Why Foundation Models in Pathology Are Failing

Hamid R. Tizhoosh
Kimia Lab, Mayo Clinic, Rochester, MN, USA

Abstract

In non-medical domains, foundation models (FMs) have revolutionized computer vision and language processing through large-scale self-supervised and multimodal learning. Consequently, their rapid adoption in computational pathology was expected to deliver comparable breakthroughs in cancer diagnosis, prognostication, and multimodal retrieval. However, recent systematic evaluations reveal fundamental weaknesses—low diagnostic accuracy, poor robustness, geometric instability, heavy computational demands, and concerning safety vulnerabilities. This short paper examines these shortcomings and argues that they stem from deeper conceptual mismatches between the assumptions underlying generic foundation modeling in mainstream AI and the intrinsic complexity of human tissue. Seven interrelated causes are identified: biological complexity, ineffective self-supervision, overgeneralization, excessive architectural complexity, lack of domain-specific innovation, insufficient data, and a fundamental design flaw related to tissue patch size. These findings suggest that current pathology foundation models remain conceptually misaligned with the nature of tissue morphology and call for a fundamental rethinking of the paradigm itself.

1. Introduction

Foundation models (FMs) have transformed image and text processing by leveraging massive data corpora and self-supervised learning [1–3]. Their success in general computer vision inspired enthusiasm for analogous applications in histopathology, where the potential for automating diagnostic reasoning and multimodal retrieval is expected to be immense [4, 5]. However, emerging evidence indicates that pathology FMs underperform relative to expectations and exhibit critical weaknesses under realistic evaluation. This short paper analyzes these limitations and explores conceptual reasons behind their failures.

2. Empirical Evidence of Weaknesses

While not exhaustive, the following summaries highlight several recent studies that have reported weaknesses in foundation models to analyze tissue images.

Low Accuracy and Inconsistent Clinical Performance – Alfasly *et al.* [6] evaluated several leading pathology foundation models—including UNI, GigaPath, and Virchow—on 11,444 whole-slide images from 23 organs and 117 cancer subtypes in TCGA using a zero-shot retrieval framework based on Yottixel's patch-embedding approach. Despite their scale, these models achieved only modest performance, with macro-averaged F1 scores around 40–42% for top-5 retrieval and pronounced organ-level variability: kidneys reached up to 68% (top-1 F1)[1], whereas lungs dropped to 21%. Aggregating patches into single WSI-level embeddings did not improve results and sometimes degraded them, suggesting loss of spatial information. Overall, while pathology foundation models outperform older CNN baselines, their absolute accuracy remains low, revealing limited generalization across tissue types and underscoring that current FMs capture texture patterns rather than true diagnostic morphology.

Lack of Robustness and Site Bias – De Jong *et al.* [7] systematically evaluated the robustness and generalization of ten leading pathology foundation models across multiple institutions and datasets using a newly defined *Robustness Index* (RI), which quantifies whether model embeddings cluster more strongly by biological class or by medical center. The RI compares within-class versus within-center similarity (with RI>1 indicating true biological robustness). Among all tested models, only Virchow2 achieved RI $\approx$ 1.2—meaning biological structure dominated site-specific bias—whereas all others had RI ¡ 1 (e.g., UNI $\approx$ 0.9, Phikon-v2 $\approx$ 0.7). Embeddings from most models, therefore, grouped primarily by hospital or scanner rather than by cancer type, leading to large performance drops on unseen centers. The study concludes that current pathology foundation models remain fragile, confounded, and insufficiently domain-robust, underscoring the urgent need for cross-institutional validation and bias-resilient architectures before clinical use.

Geometric Fragility – Elphick *et al.* [8] investigated twelve self-supervised pathology foundation models and assess how well their latent representations remain stable when image patches are rotated. The authors apply rotations in 15° increments from 0° to 360° on patches extracted from the TCGA-KIRC dataset and compute two metrics to quantify invariance: mean mutual k-nearest neighbours (m-kNN) and mean cosine distance between embeddings of non-rotated versus rotated patches. They

[1] Outlier organ performance, such as the unusually high kidney F1 scores, may partly reflect genuine morphological distinctiveness—but could also indicate **data leakage** or hidden technical bias. In pathology FMs, these effects are common and often hard to disentangle, so all high outlier results warrant careful patient-level and site-level validation.

report that the model PathDino (a small model with less than 10M parameters [9]) achieved the highest m-kNN score of 0.85, making it the most rotation-invariant by that measure, and that Hibou-L achieved the lowest cosine distance of 0.016, indicating the best alignment for that metric. In contrast, Virchow had the lowest m-kNN (≈0.53) and Phikon 2 the highest cosine distance (≈0.145), indicating much poorer invariance. Importantly, the results show statistically that models trained with explicit rotation augmentation significantly outperform those without (t = 6.91; $p < 0.0001$ for m-kNN; $t = -8.88$; $p < 0.0001$ for cosine distance). The study thus concludes that because transformer-based architectures lack an inherent rotational inductive bias, rotation augmentation in training is a necessary design choice for pathology FMs to achieve acceptable invariance.

Resource Burden and Fragile Adaptation – Mulliqi *et al.* [10] conducted a large-scale study using over 100,000 prostate biopsy slides (from 7,342 patients across 15 sites in 11 countries) to compare two pathology foundation models (FMs) against a task-specific (TS) end-to-end model for prostate cancer diagnosis and Gleason grading. They found that although the FMs offered utility in data-scarce settings, when enough labeled data were available the TS model matched or even outperformed the FMs. Critically, the FMs consumed up to 35× more energy than the TS model, raising sustainability concerns. Despite the purported universality of FMs, their performance did not substantially exceed TS models in the clinically validated setting, highlighting that heavy compute demands and fine-tuning instability limit the practical superiority of FMs in pathology.

Linear Probing: Because Fine-Tuning Does Not Work – Although foundation models are often promoted for their flexibility and "emergent" adaptability to new tasks, in computational pathology their downstream use is overwhelmingly limited to linear probing—training a shallow linear classifier on frozen embeddings rather than fine-tuning the model itself. This dependency arises because most pathology FMs are too large, memory-intensive, and unstable to fine-tune on moderate-sized data sets typical of clinical research (often hundreds to a few thousand slides). Recent studies confirm that full fine-tuning frequently degrades accuracy relative to linear probing due to overfitting and catastrophic forgetting [10, 11]. Yet this pragmatic retreat stands in stark contrast to the foundational premise of the FM paradigm, as articulated in Bommasani et al. (2021) [1] and many subsequent multimodal works that promise large pretrained systems should enable *zero-shot* and easily fine-tuned adaptation across domains. In pathology, however, this promise collapses—most "foundations" function only as static feature extractors that must be linearly probed, a situation akin to **buying a Ferrari that cannot run and then purchasing a bicycle to tow it**. The contradiction highlights that current FMs are less "foundations" than frozen front-ends, exposing the gap between theoretical universality and real-world usability in medical AI.

Figure 1: AI models can recognize dogs and even distinguish among breeds—tasks that children can perform with ease. In contrast, recognizing complex tissue patterns in pathology requires an adult with more than a decade of specialized education and training.

Security and Safety Vulnerabilities – Wang *et al.* [12] introduced *Universal and Transferable Adversarial Perturbations (UTAP)*, imperceptible noise patterns that collapse FM embeddings across architectures. These universal attacks threaten clinical reliability. In histopathology, these perturbations have a dual interpretation go beyond malicious attacks that are clearly a security risk. They have a real-world noise analogue; they approximate the small, systematic variations that arise naturally in the imaging pipeline:

- Differences in H&E staining
- Scanner optics and illumination variability
- Compression artifacts
- color normalization and rescaling
- Slide preparation imperfections

—— *dust, bubbles, section thickness, etc.*

- Downstream digital processing

—- *gamma correction, color-space conversion*

In this view, adversarial perturbations are not malicious but **diagnostic stress tests** that reveal how sensitive a model is to minor pixel-level changes that can—and do—occur in laboratory and acquisition workflows.

3. Why Are Pathology FMs Failing?

This section attempts to provide a conceptual and holistic analysis of why foundation models in pathology underperform, examining the cognitive, methodological, and epistemic assumptions underlying current AI paradigms.

Underestimating the Complexity of Human Tissue – The AI community often underestimates the semantic complexity of tissue morphology (Fig.1). A child learns to recognize dogs by age two and breeds by seven (a task at which AI performs exceptionally well) [13, 14]; A pathologist—a highly trained human adult—typically requires more than twelve years of education to distinguish cancer subtypes based on tissue morphology. Unlike natural images, tissue interpretation depends on context, scale, and clinical correlation—far beyond simple object recognition.

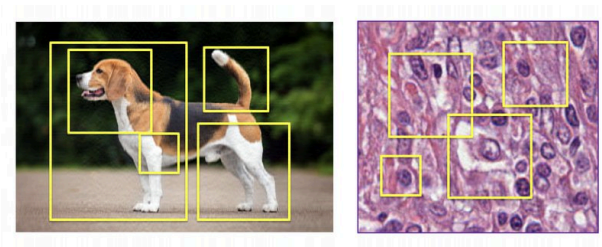

Figure 2: Self-supervised learning often rests on the implicit assumption that each image represents a single, coherent object—an assumption that fails in histopathology, where multiple heterogeneous tissue structures coexist within the same field of view.

Ineffective Self-Supervision for Tissue Images – Most early self-supervised learning frameworks were developed and validated on single-object datasets such as ImageNet, making them less suited for complex, structure-rich images like histopathology slides, which lack discrete, well-defined objects (Fig.2). The "local-global crop" assumption fails in tissue slides, where patches contain mixed or irrelevant content. As a result, models learn stain texture instead of biological patterns, weakening generalization.

The Myth of the Universal Model – According to the *No Free Lunch theorem* [15], no single model excels across all problems. Expecting one FM to generalize to all organs and cancers ignores pathology's heterogeneity. Benchmarks reveal wide organ-dependent performance swings, underscoring the limits of universal architectures.

Architectural Excess and Occam's Razor – Modern foundation models (FMs) often pursue ever-greater depth and parameter scaling without demonstrable performance benefits. In line with Occam's Razor [16], progress in pathology AI may instead depend on leaner, domain-structured architectures that embody the hierarchical and contextual organization of biological tissues, thereby enhancing both interpretability and generalization.

Lack of Domain-Specific Innovation – Many pathology foundation models (FMs) are direct adaptations of general-purpose frameworks such as CLIP, DINO, or MAE, retrained or fine-tuned with pathology data. However, few incorporate domain-specific mechanisms such as magnification awareness, stain-invariant representations, or morphology-aware pretext tasks, resulting in methodological stagnation and limited engagement with the underlying biological domain [17]. Despite the unique multi-scale and heterogeneous nature of tissue images, no major advances have been introduced in model topology, input preparation, or loss formulation to explicitly tailor foundation models to the structural and semantic characteristics of histopathology.

Data Deficit and Scaling Limits – CLIP was trained on 400 million image–text pairs; no pathology dataset approaches that scale. Even the largest multi-institutional archives offer fewer than one million WSIs, fragmented and inconsistently labeled. Privacy constraints exacerbate scarcity, capping the value of foundation-scale pretraining.

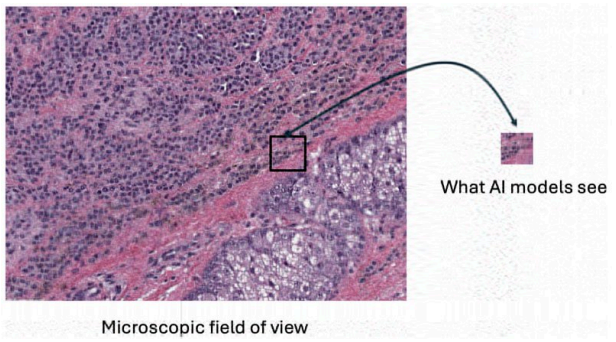

Figure 3: The field of view in light microscopy is traditionally quite large—approximately 2000 × 1500 pixels—and becomes vastly larger in whole-slide images (WSIs). In contrast, most AI models operate on small image patches, typically around 224 × 224 pixels.

Patch Size and Field-of-View Mismatch – A critical but widely overlooked issue is the patch-size mismatch between ViT architectures and diagnostic field of view. Most FMs use 224×224-pixel patches, a convention inherited from ImageNet. Yet even low-end educational microscopes produce 2048×1536-pixel (3 MP) views sufficient for diagnostic teaching (Fig. 3). Such small tiles capture fine-grained micro-texture but fail to represent mesoscale tissue architecture, glandular context, or stromal organization. This design choice prioritizes computational convenience—perhaps even favoring rapid experimentation and publication—over biological realism. Consequently, many foundation models end up encoding superficial texture statistics rather than diagnostically meaningful morphology. To bridge this gap, adaptive or hierarchical patching strategies combined with multi-scale attention mechanisms are urgently needed to model both local patterns and global structural context.

Vision Transformers (ViTs) split images into small patches (e.g., 16×16) to convert them into manageable token sequences for self-attention. This is computationally efficient but semantically costly: the model initially loses the global spatial structure (paramount for tissue morphology) and must learn it back from data. In pathology, where diagnostic meaning resides in multimagnification architecture, this design leads to **models that see textures but not tissues**.

4. Synthesis: A Paradigm Mismatch

These failures reveal a mismatch between the foundation-model paradigm—built on abundance and homogeneity—and pathology, characterized by scarcity, heterogeneity, and context dependence. Scaling existing architectures without domain adaptation amplifies fragility rather than resolving it.

5. Ethical Considerations

The rapid proliferation of foundation model publications in pathology has created both excitement and concern within the scientific community. Alongside legitimate progress, several ethical challenges have emerged that undermine transparency, reproducibility, and trust in the field.

Journals and Editors – Editorial practices play a pivotal role in shaping scientific integrity. An emerging concern is the tendency of some journals or conference committees to invite non-experts or overly sympathetic reviewers, particularly in competitive or high-visibility areas such as foundation models. This can lead to *insufficiently critical peer review* and the publication of technically or conceptually weak work. A related issue is publication bias—the reluctance to publish "negative" or non-confirmatory results. Studies reporting low performance, instability, or bias in foundation models often face editorial resistance, skewing the literature toward overstated success and suppressing valuable evidence on limitations. This imbalance reinforces unrealistic expectations and slows methodological self-correction.

Researchers and Authors – At the research level, ethical responsibility extends beyond dataset curation and model evaluation to the transparency of reporting. Some studies may employ undisclosed training "tricks" or hyperparameter tuning strategies (omitting details), obscuring the true source of performance gains and limiting reproducibility. Similarly, non-rigorous or selectively chosen metrics—for example, reporting only accuracy and AUC values while ignoring F1 scores, calibration or fairness—can may present a misleading picture of model capability. Another recurring problem is potential "data leakage", whether through overlapping patient samples, improper cross-validation, or hidden site bias. Such leakage can artificially inflate results, leading to spurious claims of generalization. In medical AI, where trust and safety are paramount, this constitutes not just methodological carelessness but an ethical lapse, as it risks translating flawed models into clinical claims.

Healthcare Management and Premature Adoption – A growing ethical risk lies beyond the research community: the premature adoption of foundation models by healthcare managers, administrators, or policy decision-makers eager to showcase "AI readiness" or institutional innovation. Driven by institutional pressure, marketing, or a desire to appear technologically progressive, some healthcare leaders may implement foundation models before their performance, generalizability, or bias have been adequately validated. Such enthusiasm, though often well-intentioned, can bypass the necessary clinical governance, risk assessment, and regulatory scrutiny. This **overzealous adoption** poses real dangers—misdiagnoses, inequitable treatment, and erosion of public trust—especially when AI systems are deployed under the illusion of universality or clinical maturity. Ethical stewardship therefore requires not only transparent science but also responsible decision-making at the administrative level, ensuring that AI implementation follows evidence, not hype.

6. Conclusion

Foundation models have undeniably transformed vision and language, yet human tissue is not the Internet. The disappointing accuracy, fragility, and ethical blind spots of current pathology FMs reveal a deeper conceptual misalignment between data-driven generalization and the structured reasoning of medicine. Their challenges—ranging from architectural oversimplification and data scarcity to publication bias, hidden leakage, and premature adoption—underscore that progress in computational pathology cannot be achieved by scaling technology alone. True advancement will require rethinking the very foundations: developing models that see tissue as pathologists do—multi-scale, contextual, biologically grounded, and transparently validated within rigorous ethical and clinical frameworks. Only then can foundation models evolve from impressive demonstrations of computation to trustworthy instruments of medical understanding.

References

[1] R. Bommasani et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

[2] Ce Zhou, Qian Li, Chen Li, Jun Yu, Yixin Liu, Guangjing Wang, Kai Zhang, Cheng Ji, Qiben Yan, Lifang He, et al. A comprehensive survey on pretrained foundation models: A history from bert to chatgpt. *International Journal of Machine Learning and Cybernetics*, pages 1–65, 2024.

[3] Muhammad Awais, Muzammal Naseer, Salman Khan, Rao Muhammad Anwer, Hisham Cholakkal, Mubarak Shah, Ming-Hsuan Yang, and Fahad Shahbaz Khan. Foundation models defining a new era in vision: a survey and outlook. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

[4] Mohsin Bilal, Manahil Raza, Youssef Altherwy, Anas Alsuhaibani, Abdulrahman Abduljabbar, Fahdah Almarshad, Paul Golding, Nasir Rajpoot, et al. Foundation models in computational pathology: A review of challenges, opportunities, and impact. *arXiv preprint arXiv:2502.08333*, 2025.

[5] Conghao Xiong, Hao Chen, and Joseph JY Sung. A survey of pathology foundation model: Progress and future directions. *arXiv preprint arXiv:2504.04045*, 2025.

[6] Saghir Alfasly, Ghazal Alabtah, Sobhan Hemati, Krishna Rani Kalari, Joaquin J Garcia, and

HR Tizhoosh. Validation of histopathology foundation models through whole slide image retrieval. *Scientific Reports*, 15(1):3990, 2025.

[7] Edwin D de Jong, Eric Marcus, and Jonas Teuwen. Current pathology foundation models are unrobust to medical center differences. *arXiv preprint arXiv:2501.18055*, 2025.

[8] Matouš Elphick, Samra Turajlic, and Guang Yang. Are the latent representations of foundation models for pathology invariant to rotation? *arXiv preprint arXiv:2412.11938*, 2024.

[9] Saghir Alfasly, Abubakr Shafique, Peyman Nejat, Jibran Khan, Areej Alsaafin, Ghazal Alabtah, and Hamid R Tizhoosh. Rotation-agnostic image representation learning for digital pathology. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11683–11693, 2024.

[10] Nita Mulliqi, Anders Blilie, Xiaoyi Ji, Kelvin Szolnoky, Henrik Olsson, Sol Erika Boman, Matteo Titus, Geraldine Martinez Gonzalez, Julia Anna Mielcarz, Masi Valkonen, et al. Foundation models–a panacea for artificial intelligence in pathology? *arXiv preprint arXiv:2502.21264*, 2025.

[11] J. Liang et al. Benchmarking foundation models as feature extractors for pathology tasks. *Nature Biomedical Engineering*, 9:1516, 2025.

[12] Yuntian Wang, Xilin Yang, Che-Yung Shen, Nir Pillar, and Aydogan Ozcan. Universal and transferable attacks on pathology foundation models. *arXiv preprint arXiv:2510.16660*, 2025.

[13] Petra Eretová, Helena Chaloupková, Marcela Hefferová, and Eva Jozífková. Can children of different ages recognize dog communication signals in different situations? *International journal of environmental research and public health*, 17(2):506, 2020.

[14] Alfredo F Pereira and Linda B Smith. Developmental changes in visual object recognition between 18 and 24 months of age. *Developmental science*, 12(1):67–80, 2009.

[15] David H Wolpert and William G Macready. No free lunch theorems for optimization. *IEEE transactions on evolutionary computation*, 1(1):67–82, 2002.

[16] Eugenio Piasini, Shuze Liu, Pratik Chaudhari, Vijay Balasubramanian, and Joshua I Gold. How occam's razor guides human decision-making. *bioRxiv*, pages 2023–01, 2025.

[17] Dong Li, Guihong Wan, Xintao Wu, Xinyu Wu, Ajit J Nirmal, Christine G Lian, Peter K Sorger, Yevgeniy R Semenov, and Chen Zhao. A survey on computational pathology foundation models: Datasets, adaptation strategies, and evaluation tasks. *arXiv preprint arXiv:2501.15724*, 2025.