



















A generalizable pathology foundation model using a unified knowledge distillation pretraining framework

Received: 2 August 2024

Accepted: 10 July 2025

Published online: 02 September 2025

 Check for updates

Jiabo Ma ^{1,18}, Zhengrui Guo ^{1,18}, Fengtao Zhou¹, Yihui Wang ¹, Yingxue Xu ¹, Jinbang Li^{2,3}, Fang Yan ⁴, Yu Cai ⁵, Zhengjie Zhu ⁶, Cheng Jin ¹, Yi Lin ¹, Xinrui Jiang¹, Chenglong Zhao^{2,3,7}, Danyi Li^{2,3}, Anjia Han ⁸, Zhenhui Li ⁹, Ronald Cheong Kin Chan¹⁰, Jiguang Wang ^{11,12}, Peng Fei ¹³, Kwang-Ting Cheng ^{1,5}, Shaoting Zhang ^{4,14} , Li Liang ^{2,3,15}  & Hao Chen ^{1,11,12,16,17} 

The generalization ability of foundation models in the field of computational pathology (CPath) is crucial for their clinical success. However, current foundation models have only been evaluated on a limited type and number of tasks, leaving their generalization ability unclear. We establish a comprehensive benchmark to evaluate the performance of off-the-shelf foundation models across six distinct clinical task types, encompassing a total of 72 specific tasks. Our findings reveal that existing foundation models excel at certain task types but struggle to effectively handle the full breadth of clinical tasks. To improve the generalization of pathology foundation models, we propose a unified knowledge distillation framework consisting of both expert and self knowledge distillation, where the former allows the model to learn from the knowledge of multiple expert models, while the latter leverages self distillation to enable image representation learning via local–global alignment. On the basis of this framework, we develop a Generalizable Pathology Foundation Model (GPFM). Evaluated on the established benchmark, GPFM achieves an average rank of 1.6, ranking first in 42 tasks, positioning it as a promising method for feature representation in CPath.

In recent decades, the shift to digital pathology, particularly through whole slide imaging, has modernized the workflow of clinicians and improved access to slide data¹. This has paved the way for computational pathology (CPath), an emerging field that leverages digital whole slide images (WSIs) and computational methods for clinical decision-making^{2–4}. Specifically, CPath introduces advanced capabilities such as gene mutation prediction^{5–7}, direct prognosis^{8–10} and treatment response assessment^{11–13} directly from WSIs, demonstrating profound clinical importance. However, the diversity of clinical

pathology tasks, combined with the limited data and annotations, poses challenges when training robust models for each individual task from scratch. This process is not only time consuming but also impractical in real-world scenarios⁴. Consequently, the CPath community is actively seeking solutions that can effectively address this diverse range of tasks simultaneously^{14–20}.

In recent years, there has been progress in the fields of computer vision and natural language processing driven by self-supervised learning on large-scale datasets. These pretrained models, commonly

referred to as foundation models (FM), have garnered widespread attention and have exhibited remarkable success across various tasks^{21–23}. In the field of CPath, some efforts^{24,24–29,30–32} have been dedicated to pretraining FMs that can learn inherent representations of histopathology images, catering to the diverse array of tasks encountered in clinical pathology practice. However, the current FMs have only been evaluated on a limited type of tasks (Fig. 2a), leaving their overall performance unclear. To comprehensively evaluate these models, we built a comprehensive benchmark spanning six major clinical task categories (Fig. 1d), comprising 72 specific tasks. Our findings revealed that the generalization ability of these models is still limited and no single model can effectively address all the tasks (Fig. 1d). It can be seen that UNI²⁶ achieves the best performance in WSI classification, image retrieval, survival analysis and patch-level (region-of-interest (ROI)) tissue classification tasks; Phikon²⁵ performs best in report generation tasks; and CONCH²⁸ obtains highest performance in visual question answering (VQA) tasks. This can be attributed to the fact that each FM is trained using distinct datasets and pretraining strategies, leading to specific advantages for each model within particular datasets. These findings highlight the need for further research to develop more generalizable FMs that can consistently perform well across the diverse types of clinical tasks encountered in CPath. By addressing this challenge, we can unlock the full potential of the FMs in CPath.

To improve the generalization of pathology FM and enhance the overall performance, an intuitive idea is to leverage the specific strengths of existing models by employing knowledge distillation techniques^{33,34}. Accordingly, we proposed a self-supervised learning framework with expert and self knowledge distillation to develop a Generalizable Pathology Foundation Model (GPFM). On the basis of the aforementioned pretraining method, we collected a dataset comprising 95,572 slides, encompassing 34 major tissue types, for the purpose of training and evaluating the GPFM. From this collection, we extracted 190 million patches derived from 72,280 slides to facilitate the pretraining (Fig. 1a). With the collected diverse tissues and indirectly using the images used to pretrain expert models (for example, UNI and CONCH), GPFM exhibits outstanding performance across the established benchmarks (Fig. 1b,c), achieving an average rank of 1.6, while the second best-performing model, UNI, achieves an average rank of 3.7 (Fig. 2c). These results demonstrate the efficacy of GPFM as a generalizable FM in CPath. The consistent performance of GPFM across a diverse range of clinical tasks underscores the advantages of employing knowledge distillation to integrate the strengths of specialized expert models.

Results

We evaluated various FMs across 72 tasks, encompassing 36 WSI classification tasks, 15 survival analysis tasks, 16 patch-level (ROI) tissue classification tasks, 2 pathological visual question answering tasks, 2 report generation tasks and 1 pathological image retrieval task (Fig. 2e–g). Since the tasks involved different types of evaluation metric, we assessed the overall performance of the FMs using an average ranking approach and reported the critical difference (CD) diagram^{35–37}. The model with the best performance was ranked 1st, while the model with the lowest performance was ranked 9th. Across all tasks, the GPFM model achieved the top average rank score of 1.6 (ranked first in 42 tasks), outperforming the second-best model, UNI, which had a ranking score of 3.7 (ranked first in 6 tasks). To evaluate the significance of GPFM's ranking score relative to other FMs, we performed the Nemenyi statistical test³⁵ (Fig. 2d). The results demonstrate that GPFM exhibited a statistically significant critical difference compared with the other eight models.

We calculated the average evaluation metrics across all 72 tasks (Fig. 2b), revealing that GPFM achieved the highest average score of 0.833, surpassing the second-best model, UNI, which scored 0.818. To assess statistical significance, we conducted a Wilcoxon signed-rank two-sided test³⁵ comparing GPFM with the second- and third-best

models. The results showed that all *P* values were below 0.001, confirming that GPFM consistently and significantly outperformed the existing FMs. Considering both the ranking perspective and the average metric aspect, the results clearly indicate that GPFM achieves state-of-the-art performance and is much more generalizable compared with the other FMs.

WSI classification

WSI classification is pivotal in accurate cancer diagnosis. It aids in categorizing the specific subtype of cancer, which can be substantially improved by using FMs. Therefore, it is important to evaluate the representation learning capabilities of different FMs. We conducted experiments on a total of 36 tasks, including 20 internal validation datasets and 16 external validation datasets. The detailed experimental results are presented in Supplementary Tables 1–18.

Across 36 WSI classification tasks, ranked according to the area under the curve (AUC) metric, GPFM achieved an outstanding average ranking score of 1.22, decisively surpassing the second-best model, UNI, which attained an average ranking score of 3.60 (Fig. 3a). We assessed overall performance using average metrics: AUC, balanced accuracy, and weighted F1 score. Specifically, GPFM achieved the highest average AUC of 0.891, a 1.6% improvement over UNI ($P < 0.001$; Fig. 3d). Similarly, GPFM outperformed UNI in balanced accuracy (0.752, +3.1%, $P < 0.001$; Fig. 3b) and weighted F1 score (0.736, +3.0%, $P < 0.001$; Fig. 3c). In addition, GPFM achieved the best performance in both internal and external tasks, with AUCs of 0.938 (+1.6% over UNI; Fig. 3e) and 0.832 (+1.5% over UNI; Fig. 3f). These results across multiple metrics highlight GPFM's strong generalization capability and potential for WSI classification tasks.

GPFM enhances diagnostic accuracy across multiple cancer types.

GPFM demonstrates superior diagnostic accuracy across a range of cancer types and tasks. In breast cancer, GPFM outperforms other models in all 6 evaluated tasks, including 5 subtyping tasks (Fig. 3j and Extended Data Fig. 1a) and 1 metastasis detection task (Extended Data Fig. 1d). For lung cancer, GPFM excels in 3 subtyping tasks, 2 metastasis detection tasks and 2 primary site prediction tasks (Extended Data Fig. 1b,f,h), except for 1 external validation for lung cancer metastasis detection, where UNI performs slightly better. In gastric cancer, GPFM achieves the best performance in 6 out of 9 tasks, including vascular invasion detection (Fig. 3h), perineural invasion detection and Lauren subtyping (Extended Data Fig. 1g,i). Furthermore, GPFM consistently delivers top performance in tasks involving other organs, such as brain tumour subtyping, head and neck cancer primary site and T stage prediction, colon lesion grading, prostate cancer grade assessment, ovarian cancer subtyping (Extended Data Fig. 2b–d, and Fig. 3i,g) and renal cell carcinoma classification (Fig. 3g). Overall, GPFM establishes itself as a leading model in cancer diagnosis across diverse tasks and cancer types.

GPFM advances gene mutation prediction. We conducted experiments on lung cancer and brain cancer slides. GPFM achieved the best results in both TP53 mutation prediction for lung cancer, with an AUC of 0.855 (+1.3% over Phikon; Extended Data Fig. 1e), and IDH1 mutation prediction for glioma, with an internal AUC of 0.986 and an external AUC of 0.943 (Extended Data Fig. 2a).

These results, along with the cancer diagnosis findings, highlight GPFM's superior generalizability compared with existing FMs. A key factor in this success is GPFM's ability to integrate knowledge from expert models through a unified knowledge distillation mechanism. Unlike previous FMs that do not employ knowledge distillation, GPFM leverages this approach to learn from a broader range of data and perspectives, dramatically enhancing its performance. This capability underscores GPFM's advanced adaptability and effectiveness across diverse tasks.

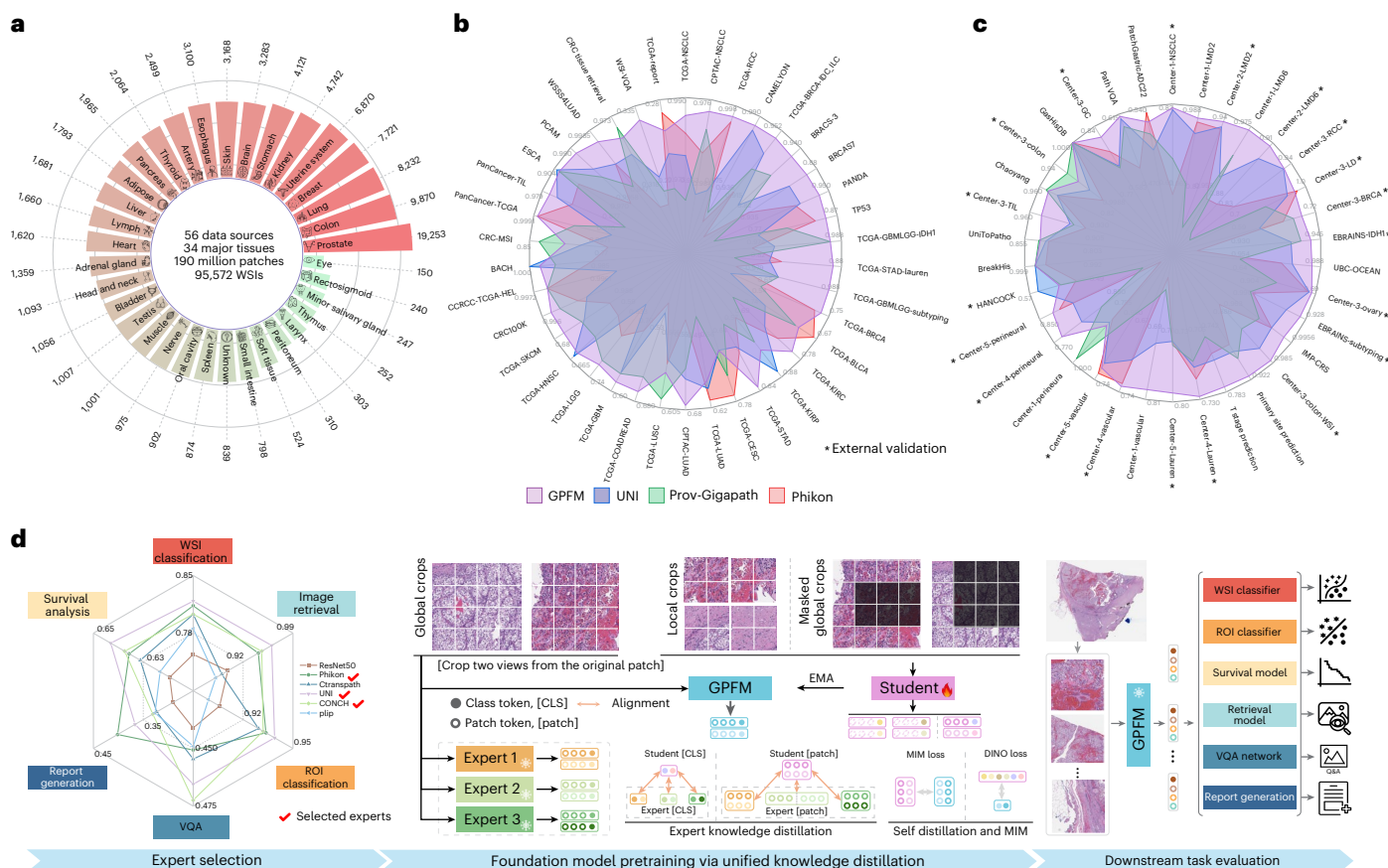


Fig. 1 | Overview of the GPFM. GPFM is a state-of-the-art pretrained FM that demonstrates exceptional performance across 72 diverse tasks. **a**, The GPFM dataset comprises a large-scale collection of 95,572 slides spanning 34 major tissue types, enabling comprehensive model training and evaluation. **b,c**, Performance evaluation of FMs across a diverse set of tasks: 52 internal tasks and 20 external tasks. Only the top 4 models are presented here. For a more comprehensive analysis, including additional FMs, please see Fig. 2.

d, Overview of unified knowledge distillation for GPFM. The experts used for expert knowledge distillation are selected on the basis of their average performance on six different clinical tasks. The pretraining algorithm includes three key components: (1) mask image modelling (MIM), (2) self distillation and (3) expert knowledge distillation. The parameters of GPFM are updated through exponential moving average (EMA).

Survival analysis

Accurate prediction of a patient's survival risk can enable more targeted and effective treatment strategies^{38–41}. A robust FM is essential for improving the precision of survival risk prediction, ultimately leading to better patient outcomes. To evaluate the performance of various FMs in survival analysis, we conducted experiments on 15 datasets. Following the methodologies of previous works^{39,41,42}, we adopted the Concordance index (C-index) as the evaluation metric to compare the performance of different FMs.

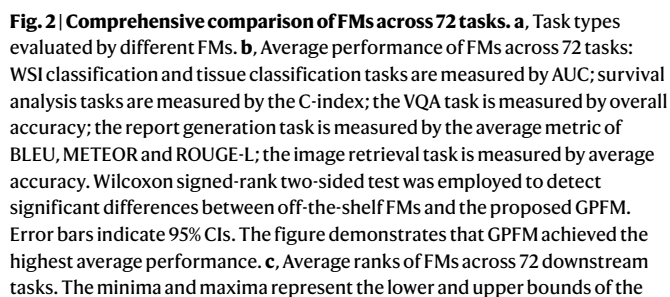
Across the 15 survival analysis tasks, GPFM achieved an impressive average ranking score of 2.1, ensuring the best or second-best performance in 13 tasks (Fig. 4a,d–f and Supplementary Tables 19–23). In comparison, the second best-performing model, UNI, attained an average ranking score of 4.6, achieving top-2 performance in only 4 tasks (Fig. 4a,d–f). Furthermore, when evaluated using the widely recognized C-index metric, GPFM emerged as the top performer, achieving an average C-index of 0.665 (Fig. 4b). This result represents a statistically significant improvement of 3.4% over UNI ($P < 0.001$), further demonstrating the superior generalization capability of GPFM for survival analysis tasks. To further validate the generalization of FMs, we conducted additional validation studies, including one external validation for head and neck cancer (TCGA-HNSC) and one internal validation for lung adenocarcinoma (TCGA-LUAD). In the head and neck cancer survival prediction task, UNI achieved the best performance in both the TCGA-HNSC and HANCOCK cohorts, while our method

ranked as the second-best performer (Fig. 4c). However, in the lung adenocarcinoma task, GPFM demonstrated a 10.6% improvement in the CPTAC-LUAD cohort (Extended Data Fig. 3h) compared with UNI.

Survival analysis tasks are inherently more challenging than WSI classification, and no single model has been able to dominate these tasks (Fig. 2e). The experimental results from both WSI classification and survival analysis highlight the limited generalization capability of existing FMs. This limitation is probably attributable to the data distribution of their training sets and the pretraining methods they employ. While existing FMs exhibit limited generalization, they demonstrate exceptional performance on specific types of tasks. By leveraging their individual strengths, it is possible to construct a more powerful and versatile model. This is precisely what we have achieved in this study: we propose a unified distillation framework to distill the capabilities of existing models—particularly in tasks where they excel—into GPFM, thereby substantially enhancing its generalization ability.

ROI classification

The performance of WSI classification is influenced by both the feature extractor (that is, FM) and the multiple instance learning (MIL) method. Unlike WSI classification, ROI classification tasks allow for a direct assessment of the FMs' feature representation capabilities, independent of MIL methods. To this end, we employed a linear probe approach, as outlined in ref. 43, to evaluate the FMs. Our assessment spanned 16 ROI classification tasks, encompassing 13 internal and 3 external



GPfM emerged as the top performer across all 16 ROI classification tasks, securing the best ranking score of 1.88, outperforming the second-ranked model, Prov-Gigapath, which scored 3.09 (Fig. 5a). In terms of conventional metrics, GPfM achieved the highest average AUC of 0.946 (+0.2% over Prov-Gigapath, $P < 0.001$; Fig. 5d), the best weighted F1 score of 0.865 (+0.9%, $P < 0.001$; Fig. 5c), and the highest

balanced accuracy of 0.866 (+1%, $P < 0.001$; Fig. 5b). GPFM exhibited outstanding performance in several tasks, including the detection of metastatic tissue in breast cancer (Fig. 5g), tissue type classification in lung cancer (Fig. 5h), the classification of tumour-infiltrating lymphocytes (TILs) (Fig. 5j), and the classification of gastric cancer tissues (Fig. 5k). In relatively simpler ROI classification tasks, GPFM shared the top rank with other FMs. For instance, in pancancer tissue classification (Extended Data Fig. 3f), breast tumour classification (Extended Data

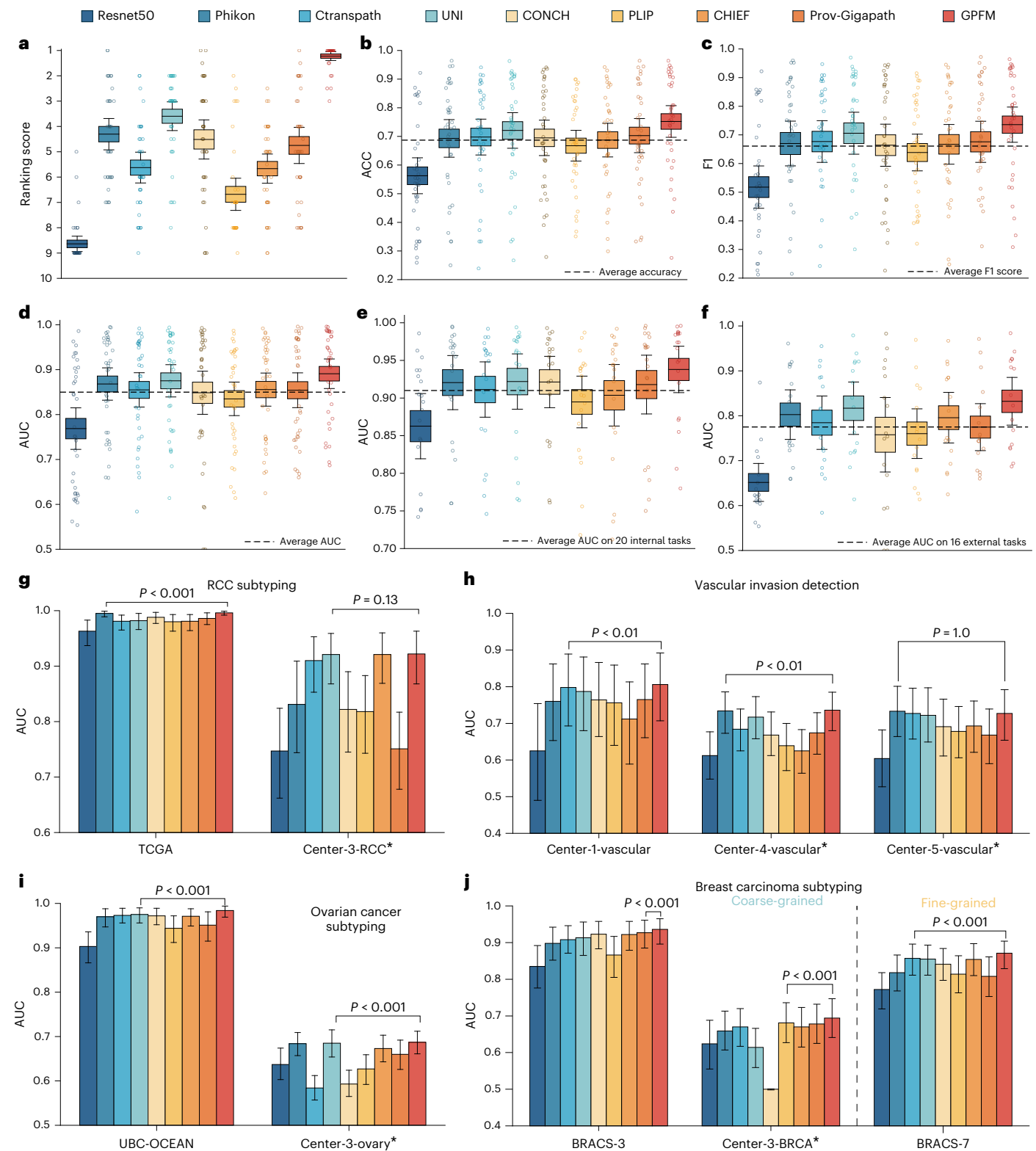


Fig. 3 | Performance of FMs on WSI classification tasks. **a**, Average ranking of FMs based on AUC across 36 WSI classification tasks. **b–d**, Average balanced accuracy (ACC) (**b**), weighted F1 score (F1) (**c**) and AUC (**d**) of FMs across 36 WSI classification tasks. **e**, Average AUC of FMs on 20 internal WSI classification tasks. **f**, Average AUC of FMs on 16 external validation cohorts. For **a–f**, the minima and maxima represent the lower and upper bounds of the 95% CIs, respectively; the centre and the bounds of the box represent the mean and standard error,

respectively. **g–j**, Model performance on specific tasks: RCC subtyping (**g**), vascular invasion detection (**h**), ovarian cancer subtyping (**i**) and breast carcinoma subtyping (**j**). The Wilcoxon signed-rank two-sided test was employed for data analysis (1,000 bootstrap replicates). * represents external validation cohorts. Error bars represent 95% CIs; the centre indicates mean. Additional results are shown in Extended Data Figs. 1 and 2.

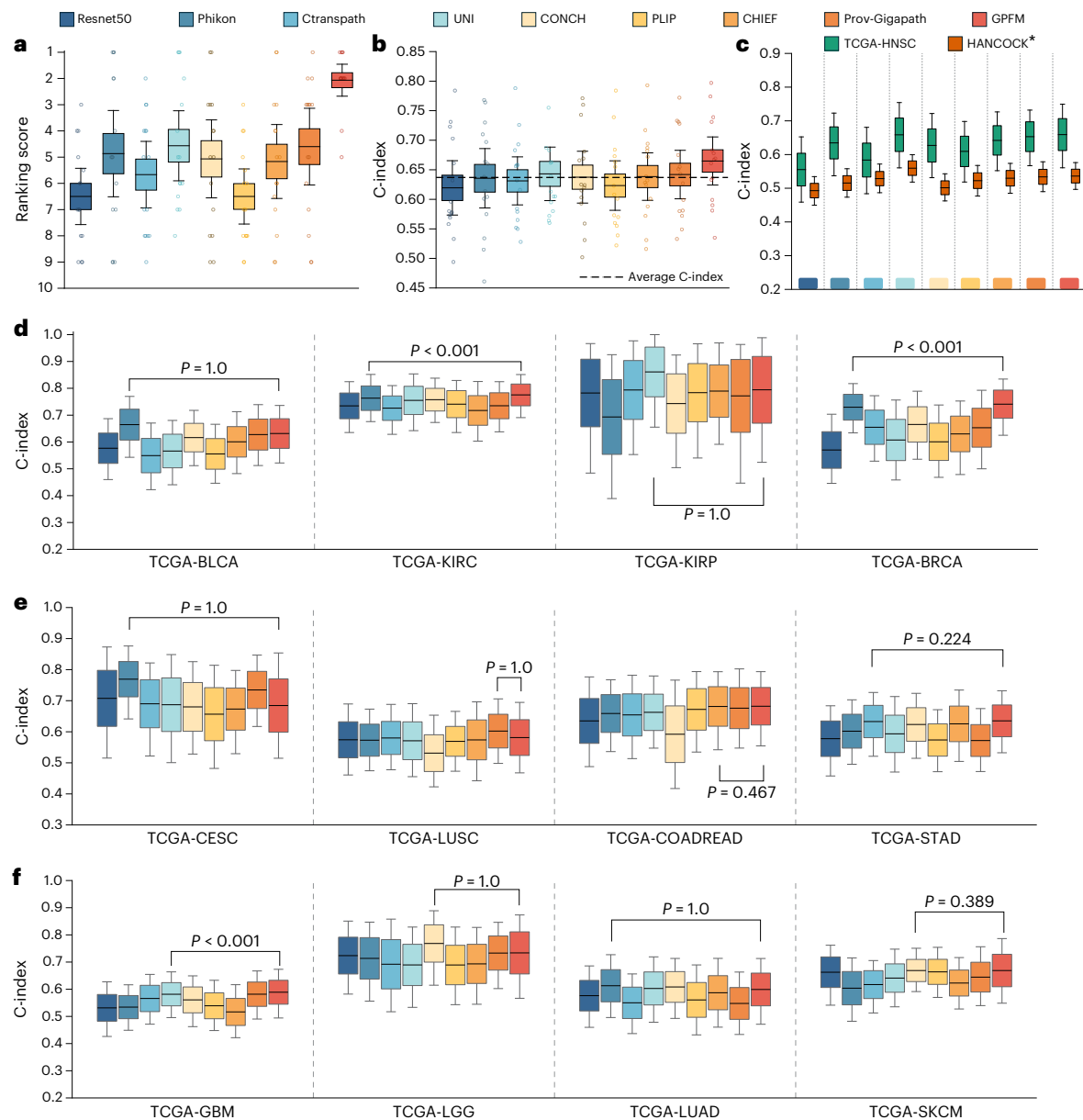


Fig. 4 | Performance of FMs across 15 survival analysis tasks. a, Average ranking of FMs in 15 survival analysis tasks. **b**, Average C-index of various FMs across 15 tasks. **c**, Results on TCGA-HNSC data and the HANCOCK cohort. The survival prediction model was trained on the TCGA-HNSC cohort and subsequently tested on the HANCOCK cohort. **d–f**, C-index of FMs across 12 survival analysis tasks.

In all boxplots, the minima and maxima represent the lower and upper bounds of the 95% CIs, the centre represents the mean, and the bounds of box represent the standard error. Wilcoxon signed-rank two-sided test was used for data analysis (1,000 bootstrap replicates).

Fig. 3b), colorectal cancer tissue classification (Fig. 5f) and kidney tissue classification (Fig. 5e), GPFM achieved performance on par with other leading FMs. In tasks where GPFM did not achieve the top performance, it consistently ranked as the second-best method (Extended Data Fig. 3a,d,e, and Fig. 5i) or the third-best method (Extended Data Fig. 3c). This consistent high ranking across diverse tasks contributed to GPFM's overall superior performance. In addition, the average ranking scores (Fig. 5a) of UNI and Prov-Gigapath are close, with ranking scores of 3.2 and 3.1, respectively. This indicates that no single existing model dominates ROI classification tasks. In contrast, by integrating knowledge from all adopted expert models, the unified knowledge distillation enables GPFM to surpass the performance of individual models, achieving a substantially lower average ranking score (that is, higher rank) of 1.88, outperforming the next-best model by more than one point. This underscores GPFM's strength as a highly generalizable FM.

Furthermore, to evaluate the robustness of GPFM in handling images with varying resolutions, we visualized the heat map of attention scores between the [patch] tokens and [CLS] tokens of the ViT transformer (Extended Data Fig. 3g). Across four resolutions: 224×224 , 448×448 , 896×896 and $1,344 \times 1,344$, we observed consistent attention patterns, highlighting GPFM's robustness in adapting to different image resolutions.

Pathological image retrieval

Image retrieval techniques could match the new patient pathology images to a curated database of previously diagnosed cases, providing pathologists with a novel tool to enhance diagnostic accuracy. Through visual inspection and comparison of similar historical cases, pathologists can leverage image search functionality to enhance their diagnostic decision-making. In this study, we employ the colorectal cancer (CRC)-100K dataset⁴⁴ for conducting pathological image retrieval tasks.

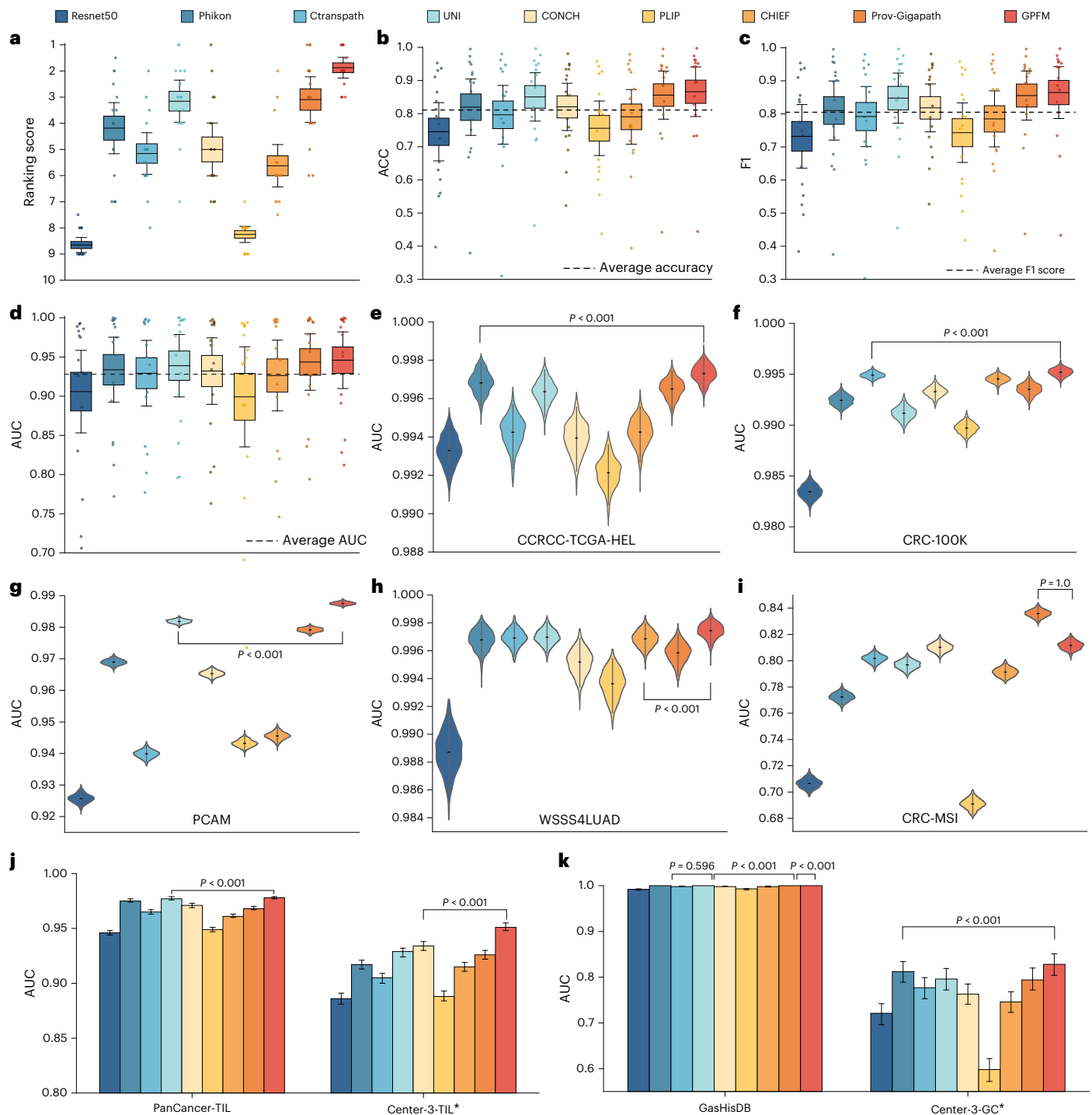


Fig. 5 | Performance of FMs on tissue classification tasks. **a**, Average ranking order of FMs based on AUC across 16 tasks. **b–d**, Average balanced accuracy (ACC) (**b**), weighted F1 score (F1) (**c**) and AUC (**d**) of FMs across 16 tasks. The centre represents the mean and the bounds of the box represent the standard error; the minima and maxima represent the lower and upper bounds of the 95% CIs. **e–i**, AUC of FMs across 5 tissue classification tasks. The centre in violin plots

represents the mean AUC. **j**, Tumour-infiltrating lymphocyte classification based on the PanCancer-TIL (internal) and Center-3-TIL (external) datasets. **k**, Gastric cancer tissue classification with GasHisDB (internal) and Center-3-GC (external) datasets. In all panels, error bars indicate 95% CIs. Wilcoxon signed-rank two-sided test was used for data analysis (1,000 bootstrap replicates). More results are presented in Extended Data Fig. 3.

The experimental results (Fig. 6a and Supplementary Table 37) show that the GPFM model achieved the second-best top-1 accuracy with a value of 0.906 (−1.9%, Prov-Gigapath). However, GPFM outperforms other models in terms of top-3 and top-5 accuracy, achieving values of 0.993 (+0.5%, Prov-Gigapath) and 0.995 (+0.2%, Prov-Gigapath), respectively. To further explore clustering effect and feature representation ability, we used *t*-distributed stochastic neighbour

embedding (*t*-SNE)⁴⁵ to project the features extracted by GPFM into a two-dimensional (2D) embedding space. The categories are well clustered, further illustrating that the features are highly discriminative (Fig. 6b). We also visualized the feature distribution of other FMs (Extended Data Fig. 4). The features extracted by GPFM are clustered more tightly and the query image is also located within the candidate cluster, indicating a better clustering effect. This observation suggests

that GPFM has superior feature representation capabilities in capturing the intrinsic patterns and structures present in the data.

Pathological image VQA

VQA is an exciting field of artificial intelligence (AI) that aims to enable machines to answer questions about visual content. In the domain of pathology, VQA systems can be particularly powerful, allowing clinicians and researchers to quickly and accurately extract relevant information from medical images.

For the patch-level VQA task, our model achieved the second-best performance, with results only slightly lower than those of CONCH (Fig. 6c and Supplementary Table 38). CONCH is a vision–language FM trained on millions of image–text pairs, which inherently provides it with an advantage in VQA tasks. Despite this, our results highlight the substantial potential of our approach compared with other pure vision FMs. To further illustrate the capabilities of our model, we visualized the query images, questions and answers generated by different FMs (Fig. 6d,e). As demonstrated in the figures, both GPFM and CONCH consistently produced more-reliable and accurate answers compared with the other models.

Moreover, in the WSI-level VQA task⁴⁶, our model achieved the best or second-best performance across 6 out of 7 metrics, demonstrating performance comparable to that of the slide-level FM CHIEF (Extended Data Fig. 5 and Supplementary Table 39). These results, combined with the patch-level findings, underscore the effectiveness of unified knowledge distillation. Specifically, the knowledge acquired by CONCH from millions of image–text pairs can be successfully distilled into GPFM without requiring access to the original image–text pair data. The strong performance of GPFM highlights the potential of leveraging textual knowledge indirectly, without the need for direct use of text data, thereby offering a promising direction for future research in VQA tasks.

Pathology report generation

Pathology reports are essential components of the healthcare system, providing critical information to clinicians and patients about the diagnosis, prognosis and treatment of various medical conditions. These reports summarize the findings from pathological examinations, such as biopsies, cytology samples and surgical specimens, and play a vital role in guiding clinical decision-making. Traditionally, pathology reports are written manually by pathologists and their teams, a time-consuming and labour-intensive process. Recent advancements in natural language processing and machine learning have enabled the development of automated pathology report generation systems, which can dramatically improve the efficiency and consistency of this critical task^{47–49}. To assess the effectiveness of FMs in this domain, we evaluated their performance on the TCGA WSI-Report dataset, curated by ref. 47, and the PatchGastricADC22 (ref. 50) dataset.

The experimental results demonstrate that Phikon achieved the best performance across all six metrics, while GPFM achieved comparable performance and ranked as the second-best model on both tasks (Fig. 6f, and Supplementary Tables 40 and 41). It is quite surprising to observe that vision FMs (for example, Phikon and GPFM) performed much better in this task than vision–language FMs such as CONCH and PLIP. This performance gap may be attributed to PLIP and CONCH's training paradigm, which relies solely on short descriptions or captions of pathological images without access to global contextual information. Consequently, these text–image pairs proved less effective for comprehensive report generation compared with their original VQA task applications. The examples of generated reports shown in Extended Data Figs. 6 and 7 certify this assumption.

To further validate these findings, we conducted stratified report generation analyses by stratifying the TCGA WSI-report dataset by major cancer types, that is, breast, lung and kidney cancers, for independent evaluation. Results (Supplementary Table 42 and Extended

Data Fig. 8a–c) reveal that Phikon keeps its superiority in breast and lung cancer report generation, yet is slightly outperformed by our GPFM in kidney cancer report generation. To leverage the complementary strengths of existing FMs, the proposed unified knowledge distillation approach can distill the capabilities of Phikon in report generation into the GPFM. This synergistic integration allows us to combine the respective strengths of these FMs, leading to the development of a more generalizable model. To further assess clinical relevance, an experienced pathologist evaluated the diagnostic reports using a 4-tier scoring system (Extended Data Fig. 8d). The blinded human-based evaluation results demonstrate GPFM's superior performance, achieving the highest average scores across breast, lung and kidney cancer reports (Supplementary Table 43 and Extended Data Fig. 8a–c). These expert-validated results underscore the potential of our unified knowledge distillation approach to generate clinically meaningful reports that align with pathologists' diagnostic standards, marking a major step toward the practical application of AI in pathology workflow automation.

The effectiveness of expert knowledge distillation

In the self-supervised learning framework proposed in this study, we introduced a unified knowledge distillation model to facilitate the transfer of knowledge from off-the-shelf FMs to GPFM during the pre-training stage. To assess the effectiveness of this module, we conducted an experiment where we removed the Expert Knowledge Distillation module, resulting in a modified self-supervised learning framework known as DINOv2 (ref. 43). We trained both DINOv2 and GPFM on the same dataset and evaluated their performance in tissue classification tasks. The experimental results clearly demonstrate the positive impact of expert knowledge distillation on the performance of the models across 12 tasks (Extended Data Fig. 9 and Supplementary Table 44). The experimental results demonstrated marked improvements not only in the performance of individual tasks but also in overall average performance, with substantial enhancements observed across all three evaluation metrics. The AUC increased by 0.6%, the weighted F1 score improved by 1.8%, and the balanced accuracy showed an increase of 1.8%. These findings provide strong evidence for the effectiveness of transferring knowledge from off-the-shelf pathology FMs through the proposed knowledge distillation learning framework. However, even with the distillation, GPFM still can not beat vanilla DINOv2 in all tasks such as Chaoyang and BreakHis, illustrating that there is still room for improving the distillation strategy.

Discussion

In this study, we construct a benchmark for CPath tasks. In addition, we introduce GPFM, a generalizable FM designed for a broad spectrum of CPath tasks. To enhance the model's versatility, we propose a unified knowledge distillation pretraining framework, which effectively consolidates expertise from a variety of existing models. This innovative approach ensures that GPFM can adapt and excel across different CPath tasks. To further maximize the diversity of data used for pretraining, we gathered 190 million images sourced from 56 sources, spanning 34 major tissue types. This rich dataset, combined with our advanced pre-training methodology, empowers GPFM to surpass current FMs in performance across 72 CPath tasks. Unlike other models that demonstrate proficiency in narrow domains—such as UNI²⁶, which specializes in WSI classification, and Phikon²⁵, which excels in report generation—GPFM showcases exceptional generalization, outperforming its counterparts across a wide array of CPath challenges by combining the strengths of expert models.

Recently, several vision–language^{28,29} and pure vision^{24–26,51} pathology FMs have been developed. However, the overall performance of these existing FMs is unclear due to the absence of a comprehensive benchmark. Our analysis reveals that no single existing model consistently exhibits the best performance. This is probably because each FM

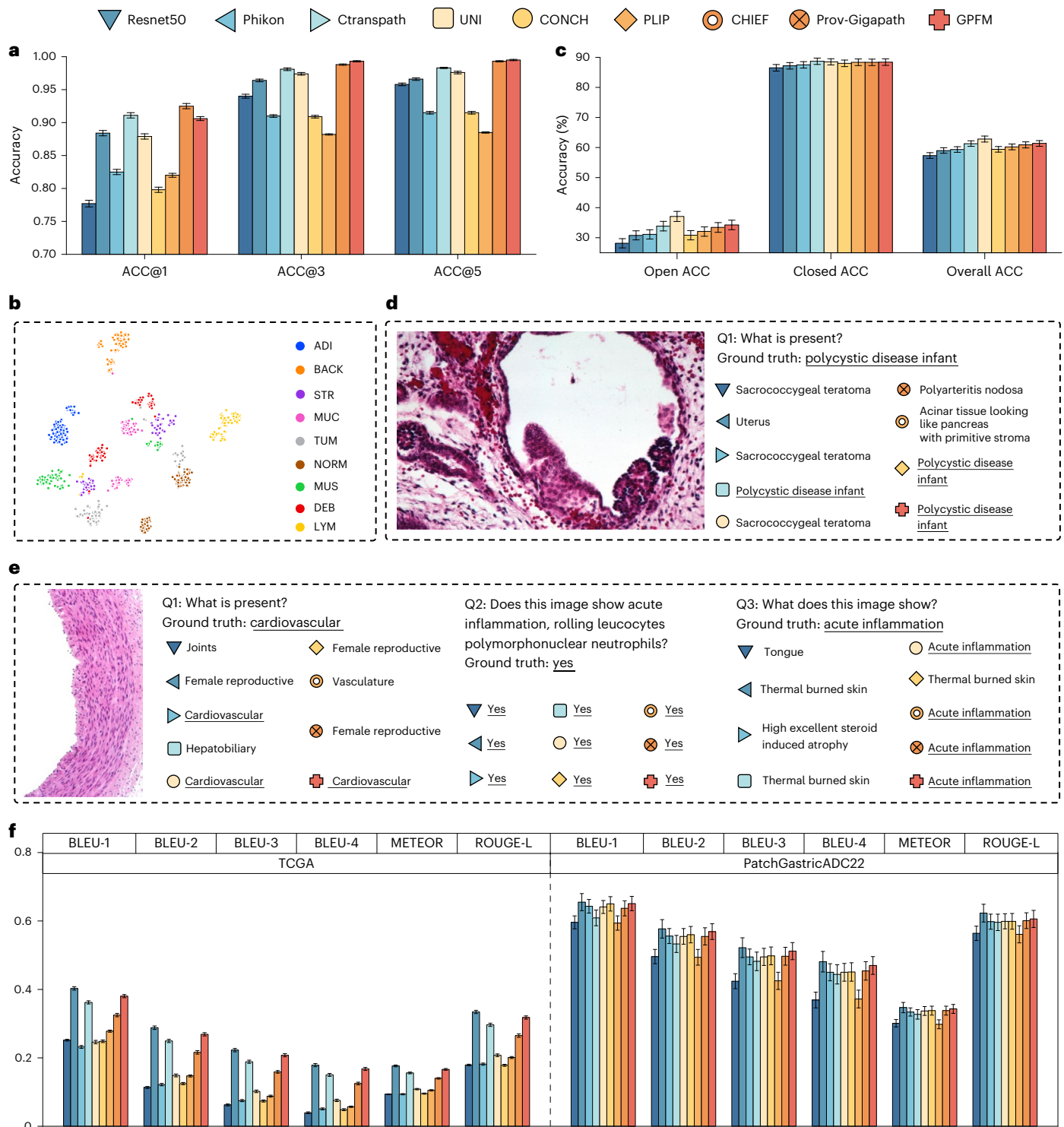


Fig. 6 | Overview of pathology tissue retrieval, VQA and report generation.

a, Top-1, top-3 and top-5 accuracies of different FMs on pathology tissue retrieval tasks. **b**, Distribution of features extracted by GPFM. For each class, 100 samples from the test set were used, and a total of 900 samples were subjected to *t*-SNE dimensionality reduction to 2D. **c**, Performance of VQA on the PathVQA dataset, measured by open-ended accuracy, closed-ended accuracy and overall accuracy,

for different FMs. **d**, An example of open-ended questions along with the answers generated by various FMs. **e**, Three example questions and the answers generated by FMs related to the query image. **f**, Performance of WSI report generation on TCGA and PatchGastricADC22 datasets. The models were measured by six different language quality metrics, that is, BLEU-1, BLEU-2, BLEU-3, BLEU-4, METEOR and ROUGE-L. In all panels, data indicate mean \pm s.d.

is trained using distinct datasets and pretraining strategies, leading to model-specific advantages for particular domains and datasets. The root of a model's generalization ability lies in the diversity of the training data. Unfortunately, gathering extremely large-scale diverse datasets, especially for sensitive medical data, is very difficult due to

security and privacy concerns. Therefore, it is almost impossible to access and use all the data used to develop the existing FM. Although accessing the original private training data is limited, the pretrained models themselves are available. Since the knowledge of the pretrained models is derived from the training data, we can indirectly leverage this

knowledge by using a unified knowledge distillation framework. It provides a feasible method to integrate knowledge from a large number of existing models under the premise of limited data and protecting data privacy, which has better feasibility and scalability in clinical practice. The substantially greater generalization ability of GPFM compared with existing FMs suggests that transferring knowledge from one existing model to another may be a more viable path to further advancing pathology FMs in the future, especially given the challenges of assembling large-scale diverse medical datasets.

This study also has some limitations. We recognize that current off-the-shelf FMs still exhibit potential in specific tasks, such as Phikon for report generation using TCGA data. This illustrates that the proposed unified knowledge distillation approach is not perfect and has room for improvement. Future research should concentrate on developing sophisticated methodologies to effectively distill and incorporate expert knowledge into one model, maximizing their potential across a broader spectrum of tasks. An example is further expanding the model's parameter size to enhance its adaptability, facilitating a more comprehensive assimilation of knowledge from diverse FMs. In addition, the current GPFM is a unimodal FM, which limits its ability to effectively handle cross-modal tasks such as VQA. Given the prevalence of multimodal data in pathology, encompassing WSIs, reports and genomic data, the development of a multimodal pathology FM is more attractive. Such a model would be more adept at integrating heterogeneous information, offering a more holistic understanding of patient data and enhancing diagnostic accuracy.

Methods

FM pretraining

CPath has emerged as a groundbreaking field that synergizes the power of AI with the expertise of pathologists, revolutionizing the practice of diagnosing and analysing diseases. At the core of this transformative discipline lies the FM, which serves as the backbone for a wide range of applications in pathology. While there exist some readily available FMs such as Ctranspath (pretrained on 32,000 TCGA slides)²⁴ and UNI²⁶ (pretrained on 100,000 private slides), the use of public data remains incomplete, and the evaluation of these models in CPath tasks is inadequate. The limited diversity of primary sites in the pretraining slides also restricts the adaptability of current FMs for public CPath benchmarks. To facilitate the advancement of CPath, we meticulously curated a comprehensive dataset comprising 56 histopathology datasets, encompassing a wide spectrum of 34 distinct tissue types for pretraining and downstream task evaluation (Supplementary Table 45). Leveraging this large-scale dataset, we developed a self-supervised learning approach with unified knowledge distillation to construct an FM that surpasses existing models.

Dataset preparation. To boost the performance of FMs, diverse datasets for various tissues are necessary. We have collected over 33 datasets as depicted in Supplementary Table 46 (from row 1 to row 33). To process WSIs, we employed the OpenSlide⁵² and CLAM toolkit⁵³ to find all non-overlapping 512×512 patches at level 0 that contain tissues. It is worth noting that we did not scale the patches to a uniform resolution, opting instead to use the original resolution of each WSI. This approach was implemented to increase the robustness of the FMs to varying resolutions. For datasets that only contain ROI images, we extracted non-overlapping 512×512 patches as well. Upon processing all 33 datasets, we obtained a comprehensive dataset, as presented in Supplementary Table 47. The pretraining data consist of 72,280 WSIs and a total of 190,212,668 patches.

Pretraining with self and expert knowledge distillation. In CPath, current FMs typically rely on state-of-the-art self-supervised pretraining (SSL) methods, such as DINOv2 (ref. 43) and iBOT⁵⁴. These methods are applied directly to either private or public datasets. For instance,

Phikon²⁵ was constructed on the basis of 6,093 TCGA slides using iBOT, while UNI was built upon ~100,000 private and public slides using DINOv2. Due to a larger training dataset and more powerful SSL methods, UNI outperforms Phikon in various tasks. However, UNI still lags behind other FMs in tasks related to text analysis and survival analysis due to its pretraining strategy and limited coverage of primary sites. To address the limitations of current FMs and further enhance their performance, we propose a novel pretraining strategy involving 'unified knowledge distillation'. The framework of the proposed pretraining method is similar to that of DINOv2; we employ teacher–student networks with masking image modelling (MIM) loss⁵⁵ and DINO (self distillation)^{54,56} loss to optimize the student network (Fig. 1c). Specifically, given an input image x , we obtain two augmented views, u and v . Random masking is then applied to both u and v , resulting in masked views, \hat{u} and \hat{v} . For the MIM objective, the student network takes \hat{u} and \hat{v} as inputs and aims to predict the masked tokens. With the DINO objective, we first crop n additional local views, w_i , and extract encoded class ([CLS]) tokens using the student network. Next, we obtain the [CLS] tokens of the global views (u and v) using the teacher network. Finally, we compute the cross-entropy loss between the local views and global views' [CLS] tokens. However, this strategy fails to leverage the knowledge from existing vision FMs, such as UNI and vision–language FMs such as CONCH²⁸, which restricts their applicability to different tissue types. To facilitate the transfer of knowledge from established pathology FMs, we propose an expert knowledge distillation module aimed at distilling knowledge into the student network^{33,57}. To maximize the generalizability of the pretrained model, it is crucial to balance the performance and diversity of expert models. We evaluated several existing models across six different tasks, selecting those that excelled in classification (UNI), report generation (Phikon) and visual question answering (CONCH) as expert models (see Fig. 1c). The [CLS] token, which represents the overall information of a patch for downstream tasks, serves as a critical component in our approach. If the [CLS] token of our model aligns well with those of the expert models, it indicates that our model can effectively assimilate the knowledge from selected experts. Similarly, the [PATCH] token also contains rich information. For example, some methods use mean pooling to perform downstream tasks⁵⁸. Therefore, aligning the [PATCH] token can further improve the effect of knowledge transfer. To achieve the above alignments, we use the student network to encode the global views u and v and extract the [CLS] and [PATCH] tokens. In addition, we employ the adopted experts to obtain their [CLS] and [PATCH] tokens. For aligning the class tokens, we use cosine similarity. As for the patch token alignment, we employ both cosine similarity and smooth L1 distance. The pseudocode for this process is outlined in Extended Data Table 1. The hyperparameters used in the pretraining phase are provided in Supplementary Table 48. Once the student network is updated, we adopt the exponential moving average (EMA) to update the teacher network (GPFM).

Baselines. To evaluate the performance of our FM, GPFM, we conducted a comprehensive evaluation by comparing it with other vision FMs, namely: Ctranspath²⁴, Phikon²⁵, UNI²⁶, slide-level FM CHIEF⁵⁹ and Prov-Gigapath⁵¹, as well as visual–language FMs PLIP²⁹ and CONCH²⁸. As a baseline, we also compared these FMs with a ResNet50 (ref. 60) pretrained on the ImageNet dataset⁶¹. The model configurations and training details for all these models are presented in Supplementary Table 49. For all downstream tasks, it should be emphasized that feature extraction was consistently performed on images resized to 224×224 resolution, except where explicitly stated otherwise in the experimental protocol.

WSI classification

In CPath, WSI classification typically employs multiple instance learning (MIL) as the underlying methodology. The MIL approach involves the following steps: (1) Non-overlapping tissue patches are cropped

from the original WSI, and features are extracted using a feature extractor. (2) A feature aggregator is applied to integrate the patch-level features into a slide-level feature, enabling classification. To preprocess the WSIs, we use the pipeline described in the CLAM toolkit⁵³. Specifically, we employ the default segmentation configuration of CLAM to extract patches with 512×512 pixels at level 0 for all slides. Slides with a limited number of patches are discarded. Once all patches are extracted, we resize them to 224×224 pixels. We then use FMs to extract features from the resized patches and save these features for subsequent MIL analysis. There are several MIL methods available, such as attention-based multiple instance learning (ABMIL)⁶² and TransMIL⁶³. After evaluating the performance of different FMs across various WSI classification tasks, we found that ABMIL consistently achieves the best results, which aligns with the findings from previous studies^{26,27}. Therefore, we adopted ABMIL to evaluate the performance of different FMs in our experiments. The architecture and training details of ABMIL are presented in Supplementary Table 50. For CHIEF⁵⁹ and Prov-Gigapath⁵¹ models, we used their pretrained slide-level FM to perform classification.

To evaluate the performance of the MIL model, we assessed the balanced accuracy, weighted F1 score, and AUC, which consider the class imbalance present in the dataset. Our experiments encompass 36 pathology WSI classification tasks, including 20 internal and 16 external validation datasets. The results of our experiments are presented in Supplementary Tables 1–18.

NSCLC subtyping on TCGA, CPTAC and Center-1 cohorts (2 classes).

To perform subtyping of non-small-cell lung cancer (NSCLC), we used data from TCGA⁶⁴, CPTAC⁶⁵ and Center-1. The TCGA cohort comprises 541 lung adenocarcinoma (LUAD) and 512 lung squamous cell carcinoma (LUSC) samples. The data are label stratified at a ratio of 7:1:2, resulting in 738 slides for training, 105 slides for validation and 210 slides for testing. For the CPTAC cohort, there are 1,077 LUSC slides and 1,136 LUAD slides. Similarly, this cohort was label stratified at a 7:1:2 ratio, yielding 1,549 slides for training, 222 slides for validation and 442 slides for testing. In addition, we included 180 LUAD slides and 30 LUSC slides from Center-1 for external validation. We directly predicted the subtype of the slides using the model trained on the TCGA cohort. The experimental results are presented in Supplementary Table 2.

Lung cancer metastatic detection and primary site prediction (2 classes and 6 classes).

For metastatic detection, we used 1,198 WSIs from Center-1, comprising 705 patients, including 391 primary cases and 314 metastatic cases. To predict the primary site of metastatic cancer, we curated a dataset with six distinct classes: LUAD (391 cases), breast (55 cases), colon (186 cases), kidney (25 cases), liver (34 cases) and carcinoma of unknown primary (CUP, 14 cases). For both tasks, the data were stratified into training, validation and test sets at a ratio of 7:1:2. In addition, we incorporated an external validation cohort consisting of 530 WSIs (431 cases) from Center-2. For the metastatic detection task, the Center-2 cohort included 238 primary cases and 193 metastatic cases. For the primary site prediction task, the Center-2 cohort comprised 238 LUAD cases, 50 breast cases, 96 colon cases, 30 kidney cases, 10 liver cases and 7 CUP cases. To facilitate distinction between the datasets, we designated the Center-1 cohort as Center-1-LMD and the Center-2 cohort as Center-2-LMD. The experimental results are presented in Supplementary Table 3.

Renal cell carcinoma (RCC) subtyping (3 classes) on TCGA and Center-3 cohorts.

This task contains kidney renal papillary cell carcinoma (KIRP), kidney chromophobe (KICH) and kidney renal clear cell carcinoma (KIRC) WSIs from the TCGA database⁶⁴. After preprocessing, 3 KIRP slides without sufficient foreground were excluded, resulting in 297 KIRP slides, 121 KICH slides and 519 KIRC slides for further analysis. For training and evaluation, we label stratified the TCGA-RCC cohort

into a 7:1:2 train–validation–test split (656:94:187 slides). In addition, we adopted 28 KICH slides, 30 KIRC slides and 30 KIRP slides from Center-3 (Center-3-RCC) as the external cohort. The experimental results are reported in Supplementary Table 4.

CAMELYON for breast metastasis detection (2 classes). This dataset consists of a total of 899 slides, sourced from the Cancer Metastases in Lymph Nodes Challenge 2016 (CAMELYON16, 399 slides)⁶⁶ and the CAMELYON17 (500 slides)⁶⁷. These slides were divided into two classes: normal and metastasis, with a distribution of 557 slides classified as normal and 341 slides classified as metastasis. After image preprocessing, a corrupted normal slide was removed, resulting in a total of 898 WSIs. For training and evaluation, we employed a label-stratified train–validation–test split at a ratio of 7:1:2. This resulted in 630 slides for training, 91 slides for validation and 180 slides for testing. The experimental result is shown in Supplementary Table 5.

Lobular and ductal carcinoma subtyping on TCGA and Center-3 cohorts (2 classes).

We used the TCGA-BRCA dataset⁶⁴ and slides from Center-3 for both internal and external experiments. The TCGA-BRCA dataset contains 787 slides of invasive ductal carcinoma (IDC) and 198 slides of invasive lobular carcinoma (ILC). For training and evaluation, the dataset was stratified by labels into training, validation and testing folds at a ratio of 7:1:2, resulting in 689 slides for training, 99 slides for validation and 197 slides for testing. We also adopted BRCA slides (Center-3-LD) from Center-3 to conduct external validation. This dataset comprises 84 ILC slides and 299 IDC slides. The subtyping results are presented in Supplementary Table 6.

BRACS for breast carcinoma subtyping (3 classes and 7 classes).

This dataset involves 547 breast carcinoma H&E slides obtained from 187 patients⁶⁸. To ensure the quality of the dataset, slides that did not meet the criteria for tumour proportion were excluded, resulting in a total of 545 slides for analysis. The dataset was derived from the breast carcinoma subtyping (BRCA) task, which encompasses both coarse-grained (benign tumours, atypical tumours and malignant tumours) and fine-grained (normal, pathological benign, usual ductal hyperplasia, flat epithelial atypia, atypical ductal hyperplasia, ductal carcinoma in situ and invasive carcinoma) subtyping tasks. For training and evaluation, a label-stratified train–validation–test split was employed, maintaining a ratio of 7:1:2 based on the fine-grained classes. This partitioning resulted in 382 slides for training, 54 slides for validation and 109 slides for testing. In addition, we also adopted 84 normal slides and 383 abnormal slides from Center-3 to perform external validation (Center-3-BRCA). The coarse-grained and fine-grained classification results are presented in Supplementary Tables 7 and 8, respectively.

PANDA for prostate cancer grade assessment (6 classes).

This dataset was designed for prostate cancer grade assessment and consists of a total of 10,616 core needle biopsies sourced from the Prostate cANcer graDe Assessment (PANDA) challenge⁶⁹. After preprocessing, slides without sufficient foreground were excluded, resulting in 10,212 slides available for further analysis. The dataset includes the following subtypes: background or unknown (2,724 slides), stroma (2,602 slides), healthy epithelium (1,321 slides), cancerous epithelium - Gleason 3 (1,205 slides), cancerous epithelium - Gleason 4 (1,187 slides) and cancerous epithelium - Gleason 5 (1,163 slides). For training and evaluation, the train–validation–test cohort was label stratified at a ratio of 7:1:2, resulting in 7,143 slides for training, 1,019 slides for validation and 2,040 slides for testing. The experimental results are reported in Supplementary Table 9.

TCGA-LUAD for lung adenocarcinoma TP53 gene mutation prediction (2 classes).

The LUAD TP53 gene mutation prediction task consists of 469 formalin-fixed paraffin-embedded H&E-stained WSIs

of lung adenocarcinoma sourced from the TCGA database, along with their TP53 gene mutation annotations. Slides without reported TP53 mutation status were excluded from the dataset. WSIs used in this task were classified into 2 classes, namely TP53 mutant (248 slides) and TP53 wildtype (221 slides). For training and evaluation, we label stratified the WSIs into a training–validation–test cohort at a ratio of 7:1:2, including 345 slides for training, 41 slides for validation and 83 slides for testing. The experimental results for TCGA-LUAD TP53 gene mutation prediction could be found in Supplementary Table 10.

The mutation status of IDH in glioma (2 classes). To predict the IDH mutational status in gliomas, we used data from TCGA-GBM and TCGA-LGG, comprising a total of 979 slides, including 722 positive slides and 257 negative slides. For model training and evaluation, the dataset was divided into training, validation and test sets at a label-stratified ratio of 7:1:2. In addition, to validate the robustness of our model, we incorporated an external validation set consisting of 852 slides (322 positives and 530 negatives) from EBRAINS⁷⁰. The detailed experimental results for this task are presented in Supplementary Table 11.

Ovarian cancer subtyping (5 classes) on UBC-OCEAN and Center-3 cohorts. To perform ovarian cancer classification, we adopted the UBC-OCEAN dataset. This dataset is a collection of 538 slides obtained from the UBCOvarian Cancer subtype classification and outlier detection (UBC-OCEAN) competition^{71,72}. The main objective of this competition is to accurately classify ovarian cancer subtypes into five distinct categories. After image preprocessing, the slides without sufficient foregrounds were excluded to reduce data noise, resulting in a total of 527 slides for further analysis. The subtypes of the dataset include clear cell (CC, 98 slides), endometrioid (EC, 122 slides), high-grade serous carcinoma (HGSC, 221 slides), low-grade serous carcinoma (LGSC, 43 slides) and mucinous carcinoma (MC, 43 slides). For training and evaluation, we label stratified the data into train–validation–test folds at a ratio of 7:1:2 (369:52:104 slides). In addition, we also adopted 100 CC, 100 HGSC, 38 LGSC, 97 EC and 35 MC slides from Center-3 as the external validation cohort (Center-3-Ovary). The experimental results are presented in Supplementary Table 12.

Brain tumour subtyping (3 classes). To conduct brain tumour subtyping, we used a dataset of 1,276 slides from TCGA-GBM and TCGA-LGG, comprising 217 oligodendroglioma slides, 164 anaplastic astrocytoma slides and 895 glioblastoma slides. For model training and evaluation, the dataset was label stratified and divided into training, validation and test sets, with 839, 200 and 237 slides, respectively. In addition, we incorporated an external validation set of 732 slides from the EBRAINS Digital Tumour Atlas⁷⁰, which includes 84 oligodendroglioma slides, 89 anaplastic astrocytoma slides and 559 glioblastoma slides. The experimental results for this task are detailed in Supplementary Table 13.

Lesion grade classification of colon cancer. To perform lesion grade classification of colon cancer, we used the IMP-CRS-2024 dataset^{73–75} for experiments. This dataset comprises 847 non-neoplastic slides, 2,847 low-grade lesion slides and 1,638 high-grade lesion slides. We adhered to the official dataset splits, using 3,300 slides from CRS2 for training, 1,132 slides from CRS1 for validation and 900 slides from CRS_Test for testing. In addition, we incorporated an external validation set from Center-3, referred to as Center-3-Colon-WSI, which includes 100 non-neoplastic slides, 121 low-grade lesion slides and 76 high-grade lesion slides. The experimental results for this task are detailed in Supplementary Table 14.

Head and neck cancer primary site prediction and tumor-node-metastasis analysis. We employed the HANCOCK dataset⁷⁶ to predict the primary site of head and neck tumours and to determine the T stage of the tumours. For primary site prediction, we used 708 slides, including 80 hypopharynx slides, 182 larynx slides, 317 oropharynx slides and

129 oral cavity slides. The dataset was label stratified and divided into 495 WSIs for training, 68 WSIs for validation and 145 WSIs for testing. For the tumor-node-metastasis analysis task, we used 705 slides from the HANCOCK dataset to predict the tumour stage (T stage). This dataset comprises 259 T1 slides, 256 T2 slides, 123 T3 slides and 67 T4 slides. The dataset was partitioned into training, validation and testing sets with 496, 67 and 142 slides, respectively. The experimental results for both tasks are presented in Supplementary Table 15.

Lauren subtyping of gastric cancer. We used the TCGA-STAD dataset to conduct Lauren classification. The TCGA-STAD cohort comprises 81 diffuse-type, 125 mixed-type and 184 intestinal-type WSIs. For model training and evaluation, we divided the dataset into training, validation and test sets in a stratified 7:1:2 ratio based on labels. Furthermore, we incorporated 141 WSIs from Center-5 and 319 WSIs from Center-4 as external validation cohorts. The Center-5 cohort consists of 77 diffuse-type, 33 mixed-type and 31 intestinal-type WSIs, while the Center-4 cohort includes 143 diffuse-type, 86 mixed-type and 90 intestinal-type WSIs. We detail the results of these three datasets for this task in Supplementary Table 16.

Vascular invasion detection in gastric cancer. To detect vascular invasion in gastric cancer, we used a dataset comprising 396 WSIs from Center-1, referred to as the Center-1-Vascular dataset. This dataset includes 197 positive cases and 168 negative cases. For the purpose of model training and evaluation, the data were partitioned into training, validation and test sets at a ratio of 7:1:2. In addition, we incorporated two external validation sets: 230 WSIs (140 positive and 90 negative) from Center-5 and 319 WSIs (122 positive and 197 negative) from Center-4. The experimental results of all three datasets for this task are shown in Supplementary Table 17.

Perineural invasion detection in gastric cancer. To detect perineural invasion in gastric cancer, we used a dataset consisting of 397 WSIs obtained from Center-1. This dataset includes 255 positive cases and 141 negative cases. For model training and evaluation, the data were divided into training, validation and test sets at a ratio of 7:1:2. Furthermore, we incorporated two additional external validation sets: 232 WSIs (156 positive and 76 negative) from Center-5 and 319 WSIs (112 positive and 207 negative) from Center-4. See Supplementary Table 18 for experimental results.

Survival analysis

Survival analysis has traditionally been employed to analyse time-to-event data in cancer studies, focusing on events such as disease progression or patient survival. When applied to WSIs, survival analysis offers new opportunities for studying various aspects of tissue behaviour and predicting patient outcomes^{42,77}. By integrating survival analysis with WSIs, researchers can investigate the correlation between specific morphological features and patient outcomes. In our study, we adopted ABMIL⁶² for survival analysis with negative log-likelihood (NLL) loss⁷⁸, following a similar model architecture and training configuration as WSI classification reported in Supplementary Table 50. For CHIEF and Prov-Gigapath models, we used their pretrained slide-level FM to perform classification.

To evaluate the effectiveness of different FMs in survival analysis, we employed a train:test split of 8:2 setting and used the C-index metric to assess performance. We conducted survival analysis on 14 TCGA datasets, including breast cancer (BRCA), bladder cancer (BLCA), kidney renal clear cell carcinoma (KIRC), kidney renal papillary cell carcinoma (KIRP), lung adenocarcinoma (LUAD), stomach adenocarcinoma (STAD), lung squamous cell carcinoma (LUSC), colon adenocarcinoma (COAD), rectum adenocarcinoma (READ), glioblastoma multiforme (GBM), low-grade glioma (LGG), skin cutaneous melanoma (SKCM), cervical squamous cell carcinoma (CESC) and head–neck

squamous cell carcinoma (HNSC). In addition, we performed external validation on the HANCOCK dataset. The number of slides for each dataset is reported in Supplementary Table 51. To ensure robust and consistent results, we maintained uniform censorship (survival status information) between the training and testing datasets. To address the challenge of imbalanced survival times, we employed a stratified approach. Specifically, we sorted the cases on the basis of survival time and divided them into four equally sized bins. We assigned the label of the bin to all cases within it. As a result, we label stratified the train–test cohort into an 8:2 ratio. The experimental results are presented in Supplementary Tables 19–23.

ROI classification

For patch-level tissue classification tasks, we evaluated the transfer performance and representation ability of different FMs using a linear probe, inspired by the approach employed in DINOv2 (refs. 43,79). Initially, we extracted features from the images using the pretrained FMs. Subsequently, we employed a linear layer for performing classification. To optimize the model, we used AdamW⁸⁰ with an initial learning rate of 5×10^{-4} and weight decay of 1×10^{-5} . In addition, we incorporated a cosine annealing scheduler to update the learning rate during training⁸¹. To obtain the best model, we set the maximum number of epochs to 3,000 and implemented early stopping with patience of 100 epochs. For ensuring fair comparison, we maintained a consistent batch size of 256 across all methods.

To evaluate the performance of patch-level tissue classification, we considered the impact of class imbalance in the dataset and assessed the metrics of balanced accuracy, weighted F1 score and AUC. These metrics provide comprehensive insights into classification performance, accounting for both accuracy and the ability to handle imbalanced class distributions. Specifically, we compared the FMs across 16 tasks. For all experiments in this section, we estimated model performance using non-parametric bootstrapping with 1,000 bootstrap replicates. We employed Torchmetrics⁸² for bootstrapping sampling and obtained the mean and standard deviation of the metrics. The experimental results are presented in Supplementary Tables 25–36. Furthermore, we report the average performance of the patch-level tissue classification results across 12 tasks in Supplementary Table 24, demonstrating the superior performance of GPFM.

CRC-100K for colorectal cancer tissue classification (9 classes). This dataset consists of NCT-CRC-HE-100K and CRC-VAL-HE-7K⁴⁴. The NCT-CRC-HE-100K comprises 100,000 non-overlapping 224×224 patches obtained from 86 human cancer tissue slides stained with H&E. These tissue slides were sourced from the NCT biobank (National Center for Tumor Diseases) and the UMM pathology archive (University Medical Center Mannheim). Concurrently, CRC-VAL-HE-7K consists of 7,180 images (224×224) extracted from 50 patients diagnosed with colorectal adenocarcinoma. The subtypes of this dataset contain: adipose (ADI, 11,745 ROIs), background (BACK, 11,413 ROIs), debris (DEB, 11,851 ROIs), lymphocytes (LYM, 12,191 ROIs), mucus (MUC, 9,931 ROIs), smooth muscle (MUS, 14,128 ROIs), normal colon mucosa (NORM, 9,504 ROIs), cancer-associated stroma (STR, 10,867 ROIs) and colorectal adenocarcinoma epithelium (TUM, 15,550 ROIs). For training and evaluation, we use the official train–test split (100,000:7,180). The experimental results are reported in Supplementary Table 25.

CCRCC-TCGA-HEL for CCRCC tissue classification (4 classes). This dataset⁸³ comprises a total of 52,713 ROI images, each with dimensions of 300×300 pixels. The dataset encompasses six distinct categories, namely: renal cancer (cancer, 13,057 ROIs), normal renal tissue (normal, 8,652 ROIs), stromal tissue (stroma, 5,460 ROIs), red blood cells (blood, 996 ROIs), empty background (empty, 16,026 ROIs) and other textures, including necrotic, torn and adipose tissue (other, 8,522 ROIs). The image tiles were selected at random from two sources: the TCGA-KIRC

WSIs and the Helsinki datasets. For training and evaluation, we focused on four specific categories: cancer, stroma, normal and blood. This decision was made due to the potential ambiguities associated with the ‘other’ category and the lack of meaningful information conveyed by the ‘empty’ category. We randomly shuffled the samples and set the train–test split at a 22,530:5,635 ratio. The experimental results are shown in Supplementary Table 26.

BACH for breast cancer tissue classification (4 classes). The dataset⁸⁴ was used for the breast cancer subtyping task and consists of 400 images with dimensions of $2,048 \times 1,536$ pixels. The dataset was labelled into four classes: normal (100 ROIs), benign (100 ROIs), in situ carcinoma (100 ROIs) and invasive carcinoma (100 ROIs). For training and evaluation, all ROIs were resized to 224×224 pixels and we label stratified the train–test data at a ratio of 8:2 (320:80 ROIs). The experimental results are summarized in Supplementary Table 27.

BreakHis for breast cancer image classification (2 classes). This dataset⁸⁵ was collected for breast cancer histopathological image classification and contains two main groups: benign tumours (2,480 ROIs) and malignant tumours (5,429 ROIs). The ROIs in this dataset have 4 different magnifications ($\times 40$, $\times 100$, $\times 200$ and $\times 400$). For training and evaluation, we resized all images to 224×224 pixels to ensure consistency and label stratified the train–test data at a ratio of 8:2 (6,327:1,582 ROIs). The experimental results are presented in Supplementary Table 27.

UniToPatho for CRC polyp classification (6 classes). This dataset is a meticulously annotated dataset comprising 9,536 H&E-stained patches extracted from 292 WSIs⁸⁶. The primary objective of this dataset is to facilitate the training of deep neural networks for the classification of colorectal polyps and the grading of adenomas. The annotations include 6 classes: normal tissue (950 ROIs), hyperplastic polyp (545 ROIs), tubular adenoma with high-grade dysplasia (454 ROIs), tubular adenoma with low-grade dysplasia (3,618 ROIs), tubulo-villous adenoma with high-grade dysplasia (916 ROIs) and tubulo-villous adenoma with low-grade dysplasia (2,186 ROIs). For training and evaluation, we used the official train–test split (6,270:2,399 ROIs). The experimental results are shown in Supplementary Table 28.

CRC-MSI for microsatellite instability (MSI) screening (2 classes). This dataset consists of 51,918 histological images (512×512) of colorectal cancer obtained from the TCGA database⁸⁷. In addition to the visual data, information regarding the MSI status of each patient was obtained. Patients were classified into two categories: those with high MSI (MSI-H) and those with either low (MSI-L) or stable (MSS) microsatellite, collectively referred to as NonMSI-H. For training and evaluation, we used the official train–test split (19,557:32,361 ROIs). The experimental results are shown in Supplementary Table 29.

PanCancer-TCGA for tissue classification (32 classes). This dataset comprises 271,170 images with dimensions of 256×256 pixels⁸⁸. The images were extracted from 8,736 histopathology WSIs obtained from the TCGA database. These images represent various cancer types and are annotated with the following 32 classes: head and neck squamous cell carcinoma (11,790 ROIs), bladder urothelial carcinoma (9,990 ROIs), uterine carcinosarcoma (2,120 ROIs), colon adenocarcinoma (8,150 ROIs), lymphoid neoplasm diffuse large B-cell lymphoma (8,40 ROIs), lung squamous cell carcinoma (16,560 ROIs), brain lower grade glioma (23,530 ROIs), esophageal carcinoma (3,380 ROIs), pheochromocytoma and paraganglioma (1,350 ROIs), sarcoma (13,480 ROIs), glioblastoma multiforme (23,740 ROIs), adrenocortical carcinoma (4,980 ROIs), uterine corpus endometrial carcinoma (12,480 ROIs), prostate adenocarcinoma (9,810 ROIs), breast invasive carcinoma (23,690 ROIs), stomach adenocarcinoma (9,670 ROIs), pancreatic

adenocarcinoma (4,090 ROIs), skin cutaneous melanoma (10,060 ROIs), ovarian serous cystadenocarcinoma (2,520 ROIs), thymoma (3,600 ROIs), lung adenocarcinoma (16,460 ROIs), kidney renal papillary cell carcinoma (6,790 ROIs), testicular germ cell tumours (6,010 ROIs), kidney renal clear cell carcinoma (11,650 ROIs), rectum adenocarcinoma (1,880 ROIs), cholangiocarcinoma (900 ROIs), cervical squamous cell carcinoma and endocervical adenocarcinoma (6,270 ROIs), thyroid carcinoma (11,360 ROIs), mesothelioma (2,090 ROIs), uveal melanoma (1,640 ROIs), liver hepatocellular carcinoma (8,370 ROIs), and kidney chromophobe (2,460 ROIs). For training and evaluation, the train–test split was set to 21,736:54,342 ROIs. The experimental results are summarized in Supplementary Table 30, indicating that GPfM outperforms other models across all three metrics.

TIL classification (2 classes). We used the PanCancer-TIL dataset^{89,90} for tumour-infiltrating lymphocyte classification. It includes 304,097 images with a size of 100×100 pixels at $0.5 \mu\text{m}$ per pixel. The images were labelled with the following two classes: TIL-positive (if there are at least two TILs present in the image; 54,910 ROIs) and TIL-negative (249,187 ROIs). For training and evaluation, we used the official training–validation–test split (209,221:38,601:56,275 ROIs). To ensure consistency, we resized all images to 256×256 pixels. We employed the validation set to select the best model and subsequently evaluated its performance on the test set. In addition, we also adopted the data from Center-3 to conduct external validation. The TIL-negative samples (8,361 ROIs) were obtained from healthy lymph nodes of pancreatic type, and TIL-positive samples (10,131 ROIs) were obtained from the marked cancerous areas on lymph nodes with metastasis. The experimental results are presented in Supplementary Table 31.

ESCA for esophageal carcinoma subtyping (11 classes). This dataset⁹¹ comprises 367,229 images with a size of 256×256 pixels. These patches were obtained from 320 H&E WSIs of esophageal adenocarcinoma and adenocarcinoma of the esophagogastric junction, specifically, 22 slides from University Hospital Cologne (UKK), 62 slides from Landesklinikum Wiener Neustadt (WNS), 22 slides from TCGA and 214 slides from the University Hospital Berlin Charité (CHA). These images were annotated and labelled with one of eleven classes: adventitia (71,131 ROIs), lamina propria mucosae (2,173 ROIs), muscularis mucosae (2,951 ROIs), muscularis propria (83,358 ROIs), regression tissue (56,490 ROIs), mucosa gastric (44,416 ROIs), mucosa oesophagus (18,561 ROIs), submucosa (22,117 ROIs), submucosal glands (1,516 ROIs), tumour (63,863 ROIs) and ulceration (753 ROIs). For training and evaluation, we adopted the CHA dataset, containing 178,187 ROIs, as the training set, and we combined the UKK, WNS and TCGA datasets as a single testing cohort consisting of 189,142 ROIs. In our experiment, all images were resized to 224×224 pixels to ensure consistency, and experimental results are shown in Supplementary Table 32.

PCAM for metastatic tissue classification (2 classes). This dataset consists of 327,680 colour images (96×96 pixels) extracted from CAMELYON16 (refs. 66,92). Each image was annotated with a binary label indicating the presence of metastatic tissue. For training and evaluation, we adopted the official train–validation–test split (262,144:32,768:32,768 ROIs) and resized all images to 224×224 in our experiment. The experimental results are presented in Supplementary Table 33.

WSSS4LUAD for lung adenocarcinoma tissue classification (3 classes). This dataset^{93,94} was collected from Guangdong Provincial People's Hospital (GDPH) and TCGA. It consists of 10,091 images with the following three common and meaningful tissue types: tumour epithelial tissue (6,579 ROIs), tumour-associated stroma tissue (1,680 ROIs) and normal tissue (1,832 ROIs). It is worth noting that in the WSSS4LUAD dataset, one image may belong to several categories. To

avoid ambiguity, we only chose one label for each image on the basis of the order of diagnosability (that is, from tumour epithelial tissue to normal tissue). For training and evaluation, all images were resized to 224×224 pixels and we label stratified the train–test data at a ratio of 8:2 (8,072:2,019 ROIs). The experimental results are presented in Supplementary Table 34.

Chaoyang for colon tissue classification (4 classes). This dataset⁹⁵ contains colon patches from Chaoyang hospital including 1,816 normal ROIs, 1,163 serrated ROIs, 2,244 adenocarcinoma ROIs and 937 adenoma ROIs. For training and evaluation, we resized all patches to 224×224 pixels and used the official train–test split (4,021:2,139 ROIs). In addition, we adopted 9,214 normal ROIs and 11,854 adenoma ROIs from Center-3 for external validation. The experimental results are presented in Supplementary Table 35.

GasHisDB for gastric tissue classification (2 classes). The dataset consists of a total of 13,124 abnormal images (160×160) and 20,160 normal images. For training and evaluation, we resized all patches to 224×224 pixels and label stratified the train–test data at a ratio of 8:2 (26,627:6,657 ROIs). In addition, we adopted the 709 normal tissues and 1,828 abnormal tissues from Center-3 to perform external validation. Results can be found in Supplementary Table 36.

Pathological tissue retrieval

In the linear probe evaluation tasks, we extracted semantically rich features using different FMs and then constructed a task-specific classifier. These features are not only applicable for supervised learning but also prove to be valuable for image-to-image retrieval. The primary goal of this application is to retrieve images that share the same class label as a given query image, thereby facilitating efficient image retrieval. The CRC-100K dataset comprises 100,000 non-overlapping 224×224 patches extracted from 86 human cancer tissue slides stained with H&E for training purposes. In addition, it includes 7,180 images with 224×224 pixels extracted from 50 patients diagnosed with colorectal adenocarcinoma for testing. The dataset consists of multiple classes, including adipose (ADI, 11,745 ROIs), background (BACK, 11,413 ROIs), debris (DEB, 11,851 ROIs), lymphocytes (LYM, 12,191 ROIs), mucus (MUC, 9,931 ROIs), smooth muscle (MUS, 14,128 ROIs), normal colon mucosa (NORM, 9,504 ROIs), cancer-associated stroma (STR, 10,867 ROIs) and colorectal adenocarcinoma epithelium (TUM, 15,550 ROIs). For training and evaluation, we used the official train–test split, with 100,000 samples for training and 7,180 samples for testing.

To initiate the pathological tissue image retrieval process, we began by embedding all images using pretrained FMs. Next, each image in the test set was treated as a query and compared against the images in the training set. To ensure that all features have a comparable impact on the computation of similarity, we independently normalized each feature component to the range $[0, 1]$ ⁹⁶. This normalization process involved calculating the mean and variance of the training set features, which were then used to normalize both the training and testing features.

To evaluate the similarity between the query image and candidate images, we employed the L2 distance metric. A lower distance value indicates a higher degree of similarity between the images. The retrieved images were subsequently ranked on the basis of their similarity scores, and the corresponding class labels were used to evaluate the success of the retrieval process. To assess retrieval performance, we employed evaluation metrics such as Acc@K, where K represents the top-K retrieved images (typically 1, 3 and 5). Similar to the patch-level classification evaluation, we estimated model performance using non-parametric bootstrapping with 1,000 bootstrap replicates. Due to the limitation in the number of classes, we primarily focused on the CRC tissue retrieval tasks, and the experimental results are presented in Supplementary Table 37.

Pathology visual question answering

The objective of this subsection is to evaluate the performance of our proposed pathology FM in the context of VQA tasks. To this end, we used the PathVQA⁹⁷ and the WSI-VQA⁴⁶ datasets as benchmark datasets for our experiments. These datasets provide a comprehensive framework for assessing the model's ability to comprehend and reason about both patch-level and WSI-level visual pathology information, enabling accurate responses to queries related to observed pathological features.

Patch-level VQA on PathVQA dataset. To evaluate the effectiveness of FMs in pathology VQA, we used the PathVQA dataset⁹⁷, which is the largest and most widely used dataset in the pathology domain for VQA tasks. The dataset consists of 32,799 image–question–answer triplets, divided into three subsets: a training set (50%) containing 16,400 triplets used for model training, a validation set (30%) comprising 9,840 triplets for hyperparameter tuning and overfitting prevention, and a test set (20%) including 6,560 triplets for final model performance evaluation. To ensure a rigorous and comparative analysis, we adopted the multimodal unified medical captioning (MUMC) method⁹⁸, which currently represents the state-of-the-art approach for the PathVQA dataset. The MUMC method has exhibited superior performance in leveraging the synergies between visual and textual information for medical image understanding tasks. The experimental results are reported in Supplementary Table 38.

The VQA model architecture consists of four main components: the image encoder, text encoder, multimodal encoder and answering decoder. The image encoder is responsible for capturing domain-specific visual features. We employed various pathology FMs as the image encoder. During the fine-tuning process, the weights of the image encoder were kept frozen to preserve the integrity of the pretrained visual representations and focus on learning task-specific multimodal interactions. The text encoder is designed to process textual inputs, specifically the questions related to the pathology images. We use a 6-layer transformer architecture for the text encoder. It is initialized with the first six layers of a pretrained BERT model, which has a strong track record in language understanding tasks and has demonstrated excellent performance in several medical and clinical applications. The multimodal encoder is responsible for fusing visual and textual features. It consists of the last six layers of the pretrained BERT model and incorporates cross-attention mechanisms at each layer. This integration enables the model to learn robust multimodal interactions, which are crucial for effectively answering questions based on the provided pathology images. The answering decoder, which comprises a 6-layer transformer, receives the multimodal embeddings generated by the previous components and generates text tokens corresponding to the answers. During the training stage, we fine tuned the model for a total of 100 epochs using a batch size of 8. To optimize the model, we employed the AdamW optimizer with an initial learning rate of 2×10^{-5} . Throughout the training process, the learning rate was decayed to 1×10^{-8} to ensure gradual convergence and stability. To evaluate the performance of the VQA models, we adopted accuracy as the metric, which is consistent with previous research studies^{98,99}. We treated VQA as a generative task by calculating similarities between the generated answers and the candidate list of answers, selecting the answer with the highest score as the final answer.

WSI-level VQA on the WSI-VQA dataset. The dataset comprises 977 WSIs and 8,671 question–answer pairs, which are divided into three subsets: training, validation and test. Specifically, the training subset consists of 804 WSIs and 7,139 pairs, while the validation subset includes 87 WSIs and 798 pairs. The test subset contains 86 WSIs and 735 pairs. In the close-ended portion of the test subset, the correct answers were distributed as follows: 151 for option A, 107 for B, 86 for C and 46 for D. For the WSI-VQA dataset, we adhered to the implementation framework proposed by ref. 46, with modifications limited to

replacing the visual features. The experimental results are reported in Supplementary Table 39.

Pathology report generation

The task of pathology report generation is inspired by existing works on chest X-ray and other medical report generation^{100–102}. In this task, the report generation model takes a WSI as input and generates the corresponding pathology report. Specifically, the input WSI is first processed by FMs to extract an initial representation. This representation is then fed into the encoder–decoder architecture of report generation models to produce the decoded pathology report. During this process, the visual encoder further processes the initial representations of WSIs through specific designs^{47,101,102} to obtain the optimal WSI features for the report decoding stage. The text decoder of the model then uses these features for report generation. A good initial representation of WSI could critically facilitate both the visual encoding and textual decoding stages. Consequently, the quality of the generated report is directly influenced by the representations provided by the FMs. In this task, we adopted the HistGen model⁴⁷ for WSI report generation and set the learning rate to 1×10^{-4} and weight decay to 0.8 per epoch. The model was trained for 40 epochs with batch size 1 using features extracted from different FMs.

To evaluate the report generation performance of FMs, we used natural language generation metrics including BLEU¹⁰³, METEOR¹⁰⁴ and ROUGE-L¹⁰⁵, in which BLEU was further split into BLEU-1, BLEU-2, BLEU-3 and BLEU-4 for evaluation of different granularities. These metrics provide a robust framework for evaluating machine-generated text, each bringing unique strengths to assess different aspects of text quality. This task was conducted on the TCGA WSI-Report dataset proposed in ref. 47, containing 7,690 WSIs and the paired diagnosis reports in total, and the PatchGastricADC dataset³⁰ which includes 991 pairs of histological descriptions and WSIs of stomach adenocarcinoma endoscopic biopsy specimens. A 7:1:2 train–validation–test split was employed and the experimental results are reported in Supplementary Tables 40 and 41.

To assess the robustness of each FM in report generation, we conducted a stratified analysis of the TCGA WSI-Report dataset based on cancer types, focusing on major organ cancers including breast, lung and kidney. The stratified evaluation results are presented in Supplementary Table 42. In addition, we collaborated with an experienced pathologist to perform a rigorous human evaluation of the reports generated by different models. The evaluation employed a 4-tier scoring system (illustrated in Extended Data Fig. 8d), and the scoring distribution and average score of each FM are summarized in Supplementary Table 43.

Computing software and hardware

In this project, we used PyTorch¹⁰⁶ (v.2.1.2 with CUDA 12.1) for both pretraining and evaluating downstream tasks. To pretrain the GPFM model, we incorporated established FMs, namely: UNI¹⁰⁷, Phikon¹⁰⁸ and CONCH¹⁰⁹, as additional teachers. It is worth noting that access to UNI and CONCH requires a previous application submission. The GPFM model was pretrained using the FullyShardedDataParallel (FSDP) technique on 2×8 80 GB NVIDIA H800 GPU nodes. All other data processing and evaluation for downstream tasks were carried out on a server equipped with $8 \times$ NVIDIA RTX 3090 GPUs. To assess the model's performance, we employed Torchmetrics (v.1.3.2)⁸² and Scikit-learn (v.1.2.2)¹¹⁰ for metric evaluation. For WSI processing, we relied on OpenSlide-Python (v.1.2.0)⁵² and the CLAM^{53,111} codebase. Pathology VQA evaluation was conducted using the MUMC^{98,112} codebase. Furthermore, for histology report generation, we used the HistGen^{47,113} codebase. Matplotlib (v.3.7.1), seaborn (v.0.13.0) and Origin 2021 were used to plot figures. Please see Supplementary Table 51 for a comprehensive list of the aforementioned models and libraries used in this study.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

This study incorporates a total of 56 datasets. Out of these, 33 datasets were used for pretraining, and a subset of them was also employed for evaluation purposes (TCGA⁶⁴, CPTAC⁶⁵, PANDA⁶⁹, NADT-Prostate¹¹⁴, BCNB¹¹⁵, CAMELYON^{66,67}, BRACS⁶⁸, TIGER2021 (ref. 116), MIDOG2022 (ref. 117), AGGC2022 (ref. 118), O.B.R.^{119,120}, ACROBAT2023 (ref. 121), AML-C-LMU¹²², ARCH¹²³, BACH⁸⁴, CAMEL¹²⁴, DiagSet¹²⁵, DLBCL¹²⁶, GTEx¹²⁷, HunCRC¹²⁸, Janowczyk¹²⁹, LC25000 (ref. 130), MIDOG2021 (ref. 117), OCELOT¹³¹, Oste. Tumor¹³², PAIP2019 (ref. 133), PAIP2020 (ref. 134), PAIP2021, Post-NAT-BRCA¹³⁵, SICAPv2 (ref. 136), SLN-Breast¹³⁷, SPIE2019 (ref. 138)). The remaining 23 datasets were specifically dedicated to downstream task evaluation. The public datasets include PatchGastricADC22 (ref. 50), UBC-OCEAN⁷¹, WSI-VQA⁴⁶, CRC-100K⁴⁴, CRC-MSI⁸⁷, CCRCC-TCGA-HEL⁸³, PanCancer-TCGA⁸⁸, PanCancer-TIL⁸⁹, ESCA⁹¹, PCAM⁹², BreakHis⁸⁵, UniToPatho⁸⁶, Chaoyang⁹⁵, PathVQA⁹⁷, HistGen⁴⁷, IMP-CRS^{73–75}, HANCOCK⁷⁶ and GasHistDB¹³⁹. For detailed information on the public data used in this project, please see Supplementary Table 45. For the data from Center-1 to Center-5, these datasets are not publicly available due to patient privacy obligations, institutional review board requirements and data use agreements. However, researchers interested in accessing de-identified data may submit a reasonable request directly to the corresponding authors, subject to obtaining the necessary ethical approvals and complying with institutional policies. The splits of the dataset can be found in our GitHub repository. Source data are provided with this paper.

Code availability

The code and weights of the GPFM have been made available on GitHub at <https://github.com/birkhoffkiki/GPFM/> (ref. 140).

References

- Niaz, M. K. K., Parwani, A. V. & Gurcan, M. N. Digital pathology and artificial intelligence. *Lancet Oncol.* **20**, 253–261 (2019).
- Deng, S. et al. Deep learning in digital pathology image analysis: a survey. *Front. Med.* **14**, 470–487 (2020).
- Srinidhi, C. L., Ciga, O. & Martel, A. L. Deep neural network models for computational histopathology: a survey. *Med. Image Anal.* **67**, 101813 (2021).
- Song, A. H. et al. Artificial intelligence for digital and computational pathology. *Nat. Rev. Bioeng.* **1**, 930–949 (2023).
- Coudray, N. et al. Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning. *Nat. Med.* **24**, 1559–1567 (2018).
- Bilal, M. et al. Development and validation of a weakly supervised deep learning framework to predict the status of molecular pathways and key mutations in colorectal cancer from routine histology images: a retrospective study. *Lancet Digit. Health* **3**, 763–772 (2021).
- Zamanitajeddin, N., Jahanifar, M., Bilal, M., Eastwood, M. & Rajpoot, N. Social network analysis of cell networks improves deep learning for prediction of molecular pathways and key mutations in colorectal cancer. *Med. Image Anal.* **93**, 103071 (2024).
- Wulczyn, E. et al. Interpretable survival prediction for colorectal cancer using deep learning. *npj Digit. Med.* **4**, 71 (2021).
- Mobadersany, P. et al. Predicting cancer outcomes from histology and genomics using convolutional networks. *Proc. Natl Acad. Sci. USA* **115**, 2970–2979 (2018).
- Courtiol, P. et al. Deep learning-based classification of mesothelioma improves prediction of patient outcome. *Nat. Med.* **25**, 1519–1525 (2019).
- Vanguri, R. S. et al. Multimodal integration of radiology, pathology and genomics for prediction of response to PD-(L)1 blockade in patients with non-small cell lung cancer. *Nat. Cancer* **3**, 1151–1164 (2022).
- Zhang, Y. et al. Histopathology images-based deep learning prediction of prognosis and therapeutic response in small cell lung cancer. *npj Digit. Med.* **7**, 15 (2024).
- Hu, J. et al. Using deep learning to predict anti-PD-1 response in melanoma and lung cancer patients from histopathology images. *Transl. Oncol.* **14**, 100921 (2021).
- Kang, M., Song, H., Park, S., Yoo, D. & Pereira, S. Benchmarking self-supervised learning on diverse pathology datasets. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition* (eds Geiger, A. et al.) 3344–3354 (IEEE/CVF, 2023).
- Li, B., Li, Y. & Eliceiri, K. W. Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition* (eds Forsyth, D. et al.) 14318–14328 (IEEE/CVF, 2021).
- Lazard, T., Lerousseau, M., Decencière, E. & Walter, T. Giga-ssl: self-supervised learning for gigapixel images. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition* (eds Geiger A., et al.) 4304–4313 (IEEE/CVF, 2023).
- Schirris, Y., Gavves, E., Nederlof, I., Horlings, H. M. & Teuwen, J. Deepsmile: contrastive self-supervised pre-training benefits msi and hrd classification directly from H&E whole-slide images in colorectal and breast cancer. *Med. Image Anal.* **79**, 102464 (2022).
- Vu, Q. D., Rajpoot, K., Raza, S. E. A. & Rajpoot, N. Handcrafted histological transformer (H2T): unsupervised representation of whole slide images. *Med. Image Anal.* **85**, 102743 (2023).
- Claudio Quiros, A. et al. Adversarial learning of cancer tissue representations. In *Proc. Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference* (eds de Bruijne, M. et al.) 602–612 (MICCAI, 2021).
- Jiang, S., Hondelink, L., Suriawinata, A. A. & Hassanpour, S. Masked pre-training of transformers for histology image analysis. *J. Pathol. Inform.* **15**, 100386 (2024).
- Bommasani, R. et al. On the opportunities and risks of foundation models. Preprint at <https://doi.org/10.48550/arXiv.2108.07258> (2021).
- Zhou, C. et al. A comprehensive survey on pretrained foundation models: a history from bert to chatgpt. *Int. J. Mach. Learn. Cybern.* 1–65 (2024).
- Moor, M. et al. Foundation models for generalist medical artificial intelligence. *Nature* **616**, 259–265 (2023).
- Wang, X. et al. Transformer-based unsupervised contrastive learning for histopathological image classification. *Med. Image Anal.* **81**, 102559 (2022).
- Filiot, A. et al. Scaling self-supervised learning for histopathology with masked image modeling. Preprint at medRxiv <https://doi.org/10.1101/2023.07.21.23292757> (2023).
- Chen, R. J. et al. Towards a general-purpose foundation model for computational pathology. *Nat. Med.* **30**, 850–862 (2024).
- Vorontsov, E. et al. A foundation model for clinical-grade computational pathology and rare cancers detection. *Nat. Med.* **30**, 2924–2935 (2024).
- Lu, M. Y. et al. A visual-language foundation model for computational pathology. *Nat. Med.* **30**, 863–874 (2024).
- Huang, Z., Bianchi, F., Yuksekogonul, M., Montine, T. J. & Zou, J. A visual-language foundation model for pathology image analysis using medical twitter. *Nat. Med.* **29**, 2307–2316 (2023).
- Chen, X., Xie, S. & He, K. An empirical study of training self-supervised vision transformers. In *Proc. IEEE/CVF International Conference on Computer Vision* (eds Damen, D. et al.) 9640–9649 (IEEE/CVF, 2021).

31. Radford, A. et al. Learning transferable visual models from natural language supervision. In *Proc. International Conference on Machine Learning* (ed. Lawrence, N.) 8748–8763 (PMLR, 2021).
32. Yu, J. et al. CoCa: Contrastive Captioners are Image–Text Foundation Models (TMLR, 2023).
33. Hinton, G., Vinyals, O. & Dean, J. Distilling the knowledge in a neural network. Preprint at <https://doi.org/10.48550/arXiv.1503.02531> (2015).
34. Gou, J., Yu, B., Maybank, S. J. & Tao, D. Knowledge distillation: a survey. *Int. J. Comput. Vis.* **129**, 1789–1819 (2021).
35. Demvsar, J. Statistical comparisons of classifiers over multiple data sets. *J. Mach. Learn. Res.* **7**, 1–30 (2006).
36. Maier-Hein, L. et al. Why rankings of biomedical image analysis competitions should be interpreted with care. *Nat. Commun.* **9**, 5217 (2018).
37. Antonelli, M. et al. The medical segmentation decathlon. *Nat. Commun.* **13**, 4128 (2022).
38. Wang, P., Li, Y. & Reddy, C. K. Machine learning for survival analysis: a survey. *ACM Comput. Surv.* **51**, 1–36 (2019).
39. Xu, Y. & Chen, H. Multimodal optimal transport-based co-attention transformer with global structure consistency for survival prediction. In *Proc. IEEE/CVF International Conference on Computer Vision* (eds Geiger, A. et al.) 21241–21251 (IEEE/CVF, 2023).
40. Zhang, Y., Xu, Y., Chen, J., Xie, F. & Chen, H. Prototypical information bottlenecking and disentangling for multimodal cancer survival prediction. In *Proc. 12th International Conference on Learning Representations* (eds Chaudhuri, S. et al.) (ICLR, 2024).
41. Zhou, F. & Chen, H. Cross-modal translation and alignment for survival analysis. In *Proc. IEEE/CVF International Conference on Computer Vision* (eds Geiger, A. et al.) 21485–21494 (IEEE/CVF, 2023).
42. Wiegrebbe, S., Kopper, P., Sonabend, R., Bischl, B. & Bender, A. Deep learning for survival analysis: a review. *Artif. Intell. Rev.* **57**, 65 (2024).
43. Oquab, M. et al. *Dinov2: Learning Robust Visual Features Without Supervision* (TMLR, 2024).
44. Kather, J. N. et al. Predicting survival from colorectal cancer histology slides using deep learning: a retrospective multicenter study. *PLoS Med.* **16**, 1002730 (2019).
45. Maaten, L. & Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **9**, 2579–2605 (2008).
46. Chen, P., Zhu, C., Zheng, S., Li, H. & Yang, L. Wsi-vqa: interpreting whole slide images by generative visual question answering. In *European Conference on Computer Vision* (eds Leonardi, E. et al.) 401–417 (Springer, 2025).
47. Guo, Z. et al. Histgen: histopathology report generation via local-global feature encoding and cross-modal context interaction. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* (eds Linguraru, M.G. et al.) 189–199 (Springer, 2024).
48. Guevara, B. C. et al. Caption generation from histopathology whole-slide images using pre-trained transformers. In *Proc. Medical Imaging with Deep Learning, Short Paper Track* (eds Oguz, I. et al.) (MIDL, 2023).
49. Chen, P. et al. Wsicap: Multiple instance generation of pathology reports for gigapixel whole-slide images. In *Proc. Medical Image Computing and Computer Assisted Intervention – MICCAI 2024* (eds Linguraru, M. G. et al.) 546–556 (Springer Nature, 2024).
50. Tsuneki, M. & Kanavati, F. Inference of captions from histopathological patches. In *International Conference on Medical Imaging with Deep Learning* (ed. Lawrence, N.) 1235–1250 (PMLR, 2022).
51. Xu, H. et al. A whole-slide foundation model for digital pathology from real-world data. *Nature* **630**, 181–188 (2024).
52. Goode, A., Gilbert, B., Harkes, J., Jukic, D. & Satyanarayanan, M. Openslide: a vendor-neutral software foundation for digital pathology. *J. Pathol. Inform.* **4**, 27 (2013).
53. Lu, M. Y. et al. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nat. Biomed. Eng.* **5**, 555–570 (2021).
54. Zhou, J. et al. ibot: image bert pre-training with online tokenizer. In *Proc. International Conference on Learning Representations* (eds Liu, Y. et al.) 716 (ICLR, 2022).
55. He, K. et al. Masked autoencoders are scalable vision learners. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition* (eds Dana, K. et al.) 16000–16009 (IEEE/CVF, 2022).
56. Zhang, H. et al. DINO: DETR with improved denoising anchor boxes for end-to-end object detection. In *The Eleventh International Conference on Learning Representations* (eds Liu, Y. et al.) 1581 (ICLR, 2022).
57. Ranzinger, M., Heinrich, G., Kautz, J. & Molchanov, P. AM-Radio: agglomerative visual foundation model – reduce all domains into one. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (eds Akata, Z. et al.) 12490–12500 (IEEE/CVF, 2024).
58. Zimmermann, E. et al. Virchow2: scaling self-supervised mixed magnification models in pathology. Preprint at <https://doi.org/10.48550/arXiv.2408.00738> (2024).
59. Wang, X. et al. A pathology foundation model for cancer diagnosis and prognosis prediction. *Nature* **634**, 970–978 (2024).
60. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition* (eds Agapito, L. et al.) 770–778 (IEEE, 2016).
61. Deng, J. et al. Imagenet: a large-scale hierarchical image database. In *Proc. 2009 IEEE Conference on Computer Vision and Pattern Recognition* (eds Essa, I. et al.) 248–255 (IEEE, 2009).
62. Ilse, M., Tomczak, J. & Welling, M. Attention-based deep multiple instance learning. In *Proc. International Conference on Machine Learning* (ed. Lawrence, N.) 2127–2136 (PMLR, 2018).
63. Shao, Z. et al. Transmil: transformer based correlated multiple instance learning for whole slide image classification. *Adv. Neural Inf. Process. Syst.* **34**, 2136–2147 (2021).
64. Weinstein, J. N. et al. The cancer genome atlas pan-cancer analysis project. *Nat. Genet.* **45**, 1113–1120 (2013).
65. Edwards, N. J. et al. The cptac data portal: a resource for cancer proteomics research. *J. Proteome Res.* **14**, 2707–2713 (2015).
66. Bejnordi, B. E. et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *JAMA* **318**, 2199–2210 (2017).
67. Bandi, P. et al. From detection of individual metastases to classification of lymph node status at the patient level: the camelyon17 challenge. *IEEE Trans. Med. Imaging* **38**, 550–560 (2018).
68. Brancati, N. et al. Bracs: a dataset for breast carcinoma subtyping in H&E histology images. *Database* **2022**, 093 (2022).
69. Bulten, W. et al. Artificial intelligence for diagnosis and gleason grading of prostate cancer: the panda challenge. *Nat. Med.* **28**, 154–163 (2022).
70. Roetzer-Pejrimovsky, T. et al. The digital brain tumour atlas, an open histopathology resource. *Sci. Data* **9**, 55 (2022).
71. Asadi-Aghbolaghi, M. et al. Machine learning-driven histotype diagnosis of ovarian carcinoma: insights from the ocean ai challenge. Preprint at *medRxiv* <https://doi.org/10.1101/2024.04.19.24306099> (2024).
72. Farahani, H. et al. Deep learning-based histotype diagnosis of ovarian carcinoma whole-slide pathology images. *Mod. Pathol.* **35**, 1983–1990 (2022).

73. Oliveira, S. P. et al. CAD systems for colorectal cancer from WSI are still not ready for clinical acceptance. *Sci. Rep.* **11**, 14358 (2021).
74. Neto, P. C. et al. imil4path: a semi-supervised interpretable approach for colorectal whole-slide images. *Cancers* **14**, 2489 (2022).
75. Neto, P. C. et al. An interpretable machine learning system for colorectal cancer diagnosis from pathology slides. *npj Precis. Oncol.* **8**, 56 (2024).
76. Dörrich, M. et al. A multimodal dataset for precision oncology in head and neck cancer. *Nat. Commun.* **16**, 7163 (2025).
77. Chen, R. J. et al. Whole slide images are 2d point clouds: context-aware survival prediction using patch-based graph convolutional networks. In *Proc. 24th International Conference on Medical Image Computing and Computer Assisted Intervention–MICCAI 2021* (eds de Bruijne, M. et al.) 339–349 (Springer, 2021).
78. Zadeh, S. G. & Schmid, M. Bias in cross-entropy-based training of deep survival networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **43**, 3126–3137 (2020).
79. Balestrieri, R. et al. A cookbook of self-supervised learning. Preprint at <https://doi.org/10.48550/arXiv.2304.12210> (2023).
80. Loshchilov, I. & Hutter, F. Decoupled weight decay regularization. In *Proc. International Conference on Learning Representations* (eds Rush, A. et al.) 939 (ICLR, 2019).
81. Loshchilov, I. & Hutter, F. SGDR: stochastic gradient descent with warm restarts. In *Proc. International Conference on Learning Representations* (eds Ranzato, M. A. et al.) (ICLR, 2017).
82. Detlefsen, N. S. et al. TorchMetrics-measuring reproducibility in PyTorch. *J. Open Source Softw.* **7**, 4101 (2022).
83. Brummer, O., Pölönen, P., Mustjoki, S. & Brück, O. Computational textural mapping harmonises sampling variation and reveals multidimensional histopathological fingerprints. *Br. J. Cancer* **129**, 683–695 (2023).
84. Aresta, G. et al. BACH: grand challenge on breast cancer histology images. *Med. Image Anal.* **56**, 122–139 (2019).
85. Spanhol, F. A., Oliveira, L. S., Petitjean, C. & Heutte, L. A dataset for breast cancer histopathological image classification. *IEEE Trans. Biomed. Eng.* **63**, 1455–1462 (2015).
86. Barbano, C. A. et al. Unitopatho, a labeled histopathological dataset for colorectal polyps classification and adenoma dysplasia grading. In *2021 IEEE International Conference on Image Processing (ICIP)* (eds Dufaux, E. & Wang, Z. J.) 76–80 (IEEE, 2021).
87. Kather, J. N. et al. Deep learning can predict microsatellite instability directly from histology in gastrointestinal cancer. *Nat. Med.* **25**, 1054–1056 (2019).
88. Komura, D. et al. Universal encoding of pan-cancer histology by deep texture representations. *Cell Rep.* **38**, 110424 (2022).
89. Abousamra, S. et al. Deep learning-based mapping of tumor infiltrating lymphocytes in whole slide images of 23 types of cancer. *Front. Oncol.* **11**, 806603 (2022).
90. Saltz, J. et al. Spatial organization and molecular correlation of tumor-infiltrating lymphocytes using deep learning on pathology images. *Cell Rep.* **23**, 181–193 (2018).
91. Tolkach, Y. et al. Artificial intelligence for tumour tissue detection and histological regression grading in oesophageal adenocarcinomas: a retrospective algorithm development and validation study. *Lancet Digit. Health* **5**, 265–275 (2023).
92. Veeling, B. S., Linmans, J., Winkens, J., Cohen, T. & Welling, M. Rotation equivariant CNNs for digital pathology. In *Proc. 21st International Conference on Medical Image Computing and Computer Assisted Intervention–MICCAI 2018* (eds Frangi, A. F. et al.) 210–218 (Springer, 2018).
93. Han, C. et al. WSSS4LUAD: grand challenge on weakly-supervised tissue semantic segmentation for lung adenocarcinoma. Preprint at <https://doi.org/10.48550/arXiv.2204.06455> (2022).
94. Han, C. et al. Multi-layer pseudo-supervision for histopathology tissue semantic segmentation using patch-level classification labels. *Med. Image Anal.* **80**, 102487 (2022).
95. Zhu, C., Chen, W., Peng, T., Wang, Y. & Jin, M. Hard sample aware noise robust learning for histopathology image classification. *IEEE Trans. Med. Imaging* **41**, 881–894 (2021).
96. Aksoy, S. & Haralick, R. M. Probabilistic vs. geometric similarity measures for image retrieval. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition* (eds Forsyth, D. & Kreigman, D.) 357–362 (IEEE, 2000).
97. He, X., Zhang, Y., Mou, L., Xing, E. & Xie, P. PathVQA: 30000+ questions for medical visual question answering. Preprint at <https://doi.org/10.48550/arXiv.2003.10286> (2020).
98. Li, P., Liu, G., He, J., Zhao, Z. & Zhong, S. Masked vision and language pre-training with unimodal and multimodal contrastive losses for medical visual question answering. In *Proc. International Conference on Medical Image Computing and Computer-Assisted Intervention* (eds Greenspan, H. et al.) 374–383 (Springer, 2023).
99. Gong, H., Chen, G., Mao, M., Li, Z. & Li, G. VQAMix: conditional triplet mixup for medical visual question answering. *IEEE Trans. Med. Imaging* **41**, 3332–3343 (2022).
100. Chen, Z., Shen, Y., Song, Y. & Wan, X. Cross-modal memory networks for radiology report generation. In *Proc. 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing* (eds Zong, C. et al.) 5904–5914 (Association for Computational Linguistics, 2021).
101. Chen, Z., Song, Y., Chang, T.-H. & Wan, X. Generating radiology reports via memory-driven transformer. In *Proc. 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (eds Webber, B. et al.) 1439–1449 (Association for Computational Linguistics, 2020).
102. Li, M. et al. Cross-modal clinical graph transformer for ophthalmic report generation. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition* (eds Dana, K. et al.) 20656–20665 (IEEE/CVF, 2022).
103. Papineni, K., Roukos, S., Ward, T. & Zhu, W.-J. BLEU: a method for automatic evaluation of machine translation. In *Proc. 40th Annual Meeting of the Association for Computational Linguistics* (eds Isabelle, P. et al.) 311–318 (ACL, 2002).
104. Denkowski, M. & Lavie, A. Meteor 1.3: automatic metric for reliable optimization and evaluation of machine translation systems. In *Proc. Sixth Workshop on Statistical Machine Translation* (eds Callison-Burch, C. et al.) 85–91 (ACL, 2011).
105. Lin, C.-Y. ROUGE: a package for automatic evaluation of summaries. In *Text Summarization Branches Out* 74–81 (ACL, 2004).
106. Paszke, A. et al. PyTorch: an imperative style, high-performance deep learning library. *Adv. Neural Inf. Process. Syst.* **32**, 721 (2019).
107. Chen, R. J. et al. Towards a general-purpose foundation model for computational pathology. *GitHub* <https://github.com/mahmoodlab/UNI> (2024).
108. Filiot, A. Scaling self-supervised learning for histopathology with masked image modeling. *GitHub* <https://github.com/owkin/HistoSSLscaling> (2023).
109. Lu, M. Y., Chen, B. & Mahmood, F. A visual-language foundation model for computational pathology. *GitHub* <https://github.com/mahmoodlab/CONCH> (2023).
110. Pedregosa, F. et al. Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
111. Lu, M. Y. et al. CLAM: A deep-learning-based pipeline for data efficient and weakly supervised whole-slide-level analysis. *GitHub* <https://github.com/mahmoodlab/CLAM> (2021).

112. Li, P. MUMC: Masked vision and language pre-training with unimodal and multimodal contrastive losses for medical visual question answering. *GitHub* <https://github.com/pengfeiliHEU/MUMC> (2023).
113. Guo, Z. HistGen: histopathology report generation via local-global feature encoding and cross-modal context interaction. *GitHub* <https://github.com/ddavid4real/HistGen> (2024).
114. Wilkinson, S. et al. Nascent prostate cancer heterogeneity drives evolution and resistance to intense hormonal therapy. *Eur. Urol.* **80**, 746–757 (2021).
115. Xu, F. et al. Predicting axillary lymph node metastasis in early breast cancer using deep learning on primary tumor biopsy slides. *Front. Oncol.* **11**, 759007 (2021).
116. Shephard, A. et al. TIAger: tumor-infiltrating lymphocyte scoring in breast cancer for the tiger challenge. Preprint at <https://doi.org/10.48550/arXiv.2206.11943> (2022).
117. Aubreville, M. et al. Domain generalization across tumor types, laboratories, and species—insights from the 2022 edition of the mitosis domain generalization challenge. *Med. Image Anal.* **94**, 103155 (2024).
118. Huo, X. et al. A comprehensive ai model development framework for consistent gleason grading. *Commun. Med.* **4**, 84 (2024).
119. Wang, C.-W. et al. Weakly supervised deep learning for prediction of treatment effectiveness on ovarian cancer from histopathology images. *Comput. Med. Imaging Graph.* **99**, 102093 (2022).
120. Wang, C.-W. et al. Histopathological whole slide image dataset for classification of treatment effectiveness to ovarian cancer. *Sci. Data* **9**, 25 (2022).
121. Weitz, P. et al. A multi-stain breast cancer histological whole-slide-image data set from routine diagnostics. *Sci. Data* **10**, 562 (2023).
122. Matek, C., Schwarz, S., Marr, C. & Spiekermann, K. A single-cell morphological dataset of leukocytes from AML patients and non-malignant controls [data set]. *The Cancer Imaging Archive* <https://doi.org/10.7937/tcia.2019.36f5o9ld> (2019).
123. Gamper, J. & Rajpoot, N. Multiple instance captioning: learning representations from histopathology textbooks and articles. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition* (eds Forsyth, D. et al.) 16549–16559 (IEEE/CVF, 2021).
124. Xu, G. et al. CAMEL: a weakly supervised learning framework for histopathology image segmentation. In *Proc. IEEE/CVF International Conference on Computer Vision* (eds Kweon, I. S. et al.) 10682–10691 (IEEE/CVF, 2019).
125. Koziarski, M. et al. DiagSet: a dataset for prostate cancer histopathological image classification. *Sci. Rep.* **14**, 6780 (2024).
126. Vrabac, D. et al. DLBCL-Morph: morphological features computed using deep learning for an annotated digital DLBCL image set. *Sci. Data* **8**, 135 (2020).
127. Carithers, L. J. et al. A novel approach to high-quality postmortem tissue procurement: the GTEx project. *Biopreserv. Biobank.* **13**, 311–319 (2015).
128. Pataki, B. Á. et al. HunCRC: annotated pathological slides to enhance deep learning applications in colorectal cancer screening. *Sci. Data* **9**, 370 (2022).
129. Janowczyk, A. & Madabhushi, A. Deep learning for digital pathology image analysis: a comprehensive tutorial with selected use cases. *J. Pathol. Inform.* **7**, 29 (2016).
130. Borkowski, A. A. et al. Lung and colon cancer histopathological image dataset (LC25000). Preprint at <https://doi.org/10.48550/arXiv.1912.12142> (2019).
131. Ryu, J. et al. OCELOT: overlapped cell on tissue dataset for histopathology. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition* (eds Geiger, A. et al.) 23902–23912 (IEEE/CVF, 2023).
132. Leavey, P. et al. Osteosarcoma data from UT Southwestern/UT Dallas for viable and necrotic tumor assessment [data set]. *The Cancer Imaging Archive* <https://doi.org/10.7937/tcia.2019.bvhjhdas> (2019).
133. Kim, Y. J. et al. PAIP 2019: liver cancer segmentation challenge. *Med. Image Anal.* **67**, 101854 (2021).
134. Kim, K. et al. PAIP 2020: microsatellite instability prediction in colorectal cancer. *Med. Image Anal.* **89**, 102886 (2023).
135. Tafavvoghi, M., Bongo, L. A., Shvetsov, N., Busund, L.-T. R. & Møllersen, K. Publicly available datasets of breast histopathology H&E whole-slide images: a scoping review. *J. Pathol. Inform.* **15**, 100363 (2024).
136. Silva-Rodríguez, J., Colomer, A., Sales, M. A., Molina, R. & Naranjo, V. Going deeper through the gleason scoring scale: an automatic end-to-end system for histology prostate grading and cribriform pattern detection. *Comput. Methods Programs Biomed.* **195**, 105637 (2020).
137. Kemaloglu, N., Aydoğan, T. & Küçüksille, E. U. in *Artificial Intelligence for Data-Driven Medical Diagnosis* (eds Gupta, D. et al.) 55–84 (De Gruyter, 2021).
138. Petrick, N. et al. SPIE-AAPM-NCI BreastPathQ challenge: an image analysis challenge for quantitative tumor cellularity assessment in breast cancer histology images following neoadjuvant treatment. *J. Med. Imaging* **8**, 034501 (2021).
139. Hu, W. et al. GasHisSDB: a new gastric histopathology image dataset for computer aided diagnosis of gastric cancer. *Comput. Biol. Med.* **142**, 105207 (2022).
140. Ma, J. et al. A generalizable pathology foundation model using a unified knowledge distillation pretraining framework. *GitHub* <https://github.com/birkhoffkiki/GPFM/> (2024).

Acknowledgements

This work was supported by the National Natural Science Foundation of China (No. 62202403), the Hong Kong Innovation and Technology Commission (Project No. PRP/034/22FX and No. ITCPD/17-9), the Research Grants Council of the Hong Kong Special Administrative Region, China (No. R6003-22) and the HKUST Frontier Technology Research for Joint Institutes with Industry (No. OKT24EG01). We thank the support of HKUST SuperPod in providing the GPU platform for foundation model pretraining.

Author contributions

J.M. conceived the study and designed the experiments. J.M., Z.G. and F.Z. collected the data for self-supervised learning and downstream task evaluation. J.M. performed model pretraining and conducted patch-level tissue classification tasks. Y.L. and X.J. participated in discussions regarding the design of the self-supervised learning framework and were responsible for reproducing the foundation models. J.M., F.Z. and Y.C. evaluated the weakly supervised WSI classification task. Z.G. performed survival analysis and report generation tasks. Y.W. and Y.X. conducted pathological image retrieval and curated the data of WSI-report pairs. Z.Z. performed pathological image visual question answering. C.J. assisted in result analysis and the creation of visualized attention maps. J.M. and Z.G. prepared the manuscript with input from all co-authors. R.C.K.C, A.H. and L.L. provided medical guidance. Z.L., J.L., C.Z. and D.L. provided and preprocessed data for some downstream tasks. S.Z. and F.Y. provided preprocessed data for external validation. P.F. and J.W. offered insightful suggestions for the experimental design and thoughtfully directing the research trajectory. K.-T.C reviewed and refined the draft. H.C. supervised the research.

Ethics declaration

This project has been reviewed and approved by the Human and Artefacts Research Ethics Committee (HAREC) of Hong Kong University of Science and Technology (protocol no. HREP-2024-0212).

Competing interests

The authors declare no competing interests.

Additional information

Extended data is available for this paper at <https://doi.org/10.1038/s41551-025-01488-4>.

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41551-025-01488-4>.

Correspondence and requests for materials should be addressed to Shaoting Zhang, Li Liang or Hao Chen.

Peer review information *Nature Biomedical Engineering* thanks Zhi Huang and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

Reprints and permissions information is available at www.nature.com/reprints.

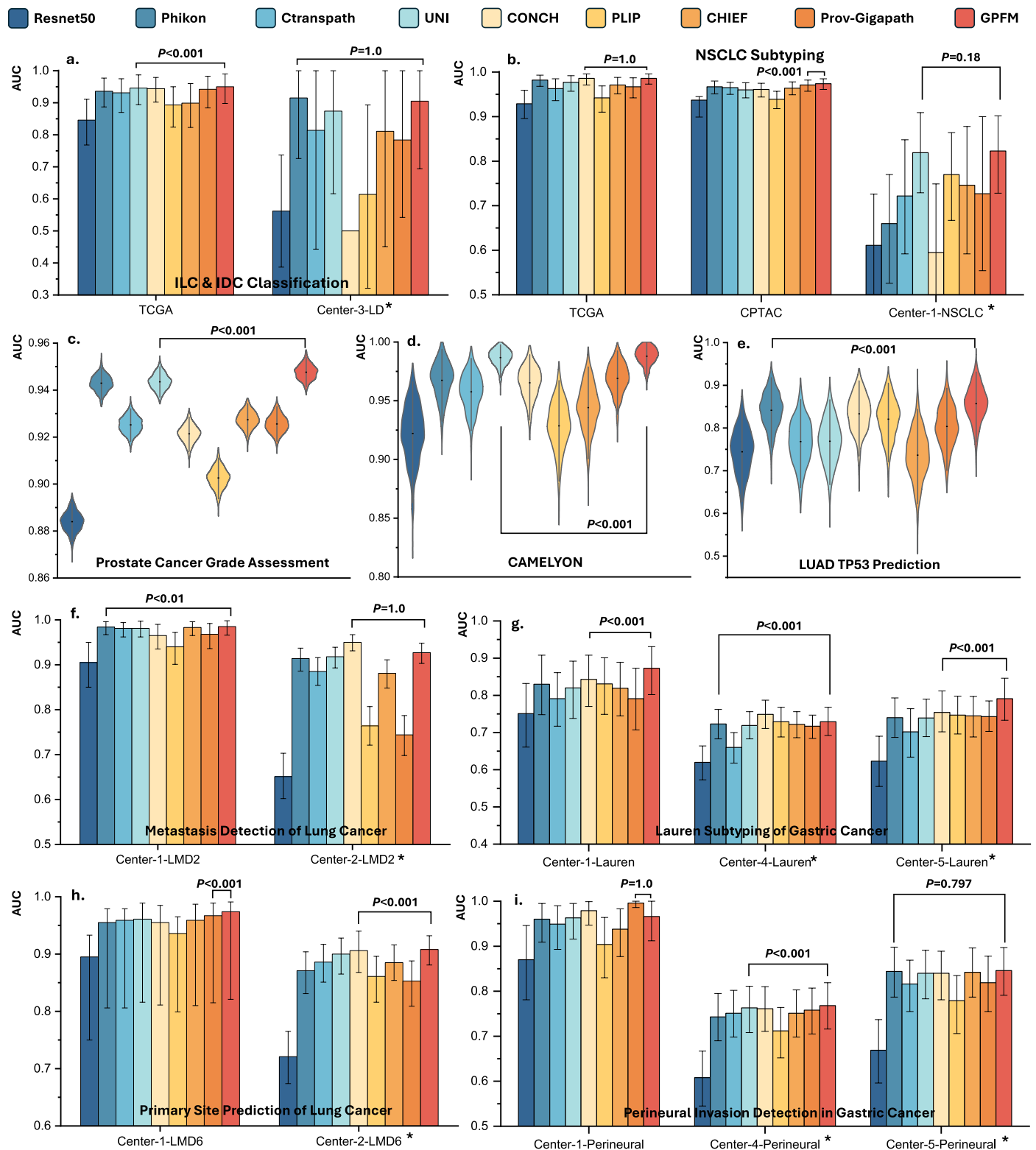
Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

© The Author(s), under exclusive licence to Springer Nature Limited 2025

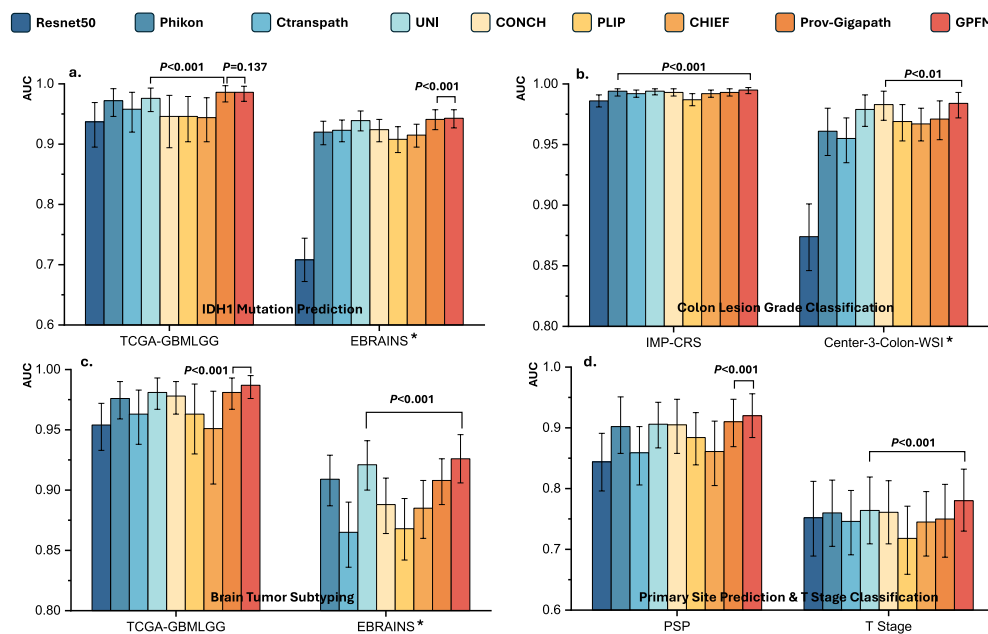
¹Department of Computer Science and Engineering, The Hong Kong University of Science and Technology, Hong Kong SAR, China. ²Department of Pathology, Nanfang Hospital and School of Basic Medical Sciences, Southern Medical University, Guangzhou, China. ³Guangdong Provincial Key Laboratory of Molecular Tumor Pathology, Guangzhou, China. ⁴Shanghai Artificial Intelligence Laboratory, Shanghai, China. ⁵Department of Electronic and Computer Engineering, The Hong Kong University of Science and Technology, Hong Kong SAR, China. ⁶Information Hub, The Hong Kong University of Science and Technology (Guangzhou), Guangzhou, China. ⁷Department of Pathology, The First Affiliated Hospital of Shandong First Medical University and Shandong Provincial Qianfoshan Hospital, Jinan, China. ⁸Department of Pathology, The First Affiliated Hospital, Sun Yat-sen University, Guangzhou, China. ⁹Department of Radiology, The Third Affiliated Hospital of Kunming Medical University, Yunnan Cancer Hospital, Kunming, China. ¹⁰Department of Anatomical and Cellular Pathology, The Chinese University of Hong Kong, Hong Kong SAR, China. ¹¹Department of Chemical and Biological Engineering, The Hong Kong University of Science and Technology, Hong Kong SAR, China. ¹²Division of Life Science, The Hong Kong University of Science and Technology, Hong Kong SAR, China. ¹³School of Optical Electronic Information, Huazhong University of Science and Technology, Wuhan, China. ¹⁴Qing Yuan Research Institute, Shanghai Jiao Tong University, Shanghai, China. ¹⁵Jinfeng Laboratory, Chongqing, China. ¹⁶State Key Laboratory of Molecular Neuroscience, The Hong Kong University of Science and Technology, Hong Kong SAR, China. ¹⁷Shenzhen–Hong Kong Collaborative Innovation Research Institute, The Hong Kong University of Science and Technology, Shenzhen, China. ¹⁸These authors contributed equally: Jiabo Ma, Zhengrui Guo.

✉ e-mail: zhangshaoting@pjlab.org.cn; lli@smu.edu.cn; jhc@cse.ust.hk



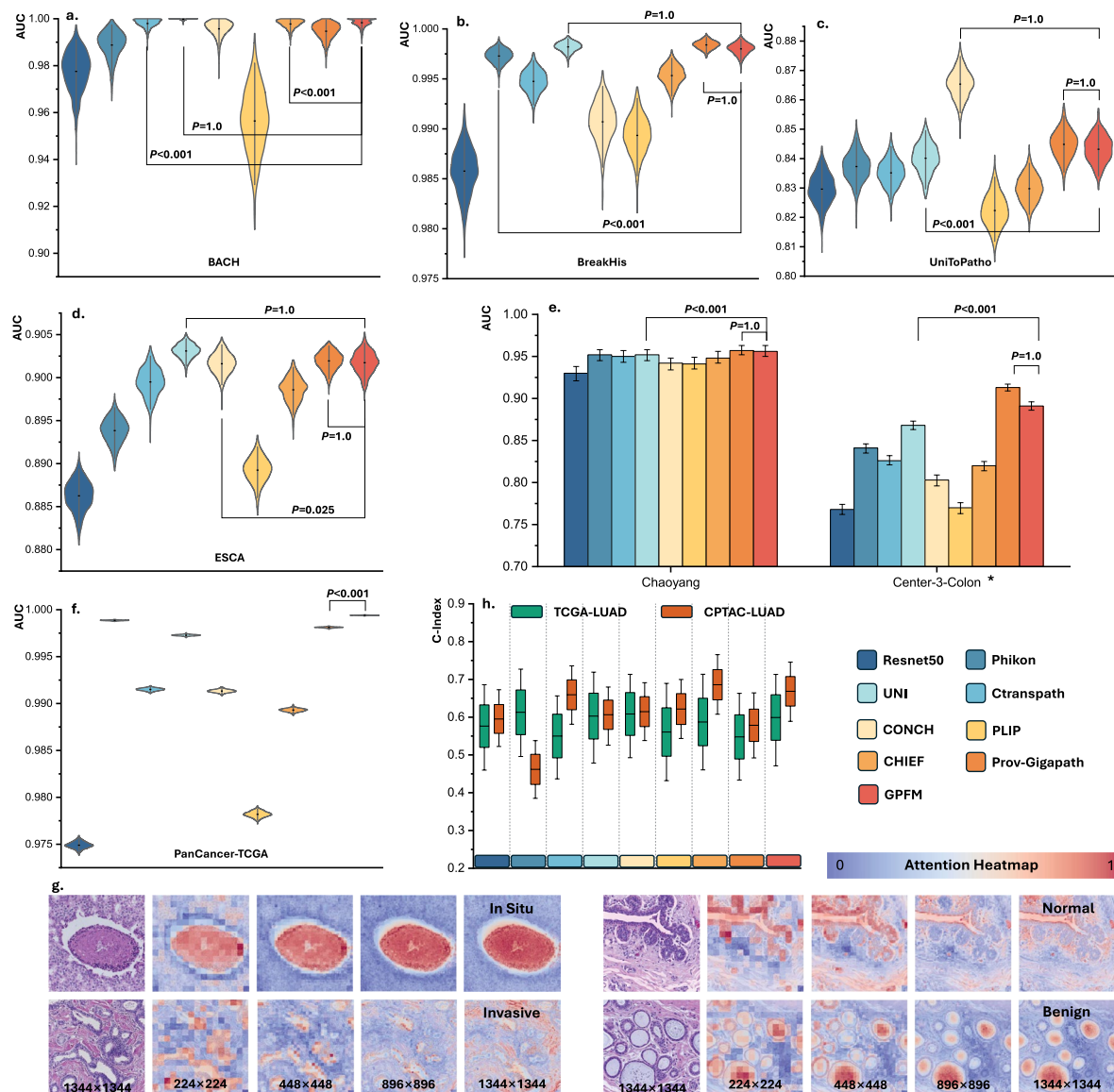
Extended Data Fig. 1 | Extended Results of WSI Classification. **a.** Performance comparison of foundation models in ILC and IDC classification. **b.** NSCLC subtyping performance across models. **c–e.** Model performance in prostate cancer grading, breast cancer metastasis detection, and LUAD TP53 mutation prediction, respectively. **f–i.** Extended evaluation including lung cancer metastasis detection, gastric cancer Lauren subtyping, lung cancer primary site

prediction, and gastric cancer perineural invasion detection. Violin plots show the distribution of 1,000 bootstrap replicates. Centre indicates mean. Error bars represent 95% CI. External validation cohorts are marked with *. 1,000 bootstrap replicates are performed for all bar plots. The Wilcoxon signed-rank two-side test is used for data analysis.



Extended Data Fig. 2 | Extended Results of WSI Classification. **a.** IDH-1 mutation prediction in brain tumors. **b.** Lesion grading in colon cancer. **c.** Brain tumor subtyping performance. **d.** Dual-task evaluation: primary site prediction and T-stage classification in head & neck cancer. Centre indicates mean. Error bars

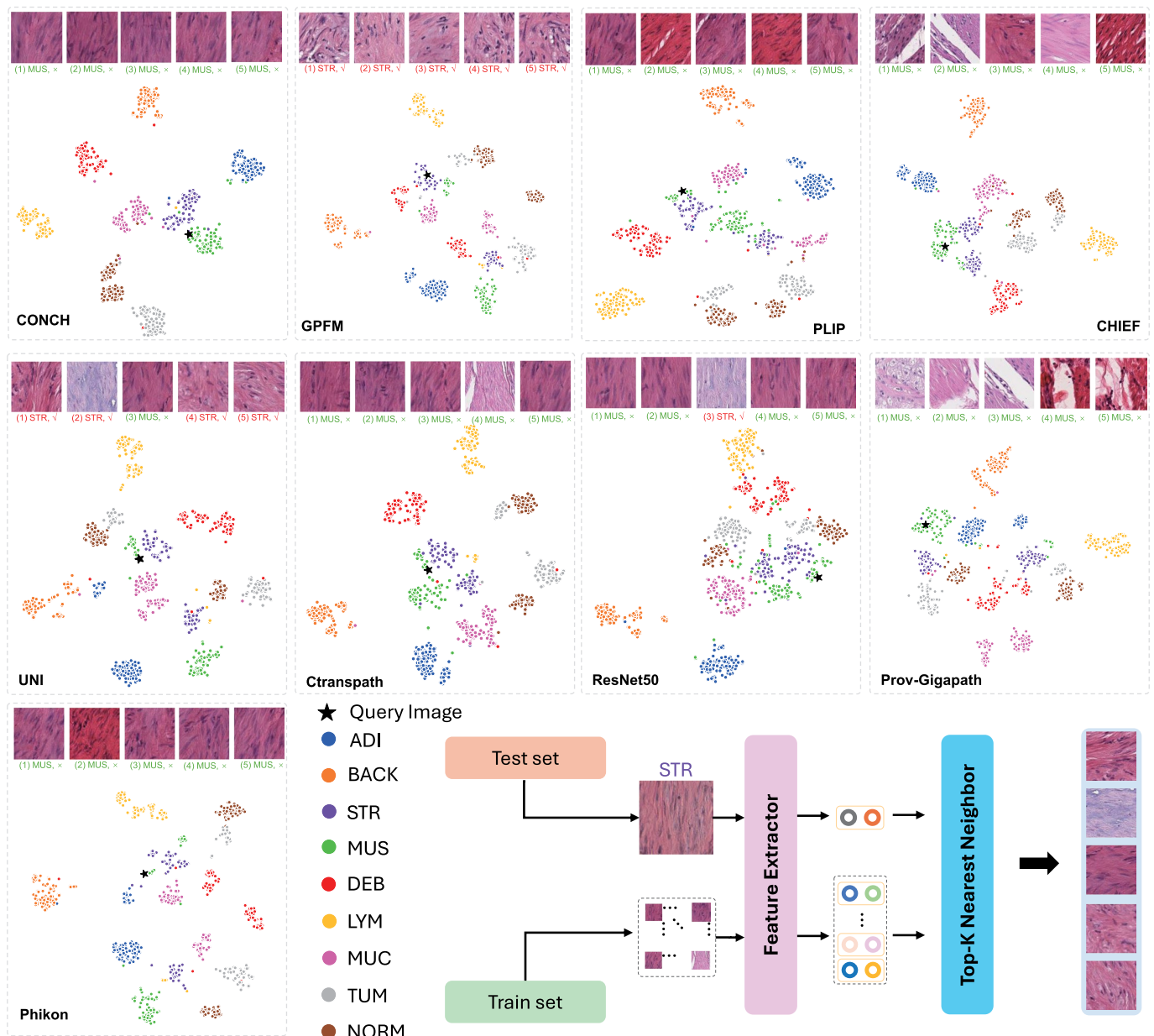
represent 95% CI. External validation cohorts are marked with *. For all bar plots, 1,000 bootstrap replicates are performed. The Wilcoxon signed-rank two-side test is used for data analysis.



Extended Data Fig. 3 | Extended Result of ROI Classification Tasks.

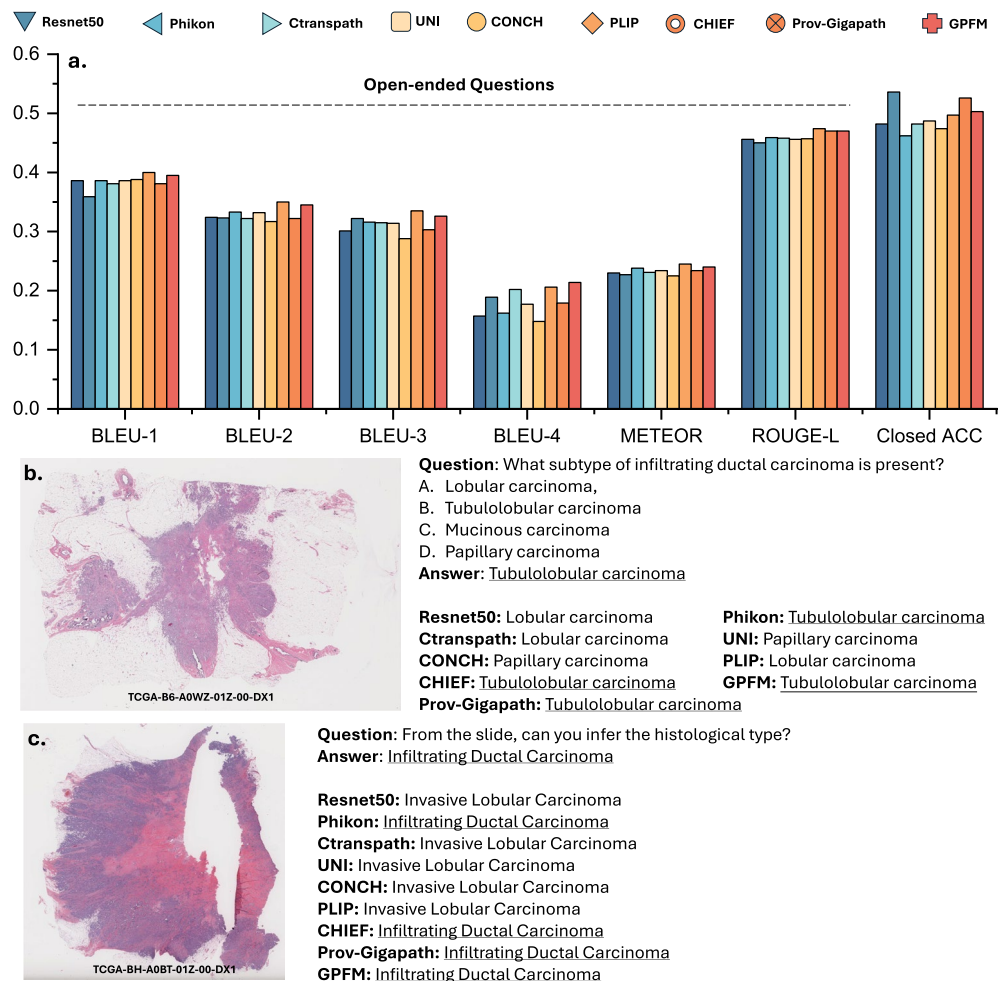
a-d. The AUC of foundation models on BACH, BreakHis, UniToPatho, and ESCA, respectively. **e.** The colon tissue classification performance. The Chaoyang and Center-3-Colon serve as internal and external, respectively. **f.** The performance of pancancer classification of different foundation models. For all subfigures **a-f**, the error bar indicates the 95% CI and the centre represents mean. **g.** Attention heatmap of GPFM across various image resolutions for BRCA subtyping in BACH dataset. The colored squares represent the 14×14 [PATCH] tokens encoded by the GPFM model. The heatmap values indicate the similarity between each [PATCH] token and the [CLS] token generated by the last layer of GPFM, measured

using Euclidean distance. The consistent attention patterns observed across varying image resolutions and tissue types underscore the robust capabilities of the GPFM model. **h.** Results on TCGA-LUAD data and the CPTAC-LUAD cohort. The survival prediction model was trained on the TCGA-LUAD cohort and subsequently tested on the CPTAC-LUAD cohort. The minima and maxima represent the lower and upper bound of the 95% CI, respectively. The centre and the bounds of box represent the mean and the standard error, respectively. For subfigures **a-f** and **h**, the 1,000 bootstrap replicates are performed. The Wilcoxon signed-rank two-side test is used for data analysis.

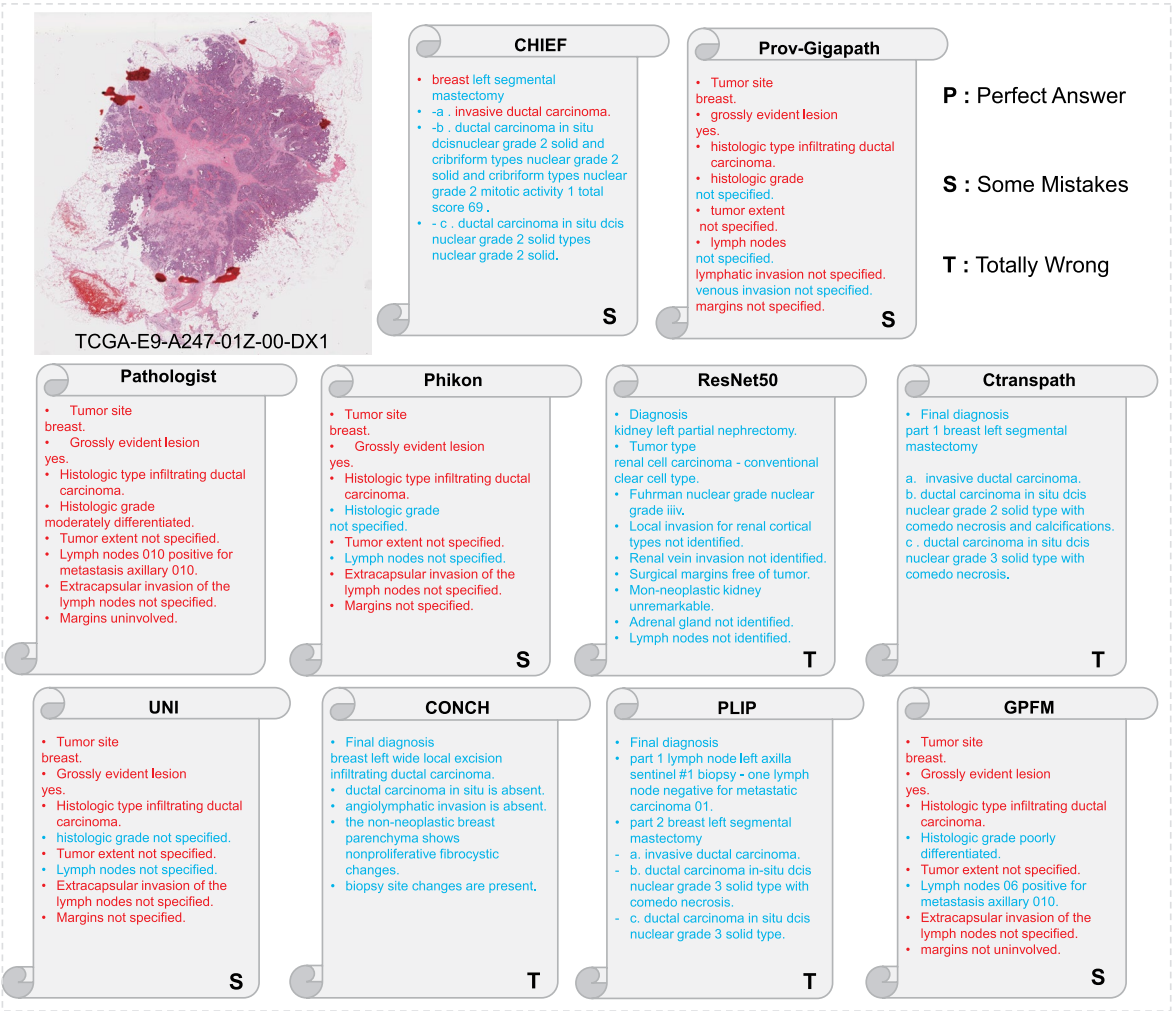


Extended Data Fig. 4 | Overview of Pathology ROI Tissue Retrieval. The central figure illustrates the framework for pathology tissue ROI retrieval. The surrounding figures visualize the distribution of features extracted by different models using t-SNE dimensionality reduction to 2D. For each class, 100 samples from the test set

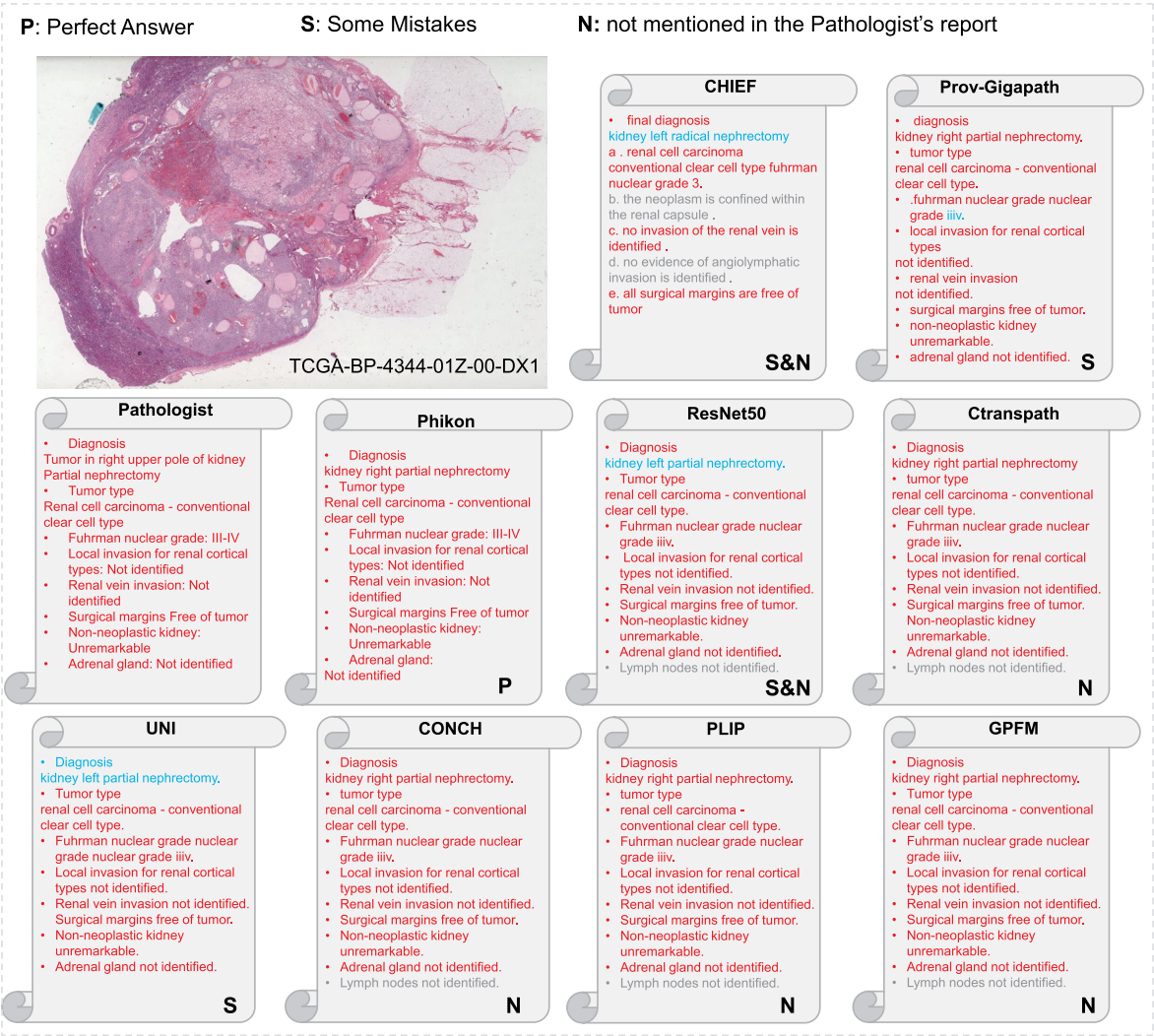
were used, and together with the query image, a total of 901 samples were subjected to the t-SNE analysis. The different classes are distinctly colored in the 2D t-SNE plot. The retrieved top-5 images for the query are also shown, demonstrating the GPFM's performance on this pathology tissue retrieval task.



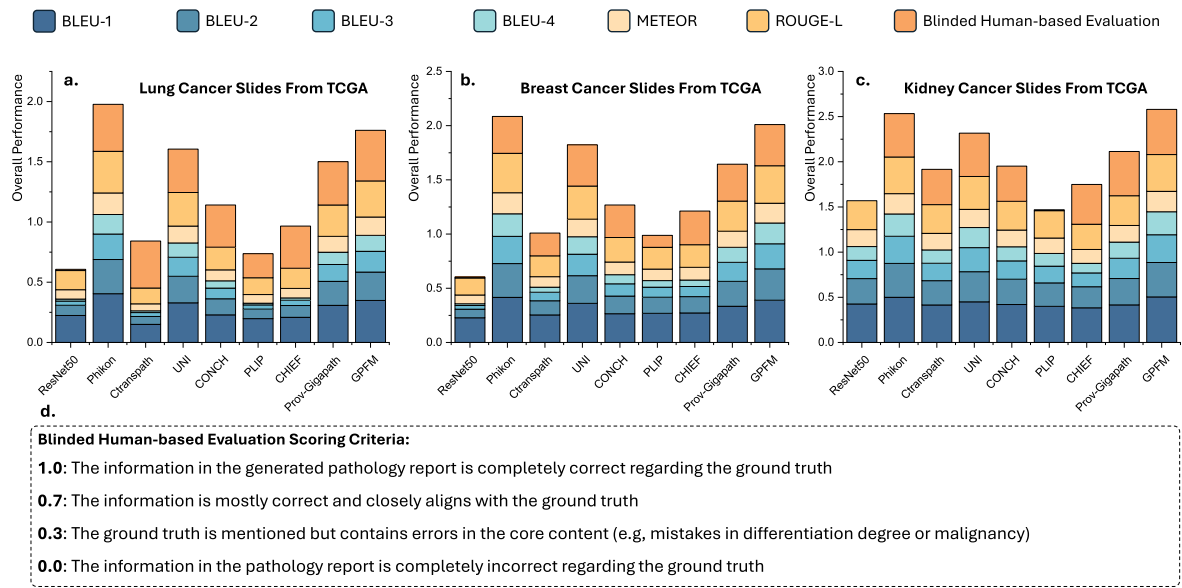
Extended Data Fig. 5 | VQA results on WSI-VQA dataset. a. Open-ended and close-ended statistical results. **b.** A close-ended question and corresponding answers. **c.** An open-ended question and corresponding answers.



Extended Data Fig. 6 | Generated Example Reports. The ground truth report is provided by pathologist. The text in red indicates correct predictions, the text in blue represents incorrect predictions.

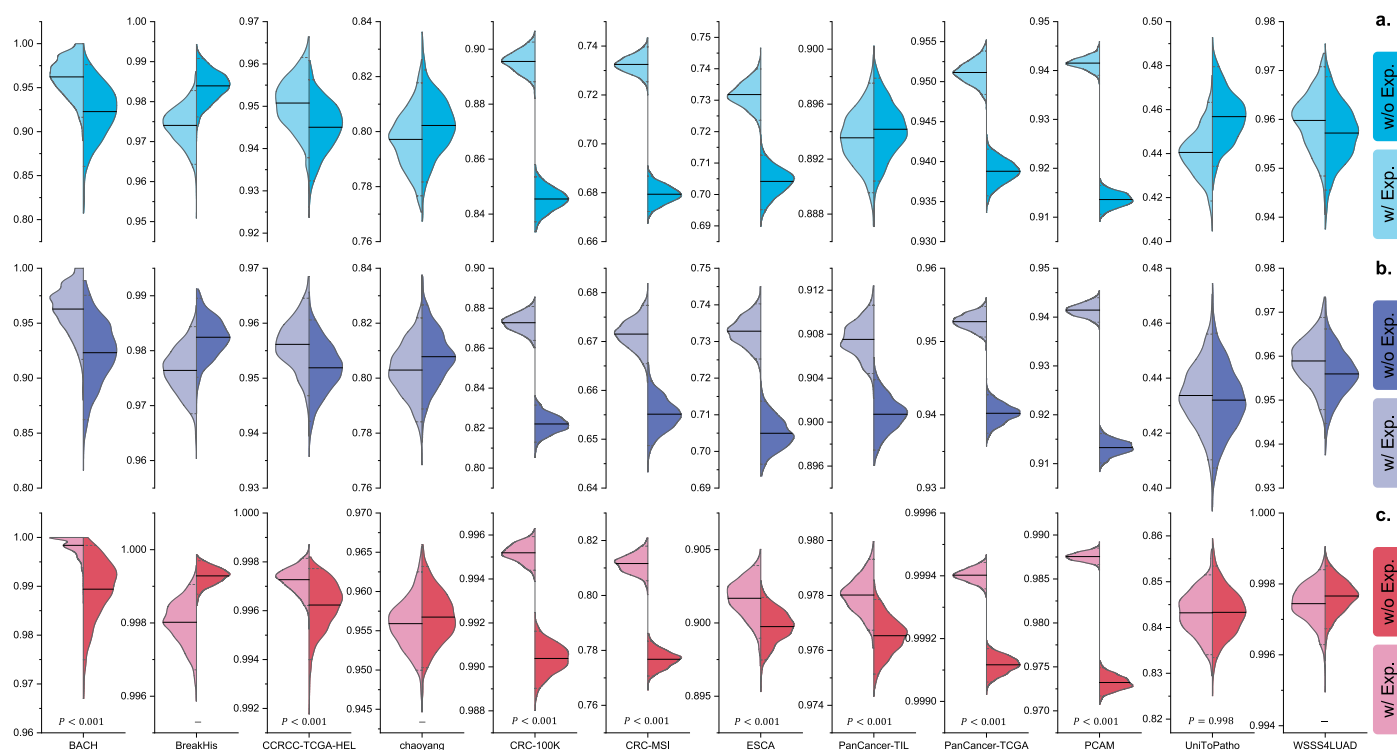


Extended Data Fig. 7 | Generated Example Reports. The ground truth report is provided by pathologist. The text in red indicates correct predictions, the text in blue represents incorrect predictions, and the text in gray is the predicted text not mentioned in the pathologist's report.



Extended Data Fig. 8 | Evaluation of Report Quality Based on Organ-Specific Analysis. **a-c.** Performance assessment of generated pathology reports for lung cancer, breast cancer, and kidney cancer, respectively. **d.** Scoring criteria for human-based blind evaluation of foundation-model-generated pathology

reports. The scoring system ranges from 0.0 to 1.0, where 1.0 indicates complete accuracy with ground truth, 0.7 represents mostly correct information, 0.3 indicates presence of core content errors, and 0.0 denotes completely incorrect information.



Extended Data Fig. 9 | The Effectiveness of Expert Knowledge Distillation.

The figure presents the performance difference between GPFM (with Expert Knowledge Distillation, that is, w/ Exp. in figure) and DINOv2 (without Expert Knowledge Distillation, that is, w/o Exp. in figure). The horizontal black lines indicate the mean AUC. If GPFM outperforms DINOv2, the p-value is also reported. **a.** The balanced accuracy of the models with and without Expert Knowledge Distillation. **b.** The weighted F1 score of the models with and without

Expert Knowledge Distillation. **c.** The AUC of the models with and without Expert Knowledge Distillation. The centre represent mean and the dashed lines indicate the 2.5-th and 97.5-th percentile, respectively. Significance testing was conducted using the Wilcoxon signed-rank one-sided test, demonstrating that Expert Knowledge Distillation consistently improves performance across the majority of tasks, highlighting the effectiveness of this technique in enhancing the GPFM.

Extended Data Table 1 | pseudocode of the Expert Knowledge Distillation module

Algorithm 1 The PyTorch-like pseudocode of the Expert Knowledge Distillation module.

Require: T_a , T_b , and T_c # off-the-shelf foundation models, we used phikon, uni, and conch in this study.

Require: S # student model

Require: v # global views

```

1:  $sc, sp = S(v)$  # [CLS] token and [patch] token encoded by student
2:  $ac, ap = T_a(v)$  # [CLS] token and [patch] token encoded by  $T_a$ 
3:  $bc, bp = T_b(v)$ 
4:  $cc, cp = T_c(v)$ 
5:  $d_{ac} = 1 - \cos(sc, ac)$ 
6:  $d_{bc} = 1 - \cos(sc, bc)$ 
7:  $d_{cc} = 1 - \cos(sc, cc)$ 
8:  $d_c = \alpha d_{ac} + \beta d_{bc} + \gamma d_{cc}$ 
9:  $d_{ap} = \eta * (1 - \cos(sp, ap)) + \theta * \text{SmoothL1}(sp, ap)$ 
10:  $d_{bp} = \eta * (1 - \cos(sp, bp)) + \theta * \text{SmoothL1}(sp, bp)$ 
11:  $d_{cp} = \eta * (1 - \cos(sp, cp)) + \theta * \text{SmoothL1}(sp, cp)$ 
12:  $d_p = \mu d_{ap} + \lambda d_{bp} + \phi d_{cp}$ 
13:  $d = d_c + d_p$ 

```

The PyTorch-like pseudocode of the Expert Knowledge Distillation module.

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our [Editorial Policies](#) and the [Editorial Policy Checklist](#).

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
<input type="checkbox"/>	<input checked="" type="checkbox"/> The exact sample size (<i>n</i>) for each experimental group/condition, given as a discrete number and unit of measurement
<input type="checkbox"/>	<input checked="" type="checkbox"/> A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
<input type="checkbox"/>	<input checked="" type="checkbox"/> The statistical test(s) used AND whether they are one- or two-sided <i>Only common tests should be described solely by name; describe more complex techniques in the Methods section.</i>
<input checked="" type="checkbox"/>	<input type="checkbox"/> A description of all covariates tested
<input checked="" type="checkbox"/>	<input type="checkbox"/> A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
<input type="checkbox"/>	<input checked="" type="checkbox"/> A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
<input type="checkbox"/>	<input checked="" type="checkbox"/> For null hypothesis testing, the test statistic (e.g. <i>F</i> , <i>t</i> , <i>r</i>) with confidence intervals, effect sizes, degrees of freedom and <i>P</i> value noted <i>Give P values as exact values whenever suitable.</i>
<input checked="" type="checkbox"/>	<input type="checkbox"/> For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
<input checked="" type="checkbox"/>	<input type="checkbox"/> For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
<input checked="" type="checkbox"/>	<input type="checkbox"/> Estimates of effect sizes (e.g. Cohen's <i>d</i> , Pearson's <i>r</i>), indicating how they were calculated

Our web collection on [statistics for biologists](#) contains articles on many of the points above.

Software and code

Policy information about [availability of computer code](#)

Data collection	To collect dataset for pretraining and downstream tasks, we use Python (version 3.10.10) along with following packages: openslide-python (version 1.2.0), pillow (version 9.4.0), opencv-python (version 4.7.0.72), and CLAM code base (http://github.com/mahmoodlab/CLAM).
Data analysis	In this project, we use Python and related open-sourced libraries for all experiments which can be repeated easily. For foundation model pretraining, we utilized PyTorch (version 2.1.2 with CUDA 12.1) and pretrained it using the FullyShardedDataParallel (FSDP) technique on 2x8 80GB NVIDIA H800 GPU nodes. All other data processing and evaluation for downstream tasks were carried out on a server equipped with 8x NVIDIA RTX 3090 GPUs. To assess the model's performance, we employed Torchmetrics (version 1.3.2) and Scikit-learn (version 1.2.2) for metric evaluation. Other pretrained encoders could be found at the following links: ResNet-50 (http://download.pytorch.org/models/resnet50-19c8e357.pth), Phikon (https://huggingface.co/owkin/phikon), Ctranspath (https://github.com/Xiyue-Wang/TransPath), UNI (https://huggingface.co/MahmoodLab/UNI), CONCH (https://huggingface.co/MahmoodLab/CONCH), PLIP (https://github.com/PathologyFoundation/plip). For WSI classification and survival analysis, we relied on openslide-python (version 1.2.0) and the CLAM (http://github.com/mahmoodlab/CLAM) codebase. Pathology VOA evaluation was conducted using the MUMC (https://github.com/pengfeiliHEU/MUMC) code base. For ROI classification, we released code at Github (https://github.com/birkhoffkiki/GPFM). For histology report generation, we utilized the HistGen (https://github.com/ddavid4real/HistGen) code base. matplotlib (version 3.7.1), seaborn (version 0.13.0) and Origin 2021 were used to plot figures.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio [guidelines for submitting code & software](#) for further information.

Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our [policy](#)

This study incorporates a total of 56 datasets (sources). Out of these, 33 datasets are utilized for pretraining, and a subset of them is also employed for evaluation purposes. The remaining 23 datasets are specifically dedicated to downstream task evaluation. The splits of the dataset can be found in our GitHub repository (<https://github.com/birkhoffkiki/GPFM>). For the data from Center-1 to Center-5, these datasets are not publicly available due to patient privacy obligations. However, researchers interested in accessing de-identified data may submit a reasonable request directly to the corresponding authors.

The links of used public data are listed at following lines:

1. TCGA (<https://portal.gdc.cancer.gov>)
2. CPTAC (<https://proteomic.datacommons.cancer.gov/pdc>)
3. PANDA (<https://www.kaggle.com/c/prostate-cancer-grade-assessment/data>)
4. NADT-Prostate (<https://www.cancerimagingarchive.net/collection/nadt-prostate/>)
5. BCNB (<https://bcnb.grand-challenge.org>)
6. CAMELYON16 (<https://camelyon16.grand-challenge.org/Data/>)
7. CAMELYON17 (<https://camelyon17.grand-challenge.org/Data/>)
8. BRACS (<https://www.bracs.icar.cnr.it/download/>)
9. TIGER2021 (<https://tiger.grand-challenge.org/>)
10. MIDOG2022 (<https://midog.deepmicroscopy.org/download-dataset/>)
11. AGGC2022 (<https://aggc22.grand-challenge.org/>)
12. O.B.R. (<https://www.cancerimagingarchive.net/collection/ovarian-bevacizumab-response/>)
13. ACROBAT2023 (<https://acrobot.grand-challenge.org/>)
14. AML-C-LMU (https://www.cancerimagingarchive.net/collection/aml-cytomorphology_lmu/)
15. ARCH (https://warwick.ac.uk/fac/cross_fac/tia/data/arch)
16. BACH (<https://zenodo.org/records/3632035>)
17. CAMEL (<https://drive.google.com/open?id=1brr8CnU6ddzAYT157wkdXjbSzoilDF9y>)
18. DiagSet (<https://ai-econsilio.diag.pl/>)
19. DLBCL (<https://github.com/stanfordmlgroup/DLBCL-Morph>)
20. GTEx (<https://gtexportal.org/home/histologyPage>)
21. HunCRC (<https://www.cancerimagingarchive.net/collection/hungarian-colorectal-screening/>)
22. Janowczyk (<https://andrewjanowczyk.com/use-case-1-nuclei-segmentation/>)
23. LC25000 (<https://academictorrents.com/details/7a638ed187a6180fd6e464b3666a6ea0499af4af>)
24. MIDOG2021 (<https://imig.science/midog2021/download-dataset/>)
25. OCELOT (<https://zenodo.org/record/7844149>)
26. Oste. Tumor (<https://www.cancerimagingarchive.net/collection/osteosarcoma-tumor-assessment/>)
27. PAIP2019 (<https://paip2019.grand-challenge.org/>)
28. PAIP2020 (<https://paip2020.grand-challenge.org/>)
29. PAIP2021 (<https://paip2021.grand-challenge.org/>)
30. Post-NAT-BRCA (<https://www.cancerimagingarchive.net/collection/post-nat-brca/>)
31. SICAPv2 (<https://data.mendeley.com/datasets/9xxm58dvs3/1>)
32. SLN-Breast (<https://www.cancerimagingarchive.net/collection/sln-breast/>)
33. SPIE2019 (<https://breastpathq.grand-challenge.org/>)
34. PatchGastricADC22 (<https://zenodo.org/records/6550925>)
35. UBC-OCEAN (<https://www.kaggle.com/competitions/UBC-OCEAN/data>)
36. WSI-VQA (<https://github.com/cpystan/WSI-VOA>)
37. CRC-100K (<https://zenodo.org/records/1214456>)
38. CRC-MSI (<https://zenodo.org/records/3832231>)
39. CCRCC-TCGA-HEL (<https://zenodo.org/records/7898308>)
40. PanCancer-TCGA (<https://zenodo.org/records/5889558>)
41. PanCancer-TIL (<https://zenodo.org/records/6604094>)
42. ESCA (<https://zenodo.org/records/7548828>)
43. PCAM (<https://github.com/basveeling/pcam>)
44. BreakHis (<https://www.kaggle.com/datasets/ambarish/breakhis>)
45. UniToPatho (<https://ieee-dataport.org/open-access/unitopatho>)
46. Chaoyang (<https://github.com/bupt-ai-cz/HSA-NRL>)
47. PathVQA (<https://huggingface.co/datasets/flaviagiammarino/path-vqa>)
48. HistGen (<https://github.com/ddavid4real/HistGen>)
49. IMP-CRS (<https://rdm.inesctec.pt/dataset/nis-2023-008>)
50. HANCOCK (https://github.com/ankilab/HANCOCK_MultimodalDataset)
51. GasHisDB (<https://figshare.com/articles/dataset/GasHisSDB/15066147>)

Research involving human participants, their data, or biological material

Policy information about studies with [human participants or human data](#). See also policy information about [sex, gender \(identity/presentation\), and sexual orientation](#) and [race, ethnicity and racism](#).

Reporting on sex and gender

We did not use gender or sex as a covariate in our study. The data used in this data are all public, we refer readers to the original source for the detailed description.

Reporting on race, ethnicity, or other socially relevant groupings

We did not collect or use any covariates regarding race, ethnicity, and other social groupings at any stage of the study.

Population characteristics

We did not collect or use any covariates pertaining to population characteristics at any stage of the study.

Recruitment

No patient recruitment was necessary for this study.

Ethics oversight

This project has been reviewed and approved by the Human and Artefacts Research Ethics Committee (HAREC) of Hong Kong University of Science and Technology. The protocol number is HREP-2024-0212.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences

☐ Behavioural & social sciences

☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/documents/nr-reporting-summary-flat.pdf](https://www.nature.com/documents/nr-reporting-summary-flat.pdf)

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size

No formal sample size calculation was performed due to the exploratory nature of this foundational model pretraining study. Instead, the sample size was determined by (1) the availability of public whole slide images (WSIs) and (2) computational feasibility within our fixed budget. For pretraining, we utilized 72,280 WSIs, from which 190,212,668 patches (512×512 pixels) were extracted. This large-scale dataset was selected to ensure broad representation of histopathological features and enhance the generalizability of the foundation model. The remaining 23,292 WSIs and additional patch-level datasets were reserved for downstream tasks to enable robust validation and fine-tuning.

Data exclusions

For pretraining data, no exclusions were performed after curation.

Replication

Attempts at replication were successful for the reported model results. Code corresponding this work can be accessed at <https://github.com/birkhoffkiki/GPFM>

Randomization

For downstream tasks that required creating train, validation, test splits, we either used official splits created by the original investigators of each dataset when available, or created them randomly. The train-validation-test splits could be accessed at <https://github.com/birkhoffkiki/GPFM>

Blinding

Blinding was not necessary for our study because our experiments were based on digitized histology slides or region-level images. Reporting

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> Antibodies
<input checked="" type="checkbox"/>	<input type="checkbox"/> Eukaryotic cell lines
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology and archaeology
<input checked="" type="checkbox"/>	<input type="checkbox"/> Animals and other organisms
<input checked="" type="checkbox"/>	<input type="checkbox"/> Clinical data
<input checked="" type="checkbox"/>	<input type="checkbox"/> Dual use research of concern
<input checked="" type="checkbox"/>	<input type="checkbox"/> Plants

Methods

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input checked="" type="checkbox"/>	<input type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging

Plants

Seed stocks

Report on the source of all seed stocks or other plant material used. If applicable, state the seed stock centre and catalogue number. If plant specimens were collected from the field, describe the collection location, date and sampling procedures.

Novel plant genotypes

Describe the methods by which all novel plant genotypes were produced. This includes those generated by transgenic approaches, gene editing, chemical/radiation-based mutagenesis and hybridization. For transgenic lines, describe the transformation method, the number of independent lines analyzed and the generation upon which experiments were performed. For gene-edited lines, describe the editor used, the endogenous sequence targeted for editing, the targeting guide RNA sequence (if applicable) and how the editor was applied.

Authentication

Describe any authentication procedures for each seed stock used or novel genotype generated. Describe any experiments used to assess the effect of a mutation and, where applicable, how potential secondary effects (e.g. second site T-DNA insertions, mosaicism, off-target gene editing) were examined.