

A clinically-oriented foundation model for intraoperative pathology

Received: 8 October 2025

Accepted: 8 September 2026

Accelerated Article Preview

Published online: 10 September 2026

Cite this article as: Zhao, Z. et al. A clinically-oriented foundation model for intraoperative pathology. *Nature Medicine* <https://doi.org/10.1038/s41591-026-04703-0> (2026).

Open access

Zihan Zhao, Fengtao Zhou, Ronggang Li, Bing Chu, Xinke Zhang, Xueyi Zheng, Ke Zheng, Xiaobo Wen, Jiabo Ma, Yihui Wang, Jiewei Chen, Chengyou Zheng, Jiangyu Zhang, Yongqin Wen, Jiajia Meng, Ziqi Zeng, Xiaoqing Li, Jing Li, Dan Xie, Yaping Ye, Yu Wang, Hao Chen & Muyan Cai

This is a PDF file of a peer-reviewed paper that has been accepted for publication. Although unedited, the content has been subjected to preliminary formatting. Nature Medicine is providing this early version of the typeset paper as a service to our authors and readers. The text and figures will undergo copyediting and a proof review before the paper is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers apply.

Editor summary:

CRISP, a vision-based pathology foundation model developed exclusively from frozen section slides, supports treatment decision-making throughout the surgical workflow with superior performance to current foundation models and extensive validation, including in a prospective cohort.

Editor recognition statement:

Primary Handling Editors: Mattia Andreoletti and Saheli Sadanand, in collaboration with the Nature Medicine team.

Peer review information:

Nature Medicine thanks Xuelei Ma, Thomas Roetzer-Pejrimovsky and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

1. Extended Data

Figure or Table #	Figure/Table title	Filename	Figure/Table Legend
Please group Extended Data items by type, in sequential order. Total number of items (Figs. + Tables) must not exceed 10.	One sentence only	Whole original file name including extension. i.e.: Smith_ED_Fig1.jpg	If you are citing a reference for the first time in these legends, please include all new references in the main text Methods References section, and carry on the numbering from the main References section of the paper. If your paper does not have a Methods section, include all new references at the end of the main Reference list.
Extended Data Fig. 1	Extended Performance of CRISP in common	exfig1.pdf	(A) Predictive performance in SYCC cohort. (B) Confusion matrices showing CRISP's predictions for common benign–malignant differentiation tasks. Data are presented as mean $\pm$ 95%

	benign-malignant differentiation tasks.		CI in (A), with cohort sizes for each site indicated in Supplementary Table S2.1.
Extended Data Fig. 2	Extended Evaluation of CRISP for intraoperative decision-making.	exfig2.pdf	(A) Data composition of the SYCC validation cohort for the PCNSLD task (top) and CNSLC task (bottom). (B) Comparison of AUROC between CRISP and other models on SYCC validation cohorts for the PCNSLD task. (C) Top-k accuracy comparison of CRISP and other models in SYCC validation cohorts for the CNSLC task. (D) Data composition of SYCC validation cohorts for the OCCGS, OCFGS, TNDD, and PGD tasks. (E) Predicted probability distributions of CRISP across classes in the OCCGS task in the SYCC validation cohort. (F) Predicted probability distributions of CRISP across classes in the TNDD task in the SYCC validation cohort. (G) Top-k accuracy comparison of CRISP and other models in the SYCC validation cohort for the OCFGS task. (H) Specificity comparison of CRISP and other models at 95% sensitivity for the PGD task in the SYCC validation cohort. Error bars denote 95% CI. For box plots in (E) and (F), center line represents median; box bounds represent 25th and 75th percentiles; whiskers extend to the non-outlier range; outliers are plotted individually. Data are presented as mean $\pm$ 95% CI in (B), (C), (G), and (H). Statistical significance was assessed using one-way ANOVA with Tukey's HSD test for pairwise comparisons (two-sided) in (E) and (F); *$P \leq 0.05$, **$P \leq$

				0.01, $^{***}P \leq 0.001$. Exact cohort sizes for each task are provided in Supplementary Table S3.1.
Extended Data Fig. 3	Confusion matrices showing CRISP's predictions for preoperative and intraoperative decision-making tasks.	exfig3.pdf		Confusion matrices for CRISP predictions across six decision-making tasks: PCNSLD, CNSLC, OCCGS, OCFGs, TNDD, and PGD.
Extended Data Fig. 4	Confusion matrices showing CRISP's predictions in late intraoperative assessment.	exfig4.pdf		Confusion matrices for CRISP predictions across three late intraoperative decision-making tasks: SLMD, CLMD and BSMA.
Extended Data Fig. 5	Detailed comparison of ROC curves between CRISP and pathologists in the reader study.	exfig5.pdf		(A-F) ROC curves of CRISP compared with diagnostic performance of three pathologists for the PNBm (A), TNBM (B), BNBm (C), CLMD (D), SLMD (E), and BSMA (F) tasks. (G) ROC curve on the sub-cohort of 214 cases with inter-observer disagreement among three pathologists. Pathologist 1 (P1) represents a senior pathologist with 10 years of clinical experience; Pathologist 2 (P2) and Pathologist 3 (P3) represent junior pathologists with 2 years of experience each.

Extended Data Fig. 6	Waterfall plots of CRISP prediction scores across subtasks in the SYCC prospective cohort.	exfig6.pdf	(A–F) Waterfall plots illustrating the distribution of CRISP-predicted scores for individual cases in the PNBm (A), TNBM (B), BNBm (C), CLMD (D), SLMD (E), and BSMA (F) tasks.
Extended Data Fig. 7	Confusion matrices of CRISP predictions in the IHC-required subgroup within the SYCC prospective cohort.	exfig7.pdf	(A–F) Confusion matrices illustrating the diagnostic performance of CRISP for the PNBm (A), TNBM (B), BNBm (C), CLMD (D), SLMD (E), and BSMA (F) tasks.
Extended Data Fig. 8	Waterfall plots of CRISP screening scores in external and simulated deployment cohorts.	exfig8.pdf	(A–C) Waterfall plots illustrating the distribution of CRISP-predicted scores in the FUSCC (A), TCIH (B), and clinical simulated deployment (C) cohorts. Each bar represents a single case. The screening thresholds for all cohorts were maintained consistent with those established for the SYCC cohort.
Extended Data Fig. 9	Comparison of diagnostic turnaround times between CRISP and pathologists in the clinical simulated deployment cohort.	exfig9.pdf	Box plot illustrating the diagnostic time for CRISP versus human pathologists. The CRISP diagnostic time encompasses the entire end-to-end workflow, from slide scanning to the generation of final predictive results. For each box, center line represents median; box bounds represent 25th and 75th percentiles; whiskers extend to the non-outlier range; outliers are plotted as individual

Extended Data Fig. 10	Comparison of diagnostic certainty scores before and after CRISP assistance in the clinical simulated deployment cohort.	exfig10.pdf	points beyond the whiskers. The cohort size is $n = 100$ cases for each group. (A) Changes in diagnostic certainty for the subgroup of typical cases (highly confident predictions by CRISP). (B) Changes in diagnostic certainty for the subgroup of uncertain cases (low-confidence predictions by CRISP). In both panels, diagnostic certainty is scored on a scale of 0 to 3, representing a gradient from "uncertain" to "completely certain." Each pair of connected points represents an individual case, illustrating the shift in an individual pathologist's confidence following CRISP integration. Black diamonds and dashed lines denote the mean certainty scores for each group.
-----------------------	--	-------------	--

14

15 **1. Supplementary Information:**

16 **A. PDF Files**

Item	Present?	Filename	A brief, numerical description of file contents.
		Whole original file name including extension. i.e.:	i.e.: <i>Supplementary Figures 1-4, Supplementary Discussion, and Supplementary Tables 1-4.</i>

		Smith_SI.pdf. The extension must be .pdf		
Supplementary Information	Yes	supplementary_info.pdf		Supplementary Figures 1-7 and Supplementary Tables 1s-6s.
Reporting Summary	Yes	NMED-A146190C Chen RS.pdf		
Peer Review Information	Yes	NMED-A146190C Chen Peer Review.pdf		

A clinically-oriented foundation model for intraoperative pathology

Zihan Zhao^{1,2†}, Fengtao Zhou^{3†}, Ronggang Li^{4†}, Bing Chu⁵, Xinke Zhang¹, Xueyi Zheng¹, Ke Zheng¹, Xiaobo Wen¹, Jiabo Ma³, Yihui Wang³, Jiewei Chen¹, Chengyou Zheng¹, Jiangyu Zhang⁶, Yongqin Wen⁷, Jiajia Meng⁸, Ziqi Zeng⁹, Xiaoqing Li⁴, Jing Li¹⁰, Dan Xie¹, Yaping Ye¹¹, Yu Wang⁸, Hao Chen^{3,12,13,14,15,16*}, Muyan Cai^{1*}

1 Department of Pathology, State Key Laboratory of Oncology in South China, Guangdong Provincial Clinical Research Center for Cancer, Sun Yat-sen University Cancer Center, Guangzhou, China.

2 Department of Gastric Surgery, State Key Laboratory of Oncology in South China, Guangdong Provincial Clinical Research Center for Cancer, Sun Yat-sen University Cancer Center, Guangzhou, China.

3 Department of Computer Science and Engineering, Hong Kong University of Science and Technology, Hong Kong, China.

4 Department of Pathology, Jiangmen Central Hospital, Jiangmen, China.

5 Department of Pathology, Zhongshan People's Hospital, Zhongshan, China.

6 Department of Pathology, Affiliated Cancer Hospital & Institute of Guangzhou Medical University, Guangzhou, China.

7 Department of Pathology, The Tenth Affiliated Hospital, Southern Medical University, Dongguan People's Hospital, Dongguan, China.

8 Department of Pathology, Huizhou Central People's Hospital, Huizhou, China.

9 Department of Pathology, Zhujiang Hospital, Southern Medical University, Guangzhou, China.

10 Zhongshan School of Medicine, Sun Yat-sen University, Guangzhou, China.

11 Department of Pathology, Nanfang Hospital, Southern Medical University, Guangzhou, China.

12 Department of Chemical and Biological Engineering, Hong Kong University of Science and Technology, Hong Kong, China.

13 Division of Life Science, Hong Kong University of Science and Technology, Hong Kong, China.

14 HKUST Shenzhen-Hong Kong Collaborative Innovation Research Institute, Futian, Shenzhen, China.

15 State Key Laboratory of Nervous System Disorders, The Hong Kong University of Science and Technology, Hong Kong, China.

16 SIAT-HKUST Joint Laboratory for Brain Science, The Hong Kong University of Science and Technology, Hong Kong, China

†Contributed equally.

*Dr. Muyan Cai and Prof. Hao Chen are the co-corresponding authors.

Correspondence: Dr. Muyan Cai, Department of Pathology, Collaborative Innovation Center for Cancer Medicine, State Key Laboratory of Oncology in South China, Sun Yat-sen University Cancer Center, Guangzhou, 510060, China. Email: caimy@sysucc.org.cn;

Prof. Hao Chen, Room 3524, Academic Building, Department of Computer Science and Engineering, Hong Kong University of Science and Technology, Hong Kong, China. Email: jhc@ust.hk.

Abstract

Intraoperative pathology is pivotal to precision surgery, yet its clinical impact is constrained by diagnostic complexity and the limited availability of high-quality frozen-section data. While computational pathology has made significant strides, the lack of large-scale, prospective validation has impeded its routine adoption in surgical workflows. Here, we introduce CRISP, a clinically oriented foundation model developed on over 100,000 frozen sections from ten medical centers, specifically designed to provide Clinically-oriented Robust Intraoperative Support for Pathology (CRISP). CRISP was comprehensively evaluated on more than 15,000 intraoperative slides across nearly 100 retrospective diagnostic tasks, including benign-malignant discrimination, key intraoperative decision-making, and pan-cancer detection, etc. The model demonstrated robust generalization across 6 institutions, 14 tumor types, and 24 anatomical sites-including previously unseen sites and rare cancers. In a prospective cohort of over 3,000 patients, CRISP sustained high diagnostic accuracy under real-world conditions, directly informing surgical decisions in 92.6% of cases. Human-AI collaboration further reduced diagnostic workload by 35%, avoided 105 ancillary tests and enhanced detection of micrometastases with 87.5% accuracy. Together, these findings suggest that CRISP represents a clinically oriented approach to AI-driven intraoperative pathology, with the potential to support surgical decision-making and facilitate the translation of computational methods into clinical practice.

Introduction

Rapid and accurate intraoperative diagnosis is a cornerstone of modern surgical practice, directly influencing patient outcomes and guiding critical decisions in real time.^{1,2} For over a century, frozen section examination has served as an indispensable modality for intraoperative pathological assessment, enabling surgeons to make informed choices regarding tumor resection, margin status, and differential diagnosis across a wide spectrum of surgical specialties.³⁻⁶ The widespread adoption of frozen sections is driven by their expedited preparation and diagnostic accuracy, which closely parallels that of formalin-fixed paraffin-embedded (FFPE) specimens in clinical practice.^{7,8} As a result, frozen section analysis has become integral to precision surgical management, supporting timely interventions that can profoundly improve patient outcomes.

Despite its clinical utility, frozen section diagnosis presents formidable challenges. The inherent morphological complexity of frozen tissue, compounded by the urgency of intraoperative decision-making, demands rapid and highly accurate interpretation. These constraints introduce significant risks, including the potential for misdiagnosis and suboptimal surgical outcomes.^{9,10} The need for robust, scalable support systems to assist pathologists and surgeons in real-time intraoperative workflows is therefore both pressing and universal.

Recent advances in computational pathology and artificial intelligence (AI) have begun to reshape the landscape of diagnostic medicine. Deep learning models have demonstrated remarkable promise in automating and augmenting pathological assessment, offering the potential to enhance diagnostic consistency and efficiency. However, the application of AI to intraoperative frozen section analysis remains in its infancy.¹¹ Prior studies have explored deep learning approaches in specific disease contexts,¹²⁻¹⁴ Yet these efforts are frequently constrained by small, homogeneous datasets that fail to capture the full spectrum of morphological and technical variability encountered in clinical practice.¹⁵ As a result, the development of generalizable and clinically practical AI models for intraoperative diagnosis has remained elusive.

The emergence of large-scale AI models, or foundation models, has further fueled optimism for the clinical translation of computational pathology.¹⁵⁻¹⁸ Pretrained on vast and diverse datasets, these models have shown strong generalizability across a range of diagnostic tasks.¹⁹ However,

most existing foundation models are built upon FFPE specimens,²⁰ which differ fundamentally from frozen sections in both preparation and tissue morphology. Direct application of FFPE-trained models to intraoperative frozen section risks compromising diagnostic accuracy as such models may lack exposure to frozen section morphology,^{10,21} while attempts to bridge this gap using virtual staining techniques raise concerns regarding reliability and clinical trustworthiness of synthetic images.²²⁻²⁴ Additionally, the lack of prospective validation in most studies limits their readiness for real-world deployment.²⁵

Addressing these challenges requires the development of a robust and generalizable foundation model specifically tailored to frozen section pathology, with demonstrated clinical validity in diverse and realistic settings. In this study, we introduce CRISP, an AI-enabled comprehensive deep learning framework designed to support a wide spectrum of intraoperative diagnostic and predictive tasks. Leveraging the largest frozen section dataset assembled to date-comprising over 100,000 slides from ten medical centers across different regions-CRISP was pretrained on 50 million image patches from 25 anatomical sites. The model's performance was rigorously evaluated across multiple downstream tasks commonly encountered in intraoperative pathology, with external validation cohorts drawn from independent institutions. To further assess its real-world clinical utility, we conducted a registered prospective observational study involving more than 3,000 patients in real-world surgical workflows. Through both rigorous retrospective and prospective analyses, our findings demonstrate the practical value and broad applicability of a frozen-specific foundation model, making a significant advance toward the integration of AI-driven decision support in precision surgery.

Results

Patient cohorts and study design

The development of a clinically robust foundation model for intraoperative diagnosis necessitates a comprehensive understanding of the biological and technical heterogeneity inherent in real-world surgical pathology.²⁶ To address this, we constructed the largest multi-center frozen section dataset to date, comprising 103,797 frozen section slides obtained primarily from ten medical centers and spanning a broad spectrum of anatomical sites (Fig. 1a). This dataset was specifically curated to reflect the diversity of clinical scenarios and laboratory

conditions encountered during intraoperative consultations, as detailed in the Methods.

For model pre-training, we utilized 81,667 frozen section slides from three major centers and the TCGA frozen section dataset.²⁷ To rigorously assess the model's robustness and generalizability, we established 98 downstream diagnostic tasks designed to emulate critical intraoperative decision-making scenarios. Validation was conducted across multiple centers, including pan-cancer analyses, rare disease subtypes, and challenging conditions (Fig. 1b-e).

To evaluate cross-modality generalizability, we further constructed a paired dataset comprising frozen section and corresponding FFPE slides from the same 685 patients across GZH, HZH, and DGH cohorts, with one-to-one matching to minimize sampling bias (Supplementary Fig. S1).

To further bridge the gap between proof-of-concept and clinical translation, we conducted a large-scale prospective study (Fig. 1f). This study enrolled 3,064 consecutive patients with intraoperative consultation at Sun Yat-sen University Cancer Center, Fudan University Shanghai Cancer Center (FUSCC) and Tianjin Medical University Cancer Institute and Hospital (TCIH) between November 11, 2024, and July 1, 2025. Based on the tissue site, patients from SYCC were assigned to three prospective cohorts: SYCC-THYR (n=1,330), SYCC-BREA (n=391), and SYCC-LUNG (n=350) (Supplementary Table S1s). To assess the practical feasibility of CRISP within clinical workflows, we conducted a simulated deployment analysis using an independent set of 100 cases underwent intraoperative consultation, which were not part of any training or evaluation cohorts (Supplementary Fig. S2).

Classification of common benign and malignant lesions

To evaluate CRISP's performance in this critical context, we validated the model using datasets from five independent centers (Fig. 2a,b), spanning six common anatomical sites (thyroid, lung, breast, brain, ovary, and head and neck), as well as complex abdominal lesion differentiation tasks (Supplementary Table S2.1).

CRISP was benchmarked against several state-of-the-art pathology foundation models^{15-17,28,29} across 21 downstream diagnostic tasks. In malignancy discrimination, CRISP consistently achieved superior performance, with an average AUROC of 0.970, an average accuracy of 0.945, and an average F1 score of 0.924-significantly surpassing the next-best model (Fig. 2c; $P <$

0.001, effect size $r = 0.62$). Notably, CRISP was the only model to achieve a median F1 score above 0.9 (0.902; Fig. 2d).

CRISP achieved consistently higher performance across the 21 downstream tasks, with modest but consistent improvements over other approaches (Fig. 2e, Extended Data Fig. 1a). In the subset of 14 tasks derived from previously unseen institutions, CRISP was the only model to achieve an average AUC above 0.95 (0.967 ± 0.018), maintaining AUC values greater than 0.9 across all tasks. Even in particularly challenging datasets—such as NFH-Abdomen and ZSH-Abdomen—or in heterogeneous multi-institutional data (ZHDG-Breast, ZHG-H&N, ZSGZ-Brain), CRISP sustained an average AUC of 0.957, outperforming Virchow2 by 3.5% ($P < 0.001$) (Supplementary Tables S2.2-2.9).

To further evaluate diagnostic sensitivity, we analyzed CRISP's performance in detecting malignant lesions across 1,729 cancer cases spanning 21 downstream tasks. The model correctly identified 1,634 cases, yielding an overall sensitivity of 94.5% (Extended Data Fig. 1b). At sensitivity thresholds of 90% and 95%, CRISP achieved median specificities of 0.917 and 0.790, respectively, again significantly outperforming the second-best model (+8.0% and +7.7% over Virchow2, $P = 0.004$ and $P = 0.013$; Fig. 2f). Comparative experiments showed that slide-level and case-level aggregation achieved similar performance across cohorts, with no consistent advantage observed for either strategy (Supplementary Table S2.10).

Visualization analyses further demonstrated CRISP's ability to accurately localize malignant regions within frozen tissue sections, supporting its capacity for sensitive and precise cancer detection across a wide array of anatomical sites, institutions, and clinical contexts (Fig. 2g, Supplementary Fig S3).

Model evaluation in key intraoperative decision-making tasks

To rigorously evaluate CRISP's generalizability and clinical utility, we established 20 multi-center downstream tasks spanning the entire surgical continuum: early planning, mid-procedure guidance, and late intraoperative assessment (Fig. 3a, Supplementary Table S3.1). These tasks were designed to reflect the diversity and complexity of real-world clinical decisions, with detailed clinical significance described in the Methods. Across both binary and multiclass tasks, CRISP consistently achieved the highest performance among all evaluated models

(Supplementary Table S3.2-3.3).

Early intraoperative decision: In both the SYCC and NFH cohorts (Fig. 3b, Extended Data Fig. 2a), CRISP achieved AUCs of 0.998 and 0.987, respectively-representing numerical improvements over the second-performing comparator (both $P < 0.001$; Fig. 3b-3c, Extended Data Fig. 2b, Supplementary Table S3.4). Notably, CRISP correctly identified 64 of 66 patients with PCNSL, whereas alternative models missed up to 2.5 times as many cases, underscoring the clinical impact of enhanced diagnostic precision (Fig. 3d). In fine-grained tumor classification, CRISP demonstrated a pronounced 20.4% gain in top-1 accuracy in the external cohort (Fig. 3e, Extended Data Fig. 2c, Supplementary Table S3.5), highlighting its potential to support nuanced early intraoperative decision-making.

Mid-procedure decision: CRISP was evaluated across eight cohorts spanning ovarian (Ovarian Cancer Coarse/Fine-Grained Subtyping, OCCGS/OCFGS), thyroid (Thyroid Nodule Differential Diagnosis, TNDD), and parathyroid tissue (Parathyroid Gland Detection, PGD), consistently achieving the highest performance in all tasks, with a mean macro-AUC of 0.954, compared to 0.934 for the second-best baseline ($P < 0.001$; Fig. 3f-3g, Extended Data Fig. 2d, Supplementary Table S3.6-3.9). Confusion matrix analyses underscored CRISP's robust discrimination between distinct tissue types (Extended Data Fig. 3). For ovarian and thyroid tumors, where accurate subtype classification directly informs surgical strategy, CRISP showed the highest top-k ($k = 1, 2, 3$) accuracies in external cohorts, with top-1 accuracy improving by 13.6% over CONCH (Fig. 3h-j). In PGD, a particularly challenging task, CRISP was the only model to maintain AUCs consistently above 0.9. In external validation, it reached a specificity of 0.559 at 95% sensitivity, outperforming all other models and demonstrating potential to mitigate inadvertent parathyroid resection (Fig. 3k). Similar results were observed in internal cohorts (Extended Data Fig. 2e-h), further supporting CRISP's robustness in fine-grained intraoperative decision-making.

Late intraoperative assessment: CRISP's performance was evaluated across eight cohorts focused on sentinel and cervical lymph node metastasis detection (SLMD, CLMD) and breast cancer surgical margin assessment (BSMA) (Fig. 4a).³⁰⁻³² The model achieved mean AUCs of 0.984, 0.992, and 0.992 across these tasks, outperforming all other evaluated models (Fig. 4b, Supplementary Table S3.10-12). Under high-sensitivity conditions, CRISP demonstrated

remarkable specificity. Among 263 positive patients, CRISP correctly identified 254 (96.6%), representing a 40% reduction in missed diagnoses compared to the second-best model (Fig. 4c, Extended Data Fig. 4). Visualization analyses further confirmed its ability to detect small clusters of residual tumor cells (Fig. 4d, Supplementary Fig. S4), supporting its utility in postoperative assessment.

Across the complete set of 20 downstream tasks, CRISP consistently outperformed all comparative models, which exhibited variable and dispersed performance (Fig. 4e). These findings collectively validate CRISP's generalizability and robust predictive power throughout the entire surgical decision-making process, establishing it as a highly effective tool for supporting precision intraoperative pathology in diverse clinical settings.

Robustness under clinical heterogeneity

A defining measure of clinically-oriented foundation model is their capacity to generalize under data-scarce conditions, particularly when confronted with unobserved samples and rare cancer types-scenarios that frequently challenge conventional diagnostic approaches.^{15,33} To rigorously evaluate this aspect, we conducted a large-scale pan-cancer detection study encompassing 13,851 intraoperative frozen slides from six independent centers (Supplementary Table S4.1). The CRISP model was deployed to extract patch embeddings from these slides, with the downstream aggregator trained exclusively on 1,573 slides from the SYCC cohort, representing six common anatomical sites: lung, breast, thyroid, head and neck, brain, and ovary (Supplementary Table S4.2).

Generalizability was systematically assessed across three external cohorts, which included both six observed and more than eleven previously unseen cancer types (Fig. 5a, b; Supplementary Table S4.3-4.5). Across these diverse datasets, CRISP consistently achieved high diagnostic accuracy, with AUCs of 0.953, 0.973 and 0.960-each exceeding the 0.95 threshold and outperforming all comparators (Fig. 5c, Supplementary Table S4.6). Notably, CRISP showed pronounced advantages in the ZSH and HDGZ cohorts, with AUC improvements of 7.7% and 6.0% over the next-best model ($P < 0.001$), respectively, underscoring its robust cross-institutional generalizability. To assess the impact of adaptation strategies, we additionally compared LoRA, full fine-tuning, IA³, and training from scratch under the same experimental

settings (Supplementary Table S4.7). The results indicated that full fine-tuning did not outperform LoRA and in some cases showed inferior generalization across cohorts, while LoRA consistently achieved the best overall performance.

To further dissect model performance, we analyzed results across 17 site-specific subgroups, encompassing both observed and unseen anatomical subtypes. CRISP consistently delivered the highest AUCs, ranging from 0.914 to 1.0 (Fig. 5d). While all models exhibited some decline in performance on unseen subgroups, CRISP maintained a distinct relative advantage, outperforming the next-best model by 10.4% in lymph nodes, 9.1% in cervix, 9.7% in bile duct, and 9.2% in pancreas. This suggests enhanced recognition and adaptability to unfamiliar cancer morphologies. Across 51 subgroup analyses from three unseen cohorts, CRISP achieved the top AUC in 47 and ranked within the top two in all, whereas Virchow2 led in only 16 unseen subgroups (Fig. 5e; Supplementary Table S4.8-4.24), highlighting CRISP's superior adaptability to novel clinical scenarios.

CRISP's performance in rare tumors was equally compelling (Fig. 5f). According to the National Cancer Institute, rare cancers are defined as those with an annual incidence below 40,000 cases in the United States.^{34,35} Across seven rare tumor types, CRISP achieved an overall AUC of 0.983, with cohort AUCs ranging from 0.963 to 0.993, consistently outperforming Virchow2. Furthermore, CRISP exhibited remarkable stability, maintaining AUCs above 0.9 in nearly all categories. In the ZSH-Else cohort, where all models performed less optimally (< 0.9), the results likely reflect the complexity and heterogeneity of tissue sources.

We further investigated the impact of unseen and rare samples on model performance (Fig. 5g, h). CRISP's diagnostic accuracy was minimally affected by either factor. Specifically, AUROC improved by 0.023, 0.017, and 0.003 for rare versus common tumor types, while decreasing by 0.025, 0.022, and 0.036 for unseen versus seen tumor types. Even when analysis was restricted to rare tumor subgroups from previously unseen anatomical sites, CRISP maintained an AUC exceeding 0.95. Under high-sensitivity conditions, CRISP also achieved markedly higher median specificity than the second-best model across both rare and unseen subgroups (Fig. 5i, j). Comprehensive performance metrics are detailed in Supplementary Tables S4.25-4.29. To evaluate cross-modality generalizability, we tested CRISP on matched FFPE slides from the HDGZ cohort. Without additional fine-tuning, CRISP achieved an AUC of 0.961 on frozen

sections and 0.972 on FFPE slides, indicating robust performance across different histological preparation modalities (Supplementary Table S4.30). Across 98 retrospective evaluations, CRISP achieved the highest AUROC in 92 tasks (93.9%), indicating a consistent performance advantage across diverse settings (Supplementary Table S4.31). At fixed sensitivities of 90% and 95%, CRISP consistently yielded higher specificities than competing foundation models, with relative improvements reaching up to 63.4% in representative tasks (Supplementary Tables S2.2, S3.2 and S4.6). Collectively, these findings demonstrate CRISP's robust adaptability to complex and heterogeneous clinical data sources, including data-scarce scenarios and rare cancer types. This level of generalization is essential for real-world deployment, ensuring reliable diagnostic support across the full spectrum of intraoperative pathology.

Prospective validation of clinical utility

Prospective validation in real-world clinical environments is the definitive benchmark for assessing the practical utility of diagnostic AI models.³⁶ Building on retrospective findings, we conducted a large-scale prospective observational study at SYCC, FUSCC and TCIH to rigorously evaluate CRISP's ability to support intraoperative decision-making under authentic surgical conditions. A total of 3,064 consecutive patients were enrolled across multiple clinically relevant frozen-section scenarios and anatomical sites. The prospective evaluation included benign-malignant discrimination of pulmonary, thyroid, and breast nodules (PNBM, TNBM, BNBM), cervical and sentinel lymph node metastasis detection (CLMD, SLMD), and breast surgical margin assessment (BSMA), together with additional multi-organ frozen-section cases from head and neck, ovary, and brain tissues. Importantly, the model's aggregators-trained exclusively on retrospective data-were deployed directly, without task-specific fine-tuning, providing an unbiased assessment of CRISP's robustness and translational potential (Fig. 6a).

Crucially, CRISP demonstrated consistent superiority over competing foundation models across independent prospective external cohorts at both FUSCC (AUROC = 0.966) and TCIH (AUROC = 0.984), confirming its robustness against site bias (Supplementary Tables S5.1-2). Across all six real-world tasks combined with the SYCC cohort, CRISP achieved an overall AUC of 0.975 and outperforming all comparator models in every task (Fig. 6b, Supplementary

Table S5.3). Compared to Virchow2, the second-best baseline, CRISP delivered AUC improvements ranging from 3.2% to 5.9% and elevated the specificity at 95% sensitivity from 0.495 to 0.809 (+63.4%), underscoring its clinical relevance in high-sensitivity scenarios. Similar performance trends were observed for sensitivity- and specificity-related metrics across external cohorts.

In linear probing experiments, CRISP showed improved performance compared to competing models, with a similar trend to aggregation-based evaluations (Supplementary Table S5.4).

CRISP maintained stable performance across varying perturbation settings, with only limited degradation observed under increasing levels of image corruption (Supplementary Table S5.5).

In a reader study conducted on the prospective SYCC cohort, the operating point of the senior pathologist fell on the ROC curve of CRISP, indicating comparable overall performance under the evaluated conditions (Fig. 6c, Extended Data Fig. 5a-f). In contrast, the performance of two junior pathologists fell below the ROC curve of CRISP, suggesting lower diagnostic accuracy under the same evaluation setting. Similar diagnostic trends were observed in the sub-analysis of 214 inter-pathologist discordant cases (Extended Data Fig. 5g). To further test resilience to laboratory variability, the cohort was stratified into 27 subgroups based on histotechnologists (Fig. 6d, Supplementary Table S5.6). CRISP maintained robust generalization, with 21 subgroups exceeding the overall AUC, 7 reaching perfect accuracy (AUC = 1.0), and all others—except the multi-doctor “Else” subgroup—remaining above 0.95. Representative histological examples and detailed subgroup compositions are provided in the Supplementary Fig. S5 and Supplementary Table S5.7. These findings highlight CRISP’s ability to sustain high diagnostic precision across diverse, prospective, and heterogeneous clinical conditions. Extended evaluation incorporating additional recently proposed models (e.g., ProvGigaPath) within the same benchmarking framework further confirmed the model's efficacy (Supplementary Tables S2.2, 3.2–3.3, 4.6, 5.1, and 5.3).³⁷

We further evaluated CRISP’s predictive accuracy in clinically challenging cases, defined by high diagnostic uncertainty in intraoperative frozen reports (n = 371). CRISP achieved an overall AUC of 0.936 in this subset (Fig. 6e, Supplementary Table S5.8), with performance largely maintained in PNB, BNB, and CLMD tasks. Modest decreases were observed in the TNB (−0.031), SLMD (−0.041), and BSMA (−0.047) tasks, yet score distributions

remained well separated between benign and malignant cases (Fig. 6f). In diagnostically elusive cases, such as sentinel lymph node micrometastases, CRISP correctly identified 7 of 8 micrometastatic and 40 of 41 macrometastatic cases (Fig. 6g), with predicted values showing clear separation from negative controls (Fig. 6h). Visualization analyses revealed that CRISP focused attention on peripheral zones-regions prone to metastatic involvement-and precisely highlighted minute malignant areas (Fig. 6i, Supplementary Fig. S6). Error analysis showed that false-positive predictions were mainly associated with rare benign tumors and frozen section artifacts, while false-negative cases were primarily observed in rare malignant tumors with ambiguous morphology (Supplementary Fig. S7).

To explore CRISP's potential to alleviate clinical workload, we examined its utility in triaging cases requiring immunohistochemistry (IHC), a common but resource-intensive diagnostic adjunct. Among 510 patients requiring additional IHC evaluation, CRISP achieved an overall AUC of 0.963, with task-specific AUCs ranging from 0.943 to 1.000 (Fig. 6j, Supplementary Table S5.9). At 95% sensitivity, CRISP correctly identified 74.5% of IHC-negative patients, underscoring its value as a screening tool to reduce unnecessary testing (Fig. 6k). We further assessed whether CRISP could safely triage typical cases without compromising pathologists' diagnostic accuracy.³⁸ When patients with predicted scores > 0.9956 or < 0.0044 were managed directly by CRISP, the combined CRISP-pathologist workflow reduced diagnostic workload by 34.9%, with 79.0% of positive cases included (Fig. 6l). Under this strategy, sensitivity and specificity remained comparable to conventional frozen-section reports (sensitivity, $P = 0.60$; specificity, $P = 1.00$; Fig. 6m), highlighting CRISP's potential as a clinical prescreening tool for streamlining the diagnosis of typical positive cases. The workload reduction and corresponding decrease in immunohistochemistry usage were further analyzed separately across all subtasks and multi-region prospective cohorts (Extended Data Fig. 6,7,8a-b). In a simulated deployment experiment conducted on 100 cases from SYCC, the average end-to-end processing time for CRISP was 196.4 seconds per case, compared to 292.6 seconds for pathologists (Extended Data Fig. 9). The pre-screening strategy identified 65 typical cases, and this approach did not result in any degradation in sensitivity or specificity (Extended Data Fig. 8c). Notably, CRISP assistance increased pathologists' diagnostic confidence by 6.0% in typical cases and 42.9% in diagnostically uncertain cases, suggesting its potential to support decision-

making particularly in challenging scenarios (Extended Data Fig. 10).

Discussion

Foundation models have redefined the landscape of artificial intelligence, but their translation into high-stakes clinical environments has remained elusive.^{39,40} Nowhere is this gap more consequential than in surgery, where more than 310 million major surgical procedures are performed worldwide each year, and intraoperative decisions must balance speed with diagnostic precision.^{41,42} The consequences of diagnostic uncertainty are profound, leading to unnecessary reoperations, overtreatment, or undertreatment, and imposing substantial economic burdens that exceed one billion dollars annually in the United States alone.⁴³ Against this backdrop, the development of a clinically reliable AI platform capable of guiding intraoperative decisions represents a paradigm shift-not only for surgical pathology, but also for the broader clinical adoption of foundation models in medicine.

In this study, we present CRISP, the first clinically oriented foundation model specifically dedicated to intraoperative pathology. Distinct from existing pathological foundation models trained on FFPE specimens, CRISP leverages large-scale frozen section data for pre-training. This strategic focus directly addresses a critical clinical gap: FFPE-trained models often falter in intraoperative settings due to frozen section-specific artifacts such as ice crystal formation and tissue compression. By constructing a large and diverse dataset sourced from ten independent medical centers and reserving data from seven centers exclusively for external validation, we ensured rigorous assessment of CRISP's diagnostic performance under authentic, heterogeneous clinical conditions. The inclusion of public frozen section data increases morphological diversity during pretraining, with potential to improve robustness to heterogeneous data. Furthermore, prospective real-time deployment provided a robust test of the model's reliability in actual surgical workflows.

CRISP's robust performance across a spectrum of intraoperative diagnostic tasks highlights its potential as a generalizable paradigm for frozen section-based decision support. Beyond achieving superior accuracy in benchmark comparisons, CRISP demonstrated the ability to address clinically critical scenarios-such as the detection of subtle malignant foci in surgical margins or lymph nodes-where diagnostic sensitivity is paramount to avoid undertreatment.^{44,45}

Importantly, CRISP maintained consistent advantages across the entire surgical workflow, from early intraoperative differential diagnosis to mid-procedure tumor classification and late assessment of resection completeness. CRISP demonstrates stable performance across FFPE data and common perturbations, indicating robustness to variations in staining and image quality encountered in clinical practice. Furthermore, CRISP's transferability to routine FFPE slides shows it captures cross-modality morphological invariants, offering potential to unify intraoperative guidance and post-operative pathology under a single framework. Given that misinterpretation of intraoperative frozen sections can lead to inappropriate surgical decisions and increased patient risk,^{46,47} CRISP's breadth of applicability suggests it could substantially reduce variability and time constraints, offering a scalable and clinically meaningful complement to routine pathology practice. While absolute improvements in AUROC are modest in high-performance regimes, such differences are expected due to ceiling effects and should be interpreted in the context of already strong baselines. Moreover, the consistent performance advantage across diverse organs, institutions, and downstream tasks suggests that the observed improvements are unlikely to arise from optimization for individual benchmarks alone, but instead reflect enhanced generalization of the pretrained representation. Importantly, these advantages were further amplified under clinically relevant high-sensitivity operating conditions. CRISP operates on standard frozen section slides and supports multi-cancer scenarios, enabling broader generalizability and easier integration into routine clinical workflows compared with prior approaches.⁴⁸

A key strength of CRISP lies in its capacity to learn generalized morphological features, enabling robust generalization even under previously unseen laboratory conditions.⁴⁹ The model maintained high predictive accuracy for tissue types with scarce or absent training samples—a crucial benchmark for foundation model generalizability. When applied to previously unseen institutional data and novel samples, CRISP-based pan-cancer detection models demonstrated stable and reliable performance, outperforming alternative models by up to 11% on challenging datasets. This provides compelling evidence that CRISP has effectively internalized universal tissue morphology from large-scale frozen sections, equipping it to navigate the complexity of diverse clinical scenarios. Notably, CRISP also excelled in the diagnosis of rare diseases, addressing a common challenge in clinical practice where labeled

data are limited and expertise may be lacking.⁵⁰⁻⁵² Such capabilities position CRISP as a valuable tool for resource-limited settings, potentially mitigating disparities in healthcare access and supporting expert-level pathology where it is most needed.⁵³

Prospective evaluation further reinforced CRISP's feasibility as a clinically oriented tool. When deployed in real-time intraoperative settings, CRISP demonstrated robust performance across diverse patient cohorts and laboratory conditions, bridging the gap between experimental development and practical implementation. The reader study further supports the clinical relevance of CRISP, demonstrating performance comparable to an experienced pathologist under time-constrained intraoperative conditions. Remarkably, CRISP alone enabled precise surgical intervention in 92.6% of patients. The diagnostic errors are concentrated in diagnostically challenging scenarios, which are also known to exhibit substantial inter-observer variability, highlighting the importance of cautious interpretation in such cases. Future work may address these limitations by incorporating more diverse training data and integrating uncertainty-aware or human-in-the-loop strategies. In contexts where clinical resources are limited, the ability to optimize resource allocation becomes essential.^{54,55} In routine cases, CRISP-based triage reduced pathologists' workload by 35% and prevented 75% of unnecessary ancillary testing, while in complex cases, it provided a standardized reference to support diagnostic reconsideration and reduce missed diagnoses. Additionally, CRISP demonstrates comparable end-to-end processing time to pathologists in simulated intraoperative settings, indicating its potential to provide timely support. These results highlight CRISP's potential for seamless integration into intraoperative diagnostic workflows, with tangible benefits for both clinical efficiency and patient safety. Beyond pre-screening, CRISP could be integrated into hybrid workflows through direct presentation of model predictions, fusion of algorithmic outputs with expert assessments, or incorporation into agent-based diagnostic systems. Given its robustness and generalizability, CRISP is well-positioned to be compatible with and readily integrated into these collaborative paradigms.

Despite these promising findings, several limitations warrant consideration. First, while extensive retrospective and prospective validation was performed on nearly 20,000 multi-center intraoperative frozen sections and over 3,000 prospective multi-site cases, a fully prospective interventional study is still necessary to assess its real-world deployment and impact on clinical

outcomes.^{56,57} Second, although CRISP demonstrated strong generalization across multiple centers, further validation in more diverse ethnic and geographic populations is essential to account for potential population-level bias and ensure global applicability.^{58,59} In addition, broader long-term multi-institutional deployment may further support continuous monitoring of real-world clinical performance. Finally, large-scale deployment and productization will require addressing technical and operational challenges common to AI clinical translation, including integration with existing workflows, regulatory compliance, and user training.⁶⁰ In summary, CRISP establishes a versatile and clinically relevant AI paradigm for intraoperative pathology, illustrating the transformative potential of large-scale, data-driven approaches to advance precision surgery. By bridging the gap between technological innovation and clinical practice, CRISP paves the way for the next generation of AI-enabled surgical decision support, with broad implications for improving patient outcomes and healthcare delivery worldwide.

Acknowledgements

We gratefully acknowledge the patients from Sun Yat-sen University Cancer Center (SYCC), Zhujiang Hospital, Southern Medical University (ZJH), Jiangmen Central Hospital (JMH), Zhongshan City People's Hospital (ZSH), Nanfang Hospital, Southern Medical University (NFH), Affiliated Cancer Hospital & Institute of Guangzhou Medical University (GZH), Huizhou Central People's Hospital (HZH), The Tenth Affiliated Hospital, Southern Medical University, Dongguan People's Hospital (DGH), Fudan University Shanghai Cancer Center (FUSCC), and Tianjin Medical University Cancer Institute and Hospital (TCIH), as well as the participants of the TCGA program, for providing tissue samples and clinical information for this study.

Funding Statement

This work was supported by grants from the National Key R&D Program of China (No. 2021YFA1300200), the Natural Science Foundation of Guangdong Province (No. 2024A1515012448), Science and Technology Program of Guangzhou (No. 2025A04J3599), Hong Kong Innovation and Technology Commission (Project No. MHP/002/22 and ITC-SKLNSD26SC01), Shenzhen Science and Technology Innovation Committee Fund (Project No. KCXFZ20230731094059008), Research Grants Council of the Hong Kong Special Administrative Region, China (Project No. R6003-22 and C4024-22GF) and Young Talents Program of Sun Yat-sen University Cancer Center (No. YTP-SYSUCC-5260025).

Author contributions

Z.H.Z., F.T.Z., Y.P.Y., B.C., R.G.L., Y.W., H.C. and M.Y.C. conceived the study and designed the experiments. Z.H.Z., Y.P.Y., B.C., X.K.Z., X.B.W., J.W.C, C.Y.Z., J.Y.Z., Y.Q.W., J.J.M., Z.Q.Z., X.Q.L., J.L., R.G.L. and Y.W. collected the data for self-supervised learning. F.T.Z. and J.B.M. performed model development for self-supervised learning. Z.H.Z., F.T.Z., X.K.Z. and M.Y.C. organized the datasets and codebases for all downstream tasks. K.Z., X.Y.Z., D.X., and M.Y.C. performed quality control of the codebase and the results. Z.H.Z., F.T.Z., H.C. and M.Y.C. prepared the paper with input from all co-authors. H.C. and M.Y.C. supervised the research.

Conflict-of-interest disclosure

The authors declare no competing interests.

Figure Legends

Figure 1. Study Cohort Characteristics and Workflow.

(A) Dataset composition and study design. A clinical-grade frozen section dataset comprising 103,797 slides was constructed to train and validate CRISP, followed by real-world deployment for prospective evaluation. (B) Data distribution across participating medical centers. (C) Distribution of slides from 25 major anatomical sites. (D) Overview of the CRISP pretraining strategy. (E) Performance comparison of CRISP with state-of-the-art pathology foundation models on

retrospective downstream tasks. For binary classification tasks, values represent the average AUROC across datasets; for multi-class tasks, values represent the average macro-AUROC. (F) Workflow of CRISP in real intraoperative settings. In a prospective cohort of over 3,000 patients, CRISP delivered robust predictions on intraoperative frozen sections, reduced pathologists' workload, and provided standardized references for challenging cases. The picture in a is from magnific.com. Icons are from flaticon.com and Biorender.

Figure 2. Performance of CRISP in common benign-malignant differentiation tasks.

(A) Task design and dataset utilization for common benign-malignant differentiation tasks. (B) Clinical intraoperative application scenario, where CRISP's diagnosis of lesion nature directly determines the surgical approach. (C) Comparison of average AUROC across tasks between CRISP and other foundation models. (D) Comparison of average F1 scores across tasks between CRISP and other foundation models. (E) Composition of external validation cohorts across anatomical sites (top) and AUROC comparison between CRISP and other models (bottom); error bars denote 95% confidence intervals. (F) Specificity comparison of CRISP and other models at 90% sensitivity (top) and 95% sensitivity (bottom) across tasks. (G) Attention visualization of CRISP predictions on malignant lung (top) and breast (bottom) WSIs, showing from left to right the original WSI, CRISP heatmap, and the most attended regions. For box plots in (D) and (F), center line represents median; box bounds represent 25th and 75th percentiles; whiskers extend to the non-outlier range; outliers are plotted individually. In (D) and (F), $n = 21$ tasks for each model comparison. Statistical significance between models across tasks was assessed using paired Wilcoxon signed-rank tests with Holm correction (two-sided); $*P \leq 0.05$, $**P \leq 0.01$, $***P \leq 0.001$. Data are presented as mean $\pm$ 95% CI in (E), with cohort sizes for each site indicated in the pie charts in (E). Representative heatmaps in (G) are shown from one of the independent test cases; consistent predictive patterns were observed across the test cohort. ZHDG, ZSH&HZH&DGH&GZH; ZHG, ZSH&HZH&GZH; ZSGZ, ZSH&GZH. Icons are from flaticon.com and Biorender.

Figure 3. Evaluation of CRISP for intraoperative decision-making.

(A) Task design and dataset utilization for full-process intraoperative decision-making. (B) Data composition of the external validation cohort for the PCNSLD task (top) and CNSLC task (bottom).

(C) Comparison of AUROC between CRISP and other models on external validation cohorts for the PCNSLD task. (D) Number of lymphoma cases missed by each model in the PCNSLD task. (E) Top-k accuracy comparison of CRISP and other models in external validation cohorts for the CNSLC task. (F) Data composition of external validation cohorts for the OCCGS, OCFGS, TNDD, and PGD tasks. (G) Performance comparison of CRISP and other models across eight PID-procedure decision tasks; Macro-AUC was used for multiclass tasks, and AUROC for binary tasks. (H) Predicted probability distributions of CRISP across classes in the OCCGS task in the external validation cohort. (I) Predicted probability distributions of CRISP across classes in the TNDD task in the external validation cohort. (J) Top-k accuracy comparison of CRISP and other models in the external validation cohort for the OCFGS task. (K) Specificity comparison of CRISP and other models at 95% sensitivity for the PGD task in the external validation cohort. Error bars denote 95% confidence intervals. Error bars denote 95% CI. For box plots in (H) and (I), center line represents median; box bounds represent 25th and 75th percentiles; whiskers extend to the non-outlier range; outliers are plotted individually. Data are presented as mean $\pm$ 95% CI in (C), (E), (J), and (K). Statistical significance was assessed using one-way ANOVA with Tukey's HSD test for pairwise comparisons (two-sided) in (H) and (I); * $P \leq 0.05$, ** $P \leq 0.01$, *** $P \leq 0.001$. Exact cohort sizes for each task are provided in Supplementary Table S3.1. NFZS, NFH&ZSH; OB, Other Benign; FL, Follicular Lesion; PTC, Papillary Thyroid Carcinoma. Icons are from flaticon.com and Biorender.

Figure. 4. Performance of CRISP in late intraoperative assessment.

(A) Data composition of downstream tasks in late intraoperative assessment. (B) Performance comparison of CRISP and other models across cohorts for the SLMD (left), CLMD (middle), and BSMA (right) tasks, evaluated by AUROC (top) and specificity at 95% sensitivity (bottom). (C) Comparison of the total number of positive cases missed by each model in the SLMD (top), CLMD (middle), and BSMA (bottom) tasks. (D) Representative positive frozen sections (top) and CRISP attention visualizations (bottom) in the SLMD (left), CLMD (middle), and BSMA (right) tasks. (E) Number of tasks among 20 full-process intraoperative decision-making tasks where each model achieved top-1 (top) or top-2 (bottom) AUROC. Error bars denote 95% CI. Data are presented as mean $\pm$ 95% CI in (B). Exact cohort sizes for each task are provided in Supplementary Table S3.1. Representative images in (D) are shown from one of the independent test cases; consistent predictive

patterns were observed across the test cohort.

Figure 5 . Pan-cancer detection performance of CRISP.

(A) Data utilization and experimental design of the pan-cancer detection task. (B) Data composition of the external validation cohorts. (C) Overall AUROC comparison across external validation cohorts. (D) Data composition (top) and AUROC comparison (bottom) across different tissue sites; in the heatmap, deeper color indicates better-performing models. (E) Number of tasks (out of 51) where each model achieved top-1 (top) or top-2 (bottom) AUROC in external validation. (F) Proportion of rare cancers in the external validation cohorts. (G) AUROC comparison between observed and unobserved tissue sites across models; crosses denote observed sites, circles denote unobserved origins. (H) AUROC comparison in rare cancers across models; crosses denote all rare cancers, circles denote the subgroup of rare cancers from unobserved sites. (I) Comparison of specificity at 95% sensitivity in rare cancers. (J) Comparison of specificity at 95% sensitivity in unobserved tissue sites. Error bars denote 95% confidence intervals. Data are presented as mean $\pm$ 95% CI in (C). For box plots in (I) and (J), center line represents median; box bounds represent 25th and 75th percentiles; whiskers extend to the non-outlier range; outliers are plotted individually. In (I) and (J), $n = 23$ and 33 tasks for each model comparison. Statistical significance between models was assessed using paired Wilcoxon signed-rank tests with Holm correction (two-sided); $*P \leq 0.05$, $**P \leq 0.01$, $***P \leq 0.001$. Exact cohort sizes for each task are provided in Supplementary Table S1.9. HDGZ, HZH&DGH&GZH. Icons are from flaticon.com and Biorender.

Figure 6. Prospective validation of the clinical utility of CRISP.

(A) Data composition and experimental design of the real-world deployment study. (B) Data composition of each downstream task and AUROC comparison across different models. (C) Overall ROC curve of CRISP compared with diagnostic performance of three pathologists. Pathologist 1 represents a senior pathologist with 10 years of clinical experience; Pathologist 2 and Pathologist 3 represent junior pathologists with 2 years of experience each. (D) AUROC comparison of CRISP predictions across histotechnologist subgroups. (E) AUROC of CRISP across all cases versus difficult cases in each downstream task; crosses denote all cases, circles denote the difficult-case subgroup. (F) Distribution of CRISP prediction scores for all difficult malignant and benign cases.

(G) Confusion matrix of CRISP predictions for negative lymph nodes, micrometastases, and macrometastases in the SLMD task. (H) Distribution of CRISP prediction scores across negative, micrometastatic, and macrometastatic cases. (I) Visualization heatmap of CRISP predictions in a representative micrometastatic case. (J) AUROC of CRISP across all cases versus cases requiring immunohistochemistry (IHC) in each downstream task; crosses denote all cases, circles denote the IHC subgroup. (K) Confusion matrix of CRISP predictions in the IHC-required subgroup. (L) Waterfall plot of CRISP predictions across all 2,071 patients from the SYCC cohort. Using a predefined threshold, 152 typical negative and 571 typical positive patients were identified. (M) Comparison of pathologists' diagnostic sensitivity and specificity before and after applying the CRISP-assisted prescreening strategy. Error bars denote 95% confidence intervals. Data are presented as mean $\pm$ 95% CI in (B) and (M). Box plots in (F) and (H) show the distribution of CRISP prediction scores; center line represents median; box bounds represent 25th and 75th percentiles; whiskers extend to the non-outlier range; outliers are plotted individually. Statistical significance between groups was assessed using two-sided Student's t-test in (F), and one-way ANOVA with Tukey's HSD test for pairwise comparisons (two-sided) in (H); * $P \leq 0.05$, ** $P \leq 0.01$, *** $P \leq 0.001$. Cohort sizes for each site were indicated in the pie charts in (B). Representative heatmap in (I) is shown from one of the independent test cases; consistent predictive patterns were observed across the test cases with micrometastases. Micro, Micrometastases; Macro, Macrometastases. Icons are from flaticon.com and Biorender.

Reference

- Gal, A.A. & Cagle, P.T. The 100-Year Anniversary of the Description of the Frozen Section Procedure. *JAMA* **294**, 3135–3137 (2005).
- Sullivan, R., *et al.* Global cancer surgery: delivering safe, affordable, and timely cancer surgery. *Lancet Oncol.* **16**, 1193–1224 (2015).
- Taxy, J.B. Frozen Section and the Surgical Pathologist: A Point of View. *Arch. Pathol. Lab. Med.* **133**, 1135–1138 (2009).
- Lyman, G.H., *et al.* American Society of Clinical Oncology guideline recommendations for sentinel lymph node biopsy in early-stage breast cancer. *J. Clin. Oncol.* **23**, 7703–7720 (2005).
- Han, Y. & Cai, G. Intraoperative frozen section diagnosis of lung specimens: An updated review. *Semin. Diagn. Pathol.* **42**, 150901 (2025).
- McIntosh, E.R., Harada, S., Drwiega, J., Brandwein-Gensler, M.S. & Gordetsky, J. Frozen section: guiding the hands of surgeons? *Ann. Diagn. Pathol.* **19**,

- 326-329 (2015).
7. Jaafar, H. Intra-operative frozen section consultation: concepts, applications and limitations. *The Malaysian journal of medical sciences: MJMS* **13**, 4 (2006).
8. Ferreiro, J.A., Myers, J.L. & Bostwick, D.G. Accuracy of Frozen Section Diagnosis in Surgical Pathology: Review of a 1-Year Experience With 24,880 Cases at Mayo Clinic Rochester. *Mayo Clin. Proc.* **70**, 1137-1141 (1995).
9. Uematsu, Y., Owai, Y., Okita, R., Tanaka, Y. & Itakura, T. The usefulness and problem of intraoperative rapid diagnosis in surgical neuropathology. *Brain Tumor Pathol.* **24**, 47-52 (2007).
10. Yoshai, E., *et al.* Enhancing frozen histological section images using permanent-section-guided deep learning with nuclei attention. *Sci. Rep.* **15**, 30594 (2025).
11. Varghese, C., Harrison, E.M., O'Grady, G. & Topol, E.J. Artificial intelligence in surgery. *Nature Medicine* **30**, 1257-1268 (2024).
12. He, T., *et al.* Pathology diagnosis of intraoperative frozen thyroid lesions assisted by deep learning. *BMC Cancer* **24**, 1069 (2024).
13. Liu, Y., *et al.* Deep learning to predict cervical lymph node metastasis from intraoperative frozen section of tumour in papillary thyroid carcinoma: a multicentre diagnostic study. *EClinicalMedicine* **60**, 102007 (2023).
14. Kim, Y.G., *et al.* Challenge for Diagnostic Assessment of Deep Learning Algorithm for Metastases Classification in Sentinel Lymph Nodes on Frozen Tissue Section Digital Slides in Women with Breast Cancer. *Cancer Res. Treat.* **52**, 1103-1111 (2020).
15. Vorontsov, E., *et al.* A foundation model for clinical-grade computational pathology and rare cancers detection. *Nature Medicine* **30**, 2924-2935 (2024).
16. Chen, R.J., *et al.* Towards a general-purpose foundation model for computational pathology. *Nature Medicine* **30**, 850-862 (2024).
17. Lu, M.Y., *et al.* A visual-language foundation model for computational pathology. *Nature Medicine* **30**, 863-874 (2024).
18. Ma, J., *et al.* PathBench: A comprehensive comparison benchmark for pathology foundation models towards precision oncology. *arXiv preprint arXiv:2505.20202* (2025).
19. Guo, F., *et al.* Foundation models in bioinformatics. *Natl Sci Rev* **12**, nwaf028 (2025).
20. Ochi, M., Komura, D. & Ishikawa, S. Pathology Foundation Models. *Jma j* **8**, 121-130 (2025).
21. Renne, S.L., Redaelli, S. & Paolini, B. Cryoembedder, automatic processor/stainer, liquid nitrogen freezing, and manual staining for frozen section examination: A comparative study. *Acta Histochem.* **121**, 761-764 (2019).
22. Ozyoruk, K.B., *et al.* A deep-learning model for transforming the style of tissue images from cryosectioned to formalin-fixed and paraffin-embedded. *Nat. Biomed. Eng.* **6**, 1407-1419 (2022).
23. Ho, M.M., Dubey, S., Chong, Y., Knudsen, B. & Tasdizen, T. F2FLDM: Latent

Diffusion Models with Histopathology Pre-Trained Embeddings for Unpaired Frozen Section to FFPE Translation. in *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)* 4382–4391 (2025).

24. Falahkheirkhah, K., *et al.* A generative adversarial approach to facilitate archival-quality histopathologic diagnoses from frozen tissue sections. *Lab. Invest.* **102**, 554–559 (2022).
25. Nagendran, M., *et al.* Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies. *BMJ* **368**, m689 (2020).
26. Tham, Y.C., *et al.* Building the world’s first truly global medical foundation model. *Nature Medicine* (2025).
27. Chang, K., *et al.* The Cancer Genome Atlas Pan-Cancer analysis project. *Nat. Genet.* **45**, 1113–1120 (2013).
28. Wang, X., *et al.* A pathology foundation model for cancer diagnosis and prognosis prediction. *Nature* **634**, 970–978 (2024).
29. Zimmermann, E., *et al.* Virchow2: Scaling self-supervised mixed magnification models in pathology. *arXiv preprint arXiv:2408.00738* (2024).
30. Dismukes, J., *et al.* Prophylactic Central Neck Dissection in Papillary Thyroid Carcinoma: All Risks, No Reward. *J. Surg. Res.* **264**, 230–235 (2021).
31. Huang, T.W., *et al.* Recommendation for axillary lymph node dissection in women with early breast cancer and sentinel node metastasis: A systematic review and meta-analysis of randomized controlled trials using the GRADE system. *Int. J. Surg.* **34**, 73–80 (2016).
32. Bundred, J.R., *et al.* Margin status and survival outcomes after breast cancer conservation surgery: prospectively registered systematic review and meta-analysis. *BMJ* **378**, e070346 (2022).
33. Yang, Z., *et al.* A foundation model for generalizable cancer diagnosis and survival prediction from histopathological images. *Nature Communications* **16**, 2366 (2025).
34. Institute, N.C. About Rare Cancers. (2019).
35. Society, A.C. New Case Estimates by Cancer Type, 2025. Vol. 2025 (2025).
36. Rajpurkar, P., Chen, E., Banerjee, O. & Topol, E.J. AI in health and medicine. *Nature Medicine* **28**, 31–38 (2022).
37. Xu, H., *et al.* A whole-slide foundation model for digital pathology from real-world data. *Nature* **630**, 181–188 (2024).
38. Campanella, G., *et al.* Real-world deployment of a fine-tuned pathology foundation model for lung cancer biomarker detection. *Nature Medicine* (2025).
39. Banerji, C.R.S., Chakraborti, T., Harbron, C. & MacArthur, B.D. Clinical AI tools must convey predictive uncertainty for each individual patient. *Nature Medicine* **29**, 2996–2998 (2023).
40. Kappen, T.H., *et al.* Evaluating the impact of prediction models: lessons learned, challenges, and recommendations. *Diagn Progn Res* **2**, 11 (2018).
41. Meara, J.G., *et al.* Global Surgery 2030: evidence and solutions for achieving health, welfare, and economic development. *The Lancet* **386**, 569–624 (2015).

- 753 42. Dobson, G.P. Trauma of major surgery: A global problem that is not going away.
754 *Int. J. Surg.* **81**, 47–54 (2020).
- 755 43. ARPA-H. The Big Question: Precision Surgical Interventions. (2024).
- 756 44. Boer, M.d., *et al.* Micrometastases or Isolated Tumor Cells and the Outcome
757 of Breast Cancer. *New England Journal of Medicine* **361**, 653–663 (2009).
- 758 45. Pleijhuis, R.G., *et al.* Obtaining adequate surgical margins in breast-
759 conserving therapy for patients with early-stage breast cancer: current
760 modalities and future directions. *Ann. Surg. Oncol.* **16**, 2717–2730 (2009).
- 761 46. Loftus, T.J., *et al.* Artificial Intelligence and Surgical Decision-making.
762 *JAMA Surgery* **155**, 148–158 (2020).
- 763 47. Stockl, T., *et al.* Accuracy of Frozen Section in the Diagnosis of
764 Gynecological Tumors: 5 Years of Experience from a Tertiary Center. *Gynecol.*
765 *Oncol.* **139**, 598 (2015).
- 766 48. Kondepudi, A., *et al.* Foundation models for fast, label-free detection of
767 glioma infiltration. *Nature* **637**, 439–445 (2025).
- 768 49. Fu, X., *et al.* A foundation model of transcription across human cell types.
769 *Nature* **637**, 965–973 (2025).
- 770 50. Wang, S., *et al.* Annotation-efficient deep learning for automatic medical
771 image segmentation. *Nature Communications* **12**, 5915 (2021).
- 772 51. Hu, X., Zeng, D., Xu, X. & Shi, Y. Semi-supervised Contrastive Learning
773 for Label-Efficient Medical Image Segmentation. in *Medical Image Computing*
774 *and Computer Assisted Intervention – MICCAI 2021* (eds. de Bruijne, M., *et*
775 *al.*) 481–490 (Springer International Publishing, Cham, 2021).
- 776 52. Lu, M.Y., *et al.* Data-efficient and weakly supervised computational pathology
777 on whole-slide images. *Nat. Biomed. Eng.* **5**, 555–570 (2021).
- 778 53. Walsh, E. & Orsi, N.M. The current troubled state of the global pathology
779 workforce: a concise review. *Diagn. Pathol.* **19**, 163 (2024).
- 780 54. Robboy, S.J., Black-Schaffer, W.S. & Wheeler, T.M. International Medical
781 Graduates and the Shortage of US Pathologists. *Arch. Pathol. Lab. Med.* **148**,
782 1181–1182 (2024).
- 783 55. Zhang, X., Lin, D., Pforsich, H. & Lin, V.W. Physician workforce in the United
784 States of America: forecasting nationwide shortages. *Human Resources for*
785 *Health* **18**, 8 (2020).
- 786 56. Cruz Rivera, S., *et al.* Guidelines for clinical trial protocols for
787 interventions involving artificial intelligence: the SPIRIT-AI extension.
788 *Nature Medicine* **26**, 1351–1363 (2020).
- 789 57. Ibrahim, H., *et al.* Reporting guidelines for clinical trials of artificial
790 intelligence interventions: the SPIRIT-AI and CONSORT-AI guidelines. *Trials*
791 **22**, 11 (2021).
- 792 58. Vyas, D.A., Eisenstein, L.G. & Jones, D.S. Hidden in Plain Sight —
793 Reconsidering the Use of Race Correction in Clinical Algorithms. *New England*
794 *Journal of Medicine* **383**, 874–882 (2020).
- 795 59. Obermeyer, Z., Powers, B., Vogeli, C. & Mullainathan, S. Dissecting racial
796 bias in an algorithm used to manage the health of populations. *Science* **366**,

447–453 (2019).

60. Lüscher, T.F., Wenzl, F.A., D'Ascenzo, F., Friedman, P.A. & Antoniadou, C. Artificial intelligence in cardiovascular medicine: clinical applications. *Eur. Heart J.* **45**, 4291–4304 (2024).

Methods

Ethical approval

This study was approved by the Ethics Committee of Sun Yat-sen University Cancer Center (Approval No. SL-B2024-708) and carried out in compliance with the Declaration of Helsinki. This was a prospective observational study, not an interventional trial. Written informed consent was obtained from all subjects enrolled in the prospective cohort, while the requirement for informed consent was waived for retrospective cohorts. The study was registered in the Chinese Clinical Trial Registry (ChiCTR2500106350) as a retrospective registration.

Large-scale intraoperative frozen section dataset

To develop a clinically-oriented foundation model for intraoperative pathology, we constructed a large-scale dataset of frozen section whole-slide images (WSIs), comprising 103,797 slides from 51,186 patients. Retrospective data were collected from eight independent medical centers, including Sun Yat-sen University Cancer Center (SYCC), Zhujiang Hospital, Southern Medical University (ZJH), Jiangmen Central Hospital (JMH), Zhongshan City People's Hospital (ZSH), Nanfang Hospital, Southern Medical University (NFH), Affiliated Cancer Hospital & Institute of Guangzhou Medical University (GZH), Huizhou Central People's Hospital (HZH), and The Tenth Affiliated Hospital, Southern Medical University, Dongguan People's Hospital (DGH). All slides collected from the participating centers were obtained from routine clinical intraoperative consultations. In total, these cohorts encompassed 38,253 patients and 81,678 intraoperative frozen WSIs collected between May 2017 and July 2024. The following inclusion criteria were applied to all cases: (1) at least one available intraoperative frozen section; (2) presence of identifiable tissue on the slide; (3) absence of scanning artifacts such as defocus or blurring; and (4) availability of complete diagnostic information, including intraoperative frozen section and corresponding postoperative paraffin section reports, as well as basic clinical characteristics (e.g., age and sex). Slides were manually reviewed by two

experienced pathologists based on thumbnail inspection to assess tissue presence and image quality, and only those with severe scanning-related artifacts (e.g., out-of-focus regions or extensive image degradation) that precluded tissue recognition were excluded, while intrinsic artifacts inherent to frozen section preparation were retained. The detailed inclusion period, number of patients, number of slides, and baseline characteristics of each cohort are summarized in Supplementary Tables S1.1-1.8. To further increase histological diversity, we additionally curated 15,975 frozen WSIs from The Cancer Genome Atlas (TCGA). The detailed selection process was as follows: diagnostic slides prepared from formalin-fixed paraffin-embedded (FFPE) tissues were excluded, resulting in 21,094 slides. These were subsequently reviewed by two experienced pathologists, and immunohistochemistry as well as other non-frozen section slides were removed. To ensure consistency, all patients were categorized into 24 major anatomical sites according to the definitions provided by the Genomic Data Commons (GDC) system.⁶¹ For model development, WSIs from ZJH, JMH, TCGA, and a subset of SYCC (43,013 slides) were used for CRISP pretraining. The remaining SYCC slides (5,178) were reserved for downstream task development and internal evaluation, ensuring no data leakage between pretraining and testing. External validation was performed using ZSH, NFH, GZH, HZH, and DGH cohorts.

For the prospective real-world deployment study of CRISP, patients undergoing thyroid, breast, or lung surgery with intraoperative consultation at Sun Yat-sen University Cancer Center were consecutively recruited between November 2024 and July 2025, representing the most common sites for intraoperative consultation. Inclusion criteria were: (1) at least one available intraoperative frozen section; (2) presence of identifiable tissue on the slide; (3) absence of scanning artifacts such as defocus or blurring; (4) complete diagnostic and clinical information; (5) age ≥ 18 years at the time of surgery; and (6) documented informed consent. Following screening, a total of 2,071 patients with 4,143 intraoperative frozen sections were included. To further assess geographical generalizability, two additional prospective cohorts from FUSCC and TCIH were included, with the same inclusion criteria applied. To simulate real-world application, 100 consecutive cases from SYCC were included for simulated clinical deployment analysis. Recruitment periods and detailed clinical characteristics for each cohort are provided in Supplementary Tables S1.9-1.11.

Data digitization and preprocessing

All clinical intraoperative frozen sections included in this study were digitized at 40× magnification (0.25 μm/pixel). Slides from the SYCC cohort were primarily scanned using an Aperio AT2 scanner (Leica Biosystems, Wetzlar, Germany) and stored in SVS format. Data from the ZJH, GZH, HZH, and DGH cohorts were scanned with a PHILIPS Ultra-Fast Scanner (Philips Electronics N.V., Amsterdam, Netherlands) and saved in iSyntax format. NFH and ZSH cohort slides were mainly scanned using a KF-PRO-120-HI scanner (Jiangfeng Biotech, Ningbo, China) and stored in KFB format. JMH cohort slides were scanned using an SQS-600P scanner (Shengqiang Technology Co., Ltd., Shenzhen, China) and saved in SDPC format. Slides from the prospective SYCC cohort were mainly scanned with a PHILIPS Ultra-Fast Scanner (Philips Electronics N.V., Amsterdam, Netherlands) and saved in iSyntax format. Slides from the FUSCC and TCIH cohorts were mainly scanned using an Aperio AT2 scanner (Leica Biosystems, Wetzlar, Germany) and stored in SVS format.

WSIs were processed using vendor-provided API tools to extract full-slide thumbnail images. Given the extensive presence of non-informative background areas, tissue regions were automatically segmented using the CLAM toolkit to identify and isolate relevant histological content for subsequent analysis.⁵² Segmented tissue regions were partitioned into a series of non-overlapping patches of 512×512 pixels at a resolution of 0.25 microns per pixel (MPP). To mitigate center-specific biases introduced by variations in optical resolution, slide preparation, and scanning settings, redundant patches from higher-resolution images were downsampled. This approach ensured a balanced representation of patches from each center, reducing the risk of model overfitting to specific institutional characteristics. In total, approximately 50 million patches were extracted from 81,667 slides for model pretraining.

No color normalization (e.g., Vahadane or Macenko) was applied in order to preserve inter-institutional variability and reflect real-world clinical conditions.

Network architecture

Building on previous studies of FFPE-based pathology foundation models, we focus here on intraoperative frozen sections and introduce CRISP, a specialized pathology foundation model

for their analysis. CRISP is built upon Virchow2, a state-of-the-art pathology foundation model pretrained on 1 million hematoxylin and eosin (H&E)-stained WSIs.^{15,29} Virchow2 was selected as the backbone model based on prior large-scale benchmarking studies demonstrating its strong overall performance across multiple pathology tasks.¹⁸ To adapt this general-purpose model to the frozen section domain while preserving its learned representations and generalizability, we employed low-rank adaptation (LoRA).⁶² LoRA is a parameter-efficient fine-tuning (PEFT) method that injects trainable rank decomposition matrices into the self-attention layers of the transformer architecture, keeping the remaining model weights frozen during training. This approach drastically reduces the number of trainable parameters, computational overhead, and risk of catastrophic forgetting. In this study, LoRA modules were integrated into the query and value projection matrices of each attention layer, initialized with a rank of 8 and a scaling factor of 16. This configuration resulted in only 2 million trainable parameters, constituting merely 0.3% of the model's total 633 million parameters.

Model training

For pretraining, we utilized DINO, a self-supervised learning strategy effective for learning visual representations without labeled data.⁶³ Specifically, DINO operates on a teacher-student paradigm in which two different augmented views of the same image are processed by the student and teacher networks, respectively. The student network learns to predict the output distribution of the teacher network, thereby encouraging the learning of semantically meaningful features without manual annotation. The teacher network's weights are updated via an exponential moving average (EMA) of the student weights, with a momentum coefficient set to 0.9995. The model was trained using the AdamW optimizer ($\beta_1 = 0.9$, $\beta_2 = 0.999$) with mixed-precision (FP16) arithmetic to accelerate training and reduce memory consumption. Pretraining was conducted using 8 NVIDIA H800 GPUs with a batch size of 384 for 640,000 iterations.

For downstream intraoperative tasks, a slide-level prediction pipeline was implemented. In computational pathology, WSIs are typically too large to be processed directly by deep learning models. Therefore, the analysis pipeline was divided into three steps: patch-level feature extraction, feature aggregation, and slide-level prediction. In this study, each 512×512 patch

(at 0.25 MPP) was resized to 224×224 pixels to serve as input to the frozen CRISP encoder; this corresponds to the same field of view as 256×256 patches at 0.5 MPP, a commonly adopted setting in baseline models, thereby ensuring consistent input resolution and fair comparison across models.^{16,37} From the final layer of the transformer, the [CLS] token embedding and the mean-pooled embeddings of all patch tokens were concatenated, forming a comprehensive 2,560-dimensional feature vector for each patch. To identify diagnostically critical regions, we employed an attention-based multiple instance learning (ABMIL) model as the feature aggregator.⁶⁴ The ABMIL module learns to assign an attention weight to each patch feature vector, effectively highlighting salient patches and generating a weighted slide-level feature representation. This aggregated feature vector was then passed through a single fully connected classification layer to produce the final slide-level prediction. For supervised downstream tasks, the cross-entropy loss function was used. Only the ABMIL aggregator and the classification layer were trained from scratch, while the CRISP encoder remained frozen to leverage the pretrained features. The model was optimized using the Adam optimizer with a learning rate of 2×10^{-4} and weight decay of 1×10^{-5} . Early stopping based on validation loss was implemented to halt training upon performance saturation and prevent overfitting. In all downstream task comparisons, identical strategies were employed across comparator models to guarantee fairness in performance assessment.

Downstream task

To comprehensively evaluate CRISP, 98 retrospective downstream tasks were established. The reference labels for all tasks were based on postoperative FFPE morphology integrated with immunohistochemical and molecular analyses, and were independently verified by a pathologist. Aggregator development for all downstream tasks was conducted using the SYCC cohort, with patients in the internal test sets kept independent from the training and validation sets to prevent data leakage. Data were randomly split at the patient level into training, validation, and test sets to prevent data leakage, with stratification by outcome labels; identical splits were used for all models to ensure fair comparison. No explicit rebalancing strategies were applied in order to preserve the original class distribution. A shared set of hyperparameters was used across all tasks and models to ensure a fair comparison. Baseline models were selected

based on their relevance to histopathology foundation modeling, as well as the availability of publicly accessible implementations or pretrained weights, to ensure reproducibility and fair comparison.

Classification of common benign and malignant lesions: Each benign-malignant classification task was formulated as an independent slide-level binary problem, with slides containing malignant regions designated as positive. Six prevalent tissue sites were included: lung, breast, thyroid, head and neck, brain, and ovary. The task encompassed commonly encountered benign and malignant lesions, such as pulmonary adenocarcinoma and inflammatory nodules, to replicate large-scale clinical pre-screening scenarios. The abdominal metastasis classification task was incorporated to evaluate CRISP's capability in distinguishing complex metastatic lesions. Abdominal metastases represent a frequent dissemination route for multiple cancers and often signal disease progression; accurate identification can therefore inform patient management and improve clinical outcomes.^{65,66} Malignant abdominal nodules comprised diverse metastatic cancers, including ovarian, intestinal, and gastric origins (Supplementary Table S6.1). Model performance was assessed on 2,771 slides across five external cohorts, with dataset partitions detailed in Supplementary Table S2.1. For comparison, a case-level aggregation strategy was implemented by concatenating slide-level features from the same case to form case-level representations. Identical data splits and evaluation protocols were used to ensure fair comparison.

Early intraoperative decision: The primary central nervous system lymphoma differentiation (PCNSLD) task was defined as a slide-level binary classification, aimed at accurately distinguishing primary central nervous system lymphoma (PCNSL) from other brain lesions. PCNSL is recognized as one of the most common primary malignant brain tumors, exhibiting poor prognosis, presenting diagnostic challenges, and being primarily managed with radiotherapy and chemotherapy, in contrast to other brain lesions.⁶⁷ Accurate intraoperative identification of PCNSL can prevent unnecessary surgical interventions and facilitate precision treatment. Slides containing PCNSL were labeled positive, whereas slides of other common brain lesions—including gliomas, medulloblastomas, central neurocytomas, metastases, and inflammatory lesions—were labeled negative. External validation was performed using 461 slides from the NFH cohort. The central nervous system lesion classification (CNSLC) task

comprised the same dataset but was formulated as a fine-grained multi-class classification to distinguish among the aforementioned lesion types. Detailed dataset partitions for intraoperative decision tasks are provided in Supplementary Table S3.1.

Mid-procedure decision: The ovarian cancer coarse-grained subtyping (OCCGS) task involved slide-level three-class classification to differentiate primary ovarian carcinomas, metastatic ovarian tumors, and benign ovarian lesions. As the ovary is a common site of metastasis from extragenital malignancies, distinguishing primary from metastatic tumors remains a critical diagnostic challenge during surgery.^{68,69} Accurate classification informs the surgical extent and guides additional treatment directed at the primary site. The ovarian cancer fine-grained subtyping (OCCFS) task shared the same dataset but focused on fine-grained classification of primary ovarian carcinoma subtypes, including serous, clear cell, and endometrioid adenocarcinomas.

The thyroid nodule differential diagnosis (TNDD) task involved a slide-level three-class classification of thyroid nodules into papillary thyroid carcinoma, follicular-patterned tumors, and benign lesions such as nodular goiter. Frozen section diagnosis of thyroid nodules is critical for intraoperative surgical decision-making, as papillary carcinoma and nodular goiter often dictate different surgical extents, while follicular-patterned tumors typically require diagnostic surgery pending definitive FFPE evaluation.⁷⁰

The parathyroid gland detection (PGD) task was a slide-level binary classification for detecting parathyroid tissue in thyroidectomy frozen sections. Inadvertent parathyroid resection, occurring in up to 16% of thyroidectomies, is the primary cause of postoperative hypocalcemia.^{71,72} Accurate frozen-section recognition of parathyroid tissue is therefore essential for preventing misresection and optimizing surgical outcomes. Labels were assigned solely based on the presence of parathyroid tissue, with all other tissues—including thyroid, lymph nodes, and thyroid carcinoma—classified as negative, reflecting real-world intraoperative settings.

Late intraoperative assessment: Sentinel lymph node metastasis detection (SLMD), cervical lymph node metastasis detection (CLMD), and breast cancer surgical margin assessment (BSMA) are three slide-level binary classification tasks designed to detect residual tumor cells in lymph nodes and surgical margins. Sentinel lymph node assessment and margin evaluation

are critical indicators of surgical completeness and directly influence patient prognosis and risk of recurrence.^{73,74} In thyroid cancer, accurate intraoperative identification of central lymph node metastasis can help avoid unnecessary dissections, whereas in breast cancer, margin and sentinel lymph node assessment form the basis of individualized treatment.^{13,75,76}

Pan-cancer detection study: The pan-cancer detection task was designed to develop a universal model capable of accurately identifying malignant lesions across intraoperative frozen samples of any source or type, thereby evaluating CRISP performance under data-limited conditions, particularly in unseen samples and rare cancers. Model development was based on 3,071 benign and malignant WSIs from six common frozen tissue sites in the SYCC cohort. Given that clinical decision-making is case-based, external validation was conducted at the case level. For each case, positivity was defined as the presence of at least one malignant slide. All slide-level features from a case were concatenated and directly input into the aggregator as case-level features. The model was validated across five external cohorts comprising 5,867 cases: ZSH and NFH served as two independent cohorts, whereas GZH, HZH, and DGH were combined into a single cohort. Detailed dataset partitions are provided in Supplementary Tables S4.2-4.5. For subgroup analyses, eight tissue sites (stomach, bone, esophagus, kidney, bladder, adrenal gland, testis, and thymus) were combined as “Else”, owing to limited sample sizes or unidirectional labels in external cohorts, which precluded reliable calculation of metrics such as AUROC.

Prospective cohort study

In the prospective cohort, the diagnostic reference standard for each patient was established based on the postoperative FFPE pathology report in combination with immunohistochemistry and molecular test results, and independently verified by a board-certified pathologist blinded to the study. Patients from SYCC were assigned to one of six intraoperative decision tasks (PNBM, TNBM, BNBM, CLMD, SLMD, BSMA) according to the surgical indication and sample type. Detailed dataset partitions are provided in Figure 2b. Aggregators used for validation in these six tasks were directly transferred from retrospective downstream task construction without any task-specific fine-tuning. For clinical applicability, validation was performed at the patient level.

Information on frozen-section slide preparation personnel was obtained from intraoperative reports, encompassing a total of 30 histotechnologists; four with single-class labels were not suitable for AUROC-based evaluation and were therefore grouped into an “Else” category for analysis. Diagnostic outcomes and confidence levels were derived from the content of the intraoperative frozen-section reports. Confidence assessment was conducted according to a prespecified rule set: reports indicating “consistent with”, “consider”, or “favor” were categorized as relatively definitive, whereas reports stating “not excluded”, “suspicious for”, or “requires differential diagnosis” were categorized as indeterminate (Supplementary Table S6.2). All classifications were independently reviewed and confirmed by a pathologist blinded to the study. Immunohistochemistry requirements were similarly inferred from the report content. The upper and lower thresholds used for coarse screening were defined based on the range of prediction scores observed in misclassified cases, such that samples within the retained range were classified with near-perfect accuracy by the model.

To evaluate clinical performance, a reader study was conducted using the entire prospective SYCC cohort. One senior pathologist (Pathologist 1; 10 years of diagnostic experience) and two junior pathologists (Pathologist 2, 3; 2 years of experience) independently reviewed all cases, blinded to model predictions and study design. Readers were asked to perform diagnostic classification across multiple clinically relevant tasks, consistent with the study endpoints. All reader predictions were recorded for subsequent analysis and comparison with model outputs.

To simulate intraoperative conditions during the reader study, each case was evaluated under a time constraint of 5 minutes. Linear probing experiments were conducted on multi-organ benign-versus-malignant classification tasks (lung, breast, and thyroid) to assess the intrinsic discriminative capacity and transferability of pretrained features. A standardized benchmark subset comprising 75 frozen-section slides from an independent SYCC cohort, not used during foundation model pretraining, was constructed for evaluation, with regions of interest annotated by an independent pathologist. For the linear probing experiments, fixed feature representations were extracted from each foundation model, and logistic regression classifiers were trained on top without updating the backbone parameters, applying identical data splits and hyperparameters across all models to ensure fair comparison. To evaluate robustness, we further assessed model performance under common perturbations, including color shifts, image blur,

additive noise, and reduced image resolution, based on the linear probing setting. The entire process was designed as a largely automated pipeline (slide scanning → data upload → model inference → result output), with minimal manual intervention required. Regarding the simulated deployment analysis, whole-slide images were digitized using six PHILIPS Ultra-Fast Scanners in parallel and automatically uploaded to a local server via the SYCC intranet. Model inference was performed on a server equipped with two NVIDIA A100 GPUs, and results were returned via an internal web-based interface. For the pre-screening strategy, the same threshold and decision criteria used in the prospective SYCC cohort were applied to this independent dataset.

Micro-metastases and macro-metastases were defined according to AJCC guidelines and determined based on both intraoperative frozen-section and postoperative FFPE pathology reports, and were independently morphologically reviewed by a pathologist blinded to the study. Micro-metastases were defined as malignant foci within sentinel lymph nodes measuring greater than 0.2 mm but not exceeding 2 mm, whereas macro-metastases were defined as malignant foci larger than 2 mm.⁷⁷

Statistical analysis

For binary classification tasks, model performance was evaluated using the area under the receiver operating characteristic curve (AUROC), F1 score, accuracy, and specificity at 90% and 95% sensitivity thresholds. For multi-class classification tasks, performance was assessed using macro-averaged AUROC and top-k accuracy. For each case, logits from multiple classes were combined using a weighted summation, followed by a sigmoid transformation to rescale the aggregated value into a unified CRISP score between 0 and 1. For binary classification tasks, optimal thresholds were determined using the Youden index on the internal validation set and subsequently applied to external test cohorts; for multi-class tasks, predictions were assigned based on the class with the highest logit. To compare overall model performance across multiple tasks, paired Wilcoxon signed-rank tests were performed, and p-values were adjusted for multiple comparisons using the Holm method.⁷⁸ The effect size r is derived from the formula $r = Z / \sqrt{N}$, with Z representing the standard normal deviate corresponding to U , and N denoting the combined sample size. Two-sided 95% confidence intervals (CIs) were computed

via bootstrap with 1,000 iterations. Differences in predicted value distributions for multi-class tasks were assessed using one-way analysis of variance (ANOVA) followed by Tukey's honestly significant difference (HSD) test for pairwise comparisons. For binary tasks, group differences in predicted score distributions were evaluated using two-sided Student's t-tests; non-parametric alternatives were considered when normality assumptions were not met. Statistical significance was defined as not significant ($P > 0.05$), * ($P \leq 0.05$), ** ($P \leq 0.01$) and *** ($P \leq 0.001$). Boxplots depict the median (50th percentile), interquartile range (25th-75th percentile); whiskers extend to the non-outlier range; outliers are plotted individually. Data analysis was performed using R version 4.2.1 (2022-06-23) and Python 3.13.5 using the numpy (version 2.2.3), scipy (version 1.15.3) and scikit-learn (version 1.7.0) packages, and deep learning models were implemented in PyTorch 2.7.1.

Data availability

The TCGA whole-slide images and associated clinical data can be accessed through the NIH Genomic Data Commons (<https://portal.gdc.cancer.gov>). Whole-slide images and clinical data from the internal and external cohorts are not publicly available owing to institutional policies and patient privacy regulations, but were collected under Institutional Review Board approvals. Access to these datasets for non-commercial academic use may be requested from the corresponding author (caimy@sysucc.org.cn). Requests will be evaluated and responded to within two weeks. A de-identified dataset comprising approximately 35,000 annotated patch-level images together with 75 corresponding whole-slide images from SYCC intraoperative frozen sections used in the linear probing experiments is publicly available to facilitate independent validation and future research (<https://huggingface.co/datasets/FTZhou/CRISP>). All other data supporting the findings of this study are available within the article and its Supplementary Information files.

Code availability

The source code for the proposed framework is available at <https://github.com/FT-ZHOU-ZZZ/CRISP>. The pretrained model weights of CRISP are available at Hugging Face (<https://huggingface.co/FTZhou/CRISP>).

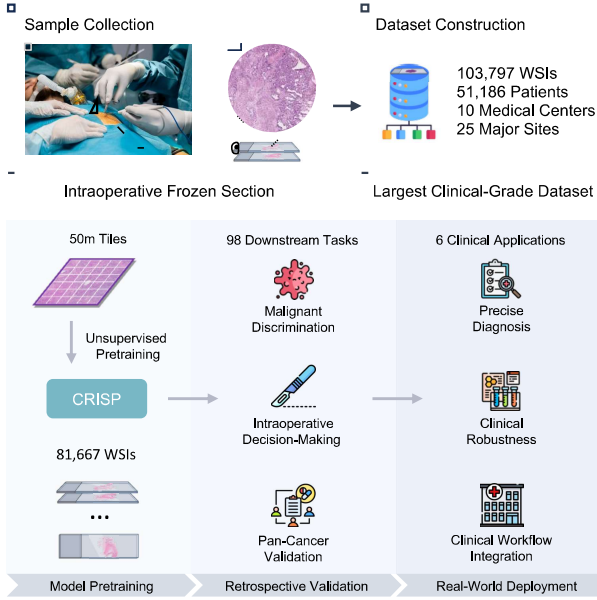
Methods-only references

61. Institute, N.C. Genomic Data Commons Data Portal. Vol. 2025.
62. Hu, E.J., *et al.* Lora: Low-rank adaptation of large language models. *ICLR* **1**, 3 (2022).
63. Caron, M., *et al.* Emerging properties in self-supervised vision transformers. in *Proceedings of the IEEE/CVF international conference on computer vision* 9650–9660 (2021).
64. Ilse, M., Tomczak, J. & Welling, M. Attention-based deep multiple instance learning. in *International conference on machine learning* 2127–2136 (PMLR, 2018).
65. Cui, H.T., Zhu, Q.Y., Zhao, H.W., Liu, H.L. & Wang, N. Pregnancy is associated with the prognosis of ovarian cancer patients with abdominal metastasis. *Am. J. Cancer Res.* **15**, 168–181 (2025).
66. Chen, H., *et al.* Artificial intelligence assisted real-time recognition of intra-abdominal metastasis during laparoscopic gastric cancer surgery. *npj Digital Medicine* **8**, 9 (2025).
67. von Baumgarten, L., *et al.* The Diagnosis and Treatment of Primary CNS Lymphoma. *Dtsch Arztebl Int* **115**, 419–426 (2018).
68. Lee, S.J., *et al.* Clinical characteristics of metastatic tumors to the ovaries. *J. Korean Med. Sci.* **24**, 114–119 (2009).
69. Prat, J. Ovarian carcinomas, including secondary tumors: diagnostically challenging areas. *Mod. Pathol.* **18**, S99–S111 (2005).
70. Trosman, S.J., Bhargavan, R., Prendes, B.L., Burkey, B.B. & Scharpf, J. The contemporary utility of intraoperative frozen sections in thyroid surgery. *Am. J. Otolaryngol.* **38**, 614–617 (2017).
71. Khairy, G.A. & Al-Saif, A. Incidental parathyroidectomy during thyroid resection: incidence, risk factors, and outcome. *Ann. Saudi Med.* **31**, 274–278 (2011).
72. Page, C. & Strunski, V. Parathyroid risk in total thyroidectomy for bilateral, benign, multinodular goitre: report of 351 surgical cases. *J. Laryngol. Otol.* **121**, 237–241 (2007).
73. Park, K.U., *et al.* Sentinel Lymph Node Biopsy in Early-Stage Breast Cancer: ASCO Guideline Update. *Journal of Clinical Oncology* **43**, 1720–1741 (2025).
74. Li, W. & Li, X. Development of intraoperative assessment of margins in breast conserving surgery: a narrative review. *Gland Surg* **11**, 258–269 (2022).
75. Yu, J., *et al.* Lymph node metastasis prediction of papillary thyroid carcinoma based on transfer learning radiomics. *Nature Communications* **11**, 4807 (2020).
76. Kopicky, L., Fan, B. & Valente, S.A. Intraoperative evaluation of surgical

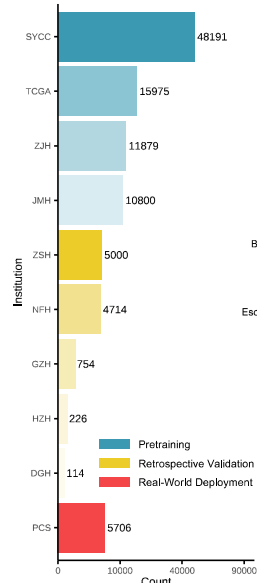
- 1168 margins in breast cancer. *Semin. Diagn. Pathol.* **41**, 293–300 (2024).
1169 77. Frederick, L., *et al.* *AJCC cancer staging manual*, (Springer Science & Business
1170 Media, 2002).
1171 78. Holm, S. A simple sequentially rejective multiple test procedure. *Scandinavian*
1172 *journal of statistics*, 65–70 (1979).
1173

ACCELERATED ARTICLE PREVIEW

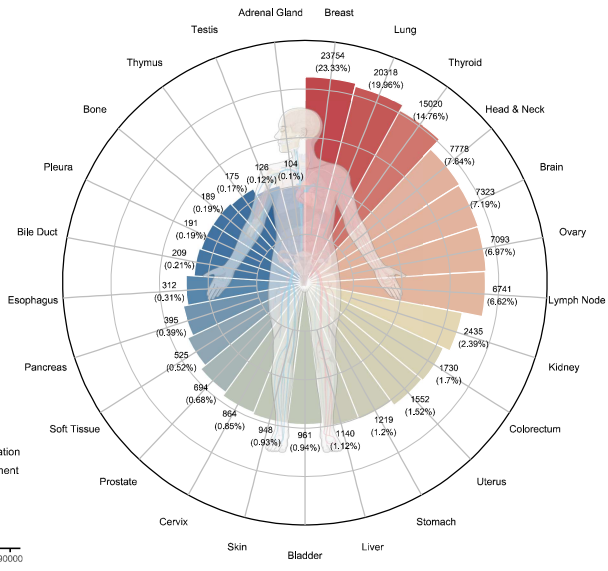
a



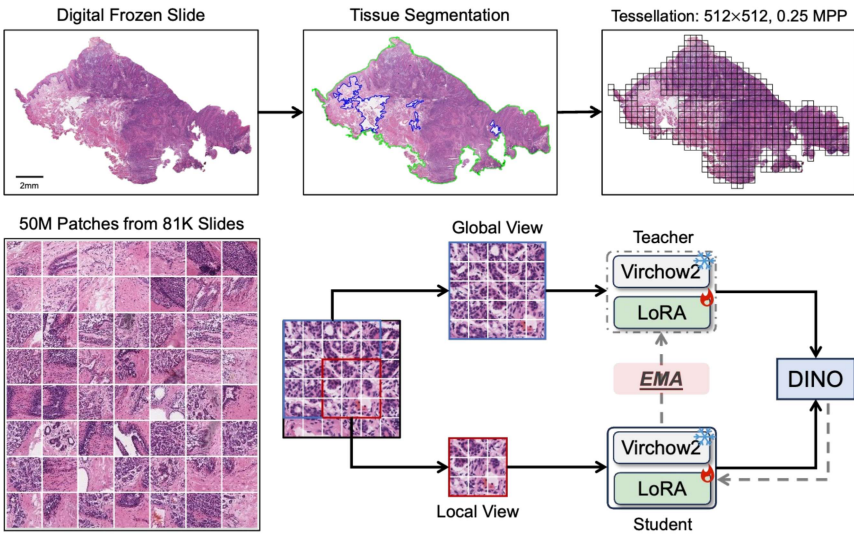
b



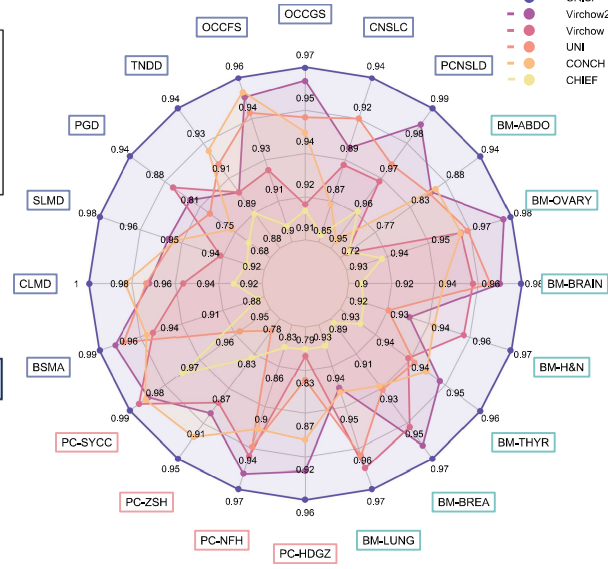
c



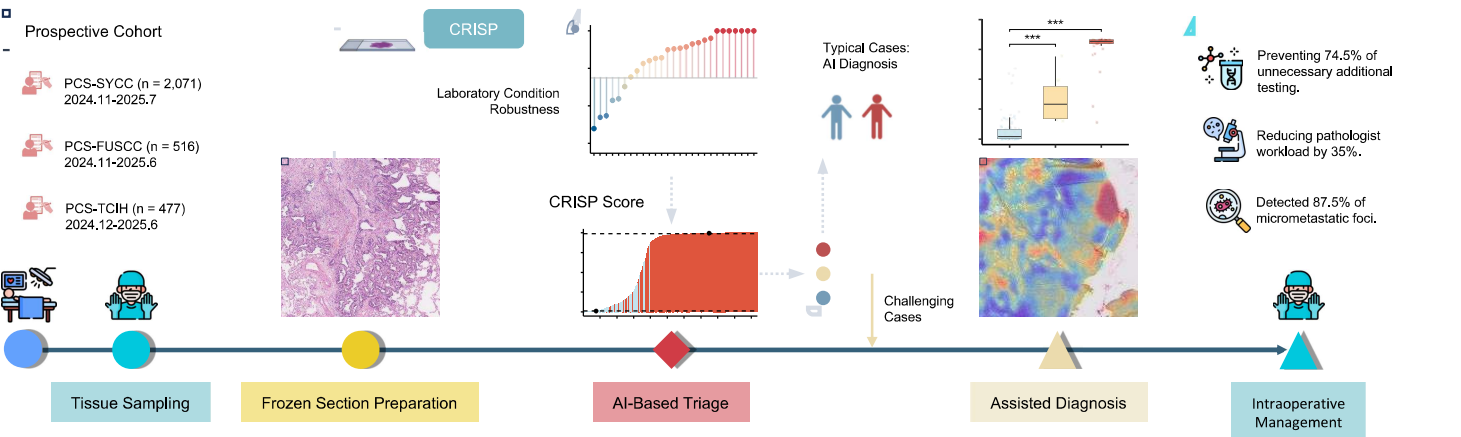
d

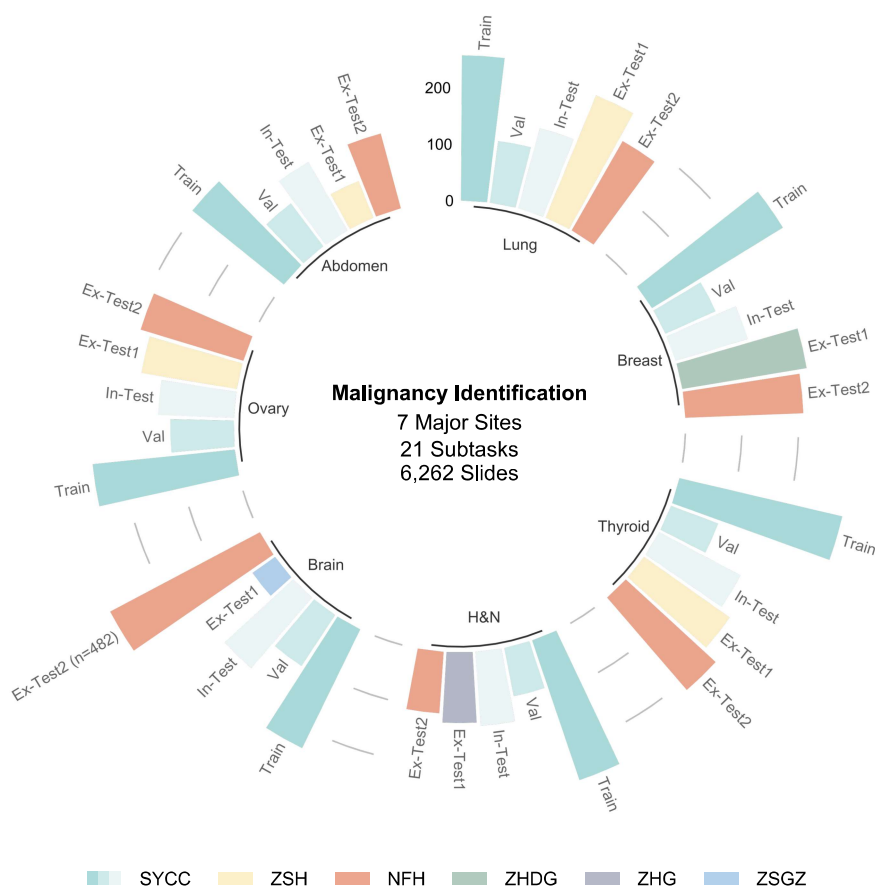
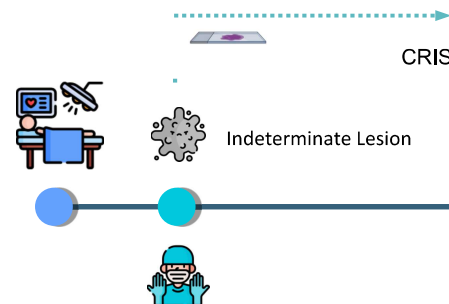
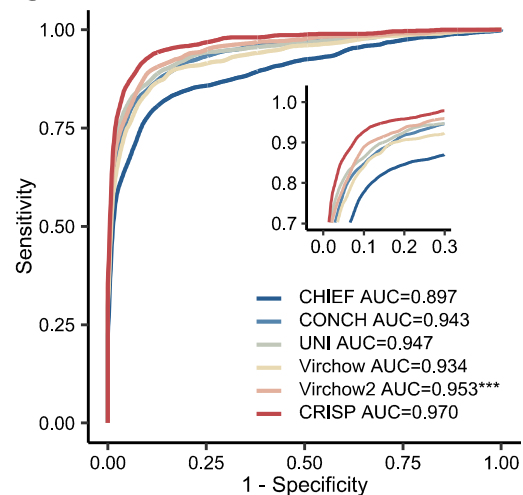
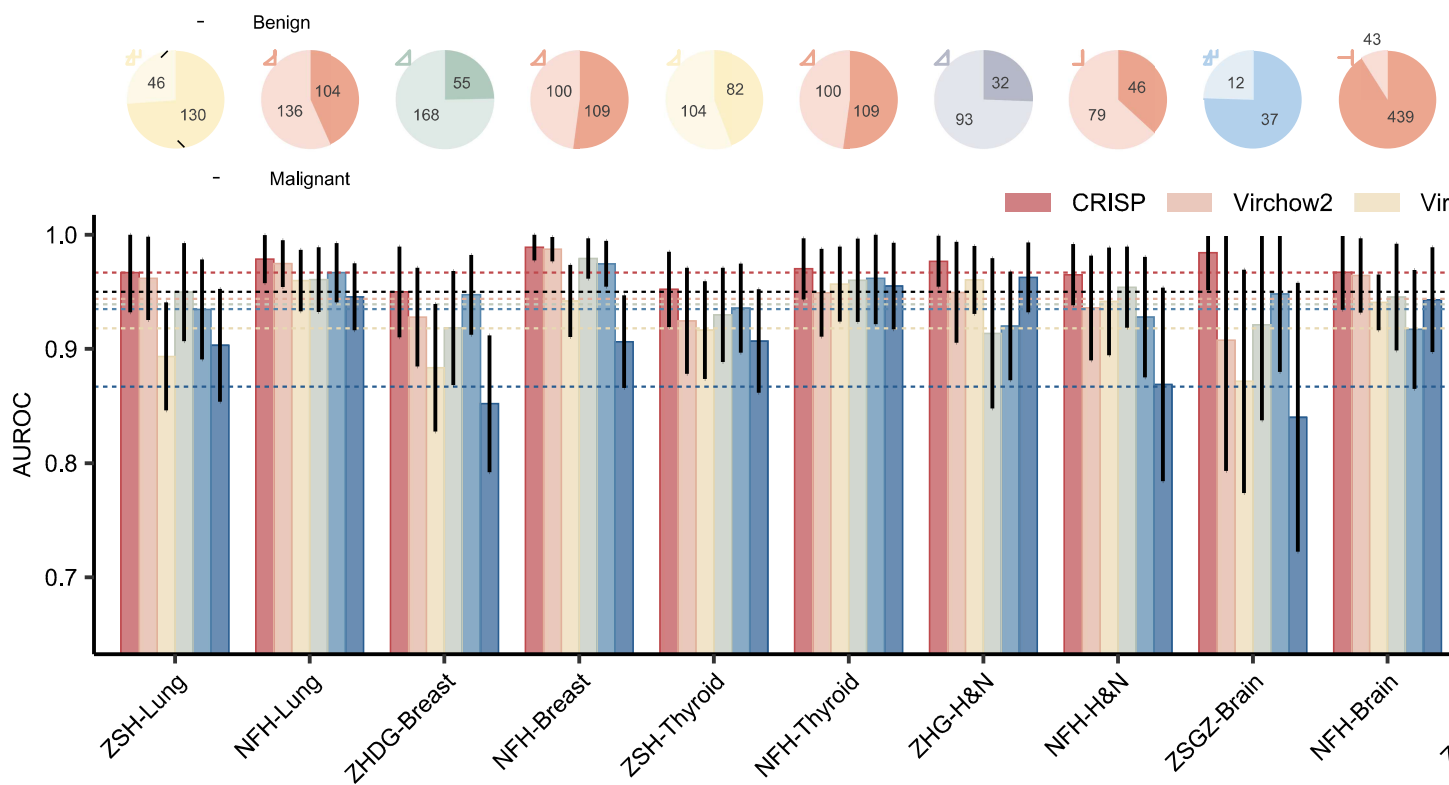
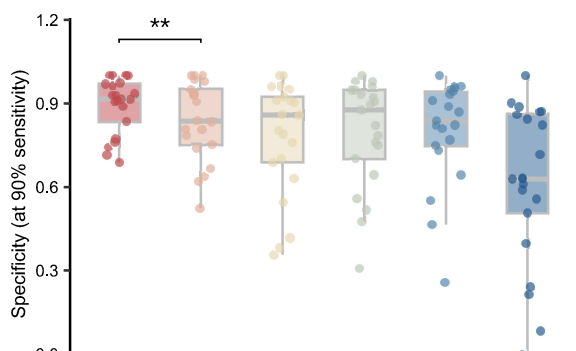
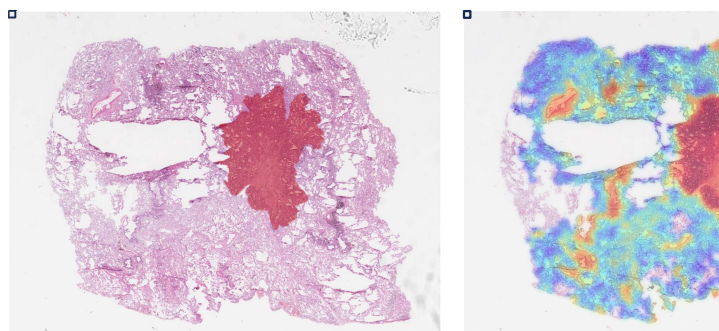


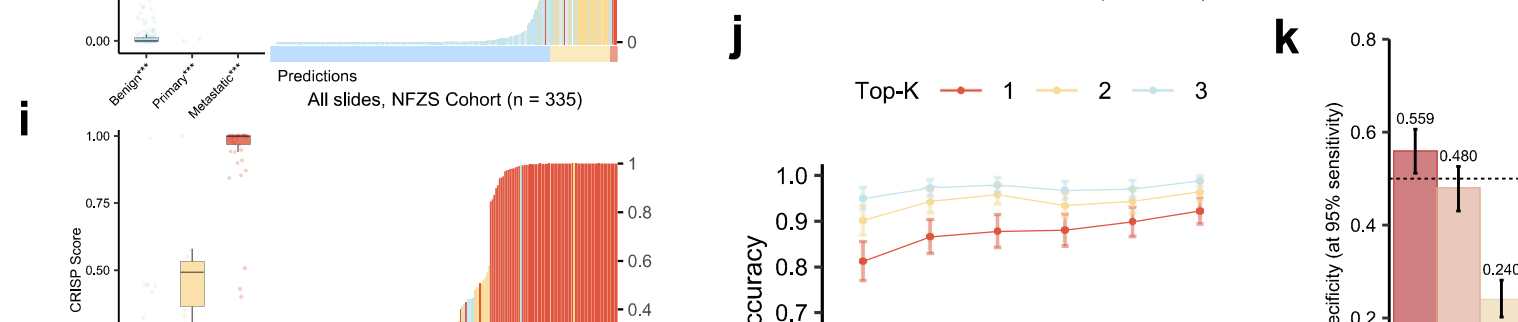
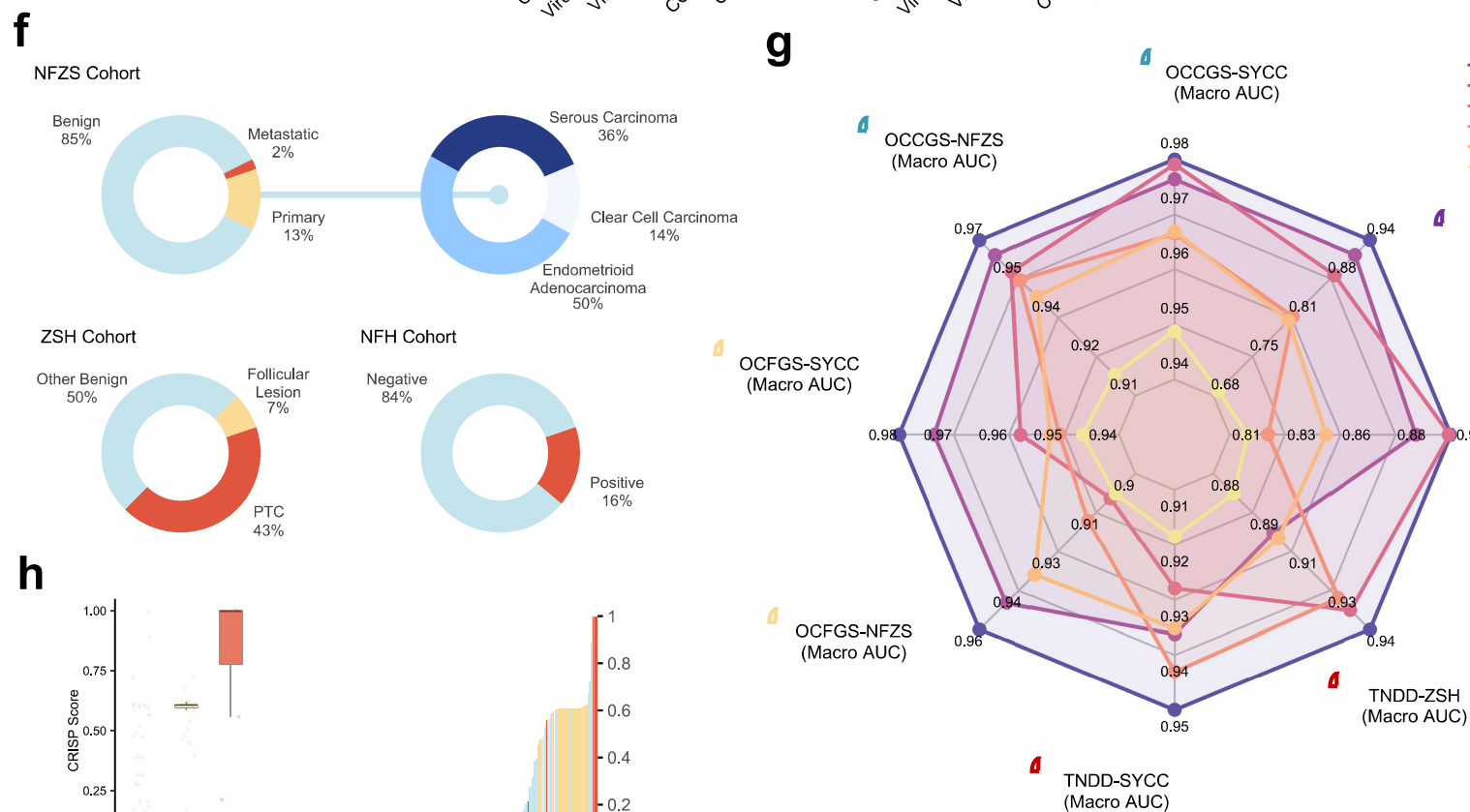
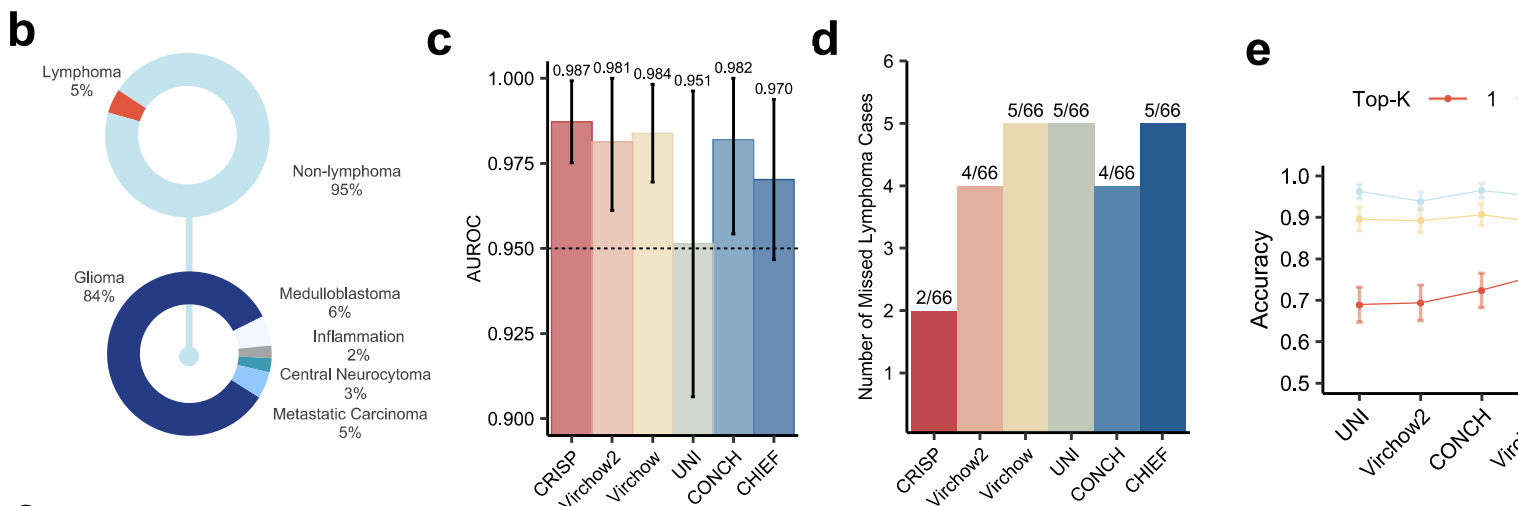
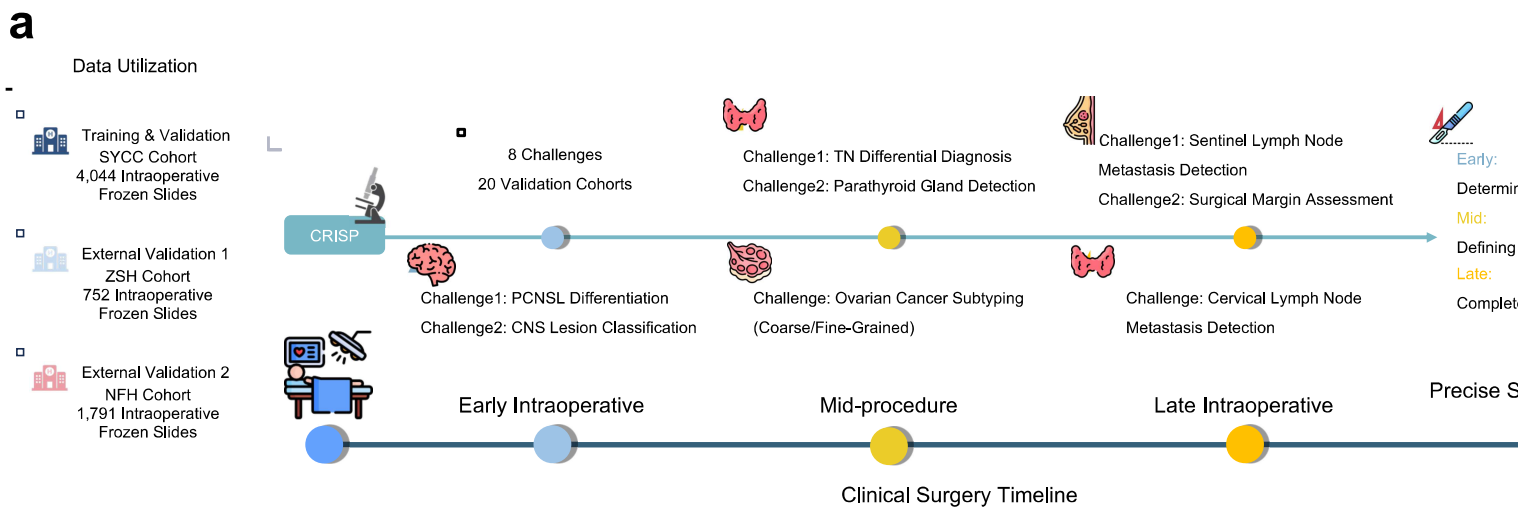
e

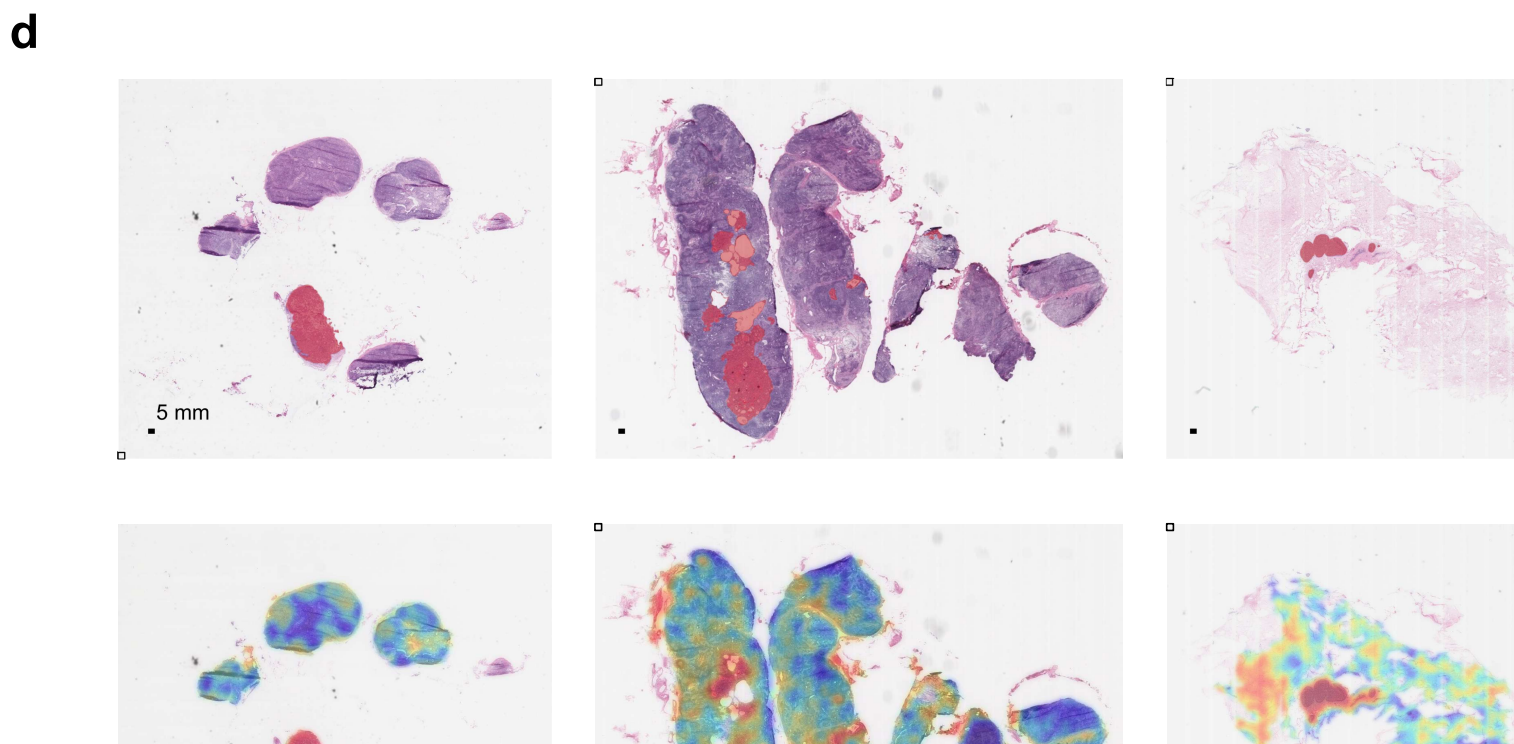
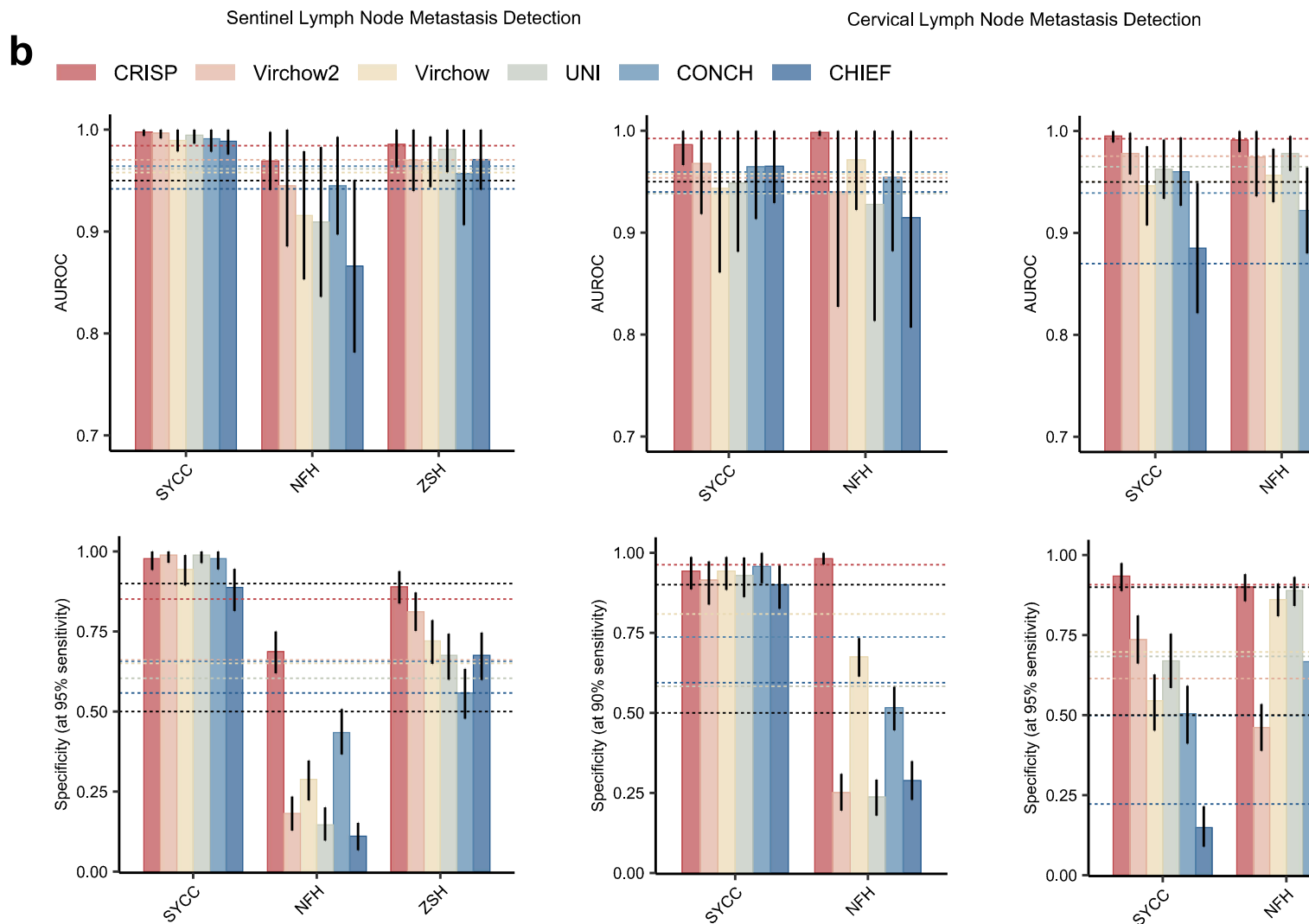
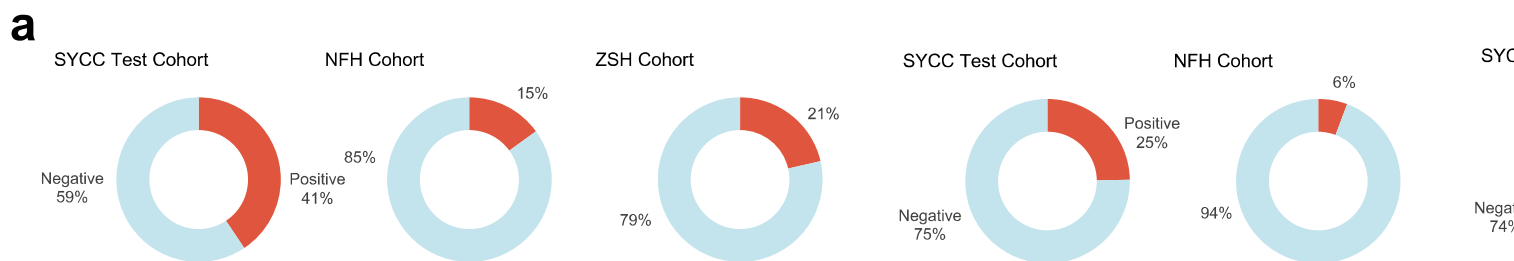


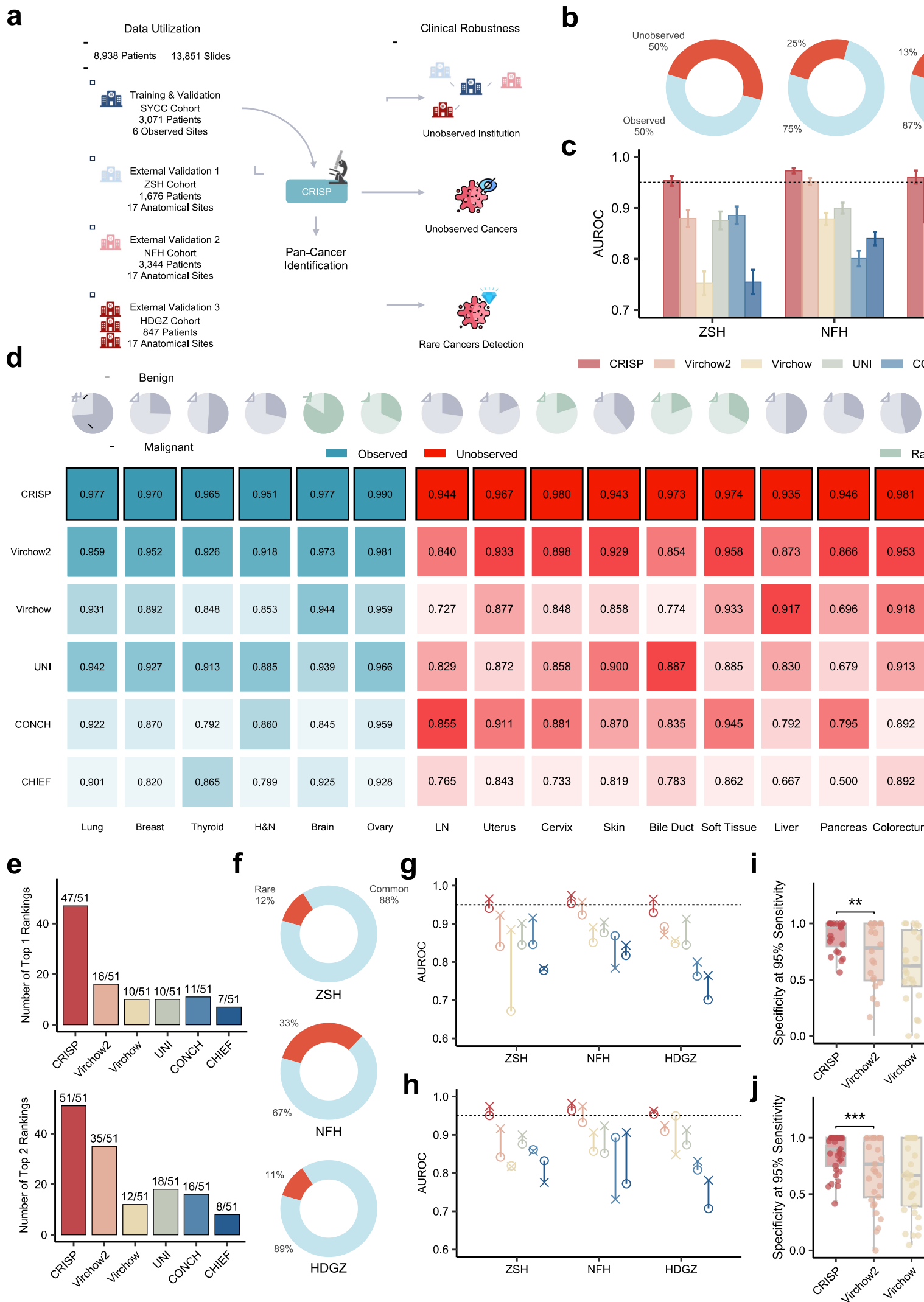
f

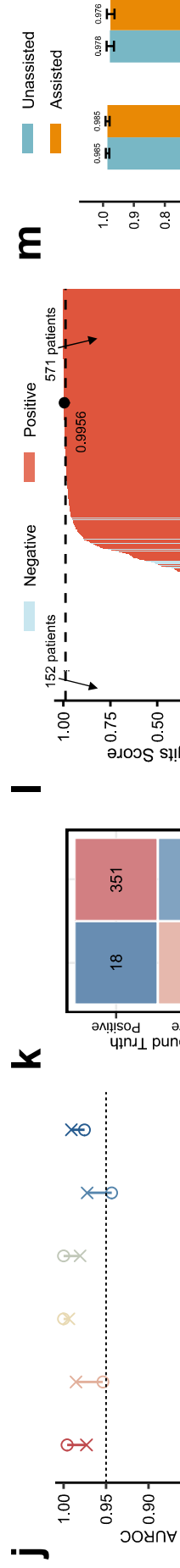
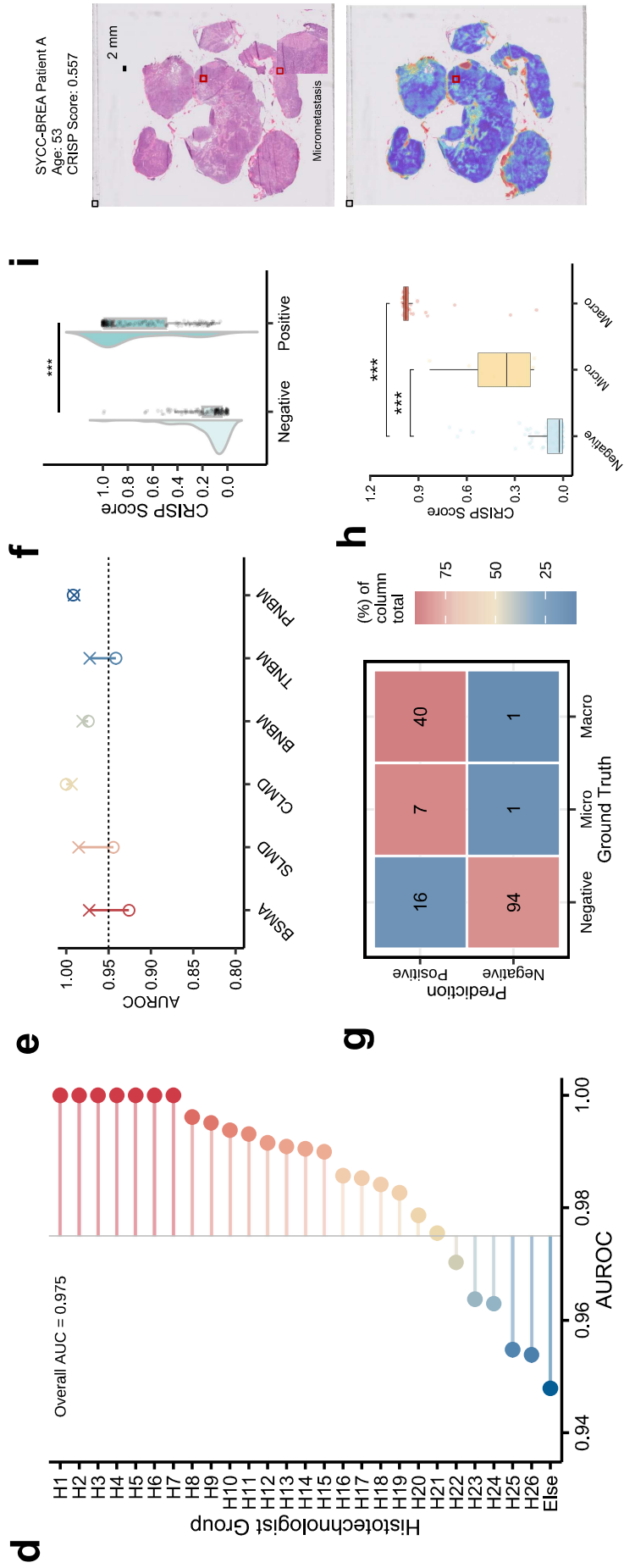
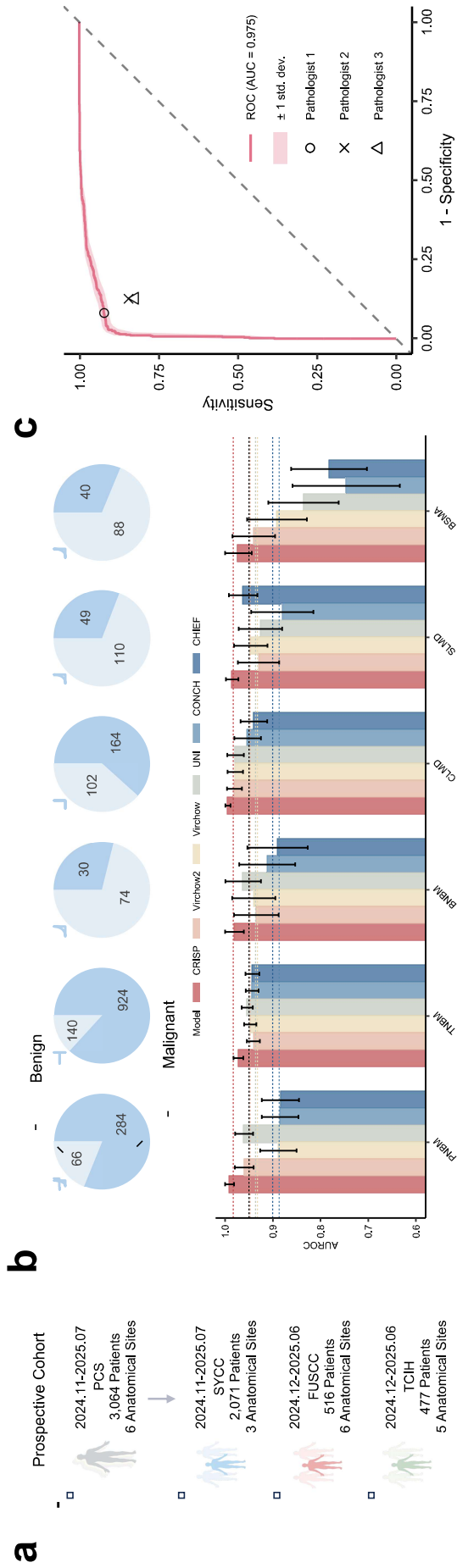


a**b****c****e****f****g**

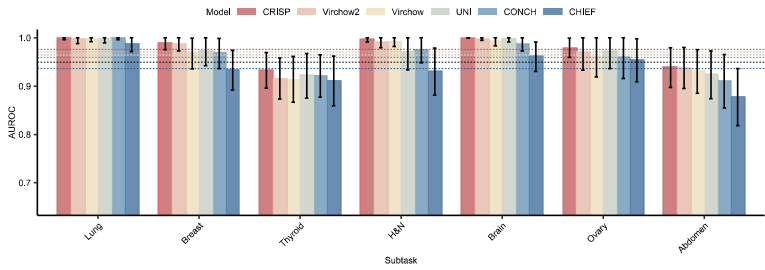




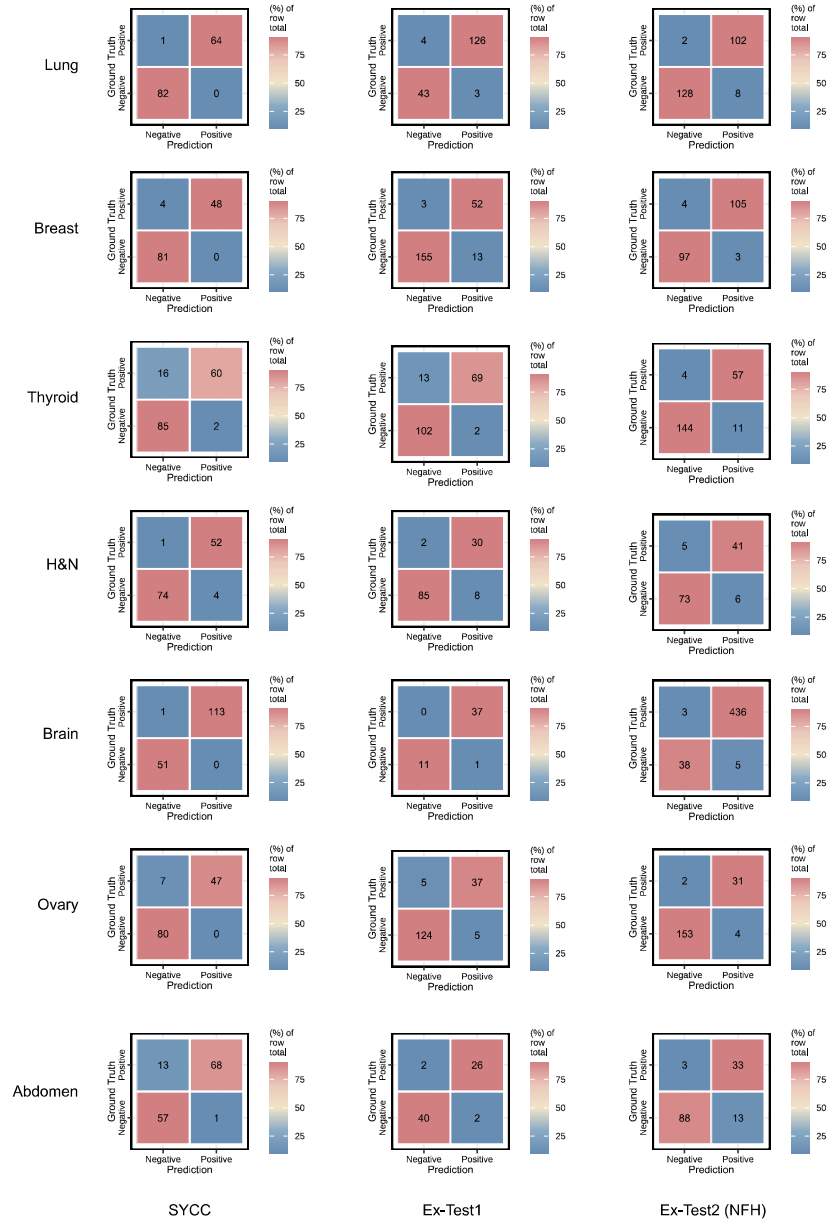


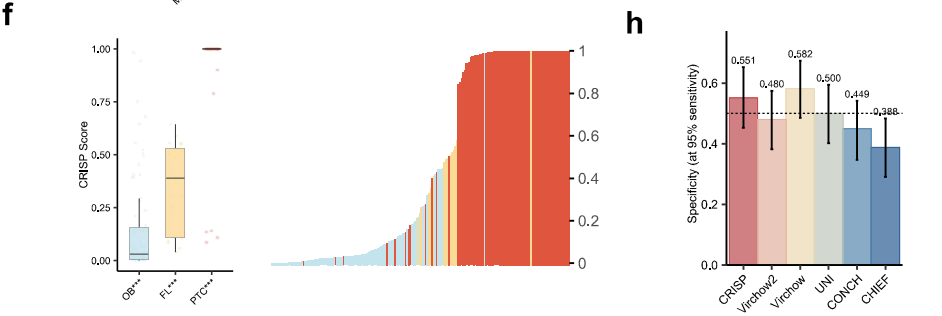
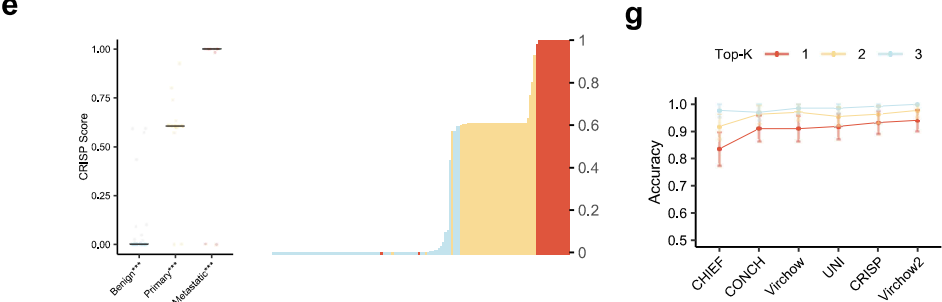
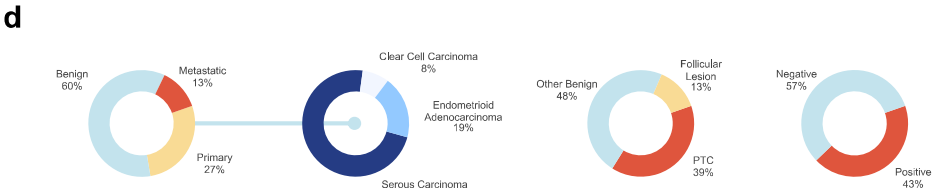
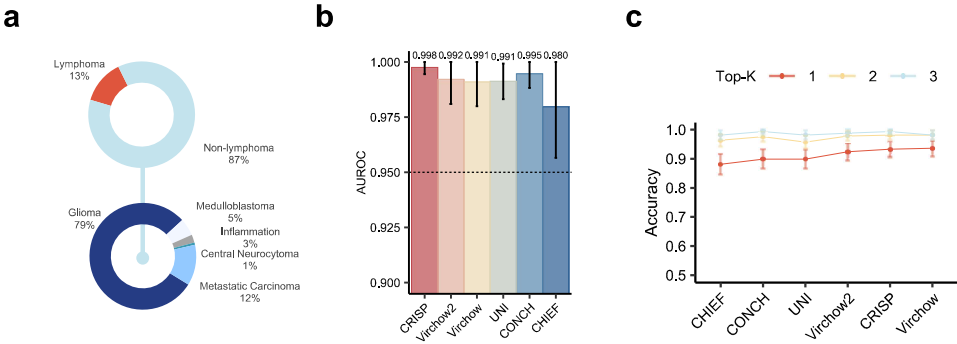


a



b

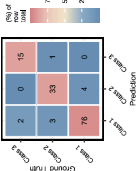




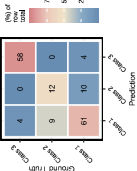
SYCC



PCNSLD

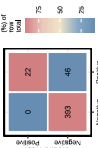


OCCGS

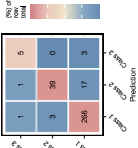


TNDD

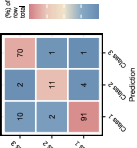
External-Test



CNSLC



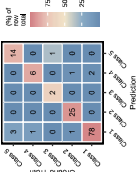
OCCFS



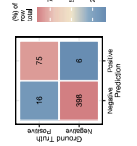
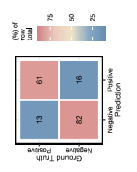
SYCC



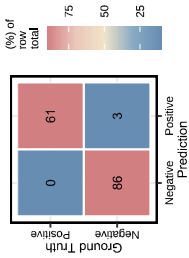
External-Test



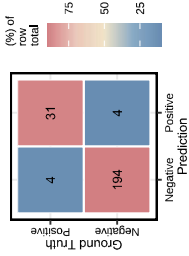
PGD



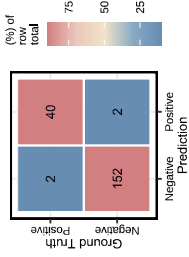
SYCC



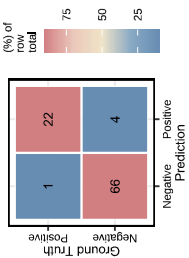
NFH



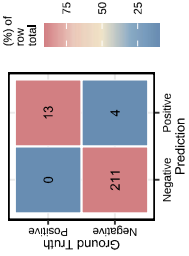
ZSH



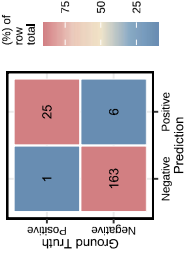
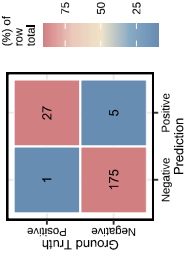
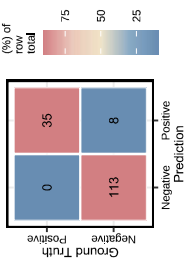
SLMD

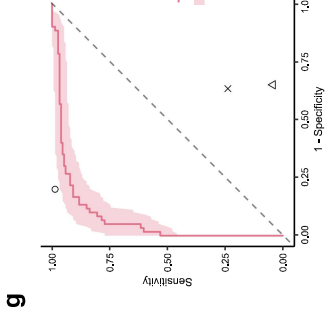
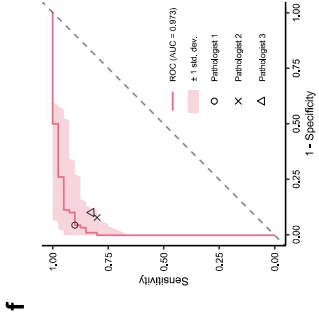
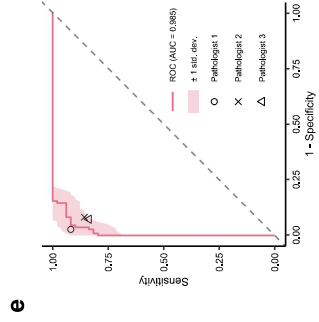
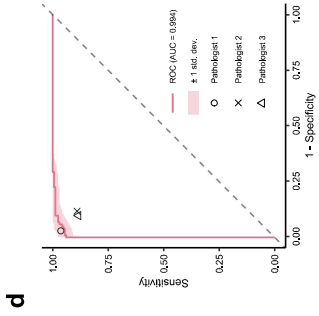
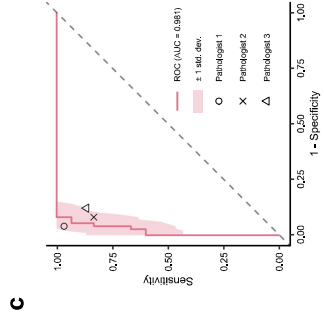
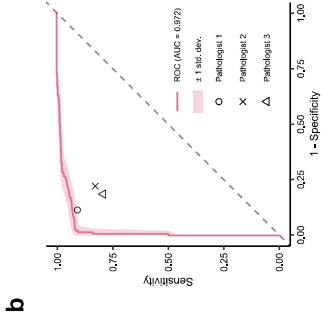
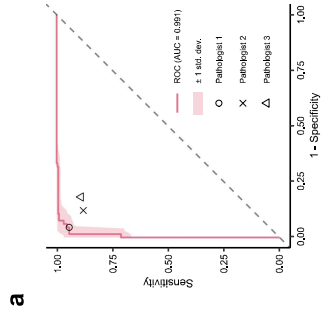


CLMD

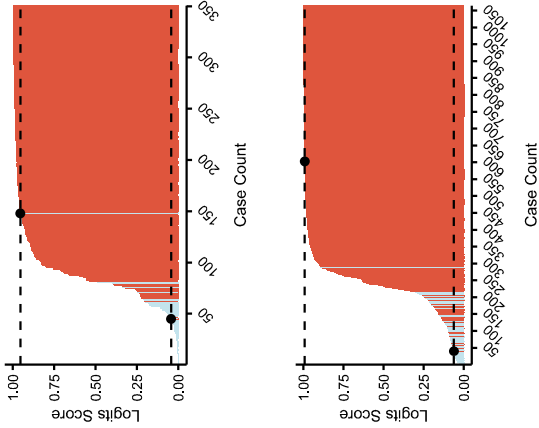


BSMA

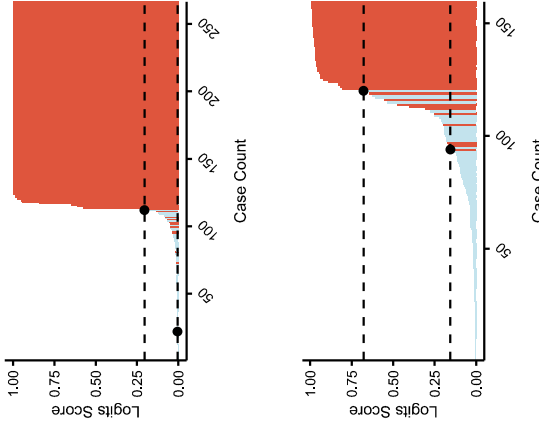




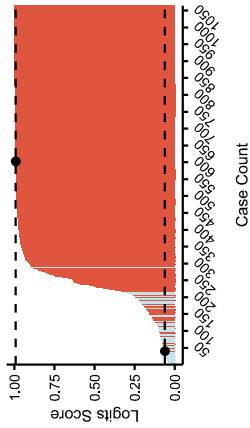
a



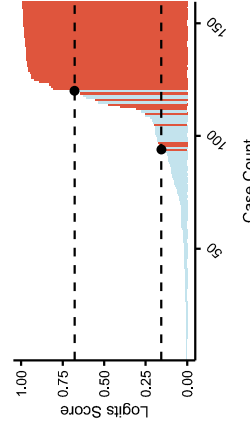
d



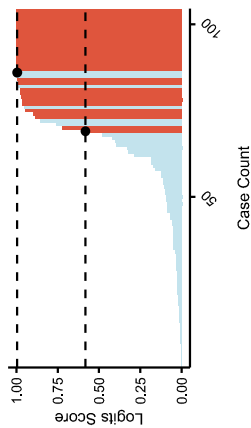
b



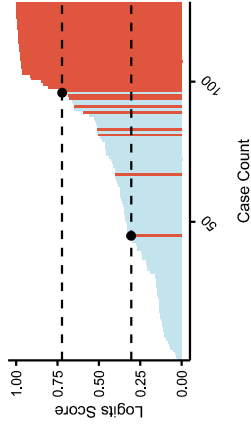
e



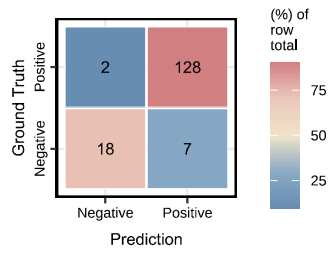
c



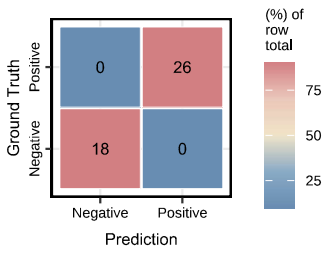
f



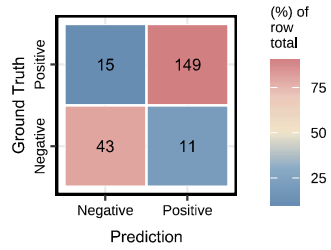
a



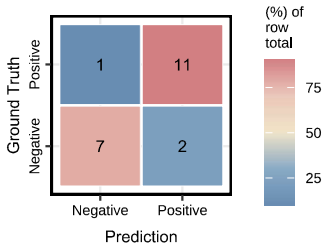
d



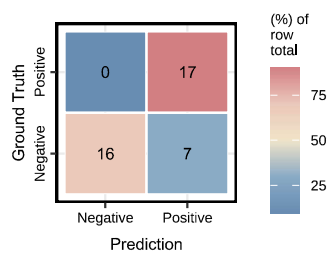
b



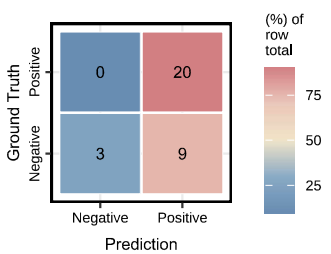
e



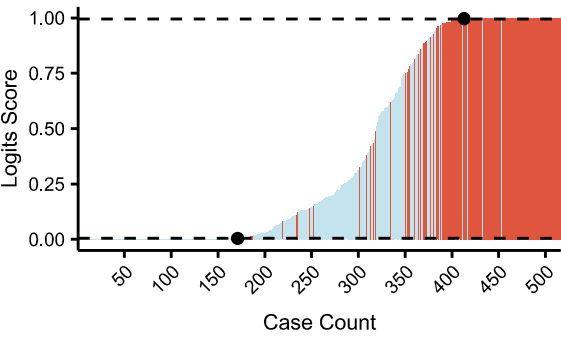
c



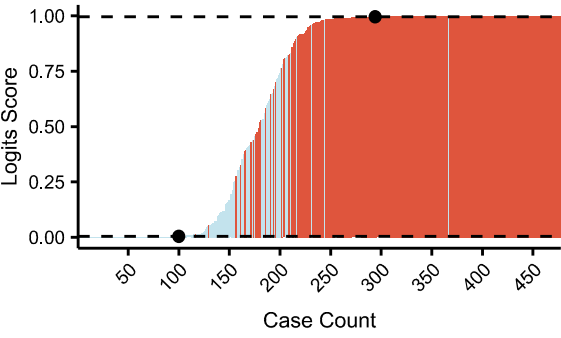
f



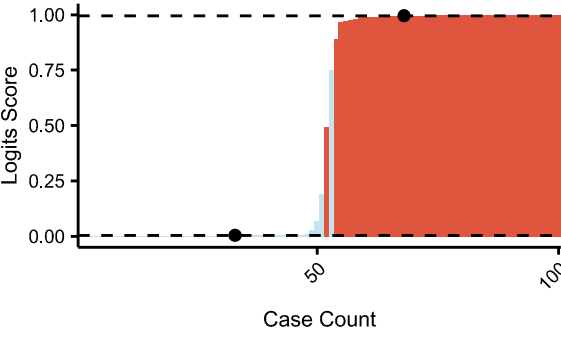
a

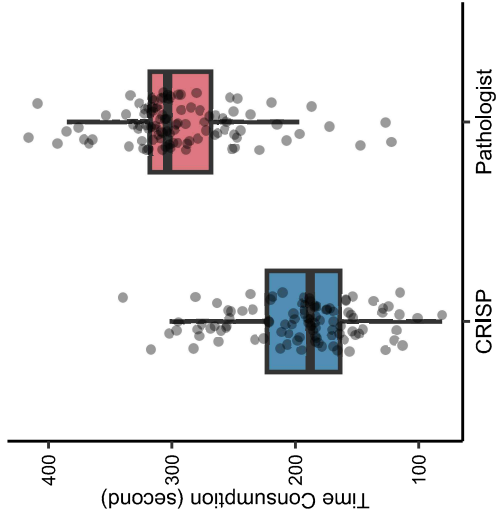


b

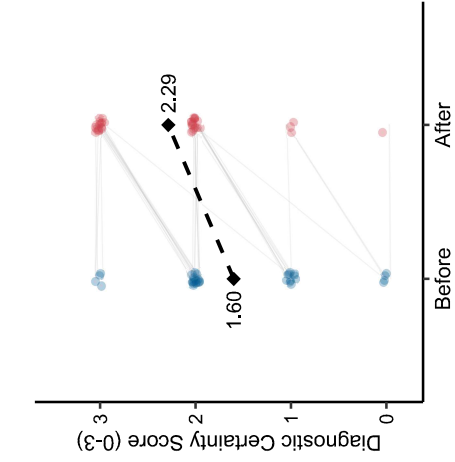
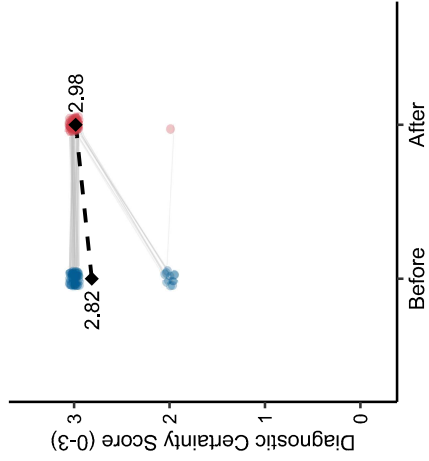


c





a



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
☐	☒ The exact sample size (<i>n</i>) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ 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>
☒	☐ A description of all covariates tested
☐	☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ 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)
☐	☒ 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>
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☐	☒ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒ 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 and process histopathological images, Python (version 3.13.5) was used along with the following libraries and tools: Aslide (https://github.com/MrPeterJin/ASlide) to handle various WSI formats, CLAM (https://github.com/mahmoodlab/CLAM) for tissue segmentation, and PyTorch (version 2.8.0) for model development.
Data analysis	We utilized Python (version 3.13.5) and PyTorch (version 2.8.0) to conduct model pretraining and downstream task evaluation. For compared foundation models, we strictly adhered to the original implementations as described in their respective publications to extract generalizable features. For task agnostic pretraining, we used 8 NVIDIA H800 GPUs, while downstream task training and evaluation were performed on NVIDIA 3090 GPUs. The vision transformer architecture was implemented using the timm library (version 1.0.16), and its weights were initialized with those from ViT. The peft library (version 0.15.2) was employed to implement LoRA modules. The weights and codes of compared foundation model were obtained from Hugging Face (https://huggingface.co/). Data analysis was conducted in R version 4.2.1 (2022-06-23) and Python 3.13.5 using the numpy (version 2.2.3), scipy (version 1.15.3) and scikit-learn (version 1.7.0) packages.

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](#)

The source code for the proposed framework is available at <https://github.com/FT-ZHOU-ZZZ/CRISP>. The pretrained model weights of CRISP are available at Hugging Face (<https://huggingface.co/FTZhou/CRISP>). The TCGA whole-slide images and associated clinical data can be accessed through the NIH Genomic Data Commons (<https://portal.gdc.cancer.gov>). Whole-slide images and clinical data from the internal and external cohorts are not publicly available owing to institutional policies and patient privacy regulations, but were collected under Institutional Review Board approvals. Access to these datasets for non-commercial academic use may be requested from the corresponding author (caimy@sysucc.org.cn). Requests will be evaluated and responded to within two weeks. A de-identified dataset comprising approximately 35,000 annotated patch-level images together with 75 corresponding whole-slide images from SYCC intraoperative frozen sections used in the linear probing experiments is publicly available to facilitate independent validation and future research (<https://huggingface.co/datasets/FTZhou/CRISP>). All other data supporting the findings of this study are available within the article and its Supplementary Information files.

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	Sex and gender characteristics of the study are available in Supplementary Tables S1.1-1.11. Sex or gender was not included as a covariate at any stage of our experimental analysis.
Reporting on race, ethnicity, or other socially relevant groupings	Race, ethnicity and other socially relevant groupings were not collected from the patients and were unrelated to model implementation or deployment.
Population characteristics	Population characteristics of this study are available in Supplementary Tables S1s.
Recruitment	Recruitment was not necessary for the retrospective datasets in this study. For the prospective real-world deployment study of CRISP, patients undergoing surgery with intraoperative consultation at Sun Yat-sen University Cancer Center, Fudan University Shanghai Cancer Center and Tianjin Medical University Cancer Institute and Hospital were recruited between November 2024 and July 2025.
Ethics oversight	This study was approved by the Ethics Committee of Sun Yat-sen University Cancer Center (Approval No. SL-B2024-708) and carried out in compliance with the Declaration of Helsinki.

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

Life sciences study design

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

Sample size	For model pre-training, we utilized 81,667 frozen section slides, generating over 50 million pathological image patches. An additional 22,130 slides were used for retrospective and prospective performance validation, resulting in a total of 103,797 frozen section slides. The sample size was determined by data availability.
Data exclusions	Cases were excluded if they met any of the following: (1) no available intraoperative frozen section; (2) absence of identifiable tissue on the slide; (3) presence of scanning artifacts such as defocus or blurring; or (4) incomplete diagnostic or clinical information.
Replication	Additional replication was unnecessary, as the models were validated using five independent medical center cohorts and three prospective cohorts, demonstrating consistent performance across all validation sets.
Randomization	Samples were randomly split into development and validation datasets. The splits were stratified by class to ensure similar class proportions across splits. For patients with multiple slides, all slides from the same patient were assigned to the same dataset to maintain patient-level independence.
Blinding	All data were anonymized prior to processing to ensure the removal of any patient-identifiable information.

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
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☐	☒ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Clinical data

Policy information about [clinical studies](#)

All manuscripts should comply with the ICMJE [guidelines for publication of clinical research](#) and a completed [CONSORT checklist](#) must be included with all submissions.

Clinical trial registration	The study was registered in the Chinese Clinical Trial Registry (ChiCTR; http://www.chictr.org.cn/) under the registration number ChiCTR2500106350.
Study protocol	Study protocols can be found in http://www.chictr.org.cn/ .
Data collection	For the prospective real-world deployment study of CRISP, patients undergoing thyroid, breast, or lung surgery with intraoperative consultation at Sun Yat-sen University Cancer Center were consecutively recruited between November 2024 and July 2025, representing the most common sites for intraoperative consultation. Inclusion criteria were: (1) at least one available intraoperative frozen section; (2) presence of identifiable tissue on the slide; (3) absence of scanning artifacts such as defocus or blurring; (4) complete diagnostic and clinical information; (5) age ≥ 18 years at the time of surgery; and (6) documented informed consent. Following screening, a total of 2,071 patients with 4,143 intraoperative frozen sections were included. To further assess geographical generalizability, two additional prospective cohorts from FUSCC and TCIH were included, with the same inclusion criteria applied. To simulate real-world application, 100 consecutive cases from SYCC were included for simulated clinical deployment analysis. Recruitment periods and detailed clinical characteristics for each cohort are provided in Supplementary Tables S1.9-1.11.
Outcomes	The primary outcome of this study was the patients' diagnostic information. The diagnostic reference standard for each patient was established based on the postoperative FFPE pathology report in combination with immunohistochemistry and molecular test results, and independently verified by a board-certified pathologist blinded to the study.

Plants

Seed stocks	NA
Novel plant genotypes	NA
Authentication	NA