A Survey of Pathology Foundation Model: Progress and Future Directions

Conghao Xiong[1] , **Hao Chen**[2] and **Joseph J. Y. Sung**[3]
[1]Department of Computer Science and Engineering, The Chinese University of Hong Kong
[2]Department of Computer Science and Engineering and Department of Chemical and Biological Engineering, The Hong Kong University of Science and Technology
[3]Lee Kong Chian School of Medicine, Nanyang Technological University
chxiong21@cse.cuhk.edu.hk, jhc@cse.ust.hk, josephsung@ntu.edu.sg

Abstract

Computational pathology, which involves analyzing whole slide images for automated cancer diagnosis, relies on multiple instance learning, where performance depends heavily on the feature extractor and aggregator. Recent Pathology Foundation Models (PFMs), pretrained on large-scale histopathology data, have significantly enhanced both the extractor and aggregator, but they lack a systematic analysis framework. In this survey, we present a hierarchical taxonomy organizing PFMs through a top-down philosophy applicable to foundation model analysis in any domain: model scope, model pretraining, and model design. Additionally, we systematically categorize PFM evaluation tasks into slide-level, patch-level, multimodal, and biological tasks, providing comprehensive benchmarking criteria. Our analysis identifies critical challenges in both PFM development (pathology-specific methodology, end-to-end pretraining, data-model scalability) and utilization (effective adaptation, model maintenance), paving the way for future directions in this promising field. Resources referenced in this survey are available at https://github.com/BearCleverProud/AwesomeWSI.

1 Introduction

Computational **Path**ology (CPath), the computational analysis of patient specimens (*i.e.*, **W**hole **S**lide **I**mages, WSIs), is increasingly important due to the critical role of histopathology. For gigapixel WSIs, **M**ultiple **I**nstance **L**earning (MIL) is the de facto framework, involving WSI patch partitioning, feature extraction via pretrained neural networks, and feature aggregation into WSI-level features [Xiong *et al.*, 2024b]. Therefore, MIL performance hinges on two components: the pretrained neural network (*extractor*) and the *aggregator*.

Pathology **F**oundation **M**odels (PFMs), neural networks pretrained on extensive pathological data that can be directly leveraged for diverse downstream tasks without retraining, such as HIPT [Chen *et al.*, 2022] and UNI [Chen *et al.*, 2024], mark a paradigm shift for MIL. Conventionally, due to the lack of PFMs, ResNet-50 [He *et al.*, 2016] pretrained on ImageNet [Deng *et al.*, 2009] serves as the extractor [Xiong *et al.*, 2024a], but struggles with pathology-specific characteristics like minimal color variation, rotation-agnosticism, and hierarchical tissue organization. While limited labeled WSIs prevented supervised pretraining, **S**elf-**S**upervised **L**earning (SSL) enables PFMs that exhibit superior generalizability in morphology recognition. This overcomes natural image pretraining limitations, in which features mainly capture general visual attributes like edges and textures, enabling better performance on downstream tasks even with limited data.

Despite their potential, PFMs face multifaceted challenges: 1) most PFMs directly adopt natural image techniques, failing to cater to the discrepancy between pathology and natural images, indicating pathology-specific methodology remains underexplored; 2) MIL, as a two-stage pipeline, traps model training in local optima, while end-to-end training of WSIs requires prohibitive computational resources; 3) undefined model and data scaling bounds and resource constraints necessitate multi-institutional federated learning, demanding efficiency; and 4) the computational demands of PFMs impede deployment and maintenance, requiring continuous adaptation to evolving WSI technologies and pathological variants.

Recent surveys on PFMs have contributed significantly to the understanding of the field; however, these works either focus primarily on the impact of PFMs on the real world rather than technical investigations of them [Ochi *et al.*, 2025], or detail the previous efforts in this field without a systematic taxonomy for technical analysis and a systematic organization of the evaluation tasks of PFMs [Chanda *et al.*, 2024]. To address these critical gaps, we introduce a comprehensive and timely survey of the current landscape. We collected papers from high-impact journals, including Nature, Nature Medicine, Nature Biomedical Engineering, Medical Image Analysis, as well as top-tier conferences such as CVPR, ICML, and AAAI. Given the rapid evolution of the field, we also incorporated preprints from repositories such as arXiv, bioRxiv, and medRxiv, acknowledging that many influential works are still under review. In total, our survey includes 27 PFM papers, 12 of which are preprints that have not yet been accepted by peer-reviewed conferences or journals.

We present this survey with three primary contributions: 1) a hierarchical taxonomy organizing PFMs based on scope, training strategy, and design to enable holistic analysis, transferable to general vision FMs; 2) a comprehensive analysis of evaluation methodologies, examining their technical mer-

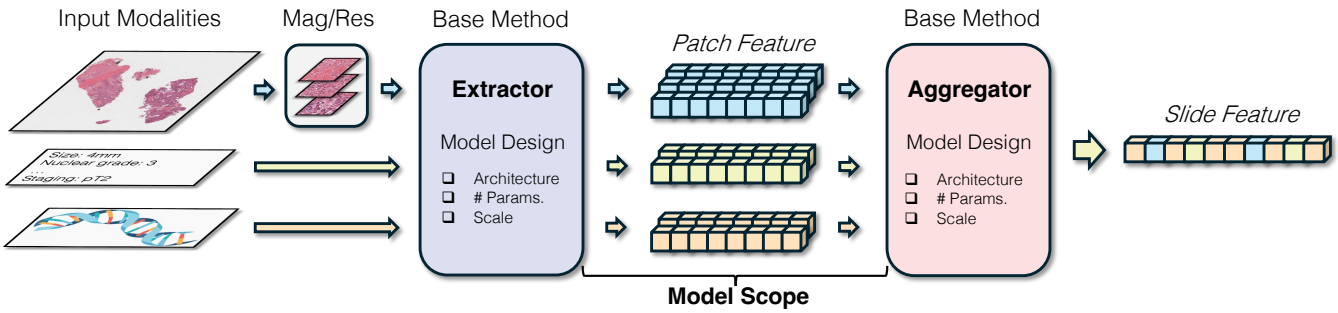

Figure 1: Schematic representation of our hierarchical taxonomy integrated within the MIL framework for PFMs.

its and limitations; and 3) a structured analysis of pathology-centric research challenges prioritizing underexplored directions. The manuscript is organized as follows: Section 2 formally formulates MIL and SSL; Section 3 introduces the proposed hierarchical taxonomy; Section 4 examines evaluation tasks for PFMs; Section 5 delineates future research directions in this field; and Section 6 concludes our survey.

2 Background and Problem Formulation

2.1 Multiple Instance Learning

In the MIL framework, a WSI is typically represented as a bag of N unordered instances (or patches). The central objective of MIL is to predict the WSI-level label $\hat{Y}$ using only the ground truth bag label Y as supervision, without access to the ground truth instance-level labels $\{y_i\}_{i=1}^N$. This setting reflects a common scenario in computational pathology, where obtaining slide-level annotations is feasible, but annotating individual patches is prohibitively expensive and impractical. The relationship between the bag label Y and the instance labels $\{y_i\}_{i=1}^N$ is typically defined under standard MIL assumptions such as the presence-based assumption, and can be formally expressed as [Xiong *et al.*, 2023],

$$Y = \begin{cases} 1 & \exists i, y_i = 1 \\ 0 & \forall i, y_i = 0 \end{cases}. \tag{1}$$

The implementation of MIL involves tessellating WSIs into non-overlapping patches $\boldsymbol{X} = \{\boldsymbol{x}_i\}_{i=1}^N \in \mathbb{R}^{N \times h \times w \times 3}$, with h, w standing for height and width, respectively. These patches undergo feature extraction through an extractor $\mathcal{M}_e(\cdot)$, generating corresponding features $\boldsymbol{Z} = \{\boldsymbol{z}_i\}_{i=1}^N \in \mathbb{R}^{N \times d}$, where each feature is computed as $\boldsymbol{z}_i = \mathcal{M}_e(\boldsymbol{x}_i)$, and d is the hidden dimension of the extractor. Subsequently, an aggregator $\mathcal{M}_g(\cdot)$ agglomerates these features to form a bag-level feature $\boldsymbol{h} = \mathcal{M}_g(\boldsymbol{Z})$ of the WSI, which finally serves as the input for the classification layer. Throughout aggregation, the extractor $\mathcal{M}_e(\cdot)$ usually remains frozen and non-trainable due to GPU memory constraints, while the aggregation network $\mathcal{M}_g(\cdot)$ is optimized during training. We refer readers unfamiliar with MIL to prior surveys for more details [Carbonneau *et al.*, 2018; Waqas *et al.*, 2024].

2.2 Self-supervised Learning

SSL leverages unlabeled data by automatically generating supervisory signals through pretext tasks [Ericsson *et al.*, 2022]. Given an input image $\boldsymbol{x}$, a transformation function $\mathcal{T}(\cdot)$ is applied to generate a modified version $\tilde{\boldsymbol{x}} = \mathcal{T}(\boldsymbol{x})$ and a corresponding pseudo-label $\tilde{y}$. An extractor $\mathcal{M}_e(\cdot)$ extracts features from $\tilde{\boldsymbol{x}}$ and generates a predicted label $\hat{y} = \mathcal{M}_e(\tilde{\boldsymbol{x}})$. The learning objective can be formalized as minimizing the difference between the predicted label $\hat{y}$ and the pseudo-label $\tilde{y}$. Common pretext tasks include contrastive learning, self-distillation, masked image modeling, *etc*., each designed to force the model to learn meaningful semantic features of the data. Through this process, the extractor can learn transferable features for downstream tasks on massive unlabeled data. We refer readers who are unfamiliar with SSL to prior surveys for more details [Ericsson *et al.*, 2022; Shurrab and Duwairi, 2022; Gui *et al.*, 2024].

3 Hierarchical Taxonomy

Our taxonomy systematically organizes PFMs through three interdependent dimensions and reflects a top-down design philosophy: 1) *Model Scope*: a categorization of the scope of the PFMs, differentiating between PFMs focused on extractors, aggregators, and both components; 2) *Model Pretraining*: a detailed examination of the spectrum of image-centric pretraining methods, including slide-level, patch-level, and multimodal techniques; and 3) *Model Design*: a rigorous analysis of architecture, categorizing PFMs according to their number of parameters and scale. This top-down structure enables systematic comparisons of PFMs, as shown in Table 1.

3.1 Model Scope

MIL consists of three parts: 1) patch partitioning, 2) feature extraction, and 3) feature aggregation. As patch partitioning has been well-established, MIL performance primarily depends on the extractor and aggregator. In addition, WSIs inherently exhibit hierarchical structures, where local histomorphological patterns captured by extractors and global hierarchical tissue organization modeled by aggregators jointly determine diagnostic accuracy. Therefore, we categorize PFMs based on their scope: *extractor-centric*, *aggregator-centric*, or *hybrid-centric*. The categorization of PFMs along this dimension is presented in the Model Scope column of Table 1.

Extractor-centric models constitute the predominant approach in PFM development, driven by two factors: the importance of high-quality features and the necessity to address domain shift brought by ImageNet-pretrained CNNs.

Model	Model Scope		Model Pretraining			Model Design		
	E.	A.	Input	Base Method	Mag/Res	Architecture	# Params.	Scale
CTransPath	✓	✗	H	MoCov3	10/224	Swin-T/14	28.3M	S
REMEDIS	✓	✗	H	SimCLR	Multi/224	ResNet-50	25.6M	S
HIPT	✓	✓	H	DINO	20/256,4096	ViT-S/16-XS/256	21.7/2.78M	S/XS
PLIP	✓	✗	P, T	CLIP	20/224	ViT-B/32	87M	B
CONCH	✓	✗	W, T	iBOT/CoCa	20/256	ViT/B-16	86.3M	B
Phikon	✓	✗	H	iBOT	20/224	ViT-S/B/L/16	21.7/85.8/307M	S/B/L
UNI	✓	✗	H	DINOv2	20/256,512	ViT-L/16	307M	L
Virchow	✓	✗	H	DINOv2	20/224	ViT-H/14	632M	H
SINAI	✓	✗	H	DINO/MAE	Unknown	ViT-S/L	21.7M/303.3M	S/L
CHIEF	✗	✓	H,T	Sup.+CLIP	10/224	CHIEF	1.2M	XS
Prov-GigaPath	✓	✓	H,I	DINOv2/MAE	20/256	ViT-g/14/LongNet	1.13B/85.1M	g/B
Pathoduet	✓	✗	H,I	MoCov3	40/256,20/1024	ViT-B/16	85.8M	B
RudolfV	✓	✗	W	DINOv2	20,40,80/256	ViT-L/14	304M	L
PLUTO	✓	✗	W	DINOv2	20,40/224	FlexiViT-S/16	22M	S
PRISM	✗	✓	H,T	CoCa	20/224	Perceiver	45.0M	S
TANGLE	✓	✓	H,G	iBOT/SimCLR	20/224	ViT-B/16/ABMIL	86.3/2.3M	B/XS
MUSK	✓	✗	H,T	MIM	10,20,40/384	BEiT-3	675M	H
BEPH	✓	✗	H	MIM	40/224	BEiTv2	192.55M	B
Hibou	✓	✗	W	DINOv2	Unknown	ViT-B/L/16	86.3/307M	B/L
mSTAR+	✓	✓	H,G,T	CLIP/ST	20/256	TransMIL/ViT-L	2.67/307M	XS/L
GPFM	✓	✗	H	UDK	40/512	ViT-L/14	307M	L
Virchow2G	✓	✗	W	DINOv2	5,10,20,40/224	ViT-G/14	1.9B	G
MADELEINE	✗	✓	W	CLIP	10,20/256	MH-ABMIL	5.0M	XS
Phikon-v2	✓	✗	W	DINOv2	20/224	ViT-L/16	307M	L
TITAN	✗	✓	W,T	iBOT/CoCa	20/8192	TITAN/TITAN$_V$	48.5/42.1M	S
KEEP	✓	✗	W,T	CLIP	20/224	UNI	307M	L
THREADS	✗	✓	H,D,R	CLIP	20/512	MH-ABMIL	11.3M	XS

Table 1: Systematic comparison of PFMs categorized based on our hierarchical taxonomy. Abbreviations used: Extractor (E.), Aggregator (A.), H&E (H), Patch (P), Text (T), WSIs with unspecified stains (W), IHC (I), Genomics (G), DNA (D), and RNA (R).

The role of the extractor aligns with established clinical practice, where pathologists emphasize cellular morphological analysis at the patch level. CTransPath [Wang *et al.*, 2022] pioneers the extractor training with a hybrid CNN-Transformer design through **S**emantic-**R**elevant **C**ontrastive **L**earning (SRCL) on 15 million patches. REMEDIS [Azizi *et al.*, 2023] demonstrates that the feature extraction capability of ResNet-50 is constrained by domain shift across different medical imaging domains. Various advancements, including Virchow [Vorontsov *et al.*, 2024] and SINAI [Campanella *et al.*, 2024], further stress the significance of robust extractors.

Aggregator-centric models play a vital role in slide-level tasks as they are the only trainable models under direct supervision of ground truth labels, yet they are relatively underexplored compared to the extractor. CHIEF [Wang *et al.*, 2024b], leveraging supervised pretraining with the anatomical site to create an anatomy-aware aggregator, first demonstrates the efficacy of aggregator pretraining. More recent research like MADELEINE [Jaume *et al.*, 2025], TITAN [Ding *et al.*, 2024], and THREAD [Vaidya *et al.*, 2025] utilizes multimodal data in aggregator pretraining with frozen patch features to enhance performance across downstream tasks. This paradigm shift reflects growing awareness that the aggregator critically impacts downstream task performance, particularly in low-resource clinical scenarios [Xu *et al.*, 2024a]. This observation aligns with transfer learning principles, wherein pretraining on large-scale datasets effectively alleviates downstream data scarcity challenges. However, empirical evidence also reveals that the pretrained CHIEF aggregator occasionally performs worse than linear probing of the extractor [Ding *et al.*, 2024], which is potentially attributable to the small model size when trained on a pretraining-scale dataset, or to the conflicts between domain bias and generic features. Consequently, further investigations are warranted to assess the advantages of pretrained larger aggregators.

Hybrid-centric models are PFMs that pretrain both the extractor and aggregator. Their advantage lies in full exploitation of the aggregator, as the aggregators can flexibly adapt to the extractor with pretraining-scale data. HIPT pioneers this approach through hierarchical pretraining of the first two layers of the extractor, excluding the last layer, which is substantiated through empirical performance. Similarly, Prov-GigaPath [Xu *et al.*, 2024a] pretrains a ViT extractor and a LongNet [Ding *et al.*, 2023] slide encoder; however, LongNet generates instance-level features rather than a single slide-level feature, necessitating integration of ABMIL [Ilse *et al.*, 2018] or non-parametric pooling strategies for slide-level tasks. TANGLE [Jaume *et al.*, 2024] pretrains both a ViT feature extractor and a transcriptomics-guided ABMIL aggregator. Finally, mSTAR [Xu *et al.*, 2024b] distinguishes itself as a fully-pretrained hybrid-centric model by an inverted pretraining sequence, contrasting with the conventional paradigm: first optimizing the multimodal aggregator, followed by pretraining the extractor with the aggregator.

Analysis of recent developments reveals two observations. First, research emphasis has progressively shifted from fea-

ture extractor pretraining toward aggregator pretraining, a transition potentially attributable to both the robust performance of existing extractors and the increasing awareness of aggregator significance, especially in limited-data scenarios. Second, current aggregators demonstrate a hierarchical dependency pattern, wherein each successive model builds upon the capabilities of prior models. For instance, TITAN utilizes features from CONCHv1.5, which in turn leverages UNI as its encoder, thereby forming a cascading performance dependency chain where the efficacy of TITAN is inherently contingent upon CONCHv1.5 and, by extension, UNI.

3.2 Model Pretraining

The pretraining methods can be categorized into supervised and SSL methods, with SSL prevailing due to their capabilities in capturing morphological patterns without labeled data, while only CHIEF opted for supervised pretraining for the aggregator. Based on our surveyed papers, SSL can be further divided into two main categories: vision-only and inter-modal methods. Vision-only methods employ three SSL techniques: *contrastive learning* (SimCLR, MoCov3), *masked image modeling* (MIM, MAE), and *self-distillation* (iBOT, DINO, DINOv2). In contrast, inter-modal methods often employ multi-stage pretraining, utilizing contrastive learning methods (CLIP, CoCa) for effective cross-modal alignment before which unimodal encoders are pretrained independently. We focus on methodology contributions in this section and present the details of each method, including input modalities, magnification, and resolution of the patches, in the Model Pretraining column of Table 1.

Contrastive Learning is an SSL branch that learns representations by maximizing similarity between positive pairs while minimizing that between negative pairs. Several seminal approaches have advanced this field: 1) SimCLR [Chen *et al.*, 2020] established foundational techniques such as aggressive data augmentation and large batch sizes; 2) MoCov3 [Chen *et al.*, 2021] advanced self-supervised learning for ViT through stabilized training techniques; 3) CLIP [Radford *et al.*, 2021] expanded the paradigm to multi-modal learning through large-scale image-caption pair training; and 4) CoCa [Yu *et al.*, 2022] proposed a unified method incorporating both contrastive and captioning objectives, enabling simultaneous visual-textual alignment and text generation capabilities. In the medical domain, REMEDIS utilizes SimCLR to enhance the robustness and data efficiency in medical imaging. TANGLE adopts a revised SimCLR method with gene expression reconstruction and slide subset alignment. Pathoduet [Hua *et al.*, 2024] enhanced MoCov3 through the integration of cross-scale positioning and cross-stain transferring tasks, specifically addressing the challenges of stain transferability and tissue-level heterogeneity. CLIP is adapted for both extractors (PLIP [Huang *et al.*, 2023]) and aggregators (Prov-GiGapath, mSTAR, MADELEINE, and THREAD), due to its versatility in aligning two or more modalities. Notably, KEEP [Zhou *et al.*, 2024] has proposed a **K**nowledge-**E**nhanced **V**ision-**L**anguage (KEVL) pretraining, further adapting CLIP for the extractor by incorporating domain expertise through knowledge-graph-cleaned image-text pairs. There are several applications of CoCa, both on the extractor and aggregator: CONCH [Lu *et al.*, 2024] adopts this framework to pretrain an extractor on 1.17 million image-caption pairs, enhancing both zero- and few-shot capabilities, while PRISM [Shaikovski *et al.*, 2024] and TITAN utilize CoCa to pretrain aggregators with multimodal capabilities.

Masked Image Modeling is an SSL method that learns representations by predicting masked portions of images from their visible regions. SimMIM [Xie *et al.*, 2022] advanced the field by simplifying existing approaches through random masking and a lightweight prediction head, and MAE [He *et al.*, 2022] introduced an asymmetric encoder-decoder design with high masking ratios. Recent investigations have demonstrated the efficacy of MIM in pretraining extractors; notably, SINAI [Campanella *et al.*, 2024] employs MAE to pretrain ViT models on a scale of 3.2 billion patches, establishing its scalability in pathological contexts. Similarly, MUSK [Xiang *et al.*, 2025] and BEPH [Yang *et al.*, 2024] further validate MIM by implementing BEiT-3 and BEiTv2 architectures, respectively. Additionally, Prov-GigaPath employs MAE to pretrain its slide encoder LongNet, demonstrating the efficacy of this method on aggregator pretraining.

Self-distillation enables model learning through its own predictions across different views, simultaneously acting as teacher and student. DINO [Caron *et al.*, 2021] pioneered the use of self-distillation by employing a teacher-student architecture with momentum encoder and multi-crop training, while iBOT [Zhou *et al.*, 2022] performs MIM via self-distillation with an online tokenizer, and DINOv2 [Oquab *et al.*, 2023] refined the DINO framework by accelerating and stabilizing the training at scale. The efficacy of self-distillation for the extractor has been demonstrated by several investigations: Phikon [Filiot *et al.*, 2023] implements iBOT on a corpus of 43 million patches spanning 16 distinct cancer sites; Phikon-v2 [Filiot *et al.*, 2024] employs DINOv2 on 456 million patches derived from 30 cancer sites; RudolfV [Dippel *et al.*, 2024] incorporates DINOv2 with pathologist knowledge on 58 tissue types and 129 stains; and Hibou [Nechaev *et al.*, 2024] further extends DINOv2 on 1.2 billion patches. Additionally, the application of self-distillation extends beyond the extractor, as evidenced by TITAN [Ding *et al.*, 2024], which utilizes iBOT for general-purpose aggregator learning. These investigations demonstrate the capacity of self-distillation in PFM pretraining. In addition, there are methodological improvements customized for pathology in this category: 1) PLUTO [Juyal *et al.*, 2024] utilizes DINOv2 together with MAE objective and Fourier losses on 195 million patches; 2) GPFM [Ma *et al.*, 2024] proposes **U**nified **K**nowledge **D**istillation (UKD), incorporating MIM, self-distillation and expert knowledge distillation together as training objectives; 3) Virchow2 [Zimmermann *et al.*, 2024] enhances DINOv2 by applying pathology-specific augmentation and reducing tissue redundancy.

3.3 Model Design

The model design refers to the following three aspects that are vital to model performance: *architecture*, *number of parameters* (*# params.*), *scale*. The scale of a model is directly determined by its number of parameters. Through quantization of the number of parameters, we establish a hierarchical scale

Venue	Model	Method	Architecture	Data Source	Data Statistics	Links
MedIA [Wang *et al.*, 2022]	CTransPath	SRCL	Swin-T/14	TCGA + PAIP	32,220 WSIs 15,580,262 Patches	
Nat. Bio. Engg. [Azizi *et al.*, 2023]	REMEDIS	SimCLR	ResNet-50	TCGA	29,018 WSIs 50 Million Patches	
CVPR [Chen *et al.*, 2022]	HIPT	DINO	ViT-S/16 ViT-XS/256	TCGA	10,678 H&E WSIs $\sim$ 104 Million Patches	
Nat. Med. [Huang *et al.*, 2023]	PLIP	CLIP	ViT-B/32	OpenPath	208,414 Image-Text Pairs	
Nat. Med. [Lu *et al.*, 2024]	CONCH	P: iBOT A: CoCa	P: ViT-B/16 A: GPT-style	In-house	21,442 WSIs 16 Million Patches $>$ 1.17M Image-Text Pairs	
MedRxiv [Filiot *et al.*, 2023]	Phikon	iBOT	ViT-S/B/L/16	TCGA	6,093 WSIs 43,374,634 Patches	
Nat. Med. [Chen *et al.*, 2024]	UNI	DINOv2	ViT-L/16	Mass-100K	100,426 H&E WSIs 100,130,900 Patches	
Nat. Med. [Vorontsov *et al.*, 2024]	Virchow	DINOv2	ViT-H/14	MSKCC	1,488,550 H&E WSIs 2 Billion Patches	
AAAI S. [Campanella *et al.*, 2024]	SINAI	DINO MAE	ViT-S ViT-L	Mount Sinai Health System	423,563 H&E WSIs 3.2 Billion Patches	
Nature [Wang *et al.*, 2024b]	CHIEF	P: Pretrained S: Sup.+CLIP	P: CTransPath S: CHIEF	Public + In-house	60,530 H&E WSIs $\sim$ 15 Million Patches	
Nature [Xu *et al.*, 2024a]	Prov-GigaPath	P: DINOv2 S: MAE A: CLIP	P: ViT-g/14 S: LongNet	Providence Health System	171,189 WSIs 1,384,860,229 Patches	
MedIA [Hua *et al.*, 2024]	Pathoduet	Enhanced MoCov3	ViT-B/16	TCGA	11,000 WSIs 13,166,437 Patches	
Arxiv [Dippel *et al.*, 2024]	RudolfV	DINOv2	ViT-L/14	TCGA + In-house	133,998 WSIs 1.25 Billion Patches	
ICML W. [Juyal *et al.*, 2024]	PLUTO	DINOv2+ MAE+Fourior	FlexiViT-S/16	TCGA + Proprietary	158,852 WSIs 195 Million Patches	
Arxiv [Shaikovski *et al.*, 2024]	PRISM	P: Pretrained S: CoCa	P: Virchow S: Perceiver	MSKCC	587,196 WSIs 195K Pathology Reports	
CVPR [Jaume *et al.*, 2024]	TANGLE	P: iBOT S: Alignment	P: ViT-B/16 S: ABMIL	TG-GATEs	47,227 WSIs 6,597 Image-Gene Pair	
Nature [Xiang *et al.*, 2025]	MUSK	UMP	BEIT-3	Quilt-1M + PathAsst	$\sim$33,000 H&E WSIs 50M Patches 1M Image-Text Pairs	
BioRxiv [Yang *et al.*, 2024]	BEPH	MIM	BEiTv2	TCGA	11,760 WSIs 11,774,353 Patches	
Arxiv [Nechaev *et al.*, 2024]	Hibou	DINOv2	ViT-L/14 ViT-B/14	Proprietary	936,441 H&E WSIs 202,464 non-H&E WSIs ViT-L: 1.2B Patches ViT-B: 512M Patches	
Arxiv [Xu *et al.*, 2024b]	mSTAR+	S: CLIP P: mSTAR	S: TransMIL P: ViT-L	TCGA	11,727 WSIs 22,127 Modality Pairs	
Arxiv [Ma *et al.*, 2024]	GPFM	UKD	ViT-L/14	33 Public Dataset	72,280 WSIs 190,212,668 Patches	
Arxiv [Zimmermann *et al.*, 2024]	Virchow2 Virchow2G	Enhanced DINOv2	ViT-H/14 ViT-G/14	MSKCC + Worldwide	3,134,922 WSIs with Diverse Stains	
ECCV [Jaume *et al.*, 2025]	MADELEINE	P: Pretrained S: CLIP + GOT	P: CONCH S:MH-ABMIL	Acrobat + BWH	16,281 WSIs with Diverse Stains	
Arxiv [Filiot *et al.*, 2024]	Phikon-v2	DINOv2	ViT-L/16	Public + In-house	58,359 WSIs 456,060,584 Patches	
Arxiv [Ding *et al.*, 2024]	TITAN	P: Pretrained Stage1: iBOT Stage2: CoCa	P: CONCHv1.5 S: ViT-T/14	Mass-340K	335,645 WSIs 423,122 Image-Text Pairs 182,862 WSI-Text Pairs	
Arxiv [Zhou *et al.*, 2024]	KEEP	KEVL	UNI	Quilt-1M + OpenPath	143K Image-Text Pairs Hierarchical Medical KG	
Arxiv [Vaidya *et al.*, 2025]	THREADS	P: Pretrained S: CLIP	P: CONCHv1.5 S: MH-ABMIL	MBTG-47K: MGH+BWH +TCGA +GTEx	47,171 H&E WSIs 125,148,770 Patches 26,615 Bulk RNA 20,556 DNA Variants	

Table 2: Technical specifications of vision-related parts of PFMs by academic preprint release date, with peer-reviewed published works in purple. Abbreviations used: Patch-level extractor (P), Alignment (A), Slide-level aggregator (S).

system, facilitating standardized cross-architectural comparisons and enabling informed model selection for practical implementations. The taxonomy of PFMs along this dimension is presented in the Model Design column of Table 1.

Architecture is a pivotal determinant of PFM capabilities. Architectures adopted by PFMs can be categorized based on scope: extractor architecture and aggregator architecture. The extractor architecture encompasses two cate-

gories: 1) CNN-based architecture such as ResNet, and 2) transformer-based architecture, including ViT [Dosovitskiy *et al.*, 2021], CNN-integrated Swin Transformer [Wang *et al.*, 2022], BEiTv2 [Peng *et al.*, 2022], FlexiViT [Beyer *et al.*, 2023], and a multimodal BEiT-3 [Wang *et al.*, 2023]. The aggregator architecture comprises two categories: ABMIL family [Ilse *et al.*, 2018; Ding *et al.*, 2024] and Perceiver [Jaegle *et al.*, 2021]. For an architectural overview of each method, we refer readers to the Architecture column in Table 1.

Scale can be derived through the number of parameters. To facilitate cross-architectural comparisons, we establish a quantization framework based on the ViT architecture, which serves as the predominant backbone across our surveyed literature. The classification includes seven categories: e**x**tra **s**mall (XS, 2.78M), **S**mall (S, 21.7M), **B**ase (B, 86.3M), **L**arge (L, 307M), **H**uge (H, 632M), **g**iant (g, 1.13B), and **G**iant (G, 1.9B). The notation ViT-B/16 indicates a ViT Base model with patch size 16. For statistics of each method, we refer readers to the # Params. and Scale columns in Table 1.

Our analysis reveals several patterns in model architectures and scaling: 1) ABMIL-derived methods demonstrate clear dominance in aggregator architectures, while the ViT family predominates in extractor architectures, which are transformer-based architectures; 2) The majority of methods utilize ViT-L as their primary backbone. Due to computational resource constraints, researchers often develop complementary smaller-scale variants (ViT-S or ViT-B) alongside their primary models, while some approaches specifically target efficiency through smaller architectures; 3) ViT-L is a popular scale for extractors, whereas ViT-XS is the primary choice for aggregators, over ViT-S. This substantial disparity in parameter counts between extractors and aggregators, despite their similar training data scale, suggests a potential data-model scale mismatch that warrants further investigation; 4) a clear trend toward larger model scales is observed: while earlier approaches frequently employed ViT-B, recent methods have increasingly standardized on ViT-L, with some extending to even larger variants such as ViT-H/g/G.

4 Evaluation Tasks for the Foundation Model

Development and evaluation constitute the two fundamental pillars of PFMs. The evaluation tasks of PFMs can be systematically categorized into four aspects: 1) slide-level tasks; 2) patch-level tasks; 3) multimodal tasks; and 4) biological tasks. A comparative analysis of these evaluation tasks is presented in Table 3, providing practitioners with comprehensive criteria for model selection based on real-world applications.

Slide-level Tasks encompass analytical tasks that utilize WSIs as primary input or output modalities. These tasks include WSI classification (Cls.), survival prediction (Surv.), WSI retrieval (Retri.), and WSI segmentation (Seg.). While survival prediction methodologically represents a classification employing specialized loss functions, its clinical application differs from standard WSI classification: the latter primarily serves for diagnosis, while the former addresses prognosis. This category is the cornerstone of CPath, enabling automated diagnosis directly from WSIs with minimal manual intervention. Consequently, the majority of methods have prioritized experimental validation in this domain.

Patch-level Tasks comprise analytical tasks on patches as inputs or outputs, including patch classification (Cls.), patch-to-patch retrieval (P2P), and patch segmentation (Seg.). These tasks effectively evaluate the efficacy of the extractor, as they operate independently of additional aggregators.

Multimodal Tasks are tasks that evaluate multimodal capabilities of PFMs. These tasks encompass cross-modal retrieval, *i.e.*, image-to-text (I2T) and text-to-image (T2I) retrieval, report generation (RG), and visual question answering (VQA). RG in our survey includes both RG and image captioning, distinguished by their input: RG utilizes WSIs to generate clinical documentation, while captioning produces concise descriptions from patches. The increasing emphasis on these tasks reflects the clinical reality that pathologists integrate multimodal data in the decision-making process.

Biological Tasks focus on biomarker detection, including genetic alteration (GA) and molecular prediction (MP). Genetic alteration includes both mutation prediction and genetic alteration, as both predict gene mutation status. Molecular prediction targets the prediction of molecular subtypes at the gene expression level, representing a distinct biomarker from genetic alteration. While these tasks can be fundamentally categorized as classification problems at either slide or patch level, their clinical applications and biological implications warrant their classification as a separate analytical category. One recently-proposed task is molecular prompting [Vaidya *et al.*, 2025], which aims to perform clinical tasks with canonical molecular profiles without requiring any task-specific model development in a similar manner to text prompting.

While the extensive scope of our evaluation tasks precludes exhaustive evaluation by any single model, several methods, notably CONCH, UNI, MUSK, GPFM, and TITAN, provide excellent evaluation benchmarks across multiple training paradigms, including zero-shot, few-shot, and complete supervised learning, thereby providing more holistic insights into model capabilities and generalization potential.

5 Future Directions

PFMs constitute an emerging paradigm with transformative potential. Future research directions bifurcate into two primary domains: effective PFM Development and Utilization.

5.1 Foundation Model Development

Pathology-specific Methodology design is essential for PFMs that effectively capture the unique characteristics of pathology data. Most PFMs are pretrained using algorithms originally developed for natural images, neglecting critical aspects of pathology images, as detailed in Sec. 3; therefore, there is an urgent need for algorithms designed to accommodate these challenges. This deficiency extends to multimodal pretraining as well, where CLIP and CoCa are employed without customization, resulting in the omission of inherent features of pathology and related data, including genomics and reports, that are vital for comprehensive analysis.

End-to-end Pretraining is critical to achieve optimal performance for PFMs. Current PFMs adopt a two-stage pretraining paradigm: extractors are trained independently, followed by the aggregator with the extractor frozen. Evidence

Model	Slide Level				Patch Level			Multimodal				Biological	
	Cls.	Surv.	Retri.	Seg.	Cls.	P2P	Seg.	I2T	T2I	RG	VQA	GA	MP
CTransPath	C	C	✗	✗	F/C	Z	C	✗	✗	✗	✗	✗	✗
REMEDIS	C	C	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
HIPT	C	C	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
PLIP	✗	✗	✗	✗	Z	Z	✗	✗	Z	✗	✗	✗	✗
CONCH	Z/F/C	✗	✗	Z	Z/F	✗	✗	Z	Z	C	✗	✗	✗
Phikon	C	C	✗	✗	C	✗	✗	✗	✗	✗	✗	C	C
UNI	F/C	✗	F	✗	F/C	Z	C	✗	✗	✗	✗	✗	✗
Virchow	C	✗	✗	✗	C	✗	✗	✗	✗	✗	✗	C	✗
SINAI	C	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	C	C
CHIEF	C	C	✗	✗	✗	✗	✗	✗	✗	✗	✗	C	C
Prov-GigaPath	Z/C	✗	C	✗	✗	✗	✗	✗	✗	✗	✗	Z/C	✗
Pathoduet	C	✗	✗	✗	F/C	✗	✗	✗	✗	✗	✗	✗	F/C
RudolfV	✗	✗	Z	✗	C	✗	C	✗	✗	✗	✗	C	C
PLUTO	C	✗	✗	✗	C	✗	C	✗	✗	✗	✗	✗	C
PRISM	Z/C	✗	✗	✗	✗	✗	✗	✗	✗	C	✗	F/C	✗
TANGLE	F	✗	C	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
MUSK	C	C	✗	✗	Z/F/C	Z	✗	Z	Z	✗	C	C	C
BEPH	Z/F/C	C	✗	✗	C	✗	✗	✗	✗	✗	✗	✗	✗
Hibou	C	✗	✗	✗	C	✗	C	✗	✗	✗	✗	C	✗
mSTAR	Z/F/C	C	✗	✗	✗	✗	✗	✗	✗	C	✗	C	C
GPFM	C	C	✗	✗	C	Z	✗	✗	✗	C	C	C	✗
Virchow2	✗	✗	✗	✗	✗	✗	✗	C	✗	✗	✗	✗	✗
MADELEINE	F	C	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	F/C
Phikon-v2	F/C	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	F/C	F/C
TITAN	Z/F/C	C	Z	✗	C	✗	✗	Z	Z	C	✗	C	C
KEEP	Z	✗	✗	Z	Z	✗	✗	Z	Z	✗	✗	✗	✗
THREADS	F/C	C	Z	✗	✗	✗	✗	✗	✗	✗	✗	C	F/C

Table 3: Comparison of the evaluation tasks between different PFMs. Abbreviations used: Zero-shot (Z), Few-shot (F), Complete (C).

suggests this complicates optimization, highlighting the need for end-to-end pretraining of PFMs, which poses significant challenges in CPath, as transitioning away from MIL requires developing extremely sophisticated and efficient architectures and algorithms capable of simultaneously integrating local and global pathology information for gigapixel images.

Data-Model Scalability is a critical direction, as performance improvements continue to demonstrate logarithmic and sub-logarithmic scaling with model and data volume, respectively, without yet reaching a clear plateau. This domain presents four sub-directions: 1) examining the relative importance of WSI and patch quantity, particularly when considering diversity, a complex concept that is yet widely acknowledged as an indicator of high-quality data; 2) exploring efficient algorithms, due to the rapid expansion in both datasets and models, evident in transitions from CONCH (ViT-B) to CONCHv1.5 (ViT-L) and from UNI (ViT-L) to UNI2 (ViT-H); 3) addressing the data-model scale mismatch problem for the aggregator, detailed in Sec. 3.1; and 4) optimizing model scale, since the giant model size poses substantial deployment challenges in both hospital and academic settings.

Federated Learning with Efficiency is essential for addressing the challenges associated with collecting massive-scale datasets across multiple institutions while preserving patient privacy, as few institutions can feasibly collect WSIs at the million-scale alone. However, current research in this area remains limited; for instance, HistoFL [Lu *et al.*, 2022] has demonstrated improved performance, yet this benefit comes at the cost of significantly increased computational overhead. As PFMs continue to grow in size, scaling federated learning further exacerbates these challenges. Consequently, there is an urgent need to develop more efficient, privacy-protected methods in such large-scale cross-institutional collaborations.

Model Robustness addresses critical challenges in multi-institutional data curation. The acquisition of data from various sites inevitably introduces technical heterogeneity across scanning equipment specifications, image magnification levels, and staining protocols, resulting in significant data variations that embed site information [de Jong *et al.*, 2025]. These disparities undermine training stability and model generalizability; recent work shows that most models encode site information more strongly than biological signals [de Jong *et al.*, 2025]. These issues will be further exacerbated in federated learning under non-IID data distributions. Consequently, developing more robust algorithms and robustness evaluation metrics for PFMs is a critical research imperative.

RAG-enhanced Pathology VLM is a trending paradigm worth investigating. Contemporary trends in **L**arge **L**anguage **M**odels (LLMs), such as Llama [Grattafiori *et al.*, 2024], with the prevalence of BERT-based architectures [Devlin *et al.*, 2019] in current multimodal PFMs suggest the potential utility of integrating LLMs with ViT architectures. Furthermore, given the demonstrated efficacy of **R**etrieval-**A**ugmented **G**eneration (RAG) [Gao *et al.*, 2023] in LLMs and the critical need for domain-specific expertise in pathol-

ogy, RAG methodology offers promising directions for representation learning in pathology VLMs. This approach transcends the limitations of existing methods such as RudolfV, which relies primarily on clustering techniques for pathologist knowledge integration, providing a potentially more sophisticated framework for incorporating domain expertise.

5.2 Foundation Model Utilization

Effective Adaptation of PFMs to downstream tasks is a critical research direction in their utilization, as these models are predominantly trained on large-scale heterogeneous datasets, resulting in general-purpose features rather than task-specific ones required for optimal performance. To address this limitation, effective adaptation methodologies are essential for task-specific optimization. The significance of this domain alignment challenge parallels the established paradigm of adapting conventional architectures, such as ResNet-50, to specialized domains like pathological image analysis, albeit with varying degrees of complexity and scope.

Model Maintenance constitutes a critical research domain in the context of PFMs, given the substantial computational resources required for their initial training. The potential diminishment of model performance due to novel diseases, tissue heterogeneity, or technological advancements necessitates efficient maintenance strategies to preserve model utility. Continual learning [Wang *et al.*, 2024a; Yu *et al.*, 2024] represents a promising approach for maintaining PFM effectiveness, as it circumvents the necessity for model retraining by learning on newly observed instances. This approach significantly reduces the required computational overhead while ensuring the model remains current with the evolving clinical, disease, and technological developments.

6 Conclusion

This survey presents a systematic analysis of the current Pathology Foundation Models through our proposed hierarchical taxonomy and comprehensive evaluation framework. Although the PFMs demonstrate significant advances in computational pathology, critical technical challenges merit further investigation. We delineate key directions that are worth exploring and might be instrumental in advancing both the theoretical foundations and practical applications of PFMs.

Acknowledgments

The authors would like to sincerely thank Professor Irwin King from the Department of Computer Science and Engineering, the Chinese University of Hong Kong, for his active involvement in the conception and development of this paper. Due to the one-submission-per-author policy, his name could only appear in the acknowledgments. The work described in this paper was partially supported by the Research Grants Council of the Hong Kong Special Administrative Region, China (CUHK 2410072, RGC R1015-23).

References

[Azizi *et al.*, 2023] Shekoofeh Azizi, Laura Culp, Jan Freyberg, Basil Mustafa, Sebastien Baur, Simon Kornblith, Ting Chen, Nenad Tomasev, Jovana Mitrović, Patricia Strachan, et al. Robust and data-efficient generalization of self-supervised machine learning for diagnostic imaging. *Nature Biomedical Engineering*, 7(6):756–779, 2023.

[Beyer *et al.*, 2023] Lucas Beyer, Pavel Izmailov, Alexander Kolesnikov, Mathilde Caron, Simon Kornblith, Xiaohua Zhai, Matthias Minderer, Michael Tschannen, Ibrahim Alabdulmohsin, and Filip Pavetic. Flexivit: One model for all patch sizes. In *CVPR*, pages 14496–14506, 2023.

[Campanella *et al.*, 2024] Gabriele Campanella, Chad Vanderbilt, and Thomas Fuchs. Computational pathology at health system scale–self-supervised foundation models from billions of images. In *AAAI Symposium*, 2024.

[Carbonneau *et al.*, 2018] Marc-André Carbonneau, Veronika Cheplygina, Eric Granger, and Ghyslain Gagnon. Multiple instance learning: A survey of problem characteristics and applications. *Pattern recognition*, 77:329–353, 2018.

[Caron *et al.*, 2021] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *CVPR*, pages 9650–9660, 2021.

[Chanda *et al.*, 2024] Dibaloke Chanda, Milan Aryal, Nasim Yahya Soltani, and Masoud Ganji. A new era in computational pathology: A survey on foundation and vision-language models. *arXiv*, 2024.

[Chen *et al.*, 2020] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, pages 1597–1607. PMLR, 2020.

[Chen *et al.*, 2021] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *CVPR*, pages 9640–9649, 2021.

[Chen *et al.*, 2022] Richard J Chen, Chengkuan Chen, Yicong Li, Tiffany Y Chen, Andrew D Trister, Rahul G Krishnan, and Faisal Mahmood. Scaling vision transformers to gigapixel images via hierarchical self-supervised learning. In *CVPR*, pages 16144–16155, 2022.

[Chen *et al.*, 2024] Richard J Chen, Tong Ding, Ming Y Lu, Drew FK Williamson, Guillaume Jaume, Bowen Chen, Andrew Zhang, Daniel Shao, Andrew H Song, et al. Towards a general-purpose foundation model for computational pathology. *Nature Medicine*, 2024.

[de Jong *et al.*, 2025] 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.

[Deng *et al.*, 2009] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.

[Devlin *et al.*, 2019] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, pages 4171–4186, 2019.

[Ding *et al.*, 2023] Jiayu Ding, Shuming Ma, Li Dong, Xingxing Zhang, Shaohan Huang, Wenhui Wang, Nanning Zheng, and Furu Wei. Longnet: Scaling transformers to 1,000,000,000 tokens. *arXiv*, 2023.

[Ding *et al.*, 2024] Tong Ding, Sophia J. Wagner, Andrew H. Song, Richard J. Chen, Ming Y. Lu, Andrew Zhang, Anurag J. Vaidya, Guillaume Jaume, Muhammad Shaban, Ahrong Kim, et al. Multimodal whole slide foundation model for pathology. *arXiv*, 2024.

[Dippel *et al.*, 2024] Jonas Dippel, Barbara Feulner, Tobias Winterhoff, Timo Milbich, Stephan Tietz, Simon Schallenberg, Gabriel Dernbach, Andreas Kunft, Simon Heinke, Marie-Lisa Eich, et al. Rudolfv: a foundation model by pathologists for pathologists. *arXiv*, 2024.

[Dosovitskiy *et al.*, 2021] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021.

[Ericsson *et al.*, 2022] Linus Ericsson, Henry Gouk, Chen Change Loy, and Timothy M Hospedales. Self-supervised representation learning: Introduction, advances, and challenges. *Signal Processing Mag.*, 2022.

[Filiot *et al.*, 2023] Alexandre Filiot, Ridouane Ghermi, Antoine Olivier, Paul Jacob, Lucas Fidon, Axel Camara, Alice Mac Kain, Charlie Saillard, and Jean-Baptiste Schiratti. Scaling self-supervised learning for histopathology with masked image modeling. *medRxiv*, 2023.

[Filiot *et al.*, 2024] Alexandre Filiot, Paul Jacob, Alice Mac Kain, and Charlie Saillard. Phikon-v2, a large and public feature extractor for biomarker prediction. *arXiv*, 2024.

[Gao *et al.*, 2023] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yixin Dai, Jiawei Sun, Haofen Wang, and Haofen Wang. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2:1, 2023.

[Grattafiori *et al.*, 2024] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[Gui *et al.*, 2024] Jie Gui, Tuo Chen, Jing Zhang, Qiong Cao, Zhenan Sun, Hao Luo, and Dacheng Tao. A survey on self-supervised learning: Algorithms, applications, and future trends. *TPAMI*, 2024.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[He *et al.*, 2022] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, 2022.

[Hua *et al.*, 2024] Shengyi Hua, Fang Yan, Tianle Shen, Lei Ma, and Xiaofan Zhang. Pathoduet: Foundation models for pathological slide analysis of h&e and ihc stains. *Medical Image Analysis*, 97:103289, 2024.

[Huang *et al.*, 2023] Zhi Huang, Federico Bianchi, Mert Yuksekgonul, Thomas J Montine, and James Zou. A visual–language foundation model for pathology image analysis using medical twitter. *Nature medicine*, 2023.

[Ilse *et al.*, 2018] Maximilian Ilse, Jakub Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *ICML*, pages 2127–2136. PMLR, 2018.

[Jaegle *et al.*, 2021] Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. Perceiver: General perception with iterative attention. In *ICML*, pages 4651–4664. PMLR, 2021.

[Jaume *et al.*, 2024] Guillaume Jaume, Lukas Oldenburg, Anurag Vaidya, Richard J Chen, Drew FK Williamson, Thomas Peeters, Andrew H Song, and Faisal Mahmood. Transcriptomics-guided slide representation learning in computational pathology. In *CVPR*, pages 9632–9644, 2024.

[Jaume *et al.*, 2025] Guillaume Jaume, Anurag Vaidya, Andrew Zhang, Andrew H Song, Richard J Chen, Sharifa Sahai, Dandan Mo, Emilio Madrigal, Long Phi Le, and Faisal Mahmood. Multistain pretraining for slide representation learning in pathology. In *ECCV*, pages 19–37, 2025.

[Juyal *et al.*, 2024] Dinkar Juyal, Harshith Padigela, Chintan Shah, Daniel Shenker, Natalia Harguindeguy, Yi Liu, Blake Martin, Yibo Zhang, Michael Nercessian, Miles Markey, et al. Pluto: Pathology-universal transformer. In *ICML Workshop*, 2024.

[Lu *et al.*, 2022] Ming Y Lu, Richard J Chen, Dehan Kong, Jana Lipkova, Rajendra Singh, Drew FK Williamson, Tiffany Y Chen, and Faisal Mahmood. Federated learning for computational pathology on gigapixel whole slide images. *Medical image analysis*, 76:102298, 2022.

[Lu *et al.*, 2024] Ming Y Lu, Bowen Chen, Drew FK Williamson, Richard J Chen, Ivy Liang, Tong Ding, Guillaume Jaume, Igor Odintsov, Long Phi Le, Georg Gerber, et al. A visual-language foundation model for computational pathology. *Nature Medicine*, 30:863–874, 2024.

[Ma *et al.*, 2024] Jiabo Ma, Zhengrui Guo, Fengtao Zhou, Yihui Wang, Yingxue Xu, Yu Cai, Zhengjie Zhu, Cheng Jin, Yi Lin, Xinrui Jiang, et al. Towards a generalizable pathology foundation model via unified knowledge distillation. *arXiv preprint arXiv:2407.18449*, 2024.

[Nechaev *et al.*, 2024] Dmitry Nechaev, Alexey Pchelnikov, and Ekaterina Ivanova. Hibou: A family of foundational vision transformers for pathology. *arXiv*, 2024.

[Ochi *et al.*, 2025] Mieko Ochi, Daisuke Komura, and Shumpei Ishikawa. Pathology foundation models. *JMA Journal*, 8(1):121–130, 2025.

[Oquab *et al.*, 2023] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv*, 2023.

[Peng *et al.*, 2022] Zhiliang Peng, Li Dong, Hangbo Bao, Qixiang Ye, and Furu Wei. Beit v2: Masked image modeling with vector-quantized visual tokenizers. *arXiv*, 2022.

[Radford *et al.*, 2021] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*. PMLR, 2021.

[Shaikovski *et al.*, 2024] George Shaikovski, Adam Casson, Kristen Severson, Eric Zimmermann, Yi Kan Wang, Jeremy D Kunz, Juan A Retamero, Gerard Oakley, David Klimstra, Christopher Kanan, et al. Prism: A multi-modal generative foundation model for slide-level histopathology. *arXiv preprint arXiv:2405.10254*, 2024.

[Shurrab and Duwairi, 2022] Saeed Shurrab and Rehab Duwairi. Self-supervised learning methods and applications in medical imaging analysis: A survey. *PeerJ Computer Science*, 8:e1045, 2022.

[Vaidya *et al.*, 2025] Anurag Vaidya, Andrew Zhang, Guillaume Jaume, Andrew H. Song, Tong Ding, Sophia J. Wagner, Ming Y. Lu, Paul Doucet, Harry Robertson, Cristina Almagro-Perez, et al. Molecular-driven foundation model for oncologic pathology. *arxiv*, 2025.

[Vorontsov *et al.*, 2024] Eugene Vorontsov, Alican Bozkurt, Adam Casson, George Shaikovski, Michal Zelechowski, Kristen Severson, Eric Zimmermann, James Hall, Neil Tenenholtz, Nicolo Fusi, et al. A foundation model for clinical-grade computational pathology and rare cancers detection. *Nature medicine*, pages 1–12, 2024.

[Wang *et al.*, 2022] Xiyue Wang, Sen Yang, Jun Zhang, Minghui Wang, Jing Zhang, Wei Yang, Junzhou Huang, and Xiao Han. Transformer-based unsupervised contrastive learning for histopathological image classification. *Medical image analysis*, 81:102559, 2022.

[Wang *et al.*, 2023] Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, et al. Image as a foreign language: Beit pretraining for vision and vision-language tasks. In *CVPR*, 2023.

[Wang *et al.*, 2024a] Liyuan Wang, Xingxing Zhang, Hang Su, and Jun Zhu. A comprehensive survey of continual learning: Theory, method and application. *TPAMI*, 2024.

[Wang *et al.*, 2024b] Xiyue Wang, Junhan Zhao, Eliana Marostica, Wei Yuan, Jietian Jin, Jiayu Zhang, Ruijiang Li, Hongping Tang, Kanran Wang, Yu Li, et al. A pathology foundation model for cancer diagnosis and prognosis prediction. *Nature*, 634(8035):970–978, 2024.

[Waqas *et al.*, 2024] Muhammad Waqas, Syed Umaid Ahmed, Muhammad Atif Tahir, Jia Wu, and Rizwan Qureshi. Exploring multiple instance learning (mil): A brief survey. *Expert Syst with Appl.*, page 123893, 2024.

[Xiang *et al.*, 2025] Jinxi Xiang, Xiyue Wang, Xiaoming Zhang, Yinghua Xi, Feyisope Eweje, Yijiang Chen, Yuchen Li, Colin Bergstrom, Matthew Gopaulchan, Ted Kim, et al. A vision–language foundation model for precision oncology. *Nature*, pages 1–10, 2025.

[Xie *et al.*, 2022] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. Simmim: A simple framework for masked image modeling. In *CVPR*, pages 9653–9663, 2022.

[Xiong *et al.*, 2023] Conghao Xiong, Hao Chen, Joseph J.Y. Sung, and Irwin King. Diagnose like a pathologist: Transformer-enabled hierarchical attention-guided multiple instance learning for whole slide image classification. In *IJCAI*, 2023.

[Xiong *et al.*, 2024a] Conghao Xiong, Hao Chen, Hao Zheng, Dong Wei, Yefeng Zheng, Joseph JY Sung, and Irwin King. Mome: Mixture of multimodal experts for cancer survival prediction. In *MICCAI*, 2024.

[Xiong *et al.*, 2024b] Conghao Xiong, Yi Lin, Hao Chen, Hao Zheng, Dong Wei, Yefeng Zheng, Joseph JY Sung, and Irwin King. Takt: Target-aware knowledge transfer for whole slide image classification. In *MICCAI*, 2024.

[Xu *et al.*, 2024a] Hanwen Xu, Naoto Usuyama, Jaspreet Bagga, Sheng Zhang, Rajesh Rao, Tristan Naumann, Cliff Wong, Zelalem Gero, Javier González, Yu Gu, et al. A whole-slide foundation model for digital pathology from real-world data. *Nature*, 2024.

[Xu *et al.*, 2024b] Yingxue Xu, Yihui Wang, Fengtao Zhou, Jiabo Ma, Shu Yang, Huangjing Lin, Xin Wang, Jiguang Wang, Li Liang, Anjia Han, et al. A multimodal knowledge-enhanced whole-slide pathology foundation model. *arXiv preprint arXiv:2407.15362*, 2024.

[Yang *et al.*, 2024] Zhaochang Yang, Ting Wei, Ying Liang, Xin Yuan, Ruitian Gao, Yujia Xia, Jie Zhou, Yue Zhang, and Zhangsheng Yu. A foundation model for generalizable cancer diagnosis and survival prediction from histopathological images. *bioRxiv*, pages 2024–05, 2024.

[Yu *et al.*, 2022] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022.

[Yu *et al.*, 2024] Dianzhi Yu, Xinni Zhang, Yankai Chen, Aiwei Liu, Yifei Zhang, Philip S Yu, and Irwin King. Recent advances of multimodal continual learning: A comprehensive survey. *arXiv preprint arXiv:2410.05352*, 2024.

[Zhou *et al.*, 2022] Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. Image bert pre-training with online tokenizer. In *ICLR*, 2022.

[Zhou *et al.*, 2024] Xiao Zhou, Luoyi Sun, Dexuan He, Wenbin Guan, Ruifen Wang, Lifeng Wang, Xin Sun, Kun Sun, Ya Zhang, Yanfeng Wang, et al. A knowledge-enhanced pathology vision-language foundation model for cancer diagnosis. *arXiv preprint arXiv:2412.13126*, 2024.

[Zimmermann *et al.*, 2024] Eric Zimmermann, Eugene Vorontsov, Julian Viret, Adam Casson, Michal Zelechowski, George Shaikovski, Neil Tenenholtz, James Hall, David Klimstra, Razik Yousfi, et al. Virchow2: Scaling self-supervised mixed magnification models in pathology. *arXiv preprint arXiv:2408.00738*, 2024.