A Survey on Computational Pathology Foundation Models

Dong Li[1] **Guihong Wan**[2] **Xintao Wu**[3] **Xinyu Wu**[1] **Yi He**[4]
Zhong Chen[5] **Ninghui Hao**[2] **Chen Zhao**[1]
[1]Baylor University [2]Harvard Medical School [3]University of Arkansas
[4]William & Mary [5]Southern Illinois University
{dong_li1, xinyu_wu1, chen_zhao}@baylor.edu
{guihong_wan@hsph, nhao1@mgh}.harvard.edu
xintaowu@uark.edu, yihe@wm.edu, zhong.chen@cs.siu.edu

Abstract

Computational pathology foundation models (CPathFMs) have emerged as a powerful approach for analyzing histopathological data, leveraging self-supervised learning to extract robust feature representations from unlabeled whole-slide images. These models, categorized into uni-modal and multi-modal frameworks, have demonstrated promise in automating complex pathology tasks such as segmentation, classification, and biomarker discovery. However, the development of CPathFMs presents significant challenges, such as limited data accessibility, high variability across datasets, the necessity for domain-specific adaptation, and the lack of standardized evaluation benchmarks. This survey provides a comprehensive review of CPathFMs in computational pathology, focusing on datasets, adaptation strategies, and evaluation tasks. We analyze key techniques, such as contrastive learning, masked image modeling and multi-modal integration, and highlight existing gaps in current research. Finally, we explore future directions from four perspectives for advancing CPathFMs. This survey serves as a valuable resource for researchers, clinicians, and AI practitioners, guiding the advancement of CPathFMs toward robust and clinically applicable AI-driven pathology solutions.

1 Introduction

Histopathology with hematoxylin and eosin (H&E) staining is central to disease diagnosis, prognosis, and treatment planning, particularly in oncology. Traditional histopathological analysis relies on manual examination of whole-slide images (WSIs) by pathologists, a process that is time-consuming, labor-intensive, and prone to inter-observer variability. The growing availability of digital WSIs has fueled the development of deep learning-based computational pathology (CPath) models that automate tasks such as tumor classification, biomarker discovery, and prognosis prediction using convolutional neural networks (CNNs) and vision transformers (ViTs). Recently, foundation models (FMs) have gained prominence in CPath [30]. Unlike conventional deep learning models that require large labeled datasets and are task-specific, computational pathology foundation models (CPathFMs) employ large backbones (often ViTs) pre-trained on diverse unlabeled histopathological data via self-supervised learning (SSL), and can be adapted to downstream tasks through transfer, few-shot, or zero-shot learning, thereby reducing reliance on expert annotations. Uni-modal CPathFMs learn from histopathological images alone, while multi-modal variants integrate images with clinical data from electronic health records (EHRs) to exploit complementary information. Despite promising advances, pre-training CPathFMs remains hindered by data scarcity, domain adaptation challenges, and inconsistent evaluation protocols.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: The 3rd Workshop on Imageomics: Discovering Biological Knowledge from Images Using AI.

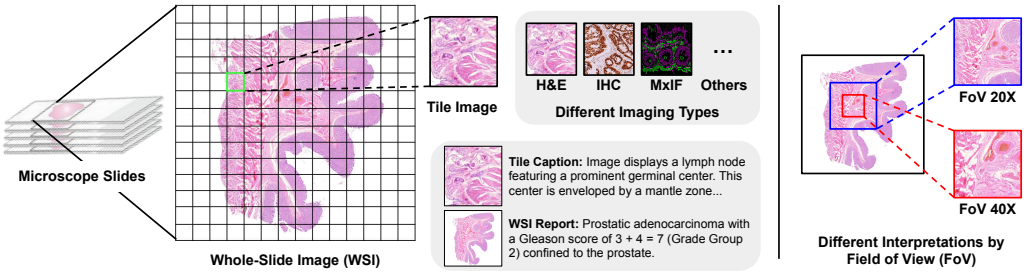

Figure 1: An illustrative example of data modalities and challenges in CPath. The figure illustrates different histopathology data types, including WSIs, tile images at multiple magnifications (Field of View, FoV), and imaging types (H&E, IHC, MxIF). These elements are critical for developing CPathFMs, highlighting the complexity of multi-scale image representation and domain-specific challenges.

Although the development of CPathFMs faces challenges related to data variability, adaptation, and evaluation, existing survey papers have not provided a sufficiently comprehensive overview of this field. Some works emphasize benchmarking but cover too few approaches and lack detailed summaries of pre-training datasets and evaluation tasks [5, 25, 29]. For example, while *Neidlinger et al.*[29] included a wide range of CPathFMs and datasets, their analysis of methods and datasets was not detailed. Similarly, *Ochi et al.*[30] and *Chanda et al.* [8] reviewed many CPathFMs, but their coverage of methods was neither comprehensive nor up to date, and they did not sufficiently discuss how these models are adapted to pathology or differentiate between adaptation strategies. Regarding evaluation, one survey merely listed tasks without providing a taxonomy, while the other offered an incomplete summary. In this survey, we aim to address these gaps by presenting a comprehensive review of CPathFMs, with particular emphasis on datasets, adaptation strategies, and evaluation tasks.

- Providing an in-depth analysis of existing pathology datasets and data curation used for pre-training CPathFMs, identifying key challenges in generalization.
- Systematically reviewing adaptation techniques in pre-training CPathFMs, covering 28 existing and up-to-date models across both uni-modal (image-based) and multi-modal (image-text) paradigms.
- For the first time, thoroughly summarizing evaluation tasks, categorizing them into six main perspectives for assessing pre-trained CPathFMs.
- Identifying key future research directions, offering insights into the challenges and opportunities for advancing CPathFM development.

2 Background

2.1 Computational Pathology (CPath)

CPath combines artificial intelligence, machine learning, and computer vision with digital pathology to support diagnosis, prognosis, and treatment planning. By leveraging whole-slide imaging (WSI) and deep learning, CPath enables scalable, automated analysis of histopathological data, reducing reliance on manual review and improving diagnostic consistency. Despite notable progress in CPath and CPathFMs, clinical deployment remains limited by challenges in performance, usability, and regulatory approval.

WSIs are gigapixel-scale digital scans of entire histology slides, capturing detailed tissue structures but requiring tiling into smaller patches for computational analysis. They serve as the foundation for both manual and automated review, with FDA approval making them standard in digital pathology workflows. As shown in Figure 1, CPath utilizes diverse data modalities, including WSIs, tile images at multiple magnifications, and imaging techniques such as H&E, immunohistochemistry (IHC), and multiplex immunofluorescence (MxIF). While H&E and IHC are routine in clinical practice, MxIF is mainly used in research. Multi-modal CPathFMs integrate these image types with clinical reports and tile captions, enhancing generalizability and interpretability for AI-assisted pathology.

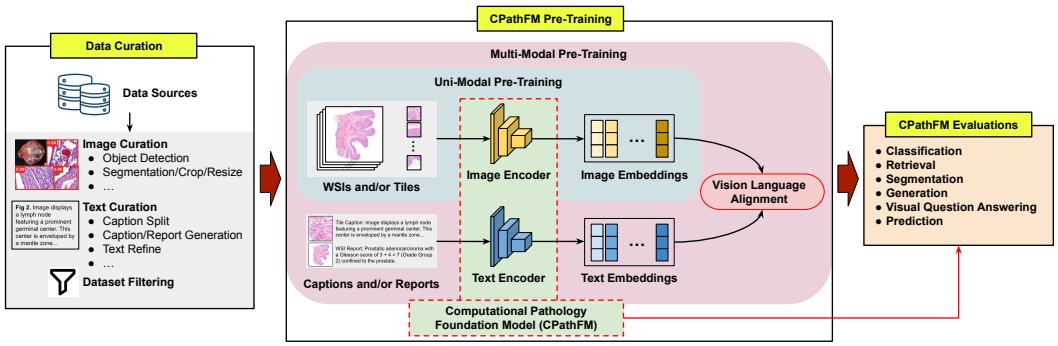

Figure 2: Overview of the pre-training pipeline for CPathFMs. The process involves data curation, including image curation, text curation, and dataset filtering, followed by uni-modal and multi-modal pre-training. The final CPathFMs are evaluated across multiple downstream tasks categorized into six main perspectives.

2.2 ViT-Based SSL Frameworks for FMs

The Vision Transformer (ViT) has become a cornerstone of foundation models due to its self-attention mechanism for capturing global dependencies, its patch-based tokenization suited for masked learning, and its scalability for large-scale multi-modal tasks. These properties have driven the design of self-supervised learning (SSL) frameworks that leverage ViT for robust representation learning in pathology and beyond.

Masked Image Modeling (MIM) is a key SSL approach. MAE [18] predicts masked patches to learn rich features, while BEiT [4] and BEiT v2 [32] refine this strategy with discrete tokenizers and knowledge distillation. In parallel, contrastive learning frameworks such as DINO [7] use self-distillation with a student–teacher paradigm, and DINOv2 [31] integrates iBOT [47] to combine MIM and contrastive learning for better generalization. Multi-modal extensions further expand ViT-based SSL: CLIP [33] aligns image and text encoders for zero-shot tasks, CoCa [45] fuses contrastive learning with caption generation, and BEiT-3 [38] enhances token prediction through multi-modal learning.

2.3 Challenges in Pre-training CPathFMs

While self-supervised contrastive learning has advanced CPathFMs, pre-training remains challenging due to limitations in data, adaptation, and evaluation, each of which directly affects model generalizability and clinical utility. As shown in Figure 2, the process involves dataset curation, training within an SSL framework, and evaluation on downstream tasks, all of which face practical obstacles.

Data scarcity and variability are major barriers. Large, diverse histopathology datasets are limited by ethical approvals and access restrictions, public datasets are rare and often single-institution, and gigapixel-scale WSIs pose storage and computational burdens. Annotation is costly and time-consuming, while variability in staining, magnification, and tissue structures introduces domain shifts, compounded by severe class imbalance. Beyond data, adapting models to heterogeneous pathology tasks is difficult, as tile-based approaches fragment global tissue context and current architectures struggle with multi-scale learning. Evaluation further complicates progress, as tasks span classification, retrieval, segmentation, and generation, yet lack standardized benchmarks and consistent protocols, making systematic comparison across datasets and institutions elusive.

3 Pre-training Datasets in CPathFMs

Although early CPathFMs used relatively small and homogeneous pre-training datasets, recent studies have shown that higher quality, larger scale, and more diverse pathology pre-training datasets are more beneficial for adapting the foundation models trained on natural image datasets or existing SSL frameworks to the pathology domain [50]. Therefore, summarizing the datasets used for pre-training CPathFMs can provide valuable insight into requirements for future research on CPathFMs. Appendix Table 2 provides a summary of pre-training datasets utilized by each method discussed in Section 4.

Table 1: Overview of architecture and adaptation strategies of CPathFMs

	Model	Reference	SSL Framework	Backbone [# Param.]†		Input	Pre-training Strategy‡		Model
				Vision	Language	Images	Vision	Language	Availability
Uni-modal	CTransPath	[39]	MoCo v3*	Swin Transformer [28M]	-	Tiles	S	-	✓
	REMEDIS	[3]	SimCLR	ResNet-152 (2×) [232M]	-	Tiles	D	-	✓
	Lunit DINO	[24]	DINO	ViT-S/(8,16) [22M]	-	Tiles	S	-	✓
	Phikon	[16]	iBOT	ViT-B/16 [86M]	-	Tiles	S	-	✗
	Virchow	[37]	DINOv2	ViT-H/14 [632M]	-	Tiles	S	-	✓
	-	[6]	DINO MAE	ViT-S [22M], ViT-B [86M] ViT-L [307M]	- -	Tiles	S	- -	✗
	RudolfV	[15]	DINOv2	ViT-L/14 [307M]	-	Tiles	D	-	✗
	Kaiko	[1]	DINO DINOv2	ViT-S/(8,16) [22M], ViT-B/(8,16) [86M] ViT-L/14 [307M]	- -	Tiles	D	-	✓
	PLUTO	[23]	DINOv2*	FlexiViT-S [22M]	-	Tiles	S	-	✓
	GigaPath	[43]	DINOv2*	ViT-G/14 [1.1B] & LongNet [125M]	-	WSIs	S, S	-	✓
	Hibou	[28]	DINOv2	ViT-B/14 [86M], ViT-L/14 [307M]	-	Tiles	S	-	✓
	BEPH	[44]	BEiTv2	ViT-B/16 [86M] & VQ-KD [86M]	-	Tiles	D, D	-	✓
	GPFM	[27]	DINOv2*	ViT-L [307M]	-	Tiles	S	-	✓
	Virchow2	[50]	DINOv2*	ViT-H/14 [632M]	-	Tiles	S	-	✓
	Phikon-v2	[17]	DINOv2	ViT-L/16 [307M]	-	Tiles	S	-	✗
	UNI	[9]	DINOv2	ViT-L/16 [307M]	-	Tiles	S	-	✓
	H-optimus-0	[34]	DINOv2	ViT-G/14 [1.1B]	-	Tiles	S	-	✓
	Atlas	[2]	DINOv2	ViT-H/14 [632M]	-	Tiles	D	-	✗
Multi-modal	PLIP	[19]	CLIP	ViT-B/32 [86M]	Transformer Layers [63M]	Tiles	D	D	✓
	PathCLIP	[36]	CLIP	ViT-B/32 [86M]	Transformer Layers [63M]	Tiles	D	D	✗
	QuiltNet	[20]	CLIP	ViT-B/(16,32) [86M]	GPT-2 [1.5B] & PubMedBERT [100M]	Tiles	D	D	✓
	CONCH	[26]	CoCa, iBOT	ViT-B/16 [86M]	Transformer Layers [∼86M]	Tiles	S	D	✓
	PRISM	[35]	CoCa	ViT-H/14 [632M] & Perceiver Net. [105M]	BioGPT [345M, 172M]	WSIs	F, S	D, F	✓
	CHIEF	[40]	CLIP*	Swin Transformer [28M]	Transformer Layers [63M]	WSIs	D	D	✓
	KEP	[49]	CLIP*	ViT-B/(16,32) [86M]	PubMedBERT [100M]	Tiles	D	S	✓
	TITAN	[14]	CoCa, iBOT*	ViT-B/16 [86M] & ViT-S [22M]	Transformer Layers [∼86M]	WSIs	F, S	D	✓
	KEEP	[48]	CLIP*	ViT-L [307M]	PubMedBERT [100M]	Tiles	D	S	✓
	MUSK	[42]	CoCa*, BEiT-3	V-FFN [202M] Shared Attention Layers [202M]	L-FFN [202M]	Tiles	S	S	✓

* Made domain-specific improvements or extensions to the SSL framework for pathology.

† For simplicity, we have streamlined some expressions. For example, "/8" denotes a patch size of 8×8 pixels, and "/(8,16)" represents "/8" and "/16", respectively. The **orange** color in [ ] represents the parameters that are being trained or tuned, while the **blue** color represents the frozen parameters.

‡ Pre-training strategies: **F**: Frozen, **S**: From Scratch, **D**: Domain-Specific Tuning.

Most CPathFMs construct large and diverse pre-training datasets from multiple sources, including public repositories such as TCGA [41], GTEx [12], and PMC OA [26], as well as internet-scale collections like Quilt-1M [20], or newly released datasets such as OpenPath [19] and PathCap [36]. The diversity of sources requires careful curation, typically involving subfigure detection and segmentation, image resizing, text refinement with LLMs, alignment of figures with captions or reports, and filtering to retain relevant pathology data. In terms of data types, uni-modal CPathFMs generally rely on WSIs and extracted tiles, while multi-modal models use task-specific inputs, with tile–caption pairs supporting tile-level training and WSI–report pairs used at the slide level; for example, CHIEF [40] employed anatomical site labels as textual features to construct WSI–text pairs.

4 Adaptation Strategies in CPathFMs

SSL has been widely applied in the development of CPathFMs to address the lack of labels. These models typically adapt SSL frameworks that have proven successful in natural images, and perform pre-training on carefully curated pathology datasets. Depending on the type of pathology data they used, these approaches can be categorized into uni-modal and multi-modal methods, as introduced in Table 1.

4.1 Uni-Modal CPathFMs

Uni-modal CPathFMs are generally trained on large, domain-specific pathology datasets using SSL frameworks to learn robust representations of pathological images without labeled data. Although there are some MIM-based methods, self-supervised contrastive learning methods play a dominant

role. Similar to the development of contrastive learning in natural images, CPathFMs were initially proposed within the MoCo [11] and SimCLR [10] frameworks. Following a transition through the DINO, DINOv2 was established as the leading framework, serving as the foundation for numerous subsequent studies.

DINO-based CPathFMs. As a successful application of SSL on ViT, DINO has been adopted as a framework for training CPathFMs. *Campanella et al.,* [6] compared the performance of DINO and MAE on different scales of pathology datasets, ultimately demonstrating the superiority of DINO for pre-training CPathFMs. *Kang et al.,* [24] focused on domain-aligned pre-training and proposed data augmentation and curation strategies specifically for pathological images.

DINOv2-based CPathFMs. Most studies using DINOv2, such as UNI [9], focus on larger ViT models and diverse pre-training datasets, with RudolfV [15] incorporating pathologist knowledge in dataset construction. Some methods adapt DINOv2 to pathology tasks: Kaiko [1] introduces Online Patching for high-throughput patch extraction, Virchow2 [50] replaces the entropy estimator with KDE, PLUTO [23] augments the loss with MAE and Fourier terms, and GPFM [27] builds a unified framework via Expert Knowledge Distillation. Distinct from these, GigaPath [43] targets whole-slide representation by treating tiles as visual tokens learned with DINOv2 and feeding them into LongNet [13] with Dilated Attention for efficient slide-level modeling.

Other Uni-Modal CPathFMs. While the majority of uni-modal methods focus on DINO and DINOv2, some methods employ other SSL frameworks. CTransPath [39] adds a branch to MoCov3 to generate queries that retrieve semantically similar samples from the memory bank as positive samples, thus guiding the network's training with a semantically relevant contrastive loss. REMEDIS [3] transfers a ResNet model, pre-trained on large-scale natural images, to the SimCLR framework for self-supervised training on pathological images. Additionally, Phikon [16] and BEPH [44] directly train a ViT model within the MIM-based SSL framework iBOT and BEiTv2, respectively.

4.2 Multi-Modal CPathFMs

Multi-modal CPathFMs enhance the model's understanding of pathological images by aligning paired image-text data under the visual-language multi-modal SSL frameworks, such as CLIP and CoCa. These methods typically train pre-trained uni-modal modules using uni-modal SSL frameworks before performing joint visual-language pre-training, which has been shown to improve the performance of downstream tasks [26].

CLIP-based CPathFMs. The success of CLIP on natural images has motivated its adaptation to pathology, where paired histopathological images and textual descriptions (*e.g.,* reports and annotations) enhance model interpretability. PLIP [19], PathCLIP [36], and QuiltNet [20] fine-tune pre-trained CLIP models on tile–caption datasets. CHIEF [40] extends this by encoding tile sequences with CTransPath to obtain WSI-level features and combining them with anatomical site information encoded by CLIP's text encoder for richer multi-modal representations. To integrate domain knowledge, Zhou *et al.* [49] introduce a pathology knowledge graph (KG) to guide visual–language pretraining, while KEEP [48] builds a disease KG and uses knowledge-guided dataset structuring to generate tile–caption pairs, incorporating positive mining and robust negative sampling strategies.

CoCa-based CPathFMs. The CoCa framework, with its multi-modal decoder, strengthens the cross-modal capabilities of CPathFMs and has been adopted in several recent models. CONCH [26] and PRISM [35] both pre-train an image encoder on pathology datasets using iBOT and DINOv2, respectively, before joint visual–language training with CoCa; PRISM further extends to the WSI-level via a Perceiver network [21] and incorporates clinical reports. MUSK [42] separately trains image and text encoders with masked data modeling in BEiT-3, then aligns them under CoCa. Building on these, TITAN [14] develops a multi-modal whole-slide foundation model, training a slide encoder in three stages: pre-training with iBOT and positional encoding, followed by CoCa-based training at both tile- and WSI-levels to enable comprehensive vision–language understanding.

5 Evaluation Tasks

CPathFMs do not target a specific task during the pre-training phase. Instead, a wide range of evaluation tasks are employed after pre-training to assess the model's ability to extract features from pathology data. These tasks are diverse, and the evaluation tasks for each CPathFM are not

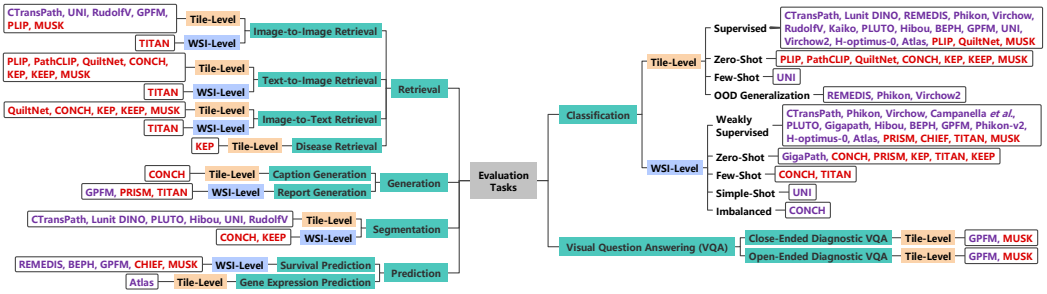

Figure 3: Taxonomy of evaluation tasks for pre-trained CPathFMs. Uni-modal and multi-modal CPathFMs are highlighted in **purple** and **red**, respectively.

standardized, making it challenging to establish a unified benchmark for CPathFMs. Therefore, we provide a summary of the evaluation tasks along with the CPathFMs performing them, as illustrated in Figure 3. We first categorized the evaluation tasks into six major perspectives based on their application objectives, followed by a further subdivision according to their specific objectives (e.g., focusing on tile-level or WSI-level). On this basis, we also considered variations in task settings (*e.g.*, supervised or zero-shot learning). Finally, we summarized which CPathFMs were used to evaluate each type of task.

Evaluation tasks in CPath cover a wide range, with classification being the most common. These tasks include cancer subtyping, biomarker detection, and mutation prediction, studied at both tile- and WSI-level under supervised, few-shot, and zero-shot settings. WSI-level classification is often weakly supervised with only global annotations, requiring aggregator networks to combine tile-level features. Beyond accuracy, models are also evaluated on out-of-distribution generalization across institutions, staining protocols, and rare disease settings. Other task types include retrieval, generation, segmentation, prediction, and VQA, many of which assess cross-modal capabilities such as aligning images with text. Some models (e.g., Virchow, RudolfV, BEPH, PLIP) also incorporate representation analysis through dimensionality reduction and clustering to qualitatively assess learned features.

6 Future Directions

Future research on CPathFMs should focus on improving their trustworthiness, extending to new imaging modalities, advancing multi-modal reasoning, and establishing standardized evaluation. Building trustworthy CPathFMs requires fairness, explainability, security, and transparency to ensure safe clinical deployment. Expanding to multiplex immunofluorescence (MxIF) imaging can provide richer insights into the tumor microenvironment but demands solutions for its high dimensionality and complex signal processing. Developing WSI-level multimodal large language models (MLLMs) for pathology VQA could enable context-aware diagnostics by integrating WSIs with clinical text, captions, and reports. Finally, standardized benchmarking datasets and evaluation metrics are needed to ensure consistent assessment of robustness, fairness, and clinical utility. A detailed discussion of these directions is provided in the Appendix Section B.

7 Conclusion

Computational pathology foundation models have emerged as a powerful approach for analyzing histopathological data, potentially playing a role in the development of robust and clinically applicable AI-driven pathology solutions. This survey provides a review of existing computational pathology foundation models, examining challenges in pre-training datasets, adaptation strategies, and evaluation tasks, while offering a comparative analysis of their strengths and limitations. Finally, we have identified key research gaps and proposed potential directions for future advancements.

References

[1] N. Aben, E. D. de Jong, I. Gatopoulos, N. Känzig, M. Karasikov, et al. Towards large-scale training of pathology foundation models. *arXiv:2404.15217*, 2024.

[2] M. Alber, S. Tietz, J. Dippel, T. Milbich, T. Lesort, P. Korfiatis, et al. A novel pathology foundation model by mayo clinic, charit\'e, and aignostics. *arXiv:2501.05409*, 2025.

[3] S. Azizi, L. Culp, J. Freyberg, B. Mustafa, et al. Robust and data-efficient generalization of self-supervised machine learning for diagnostic imaging. *Nature Biomedical Engineering*, 2023.

[4] H. Bao, L. Dong, S. Piao, and F. Wei. Beit: Bert pre-training of image transformers. *arXiv:2106.08254*, 2021.

[5] G. Campanella, S. Chen, R. Verma, J. Zeng, et al. A clinical benchmark of public self-supervised pathology foundation models. *arXiv:2407.06508*, 2024.

[6] G. Campanella, C. Vanderbilt, and T. Fuchs. Computational pathology at health system scale–self-supervised foundation models from billions of images. In *AAAI 2024 Spring Symposium on Clinical Foundation Models*, 2024.

[7] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, 2021.

[8] D. Chanda, M. Aryal, N. Y. Soltani, and M. Ganji. A new era in computational pathology: A survey on foundation and vision-language models. *arXiv:2408.14496*, 2024.

[9] R. J. Chen, T. Ding, M. Y. Lu, D. F. Williamson, G. Jaume, A. H. Song, B. Chen, A. Zhang, D. Shao, M. Shaban, et al. Towards a general-purpose foundation model for computational pathology. *Nature Medicine*, 2024.

[10] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. A simple framework for contrastive learning of visual representations. In *ICML*. PMLR, 2020.

[11] X. Chen, S. Xie, and K. He. An empirical study of training self-supervised vision transformers. In *ICCV*, 2021.

[12] G. Consortium, K. G. Ardlie, D. S. Deluca, A. V. Segrè, T. J. Sullivan, T. R. Young, E. T. Gelfand, C. A. Trowbridge, J. B. Maller, T. Tukiainen, et al. The genotype-tissue expression (gtex) pilot analysis: multitissue gene regulation in humans. *Science*, 2015.

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

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

[15] J. Dippel, B. Feulner, T. Winterhoff, T. Milbich, et al. Rudolfv: a foundation model by pathologists for pathologists. *arXiv:2401.04079*, 2024.

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

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

[18] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, 2022.

[19] Z. Huang, F. Bianchi, M. Yuksekgonul, T. J. Montine, and J. Zou. A visual–language foundation model for pathology image analysis using medical twitter. *Nature medicine*, 2023.

[20] W. Ikezogwo, S. Seyfioglu, F. Ghezloo, D. Geva, F. Sheikh Mohammed, P. K. Anand, R. Krishna, and L. Shapiro. Quilt-1m: One million image-text pairs for histopathology. *NeurIPS*, 2024.

[21] A. Jaegle, F. Gimeno, A. Brock, O. Vinyals, A. Zisserman, and J. Carreira. Perceiver: General perception with iterative attention. In *ICML*. PMLR, 2021.

[22] G. Jaume, L. Oldenburg, A. Vaidya, R. J. Chen, D. F. Williamson, T. Peeters, A. H. Song, and F. Mahmood. Transcriptomics-guided slide representation learning in computational pathology. In *CVPR*, 2024.

[23] D. Juyal, H. Padigela, C. Shah, D. Shenker, et al. Pluto: Pathology-universal transformer. *arXiv:2405.07905*, 2024.

[24] M. Kang, H. Song, S. Park, D. Yoo, and S. Pereira. Benchmarking self-supervised learning on diverse pathology datasets. In *CVPR*, 2023.

[25] J. Lee, J. Lim, K. Byeon, and J. T. Kwak. Benchmarking pathology foundation models: Adaptation strategies and scenarios. *arXiv:2410.16038*, 2024.

[26] M. Y. Lu, B. Chen, D. F. Williamson, R. J. Chen, I. Liang, T. Ding, G. Jaume, I. Odintsov, L. P. Le, G. Gerber, et al. A visual-language foundation model for computational pathology. *Nature Medicine*, 2024.

[27] J. Ma, Z. Guo, F. Zhou, Y. Wang, et al. Towards a generalizable pathology foundation model via unified knowledge distillation. *arXiv:2407.18449*, 2024.

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

[29] P. Neidlinger, O. S. El Nahhas, H. S. Muti, T. Lenz, M. Hoffmeister, H. Brenner, et al. Benchmarking foundation models as feature extractors for weakly-supervised computational pathology. *arXiv:2408.15823*, 2024.

[30] M. Ochi, D. Komura, and S. Ishikawa. Pathology foundation models. *arXiv:2407.21317*, 2024.

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

[32] Z. Peng, L. Dong, H. Bao, Q. Ye, and F. Wei. Beit v2: Masked image modeling with vector-quantized visual tokenizers. *arXiv:2208.06366*, 2022.

[33] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*. PMLR, 2021.

[34] C. Saillard, R. Jenatton, F. Llinares-López, Z. Mariet, D. Cahané, E. Durand, and J.-P. Vert. H-optimus-0, 2024.

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

[36] Y. Sun, C. Zhu, S. Zheng, K. Zhang, L. Sun, Z. Shui, Y. Zhang, H. Li, and L. Yang. Pathasst: A generative foundation ai assistant towards artificial general intelligence of pathology. In *AAAI*, 2024.

[37] E. Vorontsov, A. Bozkurt, A. Casson, G. Shaikovski, M. Zelechowski, et al. Virchow: A million-slide digital pathology foundation model. *arXiv:2309.07778*, 2023.

[38] W. Wang, H. Bao, L. Dong, J. Bjorck, Z. Peng, Q. Liu, K. Aggarwal, O. K. Mohammed, et al. Image as a foreign language: Beit pretraining for all vision and vision-language tasks. *arXiv:2208.10442*, 2022.

[39] X. Wang, S. Yang, J. Zhang, M. Wang, J. Zhang, W. Yang, J. Huang, and X. Han. Transformer-based unsupervised contrastive learning for histopathological image classification. *Medical image analysis*, 2022.

[40] X. Wang, J. Zhao, E. Marostica, W. Yuan, J. Jin, J. Zhang, R. Li, H. Tang, K. Wang, Y. Li, et al. A pathology foundation model for cancer diagnosis and prognosis prediction. *Nature*, 2024.

[41] J. N. Weinstein, E. A. Collisson, G. B. Mills, K. R. Shaw, B. A. Ozenberger, K. Ellrott, I. Shmulevich, C. Sander, and J. M. Stuart. The cancer genome atlas pan-cancer analysis project. *Nature genetics*, 2013.

[42] J. Xiang, X. Wang, X. Zhang, Y. Xi, F. Eweje, Y. Chen, Y. Li, C. Bergstrom, M. Gopaulchan, T. Kim, et al. A vision–language foundation model for precision oncology. *Nature*, 2025.

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

[44] Z. Yang, T. Wei, Y. Liang, X. Yuan, R. Gao, et al. A foundation model for generalizable cancer diagnosis and survival prediction from histopathological images. *bioRxiv*, 2024.

[45] J. Yu, Z. Wang, V. Vasudevan, L. Yeung, M. Seyedhosseini, and Y. Wu. Coca: Contrastive captioners are image-text foundation models. *arXiv:2205.01917*, 2022.

[46] S. Zhang, Y. Xu, N. Usuyama, H. Xu, J. Bagga, et al. Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. *arXiv:2303.00915*, 2023.

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

[48] X. Zhou, L. Sun, D. He, W. Guan, R. Wang, and L. o. Wang. A knowledge-enhanced pathology vision-language foundation model for cancer diagnosis. *arXiv:2412.13126*, 2024.

[49] X. Zhou, X. Zhang, C. Wu, Y. Zhang, W. Xie, and Y. Wang. Knowledge-enhanced visual-language pretraining for computational pathology. In *ECCV*. Springer, 2024.

[50] E. Zimmermann, E. Vorontsov, J. Viret, A. Casson, M. Zelechowski, et al. Virchow2: Scaling self-supervised mixed magnification models in pathology. *arXiv:2408.00738*, 2024.

A Summary Table for Pre-training Datasets in CPathFMs

Table 2 provides a summary of pre-training datasets utilized by each method discussed in Section 4.

Table 2: Statistics of pathology datasets used for pre-training CPathFMs. The "-" symbol represents the absence of relevant information.

	Reference	Data Description†		Input Image Size	Field of View	Staining Types			Data Sources		Corresponding
		# WSIs	# Tiles		(FoV)	H&E	IHC	Others	Public	Private	Method
Uni-modal	[39]	32.2K	15.6M	1024×1024	20×	✓	✗	✗	TCGA, PAIP	-	CTransPath
	[3]	29.0K	50.0M	224×224	20×	✓	✗	✗	TCGA	JFT54	REMEDIS
	[24]	36.7K	32.6M	512×512	{20, 40}×	✓	✗	✗	TCGA	TULIP	Lunit DINO
	[16]	6.1K	43.4M	224×224	20×	✓	✗	✗	TCGA	-	Phikon
	[37]	1.5M	2.0B	224×224	20×	✓	✗	✗	-	MSKCC	Virchow
	[6]	423K	1.6B, 3.2B	224×224	20×	✓	✗	✗	-	MSHS	-
	[15]	134K	1.2B	256×256	{20, 40, 80}×	✓	✓	✓	TCGA	Proprietary	RudolfV
	[1]	29.0K	256M	256×256	{5, 10, 20, 40}×	✓	✗	✗	TCGA	-	Kaiko
	[23]	158.8K	195M	224×224	{20, 40}×	✓	✓	✓	TCGA, *etc.*	Proprietary	PLUTO
	[43]	171K	1.4B	256×256	20×	✓	✓	✗	-	PHS	GigaPath
	[28]	1.1M	512M, 1.2B	224×224	20×	✓	✓	✓	-	Proprietary	Hibou
	[44]	11.7K	11.7M	224×224	20×	✓	✗	✗	TCGA	-	BEPH
	[27]	72.3K	190.2M	512×512	-	✓	✗	✗	TCGA, GTEx, *etc.*	-	GPFM
	[50]	3.1M	2.0B	392×392	{5, 10, 20, 40}×	✓	✓	✗	-	MSKCC	Virchow2
	[17]	58.0K	456M	224×224	20×	✓	✓	✓	TCGA, GTEx, *etc.*	Proprietary×4	Phikon-v2
	[9]	100K	100M	256×256, 512×512	20×	✓	✗	✗	GTEx	MGH, BWH	UNI
	[34]	500K+	100M+	-	-	✓	✗	✗	-	Proprietary	H-optimus-0
	[2]	1.2M	3.4B	256×256	{5, 10, 20, 40}×	✓	✓	✓	-	Proprietary	Atlas
Multi-modal	[19]	208K Tile-Caption Pairs		224×224	-	✓	✓	-	Twitter, PathLAION	-	PLIP
	[36]	207K Tile-Caption Pairs		-	-	✓	✓	✗	PMC OA	LBC	PathCLIP
	[20]	438K Tiles and 802K Captions		Avg. 882×1648	{10-40}×	✓	✓	-	YouTube, Twitter, *etc.*	-	QuiltNet
	[26]	21K WSIs from 16M Tile Images 1.17M Tile-Caption Pairs		256×256 448×448	20× -	✓ ✓	- ✓	- ✓	- PMC OA	Proprietary EDU	CONCH
	[35]	587K WSIs and 195K Reports		224×224	20×	✓	✗	✗	TCGA	Proprietary	PRISM
	[40]	60K WSI-Label Pairs		256×256	10×	✓	✗	✗	TCGA, GTEx, *etc.*	Proprietary	CHIEF
	[49]	A KG with 50.5K Pathology Attributes 576.6K, 138.9K Tile-Caption Pairs		- 224×224	- -	- ✓	- ✓	- -	OncoTree, *etc.* Quilt-1M, OpenPath	- -	KEP
	[14]	336K WSIs 423K Tile-Caption Pairs 183K WSI-Report Pairs		512×512 8192×8192 32768×32768	20× 20× -	✓ ✓ ✓	✓ ✓ ✓	✗ ✗ ✗	GTEx GTEx GTEx	Proprietary Proprietary Proprietary	TITAN
	[48]	A KG with 139K Disease Attributes 143K Tile-Caption Pairs		- 224×224	- -	- ✓	- ✓	- -	DO, UMLS Quilt-1M, OpenPath	- -	KEEP
	[42]	1B Text Tokens and 50M Tiles 1M Tile-Caption Pairs		384×384 384×384	{10, 20, 40}× 20×	-✓ ✓	✗ ✗	✗ ✗	PMC OA, TCGA Quilt-1M, PathCap	- -	MUSK

† The pre-training data for uni-modal CPathFMs primarily consists of the number of WSIs and tiles. However, the situation is more complex for multi-modal models, so we provide a textual description for clarification.

B Future Directions

As CPathFMs continue to evolve in recent years, several critical research directions can further enhance their reliability, applicability, and impact.

Trustworthy CPathFMs ensures fairness, explainability, security, and transparency. Fairness is especially crucial, as predicted outcomes should be independent of sensitive attributes, such as race, to avoid potential biases in clinical applications. Enhancing the explainability of CPathFMs is also essential to gaining the trust of pathologists and clinicians, as deep learning models often operate as black boxes. Furthermore, addressing security vulnerabilities in CPathFMs, such as adversarial attacks, is necessary to prevent manipulation of model predictions. Finally, transparency in model development, dataset curation, and evaluation procedures is crucial for reproducibility and regulatory approval, ensuring CPathFMs can be safely deployed in clinical workflows.

Developing CPathFMs for MxIF Imaging. Unlike H&E and IHC staining, MxIF captures spatial distributions of multiple biomarkers simultaneously, offering richer biological insights into the tumor microenvironment. However, training foundation models on MxIF images presents challenges, including higher dimensionality, complex signal processing, and the need for precise biomarker

alignment. Future research should focus on building CPathFMs that can effectively extract meaningful representations from MxIF data while addressing these computational challenges.

Development of WSI-Level MLLMs for Pathology VQA. A WSI-level MLLM would allow context-aware analysis of entire whole-slide images while integrating clinical reports, pathology captions, and other textual information. This could significantly improve AI-assisted diagnostics, enabling models to generate pathology reports, answer clinician queries, and assist in complex diagnostic decision-making.

Standardized Benchmarking Datasets and Evaluation Metrics for CPathFMs. The current landscape lacks a uniform set of evaluation metrics that can systematically compare different models across a wide range of pathology tasks. A standardized benchmark dataset incorporating diverse tissue types, staining methods, and multi-institutional sources would significantly enhance model generalization and comparability. Additionally, defining clear evaluation indicators would allow the research community to assess the robustness, fairness, and clinical utility of CPathFMs more effectively.

C Scope and Exclusions

This survey is centered on computational pathology foundation models (CPathFMs) developed primarily for histopathology image analysis, with a particular focus on models pre-trained on hematoxylin and eosin (H&E), immunohistochemistry (IHC), and multiplex immunofluorescence (MxIF) images. Our emphasis is on methods that build generalizable visual or vision–language representations from histopathological staining images and their associated clinical texts or captions.

In addition to the methods investigated in this work, there are other approaches that focus on multi-modal data beyond text and images, such as chest X-ray images [46], genetic sequences [22], *etc.,* which fall outside the scope of this survey focused on histopathology staining images. Moreover, some works focused on developing multi-modal large language models (MLLMs) as generative foundation AI assistants for pathologists [36] are not included, as these works typically align multi-modal CPathFMs mentioned above with existing LLMs. The emphasis of these models is on enhancing VQA capabilities rather than training a general feature extractor.