Development and Applications of Pathology Foundation Models in Cancer Diagnosis

Canran Xu*

Glasgow College Hainan, University of Electronic Science and Technology of China, Chengdu, Sichuan province, China, E-mail: 2959538H@student.gla.ac.uk

Abstract. Pathology foundation models are a class of deep neural networks pre-trained on massive pathology image data. These models typically obtain generic representations through self-supervised learning to support a wide range of cancer diagnostic tasks. In recent years, such models have demonstrated significant performance improvements in pan-cancer detection, rare cancer identification, cancer subtype classification, and molecular prediction. This paper focuses on the cancer diagnosis scenario, systematically summarizes the structural design and data resources of representative pathology foundation models, and emphasizes the discussion of patch-level Transformer, graph-based encoders, and visual-language models based on image-text pairing, and analyzes their roles in clinical-related tasks. Furthermore, this paper points out the main challenges currently faced in this field in terms of long-tail generalization, evaluation bias, engineering reproduction threshold, and the reliability of multimodal output, and looks forward to the future development direction of pathology foundation models for reliable clinical deployment.

1 Introduction

In conventional pathology diagnosis, tumor assessment relies on multi-scale visual cues. Pathologists need to comprehensively observe nuclear atypia, tumor growth patterns, as well as inflammatory and immune responses and other characteristics. With the development of digital pathology, traditional slides are scanned as ultra-high-resolution whole-slide images (WSI). A single slide usually contains tens of thousands of patches. However, the differences in fixation, staining, scanning and tissue processing procedures among different medical centers can lead to significant domain shifts [1, 2].

The task of cancer pathology diagnosis is highly heterogeneous, covering multiple aspects such as tumor detection, cancer classification, grading, and molecular biomarker prediction. The discrimination basis, analysis scale and annotation granularity corresponding to different tasks are not the same, which also makes model training often require a large amount of high-quality annotation data as support. Traditional supervised learning methods usually depend on region-level or patch-level fine annotations. However, in the cancer pathology scene, such annotations are not only costly and time-consuming, but also have strong subjectivity in some tasks. For instance, for grading tasks, there are often significant differences among different observers [3]. Relevant studies have shown that artificial intelligence assistance can significantly improve the consistency and accuracy of Gleason grading of prostate biopsy, which not only reflects the difficulty of manual grading itself, but also highlights the potential value of algorithm assistance [3]. Therefore, researchers have begun to explore learning frameworks that can reuse features between multiple tasks under weakly labeled or even unlabeled data.

Most of the early studies focused on convolutional neural network models with a single cancer or a single marker. The model constructed by Kather et al. can predict actionable genetic alterations and molecular subtypes of multiple cancer types directly from routine H&E slides [4]. This result indicates that there are indeed morphological patterns in conventional slides that can be learned by the model, and thus supports the hypothesis that general pathology representations can migrate across cancer species and endpoints.

On this basis, the emergence of pathology foundation models in the past two years has further promoted the development of this field. Such models are typically pre-trained with self-supervised or contrastive learning on large-scale patches or whole-slide datasets [1, 2, 5]. From the perspective of structural design, the relevant methods mainly fall into two categories: one is to use a patch encoder combined with a multi-instance learning aggregator [6], and the other is to model the long-range context at the whole-slide level through the Transformer [1, 2]. Representative models include TransMIL, CTransPath, Prov-GigaPath, Virchow, UNI, Phikon v2, as well as multimodal models CONCH and PLIP. The common goal of these models is to learn reusable embeddedness with a wide range of transfer capabilities, so that new tasks can be adapted by linear probes or lightweight fine-tuning, thereby reducing labeling dependence and improving cross-center robustness [1, 2, 5].

* Corresponding author: 2959538H@student.gla.ac.uk

2 Representative models development and research

2.1 Multi-instance learning and self-supervised contrastive learning

Many pathology foundation models are built on whole-slide level learning frameworks. In this framework, a single slide is represented as a set of multiple patches, and the learning algorithm needs to aggregate the patch features into a slide-level representation.TransMIL is one of the representative methods, and its core innovation lies in introducing the Transformer structure in the aggregation stage [6]. The model generates slide-level feature vectors by calculating the relevant attention weights between patches in a slide. The experimental results showed that in the CAMELYON16 metastasis detection task, the test AUC of TransMIL reached 93.09%. In TCGA non-small cell lung cancer and renal cancer subtype tasks, AUC was 96.03% and 98.82%, respectively [6]. These results indicate that attention-based multi-instance learning can more effectively exploit the inter-patch relationships and outperforms the simple pooling strategy as well as advanced baselines such as ABMIL and DSMIL.

On this basis, CTransPath proposed a hybrid encoder structure, combining the convolutional layer with the Transformer module, and introducing semantically relevant contrastive learning (SRCL) to improve the quality of self-supervised features [7].In this strategy, positive samples include not only enhanced versions of the same patch, but also patches that are similar in semantic space. Its pre-training data contains about 15 million patches, derived from more than 30,000 whole-slide images [7]. The authors reported that the features obtained from these self-supervised learning showed good transfer ability in several downstream tasks such as patch retrieval, patch classification, slide classification and gland segmentation [7]. Therefore, CTransPath has gradually become an important visual backbone and contrast baseline in subsequent studies of pathology foundation models [7].

Overall, this phase of research laid an important methodological foundation for the subsequent development of pathology foundation models. On one hand, multi-instance learning provides a feasible path for processing ultra-high resolution whole-slide images. On the other hand, self-supervised contrastive learning improves the ability of the model to utilize large-scale unlabeled pathological images. Together, they promote the transformation of the pathology model from task-specific training to general representation learning.

2.2 Large-scale pre-training and multi-task transfer

On the basis of the above methods, a general pathology foundation model for large-scale pre-training has been further developed in recent years.Prov-GigaPath adopts a two-stage structure consisting of a patch encoder and a slide encoder [1]. The model is pre-trained based on real-world data from a large medical network, covering 31 major tissue types and more than 30,000 patients [1]. Among them, the slide encoder is based on Transformer and models patch embedding sequences to capture the long-range context in the whole-slide range. The authors constructed 26 benchmarks including 9 cancer classification tasks and 17 pathomics prediction tasks. The results show that Prov-GigaPath achieves SOTA on 25 tasks and significantly outperforms the second-place method on 18 tasks [1]. For example, in the EGFR mutation prediction task on TCGA, this model improved 23.5% AUROC and 66.4% AUPRC compared with the second-place REMEDIS [1].

Virchow pays more attention to clinical-grade pan-cancer detection and rare cancer identification [2]. The model is a visual Transformer with 632 million parameters pre-trained on H&E whole slides and patient data in a self-supervised manner [2].In a unified downstream training setup, slide-level embeddings generated by Virchow, UNI, Phikon, and CTransPath are fed into the same classifier for comparison [2]. The results showed that Virchow had the highest specimen level AUC among the four models, reaching 0.938 and 0.934 respectively, both in the internal pan-cancer detection task and in the external institution data [2]. Under the condition of 95% sensitivity, the specificity of Virchow was 72.5%, which was still better than UNI, Phikon and CTransPath [2].In the rare cancer task of bone tumor detection, Virchow also achieved the highest AUC, indicating its strong competitiveness in pan-cancer detection and rare cancer identification [2]. These results illustrate that self-supervised pre-training on large-scale real-world slides can support clinically meaningful pan-cancer detection.

UNI is a universal foundation model of computational pathology [5]. The model is pre-trained on more than 100 million images from more than 100,000 diagnostic-grade H&E whole slides, with a total data volume of more than 77 TB, covering about 20 major tissue types [5].UNI was systematically evaluated on 34 computational pathology tasks, including tumor detection, subtype classification, grading and microenvironment phenotype recognition. In addition, the model also supports the evaluation of fine-grained cancer classification based on the OncoTree system, and the number of categories can reach 108 [5]. The results show that a single backbone network can achieve stable performance in multiple cancer diagnosis tasks.

Phikon v2 is an open-source feature extractor based on public data [8]. The authors pooled whole-slide images from multiple public cohorts to build a dataset of patches covering more than 30 cancer sites [8]. Subsequently, self-supervised learning was used to pre-train Vision Transformer on this dataset, and the encoder was publicly released. Phikon v2 was evaluated using an external validation cohort on eight slide-level tasks, with special emphasis on the complete isolation of pre-training data from the validation cohort at the patient level, and discusses the risk of data contamination [8]. Research shows that a simple and robust integrated training strategy can increase the average AUC across tasks and models by 1.75 percentage points, and the P-value is less than 0.001 [8].

 https://doi.org/10.1051/itmconf/20268801014

These results indicate that pathology foundation models trained on public resources are expected to achieve performance levels close to those of models based on private data.

In general, this phase of research marks the move from "building models for a single task" to "obtaining transferable generic representations through large-scale pre-training".Compared with the previous stage, these models not only have a larger pre-training scale but also show stronger generalization ability in tasks such as cancer detection, subtype classification, molecular prediction, and rare cancer identification, which also better reflect the core value of the foundation model in multi-task transfer.

2.3 Visual-language models and related research

Unimodal pathology foundation models have made some progress, at the same time, research has been further extended to multimodal scenes combining images and texts.CONCH is pre-trained on image-text paired samples and additional biomedical text data [9]. This model builds a shared embedding space between images and text through contrastive learning. The authors evaluated CONCH on 14 benchmark tasks, covering a variety of migration scenarios such as classification, segmentation, and image-text retrieval [9]. The results show that multimodal pre-training can significantly improve the transfer ability of the model in tasks involving images and language.

PLIP used public social media data to build the OpenPath dataset [10]. The dataset contains 208,414 pathological images and their natural language descriptions. The model uses a similar framework as CLIP to learn the alignment embedding between images and text [10].On the four external datasets, the zero-sample classification weighted F1 scores of PLIP ranged from 0.565 to 0.832, while the F1 of the contrast visual-language model ranged from 0.030 to 0.481 on the same task [10].In addition, PLIP also supports text retrieval images and image retrieval texts, providing a valuable tool for case retrieval and teaching applications [10].

Beyond the multimodal models themselves, clinically relevant studies further illustrate the potential of deep-learning models for actionable endpoints. Kather et al. used conventional H&E slides to predict actionable genetic alterations and molecular subtypes in multiple cancer species, which verified the feasibility of image prediction in multiple cancer species and cohorts [4]. Bulten et al. conducted a prospective film reading experiment to compare the performance of pathologists in Gleason grading of prostate biopsy with and without the assistance of artificial intelligence, and the results showed that the assistance of artificial intelligence could enhance the consistency with the expert reference standard and improve the grading stability [3]. Altogether, these studies show that the pathology foundation model is no longer limited to representation learning itself, but has shown practical application value in clinical scenarios such as molecular diagnosis, auxiliary imaging reading, and human-machine collaboration.

3 Data resources

The construction of pathology foundation models is highly dependent on large-scale data, and whole-slide images are the core modalities. In order to facilitate model training, slides are usually divided into patches of fixed size, which are then fed to the encoder for feature learning.Prov-GigaPath clearly states that its pre-training data contains 171,189 slides and more than 1,384,860,229 256×256 patches [1]. These slides were obtained from 28 cancer centers and covered a variety of tissue types and staining variants. Meanwhile, TCGA was used as the control data set, which contained about 30,000 slides and 208 million patches [1]. This comparison clearly demonstrates the volume gap between public resources and large-scale real-world data.

Virchow's training data included about 100,000 patients and about 1.5 million H&E whole slides [2].In the pan-cancer detection evaluation, the authors further gave the number of samples under different stratification, about 6142 cases in the overall cancer set, about 2595 cases in the rare cancer set, and about 3109 cases in the external institution set [2]. These numbers not only provide context for the AUC and specificity measures reported here but also show that the model was evaluated in a diverse and clinically relevant cohort.

UNI shows its pre-training data scale through the number of images, slides, and tissue types, emphasizing its wide coverage across organs and disease entities [5].In contrast, Phikon v2 aggregates slides from more than 100 public cohorts to construct a dataset of 460 million patches covering more than 30 cancer sites [8], and the validation cohorts used for eight slide-level tasks are not part of the pre-training data. Thus, the risk of overlap between the training set and the test set at the patient level is reduced [8].

Multimodal models also require data resources where images are paired with text.CONCH collected more than 1.17 million pairs of image-caption matched samples and combined with other biomedical corpora for pre-training [9]. These captions describe the histological type, diagnostic results, and notable morphological features. The OpenPath dataset built by PLIP comes from the public platform discussion and contains 208,414 images and their de-identified natural language descriptions [10]. These resources lay the foundation for the training and evaluation of visual-language foundation models on both image and text modalities.

4 Current challenges and future directions

Despite the outstanding performance of pathology foundation models on average, their reliable clinical application still faces many challenges.

First of all, long-tailed distribution and out-of-distribution generalization are still the core problems. Virchow achieved a high AUC in the most common cancer types, and even though it was still better than the comparison model in some tasks, its performance was still significantly decreased in some rare cancer tasks, such as bone tumor detection [2]. Other studies have found significant performance gaps between high-frequency and low-frequency entities in some tasks as well [1, 5, 8]. Therefore, future models not only need to deal with cancers with limited sample size more effectively, but also should have the ability to actively output uncertainty cues when the prediction is not reliable.

Secondly, the risk of measurement bias and data contamination cannot be ignored. As models and data sets continue to grow in size, it becomes more difficult to ensure complete independence of the pre-training cohort from the evaluation cohort.UNI and Phikon v2 are designed to avoid the use of common public datasets for both pre-training and downstream evaluation, and the training-test division is clearly explained [5, 8]. Virchow and Prov-GigaPath also employ a unified downstream training protocol when comparing different encoders to reduce the bias caused by "respective selection of favorable tasks" [1, 2]. Therefore, transparent reporting of cohort composition, patient-level division, and contamination inspection procedures is particularly important when conducting cross-sectional comparisons of underlying models.

Moreover, the threshold of project recurrence and resource is still high. Pre-training on hundreds of millions of patches or slides requires large-scale computing resources, mature data pipelines, and complex training optimization strategies, which makes it difficult for many research teams to fully reproduce the pre-training process.UNI, Phikon v2, CONCH, and PLIP alleviate this problem to some extent by exposing pre-trained encoders, feature extraction interfaces, and evaluation scripts [5, 8, 9, 10]. Nevertheless, there is a need for more standardized benchmark datasets and reference implementations in the field to enable fair comparisons between new approaches and established models.

Last but not least, the reliability and interpretability of multimodal outputs must be addressed in the future. Visual-language models such as CONCH and PLIP are capable of text generation and cross-modal retrieval [9, 10]. Such capabilities have strong appeal in teaching and research Settings, but in clinical Settings, automatically generated text must be tied to traceable graphic evidence and subject to expert review. The film reading experiment of Bulten et al shows that the evaluation criteria of a human-machine collaboration system are not equivalent to the performance indicators of a single model [3]. Future research may need to integrate visual-language foundation models with structured report templates, an interactive interface, and a human review process so that pathologists always retain control of the final report.

5 Conclusions

In general, pathology foundation models are promoting the transformation from single-task-oriented specialized systems to multi-task-oriented general feature extraction frameworks, and have shown important value in a variety of cancer diagnosis tasks. Models represented by TransMIL and CTransPath build high quality patch-level encoders by introducing Transformer-based multi-instance learning and self-supervised contrastive training. Whole-slide foundation models represented by Prov-GigaPath, Virchow, UNI, and Phikon v2 further demonstrate that large-scale pre-training on real-world and public data enables robust performance in tasks such as pan-cancer detection, rare cancer identification, subtype classification, and molecular biomarker prediction. At the same time, vision-language foundation models such as CONCH and PLIP extend pathological representation learning to the multimodal domain of image-text, and support new application scenarios such as automatic description and cross-modal retrieval. Besides, studies on image prediction of actionable genetic alterations and experiments on Gleason grade-assisted image reading further demonstrate the potential of deep learning to improve the consistency of molecular diagnosis and image reading.

However, long-tail generalization, measurement bias, engineering cost of training and deployment, as well as the safety and reliability of multimodal output, are still the key issues that must be addressed to advance the clinical application of pathology foundation models. Further research can be carried out in uncertainty modeling, rare cancer modeling, and standardized external benchmark construction. Meanwhile, the human-machine collaboration mode and workflow integration design will also become an important factor affecting the clinical implementation effect. Only when the model's ability is effectively embedded into the actual diagnostic process without weakening human supervision, the pathological basic model can truly provide credible, stable and generalizable auxiliary support for cancer diagnosis. With the continuous progress in these directions, the pathology foundation model is expected to become an important infrastructure in the future cancer pathology determination support system.

References

1. H. Xu, N. Usuyama, J. Bagga, et al., A whole-slide foundation model for digital pathology from real-world data. Nature 630, 181-188 (2024)
2. E. Vorontsov, A. Bozkurt, A. Casson, et al., A foundation model for clinical-grade computational pathology and rare cancers detection. Nat. Med. 30, 2924-2935 (2024)
3. W. Bulten, M. Balkenhol, J.-J. Awoumo Belinga, et al., Artificial intelligence assistance significantly improves Gleason grading of prostate biopsies by pathologists. Mod. Pathol. 34, 660-671 (2021)
4. J.N. Kather, L.R. Heij, H.I. Grabsch, et al., Pan-cancer image-based detection of clinically actionable genetic alterations. Nat. Cancer 1, 789-799 (2020)
5. R.J. Chen, T. Ding, M.Y. Lu, et al., Towards a general-purpose foundation model for

computational pathology. Nat. Med. 30, 850-862 (2024)

6. Z. Shao, H. Bian, Y. Chen, et al., TransMIL: Transformer based correlated multiple instance learning for whole slide image classification, in Adv. Neural Inf. Process. Syst. 34, 2136-2147 (2021)
7. X. Wang, S. Yang, J. Zhang, et al., Transformer-based unsupervised contrastive learning for histopathological image classification. Med. Image Anal. 81, 102559 (2022)
8. A. Filiot, P. Jacob, A. Mac Kain, et al., Phikon-v2, a large and public feature extractor for biomarker prediction. arXiv 2409.09173 (2024)
9. M.Y. Lu, B. Chen, D.F.K. Williamson, et al., A visual-language foundation model for computational pathology. Nat. Med. 30, 863-874 (2024)
10. Z. Huang, F. Bianchi, M. Yuksekgonul, et al., A visual-language foundation model for pathology image analysis using medical Twitter. Nat. Med. 29, 2307-2316 (2023)