5			**88.3**		79.8	81.8	Atlas-2; Pluto-4
BACH	bAcc	88.7		**92.2**	79.2			91.7		81.9	**93.8**	Atlas-2; Pluto-4
MHIST	bAcc	86.5	83.5	82.6	83.3			**88.5**		79.9	**87.9**	Atlas-2; Pluto-4
TCGA-TILS	Acc				0.939				**0.943**	**0.942**		Hibou
MSI-CRC (patch)	Acc	0.719		0.722	0.771			**0.78**	0.755	**0.793**		Hibou, Atlas-2
MSI-STAD (patch)	Acc	0.725		0.731	0.784			0.76	**0.788**	**0.829**		Hibou; Atlas-2
Gleason (Arvaniti)	bAcc	78.5		77.6	72.7			**80.7**			**79.3**	Atlas-2; Pluto-4
MSI-GC / TCGA-STAD	AUC	**0.923**	**0.915**	0.903	0.863							H-Optimus-1
KRAS-CRC / CPTAC-COAD	AUC	**0.706**	0.625	**0.649**	0.581							H-Optimus-1
TCGA Mean (7 biomarker tasks)	AUC	0.724			0.723	**0.781**	**0.784**					PRISM2
HEST-Average	Pearson r	0.403	0.422	0.43	0.405			**0.448**		0.375	**0.432**	Atlas-2; Pluto-4
IHC ER/PR assessment	AUC	**0.808**		**0.794**	0.775							UNI2-h
TCGA Mean (3 diagnostic tasks)	AUC	0.968			0.954	**0.976**	**0.975**					PRISM2
CAMELYON16 test	AUC	0.976	**0.996**	**0.996**	**0.985**							H-Optimus-1
CAMELYON16 (0.25 MPP)	bAcc	**85.9**		**85.9**	81			**86.8**		80.3	76.6	Atlas-2
PANDA-Small (test)	bAcc	66.4		66.2	65.3			**68.3**		63.2	**66.8**	Atlas-2; Pluto-4
EBRAINS (30-class)	bAcc	0.674		**0.711**	0.68	**0.735**						TITAN; UNI2-h
TCGA-UT (32-class tumor subtype)	Acc	0.83	**0.851**	**0.831**	0.804							H-Optimus-1
TCGA-UT-8K (32-class tumor subtype)	bAcc	**0.774**		0.675	0.7	**0.832**						TITAN; UNI2-h
TCGA-OT (46-class OncoTree)	bAcc	**0.643**			0.543	**0.704**						TITAN
TCGA DSS Mean (n=9,041)	C-index	**0.714**			0.677	0.692	**0.744**					PRISM2
MoNuSAC	Dice	66.9		64.5	68			**69.4**		64.9	**70.4**	Atlas-2; Pluto-4
CoNSep	Dice	64.5		63.6	63			**66.1**		63.3	**65**	Atlas-2; Pluto-4

Fig. 7. Selected source-reported comparisons of leading pathology foundation models on public benchmark tasks, including morphology, biomarker, prognosis, segmentation, and HEST morphology-molecular association results where available. Values are shown only where at least two models were evaluated on the same task with a comparable reported metric, but these results should not be interpreted as a fully controlled benchmark because adaptation protocols, pretraining cohorts, and reporting practices may differ across sources. Metrics include balanced accuracy (bAcc), top-1 accuracy (Acc), area under the curve (AUC), Pearson correlation coefficient (r), concordance index (c-index), and Dice similarity score. Green and blue cells show the best and second-best reported values within each row.

4.2. Evidence map and benchmarking gaps

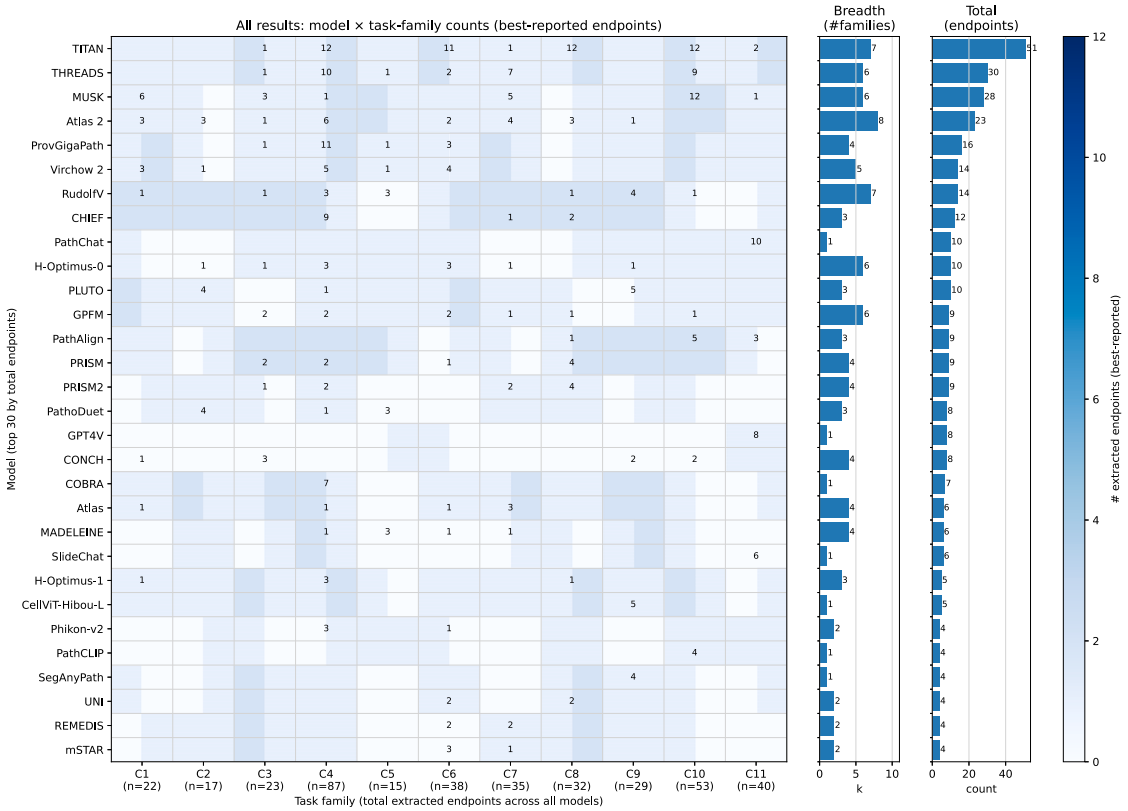
Fig. 7 summarizes selected source-reported results under partially comparable settings, focusing on tasks where at least two listed foundation models were evaluated on the same public benchmark with a comparable metric. Blank cells indicate unreported or non-comparable evidence in the reviewed sources, not evidence that a model is incapable of performing the task. Model development is continuously improving; recent models such as **Atlas-2** (Alber et al., 2026) and **TITAN** (Ding et al., 2025) appear as leading reported models in several rows, whereas other models may be absent because they were not evaluated in the same source-reported setting.

Across models, the bottleneck is methodological comparability. Because studies differ in pretraining cohorts, splits, preprocessing, adaptation protocols, and reporting conventions, and because common pretraining sources are frequently reused in downstream evaluations, many apparent model-to-model comparisons are not methodologically valid. We therefore summarize the landscape using an evidence map that emphasizes breadth, external validation, and robustness signals rather than a single leaderboard ordering. The evidence map is a structured gap-analysis of reported evidence: it shows where capability claims have been evaluated, where evidence is sparse, and where claims remain limited by heterogeneous adaptation, incomplete openness, or possible pretraining-evaluation overlap. It visualizes the current evidence gap; it does not resolve that gap or provide definitive model rankings. It does not replace controlled benchmark evaluations; rather, it complements them by showing the breadth and unevenness of the wider evidence base. The evidence map therefore plays an analytical role when interpreted alongside controlled benchmarks, showing where source-reported strengths do, or do not, persist under standardized evaluation.

Appendix (Supplementary Table S1) compiles 391 extracted results as task× metric entries and records the best-performing model(s) per endpoint according to the primary metric used in the source paper. Entries are annotated with task family, evaluation level (cell/patch/WSI), and the primary metric used (e.g., AUROC, balanced accuracy (bAcc), F1, Dice or panoptic measures (e.g., PQ/AJI); c-index for survival; and Recall@K or Acc@K retrieval; and ROUGE-L and METEOR for report generation). This structure enables a consistent view of what has been demonstrated, while keeping the limitations of best-reported extraction explicit.

Fig. 8 summarizes the evidence landscape across the 11 task families (C1-C11; Legend – Category Key, Fig. 8(c)) derived from Appendix (Supplementary Table S1). The evidence map in Fig. 8 should be read as a summary of reported capability and evidence density, not as a universal ranking, because endpoints, datasets, and adaptation protocols vary widely. Concentration of evidence (Fig. 8(a)) in a family indicates where the community has produced more reported results; sparse regions and blank cells are themselves findings, because they reveal where standardized benchmarking, external validation, or molecular and workflow-relevant evaluation remain missing. They should be interpreted as unreported or non-comparable evidence, not as model failure. We treat external validation and explicit domain-agnostic testing as higher-evidentiary signals than single-cohort improvements, because deployment failures most often emerge under acquisition or population shift.

To mitigate bias from task families that contain many extracted endpoints, Fig. 8(a) reports presence-based family coverage: a model is credited for a task family if it appears in at least one extracted best-reported entry within that family. The model-by-family heatmap provides an overview of all extracted results from Appendix (Supplementary Table S1) by summarizing the distribution of reported task × metric endpoints across models and task families. This view highlights where evidence is concentrated and where it is sparse, without implying overall dominance or universal superiority.



(a) Evidence breadth heatmap across task families.

C1	**RudolfV** CRC100k – 97,3 (Top-1 Acc)	**Atlas 2** CRC100k – 0,972 (bAcc)	**H-Optimus-1** CAM17-WILDS – 99,1 (Top-1 Acc)
C2	**PathoDuet** Cross-scale positioning – 0,95 (F1)	**Atlas 2** HEST – 0,448 (Pearson)	**Virchow 2** TCGA TILs – 0,931 (bAcc)
C3	**CONCH** BRCA/NSCLC/RCC – 0,914 (bAcc)	**GPFM** BRCA/NSCLC/RCC – 0,977 (AUC)	**PRISM2** MSK Breast – 0,980 (AUC)
C4	**TITAN** IDH mutation, LGG – 0,919 (bAcc)	**GPFM** IDH mutation, glioma – 0,998 (AUC)	**COBRA** MSI-H/MSS, CPTAC-COAD – 0,849 (bAcc)
C5	**MADELEINE** KI67 (AIDPATH) – 0,913 (AUC)	**ProvGigaPath** IHC ER/PR – 0,821 (QWK)	**Virchow 2** IHC ER/PR – 0,808 (bAcc)
C6	**ProvGigaPath** PANDA, Gleason – 0,959 (QWK)	**H-Optimus-0** LN (Camelyon16) – 1,000 (AUC)	**UNI2-g-preview** LN (WILDS) – 0,985 (bAcc)
C7	**PRISM2** RFS (MSK CRC) – 0,809 (c-index)	**DINOPath** DSS (STAD) – 0,797 (c-index)	**THREADS** Bevacizumab response – 0,863 (AUC)
C8	**H-Optimus-1** Pan-cancer-TCGA – 85,1 (Top-1 Acc)	**GPFM** Pan-cancer-TCGA – 0,999 (AUC)	**PathAlign** Biopsy vs resection – 0,942 (bAcc)
C9	**PLUTO** PanNuke – 0,671 (bPQ)	**PLUTO** Nucleus segmentation – 0,585 (AJI)	**SegAnyPath** Segmentation (BCSS) – 0,939 (Dice)
C10	**PathAlign** Image-to-text retrieval – 91,3 (Top-3 Acc)	**TITAN** Report-to-WSI retrieval – 0,900 (Mean R)	**THREADS** Molecular prompting – 0,984 (AUC)
C11	**HistoGPT-L-ER** Multi-WSI report – 0,750 (JI)	**PMPRG** Multi-organ WSI report – 0,683 (METEOR)	**GPT4V** PathQABench – 100,0 (Acc)

(b) Per-family highlights (main panel).

Legend – Category Key	
Key	**Task family**
C1	Patch- H&E images
C2	Patch- oncology/molecular/IHC
C3	Predicting subtypes
C4	Biomarker-H&E
C5	Biomarkers-other stains
C6	WSI classification
C7	Survival/response
C8	Pan-cancer/Origin
C9	Segmentation/Detection
C10	Retrieval/prompting
C11	Generative- report, VQA, co-pilot

(c) Category key.

Fig. 8. Evidence map of pathology foundation model performance across task families (Appendix, Supplementary Table S1). (a) Evidence breadth: heatmap of extracted task×metric endpoints per model and task family, including morphology-molecular association endpoints such as HEST where reported in source studies. A model is credited for a task family if it appears in at least one extracted best-reported endpoint. (b) Per-family highlights: selected representative results shown as compact cards. (c) Task-family category key.

Importantly, the heatmap in Fig. 8(a) reveals distinct *evidence profiles*. For example, **TITAN** (Ding et al., 2025) exhibits a high density of reported endpoints (51) and broad categorical coverage (7 of 11 task families), reflecting repeated evaluation across multiple endpoints within selected families. This profile is driven less by scale alone than by its distinctive WSI self-supervision paired with language alignment, a slide-level representation, and breadth-first evaluation enabling many reportable endpoints across families. In contrast, **Atlas 2** (Alber et al., 2026) demonstrates wider *functional breadth* (8/11 task families), but fewer endpoints (23), indicating broader task coverage with less depth

per family; this breadth-heavy evidence is consistent with a vision-only SSL patch encoder pretrained on pathologist-curated, multi-institution, multi-stain largest WSI corpora, yielding robust visual primitives that transfer across tasks, often via downstream aggregators. Together, these patterns show that endpoint density and categorical breadth capture complementary aspects of demonstrated capability and should not be conflated. Fig. 8(b) then provides compact, per-family summary cards reporting three peak state-of-the-art results per task family (model, endpoint, metric, and peak performance), offering a focused view of best-reported outcomes within the broader evidence context of Fig. 8(a).

Supplementary Table S1 also indicates substantial within-family variability, including differences in cohorts, label definitions, and adaptation settings. In several families, zero-/one-shot or limited-adaptation results are materially lower than fully fine-tuned settings; conversely, task-specific adaptation and fine-tuning can improve headline metrics while reducing comparability when protocols differ. These observations reinforce that evidence quality depends on dataset origin, split integrity and transparent adaptation reporting. These findings motivate a benchmarking that is *broad (task coverage)*, and *deep (standardized splits, overlap-audited partitions, and domain-agnostic operating-point reporting)*. Benchmark suites should also disclose overlap with common pretraining corpora to reduce leakage and inflated generalization.

Taken together, Fig. 8 (evidence breadth) and Fig. 2 (capability scope) suggest a consistent pattern: models with pathology-aligned pretraining and well-specified adaptation interfaces tend to surface across more task families, highlighting that impact is not scale alone, but how pretraining is matched to stain/scanner variability and weakly supervised endpoints. In the current evidence map, **Atlas 2** (Alber et al., 2026) shows the widest family coverage (8/11), followed by **TITAN** (Ding et al., 2025) and **RudolfV** (Dippel et al., 2024) (each 7/11), then **H-Optimus-0** (Saillard et al., 2024), **THREADS** (Vaidya et al., 2025), **GPFM** (Ma et al., 2025), and **MUSK** (Xiang et al., 2025) (each 6/11), and **Virchow2** (Zimmermann et al., 2024) (5/11).

Mapping these models back to Fig. 2 helps explain why they appear broadly: they combine general-purpose capability (multi-tissue and/or multi-institution exposure, with some coverage of stain/scale variability) with functional breadth (evidence across multiple task families and analysis levels). This mapping also clarifies limitations: breadth of reported endpoints is less persuasive when training-evaluation overlap exists or when domain-shift testing is absent. Accordingly, we treat "breadth without robustness" as a key gap for translation and prioritize it in the benchmarking agenda. Broad task coverage should not be interpreted as clinical maturity unless it is paired with external validation, overlap auditing, adaptation transparency, calibration, and clinically meaningful operating points.

4.3. Multitask benchmarks and their findings

Benchmarking studies evaluate pathology foundation models under more controlled conditions than original model papers by using shared datasets, standardized protocols, and consistent reporting. These studies are essential for disentangling model quality from reporting and tuning differences, and they provide the most credible basis for practical model selection. They also clarify why the evidence map should not be read as a leaderboard: source-paper results and controlled benchmark results answer different questions.

Three recurring conclusions emerge across recent benchmark efforts. First, data diversity and pathology-relevant pretraining often matter more than architecture alone. Neidlinger et al. (2026) benchmarked ten models, including **CONCH** (Lu et al., 2024a), **Virchow** (Vorontsov et al., 2024), **UNI** (R.J. Chen et al., 2024), and **CTransPath** (Wang et al., 2022), on weakly supervised tasks spanning biomarkers, morphology, and prognosis. Performance varied strongly by endpoint, and ensemble results suggested that representation diversity and task-dataset alignment can matter as much as model scale. These findings are consistent with the evidence map: high scores in one task family do not imply uniform capability across tissues or endpoints. A direct example illustrates this point. In the original **Virchow** study, averaging displayed AUROCs across 16 pan-cancer detection categories and nine biomarker tasks ranked Virchow first among the reported models, with mean AUROC 0.954 for pan-cancer detection, 0.883 for biomarkers, and 0.929 across the combined 25 results. In Neidlinger et al.'s standardized benchmark, restricted to the same four encoders available in both comparisons (**Virchow**, **UNI**, **Phikon**, and **CTransPath**) and to the corresponding morphology and biomarker task categories, **UNI** ranked above **Virchow** in morphology tasks (0.734 vs. 0.730), biomarker tasks (0.712 vs. 0.684), and the combined 24-task analysis (0.716 vs. 0.694). These are not identical task replications, and the absolute values should not be directly compared because tasks, cohorts, and downstream pipelines differ. The change in relative ordering is nevertheless informative: source-study superiority may not persist under independent standardized evaluation.

Second, rankings are sensitive to adaptation and aggregation choices. Lee et al. (2025) evaluated models in consistency and flexibility scenarios and reported that relative performance changed with few-shot, PEFT, and fine-tuning settings. Other embedding aggregation benchmarks similarly show that histology-specific encoders often outperform ImageNet feature extractors, but the best aggregation strategy remains task- and dataset-dependent. Benchmark suites should therefore disclose and, where possible, standardize the adaptation budget: labels, tuning trials, aggregation defaults, compute, model-selection rules, and calibration.

Third, external validation, robustness, and molecular grounding remain underrepresented. Benchmarks such as HEST (Jaume et al., 2024) are especially valuable for molecular pathology because they test whether foundation model representations encode biologically relevant information beyond conventional diagnostic labels. HEST-1k pairs H&E WSIs with spatial transcriptomics, and the public HEST-Benchmark evaluates gene-expression prediction from matched image regions. In this public benchmark, **H-Optimus-1** ranks first (Pearson $r = 0.4229$) using ridge regression with PCA (256 factors). HEST should still be interpreted as a morphology-molecular association benchmark under a specified protocol, not as direct evidence of clinical utility or biomarker replacement. Focused benchmarks such as ovarian subtyping (Breen et al., 2025), MIL-based biomarker and grading studies (H. Xu et al., 2025), and clinical benchmarks (Campanella et al., 2025a) add useful disease- and workflow-specific evidence, but they also show that results remain sensitive to cohort composition, preprocessing, and external validation.

The same synthesis points recur in specialized benchmark settings. First, SSL-focused benchmarks reinforce the importance of data scale and pathology-relevant pretraining: Giga-SSL (Lazard et al., 2023) reported that SSL increases the number of mutations predictable from WSIs versus supervised baselines, DPathSSL (Kang et al., 2023) showed that SSL with linear probing consistently outperforms ImageNet-pretrained models in classification and nuclei segmentation, and CPathatScale (Campanella et al., 2024) reported strong but endpoint-dependent gains from large-scale DINO-style pretraining. Second, disease-specific evaluations show that the benefit of foundation model representations depends on the endpoint and preprocessing regime. For example, Breen et al. (2025) found that histology foundation models generally outperform ImageNet counterparts for ovarian cancer subtyping, while performance still varied across extractors, preprocessing choices, and disease context (Fig. 9).

Benchmarks beyond classification support the same conclusion: model ranking is inseparable from task formulation and adaptation design. F-SEG (Gildenblat and Hadar, 2024) evaluated unsupervised factorization for H&E segmentation with **Prov-GigaPath** (Xu et al., 2024) and **UNI** (R.J. Chen et al., 2024); VLSA (P. Liu et al., 2025) combined **CONCH** (Lu et al., 2024a) with prognostic prompts for survival prediction; Mallya et al. (2025) used **CTransPath** (Wang et al.,

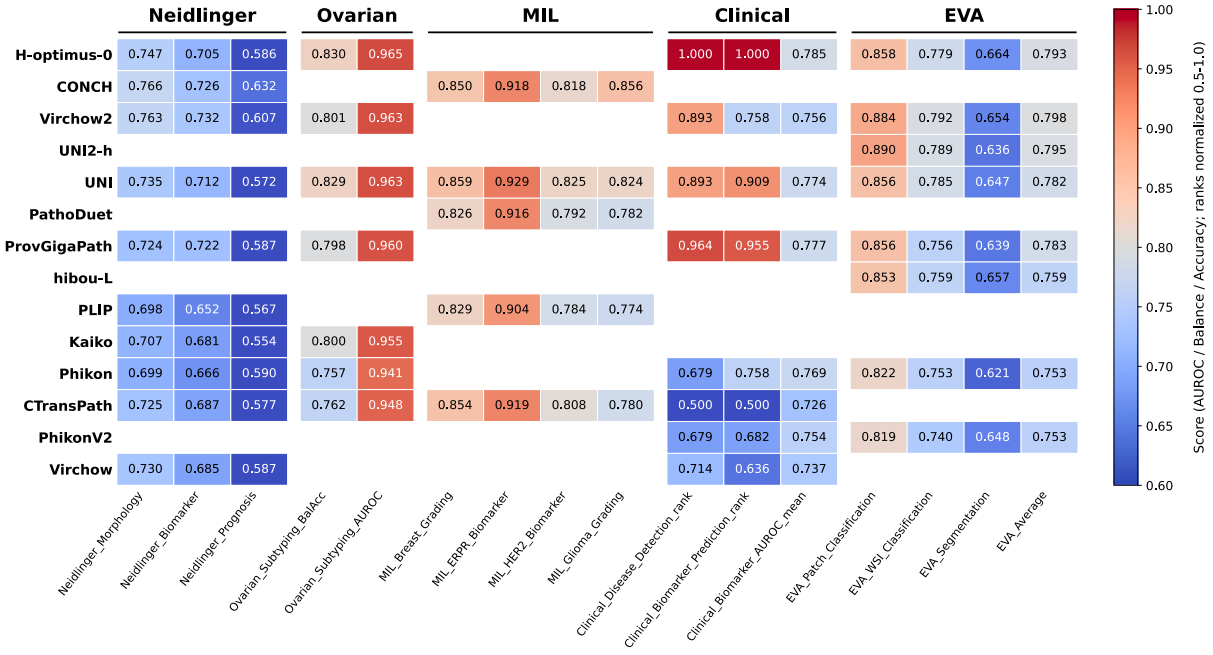


Fig. 9. AUROC, balanced accuracy, or rank-based scores as reported across selected benchmark suites (Neidlinger Neidlinger et al., 2026, ovarian Breen et al., 2025, MIL H. Xu et al., 2025, clinical Campanella et al., 2025a, and EVA). Blank cells indicate models not evaluated for a given endpoint. Normalization supports within-figure comparison but does not remove differences in cohort composition, endpoint definition, or adaptation protocol across suites.

2022) with CLAM-based MIL for bevacizumab response prediction; and X. Wang et al. (2025) proposed an integrated foundation model pipeline for GI cancer prognosis and chemotherapy-benefit prediction. These examples are best read as evidence that task-specific heads, aggregation choices, and clinical endpoint definitions remain central to benchmark interpretation, not as a simple list of model winners.

The third conclusion is clearest in robustness studies, which remain comparatively rare. PathCLIP robustness (Zheng et al., 2025) showed sensitivity to common corruptions such as brightness and blur, underscoring the need for corruption-aware validation. PathoROB (Kömen et al., 2026) proposes a robustness index tied to downstream failure modes rather than accuracy alone, treating "medical center" as a confounding label to test whether foundation models prioritize biological signal over center effects. These studies show why high mean performance is insufficient without explicit stress testing and external validation.

On the Eva[1] leaderboard (adaptation/evaluation for vision-only models), **Virchow2** leads with average balanced accuracy 0.8204 across 13 tasks, narrowly ahead of UNI2-h (0.8201), followed by H-Optimus-0 (0.8037). Community resources such as HistoBoard (Filiot, 2026) provide another useful view of model performance across reported benchmarks. Such live leaderboards accelerate iteration via standardized or semi-standardized scoring, but remain incomplete for multimodal, retrieval, and generative settings and can lag newly released models. In addition, leaderboard ordering can be sensitive to adaptation defaults, task weighting, and inclusion criteria; therefore, leaderboards are most useful when accompanied by transparency about adaptation budgets, cohort composition, overlap auditing, and uncertainty. Fig. 10 highlights coverage gaps, particularly for multimodal, retrieval, and generative settings, that remain underrepresented in controlled benchmarking.

Benchmarking also reveals structural constraints on community evidence. Several leading models including **TITAN** (Ding et al., 2025), **PRISM/PRISM2** (Shaikovski et al., 2024; Vorontsov et al., 2025), **Virchow2** (Zimmermann et al., 2024), **UNI2-h** (R.J. Chen et al., 2024), and **H-Optimus-1** (Bioptimus, 2025) are available for research under specific restrictions or non-commercial licenses. While these models support academic research, they are not fully reproducible if the data and pipelines for any downstream evaluations remain private. Other models, such as **Prov-GigaPath** (Xu et al., 2024) and **Hibou-L** (Nechaev et al., 2024), **H-Optimus-0** (Saillard et al., 2024), and **Virchow** (Vorontsov et al., 2024) have more permissive public releases. By contrast, **Atlas**, **Atlas-2** (Alber et al., 2025, 2026), and **RudolfV** (Dippel et al., 2024) are closed-weight systems; their reported results cannot be independently reproduced at the model level without released weights, embeddings, or fixed evaluation pipelines. We therefore distinguish reported evidence from independently reproducible evidence. Independent reproducibility requires access to model weights or embeddings, pretraining-data description, adaptation pipeline, downstream application data, patient-level split definitions, and evaluation scripts. Supplementary Table S2 records model access status, license/access restrictions, pretraining-data transparency, and overlap status where known. Taken together, Figs. 8 and 9, clarify why "best model" claims remain endpoint- and suite-specific: ordering shifts across task-cluster coverage breadth alongside top scores. For clinical translation, benchmark design should prioritize robustness-under-shift, calibrated operating points, and transparency about pretraining-benchmark overlap, rather than single-metric leaderboards. This design emphasis is consistent with the evidence map: families with dense results but limited external validation should be treated as maturity gaps rather than settled capability.

5. Clinical utility of pathology foundation models

Evidence for the clinical utility of pathology foundation models is emerging along three practical axes: (i) head-to-head evaluation against specialist, clinically deployed systems, (ii) prescreening/triage settings where performance must be reported at high-sensitivity operating points, and (iii) workflow-anchored studies, including prospective or "silent trial" designs, that quantify operational impact. Across all three, the defining criterion is whether the model supports a workflow decision with quantifiable operational benefit, not merely improved retrospective discrimination.

[1] https://kaiko-ai.github.io/eva/main/

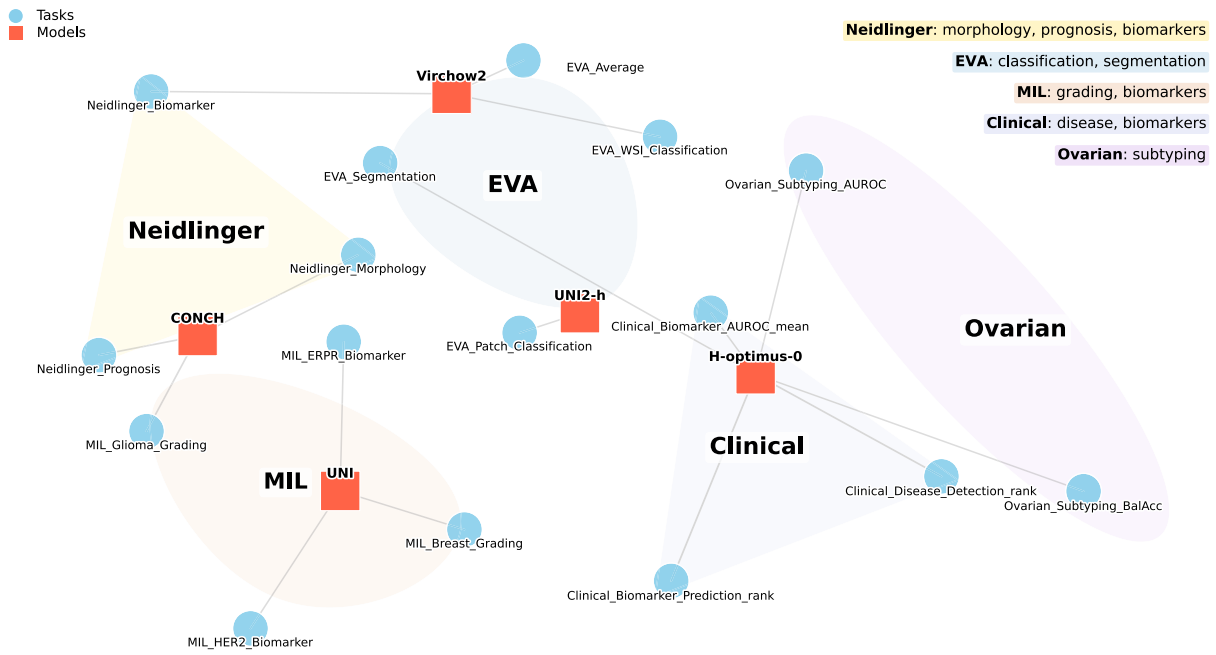


Fig. 10. Task-model linkage map for benchmarking suites. Nodes represent benchmark endpoints (tasks) and foundation models; edges indicate that a model was evaluated on the corresponding endpoint, visually summarizing model coverage across benchmark clusters.

Table 2
Clinical-grade reference comparison.

System	Metric	Reference	Prism2
Paige Prostate	AUC	0.985	0.985
Paige Prostate	bAcc	0.933	0.935
Paige Breast	AUC	0.988	0.992
Paige Breast	bAcc	0.967	0.962
Paige BLN	AUC	0.990	0.972
Paige BLN	bAcc	0.941	0.888

This section therefore uses selected clinical examples to review the deployment evidence. The key questions are not only which model scores highest, but which existing standard of care it complements, what operating point is clinically acceptable, who acts on the output, what confirmatory test or escalation pathway is triggered, and whether the model reduces workload, cost, or turnaround time without compromising safety.

A key question for adoption is whether a single reusable foundation model can approach or complement specialist clinical systems that are trained separately for each tissue and endpoint. Head-to-head evidence is therefore valuable but must be interpreted carefully: specialist products are optimized for specific workflows and operating points, while foundation model pipelines may require additional adaptation, calibration, local validation, and workflow integration.

Taken together, **Virchow** (Vorontsov et al., 2024) and **PRISM2** (Vorontsov et al., 2025) provide convergent evidence that foundation model-based pipelines can be competitive on clinically relevant endpoints, sometimes approaching specialist system performance under comparable evaluation conditions (Table 2). However, the decisive question is not superior discrimination, but whether performance is stable under site/stain/scanner shift, whether the calibrated operating point can be maintained in routine practice, and whether the output changes a real diagnostic workflow.

Prescreening tasks illustrate why clinical utility should be reported at fixed, high-sensitivity operating points rather than only with AUROC. At deployment, the question is often: how many negative or low-priority cases can be safely excluded while maintaining a target sensitivity, such as 0.95–0.99? Task-specific pipelines such as those built on IGUANA (Graham et al., 2023) and CAIMAN (Bilal et al., 2023a) operationalize this logic by optimizing specificity at high sensitivity. Foundation model pipelines should be judged against the same clinical standard.

MSI prediction in colorectal cancer is a canonical test case because it is clinically actionable, widely studied, and already embedded in clinical pathways through immunohistochemistry (IHC) and confirmatory molecular testing. Accordingly, H&E-based MSI models should not be presented as replacements for validated molecular assays unless such use is prospectively validated. In the short term, these models are most useful for prescreening or triage within current diagnostic workflows: reducing unnecessary confirmatory testing, accelerating turnaround, or flagging cases for automatic follow-up testing while maintaining a predefined sensitivity. MSI results should therefore be interpreted under the reporting controls summarized in Fig. 6 (Checklist: A–E), rather than as single-cohort headline AUROC.

In a standardized adaptation pipeline (Bilal et al., 2026) for MSS screening, **CONCH** (Lu et al., 2024a) produced comparable specificity (45% and 65%) at high-sensitivity operating points (94% and 90%, respectively), yielding a more favorable operating profile for prescreening. More broadly, reporting specificity-at-sensitivity with external validation provides a direct bridge from model performance to clinical workload reduction. Generalized MSI screening with light adaptation may offer a computationally efficient workflow compared with expensive task-specific designs that require cohort-specific calibration, additional filters, or segmentation-heavy interpretability pipelines (Baumann et al., 2026; Saillard et al., 2023).

These MSI examples illustrate why molecular grounding matters for interpreting morphology-based prediction. H&E morphology can reflect molecular state because genomic and transcriptomic alterations shape tumor architecture, immune contexture, stromal organization, necrosis, mucin production, differentiation, and cell morphology. MSI-high colorectal cancers often show recognizable features such as increased tumor-infiltrating lymphocytes, mucinous differentiation, poor differentiation, and characteristic immune/stromal patterns. These features are may not be fully validated, but they offer biological signals that computational models can use to perform molecular and biomarker tasks. Spatial-transcriptomic resources such as HEST (Jaume et al.,

2024) strengthen this bridge by directly linking image morphology with molecular readouts.

Beyond retrospective discrimination, clinically persuasive evidence links foundation model performance to measurable workflow gains. For example, **Prov-GigaPath** (Xu et al., 2024) has been evaluated in an EGFR screening pipeline (EAGLE) (Campanella et al., 2025b), where the model supports case prioritization at fixed sensitivity, translating predictive performance into reduced confirmatory testing burden and improved turnaround time. This is the type of evidence that strengthens adoption arguments: a clinically meaningful operating point, an operational benefit, multi-cohort validation, and a clear escalation path for uncertain or high-risk cases.

Workflow integration should therefore be specified before deployment. A study should define where the model enters the pathway, how it is adapted for a specific clinical task, who reviews the output, what threshold triggers action, how uncertain cases are handled, how false negatives and false positives are monitored, and whether the model reduces work without shifting unsafe burden to pathologists. To meet regulatory standards, systems must include local testing, detailed record-keeping, and performance monitoring. Organizations must also manage whether models are fixed or updated over time, watch for performance changes after launch, and ensure pathologists maintain clear oversight.

Across these exemplars, pathology foundation models appear most clinically valuable when they: (i) reduce per-task training burden through reuse and adaptation, (ii) maintain performance across heterogeneity in tissues, stains, scanners, and acquisition conditions, (iii) translate model outputs into workflow decisions such as triage, prescreening, diagnosis, biomarker screening, or decision support, and (iv) demonstrate measurable operational impact such as test reduction, throughput, turnaround time, or safer allocation of expert attention. Overall, the next decisive evidence tier is not marginal metric gains; it is prospective, operationally defined validation that demonstrates safe and scalable benefit in real laboratory workflows.

6. Conclusion and future directions

Pathology foundation models and AI copilots mark a shift in computational pathology from task-specific pipelines toward reusable representations that can be adapted across many clinical questions. Their promise is to amplify diagnostics skill by supporting faster adaptation, broader reuse, and assistant-style workflows that extend beyond classification to retrieval, reporting, and decision support. However, clinical value does not arise from pretraining alone; it emerges at the adaptation interface, where pathologists must retain control over intent, thresholds, and workflow integration.

Opportunity is inseparable from risk. Data imbalance, adaptation-induced bias, and representation shift can yield uneven performance across tissues, scanners, sites, and populations. Progress therefore depends less on isolated "best-reported" metrics and more on capability-oriented, deployment-aligned evidence: demonstrating distinctive capabilities through (i) general representation robustness (capability maps) and (ii) functional breadth across tasks and modalities (evidence maps), while evaluating reliability under real-world heterogeneity across settings. This motivates a deployment-aware agenda built around broad-and-deep benchmark bundles, standardized adaptation "budgets", and domain-agnostic evaluation at calibrated, use-case operating points, followed by validation on independent cohorts with clinically meaningful endpoints.

These requirements are amplified by whole-slide imaging realities of gigapixel scale, heterogeneous pre-analytics, and inter-site variability, which continue to challenge reproducibility and generalization. Translation further demands interpretable behavior, privacy-preserving governance, scalable infrastructure, and careful regulatory and workflow integration; without these, foundation models risk repeating the limited clinical uptake of earlier machine-learning waves.

Looking forward, the central question is not who owns the diagnostic decision, pathologists always do, but how efficiently expertise is deployed. Foundation models will be most compelling when tied to workflow-defined endpoints (e.g., triaging, prescreening at fixed high sensitivity, diagnosing, and biomarker screening) and supported by evidence breadth rather than single-number comparisons. Progress will be defined less by ever-larger models and more by deployment realism and standardization: building on the capability and evidence maps, we motivate broad-and-deep benchmark bundles with explicit overlap control, fixed adaptation budgets, calibrated decision policies at use-case operating points, and systematic stress testing under scanner/stain/site shift, ideally complemented by prospective or silent-trial evaluations aligned with clinical workflows. The decisive test is trust at scale: whether these systems can be validated, calibrated, and governed well enough to reliably protect and extend expert attention in routine practice.

CRediT authorship contribution statement

Mohsin Bilal: Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Aadam:** Writing – review & editing, Writing – original draft, Visualization, Methodology, Data curation. **Manahil Raza:** Writing – review & editing, Writing – original draft, Visualization, Methodology, Data curation. **Anas Alsuhaibani:** Writing – review & editing, Writing – original draft, Visualization, Data curation. **Youssef N. Altherwy:** Writing – review & editing, Writing – original draft, Visualization, Data curation. **Abdulrahman Alabduljabbar:** Writing – review & editing, Writing – original draft, Validation, Data curation. **Fahdah Almarshad:** Writing – review & editing, Writing – original draft, Visualization, Data curation. **Paul Golding:** Writing – review & editing, Writing – original draft, Validation, Investigation. **Nasir Rajpoot:** Writing – review & editing, Writing – original draft, Validation, Supervision, Investigation.

Declaration of competing interest

Nasir Rajpoot declares his role as a co-founder and shareholder of Histofy Ltd, a university spinout focusing on computational pathology software. All other authors declare no competing interests.

Acknowledgments

The authors extend their appreciation to Prince Sattam bin Abdulaziz University for funding this research work through project number (PSAU/2024/01/29745). The funder of the study had no role in study design, data collection, data analysis, data interpretation, or writing of the report.

Appendix A. Supplementary material

Supplementary Table S1. Reported performance of pathology foundation models across downstream tasks, extracted from the source studies and grouped by task category. For each task, we report the best-performing model, the reported performance metric and value, whether the evaluation dataset includes TCGA-derived data, and the pretraining-evaluation overlap status, categorized as known absent, possible, unknown, or not applicable.

Supplementary Table S2. A review of TCGA use in the pretraining of pathology foundation models, covering the reported TCGA pretraining category, pretraining-data transparency, and model access status.

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.mam.2026.101504.

References

Ahmed, F., Sellergren, A., Yang, L., Xu, S., Babenko, B., Ward, A., et al., 2024. PathAlign: A vision-language model for whole slide images in histopathology. In: Proceedings of the MICCAI Workshop on Computational Pathology. In: Proceedings of Machine Learning Research, vol. 254, PMLR, pp. 72–108.

Alber, M., Milbich, T., Carpen-Amarie, A., Tietz, S., Dippel, J., Muttenthaler, L., et al., 2026. Atlas 2 – foundation models for clinical deployment. http://dx.doi.org/10.48550/ARXIV.2601.05148, Version Number: 1. URL: https://arxiv.org/abs/2601.05148.

Alber, M., Tietz, S., Dippel, J., Milbich, T., Lesort, T., Korfiatis, P., et al., 2025. Atlas: A novel pathology foundation model by mayo clinic, charité, and aignostics. http://dx.doi.org/10.48550/arXiv.2501.05409, arXiv:2501.05409 [cs].

Alowais, S.A., Alghamdi, S.S., Alsuhebany, N., Alqahtani, T., Alshaya, A.I., Almohareb, S.N., et al., 2023. Revolutionizing healthcare: the role of artificial intelligence in clinical practice. BMC Med. Educ. 23 (1), 689. http://dx.doi.org/10.1186/s12909-023-04698-z, URL: https://bmcmededuc.biomedcentral.com/articles/10.1186/s12909-023-04698-z.

Atallah, N.M., Wahab, N., Toss, M.S., Makhlouf, S., Ibrahim, A.Y., Lashen, A.G., et al., 2023. Deciphering the morphology of tumor-stromal features in invasive breast cancer using artificial intelligence. Mod. Pathol. 36 (10), 100254. http://dx.doi.org/10.1016/j.modpat.2023.100254, URL: https://linkinghub.elsevier.com/retrieve/pii/S089339522300159X.

Azizi, S., Culp, L., Freyberg, J., Mustafa, B., Baur, S., Kornblith, S., et al., 2023. Robust and data-efficient generalization of self-supervised machine learning for diagnostic imaging. Nat. Biomed. Eng. 7 (6), 756–779. http://dx.doi.org/10.1038/s41551-023-01049-7, URL: https://www.nature.com/articles/s41551-023-01049-7.

Baumann, E., Schäfer, L.E., Meeuwsen, F., Kirsch, R., Nagtegaal, I.D., Berger, M.D., et al., 2026. MSAI-path: Predicting microsatellite instability from routine histology slides without reinventing the wheel. Mod. Pathol. 39 (1), 100932. http://dx.doi.org/10.1016/j.modpat.2025.100932, URL: https://linkinghub.elsevier.com/retrieve/pii/S0893395225002303.

Belagali, V., Kapse, S., Marza, P., Das, S., Li, Z., et al., 2025. TICON: A slide-level tile contextualizer for histopathology representation learning. http://dx.doi.org/10.48550/arXiv.2512.21331, URL: https://arxiv.org/abs/2512.21331, arXiv:2512.21331.

Bilal, M., Gulzar, M.A., Jaffar, N., Alabduljabbar, A., Altherwy, Y., Alsuhaibani, A., et al., 2026. Benchmarking pathology foundation models for predicting microsatellite instability in colorectal cancer histopathology. Comput. Med. Imaging Graph. 127, 102680. http://dx.doi.org/10.1016/j.compmedimag.2025.102680, URL: https://linkinghub.elsevier.com/retrieve/pii/S0895611125001892.

Bilal, M., Jewsbury, R., Wang, R., AlGhamdi, H.M., Asif, A., Eastwood, M., et al., 2023a. An aggregation of aggregation methods in computational pathology. Med. Image Anal. 88, 102885. http://dx.doi.org/10.1016/j.media.2023.102885.

Bilal, M., Nimir, M., Snead, D., Taylor, G.S., Rajpoot, N., 2023b. Role of AI and digital pathology for colorectal immuno-oncology. Br. J. Cancer 128, 3–11. http://dx.doi.org/10.1038/s41416-022-01986-1.

Bilal, M., Raza, S.E.A., Azam, A., Graham, S., Ilyas, M., Cree, I.A., et al., 2021c. Development and validation of a weakly supervised deep learning framework to predict the status of molecular pathways and key mutations in colorectal cancer from routine histology images: a retrospective study. Lancet Digit. Health 3 (12), e763–e772. http://dx.doi.org/10.1016/S2589-7500(21)00180-1, Number: 12. URL: https://linkinghub.elsevier.com/retrieve/pii/S2589750021001801.

Bilal, M., Tsang, Y.W., Ali, M., Graham, S., Hero, E., Wahab, N., et al., 2023. Development and validation of artificial intelligence-based prescreening of large-bowel biopsies taken in the UK and Portugal: a retrospective cohort study. Lancet Digit. Health 5 (11), e786–e797. http://dx.doi.org/10.1016/S2589-7500(23)00148-6, Number: 11. URL: https://linkinghub.elsevier.com/retrieve/pii/S2589750023001486.

Bioptimus, 2025. H-optimus-1. URL: https://huggingface.co/bioptimus/H-optimus-1.

Bommasani, R., Hudson, D.A., Adeli, E., Altman, R., Arora, S., von Arx, S., et al., 2022. On the Opportunities and Risks of Foundation Models. Technical Report, Stanford University, URL: https://arxiv.org/abs/2108.07258, arXiv:2108.07258.

Breen, J., Allen, K., Zucker, K., Godson, L., Orsi, N.M., Ravikumar, N., 2025. A comprehensive evaluation of histopathology foundation models for ovarian cancer subtype classification. Npj Precis. Oncol. 9 (1), 33. http://dx.doi.org/10.1038/s41698-025-00799-8, URL: https://www.nature.com/articles/s41698-025-00799-8.

Campanella, G., Chen, S., Singh, M., Verma, R., Muehlstedt, S., Zeng, J., et al., 2025a. A clinical benchmark of public self-supervised pathology foundation models. Nat. Commun. 16 (1), 3640. http://dx.doi.org/10.1038/s41467-025-58796-1, URL: https://www.nature.com/articles/s41467-025-58796-1.

Campanella, G., Fluder, E., Zeng, J., Vanderbilt, C., Fuchs, T.J., 2024. Beyond multiple instance learning: Full resolution all-in-memory end-to-end pathology slide modeling. URL: https://arxiv.org/abs/2403.04865, arXiv:2403.04865.

Campanella, G., Hanna, M.G., Geneslaw, L., Miraflor, A.P., Silva, V.W.K., Busam, K.J., et al., 2019. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. Nature Med. 25, 1301–1309. http://dx.doi.org/10.1038/s41591-019-0508-1.

Campanella, G., Kumar, N., Nanda, S., Singi, S., Fluder, E., Kwan, R., et al., 2025b. Real-world deployment of a fine-tuned pathology foundation model for lung cancer biomarker detection. Nature Med. 31 (9), 3002–3010. http://dx.doi.org/10.1038/s41591-025-03780-x, URL: https://www.nature.com/articles/s41591-025-03780-x.

Chanda, D., Aryal, M., Soltani, N.Y., Ganji, M., 2024. A new era in computational pathology: A survey on foundation and vision-language models. http://dx.doi.org/10.48550/arXiv.2408.14496.

Chen, S., Campanella, G., Elmas, A., Stock, A., Zeng, J., Polydorides, A.D., et al., 2024. Benchmarking embedding aggregation methods in computational pathology: A clinical data perspective. In: Ciompi, F., Khalili, N., Studer, L., Poceviciute, M., Khan, A., Veta, M., Jiao, Y., Haj-Hosseini, N., Chen, H., Raza, S., Minhas, I., Burlutskiy, N., Vilaplana, V., Brattoli, B., Muller, H., Atzori, M., Raza, S., Minhas, F. (Eds.), Proceedings of the MICCAI Workshop on Computational Pathology. In: Proceedings of Machine Learning Research, vol. 254, PMLR, pp. 38–50, URL: https://proceedings.mlr.press/v254/chen24a.html.

Chen, R.J., Chen, C., Li, Y., Chen, T.Y., Trister, A.D., Krishnan, R.G., Mahmood, F., 2022. Scaling vision transformers to gigapixel images via hierarchical self-supervised learning. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR, pp. 16123–16134. http://dx.doi.org/10.1109/CVPR52688.2022.01567.

Chen, R.J., Ding, T., Lu, M.Y., Williamson, D.F.K., Jaume, G., Song, A.H., et al., 2024. Towards a general-purpose foundation model for computational pathology. Nature Med. 30, 850–862. http://dx.doi.org/10.1038/s41591-024-02857-3.

Chen, T., Kornblith, S., Norouzi, M., Hinton, G., 2020. A simple framework for contrastive learning of visual representations. In: Proceedings of the 37th International Conference on Machine Learning. ICML '20, JMLR.org, pp. 1597–1607.

Chen, Y., Wang, G., Ji, Y., Li, Y., Ye, J., Li, T., Hu, M., Yu, R., Qiao, Y., He, J., 2025. SlideChat: A large vision-language assistant for whole-slide pathology image understanding. In: 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR, IEEE Computer Society, Los Alamitos, CA, USA, pp. 5134–5143. http://dx.doi.org/10.1109/CVPR52734.2025.00484, URL: https://doi.ieeecomputersociety.org/10.1109/CVPR52734.2025.00484.

Ding, T., Wagner, S.J., Song, A.H., Chen, R.J., Lu, M.Y., Zhang, A., et al., 2025. A multimodal whole-slide foundation model for pathology. Nature Med. 31, 3749–3761. http://dx.doi.org/10.1038/s41591-025-03982-3.

Dippel, J., Feulner, B., Winterhoff, T., Milbich, T., Tietz, S., Schallenberg, S., et al., 2024. RudolfV: A foundation model by pathologists for pathologists. arXiv:2401.04079 [cs, eess].

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., et al., 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations. pp. 1–22, URL: https://openreview.net/forum?id=YicbFdNTTy.

Driess, D., Xia, F., Sajjadi, M.S.M., Lynch, C., Chowdhery, A., Ichter, B., et al., 2023. PaLM-E: An embodied multimodal language model. In: Proceedings of the 40th International Conference on Machine Learning. ICML '23, JMLR.org, Honolulu, Hawaii, USA, pp. 8469–8488.

Filiot, A., 2026. HistoBoard: A unified dashboard for pathology foundation model benchmarks. https://wearewaiv.github.io/histoboard. (Accessed May 2026).

Filiot, A., Ghermi, R., Olivier, A., Jacob, P., Fidon, L., Mac Kain, A., et al., 2023. Scaling self-supervised learning for histopathology with masked image modeling. http://dx.doi.org/10.1101/2023.07.21.23292757, URL: http://medrxiv.org/lookup/doi/10.1101/2023.07.21.23292757.

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

Firoozi, R., Tucker, J., Tian, S., Majumdar, A., Sun, J., Liu, W., et al., 2025. Foundation models in robotics: Applications, challenges, and the future. Int. J. Robot. Res. 44 (5), 701–739. http://dx.doi.org/10.1177/02783649241281508, URL: https://journals.sagepub.com/doi/10.1177/02783649241281508.

Gildenblat, J., Hadar, O., 2024. Segmentation by factorization: Unsupervised semantic segmentation for pathology by factorizing foundation model features. URL: https://arxiv.org/abs/2409.05697, arXiv:2409.05697.

Graham, S., Minhas, F., Bilal, M., Ali, M., Tsang, Y.-W., Eastwood, M., et al., 2023. Screening of normal endoscopic large bowel biopsies with interpretable graph learning: a retrospective study. Gut 72, 1709–1721. http://dx.doi.org/10.1136/gutjnl-2023-329512.

Guo, Z., Ma, J., Xu, Y., Wang, Y., Wang, L., Chen, H., 2024. HistGen: Histopathology report generation via local-global feature encoding and cross-modal context interaction. In: Proceedings of Medical Image Computing and Computer Assisted Intervention – MICCAI 2024. Springer Nature Switzerland, pp. 189–199.

He, X., Cai, Z., Wei, W., Zhang, Y., Mou, L., Xing, E., et al., 2021. Towards visual question answering on pathology images. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers). Association for Computational Linguistics, Online, pp. 708–718. http://dx.doi.org/10.18653/v1/2021.acl-short.90.

Hoang, D.-T., Dinstag, G., Shulman, E.D., Hermida, L.C., Ben-Zvi, D.S., Elis, E., et al., 2024. A deep-learning framework to predict cancer treatment response from histopathology images through imputed transcriptomics. Nat. Cancer 5, 1305–1317. http://dx.doi.org/10.1038/s43018-024-00793-2, URL: https://www.nature.com/articles/s43018-024-00793-2.

Hua, S., Yan, F., Shen, T., Ma, L., Zhang, X., 2024. PathoDuet: Foundation models for pathological slide analysis of h&e and IHC stains. Med. Image Anal. 97, 103289. http://dx.doi.org/10.1016/j.media.2024.103289, URL: https://linkinghub.elsevier.com/retrieve/pii/S1361841524002147.

Huang, Z., Bianchi, F., Yuksekgonul, M., Montine, T.J., Zou, J., 2023. A visual-language foundation model for pathology image analysis using medical Twitter. Nature Med. 29, 2307–2316. http://dx.doi.org/10.1038/s41591-023-02504-3.

Jaume, G., Doucet, P., Song, A., Lu, M., Almagro-Pérez, C., Wagner, S., Vaidya, A., Chen, R., Williamson, D., Kim, A., Mahmood, F., 2024. HEST-1k: A dataset for spatial transcriptomics and histology image analysis. In: Advances in Neural Information Processing Systems 37. Neural Information Processing Systems Foundation, Inc., pp. 53798–53833. http://dx.doi.org/10.52202/079017-1704, URL: http://www.proceedings.com/079017-1704.html.

Jaume, G., Vaidya, A., Zhang, A., Song, A.H., Chen, R.J., Sahai, S., et al., 2025. Multistain pretraining for slide representation learning in pathology. In: Computer Vision – ECCV 2024. Springer Nature Switzerland, Cham, pp. 19–37. http://dx.doi.org/10.1007/978-3-031-73414-4_2.

Javed, S., Mahmood, A., Fraz, M.M., Koohbanani, N.A., Benes, K., Tsang, Y.-W., et al., 2020. Cellular community detection for tissue phenotyping in colorectal cancer histology images. Med. Image Anal. 63, 101696. http://dx.doi.org/10.1016/j.media.2020.101696, URL: https://linkinghub.elsevier.com/retrieve/pii/S136184152030061X.

Juyal, D., Padigela, H., Shah, C., Shenker, D., Harguindeguy, N., Liu, Y., et al., 2024. PLUTO: Pathology-universal transformer. In: ICML 2024 Workshop on Efficient and Accessible Foundation Models for Biological Discovery. Vienna, Austria, pp. 1–10.

kaiko.ai, Aben, N., de Jong, E.D., Gatopoulos, I., Känzig, N., Karasikov, M., Lagré, A., Moser, R., van Doorn, J., Tang, F., 2024. Towards large-scale training of pathology foundation models. arXiv:2404.15217 [cs].

Kang, M., Song, H., Park, S., Yoo, D., Pereira, S., 2023. Benchmarking self-supervised learning on diverse pathology datasets. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR, IEEE, Vancouver, BC, Canada, pp. 3344–3354. http://dx.doi.org/10.1109/CVPR52729.2023.00326, URL: https://ieeexplore.ieee.org/document/10204656/.

Kapse, S., Aygün, M., Cole, E., Lundberg, E., Song, L., Xing, E.P., 2026. GenBio-PathFM: A state-of-the-art foundation model for histopathology. http://dx.doi.org/10.64898/2026.03.17.712534, BioRxiv. URL: https://www.biorxiv.org/content/early/2026/03/20/2026.03.17.712534.

Karasikov, M., Doorn, J.v., Känzig, N., Cesur, M.E., Horlings, H.M., Berke, R., et al., 2025. Training state-of-the-art pathology foundation models with orders of magnitude less data. In: Proceedings of Medical Image Computing and Computer Assisted Intervention – MICCAI 2025. Springer Nature Switzerland, pp. 573–583.

Kather, J.N., Heij, L.R., Grabsch, H.I., Loeffler, C., Echle, A., Muti, H.S., et al., 2020. Pan-cancer image-based detection of clinically actionable genetic alterations. Nat. Cancer 1, 789–799. http://dx.doi.org/10.1038/s43018-020-0087-6.

Kather, J.N., Krisam, J., Charoentong, P., Luedde, T., Herpel, E., Weis, C.-A., et al., 2019. Predicting survival from colorectal cancer histology slides using deep learning: A retrospective multicenter study. In: Butte, A.J. (Ed.), PLOS Med. 16 (1), e1002730. http://dx.doi.org/10.1371/journal.pmed.1002730, Number: 1. URL: https://dx.plos.org/10.1371/journal.pmed.1002730.

Kather, J.N., Weis, C.-A., Bianconi, F., Melchers, S.M., Schad, L.R., Gaiser, T., et al., 2016. Multi-class texture analysis in colorectal cancer histology. Sci. Rep. 6 (1), 27988. http://dx.doi.org/10.1038/srep27988, URL: https://www.nature.com/articles/srep27988.

Kömen, J., Jong, E.D.d., Hense, J., Marienwald, H., Dippel, J., Naumann, P., et al., 2026. Towards robust foundation models for digital pathology. Nat. Commun. 17, http://dx.doi.org/10.1038/s41467-026-73923-2, URL: https://www.nature.com/articles/s41467-026-73923-2.

Lai, J., Ahmed, F., Vijay, S., Jaroensri, T., Loo, J., Vyawahare, S., et al., 2023. Domain-specific optimization and diverse evaluation of self-supervised models for histopathology. http://dx.doi.org/10.48550/arXiv.2310.13259, arXiv:2310.13259 [eess].

Lazard, T., Lerousseau, M., Gardrat, S., Vincent-Salomon, A., Stern, M.-H., Rodrigues, M., et al., 2023. Democratizing computational pathology: optimized whole slide image representations for the cancer genome atlas. http://dx.doi.org/10.1101/2023.12.04.569894, URL: http://biorxiv.org/lookup/doi/10.1101/2023.12.04.569894.

Lee, J., Lim, J., Byeon, K., Kwak, J.T., 2025. Benchmarking pathology foundation models: Adaptation strategies and scenarios. Comput. Biol. Med. 190, 110031. http://dx.doi.org/10.1016/j.compbiomed.2025.110031, URL: https://linkinghub.elsevier.com/retrieve/pii/S0010482525003828.

Lenz, T., Neidlinger, P., Ligero, M., Wölflein, G., Treeck, M.v., Kather, J.N., 2025. Unsupervised foundation model-agnostic slide-level representation learning. In: Proceedings of the Computer Vision and Pattern Recognition Conference. CVPR, pp. 30807–30817.

Liu, P., Ji, L., Gou, J., Fu, B., Ye, M., 2025. Interpretable vision-language survival analysis with ordinal inductive bias for computational pathology. In: The Thirteenth International Conference on Learning Representations. pp. 1–24, URL: https://openreview.net/forum?id=trj2Jq8riA.

Liu, J., Yang, C., Lu, Z., Chen, J., Li, Y., Zhang, M., et al., 2025. Graph foundation models: Concepts, opportunities and challenges. IEEE Trans. Pattern Anal. Mach. Intell. 47 (6), 5023–5044. http://dx.doi.org/10.1109/TPAMI.2025.3548729, URL: https://ieeexplore.ieee.org/document/10915556/.

Lu, M.Y., Chen, B., Williamson, D.F.K., Chen, R.J., Liang, I., Ding, T., et al., 2024a. A visual-language foundation model for computational pathology. Nature Med. 30, 863–874. http://dx.doi.org/10.1038/s41591-024-02856-4.

Lu, M.Y., Chen, B., Williamson, D.F.K., et al., 2024b. A multimodal generative AI copilot for human pathology. Nature 634, 466–473. http://dx.doi.org/10.1038/s41586-024-07618-3.

Lu, M., Wang, T., Zeng, Q., Lu, Z., Li, Z., Xia, Y., 2025. PathBot: A foundation model for pathological image analysis. IEEE J. Biomed. Health Informatics 29 (12), 8976–8987. http://dx.doi.org/10.1109/JBHI.2025.3619967, URL: https://ieeexplore.ieee.org/document/11198904/.

Luo, X., Xiang, J., Ji, Y., Li, R., 2025. nnMIL: A generalizable multiple instance learning framework for computational pathology. http://dx.doi.org/10.48550/arXiv.2511.14907, URL: https://arxiv.org/abs/2511.14907, arXiv:2511.14907.

Ma, J., Guo, Z., Zhou, F., Wang, Y., Xu, Y., Li, J., et al., 2025. A generalizable pathology foundation model using a unified knowledge distillation pretraining framework. Nat. Biomed. Eng. 10, http://dx.doi.org/10.1038/s41551-025-01488-4, URL: https://www.nature.com/articles/s41551-025-01488-4.

Mallya, M., Mirabadi, A.K., Farnell, D., Farahani, H., Bashashati, A., 2025. Benchmarking histopathology foundation models for ovarian cancer bevacizumab treatment response prediction from whole slide images. Discov. Oncol. 16 (1), 196. http://dx.doi.org/10.1007/s12672-025-01973-x, URL: https://link.springer.com/10.1007/s12672-025-01973-x.

Meseguer, P., Del Amor, R., Naranjo, V., 2025. MI-VisionShot: Few-shot adaptation of vision-language models for slide-level classification of histopathological images. In: Chen, Y.-W., Tanaka, S., Howlett, R.J., Jain, L.C. (Eds.), In: Innovation in Medicine and Healthcare, vol. 412, Springer Nature Singapore, Singapore, pp. 15–24. http://dx.doi.org/10.1007/978-981-97-7498-2_2, Series Title: Smart Innovation, Systems and Technologies. URL: https://link.springer.com/10.1007/978-981-97-7498-2_2.

Nechaev, D., Pchelnikov, A., Ivanova, E., 2024. Hibou: A family of foundational vision transformers for pathology. arXiv:2406.05074 [cs, eess].

Neidlinger, P., Nahhas, O.S.M.E., Muti, H.S., Lenz, T., Hoffmeister, M., Brenner, H., et al., 2026. Benchmarking foundation models as feature extractors for weakly supervised computational pathology. Nat. Biomed. Eng. 10, 1113–1123. http://dx.doi.org/10.1038/s41551-025-01516-3, URL: https://www.nature.com/articles/s41551-025-01516-3.

Nicke, T., Schäfer, J.R., Höfener, H., Feuerhake, F., Merhof, D., Kießling, F., et al., 2025. Tissue concepts: Supervised foundation models in computational pathology. Comput. Biol. Med. 186, 109621. http://dx.doi.org/10.1016/j.compbiomed.2024.109621, URL: https://linkinghub.elsevier.com/retrieve/pii/S0010482524017062.

Ochi, M., Komura, D., Ishikawa, S., 2025. Pathology foundation models. JMA J. 8 (1), 121–130. http://dx.doi.org/10.31662/jmaj.2024-0206, URL: https://www.jmaj.jp/detail.php?id=10.31662/jmaj.2024-0206.

OpenAI, 2023. GPT-4V(ision) System Card. Technical Report, OpenAI, URL: https://cdn.openai.com/papers/GPTV_System_Card.pdf.

Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., et al., 2024. DINOv2: Learning robust visual features without supervision. Featured Certification, URL: https://openreview.net/forum?id=a68SUt6zFt.

Ozlugedik, F.I., Dasdelen, M.F., Umer, R.M., Marr, C., 2026. MoA: Mixture of aggregators improves slide-level diagnosis in computational pathology. MIDL 2026 Poster. URL: https://openreview.net/forum?id=OEFIRrE6iT.

Padigela, H., Nofallah, S., Chilaparasetti, A.N., Han, R., Walker, A., Shen, J., et al., 2025. PLUTO-4: Frontier pathology foundation models. http://dx.doi.org/10.48550/arXiv.2511.02826, arXiv:2511.02826 [cs].

Perez-Lopez, R., Ghaffari Laleh, N., Mahmood, F., Kather, J.N., 2024. A guide to artificial intelligence for cancer researchers. Nat. Rev. Cancer 24 (6), 427–441. http://dx.doi.org/10.1038/s41568-024-00694-7, URL: https://www.nature.com/articles/s41568-024-00694-7.

Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., et al., 2021. Learning transferable visual models from natural language supervision. In: Proceedings of the 38th International Conference on Machine Learning. In: Proceedings of Machine Learning Research, vol. 139, PMLR, pp. 8748–8763, URL: https://proceedings.mlr.press/v139/radford21a.html.

Ray, P.P., 2023. Benchmarking, ethical alignment, and evaluation framework for conversational AI: Advancing responsible development of ChatGPT. BenchCouncil Trans. Benchmarks Stand. Eval. 3 (3), 100136. http://dx.doi.org/10.1016/j.tbench.2023.100136, URL: https://linkinghub.elsevier.com/retrieve/pii/S2772485923000534.

Saillard, C., Dubois, R., Tchita, O., Loiseau, N., Garcia, T., Adriansen, A., et al., 2023. Validation of MSIntuit as an AI-based pre-screening tool for MSI detection from colorectal cancer histology slides. Nat. Commun. 14, 6695. http://dx.doi.org/10.1038/s41467-023-42453-6.

Saillard, C., Jenatton, R., Llinares-López, F., Mariet, Z., Cahané, D., Durand, E., et al., 2024. H-optimus-0. URL: https://github.com/bioptimus/releases/tree/main/models/h-optimus/v0.

Saldanha, O.L., Loeffler, C.M.L., Niehues, J.M., Van Treeck, M., Seraphin, T.P., Hewitt, K.J., et al., 2023. Self-supervised attention-based deep learning for pan-cancer mutation prediction from histopathology. Npj Precis. Oncol. 7 (1), 35. http://dx.doi.org/10.1038/s41698-023-00365-0, URL: https://www.nature.com/articles/s41698-023-00365-0.

Sellergren, A., Kazemzadeh, S., Jaroensri, T., Kiraly, A., Traverse, M., Kohlberger, T., et al., 2025. MedGemma technical report. http://dx.doi.org/10.48550/arXiv.2507.05201, arXiv:2507.05201 [cs].

Shaikovski, G., Casson, A., Severson, K., Zimmermann, E., Wang, Y.K., Kunz, J.D., et al., 2024. PRISM: A multi-modal generative foundation model for slide-level histopathology. arXiv:2405.10254 [cs, eess].

Shao, D., Runevic, J., Chen, R.J., Williamson, D.F.K., Kim, A., Song, A.H., Mahmood, F., 2026. Mixture of mini experts: Overcoming the linear layer bottleneck in multiple instance learning. URL: https://openreview.net/forum?id=S5Io33pc78.

Singhal, K., Tu, T., Gottweis, J., Sayres, R., Wulczyn, E., Amin, M., et al., 2025. Toward expert-level medical question answering with large language models. Nature Med. 31, 943–950. http://dx.doi.org/10.1038/s41591-024-03423-7.

Song, A.H., Jaume, G., Williamson, D.F.K., Lu, M.Y., Vaidya, A., Miller, T.R., et al., 2023. Artificial intelligence for digital and computational pathology. Nat. Rev. Bioeng. 1 (12), 930–949. http://dx.doi.org/10.1038/s44222-023-00096-8, URL: https://www.nature.com/articles/s44222-023-00096-8.

Sun, Y., Wu, H., Zhu, C., Zheng, S., Chen, Q., Zhang, K., et al., 2024a. PathMMU: A massive multimodal expert-level benchmark for understanding and reasoning in pathology. In: Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part LXII. Springer-Verlag, Berlin, Heidelberg, pp. 56–73. http://dx.doi.org/10.1007/978-3-031-73033-7_4.

Sun, Y., Zhang, Y., Si, Y., Zhu, C., Shui, Z., Zhang, K., et al., 2025. PathGen-1.6M: 1.6 million pathology image-text pairs generation through multi-agent collaboration. In: The Thirteenth International Conference on Learning Representations. ICLR 2025, Singapore, pp. 1–43.

Sun, Y., Zhu, C., Zheng, S., Zhang, K., Sun, L., Shui, Z., et al., 2024b. PathAsst: A generative foundation AI assistant towards artificial general intelligence of pathology. In: Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence. In: AAAI'24/IAAI'24/EAAI'24, AAAI Press, http://dx.doi.org/10.1609/aaai.v38i5.28308.

Tamkin, A., Liu, V., Lu, R., Fein, D., Schultz, C., Goodman, N., 2021. DABS: A domain-agnostic benchmark for self-supervised learning. In: Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, Virtual. pp. 1–12.

Tan, J.W., Kim, S., Kim, E., Lee, S.H., Ahn, S., Jeong, W.-K., 2024. Clinical-grade multi-organ pathology report generation for multi-scale whole slide images via a semantically guided medical text foundation model. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2024. Springer Nature Switzerland, Cham, pp. 25–35. http://dx.doi.org/10.1007/978-3-031-72083-3_3.

Tran, M., Schmidle, P., Guo, R.R., Wagner, S.J., Koch, V., Lupperger, V., et al., 2025. Generating dermatopathology reports from gigapixel whole slide images with HistoGPT. Nat. Commun. 16 (1), 4886. http://dx.doi.org/10.1038/s41467-025-60014-x, URL: https://www.nature.com/articles/s41467-025-60014-x.

Vaidya, A., Zhang, A., Jaume, G., Song, A.H., Ding, T., Wagner, S.J., et al., 2025. Molecular-driven foundation model for oncologic pathology. http://dx.doi.org/10.48550/arXiv.2501.16652, arXiv:2501.16652 [cs].

Van Calster, B., Collins, G.S., Vickers, A.J., Wynants, L., Kerr, K.F., Barreñada, L., et al., 2025. Evaluation of performance measures in predictive artificial intelligence models to support medical decisions: overview and guidance. Lancet Digit. Health 7 (12), 100916. http://dx.doi.org/10.1016/j.landig.2025.100916, URL: https://linkinghub.elsevier.com/retrieve/pii/S2589750025000986.

Vorontsov, E., Bozkurt, A., Casson, A., Shaikovski, G., Zelechowski, M., Severson, K., et al., 2024. A foundation model for clinical-grade computational pathology and rare cancers detection. Nature Med. 30, http://dx.doi.org/10.1038/s41591-024-03141-0, URL: https://www.nature.com/articles/s41591-024-03141-0.

Vorontsov, E., Shaikovski, G., Casson, A., Viret, J., Zimmermann, E., Tenenholtz, N., et al., 2025. PRISM2: Unlocking multi-modal general pathology AI with clinical dialogue. http://dx.doi.org/10.48550/arXiv.2506.13063, arXiv:2506.13063 [cs].

Wahab, N., Toss, M., Miligy, I.M., Jahanifar, M., Atallah, N.M., Lu, W., et al., 2023. AI-enabled routine h&e image based prognostic marker for early-stage luminal breast cancer. Npj Precis. Oncol. 7 (1), 122. http://dx.doi.org/10.1038/s41698-023-00472-y, URL: https://www.nature.com/articles/s41698-023-00472-y.

Wang, X., Du, Y., Yang, S., Zhang, J., Wang, M., Zhang, J., et al., 2023. RetCCL: Clustering-guided contrastive learning for whole-slide image retrieval. Med. Image Anal. 83, 102645. http://dx.doi.org/10.1016/j.media.2022.102645, URL: https://linkinghub.elsevier.com/retrieve/pii/S1361841522002730.

Wang, X., Jiang, Y., Yang, S., Wang, F., Zhang, X., Wang, W., et al., 2025. Foundation model for predicting prognosis and adjuvant therapy benefit from digital pathology in GI cancers. J. Clin. Oncol. 43 (32), 3468–3481. http://dx.doi.org/10.1200/JCO-24-01501, URL: https://ascopubs.org/doi/10.1200/JCO-24-01501.

Wang, Y.K., Tydlitatova, L., Kunz, J.D., Oakley, G., Chow, B.K.B., Godrich, R.A., Lee, M.C.H., Haothers, 2025. Screen them all: High-throughput pan-cancer genetic and phenotypic biomarker screening from h&e whole slide images. URL: https://arxiv.org/abs/2408.09554, arXiv:2408.09554.

Wang, X., Yang, S., Zhang, J., Wang, M., Zhang, J., Yang, W., et al., 2022. Transformer-based unsupervised contrastive learning for histopathological image classification. Med. Image Anal. 81, 102559. http://dx.doi.org/10.1016/j.media.2022.102559, URL: https://linkinghub.elsevier.com/retrieve/pii/S1361841522002043.

Wang, X., Zhao, J., Marostica, E., Yuan, W., Jin, J., Zhang, J., et al., 2024. A pathology foundation model for cancer diagnosis and prognosis prediction. Nature 634, 970–978. http://dx.doi.org/10.1038/s41586-024-07894-z, URL: https://www.nature.com/articles/s41586-024-07894-z.

Wu, Y., Li, S., Du, Z., Zhu, W., 2023. BROW: Better features for whole slide image based on self-distillation. arXiv:2309.08259 [cs].

Xiang, J., Wang, X., Zhang, X., Xi, Y., Eweje, F., Chen, Y., et al., 2025. A vision–language foundation model for precision oncology. Nature 638, http://dx.doi.org/10.1038/s41586-024-08378-w, URL: https://www.nature.com/articles/s41586-024-08378-w.

Xie, Z., Zhang, Z., Cao, Y., Lin, Y., Bao, J., Yao, Z., et al., 2022. SimMIM: a simple framework for masked image modeling. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR, IEEE, New Orleans, LA, USA, pp. 9643–9653. http://dx.doi.org/10.1109/CVPR52688.2022.00943, URL: https://ieeexplore.ieee.org/document/9880205/.

Xu, H., Usuyama, N., Bagga, J., Zhang, S., Rao, R., Naumann, T., et al., 2024. A whole-slide foundation model for digital pathology from real-world data. Nature 630 (8015), 181–188. http://dx.doi.org/10.1038/s41586-024-07441-w, URL: https://www.nature.com/articles/s41586-024-07441-w.

Xu, H., Wang, M., Shi, D., Qin, H., Zhang, Y., Liu, Z., et al., 2025. When multiple instance learning meets foundation models: Advancing histological whole slide image analysis. Med. Image Anal. 101, 103456. http://dx.doi.org/10.1016/j.media.2025.103456, URL: https://linkinghub.elsevier.com/retrieve/pii/S1361841525000040.

Xu, Y., Wang, Y., Zhou, F., Ma, J., Jin, C., Yang, S., et al., 2025. A multimodal knowledge-enhanced whole-slide pathology foundation model. Nat. Commun. 16 (1), 11406. http://dx.doi.org/10.1038/s41467-025-66220-x, URL: https://www.nature.com/articles/s41467-025-66220-x.

Yan, F., Wu, J., Li, J., Wang, W., Chen, Y., et al., 2025. PathOrchestra: a comprehensive foundation model for computational pathology with over 100 diverse clinical-grade tasks. Npj Digit. Med. http://dx.doi.org/10.1038/s41746-025-02027-w, URL: https://www.nature.com/articles/s41746-025-02027-w.

Zhang, L., Song, J., Gao, A., Chen, J., Bao, C., Ma, K., 2019. Be your own teacher: Improve the performance of convolutional neural networks via self distillation. In: 2019 IEEE/CVF International Conference on Computer Vision. ICCV, pp. 3712–3721. http://dx.doi.org/10.1109/ICCV.2019.00381.

Zheng, S., Cui, X., Sun, Y., Li, J., Li, H., Zhang, Y., et al., 2025. Benchmarking PathCLIP for pathology image analysis. J. Imaging Informatics Med. 38, 422–438. http://dx.doi.org/10.1007/s10278-024-01128-4, URL: https://link.springer.com/10.1007/s10278-024-01128-4.

Zhong, W., Cui, R., Guo, Y., Liang, Y., Lu, S., Wang, Y., et al., 2024. AGIEval: A human-centric benchmark for evaluating foundation models. In: Findings of the Association for Computational Linguistics: NAACL 2024. Association for Computational Linguistics, Mexico City, Mexico, pp. 2299–2314. http://dx.doi.org/10.18653/v1/2024.findings-naacl.149, URL: https://aclanthology.org/2024.findings-naacl.149.

Zimmermann, E., Vorontsov, E., Viret, J., Casson, A., Zelechowski, M., Shaikovski, G., et al., 2024. Virchow 2: Scaling self-supervised mixed magnification models in pathology. http://dx.doi.org/10.48550/arXiv.2408.00738, arXiv:2408.00738 [cs].