



PathOrchestra: a comprehensive foundation model for computational pathology with over 100 diverse clinical-grade tasks



Fang Yan^{1,14}✉, Jianfeng Wu^{2,14}, Jiawen Li^{3,14}, Wei Wang^{4,5,14}, Yirong Chen¹, Linda Wei⁶, Jiaxuan Lu¹, Wen Chen⁷, Zizhao Gao², Jianan Li², Heng Li^{4,5}, Hong Yan^{4,5}, Jiabo Ma⁸, Minda Chen⁷, Yang Lu⁷, Qing Chen⁷, Yizhi Wang³, Xitong Ling³, Xuenian Wang³, Zihan Wang^{9,10}, Qiang Huang^{9,10}, Shengyi Hua¹¹, Mianxin Liu¹, Lei Ma¹², Tian Shen⁷, Xiaofan Zhang^{1,11}✉, Yonghong He³✉, Hao Chen⁸✉, Shaoting Zhang^{1,11,13}✉ & Zhe Wang^{2,4,5}✉

The complexity and variability of high-resolution pathological images present significant challenges in computational pathology. While AI-driven pathology foundation models have advanced the field, they require large-scale datasets, substantial storage, and significant computational resources, as well as rigorous validation for clinical applicability. We present PathOrchestra, a versatile pathology foundation model trained on 287,424 slides from 21 tissue types across three centers. Evaluated on 112 tasks from 61 private and 51 public datasets, covering digital slide preprocessing, pan-cancer classification, lesion identification, multi-cancer subtype classification, biomarker assessment, gene expression prediction, and structured report generation. Across 27,755 whole slide images and 9,415,729 region-of-interest images, it achieved over 0.950 accuracy in 47 tasks, including pan-cancer classification, lymphoma subtyping, and bladder cancer screening. It is the first to generate structured reports for colorectal cancer and lymphoma. Overall, PathOrchestra demonstrates the clinical readiness of large-scale self-supervised pathology foundation models, achieving high accuracy and offering potential to digital medicine integration.

Pathology is heralded as the gold standard for disease diagnosis, encompassing a broad range of tasks including tumor detection^{1,2}, typing^{3,4}, grading^{5,6}, molecular expression analysis^{7–10}, prognosis assessment^{11,12}, and prediction of treatment response^{13–15}. These tasks require handling high-resolution and morphologically diverse pathological images, posing significant diagnostic resource challenges. Computational pathology (CPath) leverages powerful algorithms to interpret and quantify pathological images, providing a more precise understanding of disease mechanisms. However, traditional AI training paradigms relying on large annotated datasets face critical limitations, as accurately labeling vast pathological data across diverse disease types is often impractical, which underscores the urgent need for generalizable models that can efficiently handle multicenter data while reducing reliance on manual annotation.

The advent of self-supervised learning in computer vision (e.g., DINOv2¹⁶, MoCov3¹⁷, MAE¹⁸) and the success of large language models like the GPT series^{19,20} have propelled the development of data-driven pre-training methodologies. These models^{21–24} leverage vast collections of whole slide images (WSIs) to learn high-quality feature representations from unlabeled data, demonstrating state-of-the-art performance in downstream tasks using limited labeled data and fewer training parameters. Despite these advancements, a comprehensive evaluation of foundational pathology models is still lacking. Furthermore, the impact of pre-training data and model architecture on performance in downstream tasks remains under-explored. While technical assessments have been carried out, the clinical applicability and potential of these models remain insufficiently addressed, limiting their full integration into real-world clinical settings.

A full list of affiliations appears at the end of the paper. ✉e-mail: yanfang@pjlab.org.cn; zhangxiaofan@pjlab.org.cn; heyh@sz.tsinghua.edu.cn; jhc@cse.ust.hk; zhangshaoting@pjlab.org.cn; zhwang@fmmu.edu.cn

In this paper, we present PathOrchestra (Fig. 1), a pathology foundation model that sets new performance benchmarks through a rigorous evaluation across 112 clinically relevant downstream tasks. PathOrchestra is implemented as a self-supervised vision encoder, pretrained on an extensive corpus of 287,424 hematoxylin and eosin (H&E)-stained whole-slide images encompassing 21 distinct tissues, organs, and systems, sourced from both proprietary archives and publicly available datasets from three independent centers. We systematically examine its translational potential in diverse clinical scenarios (Fig. 2), demonstrating robust and consistent performance on a large-scale evaluation comprising 27,755 WSIs and 9,415,729 region-of-interest (ROI) images (Supplementary Tables 1 and 2). PathOrchestra achieved an accuracy exceeding 0.950 in 47 tasks, thereby establishing state-of-the-art benchmarks in real-world applications, including digital slide preprocessing, pan-cancer classification, lesion identification and detection, multi-cancer subtype classification, biomarker assessment, gene expression prediction, and structured pathology reporting.

Results

Dataset curation

The training dataset consists of 287,424 WSIs, covering 21 distinct types of tissues, organs, and systems, including Lymph Node (68,054), Stomach (59,598), Breast (32,890), Cervix (27,786), Lung (26,842), Bladder (16,326), Prostate (10,188), Kidney (9,328), Brain (7,240), Blood System (4,232), Colon (4,088), Liver (3,344), Uterus (3,086), Ovary (2,962), Skin (2,550), Head and Neck (2,526), Esophagus (2,492), Bone (1,890), Pancreas (932), Nerve (770), and Eye (300). These WSIs were digitized from frozen and formalin-fixed paraffin-embedded (FFPE) tissue slides stained with H&E. The dataset consists of 268,859 WSIs derived from FFPE samples and 18,565 from frozen samples, resulting in a FFPE-to-frozen sample ratio of ~14:1. The scanning process utilized equipment from four different manufacturers, including the SQS-600P scanner (Shenzhen Shengqiang Technology Co., Ltd., China), using a 20 × objective and generating images in .sdpc format; the KF-PRO-005 scanner (Ningbo Jiangfeng Bioinformatics Technology Co., Ltd., China), employing a 40 × objective and producing .kfb files; the Aperio ScanScope GT 450 scanner (Aperio, USA), also using a 40 × objective to generate .svs files; and the Panoramic MIDI II scanner (3DHISTECH, Hungary), utilizing a 40 × objective and producing .mdsx files. We sampled 256 × 256 patches at 20 × magnification, generating a total of 141,471,591 patches, all of which were used in the actual training.

The training dataset includes data from three sources, including the Xijing Hospital, Fourth Military Medical University, Xi'an, China (2023–2024) with 123,169 WSIs and 13,303 cases; The First Affiliated Hospital of USTC, Hefei, China (2022–2024) with 133,999 WSIs and 16,749 cases; and The Cancer Genome Atlas (TCGA), USA (2006–2016), which encompasses contributions from Memorial Sloan Kettering Cancer Center, Johns Hopkins University, University of California Los Angeles (UCLA), University of Michigan, Mayo Clinic, University of Washington, and University of Texas MD Anderson Cancer Center, totaling 30,256 WSIs and 10,990 cases. This study was approved by the Institutional Review Boards of Xijing Hospital (IRB No. KY20252335) and The First Affiliated Hospital of USTC (IRB No. 2025-RE-204) for the retrospective analysis of internal pathological images. Prior to computational analysis and model development, all internal data including WSIs and pathology reports were de-identified to ensure patient confidentiality. Since the study involved retrospective analysis of archival pathological slides without direct patient recruitment, informed consent was not required. The training data consisted solely of pathological images without any associated patient diagnoses or personal identifiers. Details were shown in Supplementary Table 4.

Pathology image preprocessing and quality control tasks

We classified the preprocessing tasks in CPath into two main categories which are pathology quality control tasks and general analysis tasks. The quality control tasks include wrinkle detection, bubble and adhesive identification, contamination detection, and blur detection. The general analysis tasks encompass pathology image identification, recognition of H&E and

IHC staining, identification of IHC marker types, magnification recognition, differentiation between frozen and FFPE sections, identification of biopsy versus large specimens, and recognition of IHC membrane and nuclear positivity. Preprocessing is a critical step in the CPath workflow, directly impacting the performance of subsequent algorithms. Developing a feature extractor with strong generalization capabilities can significantly enhance model performance while reducing reliance on extensive training data and annotations. All preprocessing tasks were conducted using supervised ROI classification based on the linear probing method, with model performance evaluated using ACC and F1 scores. Detailed information on the preprocessing tasks and experimental setups is provided in the Methods section.

Among the 12 preprocessing-related tasks (Fig. 3), the PathOrchestra model achieved ACC and F1 scores exceeding 0.950 in 7 subtasks. The general analysis tasks demonstrated superior performance compared to the quality control tasks, with ACC and F1 scores surpassing 0.970 for pathology image identification, recognition of H&E and IHC staining, recognition of H&E and fluorescence staining, magnification discrimination, and differentiation between frozen and FFPE sections. In quality control tasks such as bubble and adhesive identification, the model achieved ACC and F1 scores above 0.980. These results demonstrate the model's robust capability to effectively execute a wide range of preprocessing tasks without extensive retraining.

Pan-cancer classification tasks

We evaluated PathOrchestra on slide-level pan-cancer classification tasks, including a 17-class pan-cancer tissue classification using an in-house FFPE dataset, a 32-class classification using data from the TCGA FFPE dataset, and a 32-class classification using data from the TCGA frozen tissue dataset. These tasks assessed the model's performance across different cancer types, data sources, and specimen preparation methods. Pan-cancer classification models are designed to recognize and distinguish multiple cancer types, necessitating strong generalization capabilities crucial for the early detection of various cancers. All tasks were performed using weakly supervised WSI classification based on the attention-based ABMIL, with model performance evaluated using AUC, ACC and F1 scores. Detailed information on these tasks and experimental setups is provided in the Methods section.

In the 17-class pan-cancer tissue classification task (Fig. 4a), the model achieved an average value for AUC of 0.988, ACC of 0.879, and F1 score of 0.863. Notably, the Prostate classification task achieved perfect scores (ACC, AUC, and F1 = 1.0). Upon inspection, all prostate samples were needle biopsies, which have consistent and visually distinctive features compared to other organ types that are mostly large surgical specimens. We speculate that this uniform sampling modality contributes to the model's strong performance in this task. For the 32-class classification using the TCGA FFPE dataset, the model attained the AUC of 0.964, ACC of 0.666, and F1 score of 0.667 (Fig. 4b). We observed that the classification performance for frozen sections was 1.4% lower in AUC, 8.9% lower in ACC, and 9% lower in F1 compared to FFPE sections under similar conditions (Fig. 4c). This discrepancy may be attributed to the better preservation of tissue structure and morphology in FFPE sections, which facilitates feature extraction and classification. These results underscore the model's robust generalization capabilities in pan-cancer classification tasks, highlighting its potential as a foundational pathology model for screening multiple cancer types and improving diagnostic accuracy.

Multi-organ lesion identification and analysis tasks

Lesion identification and detection encompass 15 tasks, including WSI-level breast cancer metastasis detection (CAMELYON16, CAMELYON17), ROI-level lymphoma tumor region segmentation (PD-L1), multi-organ tumor-infiltrating lymphocyte (TIL) positive/negative screening (TCGA-TIL), breast tumor classification (PCam), colorectal malignant lesion classification (DigestPath), multi-organ cell semantic segmentation (SegPath), colorectal gland segmentation (GlaS), multi-organ nuclear instance segmentation (PanNuke), colorectal cell instance segmentation (CoNSEP),

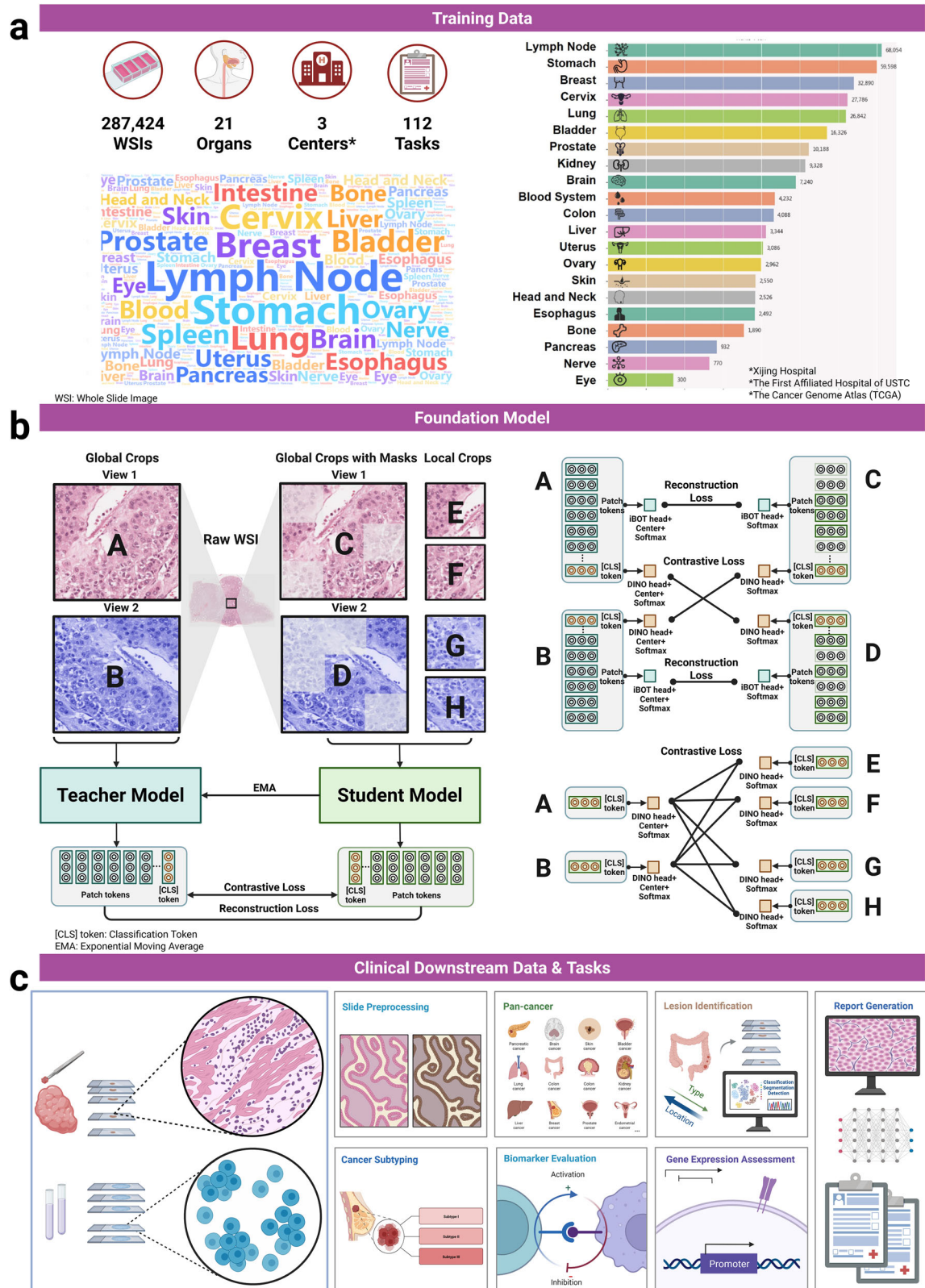


Fig. 1 | Overview of PathOrchestra. **a** The foundational model was trained on 287,424 WSIs sourced from three centers, covering 21 human organs, tissues and systems. **b** The training strategy followed an unsupervised contrastive learning framework, utilizing a teacher model as the foundation for testing downstream tasks.

c Diverse downstream task data encompasses both histology and cytology, enabling the model to support 112 clinical tasks across seven major categories. Figure c was created using the BioRender platform (biorender.com), in accordance with BioRender's academic licensing terms.

Fig. 2 | Overview of downstream tasks and dataset statistics. It covers a total of 112 tasks, including 61 WSI tasks and 51 ROI tasks, with a vast dataset comprising 27,755 WSIs and 9,415,729 ROIs. The datasets used span both public and private sources, with 51 public datasets and 61 private datasets, incorporating over 35 types of stains and analyzing >50 genes. The downstream tasks are categorized into seven groups: Slide Preprocessing (12 tasks), Pan-Cancer (3 tasks), Lesion Identification (15 tasks), Cancer Subtyping (36 tasks), Biomarker Evaluation (36 tasks), and Gene Expression Assessment (10 tasks). The methodology employed for these tasks includes detection, retrieval, segmentation, linear classification, and weakly supervised classification approaches, utilizing models like ViTDet³⁵, KNN⁶¹, Mask2Former²⁵, Linear Probing¹⁸, and ABMIL³³.

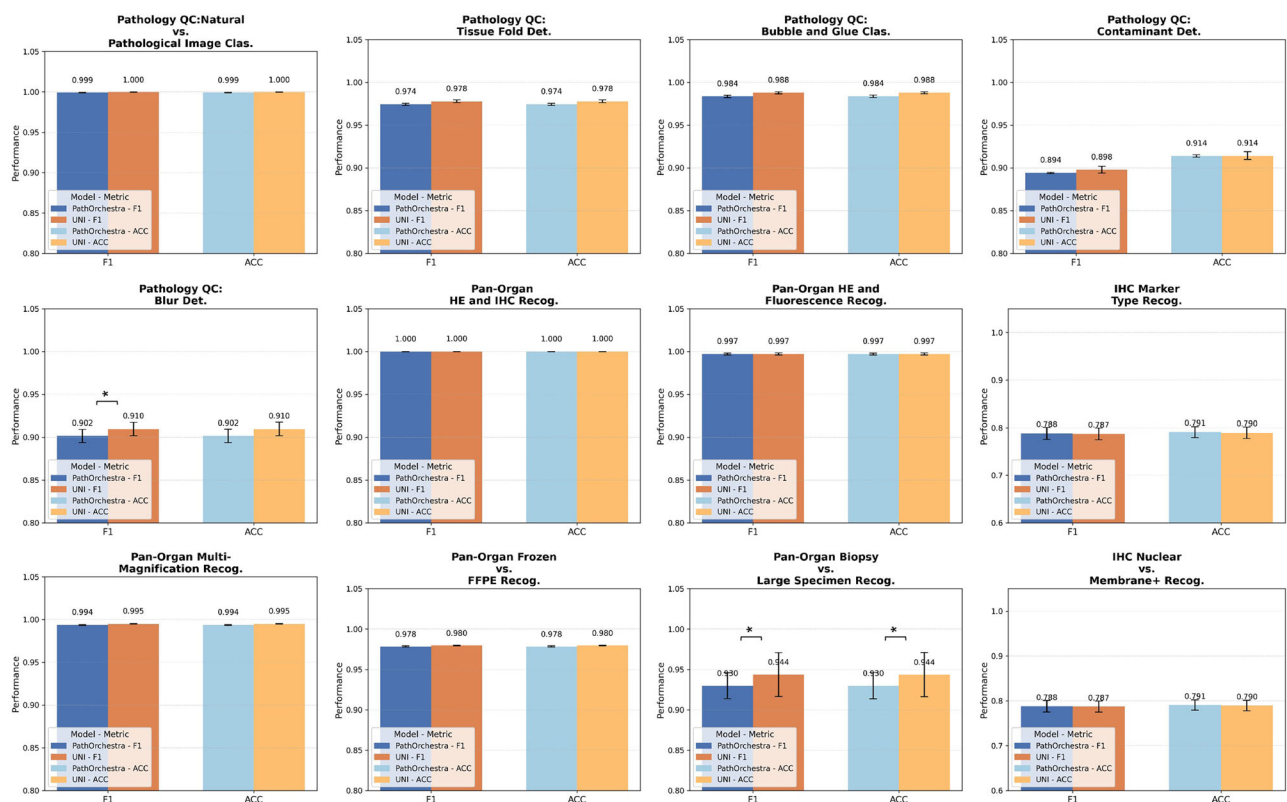
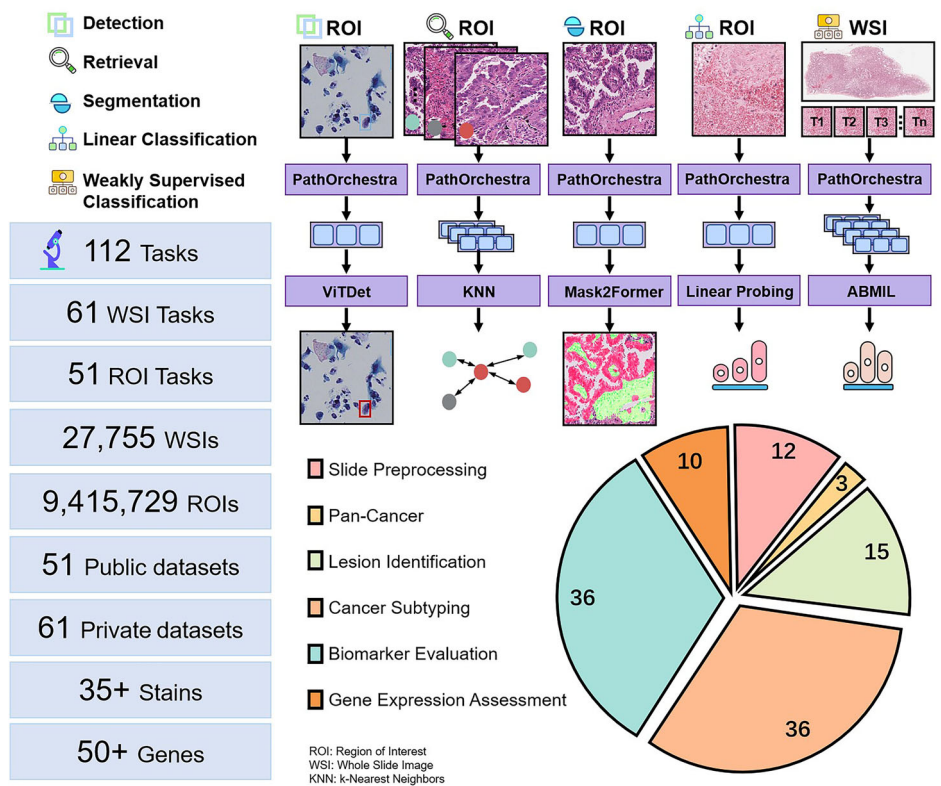


Fig. 3 | Performance on pathology image preprocessing and quality control tasks. Model performance is shown for 12 region-of-interest (ROI)-level classification sub-tasks, encompassing pathology image quality assessment, magnification

recognition, staining type identification, specimen collection method classification, and slide preparation form identification.

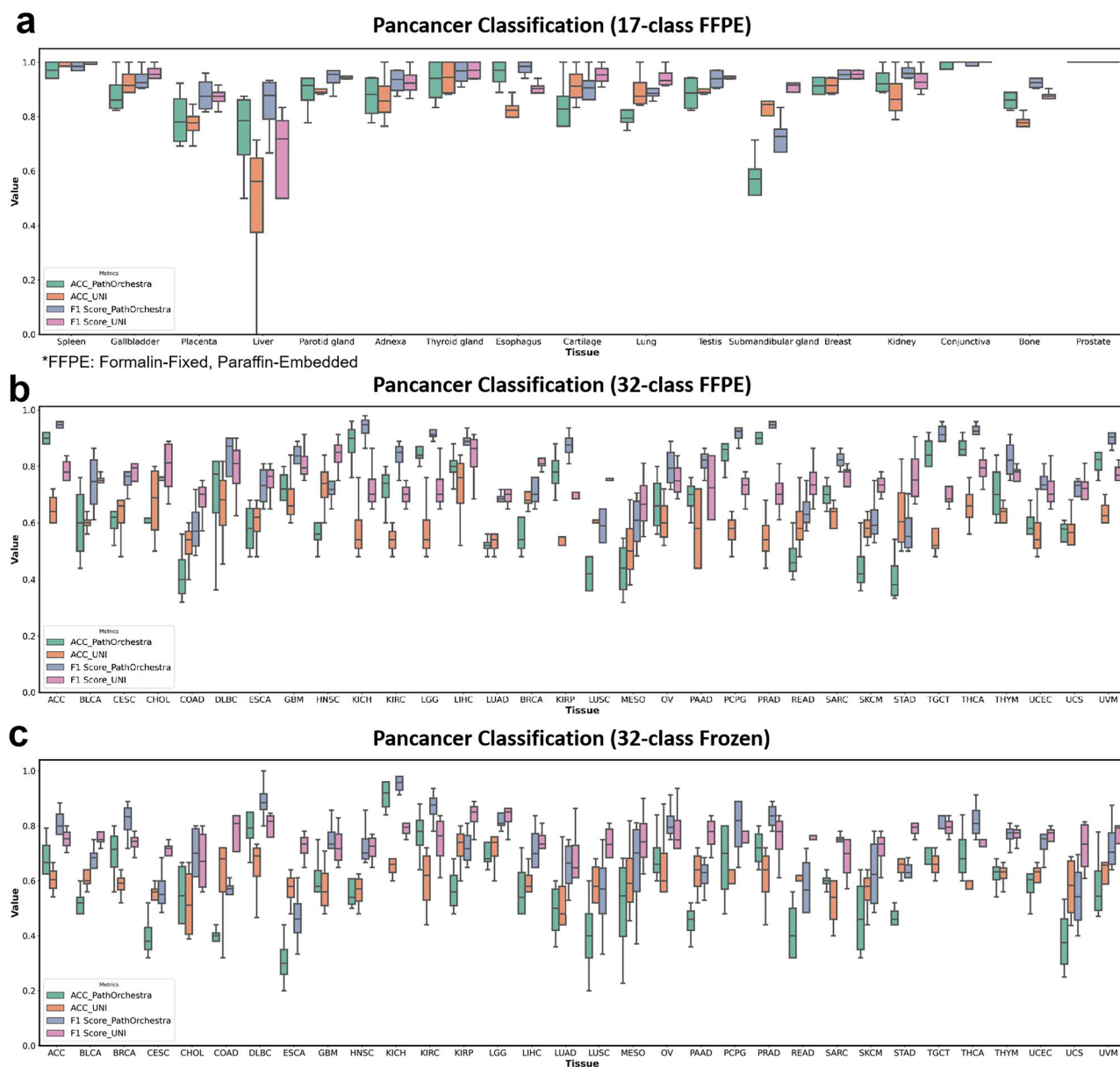


Fig. 4 | Results of pan-cancer classification tasks. a Model performance in weakly supervised whole-slide image (WSI)-level classification tasks for 17 cancer types using in-house center data. The bar chart presents classification accuracy for each cancer type. **b, c** Evaluation of model performance in pan-cancer prediction for 32

high-incidence cancer types from the TCGA database across two slide formats: formalin-fixed paraffin-embedded (FFPE) and frozen sections. The bar charts display prediction accuracy for each format. Abbreviations: WSI, whole-slide image; FFPE, formalin-fixed paraffin-embedded; TCGA, The Cancer Genome Atlas.

lung cancer tissue segmentation (WSSS4LUAD), cross-organ and cross-scanner adenocarcinoma segmentation (COSAS-Seg) and tissue classification (COSAS-Clas), multi-organ mitosis detection (SegPath), cervical cancer TCT inflammatory cell detection, and multi-organ PE&A inflammatory cell detection. For WSI-level classification tasks, we utilized ABMIL for weakly supervised learning, while ROI classification was evaluated using linear probing. Cell segmentation tasks were conducted using Mask2Former²⁵ architectures. Classification performance was assessed using AUC, ACC and F1 scores, while segmentation performance was evaluated using mAP, CPA, IoU, and Dice coefficients. Detailed descriptions of these tasks and experimental setups are provided in the Methods section.

We evaluated the performance of PathOrchestra across multiple tasks under the multi-organ lesion identification and analysis category, with segmentation results shown in Fig. 5. For the CoNSEp task, which is a multi-organ nuclear instance segmentation task, we used three

evaluation metrics: CPA, IoU, and Mean Dice. PathOrchestra performed poorly in this task, with an CPA of 0.556, IoU of 0.399, and Mean Dice of 0.550, indicating substantial room for improvement in nuclear instance segmentation. Similarly, for the WSSS4LUAD task, which is a semantic segmentation task, the CPA was 0.877, IoU was 0.771, and Mean Dice was 0.871. For the SegPath task, we segmented lesion regions in images using evaluation metrics including Precision, Recall, and Mean Dice. The experimental results showed that PathOrchestra achieved a Precision of 0.756, Recall of 0.720, and Dice of 0.717 in the SegPath task, demonstrating balanced segmentation performance and the ability to effectively identify and segment lesion regions. For the Gland segmentation task (based on the GlaS dataset), we evaluated the performance of PathOrchestra in glandular region segmentation, using Precision, Recall, and Dice as evaluation metrics. The results showed that the model achieved a Precision of 0.929, Recall of 0.868, and Dice of 0.931, indicating good performance in glandular region segmentation. For the

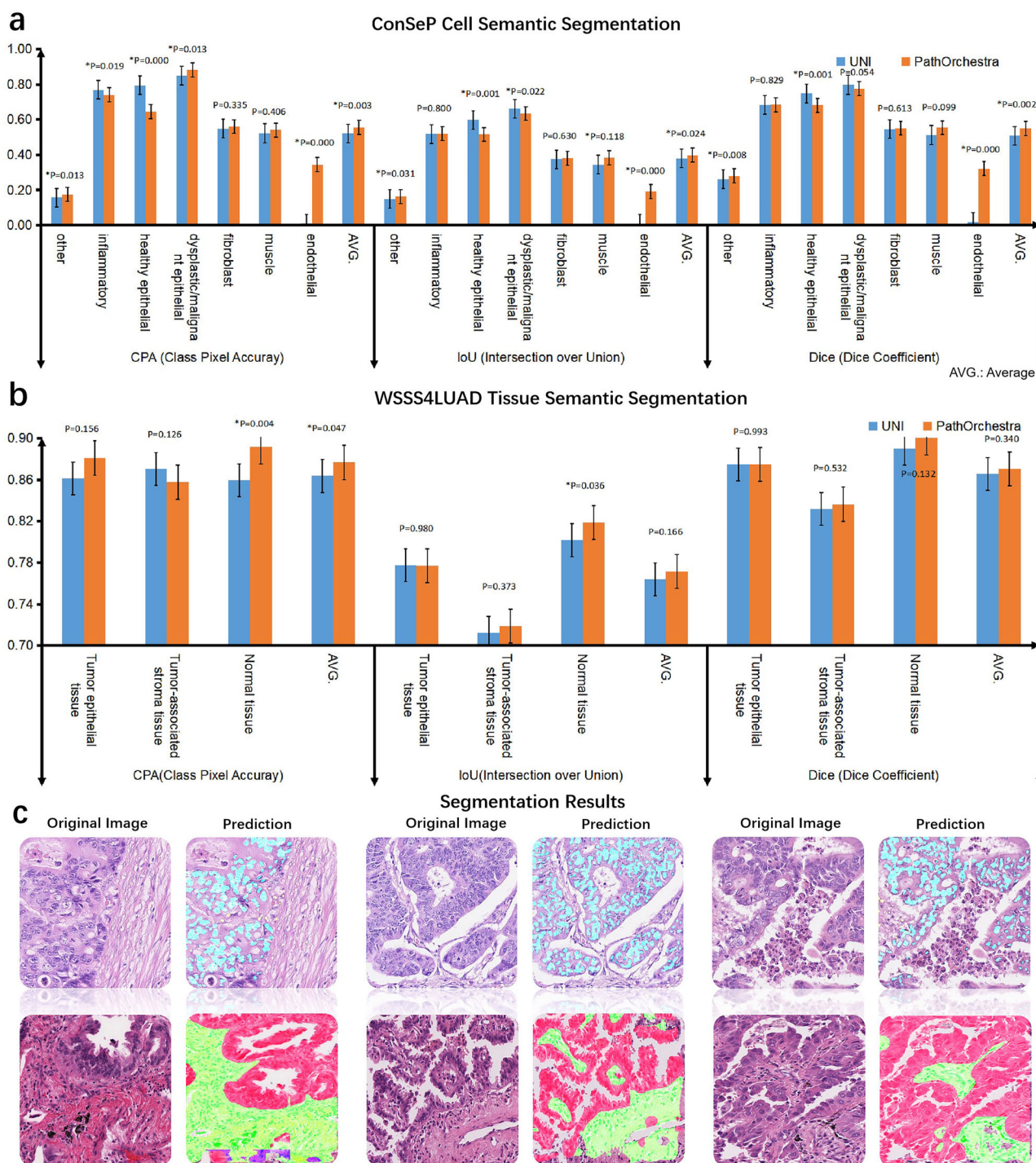


Fig. 5 | Cell and tissue segmentation results. **a** Comparative analysis of semantic segmentation performance between PathOrchestra and UNI on the ConSeP dataset for cell segmentation. The bar chart summarizes results for three metrics-Class Pixel Accuracy (CPA), Intersection over Union (IoU), and Dice coefficient-across multiple cell types. **b** Performance comparison on the WSSS4LUAD dataset for tissue segmentation, evaluated across three tissue categories: tumor epithelial tissue,

tumor-associated stromal tissue, and normal tissue. Results are reported for CPA, IoU, and Dice coefficient. **c** Representative qualitative examples of model predictions. Each pair shows the original histopathology image (left) alongside the corresponding segmentation output generated by the model (right). Abbreviations: CPA, Class Pixel Accuracy; IoU, Intersection over Union.

PanNuke task, PathOrchestra achieved a Precision of 0.923, Recall of 0.833, and Dice of 0.849, demonstrating excellent performance in nuclear segmentation.

In the mitosis detection task, we also evaluated the performance of PathOrchestra under the SegPath task, using mAP (bbox, IoU = 0.5), mAP (seg, IoU = 0.5), and mAP (seg, IoU = 0.75) as evaluation metrics. The

experimental results showed that the model's mAP (bbox, IoU = 0.5) and mAP (seg, IoU = 0.5) were 0.885 and 0.883, respectively, while mAP (seg, IoU = 0.75) was 0.658. These results indicate that PathOrchestra has limitations in accurately locating and segmenting mitotic cells, especially at higher segmentation precision requirements. For two cell detection tasks, namely TCT inflammatory cell detection and PE&A inflammatory cell

detection, the results showed that PathOrchestra achieved an mAP (bbox, IoU = 0.5) of 0.351 in the TCT inflammatory cell detection task, while the mAP in the PE&A inflammatory cell detection task was only 0.634, indicating considerable performance variation across different cell type detection tasks.

For other classification tasks, we used ACC, F1, and AUC as evaluation metrics. PathOrchestra performed excellently in the breast cancer metastasis detection task (WSI level), achieving 1.00 for ACC, F1, and AUC on the CAMELYON16 dataset, while its performance on CAMELYON17 was subpar, with an ACC of 0.840, F1 of 0.588, and AUC of 0.784. However, on the Pcam dataset, which is also a breast cancer classification model, PathOrchestra performed well, achieving an ACC of 0.916, F1 of 0.916, and AUC of 0.975. In other classification tasks, such as lymphoma PD-L1 tumor region detection, TIL positivity detection, and DigestPath tumor cell classification, PathOrchestra's ACC values were 0.851, 0.931, and 0.916, respectively, with AUCs of 0.927, 0.959, and 0.975, demonstrating good classification performance. However, in some tasks, such as COSAS screening and COSAS tissue classification, the model's AUC values were relatively low, at 0.897 and 0.624, respectively, indicating its shortcomings in complex tissue type classification tasks. Overall, PathOrchestra demonstrated some potential in multi-organ lesion identification and analysis tasks, but improvements are still needed in certain tasks to enhance the model's accuracy and robustness, particularly in more complex segmentation and cell detection tasks.

Multi-cancer subtype classification tasks

Tumor classification and subtyping are critical in enhancing diagnostic accuracy and personalizing treatment plans, thereby optimizing patient prognosis and clinical outcomes. We evaluated the performance of PathOrchestra across tumor classification and subtyping tasks, including cervical cancer subtyping, bladder cancer screening (WSI and ROI), cervical TCT screening (WSI and ROI), gastrointestinal benign and malignant screening, lymphoma and reactive hyperplasia screening, lymphoma cell origin classification, colorectal cancer detection, cervical epithelial lesion detection, colorectal and lung tissue screening (LC25K), lung cancer subtyping (LC25K-lung), colorectal subtyping (LC25K-colon), breast cancer screening (BreastHis), lung cancer subtyping (TCGA-NSCLC), colorectal subtyping (TCGA-RCC), lymphoma subtyping, breast cancer subtyping (BACH), esophageal cancer screening (TCGA-ESCA), colorectal tissue classification (HunCRC-patch), prostate ISUP grading (PANDA), prostate Gleason grading, colorectal high and low-grade screening, gastric cancer subtyping (PatchGastricAD22), prostate cancer tissue classification (AGGC), glioma IDH1 screening (TCGA-IDH1), multi-organ lymph node metastasis screening, colorectal polyp screening, colorectal subtyping (HunCRC-WSI), renal cancer tissue classification (TCGA-RCC), colorectal tissue classification (CRC-100K), colorectal subtyping (institutional data), colorectal tissue classification (Kather), and brain subtyping (Ebrains). Detailed information on these tasks and experimental setups is provided in the Methods section.

In the evaluation of 34 tumor classification and subtype identification tasks (Fig. 6), PathOrchestra demonstrated broad adaptability and high performance levels. Notably, in the classification tasks for cervical and bladder cancers, PathOrchestra excelled. For instance, the ACC for benign and malignant bladder cancer classification reached 0.954, with an F1 score of 0.943 and an AUC value of 0.989, showcasing the model's exceptional performance in this task. Furthermore, in the patch classification task for bladder cancer, the model's F1 score improved to 0.945, while the ACC and AUC values also reached 0.945 and 0.988, respectively, indicating that the model maintains efficiency even in more granular image processing scenarios. In the screening for benign and malignant gastrointestinal tumors, PathOrchestra exhibited remarkable precision, achieving an ACC of 0.993, an F1 score of 0.976, and an AUC as high as 0.988. These results highlight the model's strong diagnostic potential in the detection and classification of gastrointestinal cancers. In lymphoma-related classification tasks, PathOrchestra also demonstrated robust performance. For example, in the

task of classifying lymphoma versus reactive hyperplasia, the model achieved an ACC of 0.975, an F1 score of 0.941, and an AUC of 0.994. In the more refined classification of B/T-cell lymphomas, PathOrchestra recorded an F1 score of 0.918 and an AUC of 0.993, reflecting the model's high accuracy in subtype classification. Similarly, in the subtype classification tasks for colorectal and lung cancers, PathOrchestra performed admirably. In the colorectal subtype classification (LC25K-2), all evaluation metrics (ACC, F1, and AUC) achieved 1.000, demonstrating absolute classification accuracy. Likewise, for the lung cancer subtype classification (LC25K-3), the model attained a high ACC and F1 score of 0.999, with an AUC of 0.999, indicating strong generalization capabilities across different tissue types.

However, despite PathOrchestra's outstanding performance in most tasks, there are areas for improvement. For instance, in the IDH1 screening task for gliomas, although the AUC reached 0.940, the accuracy and F1 scores were relatively low at 0.896 and 0.894, suggesting that the model's precision in screening this specific subtype may not match its performance in other cancers. In more complex tasks, PathOrchestra's performance was comparatively weaker. For example, in the prostate cancer grading (AGGC) task, the model achieved an ACC of only 0.660, an F1 score of 0.828, and an AUC of 0.873, reflecting limitations in handling highly complex pathological tasks. Additionally, in glioma screening (Ebrains), although the AUC reached 0.871, the ACC and F1 scores were notably lower at 0.386 and 0.278, indicating a need for further optimization in brain tumor screening. In summary, PathOrchestra demonstrated exceptional accuracy and robustness across various cancer classification and subtype identification tasks, particularly excelling in bladder cancer, gastrointestinal cancers, and colorectal cancer tasks. Nevertheless, there remains room for improvement in certain complex tasks, such as prostate cancer grading and glioma screening. This indicates that while PathOrchestra performs excellently in most tasks, further optimization is necessary for specific tasks to enhance its overall diagnostic capability.

Biomarker assessment tasks

Biomarker evaluation is critical for clinical diagnosis, treatment selection, and prognostic assessment. We evaluated PathOrchestra across 36 biomarker-related tasks involving the qualitative assessment and scoring of commonly used IHC markers. These included over 30 protein markers with varied biological functions, expression patterns, and clinical significance. The model demonstrated robust performance across diverse markers and datasets, highlighting its adaptability to complex molecular evaluation tasks. Full task descriptions and experimental details are provided in the Methods section.

PathOrchestra achieved varying levels of performance across the 36 tasks as shown in Fig. 7. High accuracy was observed in markers such as HER2 scoring and CD20 qualitative analysis, with performance exceeding 0.920. For more complex tasks involving combinations of IHC markers, accuracy ranged from 0.800 to 0.880. The model showed promising capability in analyzing nuanced histopathological features, with markers like PD-L1 and CyclinD1 yielding strong results. In contrast, tasks involving subtle expression differences, such as CD56 and CXCL-13, were more difficult, with accuracies around 0.700. These results suggest the need for further model refinement and more comprehensive training datasets. Evaluation of hormone receptor-related biomarkers, including ER and PR, also showed solid performance, with accuracies around 0.900. Overall, these findings demonstrate the strong potential of PathOrchestra for biomarker evaluation, while also indicating areas where future improvements are needed.

Gene expression prediction tasks

Gene expression evaluation in computational pathology significantly enhances both clinical and research applications by enabling personalized treatments, monitoring disease progression, and driving innovation. We evaluated the performance of PathOrchestra across 10 gene expression prediction tasks, including multi-disease gene expression and specific assessments for invasive ductal carcinoma (IDC) in breast cancer, prostate

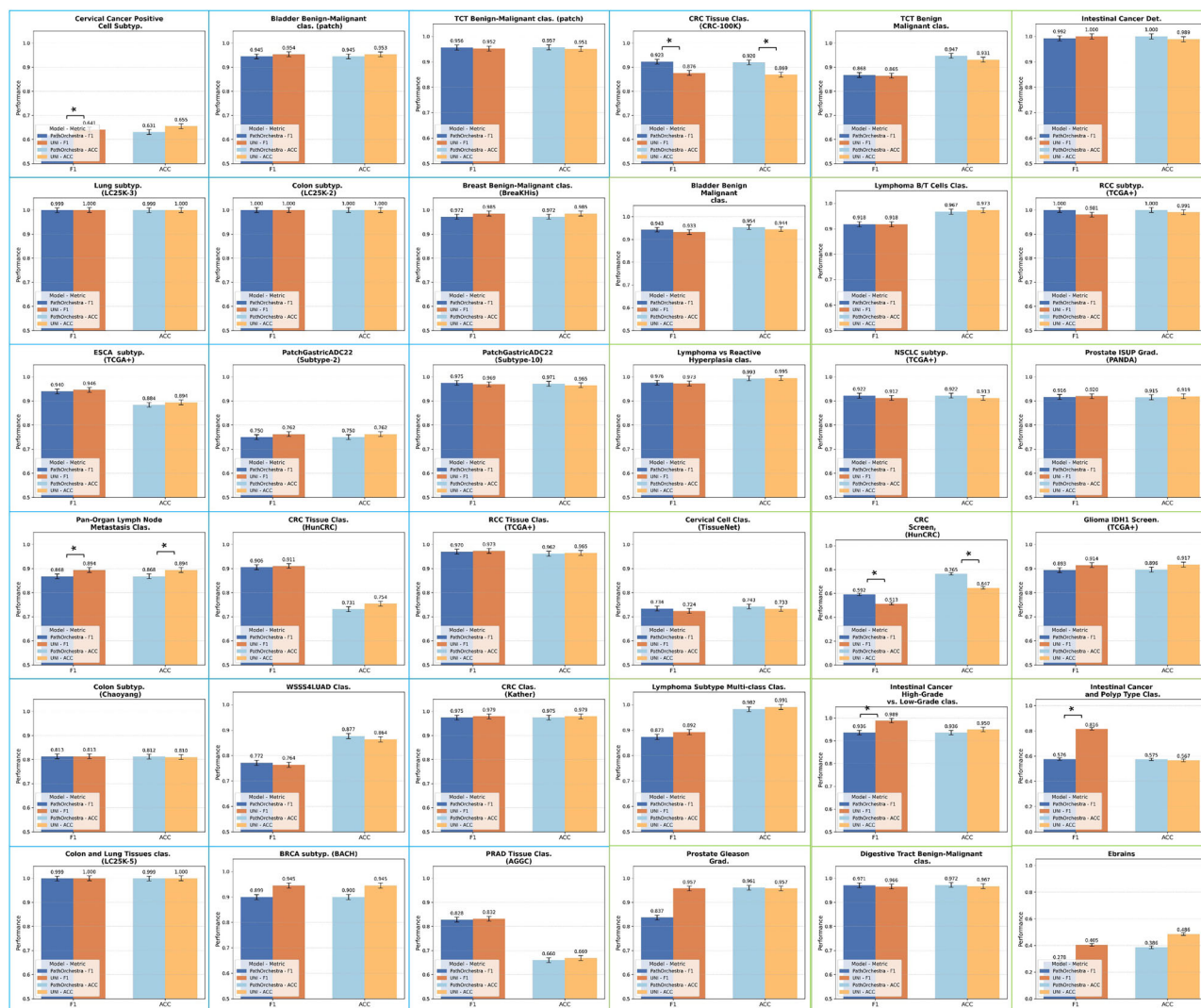


Fig. 6 | Results of multi-cancer subtype classification. The bar charts show PathOrchestra's performance across region-of-interest (ROI)-level (blue outlines) and whole-slide image (WSI)-level (green outlines) multi-cancer subtype classification tasks. Key performance metrics, including accuracy (ACC) and F1 score, are presented for each task. Results are compared between public and private datasets,

demonstrating the model's robustness across diverse classification challenges. Each panel displays a separate classification setting, with ROI-level and WSI-level curves overlaid for direct comparison. All colors and abbreviations are defined as follows: ROI, region of interest; WSI, whole-slide image; ACC, accuracy.

cancer (PRAD), skin cancer (SKCM), colon cancer (COAD), rectal cancer (READ), clear cell renal cell carcinoma (CCRCC), liver cancer (HCC), lung adenocarcinoma (LUAD), and lymph node IDC. Detailed information regarding the gene datasets and experimental setups is provided in the Methods section.

As illustrated in Fig. 8, we compared the performance of PathOrchestra with other foundational models, GigaPath²¹ and UNI²², in predicting gene expression across multiple cancers. PathOrchestra demonstrated superior performance in five cancer types: LUAD, READ, SKCM, PRAD, and IDC. On average, the model outperformed UNI by 1% and GigaPath by 5.7% across these five cancer types. In the PRAD dataset, PathOrchestra achieved 6.5% higher accuracy compared to UNI and 6.1% higher than GigaPath. For the IDC dataset, although all three models performed above 0.700, PathOrchestra exceeded UNI by 4.4% and GigaPath by 1%. Additionally, we conducted a detailed comparison of gene prediction capabilities between PathOrchestra and UNI across LUAD, PRAD, READ, and SKCM. The proposed model showed relatively stable performance across multiple gene categories, with particularly strong results in the PRAD dataset, significantly surpassing UNI for genes such as MT-ATP6, PRL34, PRL12, PRL26, and

PRL29. These findings demonstrate that PathOrchestra is capable of predicting gene expression profiles from H&E-stained images, identifying distinct gene expression patterns across various cancers.

Structured report generation tasks

Generating pathology reports is one of the most challenging clinical tasks in computational pathology, primarily due to the large number of WSIs per patient, the complexity of diagnostic content, and the difficulty in predicting subtasks. By selecting appropriate disease types and breaking down the pathology report into a structured format, we can use artificial intelligence algorithms to generate pathology reports from diagnostic and auxiliary examination perspectives. This approach can effectively save time of pathologists while providing reasonable diagnostic explanations, thus holding significant value for clinical practice. This study focuses on two diseases, lymphoma and colorectal cancer, employing a subtask combination method to generate structured reports for each patient based on the PathOrchestra model.

In the task of determining lymphoma subtypes, conventional HE image diagnoses often do not provide sufficient evidence to identify specific

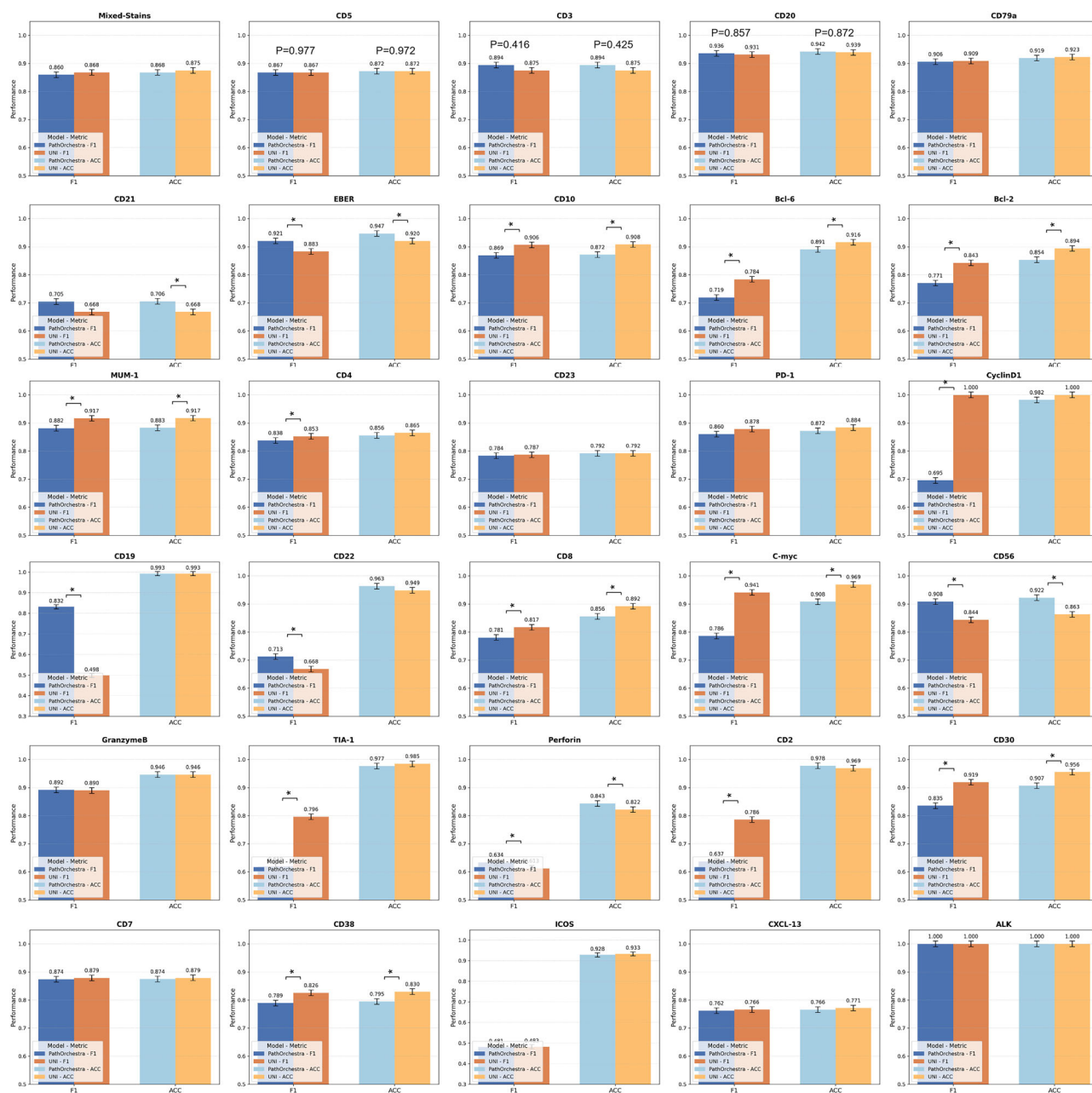


Fig. 7 | Performance on biomarker expression prediction. The bar charts report accuracy (ACC) and F1 score for each downstream clinical task, comparing PathOrchestra and UNI side by side. Each panel corresponds to a distinct clinical

task; bars are grouped by model and metric to enable direct comparison. Abbreviations: ACC, accuracy; F1, F1 score.

subtypes, usually requiring quantitative analysis from IHC, with different subtype suspected patients corresponding to different immunohistochemical panels. This paper emphasizes the AITL and DLBCL subtypes, combining the HE subtype prediction task with the qualitative assessment of 29 IHC markers, and diagnosing multiple differently stained pathology slides for each patient, resulting in structured reports. As seen in example from Fig. 9a, the model demonstrates the ability to diagnose lymphoma subtypes and provides detailed diagnostic explanations via immunohistochemistry. For instance, in Patient 1, the algorithm incorrectly predicted CD21 as negative, and due to data loss during the collection process, it was unable to provide effective diagnoses for CD20, CXCL-13, and CD10, while the qualitative assessments of the other IHC markers were consistent with the actual diagnoses by pathologists. The results of subtype diagnosis and

accompanying IHC reports indicate that our model is capable of making comprehensive diagnoses for this complex disease.

Colorectal cancer is a prevalent disease with a complex initial report. This study focuses solely on the assessment of HE images to distinguish between tumor and non-tumor, as well as to classify cancer grades (cancerous, low-grade cancer, high-grade cancer). Additionally, for negative samples, we further classify polyp types (including hyperplastic polyps, inflammatory polyps, polypoid hyperplasia, and no polyps). As illustrated in example from Fig. 9b, the model exhibits the ability to screen for colorectal cancer and determine the severity of the tumor. For instance, Patient 1 was diagnosed by the algorithm as a tumor type and identified the tumor grade as high-grade. This diagnosis aligns with the actual result of adenoma with focal high-grade intraepithelial neoplasia. Thus, our model demonstrates

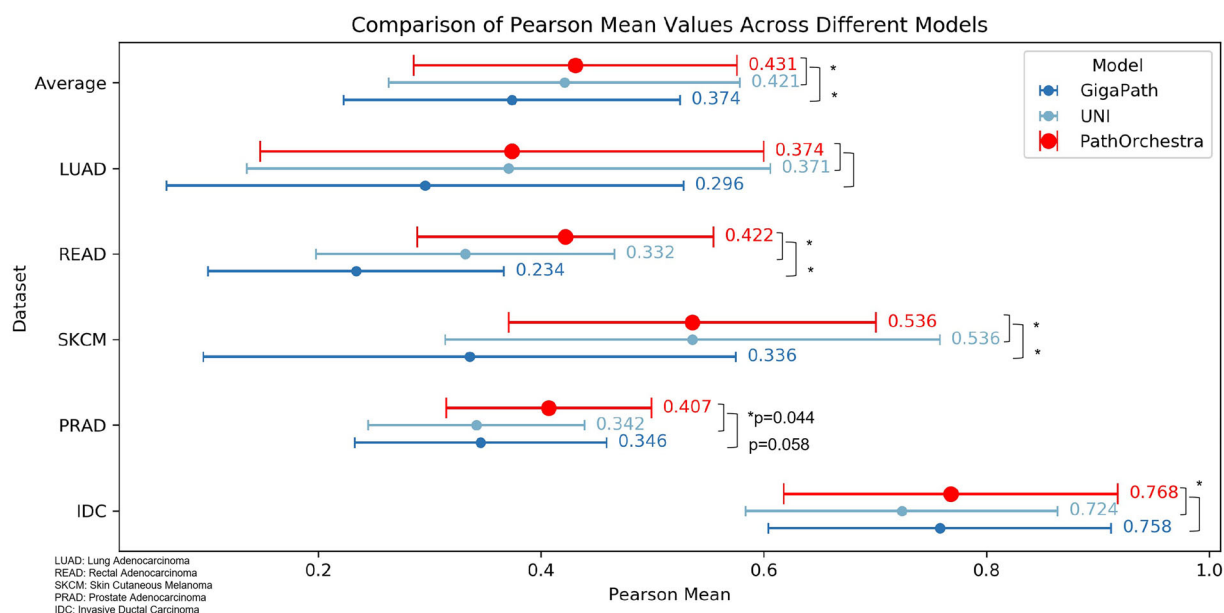
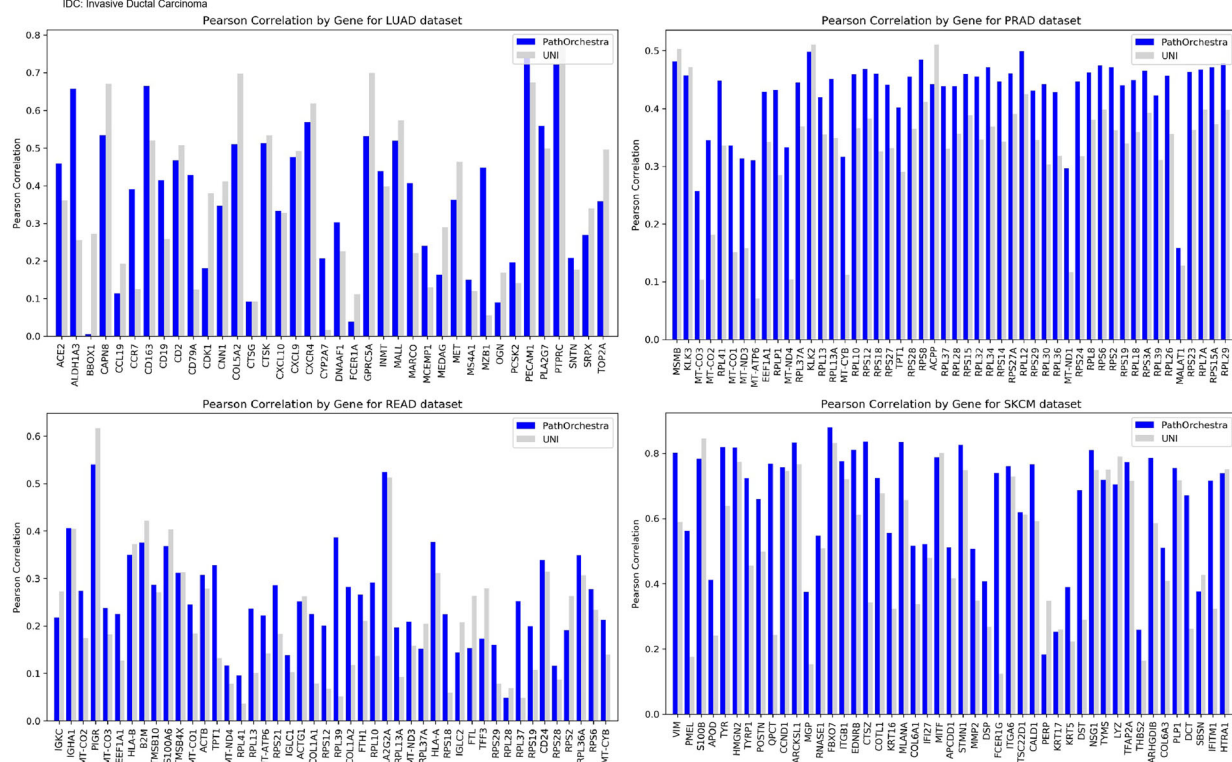
a**b**

Fig. 8 | Comparison of gene expression evaluation results. **a** The performance advantage of the model over GigaPath and UNI in gene expression prediction was evaluated across five cancer types: LUAD, READ, SKCM, PRAD, and IDC. **b** A detailed comparison of the model and UNI was conducted on LUAD, PRAD, READ,

and SKCM, focusing on their predictive capabilities across multiple gene categories within each cancer type. Abbreviations: LUAD, Lung Adenocarcinoma; READ, Rectal Adenocarcinoma; SKCM, Skin Cutaneous Melanoma; PRAD, Prostate Adenocarcinoma; IDC, Invasive Ductal Carcinoma.

the capability to generate preliminary diagnostic reports for this common disease.

Discussion

The recent emergence of foundational models has transformed computational pathology by enabling the analysis of rare diseases, reducing reliance on extensive manual annotations, improving the efficiency of downstream training, and supporting the development of complex structured tasks. In this study, we present PathOrchestra, a self-supervised vision encoder trained on a scale of 287,424 WSIs

spanning 21 distinct human tissues, organs and systems. PathOrchestra demonstrates generalization across 27,755 WSIs and 9,415,729 ROI images, achieving over 0.950 accuracy in 47 tasks, including pan-cancer classification, lymphoma subtyping, and bladder cancer screening. Importantly, it is the first model to generate structured diagnostic reports for high-incidence colorectal cancer and diagnostically complex lymphoma. These tasks are rarely addressed by existing foundational models, yet they hold substantial clinical significance. These results highlight the translational potential of PathOrchestra in computational pathology.

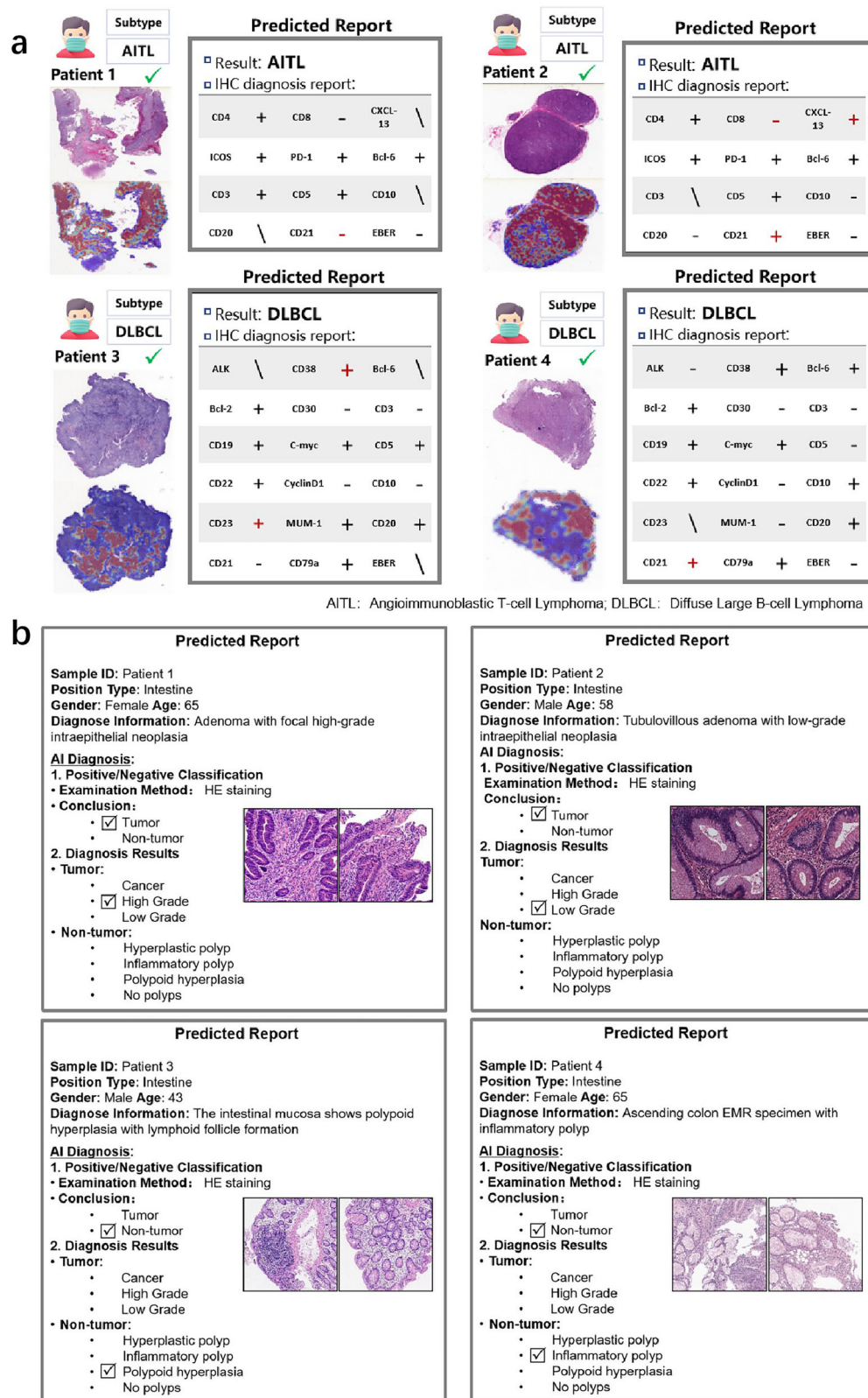


Fig. 9 | Structured report results for lymphoma and intestine. **a** Four patients diagnosed with lymphoma, including AITL and DLBCL subtypes, were selected to demonstrate the structured report generated by the PathOrchestra model. This report integrates predictions for lymphoma subtypes and qualitative IHC tasks across 30 combined tasks. It includes subtype diagnosis based on H&E-stained images and the corresponding IHC results, with the IHC panels varying for each

subtype. **b** The pathology report for colorectal cancer was streamlined to focus on predicting malignancy (positive and negative) and tumor grading (cancer, low-grade, high-grade) based on H&E-stained images. Additionally, for negative samples, the model further classifies the type of polyps. The patient logo in figure a was generated using GPT.

There is no effective consensus on benchmarks for evaluating foundational pathology models, with unclear definitions on the perspective (technical or clinical), specific tasks, and the scope of evaluation (number of tasks, classifications, representative tasks, etc.). Although previous studies have performed multi-faceted evaluations, such as the UNI model covering 34 types of downstream tasks, these evaluations are predominantly from a technical standpoint. How to effectively utilize foundational pathology models and integrate them with actual clinical needs to maximize their value and promote their development at both technical and application levels remains a critical issue. A significant contribution of this study is the in-depth exploration of how large pathology models can be integrated with clinical applications to more effectively promote their practical implementation. By addressing a wide array of tasks, including digital slide pre-processing, pan-cancer classification, lesion identification, multi-cancer subtyping, biomarker assessment, gene expression prediction, and structured report generation, we have showcased the model's versatility and its potential to meet real-world clinical needs.

The multifaceted capabilities of PathOrchestra in structured report generation, pan-cancer classification, and gene expression prediction signify its vital role in advancing clinical integration and enhancing diagnostic practices for complex diseases. First, our exploration into structured report generation for complex diseases such as lymphoma and colorectal cancer represents a significant step toward practical clinical integration. By decomposing pathology reports into structured formats and combining multiple subtasks including subtype diagnosis and qualitative IHC marker analysis, PathOrchestra effectively generates comprehensive reports that can enhance diagnostic accuracy and efficiency. This approach not only alleviates the workload of pathologists but also contributes to standardized reporting practices. Second, the impressive performance of PathOrchestra in pan-cancer classification tasks underscores its ability to generalize across diverse cancer types, specimen preparation methods, and data sources. Achieving AUCs exceeding 0.930 in both in-house and TCGA datasets, the model demonstrates robust diagnostic capabilities that are crucial for early cancer detection and personalized treatment planning. Third, its superior performance in gene expression prediction tasks outperforming models like GigaPath and UNI by significant margins, highlights its potential in advancing molecular pathology by providing insights into gene expression profiles directly from H&E-stained images.

We observed relatively lower performance in certain fine-grained tasks, especially pan-organ IHC marker classification and multi-class ROI-level segmentation. The difficulty stems from three main factors: first, visual similarity among markers (e.g., nuclear markers like PR, ER, Ki67; membrane markers like HER2), second, limited marker diversity and class imbalance, and third, the use of linear probing, which constrains task-specific adaptation. Additionally, in multi-slide colorectal structured reporting, the model faced challenges in integrating dispersed lesion information and inferring detailed attributes such as vascular invasion or margin status. These findings highlight the need for multimodal integration, hierarchical modeling, and task-specific fine-tuning.

To better understand the impact of architectural design and training dynamics on downstream performance, we conducted ablation experiments focusing on two key factors: backbone size (ViT-Large vs. ViT-Giant) and training iteration count. As shown in Supplementary Fig. 1a, models using the ViT-Giant backbone consistently outperformed those using ViT-Large across both linear probing and KNN evaluation schemes. This suggests that increasing model capacity leads to richer feature representations that generalize better across clinical tasks—particularly for nuanced classification problems where subtle morphological differences must be captured. Moreover, Supplementary Fig. 1b shows the effect of training iteration count on model performance. With longer pretraining (more epochs), both accuracy and F1 score improved notably, especially under the linear probing evaluation. These results indicate that model architecture and training schedule play a significant role in determining downstream generalization. Importantly, they suggest that foundational models should not be treated as one-size-fits-all encoders, and thoughtful adjustments to backbone capacity

and pretraining length are crucial when targeting different downstream clinical tasks.

The reasons for the model's enhanced performance in downstream tasks can be summarized in three key aspects. First, the model was trained on 287,424 high-quality WSI pathology images ensuring rich feature learning with a 304M parameter model. Second, the data was sourced from three major centers, encompassing diverse populations from Asia, the Americas, and other regions, enhancing the model's adaptability and generalization, leading to more reliable performance across various clinical contexts. Finally, the implementation of a self-supervised learning strategy effectively mined features from unlabeled data, improving the model's representation capabilities for pathology images and its transfer learning efficiency across different tasks.

Despite above advancements, several limitations warrant discussion. First, while the model demonstrates strong performance in tumor-related tasks, its capability in handling inflammatory conditions and distinguishing between neoplastic and reactive processes remains limited. The pan-cancer evaluations primarily focused on tumor types, and future work should incorporate a broader spectrum of pathological conditions, including inflammatory and infectious diseases, to enhance the model's diagnostic utility. Second, the imbalance in training data across different tissue types and disease categories may influence the model's performance. Although we utilized a vast dataset, ensuring balanced representation is essential for unbiased learning and generalization. Future studies will investigate the effects of data balance and explore strategies such as data augmentation or synthetic data generation to address this issue. Third, the unique characteristics of pathological data such as the gigapixel scale of WSIs and the necessity for multi-resolution analysis, call for specialized model architectures and training strategies. Future work should focus on optimizing the model for WSI-level analysis, incorporating hierarchical and multi-scale representations to capture both global tissue architecture and cellular-level details. Fourth, this study represents the exploration of applying vision foundation models for generating structured pathology reports. We focused on gastrointestinal cancer and lymphoma, generating reports that cover basic diagnostic classifications, demographic data, and IHC marker expression. However, generating structured reports for complex cancers, such as colorectal cancer, presents significant challenges, particularly because it requires the integration of up to 40 subtasks and the analysis of multiple slides from the same patient. Additionally, critical tasks such as vascular invasion and margin assessment are essential for accurate prognosis and treatment planning, yet these are not covered by the current model. Future work will address these missing tasks, integrating multimodal data to enhance the reports with microscopic descriptions and improving the standardization and accuracy of AI-assisted pathology diagnostics. Additionally, The model's lower performance in pan-organ IHC classification stems from the inherent visual similarity of IHC markers and insufficient dataset diversity. While we employed a pretrained vision model and evaluated it using Linear Probing, this method has limitations in handling multiple markers. To address these issues, future work will focus on increasing dataset diversity, implementing multi-task learning, and adopting more complex architectures. Incorporating clinical data and enhancing model interpretability will further increase its applicability and reliability in clinical settings. Finally, the exploration of multimodal approaches represents another promising avenue. Integrating data from various modalities, such as genomics, proteomics, and radiology, alongside pathology images, could provide a more comprehensive understanding of disease processes. Additionally, developing models capable of natural language processing tasks such as interpreting clinical notes or generating narrative reports, could further enhance the model's clinical applicability.

In conclusion, PathOrchestra represents a substantial advancement in the development of foundational pathology models, demonstrating exceptional performance across a wide range of clinically relevant tasks. By addressing both technical and clinical perspectives, this work contributes to bridging the gap between computational pathology research and real-world clinical practice. Ongoing efforts to refine the model, expand its capabilities,

and facilitate its integration into healthcare settings hold the promise of significantly improving diagnostic accuracy, personalized treatment, and ultimately patient outcomes in the future.

Methods

Network architecture and pretraining protocol

To achieve robust performance across 112 clinical-grade downstream tasks, we pre-trained our model on the curated dataset using the DINOv2 architecture¹⁶, which is designed to extract high-quality visual features through self-supervised learning on unlabeled data. DINOv2 employs a teacher-student framework based on Vision Transformers (ViT)²⁶, leveraging an alignment mechanism that enables the student network to learn meaningful feature representations without relying on labeled data. By eliminating the dependence on manual annotations, DINOv2 allows for the extraction of rich features from large-scale unlabeled images, significantly reducing the need for human intervention. The architecture enhances feature generalization through multi-scale, multi-view data augmentation techniques. By incorporating image crops at different scales and applying diverse augmentations, the model learns feature representations that are robust and generalize well across various downstream tasks, including classification, detection, and segmentation.

We adopt the pretraining strategy of DINOv2 to develop a pathology-oriented visual foundation model on a large-scale unlabeled histopathology dataset. DINOv2 employs a teacher-student framework trained via multi-view self-distillation. Each training image is augmented into a set of global and local crops. The global views encourage the model to capture high-level tissue architecture, while the local crops focus on fine-grained features such as cellular morphology and nuclear atypia. A key component of DINOv2 is its implicit sample selection mechanism: only a subset of local views is selected to compute the loss against the global teacher outputs, avoiding explicit negative sampling. This contrastive-like design is well-suited for histopathology, where structural hierarchies are consistent within images but highly variable across samples. This strategy avoids reliance on negative sampling while encouraging the student network to align with meaningful and diverse representations from different spatial scales and contexts. Furthermore, all image crops undergo standardized data augmentation including random cropping, color jittering, grayscale conversion, Gaussian blurring, and horizontal flipping. These augmentations simulate domain-specific variations such as differences in staining protocols, section thickness, and scanner types. The combination of multi-scale view sampling and strong augmentation enables the model to focus on structure-dominated features, improving the robustness and generalizability of learned representations across diverse pathology sources.

The training process of DINOv2 primarily consists of the following three steps. The first step is data preprocessing and augmentation, which involves applying various augmentations such as random cropping, horizontal flipping, color jittering, grayscale conversion, and Gaussian blur to the original images. These operations generate multiple global and local crops, which are simultaneously fed into the student network for feature extraction. The second step is feature extraction and alignment. The student network extracts features from both global and local crops, while the teacher network extracts features solely from the global crops. The DINO and iBOT heads are then used to compute the alignment loss between the features extracted by the student and teacher networks. The third step involves computing the loss and updating the network. The total loss is calculated using the DINO²⁷ loss, iBOT²⁸ loss, and KoLeo regularization²⁹, after which the student network's parameters are updated through backpropagation. Simultaneously, the teacher network's parameters are updated using the Exponential Moving Average (EMA) mechanism³⁰, ensuring that the teacher network remains smoother and more stable. In the experiments conducted in this paper, we chose to use the teacher network as the encoder for downstream tasks. This decision is grounded in the observation that the teacher network, which is updated through EMA, exhibits smoother and more stable parameters. Consequently, this leads to enhanced generalization capabilities. Therefore, the encoder of the teacher network is more

adept at handling downstream tasks that demand high robustness and applicability across different domains.

Clinical task assessment methodology

We evaluated the pre-trained model on a diverse set of 112 clinically relevant downstream tasks, including pathology quality control, pan-cancer classification, lesion detection, multi-cancer subtyping, biomarker assessment, gene expression prediction, and structured report generation.

Pathology Image Preprocessing and Quality Control Tasks. In clinical pathology, image quality is crucial for the accuracy of the final diagnosis. Image preprocessing eliminates factors that affect image quality, such as focus blur, tissue folding, bubbles, and impurities, ensuring clarity and diagnostic value. Quality control tasks automate image screening to exclude low-quality images that may compromise subsequent analysis. This approach significantly reduces manual intervention, enhancing pathologists' efficiency and diagnostic accuracy. To assess the model's ability to handle common artifacts and variations in pathological images, we conducted several quality control tasks focusing on image quality assessment and recognition of staining methods, magnifications, and specimen types. For image quality assessment, we developed classifiers to detect various issues that can impede accurate pathological analysis. These issues included differentiating between natural images (sourced from the ImageNet-1K validation set) and pathological images (random patches from gastrointestinal H&E and immunohistochemistry data), identifying tissue folds, bubbles, glue artifacts, contaminants such as dust, hair, and fibers, and detecting blurriness using the FocusPath-UofT dataset³¹. In terms of staining and specimen recognition, we evaluated the model's capability to recognize different staining methods, magnifications, and specimen types. We classified images based on staining type, distinguishing between H&E and IHC stained images, as well as between H&E stained images and immunofluorescence images from the DeepCell dataset³². We also classified four types of IHC markers (PR, ER, Ki67, and HER2) to assist in molecular characterization, differentiating IHC markers based on nuclear (PR, ER, Ki67) and membrane (HER2) positivity. Furthermore, we developed models to distinguish images at different magnifications (10 × and 20 ×) to improve model robustness and to identify whether H&E images were from frozen sections or FFPE samples. Specimen type recognition included classifying images based on specimen size to enhance preprocessing workflows. For these tasks, we utilized datasets comprising thousands to hundreds of thousands of images, employing linear probing strategies.

Pan-cancer Classification Tasks. Pan-cancer classification tasks involve unified modeling of image data from different cancer types to assist pathologists in quickly identifying the cancer type during pathological examination, preventing misdiagnosis due to incorrect cancer type assumptions. Clinically, pathology departments often face cases where the cancer origin is unclear, especially during initial assessments when pathologists may not be able to immediately identify the cancer type. By training a pan-cancer classification model, the system can provide preliminary diagnostic suggestions, supporting early screening. Additionally, multi-cancer classification improves the recognition of rare cancers, enhancing the model's generalization ability in cases of data scarcity. To assess the model's ability to distinguish between major cancer types across organs, we defined a set of pan-cancer classification tasks, which involve coarse-grained, multi-class classification at the cancer-type level. These tasks simulate real-world diagnostic scenarios where slides from different organs are mixed, and prior knowledge of cancer type may be unavailable. Such models can assist in initial triage, cross-organ screening, and error detection. We implemented three pan-cancer tasks: (1) a 17-class classification task using 1071 WSIs from an in-house dataset spanning cancers of the lung, kidney, breast, prostate, and other organs. (2) a 32-class FFPE pan-cancer classification task based on 3,036 WSIs from TCGA. (3) a 32-class frozen pan-cancer classification task based on 3,038 TCGA WSIs. All tasks were formulated using the weakly supervised ABMIL framework³³ with consistent training parameters to ensure comparability. These pan-cancer tasks are distinct from cancer subtyping tasks, which focus on finer-grained

classification within a single cancer type (e.g., lung adenocarcinoma vs. squamous cell carcinoma). In contrast, pan-cancer classification emphasizes organ-level differentiation and benefits from shared representation learning across diverse cancer types.

Multi-organ Lesion Identification and Analysis Tasks. The core of this task is to automatically identify and analyze lesion areas in pathology slides using deep learning techniques. This technology has broad clinical applications, particularly in the early detection and accurate localization of tumors. For pathology images from different organs, the model can identify lesion regions within seconds and automatically segment potential tumor tissue. Pathologists can use these analyses for further diagnosis, improving diagnostic efficiency and accuracy while reducing the risk of missed or incorrect diagnoses. We assessed the model's capability in lesion detection and segmentation across multiple organs and modalities. For lesion detection, we trained models to detect breast cancer metastasis in lymph node WSIs using datasets from the CAMELYON16 and CAMELYON17 challenges. We developed models to segment regions for PD-L1 assessment in lymphoma, enhancing immunotherapy decision-making, and classified images based on the presence of tumor-infiltrating lymphocytes using the TCGA-TILs dataset. Additionally, we performed binary classification of metastatic cancer in breast tumor regions using the PCam dataset, which contains over 2 million image patches, and identified signet ring cell carcinoma in gastrointestinal sections using the DigestPath dataset. Training strategies included linear probing and fine-tuning, with images processed at appropriate magnifications. For segmentation tasks, we employed the Mask2Former model²⁵, fine-tuned with a Vision Transformer Adapter (ViT-Adapter)³⁴, to segment eight major cell types within tumor tissues using the SegPath dataset, segment glands in colorectal adenocarcinoma tissue sections using the GlaS dataset, and segment nuclei across 19 tissue types using the PanNuke dataset. We also utilized the ViTDet model³⁵ for few-shot detection tasks, detecting and classifying positive cells in cervical cell pathology ROIs with over 1.3 million images and identifying abnormal cells in pleural and peritoneal fluid pathology slides using 36,544 ROI images.

Multi-cancer Subtype Classification Tasks. We define cancer subtyping tasks as those aimed at capturing finer-grained distinctions within a single cancer type, including histological subtypes, tumor grades, and benign-malignant classification. These distinctions are crucial for tailoring treatment strategies and making accurate prognostic assessments, ultimately improving treatment outcomes and patient survival. We evaluated the model's capacity to perform both binary and multi-class classification across diverse organs. Binary classification tasks included distinguishing cancer from benign tissues in bladder, cervical, and digestive tract cancers, as well as differentiating lymphoma from reactive hyperplasia and B-cell from T-cell lymphomas. For multi-class subtyping, we conducted organ-specific classification tasks: detecting cervical precancerous lesions using the TissueNet dataset, categorizing tissue types with LC25K subsets, and classifying benign vs. malignant breast tumors at high magnification using BreakHis. We further subtyped non-small cell lung cancer and renal cell carcinoma using TCGA data, performed detailed lymphoma subtyping to guide treatment, graded prostate cancer using core needle biopsies and Gleason scores, and identified gastric adenocarcinoma subtypes using the PatchGastricADC22 dataset. Model training employed both linear probing and the ABMIL framework to optimize classification accuracy.

Biomarker Assessment Tasks. We assessed the model's ability to evaluate the expression of biomarkers critical for diagnosis and treatment planning, particularly in the context of IHC. Biomarkers play a crucial role in tumor subtyping, prognostic assessment, and treatment decision-making. The expression of each biomarker reflects the biological characteristics of the tumor, which in turn influences clinical decisions. For example, HER2, ER, and PR markers are critical for targeted therapy in breast cancer. By automating the evaluation of these biomarkers, we significantly enhance diagnostic efficiency and reduce human interpretation bias, especially in cases where marker expression is low or the samples are complex. The selection of biomarkers was based on their established clinical

significance in tumor diagnosis, subtype classification, prognostic assessment, and treatment decisions. We performed multi-IHC marker qualification by assessing the expression of 30 different IHC markers in lymphomas using 3,976 WSIs. These markers were chosen for their pivotal role in identifying lymphoma subtypes, evaluating prognosis, and guiding treatment decisions. Specific marker analysis was conducted for markers such as CD5, CD3, CD20, CD79a, CD21, EBER, CD10, Bcl-6, Bcl-2, MUM-1, CD4, CD23, PD-1, Cyclin D1, CD19, CD22, CD8, C-myc, CD56, Granzyme B, TIA-1, Perforin, CD2, CD30, CD7, CD38, ICOS, CXCL-13, ALK, PD-L1, HER2, ER, PR, Ki67, and others. These markers were selected for their known utility in lymphoma diagnosis, immunophenotype characterization, and prognostic value. In addition, we evaluated breast cancer biomarkers by assessing HER2 scoring and ER, PR, and Ki67 status, all of which are critical for breast cancer subtyping and treatment planning. Each task involved datasets ranging from 11 to 313 WSIs, ensuring robust and diverse evaluations. For all tasks, we consistently employed a weakly supervised classification method using the ABMIL framework with standardized parameters, ensuring consistency and reproducibility across all evaluations.

Gene Expression Prediction Tasks. Gene expression data is crucial for analyzing the molecular characteristics of cancers and for enabling precision treatment. By predicting gene expression, the model can provide pathologists with molecular-level diagnostic information, complementing traditional histological examination. This task holds significant potential in the molecular diagnosis of tumors and the identification of drug targets. To evaluate the model's ability to predict gene expression levels from pathological images using the HEST-Benchmark dataset¹⁰. We predicted expression levels of the top 50 genes across nine cancer types, including Invasive Ductal Carcinoma (IDC), Prostate Adenocarcinoma (PRAD), Skin Cutaneous Melanoma (SKCM), Colonic Adenocarcinoma (COAD), Rectal Adenocarcinoma (READ), Clear Cell Renal Cell Carcinoma (ccRCC), Hepatocellular Carcinoma (HCC), Lung Adenocarcinoma (LUAD), and Axillary Lymph Nodes in IDC (LYMPH_IDC). By integrating and summarizing these tasks according to their major categories, we provide a cohesive overview of our evaluation of the pre-trained model across a diverse range of clinically relevant applications. This approach emphasizes the model's versatility and potential impact in various domains of pathology. Detailed descriptions of each task, including datasets, experimental setups, and training parameters, are available in the Supplementary Information for further reference.

Structured Report Generation Tasks. We explored the potential of directly applying vision foundation models for structured report generation. In clinical practice, generating reports for gastrointestinal cancer is complex, primarily involving the diagnostic classification and qualitative assessment of multiple slides. AI has the potential to assist pathologists in generating structured reports, thereby improving work efficiency. Similarly, lymphoma, with its numerous subtypes and IHC markers, presents significant complexity in both subtype diagnosis and immunohistochemical evaluation. Moreover, regional disparities in healthcare resources exacerbate development imbalances. Compared to traditional report generation methods, we opted for a structured approach, as it provides more accurate and reliable results. In this study, we utilized the PathOrchestra model to generate structured reports for each patient, combining multiple task subsets. For lymphoma report generation, we first analyze conventional H&E stained images to diagnose lymphoma subtypes. However, H&E staining often fails to definitively identify specific subtypes, necessitating the integration of IHC for more precise analysis. Different lymphoma subtypes correspond to distinct IHC panels. This study focuses particularly on diagnosing Angioimmunoblastic T-cell Lymphoma (AITL) and Diffuse Large B-Cell Lymphoma (DLBCL), combining H&E subtype prediction tasks with qualitative assessments of 29 IHC markers. We employ weakly supervised learning to diagnose H&E stained pathology slides and conduct qualitative analyses for each IHC type, ultimately generating the structured reports. In colorectal cancer, we focus on distinguishing tumor from non-tumor tissues through H&E images and classifying cancer grades

(cancerous, low-grade, high-grade). For non-malignant samples, we further categorize polyp types, including hyperplastic polyps, inflammatory polyps, polypoid hyperplasia, and cases with no polyps. For each patient, we utilize weakly supervised learning to train classification models for various tasks, resulting in the generation of structured reports.

Evaluation protocol

Given that PathOrchestra is a patch-level vision foundation model, we adopted different evaluation strategies depending on the label granularity of each downstream task. For tasks with patch-level or region-level supervision, we applied linear probing on frozen features to evaluate the discriminative capacity of the encoder. For whole-slide classification tasks with only slide-level labels, such as biomarker expression assessment, we employed weakly supervised learning using the ABMIL framework to aggregate patch-level features into slide-level representations. This protocol ensures a consistent evaluation of the model across both localized and global supervision settings, following practices from previous foundation models such as UNI. For segmentation tasks, we employed the widely used Mask2Former architecture as a decoder on top of the frozen patch-level features extracted by the PathOrchestra encoder. The features were arranged into spatial grids and passed to Mask2Former, which was trained using standard cross-entropy loss to predict pixel-wise labels. This setup enables effective evaluation of the encoder's adaptability to dense prediction tasks while keeping the encoder fixed. For ROI-level cell detection tasks, we adopted a transformer-based detection head following the ViTDet framework, which operates on frozen patch-level features extracted from each ROI and predicts bounding boxes and class labels. This setup allows accurate localization of cellular instances without modifying the backbone encoder. For region-level clustering tasks, we used a simple KNN approach on the extracted features to evaluate the structural separation of phenotypes in the learned embedding space.

We adopted a range of metrics to evaluate classification, detection, and segmentation tasks. For classification tasks, we reported balanced accuracy (ACC), weighted F1 score, and area under the receiver operating characteristic curve (AUC). Balanced accuracy measures the average recall across classes to address class imbalance. The weighted F1 score integrates precision and recall, weighted by class frequency. AUC reflects the model's capability to discriminate between classes over varying thresholds. For cell detection tasks, we used mean average precision (mAP), which summarizes detection performance across multiple confidence thresholds. For segmentation tasks, we reported class pixel accuracy (CPA), intersection over union (IoU), and Dice coefficient. CPA reflects per-class pixel accuracy. IoU measures the overlap between predicted masks and ground truth. Dice coefficient evaluates overall segmentation agreement. Detailed experimental setups and task-specific results are provided in the Results section.

The ground truth for each clinical task was derived from different sources depending on the task type, including expert annotations, pathology reports, image metadata, and publicly available dataset labels. A comprehensive summary of the ground truth sources and corresponding label formats is provided in Supplementary Table 2. Model performance was quantitatively evaluated by comparing predictions to the corresponding ground truth labels for each task using standard metrics.

Hyperparameters for pretraining

The pretraining of the PathOrchestra model was carried out using the following hyperparameters, optimized for large-scale computational pathology tasks. The model architecture comprises 24 layers, with 16 attention heads and a patch size of 16. The feed-forward network (FFN) utilizes a multi-layer perceptron (MLP) structure, with an embedding dimension of 1024. Stochastic dropout is applied at a rate of 0.1 to prevent overfitting. For data augmentation, global crop scales range from 0.48 to 1.0, with a crop number and size of 2 at 224. Local crop scales range from 0.16 to 0.48, with 8 crops of size 96. The maximum masking ratio is set at 0.5, and the minimum masking ratio is set at 0.1, optimizing the model's robustness. Gradient clipping is applied with a max norm of 3.0, and normalization is

performed on the last layer. The training employed a batch size of 3072 with the learning rate starting at 0 and increasing to $2e-3$ after warmup. The learning rate gradually decays to $1e-6$. The model also utilizes a teacher model with an initial temperature of 0.04, gradually increasing to 0.4, and a teacher momentum that transitions from 0.992 to 1.000. Training was conducted for 20 epochs, with warmup epochs set to 2, and teacher temperature warmup lasting for 6 epochs. The entire process is carried out using automatic mixed precision (FP16) to improve computational efficiency. A detailed list of these hyperparameters is provided in Supplementary Table 3.

Computing hardware and software

All experiments and analyses conducted in this study (unless otherwise specified), we utilized Python (version 3.9) and PyTorch (version 2.0.0, CUDA 11.8) (pytorch.org). To train PathOrchestra via DINOv2, we adapted the Vision Transformer implementation from the Timm library (version 0.9.2) maintained by Hugging Face (huggingface.co) for the encoder backbone, and employed the original DINOv2 self-supervised learning algorithm (github.com/facebookresearch/dinov2) for pretraining. This setup utilized 4×8 80GB NVIDIA A100 GPU nodes configured for multi-GPU, multi-node training with distributed data-parallel (DDP). All additional computations for downstream experiments were conducted on single 24GB NVIDIA 3090 GPUs, 4090 GPUs, and 80GB NVIDIA A100 GPUs. WSI processing was supported by OpenSlide (version 4.3.1) and openslide-python (version 1.2.0). For training weakly-supervised models, we adapted the training scaffold code from the ABMIL codebase (github.com/AMLab-Amsterdam/AttentionDeepMIL). For semantic segmentation training, we used the original Mask2Former implementation (github.com/facebookresearch/Mask2Former). Pillow (version 9.3.0) and OpenCV-python were employed for basic image processing tasks. Matplotlib (version 3.7.1) and Seaborn (version 0.12.2) were used for creating plots and figures.

Data availability

All public datasets used in this study are publicly available from their respective publications, as summarized in Supplementary Table 5. The following datasets were used for downstream evaluation: FocusPath-UofT³¹, CAMELYON16³⁶, CAMELYON17³⁷, TCGA-TILs³⁸, PatchCamelyon (PCam)³⁹, Glas⁴⁰, PanNuke⁴¹, CoNSeP⁴², COSAS⁴³, TissueNet⁴⁴, LC25K⁴⁵, BreakHis⁴⁶, TCGA-NSCLC, TCGA-RCC, TCGA-ESCA, TCGA-IDH1⁴⁷, BACH⁴⁸, HunCRC⁴⁹, PANDA⁵⁰, PatchGastricADC22⁵¹, AGGC⁵², CRC-100K⁵³, Chaoyang⁵⁴, WSSS4LUAD⁵⁵, Kather⁵⁶, Ebrains⁵⁷, HEST¹⁰, DeepCell³², DigestPath⁵⁸, CoNSeG⁵⁹, and SegPath⁶⁰. All datasets were used in accordance with their respective usage licenses and terms. Additionally, pre-extracted image features are accessible at huggingface.co/datasets/AI4Pathology/pathorchestra-image-features.

Code availability

All training pipelines and inference interfaces for PathOrchestra are publicly available on GitHub for non-commercial academic use at github.com/yanfang-research/PathOrchestra. The PIANO framework for data pre-processing, feature extraction, and downstream task evaluation is available at github.com/WonderLandxD/PIANO. Pre-trained model weights are hosted at huggingface.co/AI4Pathology/PathOrchestra.

Received: 24 March 2025; Accepted: 23 September 2025;

Published online: 19 November 2025

References

1. 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).
2. Ding, K. et al. A large-scale synthetic pathological dataset for deep learning-enabled segmentation of breast cancer. *Sci. Data* **10**, 231–241 (2023).

3. Lu, M. Y. et al. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nat. Biomed. Eng.* **5**, 555–570 (2021).
4. Anaya, J., Sidhom, J.-W., Mahmood, F. & Baras, A. S. Multiple-instance learning of somatic mutations for the classification of tumour type and the prediction of microsatellite status. *Nat. Biomed. Eng.* **8**, 57–67 (2024).
5. Nagpal, K. et al. Development and validation of a deep learning algorithm for gleason grading of prostate cancer from biopsy specimens. *JAMA Oncol.* **6**, 1372–1380 (2020).
6. Madabhushi, A., Feldman, M. D. & Leo, P. Deep-learning approaches for gleason grading of prostate biopsies. *Lancet Oncol.* **21**, 187–189 (2020).
7. Shama, G. et al. Deep learning-based image analysis predicts pd-l1 status from h&e-stained histopathology images in breast cancer. *Nat. Commun.* **13**, 6753 (2022).
8. He, B. et al. Integrating spatial gene expression and breast tumour morphology via deep learning. *Nat. Biomed. Eng.* **4**, 827–834 (2020).
9. Kather, J. N. et al. Pan-cancer image-based detection of clinically actionable genetic alterations. *Nat. Cancer* **1**, 789–799 (2020).
10. Jaume, G. et al. Hest-1k: A dataset for spatial transcriptomics and histology image analysis. *Adv. Neural. Inf. Process. Syst.* **37**, 53798–53833 (2024).
11. Wulczyn, E. et al. Interpretable survival prediction for colorectal cancer using deep learning. *NPJ Digit. Med.* **4**, 71–81 (2021).
12. Jiang, X. et al. End-to-end prognostication in colorectal cancer by deep learning: a retrospective, multicentre study. *Lancet Digit. Health* **6**, 33–43 (2024).
13. Foersch, S. et al. Multistain deep learning for prediction of prognosis and therapy response in colorectal cancer. *Nat. Med.* **29**, 430–439 (2023).
14. Yu, K.-H. et al. Predicting non-small cell lung cancer prognosis by fully automated microscopic pathology image features. *Nat. Commun.* **7**, 12474 (2016).
15. Lipkova, J. et al. Deep learning-enabled assessment of cardiac allograft rejection from endomyocardial biopsies. *Nat. Med.* **28**, 575–582 (2022).
16. Oquab, M. et al. Dinov2: Learning robust visual features without supervision. *arXiv* <https://doi.org/10.48550/arXiv.2304.07193> (2023).
17. Chen, X., Xie, S. & He, K. An empirical study of training self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (CVPR)*, 9640–9649 (IEEE, 2021).
18. He, K. et al. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 16000–16009 (IEEE, 2022).
19. Achiam, J. et al. Gpt-4 technical report. *arXiv* <https://doi.org/10.48550/arXiv.2303.08774> (2023).
20. Brown, T. et al. Language models are few-shot learners. *Adv. Neural Inform. Process. Syst.* **33**, 1877–1901 (2020).
21. Xu, H., Usuyama, N. & Bagga, J. et al. A whole-slide foundation model for digital pathology from real-world data. *Nature* **630**, 181–188 (2024).
22. Chen, R. J., Ding, T. & Lu, M. Y. et al. Towards a general-purpose foundation model for computational pathology. *Nat. Med.* **30**, 850–862 (2024).
23. Wang, X., Zhao, J. & Marostica, E. et al. A pathology foundation model for cancer diagnosis and prognosis prediction. *Nature* **634**, 970–978 (2024).
24. Ma, J. et al. A generalizable pathology foundation model using a unified knowledge distillation pretraining framework. *Nat. Biomed. Eng.* 1–20 <https://doi.org/10.1038/s41551-025-01488-4> (2025).
25. Cheng, B., Misra, I., Schwing, A. G., Kirillov, A. & Girdhar, R. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 1290–1299 (IEEE, 2022).
26. Dosovitskiy, A. An image is worth 16 x 16 words: transformers for image recognition at scale. *Internat. Conf. Learn. Repres.* <https://doi.org/10.48550/arXiv.2010.11929> (2021).
27. Caron, M. et al. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 9650–9660 (IEEE, 2021).
28. Zhou, J. et al. ibot: Image bert pre-training with online tokenizer. *arXiv* <https://doi.org/10.48550/arXiv.2111.07832> (2021).
29. Sablayrolles, A., Douze, M., Schmid, C. & Jégou, H. Spreading vectors for similarity search. *Int. Conf. Learn. Represent.* <https://doi.org/10.48550/arXiv.1806.03198> (2019).
30. He, K., Fan, H., Wu, Y., Xie, S. & Girshick, R. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 9729–9738 (IEEE, 2020).
31. Hosseini, M. S., Zhang, Y. & Plataniotis, K. N. Encoding visual sensitivity by maxpool convolution filters for image sharpness assessment. *IEEE Trans. Image Process.* **28**, 4510–4525 (2019).
32. Greenwald, N. F. et al. Whole-cell segmentation of tissue images with human-level performance using large-scale data annotation and deep learning. *Nat. Biotechnol.* **40**, 555–565 (2022).
33. Ilse, M., Tomczak, J. & Welling, M. Attention-based deep multiple instance learning. In *International Conference on Machine Learning (ICML)*, 2127–2136 (PMLR, 2018).
34. Chen, Z. et al. Vision transformer adapter for dense predictions. In *International Conference on Learning Representations (ICLR)* (ICLR, 2023).
35. Li, Y., Mao, H., Girshick, R. & He, K. Exploring plain vision transformer backbones for object detection. In *Proceedings of the IEEE/CVF Conference on European Conference on Computer Vision (ECCV)*, 280–296 (IEEE, 2022).
36. 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).
37. 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. Imag.* **38**, 550–560 (2018).
38. 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).
39. Veeling, B. S., Linmans, J., Winkens, J., Cohen, T. & Welling, M. Rotation equivariant CNNs for digital pathology. *arXiv* <https://doi.org/10.48550/arXiv.1806.03962> (2018).
40. Sirinukunwattana, K. et al. Gland segmentation in colon histology images: the glas challenge contest. *Med. Image Anal.* **35**, 489–502 (2017).
41. Gamper, J. et al. Pannuke dataset extension, insights and baselines. *arXiv preprint arXiv:2003.10778* (2020).
42. Graham, S. et al. Hover-net: Simultaneous segmentation and classification of nuclei in multi-tissue histology images. *Med. Image Anal.* **58**, 101563 (2019).
43. Jiayan, H., Zheng, J., Jinbo, K. & Shuoyu, X. Domain-stratified training for cross-organ and cross-scanner adenocarcinoma segmentation in the cosas 2024 challenge. *arXiv* <https://doi.org/10.48550/arXiv.2409.12418> (2024).
44. Greenwald, N. F. et al. Whole-cell segmentation of tissue images with human-level performance using large-scale data annotation and deep learning. *Nat. Biotechnol.* **40**, 555–565 (2022).
45. Borkowski, A. A., Bui, M. M., Thomas, L. B., Wilson, C. P. & Mastorides, S. M. Lung and colon cancer histopathological image dataset (lc25000). *arXiv* <https://doi.org/10.48550/arXiv.1912.12142> (2019).
46. 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).

47. Tomczak, K., Czerwińska, P. & Wiznerowicz, M. Review the cancer genome atlas (tcga): an immeasurable source of knowledge. *Contemp. Oncol. Współczesna Onkologia* **2015**, 68–77 (2015).
48. Aresta, G. et al. Bach: Grand challenge on breast cancer histology images. *Med. Image. Anal.* **56**, 122–139 (2019).
49. Pataki, B. Á. et al. Huncrc: annotated pathological slides to enhance deep learning applications in colorectal cancer screening. *Sci. Data* **9**, 370 (2022).
50. Bulten, W. et al. Artificial intelligence for diagnosis and gleason grading of prostate cancer: the panda challenge. *Nat. Med.* **28**, 154–163 (2022).
51. Tsuneki, M. & Kanavati, F. Patchgastricadc22. *Zenodo* <https://doi.org/10.5281/zenodo.6021442> (2021).
52. Huo, X. et al. A comprehensive ai model development framework for consistent gleason grading. *Commun. Med.* **4**, 84 (2024).
53. Kather, J. N., Halama, N. & Marx, A. 100,000 histological images of human colorectal cancer and healthy tissue. *Zenodo* <https://doi.org/10.1109/ISBI.2009.5193250> (2018).
54. Zhu, C., Chen, W., Peng, T., Wang, Y. & Jin, M. Hard sample aware noise robust learning for histopathology image classification. *IEEE Trans. Med. Imag.* **41**, 881–894 (2021).
55. Han, C. et al. Wss4luad: Grand challenge on weakly-supervised tissue semantic segmentation for lung adenocarcinoma. *arXiv* <https://doi.org/10.48550/arXiv.2204.06455> (2022).
56. Kather, J. N. et al. Multi-class texture analysis in colorectal cancer histology. *Sci. Rep.* **6**, 27988 (2016).
57. Roetzer-Pejrimovsky, T. et al. The digital brain tumour atlas, an open histopathology resource. *Sci. Data* **9**, 55 (2022).
58. Da, Q. et al. Digestpath: A benchmark dataset with challenge review for the pathological detection and segmentation of digestive-system. *Med. Image. Anal.* **80**, 102485 (2022).
59. Wu, H., Wang, Z., Zhao, Z., Chen, C. & Qin, J. Continual nuclei segmentation via prototype-wise relation distillation and contrastive learning. *IEEE Trans. Med. Imag.* **42**, 3794–3804 (2023).
60. Komura, D. et al. Restaining-based annotation for cancer histology segmentation to overcome annotation-related limitations among pathologists. *Patterns* **4**, 100688 (2023).
61. Peterson, L. E. K-nearest neighbor. *Scholarpedia* **4**, 1883 (2009).

Acknowledgements

This work was supported by Shanghai Artificial Intelligence Laboratory and National Natural Science Foundation of China (Youth Program, No. 62506188). This study was also supported in part by the Centre for Perceptual and Interactive Intelligence (CPII) Ltd., a CUHK-led InnoCentre under the InnoHK initiative of the Innovation and Technology Commission of the Hong Kong SAR Government. Shaoting Zhang is a PI of CPII under the InnoHK.

Author contributions

This paper was primarily written by FY, JFW, JWL, and WW. FY, JWL, WC, JXL, and TS conceived the study and designed the experiments. FY, WC, MDC, YL, QC, TS performed data collection and processing. FY, SYH performed model development. YRC, LDW, ZZG, JNL, HL, HY, YZW, XTL, XNW, ZHW, QH performed experimental analysis and interpreted the results. JFW, WW, MXL, and LM reviewed and revised the manuscript. XFZ, YHH, HC, STZ and ZW supervised the research.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41746-025-02027-w>.

Correspondence and requests for materials should be addressed to Fang Yan, Xiaofan Zhang, Yonghong He, Hao Chen, Shaoting Zhang or Zhe Wang.

Reprints and permissions information is available at <http://www.nature.com/reprints>

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

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025

¹Shanghai Artificial Intelligence Laboratory, Shanghai 200030, China. ²State Key Laboratory of Holistic Integrative Management of Gastrointestinal Cancers, Department of Pathology, School of Basic Medicine and Xijing Hospital, Fourth Military Medical University, Xi'an 710032, China. ³Institute of Biopharmaceutical and Health Engineering, Tsinghua Shenzhen International Graduate School, Tsinghua University, Shenzhen 518055, China. ⁴Department of Pathology, The First Affiliated Hospital of USTC, Division of Life Sciences and Medicine, University of Science and Technology of China, Hefei 230036, China. ⁵Intelligent Pathology Institute, Division of Life Sciences and Medicine, University of Science and Technology of China, Hefei 230036, China. ⁶Multimedia Laboratory, The Chinese University of Hong Kong, Hong Kong 999077, China. ⁷SenseTime Research, Shanghai 200030, China. ⁸Department of Computer Science and Engineering, Department of Chemical and Biological Engineering, Division of Life Science, Hong Kong, University of Science and Technology, Hong Kong, SAR 999077, China. ⁹Shengqiang Technology Co. Ltd., Longgang District, Shenzhen 518055, China. ¹⁰Institute of Materials Research, Shenzhen International Graduate School, Tsinghua University, Shenzhen 518055, China. ¹¹Qing Yuan Research Institute, Shanghai Jiao Tong University, Shanghai 200240, China. ¹²National Biomedical Imaging Center, College of Future Technology, Peking University, Beijing 100871, China. ¹³Centre for Perceptual and Interactive Intelligence (CPII) Ltd., Hong Kong, SAR 999077, China. ¹⁴These authors contributed equally: Fang Yan, Jianfeng Wu, Jiawen Li, and Wei Wang. ✉ e-mail: yanfang@pjlab.org.cn; zhangxiaofan@pjlab.org.cn; heyh@sz.tsinghua.edu.cn; jhc@cse.ust.hk; zhangshaoting@pjlab.org.cn; zhwang@fmmu.edu.cn