Molecular-driven Foundation Model for Oncologic Pathology

Anurag Vaidya[1,2,3,4,‡], Andrew Zhang[1,2,3,4,‡], Guillaume Jaume[1,2,3,‡], Andrew H. Song[1,2,3,*], Tong Ding[1,2,3,5,*], Sophia J. Wagner[1,6,7,*], Ming Y. Lu[1,2,3,8], Paul Doucet[1], Harry Robertson[1,9], Cristina Almagro-Pérez[1,2,3,4], Richard J. Chen[1,2,3], Dina ElHarouni[3,10], Georges Ayoub[3,10], Connor Bossi[3,10], Keith L. Ligon[1,3,10,11], Georg Gerber[1], Long Phi Le[2,+], Faisal Mahmood[1,2,3,12,+]

[1]*Department of Pathology, Brigham and Women's Hospital, Harvard Medical School, Boston, MA*
[2]*Department of Pathology, Massachusetts General Hospital, Harvard Medical School, Boston, MA*
[3]*Cancer Program, Broad Institute of Harvard and MIT, Cambridge, MA*
[4]*Health Sciences and Technology, Harvard-MIT, Cambridge, MA*
[5]*Harvard John A. Paulson School of Engineering and Applied Sciences, Harvard University, Cambridge, MA*
[6]*Helmholtz Munich – German Research Center for Environment and Health, Munich, Germany*
[7]*School of Computation, Information and Technology, TUM, Munich, Germany*
[8]*Electrical Engineering and Computer Science, Massachusetts Institute of Technology (MIT), Cambridge, MA*
[9]*Sydney Precision Data Science Center, The University of Sydney, Camperdown, New South Wales, Australia*
[10]*Department of Oncologic Pathology, Dana-Farber Cancer Institute, Boston, MA 02215, USA*
[11]*Department of Pathology, Boston Children's Hospital, Boston, MA 02115, USA*
[12]*Harvard Data Science Initiative, Harvard University, Cambridge, MA*
‡ *Co-first authors,* ∗ *Co-second authors,* + *Co-senior authors*
Lead Contact*: Faisal Mahmood (faisalmahmood@bwh.harvard.edu)*

Foundation models are reshaping computational pathology by enabling transfer learning, where models pre-trained on vast datasets can be adapted for downstream diagnostic, prognostic, and therapeutic response tasks. Despite these advances, foundation models are still limited in their ability to encode the entire gigapixel whole-slide images without additional training and often lack complementary multimodal data. Here, we introduce THREADS, a slide-level foundation model capable of generating universal representations of whole-slide images of any size. THREADS was pretrained using a multimodal learning approach on a diverse cohort of 47,171 hematoxylin and eosin (H&E)-stained tissue sections, paired with corresponding genomic and transcriptomic profiles—the largest such paired dataset to be used for foundation model development to date. This unique training paradigm enables THREADS to capture the tissue's underlying molecular composition, yielding powerful representations applicable to a wide array of downstream tasks. In extensive benchmarking across 54 oncology tasks, including clinical subtyping, grading, mutation prediction, immunohistochemistry status determination, treatment response prediction and survival prediction THREADS outperformed all baselines while demonstrating remarkable generalizability and label efficiency. It is particularly well-suited for predicting rare events, further emphasizing its clinical utility. We intend to make the model publicly available for the broader community.
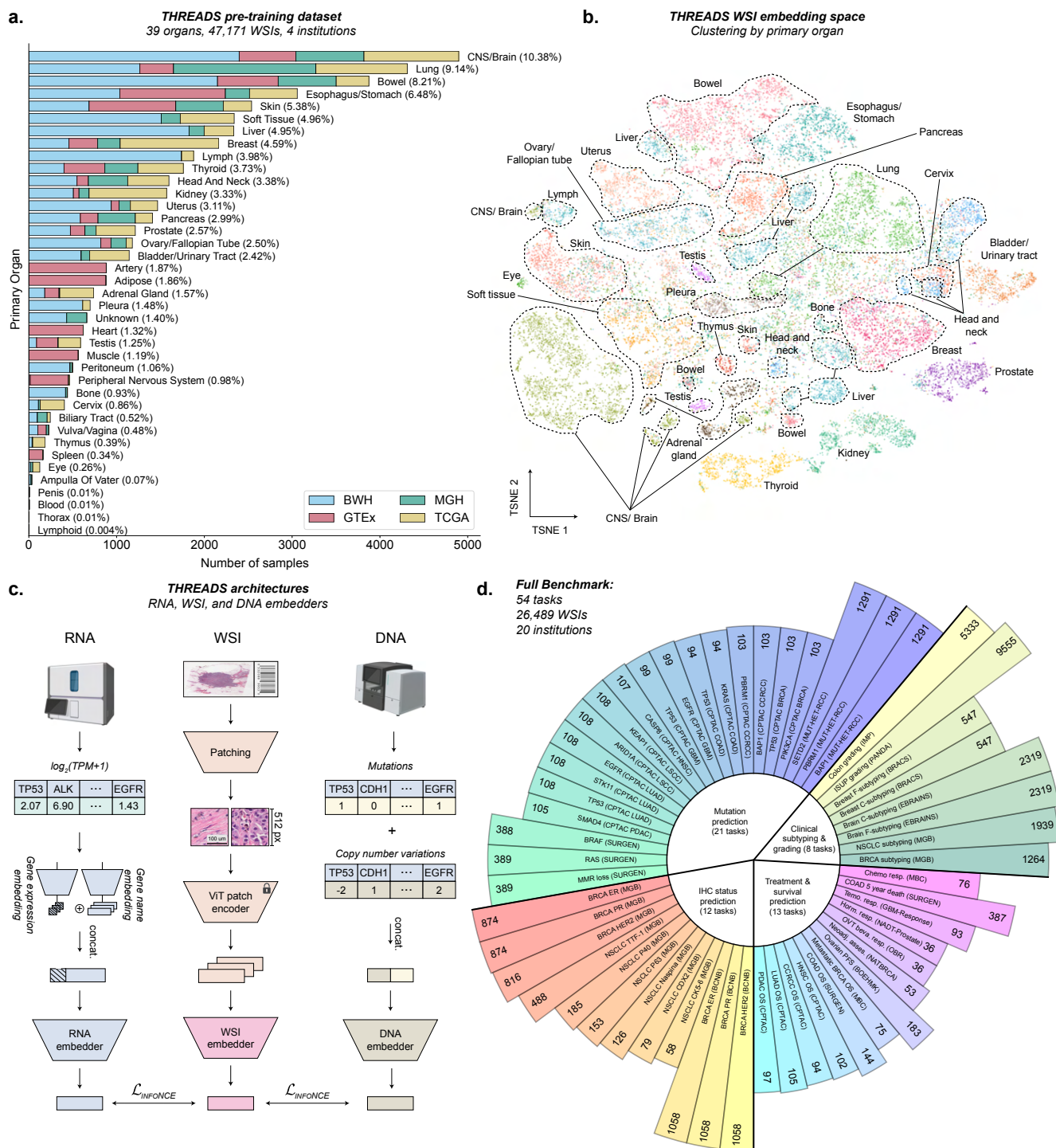
Introduction

With the advancement of precision medicine and targeted therapies, problem statements in oncology focus increasingly on rare conditions and targeted populations. As research questions become more specific, the assumption of data abundance, which underpinned early successes in AI for pathology[1,2], such as Gleason grading[3] and metastasis detection[4], no longer applies. Many current problem statements in oncology, especially for patient prognostication, and treatment response prediction, involve small patient cohorts, frequently fewer than 100 patients. The limitation of data scarcity is compounded by the size of digitized tissue sections (whole-slide images, WSIs), which can be several gigabytes each. Consequently, most computational pathology predictive models operate in a scenario where the input data size vastly exceeds the number of available samples for model training, making model training incredibly complex.

In response to these challenges, numerous foundation models specifically designed for pathology have been developed[5–7]. These models enable transfer learning from large pretraining data with hundreds of thousands of WSIs and billions of cells to narrow applications, such as biomarker prediction. However, most of these models are patch encoders and, by design, are restricted to encoding small regions of interest, orders of magnitude smaller than clinical whole-slide imaging data, which can be several gigabytes. Existing models address the limitation by training an additional model, which can be computationally expensive to train and may require a lot of downstream labels. Addressing this limitation is critical for advancing foundation models in pathology to more varied tasks and overcoming the data abundance requirement. Some models have explored whole-slide image representation learning to derive off-the-shelf slide embeddings that can be used for various downstream tasks at minimal cost[6,8–11]. However, they remain limited in scope by the diversity of training data with organ and disease-specific models[9,10], and the predictive capabilities of the resulting representations.

Here, we introduce a new foundation model for pathology, THREADS, a general-purpose encoder model that can generate WSI embeddings. THREADS was pretrained through multimodal contrastive learning, where molecular profiles obtained with next-generation sequencing are used as a guide for learning the slide representation. We posit that molecular data brings a holistic and unbiased view of the tissue morphology that encapsulates biologically and clinically relevant information[9,10]. To train THREADS, we assembled the most extensive multimodal training dataset to date, named MBTG-47K, consisting of more than 47,000 samples. Each sample includes a WSI and its corresponding molecular profile obtained from an adjacent tissue section (**Figure 1.a**). MBTG-47K was curated from Massachusetts General Hospital (MGH, 6,899 samples, or 14.6%), Brigham and Women's Hospital (BWH, 20,556 samples or 43.6%), The Cancer Genome Atlas Program (TCGA, 10,209 samples or 21.6%), and the Genotype–Tissue Expression (GTEx, 9,507 samples or 20.2%) consortium[12] (**Extended Data Table 1**). This pretraining strategy leads to a slide embedding space that encodes rich information about tissue morphology, disease, and composition (**Figure 1.b**).

We validate THREADS across a wide range of tasks in oncology, covering clinical tasks for cancer subtyping and grading, gene mutation prediction, immunohistochemistry status prediction, and treatment response and survival prediction (**Figure 1.d**). In total, our model is evaluated on 54 pathology tasks from 23 cohorts across 17 different sources. THREADS achieves state-of-the-art performance, significantly outperforming three whole-slide encoder models PRISM [13] (P-value<0.001), GIGAPATH [6] (P-value<0.001) and CHIEF [11] (P-value<0.001), and attention-based multiple instance learning classification baselines (P-value<0.001). THREADS can also serve as an effective initialization for additional model fine-tuning, which brings a significant improvement over training a model from scratch (P-value<0.001). This establishes THREADS as a foundational model that can drive AI advancements in histopathology.

Figure 1: **Study overview. a.** Tissue site distribution of MBTG-47K used for THREADS pretraining. **b.** 2-dimensional tSNE [14] representation of THREADS WSI embedding space on MBTG-47K colored by primary organ. Each point is a WSI. **c.** Block diagram of THREADS architecture for WSI representation learning. **d.** Overview of THREADS downstream evaluation composed of 54 tasks. Tasks are grouped into four families: clinical subtyping and grading (n=8 tasks), gene mutation prediction (n=21 tasks), immunohistochemistry status prediction (n=12 tasks), and treatment response and survival prediction (n=13 tasks). WSI: whole-slide image; tSNE: t-distributed stochastic neighbor embedding; MGH: Mass General Hospital; BWH: Brigham and Women's Hospital.

Results

Whole-slide image classification with THREADS

THREADS design. THREADS consists of two components. An ROI encoder model (CONCHV1.5 [15]) consisting of a Vision Transformer-Large[16,17] (ViT-L) model trained on millions of image patches via multimodal learning between ROIs and text captions; and a slide encoder that aggregates tile embeddings into a slide representation using attention-based modeling (**Figure 1.c** and **Extended Data Figure 1**). We use two types of next-generation sequencing (NGS) data for THREADS pretraining: transcriptomic profiles obtained with bulk RNA sequencing (MGH, TCGA, and GTEx samples), and genomic profiles capturing single nucleotide variants (SNVs), insertions and deletions (indels), and copy number variants (CNVs) of a targeted gene panel (BWH samples). The transcriptomic profiles are encoded using a single-cell foundation model pretrained on 5.7 million cells of various cancer types[18], and the genomic profiles using a multi-layer perceptron model[19]. We employ cross-modal contrastive learning to align the slide representation with the corresponding molecular embedding. Additional information is provided in the **Online Methods, Pretraining dataset curation**.

Downstream evaluation. We propose a large benchmark for assessing foundation models in hematoxylin and eosin (H&E)-stained whole-slide imaging. Our evaluation includes 54 tasks from 23 different cohorts, covering four families of tasks: clinical subtyping and grading (n=8 tasks, 20,427 WSIs), gene mutation prediction (n=21 tasks, 3,503 WSIs), immunohistochemistry (IHC) status prediction (n=12 tasks, 2,469 WSIs), and patient prognostication including treatment response and survival prediction (n=13 tasks, 2,857 WSIs). We curated tasks from a set of in-house data (n=12 tasks including 3,550 WSIs from three cohorts) and publicly available data (n=42 tasks including 23,161 WSIs from 17 cohorts). The diversity of tasks makes our benchmark suitable for assessing the predictive performance of slide encoders under different scenarios, from well-established clinical tasks with data abundance, such as colorectal cancer grading and breast cancer subtyping, to specific problem statements in treatment response prediction typically characterized by small patient cohorts. Our evaluation constitutes, to date, the most comprehensive benchmark introduced in computational pathology. All tasks follow a unified evaluation with either five-fold cross-validation into 80:20 splits or 50-train-test splits, depending on cohort size with label- and patient-stratified splits. An overview of each evaluation task is provided in **Figure 1.d**, with additional descriptions in the **Online Methods**, **Downstream tasks and datasets**, statistics for each task in **Extended Data Table 2,3,4,5,6,7**, and links to access public cohorts detailed in **Extended Data Table 8**.

Baselines. We compare THREADS against three foundation models for encoding WSIs: PRISM, GIGAPATH, and CHIEF. PRISM is based on the Virchow[7] patch encoder (ViT-Huge, 632 million parameters) followed by a Perceiver model[20] (45 million parameters) pretrained using contrastive learning with matched patient-level pathology reports (195,344 specimens). GIGAPATH is based on a ViT-Giant patch encoder (1.13 billion parameters) pretrained on 171,000+ WSIs (>30,000 patients) using DINOv2[21], and a LongNet[22] slide encoder pretrained using masked autoencoding. Finally, CHIEF is based on the CTransPath[23,24] patch encoder (Swin-Transformer, 28 million parameters) followed by an attention-based multiple instance learning (AB-MIL) model[25] pretrained on 60,530 WSIs using contrastive learning with the tissue site. Additional information is provided in the **Online Methods** and **Baselines**.

We evaluate THREADS and baselines using linear probing (i.e., by learning a logistic regression model) to classify slide embeddings into the downstream task label. To prevent overfitting, we avoid hyperparameter tuning and set a fixed cost, number of iterations, and solver in linear probe models. We evaluate model performance using the area under the receiver operating characteristic (AUC) for all binary classification tasks,

quadratic Cohen's kappa for grading, balanced accuracy for multi-class clinical subtyping, and concordance-index (c-index) for survival tasks. We provide additional information in the **Online Methods**, **Evaluation metrics**.

Linear probing. THREADS provides state-of-the-art performance in linear probing evaluation. THREADS leads to an absolute performance gain over PRISM, GIGAPATH, and CHIEF of 6.3%, 9.9%, 6.7%, respectively (**Figure 2.a**). Employing a mixed-effects statistical model to compare the overall performance (**Online Methods, Statistical analysis**), we showed that THREADS significantly outperformed PRISM (P<0.001), CHIEF (P<0.001), and GIGAPATH (P<0.001). When investigating each family of tasks, THREADS demonstrates absolute performance improvements of 2.1% in clinical subtyping and grading over PRISM, the second-best model (P<0.001, **Figure 2.b**), 6.1% in mutation prediction over CHIEF (P<0.001, **Figure 2.c**), 4.6% in IHC status prediction over PRISM (P<0.01, **Figure 2.d**), and 8.9% in prognostication over PRISM (P<0.001, **Figure 2.e**). At the individual task level, THREADS outperforms PRISM in 44/54 tasks, CHIEF in 49/54 tasks, and GIGAPATH in 54/54 tasks. The performance per task is detailed in **Extended Data tables 9 to 37**.

When investigating performance in diagnostic tasks (cancer subtyping and grading), THREADS with a linear model achieves performance levels that are competitive with specialist models[3,26]. For instance, THREADS reaches 98.3% AUC in breast cancer subtyping (invasive lobular carcinoma *vs.* invasive ductal carcinoma), and 98.2% AUC in lung cancer subtyping (lung adenocarcinoma *vs.* lung squamous cell carcinoma). In comparison, an attention-based MIL model[25] trained on the same data with >1.5 million parameters reaches 98.3% and 98.0%, respectively (**Extended Data Table 9 and 10**). In colorectal cancer grading, THREADS with linear probing reaches 91.9% quadratic Cohen's kappa score, just 2.3% lower than training a dedicated ABMIL model (94.2%) (**Extended Data Table 13**). These findings underscore the extensive capabilities of THREADS in providing rich slide representations for clinical use.

To validate the superior performance of THREADS in linear probing evaluation, we additionally benchmarked THREADS and baselines with varying the regularization cost (**Extended Data Figure 3 and Extended Data Table 38**). THREADS provides significant performance gain over all baselines for all the regularization costs explored: 4.3% absolute performance gain over PRISM (second-best performer) with large regularization (C=0.01), and 5.7% over CHIEF (second-best performer) with small regularization (C=10). THREADS is also less sensitive to changes in regularization than baselines across all regularization strengths, showing the model's robustness and versatility.

Transferability of THREADS. We also investigated whether linear models trained with THREADS embedding show generalization properties when tested in external cohorts. To this end, we selected a subset of tasks from our evaluation pipeline for which we have an external test set. Specifically, we first train a linear probe classifier on the entirety of one dataset and test on the entirety of the external set. We studied transferability in six different types of cancer: prediction of BAP1 and PRMB1 mutations in clear cell renal cell carcinoma, IDH mutation in a cohort of glioblastoma and low-grade glioma, prediction of ER/PR status in invasive breast cancer, subtyping in lung cancer, and survival prediction in pancreatic adenocarcinoma. Overall, THREADS provides strong transferability properties that lead to significantly better performance than PRISM (P-value<0.001), GIGAPATH (P-value<0.001), and CHIEF (P-value<0.001) as shown in **Extended Data Figure 5** and **Extended Data Table 39**. THREADS outperforms all baselines in 8/9 tasks (task-wise P-values <0.001 in 7/9 tasks in comparison to the second-best baseline). In lung and breast cancer subtyping, THREADS preserves a high predictive performance of 98.4% and 96.5% AUC, respectively, on the external cohort. Similarly, in ER and PR status prediction, THREADS leads to 88.5% and 79.4% AUC, maintaining high predictive performance. These results highlight the ability of THREADS to capture clinically and biologically relevant information without overfitting on cohort- and institution-specific features.
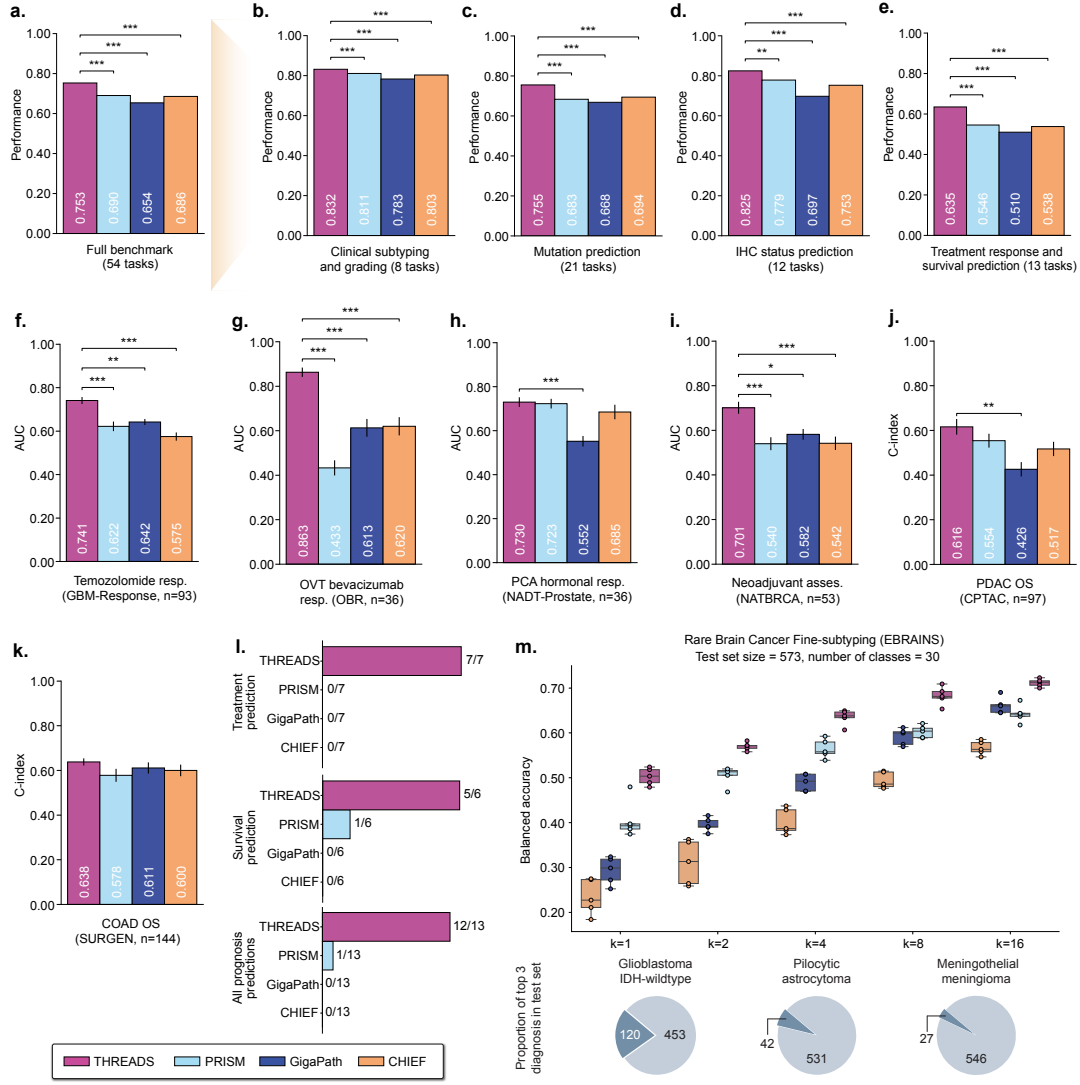
Data and label efficiency of THREADS

As pathology and oncology progress, problem statements become increasingly specific, resulting in scenarios that inherently face constraints in data availability. Such limitations are particularly prevalent in predicting patient treatment response and resistance. As part of our evaluation pipeline, we curated six treatment prediction tasks and one treatment assessment task covering several cancer types. Specifically, we tested THREADS to predict response in patients treated with temozolomide in glioblastoma (93 patients, **Figure 2.f**), bevacizumab in ovarian cancer (36 patients, **Figure 2.g**), hormonal therapy in prostate adenocarcinoma (53 patients, **Figure 2.h**), and neoadjuvant chemotherapy in high-grade serous ovarian cancer (183 patients, **Extended Data Table 37**), and platinum (and taxane for a subset) in ovarian metastasis of metastatic breast cancer (75 patients, **Extended Data Table 36**). THREADS also provides a tool for response assessment to detect signs of vascular invasion in patients with breast cancer treated with neoadjuvant chemotherapy (53 WSIs, **Figure 2.i**). THREADS provides better performance than baselines in all seven tasks, overall significantly outperforming baselines as assessed using a mixed-effect model statistical analysis (P-value<0.001, **Figure 2.l**). When considering individual tasks, THREADS significantly outperforms all baselines in 4/7 (P-value<0.05).

THREADS embeddings can also be used for patient survival prediction. We employ this approach for predicting overall survival in patients with pancreatic adenocarcinoma (**Figure 2.k**), colon and rectum adenocarcinoma (**Figure 2.l**), clear cell renal cell carcinoma (**Extended Data Table 33**), head and neck squamous carcinoma (**Extended Data Table 34**), and lung adenocarcinoma (**Extended Data Table 32**). Across all six survival tasks from our evaluation, THREADS provides the best predictive performance in five of them, overall providing significantly better performance than all baselines (P-value<0.001, **Figure 2.m**). The survival analysis using Kaplan Meier estimators also reveals the superior stratification capabilities of THREADS, which provide better separation between groups of patients considered as low and high risk than all baselines (**Extended Data Figure 6**).

To complement our analysis in data-scarce problem statements, we benchmarked THREADS and baselines in few-shot learning experiments, where we monitor the test performance when training on an increasing number of samples: $k = 1, 2, 4, 8, 16, 32$, where k is the number of training samples per class. We use the GBM-Treatment response dataset for treatment response prediction (**Extended Data Figure 4.a** and **Extended Data Table 43**)), EBRAINS dataset[27] for fine-grained (n=30 classes) and coarse-grained (n=12) brain tumor subtyping (**Figure 2.m**, **Extended Data Figure 4.b**, and **Extended Data Table 42**), the BRACS dataset[28] for fine-grained (n=7) and coarse-grained (n=3) breast tumor subtyping (**Extended Data Figure 4.c,d** and **Extended Data Table 41**), and the BCNB dataset[29] for ER status prediction (**Extended Data Figure 4.e,f** and **Extended Data Table 40**). THREADS provides the best linear probing performance, outperforming baselines for most values of k. The predictive capabilities of THREADS are particularly highlighted in subtyping rare brain tumors, where THREADS performance with k=4 is superior to PRISM performance (second best performer) with k=16 (**Figure 2.n**).

THREADS fine-tuning

THREADS can also serve as weight initialization for further finetuning on a downstream task. This approach combines the benefit of large-scale pretraining while letting the model adapt to the nuances of the downstream application. Here, we fine-tuned a THREADS-initialized model on every downstream task in our evaluation pipeline. We employ a unified fine-tuning recipe that is applied to all tasks (**Online Methods**, **Baselines**). To mitigate overfitting and costly hyperparameter searches, we did not apply layer-wise learning rate decay, weight decay, or gradient accumulation. We apply a similar strategy to fine-tune CHIEF. For GIGAPATH, we

Figure 2: **Evaluation of THREADS and baselines with linear probing. a.** Average performance of THREADS and baselines on 54 tasks. THREADS is compared against PRISM, GIGAPATH, and CHIEF. Average performance per family of tasks: **b.** clinical subtyping and grading (8 tasks), **c.** mutation prediction (21 tasks), **d.** IHC status prediction (12 tasks), and **e.** treatment response and survival prediction tasks (13 tasks). **f–k** THREADS performance on treatment response and prognostication tasks characterized by label scarcity (n=36 to n=144 patients). Binary tasks (**f–i**) are measured with AUC. Survival tasks (**j,k**) are measured with concordance-index. **f.** Temozolomide treatment response in glioblastoma (GBM). **g.** Bevacizumab treatment response in ovarian cancer (OV). **h.** Neoadjuvant response assessment in invasive breast cancer (BRCA). **i.** Hormonal therapy response in prostate adenocarcinoma (PRAD). **j.** Overall survival (OS) prediction in pancreatic ductal adenocarcinoma (PDAC). **k.** Overall survival prediction in colon adenocarcinoma (COAD). **l.** Number of tasks where each model (THREADS and baselines) reaches highest performance across all tasks (n=54 tasks), treatment response (n=6 tasks) and survival tasks (n=7 tasks). **m.** Few-shot learning performance of THREADS against baselines in brain tumor subtyping. k refers to the number of training samples per class. Error bars represent the standard error measured across multiple folds. Boxes indicate quartile values of model performance (n=5 runs), and whiskers extend to data points within 1.5-fold the interquartile range. Task-wise P-values were determined using two-sided Tukey Honest Significance Difference tests accounting for multiple comparisons following a positive result (P<0.05) of a two-way ANOVA. Statistical significance across multiple tasks (e.g., for each family) was assessed using a mixed-effects model. P<0.05: *, P<0.01: **, P<0.001: ***.

follow the recommended recipe of gradient accumulation, weight decay, and layer-wise learning rate decay. Fine-tuning recipe for PRISM is not provided.

THREADS leads to significantly better performance than CHIEF fine-tuning (absolute gain of 17.9% and P-value<0.001 assessed with mixed-effects statistical modeling) and GIGAPATH (absolute gain of 7.3%, P-value<0.001) in our 54-task evaluation pipeline. When inspecting individual tasks, THREADS leads to significantly better performance than CHIEF and GIGAPATH in 54/54 tasks and 40/54 tasks (P-value<0.001), respectively (**Figure 3.a**). In addition, THREADS fine-tuning leads to a 4.3% absolute performance boost over an attention-based MIL baseline trained from scratch (P-value<0.001 assessed with mixed-effects modeling). THREADS leads to the largest performance gain in challenging tasks characterized by small to medium-size cohorts, such as mutation prediction (absolute gain of 5.5% over ABMIL and 7.7% over GIGAPATH, **Figure 3.b**), treatment response and survival prediction (gain of 4.3% over ABMIL and 9.4% over GIGAPATH, **Figure 3.c**) and IHC status prediction (gain of 4.5% over ABMIL and 9.8% over GIGAPATH, **Figure 3.d**). In clinical subtyping and grading tasks, characterized by larger cohorts, THREADS performance is comparable to GIGAPATH fine-tuning and ABMIL training **Figure 3.e**. Additional results for specific tasks are provided in **Figure 3.f,g,h,i,j**.

We additionally compared THREADS fine-tuning with a randomly initialized THREADS model. Overall, fine-tuning leads to an average absolute gain of 2.2% across all 54 tasks (P-value<0.001 assessed with mixed-effects statistical modeling as shown in **Figure 3.k**). In examining task performance across different families of tasks, we find that fine-tuning yields the most significant improvement, a 2.8% increase, in mutation prediction tasks, which typically involve challenging tasks and cohorts of small to medium size. Conversely, it shows the smallest improvement, a 0.5% increase, in clinical subtyping and grading, where the training cohorts tend to be larger and the tasks more objective. Additional results for specific tasks are provided in (**Figure 3.l**).
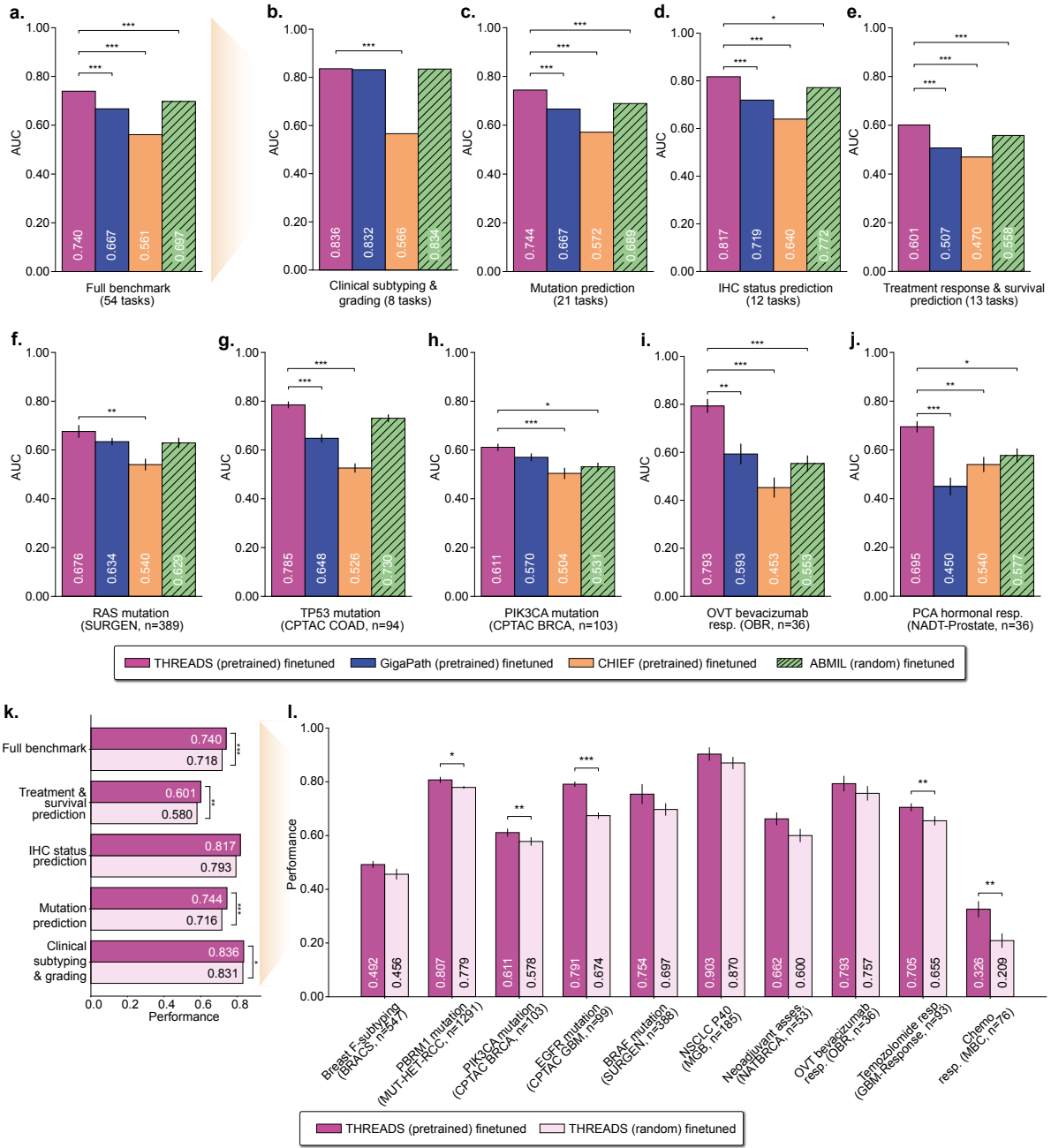
Retrieval capabilities of THREADS

THREADS is designed to provide off-the-shelf slide and patient embeddings. This property enables case and patient retrieval without additional model training or fine-tuning. To this end, we extract THREADS slide (and patient) embeddings for a collection of samples for which a diagnosis has already been made. We use these samples as a reference database to compare new query cases, which are first embedded using THREADS, and then compared to the k most similar embeddings (**Figure 3.a**). The other three slide encoders (PRISM, GIGAPATH, and CHIEF) are processed and evaluated in a similar manner. We evaluate retrieval performance using mean Average Precision at k (mAP@k), which measures the average number of relevant results within the top k retrieved items, weighted by their rank and averaged over all queries.

In rare brain tumor retrieval assessed with the EBRAINS dataset (30 classes, n=2,319 cases), THREADS provides the best overall performance for all values of k, significantly outperforming all baselines (P-value<0.001 for 3/3 baselines at all k) (**Figure 3.b**). We additionally study retrieval performance on the CPTAC consortium data, which aggregates cases from 10 cancer types for a total of 2,115 slides. THREADS outperforms all three baselines for $k = 1, 5,$ and 10 (P-value<0.05 for 3/3 baselines at mAP@1, P-value<0.001 for 2/3 baselines at mAP@5, and P-value<0.001 for 2/3 baselines at mAP@10) (**Figure 3.c**). These results highlight how THREADS can encode clinically relevant information and retrieve similar cases for comparison and investigation. Additional results are provided in **Extended Data Table 45 and 44**.

Molecular prompting with THREADS

Figure 3: **THREADS fine-tuning. a.** Average performance of THREADS and baselines finetuned on 54 benchmarking tasks, along with average performance for each family of tasks: **b.** clinical subtyping and grading (8 tasks), **c.** mutation prediction (21 tasks), **d.** IHC status prediction (12 tasks), and **e.** treatment response and survival prediction (13 tasks). Task-wise comparison of THREADS and baselines finetuned on individual tasks: **f.** RAS status prediction in SURGEN colorectal adenocarcinoma (COAD). **g.** TP53 mutation prediction in CPTAC-COAD. **h.** PIK3CA mutation prediction in CPTAC breast invasive carcinoma (BRCA). **i.** Bevacizumab response prediction in ovarian cancer with fine-tuning. **j.** Temozolomide response prediction in MGB glioblastoma (GBM). **k.** Comparison of THREADS fine-tuning *vs.* training a THREADS model from scratch on our benchmark and families of tasks. **l.** Task-wise performance of THREADS fine-tuning *vs.* THREADS randomly initialized on ten representative tasks. Error bars represent standard error, and the centers correspond to the mean computed values of each metric. Task-wise P-values were determined using two-sided Tukey Honest Significance Difference tests accounting for multiple comparisons following a positive result (P<0.05) of a two-way ANOVA. Statistical significance across multiple tasks (e.g., for each family) was assessed using a mixed-effects model. P<0.05: *, P<0.01: **, P<0.001: ***.

9

A hallmark characteristic of multimodal foundation models is to enable transfer and generalization without task supervision. In vision-language models like CLIP [30] and CONCH [15], such capabilities include zero-shot classification, in which by formulating class labels (*e.g.*, "Lung Adenocarcinoma", "Lung Squamous Cell Carcinoma") as text prompts via natural language, tasks such as NSCLC subtyping can be performed without requiring training data. In THREADS, we introduce a novel multimodal capability known as "molecular prompting" (**Figure 4.d**), in which canonical molecular profiles (*e.g.*, molecular representations of their corresponding disease states) can be leveraged to perform clinical tasks without requiring any task-specific model development. To perform molecular prompting, for each class, we average the representations of molecular profile data from a support dataset (encoded using the molecular branch of THREADS) to create class-wise molecular prototypes, which can then be used for cross-modal slide retrieval and classification. At inference time, we classify a query WSI by assigning it to the class of the nearest molecular prompt based on $L2$ distance.

We evaluated molecular prompting across eight tasks, including clinical cancer subtyping, gene mutation and IHC status prediction, and prognostication (**Figure 4.e**). Classification with molecular prompts achieves competitive performance across diverse tasks. Building IDH wild-type and mutant prompts from TCGA-GBMLGG and testing on EBRAINS yields a high AUC of 0.960, comparable to linear probing with THREADS WSI embedding (0.961). Molecular prompts can also represent typical high- and low-risk profiles, allowing patient survival to be estimated based on similarity to these prompts (additional information is provided in **Online Methods, Evaluation**). For instance, high- and low-risk prompts generated from TCGA-CCRCC and applied to prognosis prediction on CPTAC-CCRCC achieve a competitive C-index of 0.687. Additional results can be found in **Extended Data Table 46**.
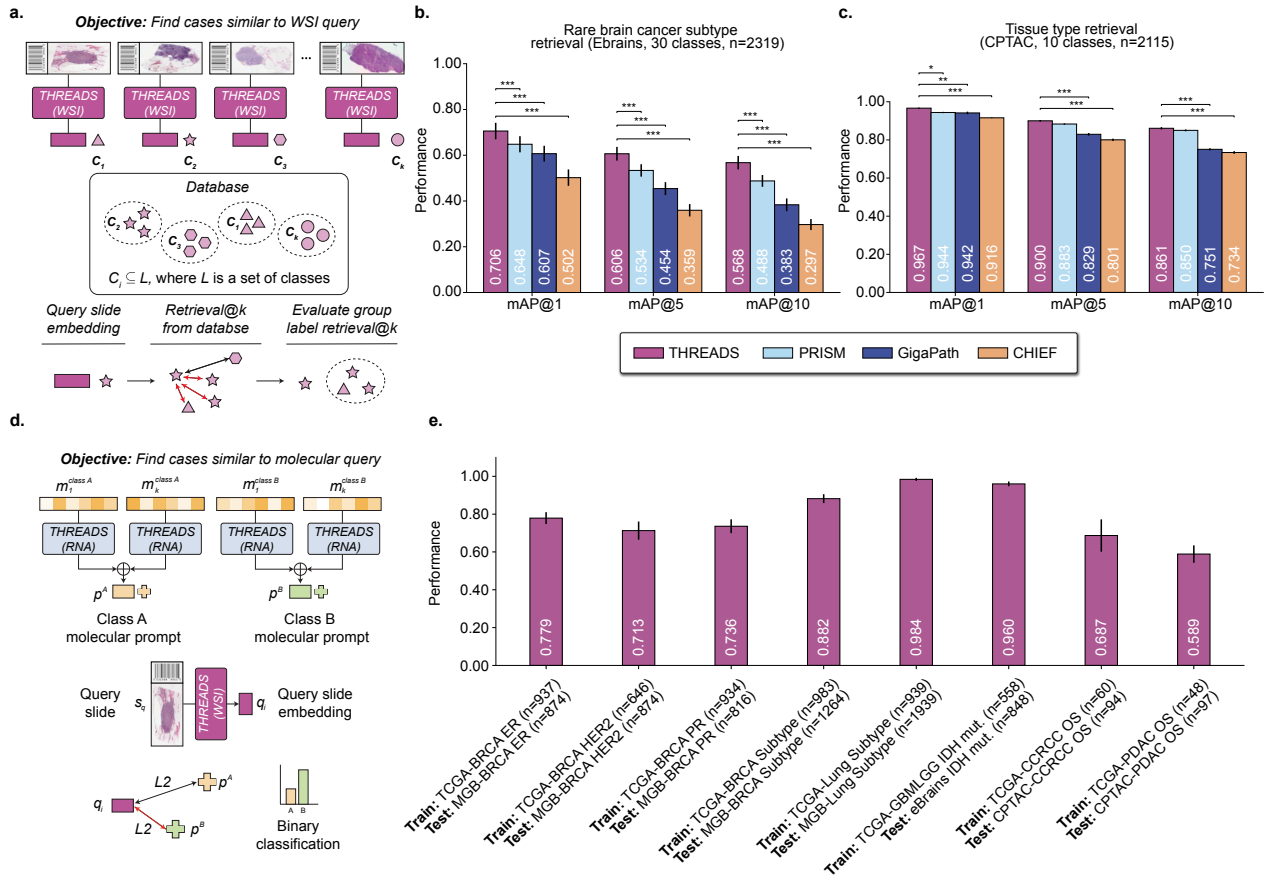
Insights into THREADS.

Scaling laws. We additionally study scaling laws in THREADS, building on existing works in foundation models that highlighted the benefits of larger training datasets and model sizes[5,7]. To this end, we pretrained THREADS using subsets of MBTG-47K of varying sizes. We sampled 1%, 5%, 25%, 50%, and 75% of the data from each source, ensuring uniform sampling across major tissue sites. This resulted in the creation of MBTG-1 (473 histomolecular pairs), MBTG-5 (2,356 histomolecular pairs), MBTG-25 (11,791 histomolecular pairs), MBTG-50 (23,584 histomolecular pairs), and MBTG-75 (35,377 histo-molecular pairs). THREADS highlights a data scaling law, as shown in **Figure 5.a**. Across all tasks, we observe a +3.9% performance increase when using 1% to 100% of MBTG-47K. All families of tasks benefit from data scaling, with treatment response and survival prediction tasks showing the largest performance gain (+5.2%). When comparing THREADS against baselines, we also observe that our approach is more data-efficient than PRISM (trained on 195,344 specimens), GIGAPATH (trained on 171,189 whole-slide images), and CHIEF (trained on 60,530 whole-slide images). Additional information is provided in **Extended Data Table 47**.
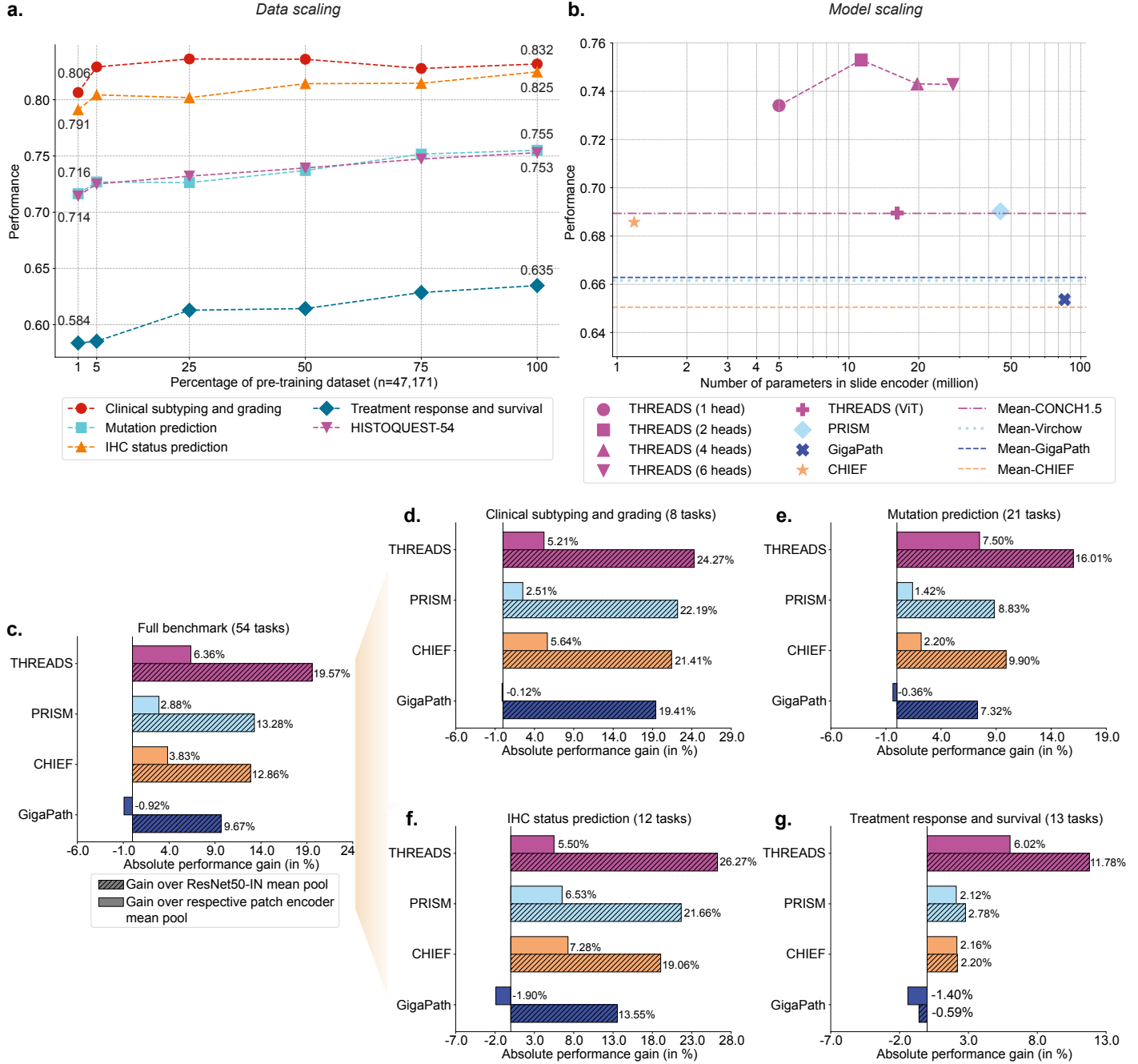
We also assessed model scaling laws by ablating THREADS using a varying number of attention heads, resulting in models with a single head (5.0 million parameters in the slide encoder), two heads (proposed approach, 11.3 million parameters), four heads (19.7 million parameters), six heads (28.1 million parameters) and a ViT model with two Transformer layers (16.1 million parameters). We observe that model scaling peaks with a two-head model and then plateaus or leads to decreased performance (-1.0% when using six vs. two attention heads). A ViT baseline trained with THREADS leads to lower performance than our proposed architecture by 6.3%. Comparing THREADS against PRISM and GIGAPATH highlights the parameter-efficiency of THREADS. Despite being 4.0× and 7.5× smaller than PRISM and GIGAPATH slide encoders, THREADS leads to significantly better performance on our benchmark. CHIEF is lightweight due to its compact architecture but provides significantly lower performance than our single-head model. This analysis highlights that scaling model size in slide encoder does not necessarily lead to better performance and that other factors are more important.

Additional information is provided in **Extended Data Table 48**.

Mean pooling. To better understand the superior performance of THREADS over baselines, we conducted additional ablations. First, we compared the quality of THREADS patch embeddings (based on CONCHV1.5) against GIGAPATH patch encoder, PRISM patch encoder (based on VIRCHOW), and CHIEF patch encoder (based on CTRANSPATH). To this end, we adopt mean pooling to derive a slide embedding, which we then use for linear probing classification. CONCHV1.5 with mean pooling reaches an average of 68.9% across all tasks outperforming VIRCHOW, GIGAPATH and CHIEF by 2.7%, 2.6% and 3.8%, respectively (**Figure 5.a** and **Extended Data Table 47**). We hypothesize that this gain stems from (i) vision-language fine-tuning in CONCHV1.5, whereas VIRCHOW, GIGAPATH, and CTRANSPATH are vision-only models, and (ii) from extracting patch features on larger regions (512×512-pixel regions vs. 256×256-pixel in baselines) which can better capture morphological context. We also note that CONCHV1.5 is a ViT-Large model (307 million vision parameters), whereas PRISM uses a ViT-H (632 million parameters, 2.0× more than CONCHV1.5), and GIGAPATH used a ViT-G (1.13 billion parameters, 3.7× more than CONCHV1.5), highlighting the parameter-efficiency of our pipeline.

Figure 4: **Retrieval and prompting capabilities of THREADS. a.** Method overview for case retrieval. **b.** Rare brain tumor subtype retrieval in EBRAINS (n=30 subtypes) evaluated using mean Average Precision (mAP). **c.** Cancer subtype retrieval in CPTAC (n=10 cancer subtypes) evaluated using mAP. **d.** Method overview for molecular prompting. **e.** Molecular prompting performance on eight tasks. Error bars represent 95% confidence intervals and the centers correspond to computed values of each metric. P-values were determined using two-sided Tukey Honest Significance Difference tests accounting for multiple comparisons following a positive result (P<0.05) of a two-way ANOVA. P<0.05: *, P<0.01: **, P<0.001: ***.

Figure 5: **Properties of THREADS. a.** Data scaling law of THREADS across all tasks and families of tasks. Percentage of training slides varies from 1% to 100% of MBTG-47K. **b.** Model scaling law of THREADS across all tasks and families of tasks. **c.** Absolute performance gain of THREADS and slide encoder baselines over the mean pooling baselines from their respective patch encoder and RESNET50-IN. Performance averaged on our 54-task benchmark. Absolute performance gain of THREADS and slide encoder baselines over the mean pooling baselines from their respective patch encoder and RESNET50-IN for each family of tasks: **d.** clinical subtyping and grading (8 tasks), **e.** mutation prediction (21 tasks), **f.** immunohistochemistry status (IHC) prediction (12 tasks), **g.** treatment response and survival prediction (13 tasks). RESNET50-IN is ResNet50 model pretrained on ImageNet (IN).

We additionally compare slide encoders with their respective mean pooling baselines, i.e., CONCHv1.5 mean pooling and THREADS, GIGAPATH mean pooling and GIGAPATH, CTRANSPATH mean pooling and CHIEF, and VIRCHOW mean pooling and PRISM (**Figure 5.c**). Using linear probing evaluation across all tasks, THREADS leads to a gain of 6.36% over CONCHv1.5 mean pooling (P-value<0.001) and 19.3% over RESNET50-IN mean pooling (P-value<0.001). In contrast, GIGAPATH slide encoder leads to lower performance than mean pooling (-0.83%). Both PRISM and CHIEF lead to a performance gain over their respective mean pooling (2.86% in PRISM and 4.40% in CHIEF), lower than THREADS performance gain. This observation highlights the complexity of whole-slide representation learning in capturing information-dense off-the-shelf embeddings.

Clustering in THREADS embedding space. To study the superior performance of THREADS in predicting clinically and biologically relevant information from whole slide images, we explored the clustering capabilities of the latent space compared to baselines. To this end, we embedded all slides from the ten CPTAC cohorts (n=2,115 WSIs) using THREADS and baselines. We then applied K-means clustering, where the number of clusters was set to the number of cancer types (n=10). From there, we computed two clustering metrics: the adjusted Rand index (ARI) and mutual information (MI) (**Extended Data Figure 2**). THREADS highlights better clustering capabilities than PRISM, GIGAPATH, and CHIEF (for instance ARI=0.654 for THREADS *vs.* 0.354 for CHIEF). In addition, we derived tSNE visualization of the latent space, which further highlights the separability of THREADS compared to baselines. We conducted a similar analysis with EBRAINS for fine brain tumor subtyping (30 classes). We observe similar trends with THREADS highlighting better clustering and linear separability than baselines (for instance, MI=2.104 for THREADS *vs.* 1.44 for GIGAPATH).

Discussion

In this study, we introduced THREADS, a foundation model for pathology that can provide biologically and clinically relevant representations of H&E-stained whole-slide images. THREADS uses a novel multimodal pretraining strategy, where the learned slide representation is guided by its corresponding molecular profile. Using this strategy, the resulting slide representations can capture morphological features reflective of the underlying molecular composition of the tissue. THREADS was thoroughly tested on a wide benchmark of 54 tasks, covering four families of tasks: clinical cancer subtyping and grading, gene mutation prediction, immunohistochemistry status prediction, and treatment response and survival prediction. THREADS consistently shows state-of-the-art performance under several evaluation scenarios, including in- and out-of-domain generalization, few-shot learning, and case retrieval. Importantly, THREADS can reach clinical-grade performance on subtyping and grading tasks using simple linear models built upon our slide embeddings. THREADS also highlights great potential for patient outcome prediction and can help identify patients who will respond to certain treatments.

THREADS sets apart from existing methods using its unique pretraining strategy based on multimodal alignment with molecular profiles. Unlike PRISM [13], which relies on matching pathology reports, molecular data provide an unbiased, objective perspective on cellular and tissue states, free from the subjective influences inherent in written reports. On the other hand, CHIEF [11] and GIGAPATH [6] employ weaker pretraining signals, relying on contrastive alignment with tissue sites and masked autoencoding, respectively. We hypothesize that these approaches lack the capacity to capture the subtle morphological features essential for addressing most clinical tasks. Our investigation into scaling laws of THREADS further reveals that the saturation point of model and data scale remains an open question in slide representation learning. We found that simpler clinical tasks, such as cancer grading, do not benefit significantly from larger pretraining datasets. However, more challenging tasks–particularly those involving treatment response prediction and molecular predictions–show

substantial performance gains when models are trained on larger and more diverse datasets. This indicates that data diversity and pretraining strategy are critical factors influencing the efficacy of the resulting model. THREADS is significantly smaller than existing models, being $7.5 \times$ smaller than GigaPath and $4.0 \times$ smaller than PRISM. This suggests that in slide representation learning, simply scaling model size may not be the most influencing factor for building general-purpose models. In addition, THREADS is trained on a highly diverse dataset that includes 39 main human tissue sites following the highest level of the OncoTree cancer classification system. In contrast, GIGAPATH, CHIEF, and PRISM are trained on less diverse tissue sites, often with a skewed distribution toward skin, breast, and lung cases. We hypothesize that the broader diversity in THREADS likely contributes to its enhanced generalizability and robustness across a wide range of tasks and tissue types.

Despite these advancements, certain limitations remain. Although THREADS was pretrained on an unprecedented cohort of over 47,000 histomolecular pairs, it cannot encompass the full spectrum of molecular and morphological heterogeneity. As next-generation sequencing becomes more widely deployed in clinical settings, the potential to scale THREADS ' pretraining dataset by orders of magnitude may reshape this landscape, potentially uncovering new scaling laws that are currently beyond reach with our existing cohorts. Additionally, extending our molecular-guided approach to include other molecular assays, such as immunohistochemistry and special stains, could broaden the scope of THREADS [10]. In addition, THREADS architecture uses a multihead attention-based model, which treats patch embeddings independently. Attempts to replace our backbone with a Vision Transformer model fail to match the performance of simpler models, even the ones with a single attention head. The use of larger image patches (512×512 pixels) instead of the typical 256×256 pixels in most patch encoders may reduce the need for explicit context modeling. Alternatively, slide encoders based on Transformers may need a larger pretraining cohort size for significant performance improvements.

THREADS has the potential to impact various aspects of computational pathology and oncology. First, it can bring off-the-shelf integration in data-scarce scenarios. THREADS can be readily used to prototype new tasks and assess predictive performance at a minimal cost. In addition, the reduced training data requirements enable the development of clinical-grade predictive systems for rare diseases. Researchers and clinicians can also utilize THREADS pretrained weights as initialization for additional fine-tuning on specific tasks. This transfer learning approach can accelerate model development and can improve performance on specialized tasks, such as rare molecular alteration classification or treatment response and resistance prediction. Finally, the case retrieval capabilities of THREADS make it well-suited for identifying rare conditions in clinical settings. Overall, our study highlights the rich biological information contained in molecular assays, which can be transferred to slide encoders to advance the development of diagnostic and prognostic tools. Future work will focus on scaling the pretraining dataset size and increasing the biological richness of the training signal by including additional modalities.

Online Methods

The Mass General Brigham (MGB) institutional review board approved the retrospective analysis of pathology slides (whole-slide images or WSIs), corresponding next-generation sequencing (NGS) assays, and corresponding reports used in this study. Research conducted in this study involved a retrospective analysis of pathology slides and NGS assays, and the participants were not directly involved or recruited for the study. The requirement for informed consent to analyze archival pathology slides and NGS assays was waived. Before scanning and digitization, all pathology slides were de-identified to ensure anonymity. The sample sizes were determined by the availability of the data.

Pretraining dataset curation

We present THREADS pretraining dataset, MBTG-47K, a large and diverse dataset composed of paired formalin-fixed paraffin-embedded (FFPE) haematoxylin and eosin (H&E) whole-slide images (WSIs), tissue bulk RNA expression, and DNA variant data including single nucleotide variations (SNV), insertions and deletions (indels), and copy number variations (CNV). Sourced from the Massachusetts General Hospital (MGH), the Brigham and Women's Hospital (BWH), the Genotype-Tissue Expression (GTEx) consortium, and The Cancer Genome Atlas (TCGA), the dataset comprises 47,171 WSIs from 39 major organs, totaling 125,148,770 512×512-pixel histology images tiled at $20\times$ magnification. 26,615 WSIs have associated RNA expression data, and 20,556 WSIs have associated DNA variant data. The total size of the dataset is 40.7 TB. We describe each data source contributing to the MBTG-47K dataset. Additional information is provided in **Extended Data Table 1**.

MGH. Anchored multiplex PCR[31] and next-generation sequencing (NGS) of total nucleic acid were applied to generate bulk RNA expression data using FusionPlex (Integrated DNA Technologies, Coralville, IA). The Solid Fusion Assay V2 is a clinically validated pan-cancer RNA assay that targets canonical exons involved in fusion variants of 59 genes and control genes. The assay generates predominantly RNA reads to detect gene fusions, splicing, exon-skipping events, and gene expression. In this study, we focused on bulk gene expression data. Transcript abundance for each gene was quantified using the Kallisto 0.50.1 pseudo alignment software, with an index built from GENCODE Human Release v45 (genome assembly GRCh38.p14). Measurements from different isoforms were summed to acquire a single total per gene. RNA expression was summarized as transcripts per million (TPM) and further normalized by taking the $\log_2$ of TPM across 54 genes. No additional batch effect normalization techniques were applied. The associated H&E glass slides were scanned using an Aperio GT450 scanner at $40\times$. In total, 6,899 FFPE H&E WSI and bulk RNA expression pairs from 25 tissue sites were utilized. The final MGH dataset amounted to 11.0 TB.

BWH. OncoPanel is an Agilent SureSelect hybrid capture targeted DNA NGS assay designed to detect SNVs, indels, CNVs, and some structural variations. Library preparation, sequencing, and bioinformatics analysis have been previously described[32]. Similar to the MGH Solid Fusion Assay V2, OncoPanel testing was performed according to a routine clinical workflow with expert molecular pathologist review. For THREADS, we sourced SNV, CNV, and indel data associated with 20,556 FFPE H&E slides from 32 tissue sites. We subsetted the data to 239 common genes across the OncoPanel versions used in this study (ranging from 2012 to 2020). CNVs were categorized into four groups: two-copy deletion, loss, gain, and amplification. SNVs and indels were categorized into three groups: small coding change, large coding change, and non-coding change. The mutation status of each gene was multi-hot encoded and concatenated together to form a vector of length 7; therefore, across the 239 genes, the variant status vector has a length of 1,673. The H&E WSIs associated with the OncoPanel testing were scanned using an Aperio GT450 scanner at $40\times$ magnification. The final BWH dataset amounted to 12.0 TB.

TCGA. The Cancer Genome Atlas (TCGA) contributed 10,209 FFPE H&E WSIs from 32 cancer types, paired with bulk RNA expression. TCGA includes histology data from over 11,000 cancer patients, with tissue samples scanned at $40\times$ and $20\times$ magnification using Aperio and Hamamatsu scanners. The expression data comprises bulk whole-transcriptomic RNA sequencing analysis from approximately 20,000 samples across 33 cancer types, with a sequencing depth of 50-200 million reads per sample, using Illumina HiSeq platforms. Exclusion criteria for WSIs were frozen tissue, benign and non-diagnostic, missing appropriate magnification information and metadata necessary for processing, and lacking associated RNA expression. While TCGA provides whole-transcriptome sequencing, we selected a set of 4,917 cancer-related genes[33], which was reduced to 4,848 genes present in the vocabulary of our transcriptomic encoder (scGPT encoder[18]) used to generate

molecular embeddings (**Online Methods, Model Design and Development**). RNA expression was measured in transcripts per million (TPM) and further normalized by taking the $\log_2$ of TPM. No additional batch effect normalization techniques were applied. The final TCGA dataset amounted to 10.5 TB.

GTEx. The Genotype-Tissue Expression (GTEx)[12] contributed 9,507 FFPE H&E WSIs and bulk RNA expression pairs from 29 tissue sites to MBTG-47ĸ. GTEx includes FFPE WSIs from 24,782 non-cancerous samples scanned at $40\times$ magnification using either Aperio AT2 or Hamamatsu NanoZoomer-XR. The expression data comprises whole-transcriptome bulk RNA sequencing from 17,382 samples, with a depth of 50-100 million reads per sample, using Illumina HiSeq 2000/2500 platforms. Exclusion criteria for WSIs include missing appropriate magnification information and metadata necessary for processing and lacking associated RNA expression. Even though GTEx provides whole-transcriptome sequencing, we selected a set of 5,000 genes showing maximum variation (as measured by the standard deviation of $\log_2$ of TPM) across all organs. This gene set was reduced to 4,932 genes found in the vocabulary of our transcriptomic encoder (scGPT[18]). RNA expression was measured in transcripts per million (TPM) and further normalized by taking the $\log_2$ of TPM. No additional batch effect normalization techniques were applied. The final GTEx dataset amounted to 7.2 TB.

Downstream tasks and datasets

We provide a description of each task and cohort in our benchmark, which includes 54 tasks from 23 datasets across nine major organs. Our benchmark covers six types of tasks: morphological tumor subtyping (**Extended Data Table 2, 52, 53, 56, 57, 58, 59**), tumor grading (**Extended Data Table 3, 54, 55**), immunohistochemistry status prediction (**Extended Data Table 4**), prediction of gene-level mutations (**Extended Data Table 5**), treatment response prediction (**Extended Data Table 6**), and survival prediction (**Extended Data Table 7**).

MGB-Breast. We used an internal cohort of invasive breast cancer (BRCA) for morphological and immunohistochemistry status prediction[10,34]. MGB-Breast comprises 1,264 WSIs (mix of biopsies and resections) scanned from Brigham and Women's Hospital (one WSI per patient). We curated one morphological subtyping task (**Extended Data Table 52**) and three immunohistochemistry (IHC) status prediction tasks: estrogen receptor (ER) status prediction, progesterone receptor (PR) status prediction, and human epidermal growth factor receptor 2 (HER2) status prediction. ER, PR, and HER2 status were manually extracted from pathology reports. Additional information is provided in **Extended Data Table 2** and **Extended Data Table 4**.

MGB-Lung. We used an internal cohort of lung cancer cases for morphological and IHC status prediction[10,34]. MGB-Lung comprises 1,939 WSIs scanned from Brigham and Women's Hospital (one WSI per patient). We curated one morphological subtyping task (**Extended Data Table 53**) and six immunohistochemistry tasks: (1) thyroid transcription factor-1 (TTF-1) status prediction, (2) protein 40 (P40) status prediction, (3) protein 63 (P63) status prediction, (4) Napsin A status prediction, (5) caudal type homeobox 2 (CDX2) status prediction, and (6) cytokeratin 5 and 6 (CK5-6) status prediction. Additional information is provided in **Extended Data Table 2** and **Extended Data Table 4**.

BCNB. We used the public BCNB dataset[29] (Breast Cancer Core-Needle Biopsy) for IHC status prediction in breast cancer. BCNB comprises 1,058 WSIs (one WSI per patient) which we use for ER status prediction, PR status prediction, and HER2 status prediction. Additional information is provided in **Extended Data Table 2**.

MUT-HET-RCC. We used the MUT-HET-RCC dataset[35] for mutation prediction in renal cell carcinoma. MUT-HET-RCC comprises 1,291 WSIs (one WSI per patient) which we use for (1) BAP1 mutation prediction,

(2) PBRM1 mutation prediction, and (3) SETD2 mutation prediction. Additional information is provided in **Extended Data Table 5**.

IMP. We used the public IMP-CRS 2024 dataset (IMP)[36] for colorectal cancer grading. IMP consists of 5,333 WSIs collected from colorectal biopsies and polypectomies. IMP is used for 3-class tumor grading into non-neoplastic lesions, low-grade lesions (adenomas with low-grade dysplasia), and high-grade lesions (adenomas with high-grade dysplasia and invasion) (see **Extended Data Table 54**). Additional information is provided in **Extended Data Table 3**.

Prostate cANcer graDe Assessment (PANDA). We used the public PANDA data for prostate cancer grading (ISUP grading)[3]. PANDA comprises 10,616 core needle biopsies from Radboud University Medical Center and Karolinska Institute, each annotated with an ISUP grade (6-class classification task). We follow prior work[5,37] in excluding slides with equivocal labels (`https://www.kaggle.com/competitions/prostate-cancer-grade-assessment/discussion/169230`), which resulted in 9,555 slides with the label breakdown shown in **Extended Data Table 55**. We used the train split (7,647 WSIs) and test split (954 WSIs) from GIGAPATH [6], and we did not use their validation split (954 WSIs). Additional information is provided in **Extended Data Table 3**.

Clinical Proteomic Tumor Analysis Consortium (CPTAC). We used the public CPTAC data for pan-cancer mutation prediction[38,39]. Specifically, we included (1) CPTAC-BRCA (invasive breast cancer) for PIK3CA and TP53 mutation prediction, (2) CPTAC-CCRCC (clear-cell renal-cell carcinoma) for BAP1 and PBRM1 mutation prediction, (3) CPTAC-COAD (colon adenocarcinoma) for KRAS and TP53 mutation prediction, (4) CPTAC-GBM (glioblastoma) for EGFR and TP53 mutation prediction, (5) CPTAC-HNSC (head and neck squamous cell carcinoma) for CASP8 mutation prediction, (6) CPTAC-LSCC (squamous cell lung carcinoma) for KEAP1 and ARID1A mutation prediction, (7) CPTAC-LUAD (lung adenocarcinoma) for EGFR, STK11 and TP53 mutation prediction, and (8) CPTAC-PDAC (pancreatic ductal adenocarcinoma) for SMAD4 mutation prediction. We also used overall survival data for CPTAC-CCRCC, CPTAC-PDAC, CPTAC-LUAD, and CPTAC-HNSC[40]. Additional information is provided in **Extended Data Table 5**.

BReAst Carcinoma Subtyping (BRACS). We used the public BRACS dataset[28] for coarse- and fine-grained breast morphological subtyping. BRACS consists of 547 breast carcinoma WSIs from 189 patients sourced from IRCCS Fondazione Pascale. Each WSI is used for coarse-grained (**Extended Data Table 56**) and fine-grained (**Extended Data Table 57**) morphological subtyping. Due to the limited size of the official test set, we redefined train-test splits with an 80:20 ratio. Because BRACS-Fine and BRACS-Coarse are slide-level prediction tasks with multiple slides per case, we kept all slides belonging to the same patient together, ensuring one patient does not end up in both train and test. Therefore, this dataset was not explicitly label-stratified (as each patient would have multiple labels). Additional information is provided in **Extended Data Table 2**.

EBRAINS. We used the EBRAINS dataset[27] for coarse- and fine-grained brain tumor subtyping. EBRAINS consists of 3,114 WSIs acquired by the EBRAINS Digital Tumor Atlas at the University of Vienna. We reused splits from UNI[5], which kept categories with at least 30 samples, resulting in 2,319 slides. Each WSI is used for coarse-grained (**Extended Data Table 58**) and fine-grained (**Extended Data Table 59**) morphological subtyping. Splits are stratified by patients to ensure slides from a patient are not found in both train and test splits. Additional information is provided in **Extended Data Table 2**.

OV-Bevacizumab. We used the OV-Bevacizumab dataset[41] for treatment response prediction in ovarian cancer. OV-Bevacizumab consists of 288 WSIs from 78 patients. Non-responders are defined as having a measurable

regrowth of the tumor or as a serum CA-125 concentration of more than twice the value of the upper limit of normal during the treatment course for the bevacizumab therapy. We kept all patients who received bevacizumab as their first-line treatment and additionally removed four cases (case IDs: P00181938C, 2630938, 2224393, 2937351), which were labeled as both responders and non-responders, yielding 85 WSIs from 36 patients. Additional information is provided in **Extended Data Table 6**.

NADT-Prostate. We used the neoadjuvant androgen deprivation therapy (NADT)-Prostate dataset[42] for hormonal therapy response prediction in prostate adenocarcinoma. Baseline tumor volumes were estimated using multiparametric magnetic resonance imaging (mpMRI). After 6 months of NADT combined with enzalutamide, patients underwent a second mpMRI before radical prostatectomy (RP). The final pathologic response to treatment was defined by a residual cancer burden of 0.05 cubic centimeters, distinguishing responders from non-responders. While the entire dataset consists of 1,401 WSIs with various stains, we only used the H&E stained WSIs, yielding 449 WSIs from 36 patients (20× magnification). Additional information is provided in **Extended Data Table 6**.

Treatment response in glioblastoma (GBM-Treatment). We collected an internal cohort of 93 glioblastoma patients, accounting for 347 H&E-stained slides, who received radiotherapy and temozolomide[43,44]. Based on patient survival in months following treatment initiation (all patients deceased) and a cutoff of 15 months [45], we stratified the patients into responders and non-responders. Additional information is provided in **Extended Data Table 6**.

Post-NAT-BRCA. We used the post-neoadjuvant therapy (NAT) breast invasive carcinoma (Post-NAT-BRCA) dataset[46] to assess the presence of lymphovascular invasion in post-NAT WSIs. The dataset contains 53 H&E-stained WSIs from 50 patients (20× magnification). Additional information is provided in **Extended Data Table 6**.

SURGEN. We used public cases from the SR386 cohort of SurGen[47], which includes 389 patients with colon and rectum adenocarcinoma. For each patient, we predict mismatch repair (MMR) loss, BRAF mutation, KRAS mutation, 5-year death, and overall survival. Treatment information is available only for a subset of patients. Additional information is provided in **Extended Data Table 7 and 5**.

MBC. We used the public Bergstrom dataset[48,49] from which we retrieved 77 metastatic breast cancer patients (MBC) with corresponding H&E WSIs (n=99 WSIs, 1 to 2 WSI per patient). All 77 patients were treated with platinum, with a subset of 54 who were additionally treated with taxane. We predict Response Evaluation Criteria in Solid Tumors (RECIST1.1) and overall survival. Since all patients in MBC received the same treatment, predicting survival can be considered as predicting response to the treatment given. Additional information is provided in **Extended Data Table 7 and 5**.

BOEHMK. We used the public BOEHMK[50] dataset comprising 183 patients for which we could retrieve the H&E WSI and corresponding metadata, including overall and progression free survival. Patients were diagnosed with high-grade serous ovarian cancer and treated with neoadjuvant chemotherapy followed by interval debulking surgery, or underwent primary debulking surgery. Since all patients in BOEHMK received the same treatment, predicting progression free survival can be considered as predicting response to the treatment given. Additional information is provided in **Extended Data Table 6 and 7**.

TCGA (generalizability). TCGA-GBMLGG consists of 1,123 WSIs from 558 patients with glioblastomas multiforme (GBM) and lower-grade gliomas (LGG). The WSIs are classified into two classes: isocitrate de-

hydrogenase (IDH) mutation (425 WSIs) and no IDH mutation (698 WSIs). **EBRAINS** serves as an external cohort for this task (IDH MUT: 333 WSIs, IDH WT 540 WSIs). **TCGA-BRCA** (invasive breast carcinoma) consists of 1,048 WSIs from 983 patients. The WSIs are classified into two cancer subtypes: invasive ductal carcinoma (IDC) (838 WSIs) and invasive ductal carcinoma (ILC) (210 WSIs). MGB-Breast subtyping serves as external cohort for this task. **TGCA-BRCA** is also used for IHC status prediction: ER (996 WSIs total, 78.3% WSIs positive), PR (993 WSIs total, 68.2% WSIs positive), HER2 (692 WSIs total, 22.8% WSIs positive)[10]. **BCNB** serves as external cohort IHC status prediction. **TCGA-Lung** (lung cancer) consists of 1,043 WSIs from 946 patients with non-small cell lung cancer (NSCLC). The WSIs are classified into two classes: lung adenocarcinoma (LUAD, 531 slides) and lung squamous cell carcinoma (LUSC, 512 slides). MGB-Lung subtyping serves as the external cohort for this task. We define **TCGA-LUAD** as only the adenocarcinomas and use this dataset for overall survival prediction (509 WSIs from 446 patients, 60.1% censored). **CPTAC-LUAD** is used as the external test cohort for this task. **TCGA-PDAC** (pancreatic ductal adenocarcinoma) consists of 180 WSIs from 166 patients with PDAC. We use overall survival labels (47.2% censored) while testing on **CPTAC-PDAC** for external validation.

Data splits. We created two types of splits: $k =$ All splits, which distribute all available samples between train and test, and fewshot splits, which restrict the size of the training set to only a few examples ("shots"). For certain datasets (EBRAINS, PANDA, IMP), we use "official" single-fold $k =$ All splits that have been publicly released. Otherwise, we create 80:20 train:test splits using 5-fold cross-validation or 50-fold bootstrapping. We also create fewshot splits with $k \in \{1, 2, 4, 8, 16, 32\}$ examples per class. For fewshot splits, if there is more than one $k =$ All fold, then corresponding fewshot splits are created by sampling k items from the training set of each $k =$ All fold, and masking the remainder. Otherwise, bootstrapped fewshot splits are created by repeatedly sampling k items from the single training fold. Note that the test set of all splits for each task is the same. For certain tasks with classes containing too few labeled examples, we omit $k = 32$ fewshot splits.

Model design and development

Each WSI goes through three steps: (1) tissue detection and patching, (2) feature extraction from each patch, and (3) slide encoding using THREADS.

Tissue segmentation and patching. Each slide is tiled into fixed-size image patches and processed using a pretrained vision model to extract patch-level feature embeddings. For compute efficiency, we only process patches overlapping with tissue and ignore background regions. Background *vs.* tissue segmentation is performed using a deep learning feature pyramidal network (FPN) fine-tuned from the `segmentation-models-pytorch` package[51] on an in-house dataset of mask annotations. Non-overlapping 512×512-pixel patches are extracted at 20× magnification (~ 0.5 µm/px) for each slide.

Patch encoder. We use the CONCHV1.5 patch encoder, the next iteration of CONCH[15]. CONCHV1.5 was trained by initializing UNI weights[5] followed by full multimodal fine-tuning using image captions. UNI is a state-of-the-art vision-only foundation model for pathology trained on 100 million image patches of size 224×224 pixels using a Vision Transformer Large (ViT-L)[16]. Fine-tuning was conducted with 1.17 million vision-language pairs (pathology image/caption pairs) using CoCa[52] on 448×448-pixel patches, as described in [15,53]. 512×512-pixel patches were resized to 448×448, and normalized using default ImageNet mean and standard deviation parameters before being passed to CONCHV1.5. An overview of CONCHV1.5 training hyperparameters is provided in **Extended Data Table 49**.

Slide encoder. THREADS consists of an attention-based multiple instance learning (ABMIL) model [25,54] with

single or multiple attention heads, depending on the configuration. For the single-headed configuration, raw patch embeddings are projected from 768-dimensional CONCHV1.5 features to 1024-dimensional features $\mathbf{X}$ using a pre-attention network with three hidden layers, layer normalization, GELU activation, and 0.1 dropout. The attention head processes batched input patch features $\mathbf{X} \in \mathbb{R}^{N \times 1024}$, where N is the number of patches. The gated attention mechanism comprises three fully connected layers: two parallel layers (a and b) with a hidden dimension of 1024 and 25% dropout, followed by a final layer (c). The attention weights $\boldsymbol{\alpha}$ are computed as:

$$\boldsymbol{\alpha} = c(\tanh(a\mathbf{X}) \odot \sigma(b\mathbf{X})) \tag{1}$$

where $\odot$ denotes element-wise multiplication, $\tanh$ and σ represent the hyperbolic tangent and sigmoid functions, respectively. The final slide-level features s are computed by multiplying the softmax-normalized attention scores by the patch features:

$$\mathbf{s} = \mathrm{softmax}(\boldsymbol{\alpha})^{\top}\mathbf{X} \tag{2}$$

In the multi-headed configuration, the pre-attention network includes a third hidden layer (GELU activation, 0.1 dropout), which projects from hidden dimension 1024 to $M \times 1024$, where M is the number of heads. The output of this layer is chunked into M feature vectors, each processed separately by its corresponding attention head. The aggregated slide-level features from each head are concatenated and projected with a post-attention linear layer $L : \mathbb{R}^{(M \times 1024)} \to \mathbb{R}^{1024}$ to derive the final 1024-dimensional slide embedding.

Gene expression encoder. We encode gene expression data using a modified scGPT model[18]. scGPT is a single-cell foundation model based on the Transformer architecture[17]. While originally developed for single-cell gene expression encoding, we adapted it to operate on bulk RNA expression data[55]. Expression data preprocessing was performed for each data source as described in **Online Methods, Pretraining dataset curation**. scGPT consists of three encoders: a gene-identity encoder G, an expression-value encoder E, and a transformer encoder T. G is a lookup table of learned 512-dimensional gene identity embeddings followed by layer normalization. E is a 2-layer MLP that expands each 1-dimensional continuous expression value into a 512-dimensional vector, and is preceded by 0.2 dropout and followed by layer normalization. The outputs of G and E are summed and passed into T, which consists of 12 stacked Transformer blocks, each with eight attention heads. We bypass the final decoder layers of scGPT and pass the mean of all tokens (including CLS) from the last transformer layer into a 2-layer projection head $P : \mathbb{R}^{512} \to \mathbb{R}^{1024}$. During THREADS pretraining, all layers of the gene expression encoder are fine-tuned. G, E, and T are initialized from the `pancancer` checkpoint (pretrained on 5.7 million cells of various cancer types), while P is randomly initialized.

SNV and CNV encoder. SNV and CNV data represent unstructured data that can be challenging to encode. Here, we adopt a simple strategy of using a multi-hot encoding passed through a 4-layer MLP (1 input, 2 hidden, and 1 output) with ReLU activation and 0.2 dropout[19]. Details about the multi-hot encoding strategy are presented in the **Online Methods, Pretraining dataset curation**. The hidden dimension of the MLP is set to the input size (ie., 1673), and mapped to the final output dimension, 1024. Our gene mutation encoder has 10128068 parameters in total.

Pretraining protocol. We pretrained THREADS using $4 \times 80\text{GB}$ A100 GPUs. The model was trained with a batch size of 300 per GPU for a maximum of 101 epochs, with early stopping based on RankMe[56] (for details, see **Model Selection** below). We start with a 5-epoch linear warmup, gradually increasing the learning rate from 0 to 1×10^{-5}. After the warmup, we apply a cosine scheduler that decays the learning rate from 1×10^{-5} down to 1×10^{-8} by the end of training. The weight decay is set to 0.0001 throughout. To improve training efficiency and data diversity, we sample 512 patches per slide during each training iteration. The AdamW optimizer was employed with β values of (0.9, 0.999). Hyperparameters and training settings are provided in

Extended Data Table 50 and 51.

Model selection. During pretraining, assessing the quality of the latent space and knowing when to stop training can be challenging. Previous works have relied on monitoring the downstream task performance at regular intervals, e.g., at the end of each epoch. However, this evaluation can be computationally intensive and result in optimizing testing performance during training, which risks artificially inflating the results. Following prior work[10, 10], we instead assess the expressivity of the embedding space by computing the smooth rank[56] of all slide embeddings within the training dataset after each epoch. Intuitively, a higher rank indicates greater diversity among the patch embeddings, ensuring that the representations have not collapsed into a limited number of modes, *i.e.,* a low-rank space. Here, we compute the rank as the entropy of the d (assuming $d < n$) L1-normalized singular values of the slide embedding matrix $H \in \mathcal{R}^{n \times d}$, which can be expressed as:

$$\text{RankMe}(H) = \exp(-\sum_{k=1}^{d} p_k \log(p_k)), \ \ p_k = \frac{\sigma_k(H)}{\sum_{k=1}^{n} |\sigma_k(H)|} + \epsilon \quad (3)$$

where σ_k denotes the kth singular value of H (sorted from large to low), and ϵ is small constant set to $1e-7$ for numerical stability. A model checkpoint is saved if the rank of the training dataset increases. Rank monitoring is started after the initial learning rate ramp-up.

Slide embedding extraction. For evaluation, we take the final output of the slide encoder with dimensionality 1024 (**Eq. 2**). For patients with multiple WSIs, we provided the union of all patch embeddings from all WSIs (belonging to that patient) to THREADS, resulting in a patient-level embedding. All patches in a slide are used while extracting slide embedding, *i.e.,* no patch sampling is done Slide embeddings were extracted using bf16 precision on 1×24GB NVIDIA 3090Ti.

Finetuning. We finetune THREADS slide encoder using the AdamW optimizer with a base learning rate of 0.000025. We do not apply weight decay, layer-wise learning rate decay, or gradient accumulation. All THREADS finetuned models are trained with a weighted cross-entropy loss for five epochs with a batch size of 1, sampling 2048 patches per batch. We use the same learning rate scheduler as GIGAPATH. No early stopping was applied; the final model used is the one obtained after five epochs. All fine-tuning experiments were conducted on a 1×24GB NVIDIA 3090Ti with bf16 precision.

Baselines

THREADS is compared against four types of baselines: GIGAPATH, PRISM, CHIEF and attention-based multiple instance learning (ABMIL).

GIGAPATH

Slide encoder. GIGAPATH [6] is a slide encoder model consisting of a pretrained patch encoder and a pretrained slide encoder. The patch encoder is a Vision Transformer (ViT) pretrained on 171,000+ pan-cancer WSIs from over 30,000 patients using DINOv2[21]. The slide encoder was trained using a LongNet[22] model with masked autoencoding. 1536-dimensional patch features were extracted using the GIGAPATH patch encoder on 256×256 patches at 20× magnification. We employed the official demo of GIGAPATH for extracting slide embeddings[1] and used the slide encoder checkpoint from Huggingface, resulting in a 768-dimensional

---

[1]https://github.com/prov-gigapath/prov-gigapath/blob/main/gigapath/slide_encoder.py

pooled embedding. For patients with multiple WSIs, we processed each WSI separately and averaged the WSI embeddings. Slide embeddings were extracted using fp16 precision on 1×24GB NVIDIA 3090Ti.

Finetuning. To finetune GIGAPATH slide encoder, we follow the official recipe: The patch encoder is kept frozen, and we initialized GIGAPATH slide encoder using the HuggingFace checkpoint and added a randomly initialized linear classification head. As per the official codebase[2], we finetuned GIGAPATH for all tasks using a batch size of 1, the AdamW optimizer, effective base learning rate of 0.00025, gradient accumulation over 32 steps, weight decay of 0.05, layer-wise learning rate decay of 0.95, and 5 epochs, without early stopping. We use a learning rate scheduler with half-cycle cosine decay after a one-epoch linear warmup up to the base learning rate. All patches were provided (no sampling) during training or testing. GIGAPATH finetuning was performed using 1×80GB NVIDIA A100 with fp16 precision.

Mean pooling. We additionally define a baseline where we average GIGAPATH patch embeddings, resulting into a slide embedding.

PRISM

Slide encoder. PRISM is a vision-language slide encoder that uses Virchow (ViT-H/14)[7] patch encoding. Virchow was trained on $> 2 \times 10^9$ patches (from over 100,000 patients) using DINOv2[21]. PRISM was then trained using a Perceiver[20] model with CoCa[52] on 587,000 WSIs-clinical reports pairs. Following the official Virchow demo[3], we first tiled all WSIs into 256×256-pixel patches at 20× and then used the publicly available patch encoder from HuggingFace to extract 2560-dimensional patch features. We then used the official PRISM codebase[4] to aggregate the patch embeddings of each WSI into a 1280-dimensional slide encoding. For patients with multiple WSIs, we provided the union of all patch embeddings from all WSIs (belonging to that patient) to PRISM. Slide embeddings are extracted using fp16 precision on 1×24GB NVIDIA 3090Ti.

Mean pooling. We additionally define a baseline where we average VIRCHOW patch embeddings into a slide embedding.

CHIEF

Slide embedding extraction. CHIEF [11] is an ABMIL-based slide encoder model that uses 768-dimensional CTRANSPATH patch embeddings[23]. CHIEF was trained using contrastive learning by aligning the slide representation with a text embedding of the tissue site. Following the official implementation, we compute 768-dimensional CHIEF pooled embeddings by passing CTransPath patch embeddings into the CHIEF slide encoder. For patients with multiple WSIs, we provided the union of all patch embeddings from all WSIs (belonging to that patient) to CHIEF. Slide embeddings are extracted using fp32 precision on 1×24GB NVIDIA 3090Ti.

CHIEF finetuning We employ the same finetuning recipe as in THREADS as both models are based on attention-based multiple instance learning. CHIEF finetuning was performed using 1×24GB NVIDIA 3090Ti with fp32 precision.

---

[2]https://github.com/prov-gigapath/prov-gigapath/blob/main/scripts/run_panda.sh
[3]https://huggingface.co/paige-ai/Virchow
[4]https://huggingface.co/paige-ai/Prism

Mean pooling. We additionally define a baseline where we average CTRANSPATH patch embeddings into a slide embedding.

RESNET50-IN

Mean pooling. We first extract patch embeddings using a ResNet50[57] model trained on ImageNet[58] (IN). We define a baseline where we average RESNET50-IN patch embeddings into a slide embedding.

Attention-based multiple instance learning

We use the widely employed attention-based multiple instance learning (ABMIL) architecture. ABMIL assigns patch-level importance scores using a single-headed non-gated attention mechanism. These attention scores are used to weight the patch embeddings, which are then summed to derive a slide embedding used for classification. The model is designed with a pre-attention linear layer which preserves the dimensionality of the input patch features (768 for CONCHv1.5) with GELU activation and 0.1 dropout, an attention network with two layers (hidden dimension 512) where the first layer has `tanh` activation and 0.25 dropout, and a post-attention linear layer with GELU activation and 0.1 dropout. The ABMIL models were trained for 20 epochs with a batch size of 1, a learning rate of 0.0003 using a cosine scheduler, and the AdamW optimizer with a weight decay of 1×10^{-5}. We use the final checkpoint for evaluation without early stopping. During training, we randomly sample 2048 patch features from each WSI. In patient-level tasks, if a patient has more than one WSI, then sampling is done from the union of all patches from those WSIs. During testing, all patches are provided to the model. For classification problems, we use a balanced cross-entropy loss. For survival prediction, we use a negative log-likelihood (NLL) loss adapted for survival prediction [19].

Evaluation

Linear probing

In-domain. In classification tasks, we employ linear probing evaluation based on scikit-learn. We use a fixed cost set to 0.5, `lbfgs` solver, a maximum of 10,000 training iterations, and balanced class weights. Since we do not include a validation set, we did not perform any hyperparameter search, e.g. over the cost. Post-hoc evaluation of the impact of the cost is reported in **Extended Data Table 38** and **Extended Data Figure 4**. To ensure fair comparisons, this evaluation recipe is applied to all sets of pooled features (THREADS, PRISM, GIGAPATH, CHIEF, and respective MEAN POOLING baselines). In survival tasks, we use the CoxNet model from `sksurv`[59], training all models for 10,000 iterations. We set the α parameter of the CoxNet model to 0.07 for overall survival prediction tasks and to 0.01 for progression-free survival prediction tasks. We make the following exceptions to ensure convergence: in CPTAC-CCRCC overall survival prediction, we set α to 0.01 in CHIEF, and in BOEHMK progression-free survival prediction, we set α to 0.02 in PRISM.

Out-of-domain (transferability). We run several sets of experiments to evaluate whether linear classifiers trained on one dataset can generalize to the same task on another dataset (**Extended Data Table 39**. We use the same setup as above and train on all samples of the "train" dataset and evaluate on a single fold containing all samples of the "test" dataset. We evaluate the performance over 100 bootstraps of the test set outputs.

Retrieval

We evaluate retrieval on tasks: cancer type retrieval (**Extended Data Table 45**) and EBRAINS fine and coarse subtype retrieval (**Extended Data Table 44**). For cancer type retrieval, we use 10 CPTAC cohorts: BRCA, CCRCC, COAD, HNSC, LUAD, LUSC, PDAC, GBM, OV, and UCEC. Each WSI within a cohort is labeled as the cancer type associated with that cohort. We use L2 distance as the similarity metric and simply compare the raw slide embedding of each slide in the test set with that of each slide in the training set. We consider up to the top 10 most similar retrieved slides and assess whether their class label matches that of the query slide. We compute mean average precision at k (mAP@k) for $k \in \{1, 5, 10\}$. Consider a set of n queries where for each query i (where $i = 1, 2, \ldots, n$), y_i is the true label of the i-th query and r_{ij} is the label of the retrieved item at rank j for query i (where $j = 1, 2, \ldots, k$). Then, mAP@k is computed as:

$$\text{mAP@}k = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{1}{k} \sum_{j=1}^{k} \delta_{r_{ij}, y_i} \cdot \frac{\sum_{s=1}^{j} \delta_{r_{is}, y_i}}{j} \right), \quad \text{where}$$

$$\delta_{a,b} = \begin{cases} 1 & \text{if } a = b, \\ 0 & \text{if } a \neq b. \end{cases}$$

Prompting

In prompting, a training dataset of RNA expression profiles with task labels is encoded using an RNA expression encoder (scGPT from THREADS pretraining), producing class prompts by averaging profiles within each class. For testing with an independent dataset, a WSI is encoded via the THREADS WSI encoder, and the L2 distance to all class prompts from the training set is computed. The final prediction is the class of the prompt nearest to the WSI embedding. For classification, all samples within each class are used to construct prompts. In survival prediction, prompts are based on the top and bottom 25% of uncensored patients ranked by survival time, representing low- and high-risk categories. The risk score at test time is defined as the distance between the WSI query and high-risk prompt minus the distance between the WSI query and low-risk prompt. In prompting with CONCHv1.5 MEAN POOLING, WSI embeddings are formed by averaging patch embeddings and further averaged by class to create prompts. At test time, classification uses the distance between the WSI prompt and query slide embedding, while survival prediction applies the same approach, substituting high- and low-risk molecular prompts with MEAN POOLING WSI prompts.

Metrics

For binary classification tasks, we report macro-AUC. For multi-class subtyping tasks, we report balanced accuracy, and for multi-class grading tasks, we report quadratic weighted kappa score. For survival prediction tasks, we report the concordance index (c-index). **macro-AUC** is a threshold-free measure that computes the area under the receiver operating curve that plots the true positive rate against the false positive rate as the classification threshold changes. **Balanced accuracy** takes the class imbalance in the evaluation set into account by computing the unweighted average of the recall of each class. **Quadratic weighted Cohen's kappa** quantifies the agreement between two annotators (e.g., ground truth and model prediction) on a classification problem, adjusting for chance agreement and penalizing based on the distance between categories. The score ranges from -1 (complete disagreement) to 1 (perfect agreement), with 0 indicating chance-level agreement. **Concordance index** (C-index) evaluates the predictive accuracy of a risk model in survival analysis by considering the order of predicted risks and actual event times. It calculates the proportion of pairs where the individual with higher predicted risk either experiences the event earlier or is censored later. The C-index ranges from 0.5 (random) to 1 (perfect prediction). For a clear overview of the metric used for each task, see column "Metric" in **Extended Data Tables 2 to 7**).

Statistical analysis

For all tasks with more than one test fold, we report the mean and standard error across all folds of the corresponding evaluation metric. For tasks with only a single test fold, we estimate 95% confidence intervals with non-parametric bootstrapping using 100 bootstrap replicates.

To assess the performance of all baselines on a specific task, we first performed a two-way Analysis of Variance (ANOVA), where the null hypothesis was that mean performance values did not differ across methods. We leveraged consistent evaluation folds that enabled direct comparisons across methods. If the ANOVA showed there was a statistically significant result (*i.e.,* P-value<0.05), a post-hoc two-sided, one-way Tukey's Honest Significant Difference (HSD) test was conducted to determine which pairs of methods differed significantly. Tukey's HSD test performs adjustment for multiple comparisons by default, so reported P-values have been adjusted for multiple comparisons[60]. Log-rank tests were used to compare Kaplan-Meier curves for statistically significant differences[19].

In addition to comparing performance on individual tasks, we aimed to assess which model performs best across all tasks (the full benchmark and each family of tasks). To compare models on the full benchmark, we fit a mixed-effects model on the data, estimating the baseline performance while accounting for the random effect of each dataset. Contrasts between individual baselines were made using pairwise comparisons using the estimated marginal means approach [61,62]. Briefly, after fitting our mixed-effects model with dataset as a random effect and model type as a fixed effect, we calculate the estimated marginal means for each model. We then performed pairwise comparisons between models using Tukey's method to adjust for multiple comparisons. A similar methodology was applied for each of the four families of tasks.

For all few-shot settings (**Extended Data Fig. 6**), we report results using box plots that indicate quartile values of model performance with whiskers extending to data points within 1.5× the interquartile range.

Computing hardware and software

We used Python (version 3.10.12) and PyTorch (version 2.3.0, CUDA 12.3) (`https://pytorch.org/`) for all experiments and analyses in the study (unless otherwise specified), which can be replicated using open-source libraries as outlined below. To train THREADS in a CLIP-style manner, we modified the original CLIP algorithm implemented by `https://github.com/mlfoundations/open_clip`. We used the implementation of scGPT from `https://github.com/bowang-lab/scGPT`. For pretraining, we used 4× 80GB NVIDIA A100 GPUs configured for multi-GPU training using distributed data-parallel (DDP). All other computations for downstream experiments were performed on single 24GB NVIDIA 3090 GPUs. All WSI processing was supported by OpenSlide (version 4.3.1), openslide-python (version 1.3.1), and Trident (`https://github.com/mahmoodlab/trident`), which additionally requires Pillow (version 10.2.0), segmentation-models-pytorch (version 0.0.3), and opencv-python (version 4.10.0.84). We use Scikit-learn[63] (version 1.5.0), Scikit-survival (version 0.23.0), and faiss (version 1.8.0) for training downstream machine learning models, specifically Logistic regression, and Cox PH. Implementations of other slide encoders benchmarked in the study are found at the following links: GigaPath (`https://github.com/prov-gigapath/prov-gigapath`), which additionally required fairscale (version 0.4.13), flash-attn (version 2.5.8), and ninja (version 1.11.1.1), PRISM (`https://huggingface.co/paige-ai/Prism`), and CHIEF (`https://github.com/hms-dbmi/CHIEF`). Matplotlib (version 3.8.4) and Seaborn (version 0.13.2) were used to create plots and figures. Usage of other miscellaneous Python libraries is listed in the **Reporting Summary**.

Code availability

Preprocessing code to (i) segment tissue from background, (ii) whole-slide image patching, and (iii) patch embedding extraction for CONCHv1.5, CTRANSPATH, GIGAPATH, and VIRCHOW can be accessed at `https://github.com/mahmoodlab/trident`. Code to run our benchmark can be accessed from `https://github.com/mahmoodlab/patho-bench`. Access to curated labels of publicly available cohorts, and data splits employed in the study can be found at `https://huggingface.co/datasets/MahmoodLab/patho-bench`. THREADS model weights and code to extract slide embeddings will be released upon publication.

Data availability

MBTG-47K: TCGA imaging data can be accessed through the NIH genomic data commons (`https://portal.gdc.cancer.gov`). TCGA transcriptomics data can be accessed through the Xena Hub (`https://xenabrowser.net/`). GTEx imaging and transcriptomics data can be accessed through the GTEx portal (`https://www.gtexportal.org/home/`). Pretraining data from BWH and MGH are proprietary patient data, and cannot be made publicly available.

Benchmark: Download links to access publicly available cohorts included as part of our benchmark are reported in **Extended Data Table 8**. Curated labels can be accessed via the THREADS-Benchmarking GitHub repository. In-house cohorts cannot be made publicly available.

Author contributions

A.V., A.Z., G.J. conceived the study and designed the experiments. A.V., A.Z., G.J., A.H.S., C.A.P., P.D., S.W., collected the data for pretraining. D.L. and K.L. assisted with curating downstream tasks. A.V., A.Z., G.J. T.D., M.Y.L. performed model development. A.V., A.Z., G.J., P.D., M.Y.L., T.D. R.C organized the datasets and codebase for all downstream tasks. A.V., A.Z., G.J., P.D., S.W., T.D. performed experimental analysis. A.V., A.Z., G.J., A.H.S., D.L., K.L. G.G., L.P.L. interpreted the experimental results and provided feedback on the study. A.V., A.Z., G.J. prepared the manuscript with input from all co-authors. H.R. performed statistical analysis. F.M. supervised the research.

Acknowledgments

Extended Data Figure 1: **Detailed architecture of THREADS.** THREADS employs a multimodal contrastive learning approach to align a whole-slide image representation with its corresponding molecular profile, obtained either using a DNA or RNA assay. **a.** The vision encoding branch uses a multihead attention-based model to pool patch embeddings into a slide embedding. **b.** The RNA encoding branch uses an scGPT model pretrained on 5.7 million cells of various cancer types, which is fully fine-tuned to yield a transcriptome embedding. **c.** The DNA encoding branch uses a multilayer perceptron (MLP) to transform copy number variations (CNV), insertions and deletions (indels), and single nucleotide variants (SNV) into a genomic embedding. WSI: whole-slide image; ViT: vision transformer; concat.: concatenations; TPM: transcripts per million.

Extended Data Figure 2: **Clustering capabilities of THREADS.** 2-dimensional tSNE representation of CPTAC cohort stratified by cancer type (n=10 cancer types) using THREADS (**a.**), PRISM (**b.**), GigaPath (**c.**), and CHIEF (**d.**). 2-dimensional tSNE representation of EBRAINS cohort stratified by tumor type (n=12 tumor types) using THREADS (**e.**), PRISM (**f.**), GigaPath (**g.**), and CHIEF (**g.**). ARI: Adjusted random index; MI: Mutual information; tSNE: t-distributed stochastic neighbor embedding.

Extended Data Figure 3: **Impact of regularization in linear probing evaluation in our benchmark (54 tasks). a.** Evolution of the average performance when varying the cost (inverse of regularization strength) in linear probing evaluation. **b.** Percentage of tasks where each baseline performs best based on the cost in linear probing evaluation. Adaptive regularization computes regularization cost by taking the embedding dimension times the number of classes normalized by 100[64].

Extended Data Figure 4: **Few shot performance of THREADS compared to baselines.** k refers to the number of training samples per class. **a.** Temozolomide response prediction in glioblastoma. **b.** Coarse-grained brain tumor subtyping in EBRAINS dataset. **c.** Fine-grained breast tumor subtyping in BRACS dataset. **d.** Coarse-grained breast tumor subtyping in BRACS dataset. **e.** Estrogen receptor status prediction in BCNB. **f.** Progesterone receptor status prediction in BCNB. Boxes indicate quartile values of model performance (n=5 runs), and whiskers extend to data points within 1.5-fold the interquartile range.

Extended Data Figure 5: **Transferability of THREADS.** THREADS and baselines are trained on one cohort and tested on an independent unseen test cohort. All test cohorts are independent of the MBTG-47K pretraining cohort. **a.** Clear cell renal cell carcinoma (ccRCC) BAP1 mutation prediction (CPTAC → MUT-HET-RCC). **b.** ccRCC PBRM1 mutation prediction (CPTAC → MUT-HET-RCC). **c.** Glioblastoma and low-grade glioma (GBMLGG) IDH mutation prediction (TCGA → EBRAINS). **d.** Invasive breast cancer (BRCA) estrogen receptor (ER) status prediction (TCGA → BCNB). **e.** BRCA progesterone receptor (ER) status prediction (TCGA → BCNB). **f.** BRCA subtype prediction (TCGA → MGB-Breast). **g.** NSCLC (lung adenocarcinoma, LUAD and lung squamous cell carcinoma LUSC) subtyping prediction (TCGA → MGB-Lung). **h.** LUAD overall survival prediction (TCGA → CPTAC). **i.** Pancreatic adenocarcinoma (PDAC) overall survival prediction (TCGA → CPTAC). **j.** Kaplan Meier (KM) curve of THREADS for PDAC overall survival prediction (TCGA → CPTAC). **k.** KM curve for GigaPath overall survival prediction (TCGA → CPTAC). Error bars represent 95% confidence intervals and the centers correspond to computed values of each metric. In Kaplan-Meier curves, line shows value and shaded region shows 95% confidence interval. Task-wise P-values were determined using two-sided Tukey Honest Significance Difference tests accounting for multiple comparisons following a positive result (P<0.05) of a two-way ANOVA. Statistical significance across multiple tasks (e.g., for each family) was assessed using a mixed-effects model. In Kaplan-Meier curves, P-values correspond to log-rank tests. P<0.05: *, P<0.01: **, P<0.001: ***.

31

Extended Data Figure 6: **Survival analysis of THREADS and baselines.** Kaplan Meier survival plots of THREADS (**a,b,c**), PRISM (**d,e,f**), GigaPath (**g,h,i**), and CHIEF (**j,k,l**) tested on pancreatic adenocarcinoma (PDAC), clear cell renal cell carcinoma (ccRCC), and colon adenocarcima (COAD). The shaded region highlights 95% confidence intervals. P-values for Kaplan Meier curves were obtained using log-rank statistical testing.

Extended Data Table 1: **Tissue Type Distribution in MBTG-47k.** THREADS-pretraining consists of 47,171 WSIs across 40 major tissue types collected from Massachusetts General Hospital (MGH), Brigham & Women's Hospital (BWH), The Cancer Genome Atlas (TCGA), and the Genotype-Tissue Expression (GTEx) consortium. Primary organs defined according to the highest level of the oncotree classification.

Primary organ	Number of slides				
	BWH	GTEx	MGH	TCGA	Total
Adipose	0	879	0	0	879
Adrenal Gland	182	162	8	387	739
Ampulla Of Vater	23	0	12	0	35
Artery	0	882	0	0	882
Biliary Tract	97	0	112	36	245
Bladder/Urinary Tract	596	0	97	451	1144
Blood	5	0	0	0	5
Bone	419	0	18	4	441
Bowel	2150	691	661	373	3875
Breast	457	327	256	1124	2164
Cervix	110	0	20	275	405
CNS/Brain	2394	646	774	1080	4894
Esophagus/Stomach	0	1202	276	0	1478
Eye	20	0	24	80	124
Head And Neck	549	129	449	470	1597
Heart	0	620	0	1	621
Kidney	508	64	114	885	1571
Liver	1825	0	172	337	2334
Lung	1264	383	1620	1042	4309
Lymph	1740	0	0	138	1878
Lymphoid	0	0	2	0	2
Muscle	0	563	0	0	563
Ovary/Fallopian Tube	820	116	173	71	1180
Pancreas	587	201	425	196	1409
Penis	6	0	0	0	6
Peripheral Nervous System	16	436	0	12	464
Peritoneum	469	0	27	2	498
Pleura	613	0	0	86	699
Prostate	474	165	127	447	1213
Skin	686	988	542	322	2538
Soft Tissue	1509	0	217	614	2340
Spleen	0	158	2	2	162
Stomach	1037	0	0	544	1581
Testis	88	245	1	256	590
Thorax	0	0	0	3	3
Thymus	38	0	6	142	186
Thyroid	401	465	377	518	1761
Unknown	431	0	230	0	661
Uterus	939	89	129	310	1467
Vulva/Vagina	103	96	28	1	228
Total	20556	9507	6899	10209	47171

Extended Data Table 2: **Summary of morphological subtyping prediction tasks.** All are WSI-level prediction tasks. **BA**: balanced accuracy, **AUC**: area under the receiver operating characteristic curve.

| Datasource | Subtyping Task | Organ | Unit | # Patients | # WSIs | # Classes | Metric | *k*=All Splits | | |
								Train:Test	Official?	# Folds
MGB-BRCA	IDC vs. ILC	Breast	WSI	1264	1264	2	AUC	1011:253	No	5
MGB-Lung	LUAD vs. LUSC	Lung	WSI	1939	1939	2	AUC	1551:388	No	5
EBRAINS	Coarse	Brain	WSI	2147	2319	12	BA	1746:573	Yes	1
EBRAINS	Fine	Brain	WSI	2147	2319	30	BA	1746:573	Yes	1
BRACS	Coarse	Breast	WSI	189	547	3	BA	396:151	No	5
BRACS	Fine	Breast	WSI	189	547	7	BA	396:151	No	5

Extended Data Table 3: **Summary of tumor grading prediction tasks.** Both are WSI-level prediction tasks. **QWK**: quadratic weighted kappa.

| Datasource | Organ | Unit | # Patients | # WSIs | # Classes | Metric | *k*=All Splits | | |
							Train:Test	Official?	# Folds
PANDA	Prostate	WSI	9555	9555	6	QWK	7647:954	Yes	1
IMP	Colon	WSI	5333	5333	3	QWK	4433:900	Yes	1

Extended Data Table 4: **Summary of molecular subtyping tasks.** All are patient-level prediction tasks. % Positive refers to percentage of samples with positive marker. **AUC**: area under the receiver operating characteristic curve.

| Datasource | Marker | Organ | Unit | # Patients | # WSIs | % Positive | Metric | *k*=All Splits | | |
								Train:Test	Official?	# Folds
MGB-BRCA	ER	Breast	Patient	874	874	70.1%	AUC	699:175	No	5
MGB-BRCA	PR	Breast	Patient	874	874	57.7%	AUC	699:175	No	5
MGB-BRCA	HER2	Breast	Patient	816	816	18.5%	AUC	652:164	No	5
BCNB	ER	Breast	Patient	1058	1058	78.5%	AUC	846:212	No	5
BCNB	PR	Breast	Patient	1058	1058	74.7%	AUC	846:212	No	5
BCNB	HER2	Breast	Patient	1058	1058	26.2%	AUC	846:212	No	5
MGB-Lung	TTF-1	Lung	Patient	488	488	67.0%	AUC	390:98	No	5
MGB-Lung	P40	Lung	Patient	185	185	38.9%	AUC	148:37	No	5
MGB-Lung	P63	Lung	Patient	153	153	52.9%	AUC	122:31	No	5
MGB-Lung	Napsin A	Lung	Patient	126	126	52.4%	AUC	100:26	No	5
MGB-Lung	CDX-2	Lung	Patient	79	79	30.4%	AUC	63:16	No	5
MGB-Lung	CK5/6	Lung	Patient	58	58	50.0%	AUC	46:12	No	5

Extended Data Table 5: **Summary of mutation prediction tasks.** All are patient-level prediction tasks. **AUC:** area under the receiver operating characteristic curve.

| | | | | | | | | | k=All Splits | | |
Datasource	Gene	Organ	Unit	# Patients	# WSIs	% Mutated	Metric	Train:Test	Official?	# Folds
CPTAC-BRCA	PIK3CA	Breast	Patient	103	112	35.9%	AUC	83:20	No	50
CPTAC-BRCA	TP53	Breast	Patient	103	112	40.8%	AUC	83:20	No	50
CPTAC-CCRCC	BAP1	Kidney	Patient	103	245	16.5%	AUC	83:20	No	50
CPTAC-CCRCC	PBRM1	Kidney	Patient	103	245	45.6%	AUC	83:20	No	50
CPTAC-COAD	KRAS	Colon	Patient	94	98	38.3%	AUC	76:18	No	50
CPTAC-COAD	TP53	Colon	Patient	94	98	66.0%	AUC	76:18	No	50
CPTAC-GBM	EGFR	Brain	Patient	99	243	24.2%	AUC	80:19	No	50
CPTAC-GBM	TP53	Brain	Patient	99	243	32.3%	AUC	80:19	No	50
CPTAC-HNSC	CASP8	Head & Neck	Patient	107	256	10.3%	AUC	86:21	No	50
CPTAC-LSCC	KEAP1	Lung	Patient	108	304	12.0%	AUC	87:21	No	50
CPTAC-LSCC	ARID1A	Lung	Patient	108	304	12.0%	AUC	87:21	No	50
CPTAC-LUAD	EGFR	Lung	Patient	108	324	36.1%	AUC	87:21	No	50
CPTAC-LUAD	STK11	Lung	Patient	108	324	16.7%	AUC	87:21	No	50
CPTAC-LUAD	TP53	Lung	Patient	108	324	59.3%	AUC	87:21	No	50
CPTAC-PDAC	SMAD4	Pancreas	Patient	105	242	19.0%	AUC	84:21	No	50
MUT-HET-RCC	BAP1	Kidney	Patient	1291	1291	12.5%	AUC	1032:259	No	5
MUT-HET-RCC	PBRM1	Kidney	Patient	1291	1291	51.8%	AUC	1032:259	No	5
MUT-HET-RCC	SETD2	Kidney	Patient	1291	1291	27.0%	AUC	1032:259	No	5
SURGEN	BRAF	Colon	Patient	388	388	10.8%	AUC	310:78	No	5
SURGEN	RAS	Colon	Patient	389	389	35.5%	AUC	311:78	No	5
SURGEN	MMR	Colon	Patient	389	389	7.7%	AUC	311:78	No	5

Extended Data Table 6: **Summary of treatment response and assessment tasks.** Unit refers to whether a task is a patient-level or WSI-level prediction task. % Positive refers to the fraction of cases either exhibiting positive (favorable) response or the criterion specified. Type refers to whether slides are resections, biopsies, or both. **AUC:** area under the receiver operating characteristic curve.

| | | | | | | | | | k=All Splits | | |
Datasource	Criterion	Organ	Type	Unit	# Patients	# WSIs	% Positive	Metric	Train:Test	Official?	# Folds
NADT-Prostate	Radiological response	Prostate	Resection	Patient	36	449	41.7%	AUC	29:7	No	50
OV-Bevacizumab	Biomarker response	Ovary	Resection	Patient	36	85	83.3%	AUC	29:7	No	50
GBM-Treatment	Clinical outcome	Brain	Biopsy	Patient	93	347	73.1%	AUC	75:18	No	50
POST-NAT-BRCA	Lymphovascular invasion	Breast	Resection	WSI	50	53	30.2%	AUC	43:10	No	50
MBC	Recist	Breast	Both	Patient	76	97	46.1%	QWK	61:15	No	50
BOEHMK	PFS	Ovary	Both	Patient	183	183	–	C-Index	146:37	No	5
MBC	OS	Ovary	Biopsy	Patient	75	96	–	C-Index	60:15	No	5

Extended Data Table 7: **Summary of survival prediction tasks.** All tasks are patient-level tasks. Censorship refers to incomplete observations due to end of study period or loss of follow-up. **OS:** duration of overall survival, **C-Index:** concordance index.

| | | | | | | | | | k=All Splits | | |
Datasource	Task	Organ	Unit	# Patients	# WSIs	% Censored	Survival (days)	Metric	Train:Test	Official?	# Folds
CPTAC-CCRCC	OS	Kidney	Patient	94	218	78.7%	1064 ± 634	C-Index	75:19	No	5
CPTAC-HNSC	OS	Head & Neck	Patient	102	243	67.6%	833 ± 423	C-Index	81:21	No	5
CPTAC-LUAD	OS	Lung	Patient	105	313	78.1%	753 ± 540	C-Index	84:21	No	5
CPTAC-PDA	OS	Pancreas	Patient	97	227	26.8%	561 ± 379	C-Index	77:20	No	5
SURGEN	OS	Colon	Patient	144	144	0.0%	854 ± 566	C-Index	115:29	No	5
SURGEN	Died within 5 years	Colon	Patient	387	387	–	–	AUC	309:78	No	5

Extended Data Table 8: **Publicly available datasets used for THREADS evaluation.**

Dataset	Link
EBRAINS[27]	https://doi.org/10.25493/WQ48-ZGX
BRACS[28]	https://www.bracs.icar.cnr.it/
PANDA[3]	https://panda.grand-challenge.org/data/
IMP[36]	https://rdm.inesctec.pt/dataset/nis-2023-008
BCNB[29]	https://bupt-ai-cz.github.io/BCNB/
CPTAC-BRCA[38]	https://www.cancerimagingarchive.net/collection/cptac-brca/
CPTAC-CCRCC[38]	https://www.cancerimagingarchive.net/collection/cptac-ccrcc/
CPTAC-COAD[38]	https://www.cancerimagingarchive.net/collection/cptac-coad/
CPTAC-GBM[38]	https://www.cancerimagingarchive.net/collection/cptac-gbm/
CPTAC-HNSC[38]	https://www.cancerimagingarchive.net/collection/cptac-hnsc/
CPTAC-LSCC[38]	https://www.cancerimagingarchive.net/collection/cptac-lscc/
CPTAC-LUAD[38]	https://www.cancerimagingarchive.net/collection/cptac-luad/
CPTAC-PDAC[38]	https://www.cancerimagingarchive.net/collection/cptac-pda/
MUT-HET-RCC	https://doi.org/10.25452/figshare.plus.c.5983795
OV-Bevacizumab[41]	https://www.nature.com/articles/s41597-022-01127-6
NADT-Prostate[42]	https://www.medrxiv.org/content/10.1101/2020.09.29.20199711v1.full
POST-NAT-BRCA	https://onlinelibrary.wiley.com/doi/10.1002/cyto.a.23244
BOEHMK	https://www.synapse.org/Synapse:syn25946117/wiki/611576
MBC	https://www.synapse.org/Synapse:syn59490671/wiki/628046
SURGEN	https://www.ebi.ac.uk/biostudies/bioimages/studies/S-BIAD1285

Extended Data Table 9: **Performance comparison between THREADS and baselines on MGB Breast tasks.** Best in **bold**, second best is underlined. **Subtype**: This is a slide-level classification task evaluated using AUROC, mean and standard error reported over 5-fold cross-validation. **ER**: This is a patient-level classification task evaluated using AUROC, mean and standard error reported over 5-fold cross-validation. **PR**: This is a patient-level classification task evaluated using AUROC, mean and standard error reported over 5-fold cross-validation. **HER2**: This is a patient-level classification task evaluated using AUROC, mean and standard error reported over 5-fold cross-validation.

	Model	Tasks			
		Subtype (↑) (n=1264)	ER (↑) (n=874)	PR (↑) (n=874)	HER2 (↑) (n=816)
Linear Probe	VIRCHOW [7] MEAN POOLING	0.971 ± 0.003	0.705 ± 0.016	0.672 ± 0.012	0.599 ± 0.020
	GIGAPATH [6] MEAN POOLING	0.979 ± 0.002	0.711 ± 0.012	0.682 ± 0.018	0.575 ± 0.014
	CHIEF [11] MEAN POOLING	0.955 ± 0.008	0.713 ± 0.024	0.721 ± 0.016	0.637 ± 0.019
	CONCHv1.5 MEAN POOLING	0.973 ± 0.005	0.728 ± 0.026	0.720 ± 0.012	0.656 ± 0.042
	PRISM [13]	0.985 ± 0.002	0.727 ± 0.022	0.639 ± 0.001	0.710 ± 0.021
	GIGAPATH [6]	0.975 ± 0.003	0.708 ± 0.014	0.689 ± 0.020	0.597 ± 0.010
	CHIEF [11]	0.978 ± 0.004	0.749 ± 0.026	0.732 ± 0.020	0.696 ± 0.018
	THREADS	0.983 ± 0.004	**0.784 ± 0.016**	**0.748 ± 0.021**	0.694 ± 0.031
Supervised	ABMIL	0.983 ± 0.003	0.700 ± 0.024	0.694 ± 0.005	0.648 ± 0.016
	GIGAPATH [6]	**0.989 ± 0.001**	0.747 ± 0.018	0.730 ± 0.016	0.663 ± 0.019
	CHIEF [11]	0.951 ± 0.005	0.696 ± 0.019	0.675 ± 0.017	0.567 ± 0.019
	THREADS Random Init	0.986 ± 0.003	0.771 ± 0.021	0.740 ± 0.014	0.685 ± 0.041
	THREADS	0.986 ± 0.002	0.771 ± 0.025	0.739 ± 0.022	**0.719 ± 0.030**

Extended Data Table 10: **Performance comparison between THREADS and baselines on MGB Lung tasks.** Best in **bold**, second best is underlined. **Subtype**: This is a slide-level classification task evaluated using AUROC, mean and standard error reported over 5-fold cross-validation. **TTF-1**: This is a patient-level classification task evaluated using AUROC, mean and standard error reported over 5-fold cross-validation. **P40**: This is a patient-level classification task evaluated using AUROC, mean and standard error reported over 5-fold cross-validation. **P63**: This is a patient-level classification task evaluated using AUROC, mean and standard error reported over 5-fold cross-validation. **Napsina**: This is a patient-level classification task evaluated using AUROC, mean and standard error reported over 5-fold cross-validation. **CDX-2**: This is a patient-level classification task evaluated using AUROC, mean and standard error reported over 5-fold cross-validation. **CK5-6**: This is a patient-level classification task evaluated using AUROC, mean and standard error reported over 5-fold cross-validation.

	Model	Tasks						
		Subtype (↑) (n=1939)	TTF-1 (↑) (n=488)	P40 (↑) (n=185)	P63 (↑) (n=153)	Napsina (↑) (n=126)	CDX-2 (↑) (n=79)	CK5-6 (↑) (n=58)
Linear Probe	VIRCHOW [7] MEAN POOLING	0.960 ± 0.002	0.837 ± 0.017	0.750 ± 0.033	0.640 ± 0.036	0.678 ± 0.050	0.529 ± 0.050	0.742 ± 0.051
	GIGAPATH [6] MEAN POOLING	0.969 ± 0.002	0.866 ± 0.008	0.739 ± 0.019	0.681 ± 0.027	0.693 ± 0.031	0.506 ± 0.092	0.697 ± 0.050
	CHIEF [11] MEAN POOLING	0.950 ± 0.002	0.750 ± 0.018	0.627 ± 0.013	0.642 ± 0.029	0.593 ± 0.024	0.487 ± 0.074	0.634 ± 0.087
	CONCHv1.5 MEAN POOLING	0.969 ± 0.003	0.836 ± 0.009	0.867 ± 0.015	0.748 ± 0.034	0.773 ± 0.044	0.588 ± 0.073	<u>0.884 ± 0.047</u>
	PRISM [13]	0.978 ± 0.005	0.858 ± 0.014	0.863 ± 0.017	0.809 ± 0.013	0.704 ± 0.015	**0.738 ± 0.032**	0.877 ± 0.022
	GIGAPATH [6]	0.957 ± 0.003	0.839 ± 0.013	0.696 ± 0.022	0.653 ± 0.034	0.634 ± 0.041	0.485 ± 0.093	0.670 ± 0.065
	CHIEF [11]	0.979 ± 0.003	0.822 ± 0.020	0.809 ± 0.039	0.792 ± 0.026	0.662 ± 0.022	0.558 ± 0.051	0.791 ± 0.049
	THREADS	0.982 ± 0.004	**0.895 ± 0.011**	<u>0.898 ± 0.027</u>	**0.879 ± 0.020**	0.817 ± 0.033	<u>0.725 ± 0.062</u>	**0.933 ± 0.026**
Supervised	ABMIL	0.980 ± 0.002	0.866 ± 0.016	0.860 ± 0.027	0.811 ± 0.023	**0.851 ± 0.013**	0.599 ± 0.047	0.813 ± 0.047
	GIGAPATH [6]	**0.988 ± 0.002**	0.863 ± 0.014	0.642 ± 0.042	0.664 ± 0.031	0.624 ± 0.036	0.475 ± 0.041	0.676 ± 0.088
	CHIEF [11]	0.951 ± 0.009	0.682 ± 0.018	0.632 ± 0.049	0.525 ± 0.031	0.510 ± 0.074	0.498 ± 0.090	0.693 ± 0.068
	THREADS Random Init	<u>0.987 ± 0.002</u>	0.882 ± 0.008	0.870 ± 0.023	0.796 ± 0.046	0.798 ± 0.020	0.610 ± 0.072	0.803 ± 0.061
	THREADS	0.985 ± 0.001	<u>0.887 ± 0.010</u>	**0.903 ± 0.025**	<u>0.859 ± 0.026</u>	<u>0.825 ± 0.018</u>	0.672 ± 0.031	0.880 ± 0.031

Extended Data Table 11: **Performance comparison between THREADS and baselines on BCNB[29] tasks.** Best in **bold**, second best is underlined. **ER**: This is a patient-level classification task evaluated using AUROC, mean and standard error reported over 5-fold cross-validation. **PR**: This is a patient-level classification task evaluated using AUROC, mean and standard error reported over 5-fold cross-validation. **HER2**: This is a patient-level classification task evaluated using AUROC, mean and standard error reported over 5-fold cross-validation.

	Model	Tasks		
		ER (↑) (n=1058)	PR (↑) (n=1058)	HER2 (↑) (n=1058)
Linear Probe	VIRCHOW [7] MEAN POOLING	0.888 ± 0.007	0.813 ± 0.010	0.707 ± 0.011
	GIGAPATH [6] MEAN POOLING	0.901 ± 0.013	0.828 ± 0.014	0.719 ± 0.015
	CHIEF [11] MEAN POOLING	0.852 ± 0.014	0.799 ± 0.021	0.703 ± 0.016
	CONCHv1.5 MEAN POOLING	0.889 ± 0.010	0.810 ± 0.022	0.737 ± 0.013
	PRISM [13]	0.892 ± 0.014	0.815 ± 0.019	0.711 ± 0.010
	GIGAPATH [6]	0.886 ± 0.016	0.811 ± 0.013	0.702 ± 0.016
	CHIEF [11]	0.883 ± 0.014	0.818 ± 0.017	0.719 ± 0.021
	THREADS	0.921 ± 0.009	0.837 ± 0.020	0.765 ± 0.008
Supervised	ABMIL	0.877 ± 0.009	0.804 ± 0.017	0.737 ± 0.011
	GIGAPATH [6]	<u>0.925 ± 0.012</u>	<u>0.856 ± 0.016</u>	0.766 ± 0.016
	CHIEF [11]	0.786 ± 0.007	0.753 ± 0.013	0.658 ± 0.029
	THREADS Random Init	**0.926 ± 0.010**	**0.859 ± 0.019**	<u>0.780 ± 0.008</u>
	THREADS	0.919 ± 0.011	0.848 ± 0.016	**0.786 ± 0.007**

Extended Data Table 12: **Performance comparison between THREADS and baselines on MUT-HET-RCC tasks.** Best in **bold**, second best is underlined. **BAP1 Mutation**: This is a patient-level classification task evaluated using AUROC, mean and standard error reported over 5-fold cross-validation. **PBRM1 Mutation**: This is a slide-level classification task evaluated using AUROC, mean and standard error reported over 5-fold cross-validation. **SETD2 Mutation**: This is a slide-level classification task evaluated using AUROC, mean and standard error reported over 5-fold cross-validation.

	Model	Tasks		
		BAP1 Mutation (↑) (n=1291)	PBRM1 Mutation (↑) (n=1291)	SETD2 Mutation (↑) (n=1291)
Linear Probe	VIRCHOW [7] MEAN POOLING	0.879 ± 0.003	0.810 ± 0.008	0.712 ± 0.012
	GIGAPATH [6] MEAN POOLING	0.880 ± 0.007	0.806 ± 0.007	0.701 ± 0.012
	CHIEF [11] MEAN POOLING	0.850 ± 0.015	0.746 ± 0.003	0.722 ± 0.007
	CONCHv1.5 MEAN POOLING	0.846 ± 0.011	0.782 ± 0.005	0.728 ± 0.011
	PRISM [13]	0.848 ± 0.017	0.797 ± 0.008	0.734 ± 0.009
	GIGAPATH [6]	0.857 ± 0.007	0.799 ± 0.003	0.708 ± 0.013
	CHIEF [11]	0.852 ± 0.013	0.781 ± 0.011	0.743 ± 0.012
	THREADS	0.873 ± 0.006	**0.826 ± 0.003**	**0.756 ± 0.009**
Supervised	ABMIL	0.858 ± 0.015	0.778 ± 0.011	0.717 ± 0.008
	GIGAPATH [6]	0.885 ± 0.005	0.816 ± 0.011	0.727 ± 0.009
	CHIEF [11]	0.802 ± 0.025	0.697 ± 0.005	0.669 ± 0.019
	THREADS Random Init	0.868 ± 0.015	0.779 ± 0.005	0.729 ± 0.006
	THREADS	**0.898 ± 0.005**	0.807 ± 0.011	0.738 ± 0.007

Extended Data Table 13: **Performance comparison between THREADS and baselines on IMP[36] tasks.** Best in **bold**, second best is underlined. **Grade**: This is a slide-level classification task evaluated using quadratic weighted kappa, mean and 95% CI reported over 100 bootstraps of a single fold.

	Model	Tasks
		Grade (↑) (n=5333)
Linear Probe	VIRCHOW [7] MEAN POOLING	0.913 (0.894-0.929)
	GIGAPATH [6] MEAN POOLING	0.903 (0.884-0.921)
	CHIEF [11] MEAN POOLING	0.877 (0.847-0.908)
	CONCHv1.5 MEAN POOLING	0.889 (0.862-0.912)
	PRISM [13]	0.935 (0.919-0.951)
	GIGAPATH [6]	0.903 (0.880-0.925)
	CHIEF [11]	0.914 (0.893-0.937)
	THREADS	0.919 (0.896-0.937)
Supervised	ABMIL	0.942 (0.926-0.957)
	GIGAPATH [6]	**0.956 (0.941-0.968)**
	CHIEF [11]	0.917 (0.893-0.937)
	THREADS Random Init	0.944 (0.922-0.964)
	THREADS	0.946 (0.929-0.963)

Extended Data Table 14: **Performance comparison between THREADS and baselines on PANDA[3] tasks.** Best in **bold**, second best is <u>underlined</u>. **ISUP Grade**: This is a slide-level classification task evaluated using quadratic weighted kappa, mean and 95% CI reported over 100 bootstraps of a single fold.

	Model	Tasks
		ISUP Grade (↑) (n=9555)
Linear Probe	VIRCHOW [7] MEAN POOLING	0.895 (0.875-0.913)
	GIGAPATH [6] MEAN POOLING	0.894 (0.871-0.915)
	CHIEF [11] MEAN POOLING	0.799 (0.765-0.827)
	CONCHV1.5 MEAN POOLING	0.845 (0.814-0.864)
	PRISM [13]	0.919 (0.901-0.935)
	GIGAPATH [6]	0.873 (0.850-0.895)
	CHIEF [11]	0.898 (0.878-0.918)
	THREADS	0.915 (0.899-0.929)
Supervised	ABMIL	<u>0.932 (0.919-0.944)</u>
	GIGAPATH [6]	**0.959 (0.951-0.967)**
	CHIEF [11]	0.798 (0.765-0.830)
	THREADS Random Init	0.926 (0.913-0.938)
	THREADS	0.930 (0.917-0.943)

Extended Data Table 15: **Performance comparison between THREADS and baselines on CPTAC-BRCA[38] tasks.** Best in **bold**, second best is <u>underlined</u>. **PIK3CA Mutation**: This is a patient-level classification task evaluated using AUROC, mean and standard error reported over 50-fold Monte Carlo. **TP53 Mutation**: This is a patient-level classification task evaluated using AUROC, mean and standard error reported over 50-fold Monte Carlo.

	Model	Tasks	
		PIK3CA Mutation (↑) (n=103)	TP53 Mutation (↑) (n=103)
Linear Probe	VIRCHOW [7] MEAN POOLING	0.569 ± 0.015	0.766 ± 0.012
	GIGAPATH [6] MEAN POOLING	0.554 ± 0.016	0.754 ± 0.012
	CHIEF [11] MEAN POOLING	<u>0.615 ± 0.015</u>	0.786 ± 0.012
	CONCHV1.5 MEAN POOLING	0.513 ± 0.017	0.796 ± 0.011
	PRISM [13]	0.575 ± 0.014	0.787 ± 0.013
	GIGAPATH [6]	0.531 ± 0.017	0.749 ± 0.012
	CHIEF [11]	**0.647 ± 0.013**	0.832 ± 0.012
	THREADS	0.571 ± 0.017	**0.876 ± 0.012**
Supervised	ABMIL	0.531 ± 0.017	0.791 ± 0.012
	GIGAPATH [6]	0.570 ± 0.016	0.775 ± 0.014
	CHIEF [11]	0.504 ± 0.023	0.620 ± 0.020
	THREADS Random Init	0.578 ± 0.016	0.833 ± 0.012
	THREADS	0.611 ± 0.015	<u>0.846 ± 0.012</u>

Extended Data Table 16: **Performance comparison between THREADS and baselines on CPTAC-CCRCC[38] tasks.** Best in **bold**, second best is underlined. **BAP1 Mutation**: This is a patient-level classification task evaluated using AUROC, mean and standard error reported over 50-fold Monte Carlo. **PBRM1 Mutation**: This is a patient-level classification task evaluated using AUROC, mean and standard error reported over 50-fold Monte Carlo.

	Model	Tasks	
		BAP1 Mutation (↑) (n=103)	PBRM1 Mutation (↑) (n=103)
Linear Probe	VIRCHOW [7] MEAN POOLING	0.624 ± 0.022	0.501 ± 0.016
	GIGAPATH [6] MEAN POOLING	0.688 ± 0.023	0.482 ± 0.015
	CHIEF [11] MEAN POOLING	0.717 ± 0.018	0.457 ± 0.016
	CONCHv1.5 MEAN POOLING	0.655 ± 0.016	0.518 ± 0.015
	PRISM [13]	0.664 ± 0.022	0.598 ± 0.019
	GIGAPATH [6]	0.670 ± 0.023	0.466 ± 0.016
	CHIEF [11]	0.745 ± 0.019	0.522 ± 0.015
	THREADS	**0.809 ± 0.014**	**0.668 ± 0.016**
Supervised	ABMIL	0.729 ± 0.020	0.576 ± 0.018
	GIGAPATH [6]	0.691 ± 0.020	0.512 ± 0.017
	CHIEF [11]	0.641 ± 0.024	0.514 ± 0.018
	THREADS Random Init	0.772 ± 0.015	0.595 ± 0.018
	THREADS	0.802 ± 0.015	0.629 ± 0.017

Extended Data Table 17: **Performance comparison between THREADS and baselines on CPTAC-COAD[38] tasks.** Best in **bold**, second best is underlined. **KRAS Mutation**: This is a patient-level classification task evaluated using AUROC, mean and standard error reported over 50-fold Monte Carlo. **TP53 Mutation**: This is a patient-level classification task evaluated using AUROC, mean and standard error reported over 50-fold Monte Carlo.

	Model	Tasks	
		KRAS Mutation (↑) (n=94)	TP53 Mutation (↑) (n=94)
Linear Probe	VIRCHOW [7] MEAN POOLING	0.661 ± 0.015	0.684 ± 0.017
	GIGAPATH [6] MEAN POOLING	0.642 ± 0.015	0.656 ± 0.018
	CHIEF [11] MEAN POOLING	0.652 ± 0.015	0.649 ± 0.019
	CONCHv1.5 MEAN POOLING	**0.742 ± 0.016**	0.690 ± 0.018
	PRISM [13]	0.554 ± 0.015	0.578 ± 0.019
	GIGAPATH [6]	0.654 ± 0.013	0.642 ± 0.016
	CHIEF [11]	0.649 ± 0.016	0.659 ± 0.018
	THREADS	0.704 ± 0.014	0.742 ± 0.016
Supervised	ABMIL	0.623 ± 0.017	0.730 ± 0.016
	GIGAPATH [6]	0.622 ± 0.015	0.648 ± 0.017
	CHIEF [11]	0.531 ± 0.017	0.526 ± 0.019
	THREADS Random Init	0.696 ± 0.015	0.698 ± 0.018
	THREADS	0.670 ± 0.016	**0.785 ± 0.014**

Extended Data Table 18: **Performance comparison between THREADS and baselines on CPTAC-GBM[38] tasks.** Best in **bold**, second best is <u>underlined</u>. **EGFR Mutation**: This is a patient-level classification task evaluated using AUROC, mean and standard error reported over 50-fold Monte Carlo. **TP53 Mutation**: This is a patient-level classification task evaluated using AUROC, mean and standard error reported over 50-fold Monte Carlo.

	Model	Tasks	
		EGFR Mutation (↑) (n=99)	TP53 Mutation (↑) (n=99)
Linear Probe	VIRCHOW [7] MEAN POOLING	0.509 ± 0.017	0.816 ± 0.012
	GIGAPATH [6] MEAN POOLING	0.623 ± 0.016	0.805 ± 0.012
	CHIEF [11] MEAN POOLING	0.662 ± 0.015	0.849 ± 0.011
	CONCHv1.5 MEAN POOLING	0.639 ± 0.014	0.828 ± 0.014
	PRISM [13]	0.619 ± 0.015	0.858 ± 0.015
	GIGAPATH [6]	0.634 ± 0.016	0.785 ± 0.014
	CHIEF [11]	0.743 ± 0.015	<u>0.862 ± 0.010</u>
	THREADS	<u>0.782 ± 0.011</u>	0.842 ± 0.013
Supervised	ABMIL	0.713 ± 0.016	0.836 ± 0.013
	GIGAPATH [6]	0.624 ± 0.014	0.698 ± 0.015
	CHIEF [11]	0.480 ± 0.020	0.519 ± 0.021
	THREADS Random Init	0.674 ± 0.012	0.832 ± 0.016
	THREADS	**0.791 ± 0.010**	**0.864 ± 0.012**

Extended Data Table 19: **Performance comparison between THREADS and baselines on CPTAC-HNSC[38] tasks.** Best in **bold**, second best is <u>underlined</u>. **CASP8 Mutation**: This is a patient-level classification task evaluated using AUROC, mean and standard error reported over 50-fold Monte Carlo.

	Model	Tasks
		CASP8 Mutation (↑) (n=107)
Linear Probe	VIRCHOW [7] MEAN POOLING	0.397 ± 0.025
	GIGAPATH [6] MEAN POOLING	0.497 ± 0.028
	CHIEF [11] MEAN POOLING	0.482 ± 0.025
	CONCHv1.5 MEAN POOLING	0.614 ± 0.028
	PRISM [13]	0.601 ± 0.027
	GIGAPATH [6]	0.474 ± 0.025
	CHIEF [11]	0.493 ± 0.027
	THREADS	**0.754 ± 0.019**
Supervised	ABMIL	0.681 ± 0.029
	GIGAPATH [6]	0.619 ± 0.029
	CHIEF [11]	0.558 ± 0.035
	THREADS Random Init	0.673 ± 0.022
	THREADS	<u>0.736 ± 0.023</u>

Extended Data Table 20: **Performance comparison between THREADS and baselines on CPTAC-LSCC[38] tasks.** Best in **bold**, second best is underlined. **KEAP1 Mutation**: This is a patient-level classification task evaluated using AUROC, mean and standard error reported over 50-fold Monte Carlo. **ARID1A Mutation**: This is a patient-level classification task evaluated using AUROC, mean and standard error reported over 50-fold Monte Carlo.

	Model	Tasks	
		KEAP1 Mutation (↑) (n=108)	ARID1A Mutation (↑) (n=108)
Linear Probe	VIRCHOW [7] MEAN POOLING	0.671 ± 0.020	0.417 ± 0.020
	GIGAPATH [6] MEAN POOLING	0.640 ± 0.017	0.437 ± 0.019
	CHIEF [11] MEAN POOLING	0.676 ± 0.016	0.524 ± 0.022
	CONCHv1.5 MEAN POOLING	0.574 ± 0.018	0.411 ± 0.020
	PRISM [13]	0.492 ± 0.018	0.449 ± 0.020
	GIGAPATH [6]	0.663 ± 0.018	0.465 ± 0.016
	CHIEF [11]	0.656 ± 0.016	0.535 ± 0.024
	THREADS	**0.685 ± 0.019**	**0.658 ± 0.023**
Supervised	ABMIL	0.459 ± 0.023	0.432 ± 0.023
	GIGAPATH [6]	0.614 ± 0.022	0.539 ± 0.022
	CHIEF [11]	0.446 ± 0.026	0.467 ± 0.024
	THREADS Random Init	0.629 ± 0.018	0.487 ± 0.020
	THREADS	0.608 ± 0.017	0.514 ± 0.022

Extended Data Table 21: **Performance comparison between THREADS and baselines on CPTAC-LUAD[38] tasks.** Best in **bold**, second best is underlined. **EGFR Mutation**: This is a patient-level classification task evaluated using AUROC, mean and standard error reported over 50-fold Monte Carlo. **STK11 Mutation**: This is a patient-level classification task evaluated using AUROC, mean and standard error reported over 50-fold Monte Carlo. **TP53 Mutation**: This is a patient-level classification task evaluated using AUROC, mean and standard error reported over 50-fold Monte Carlo.

	Model	Tasks		
		EGFR Mutation (↑) (n=108)	STK11 Mutation (↑) (n=108)	TP53 Mutation (↑) (n=108)
Linear Probe	VIRCHOW [7] MEAN POOLING	0.791 ± 0.015	0.864 ± 0.013	0.734 ± 0.014
	GIGAPATH [6] MEAN POOLING	0.782 ± 0.014	0.852 ± 0.013	0.715 ± 0.014
	CHIEF [11] MEAN POOLING	0.722 ± 0.017	0.765 ± 0.019	0.663 ± 0.014
	CONCHv1.5 MEAN POOLING	0.789 ± 0.014	0.824 ± 0.018	0.682 ± 0.015
	PRISM [13]	0.809 ± 0.013	0.854 ± 0.015	**0.755 ± 0.013**
	GIGAPATH [6]	0.791 ± 0.014	0.822 ± 0.015	0.737 ± 0.013
	CHIEF [11]	0.717 ± 0.017	0.824 ± 0.015	0.704 ± 0.014
	THREADS	**0.822 ± 0.011**	0.889 ± 0.015	0.752 ± 0.014
Supervised	ABMIL	0.748 ± 0.014	0.856 ± 0.017	0.686 ± 0.015
	GIGAPATH [6]	0.749 ± 0.017	0.791 ± 0.019	0.699 ± 0.015
	CHIEF [11]	0.495 ± 0.025	0.517 ± 0.024	0.524 ± 0.017
	THREADS Random Init	0.805 ± 0.015	0.862 ± 0.016	0.704 ± 0.015
	THREADS	0.798 ± 0.014	**0.891 ± 0.011**	0.752 ± 0.014

Extended Data Table 22: **Performance comparison between THREADS and baselines on CPTAC-PDAC[38] tasks.** Best in **bold**, second best is underlined. **SMAD4 Mutation**: This is a patient-level classification task evaluated using AUROC, mean and standard error reported over 50-fold Monte Carlo.

	Model	Tasks
		SMAD4 Mutation (↑) (n=105)
Linear Probe	VIRCHOW [7] MEAN POOLING	0.488 ± 0.020
	GIGAPATH [6] MEAN POOLING	0.418 ± 0.020
	CHIEF [11] MEAN POOLING	0.439 ± 0.018
	CONCHv1.5 MEAN POOLING	0.576 ± 0.019
	PRISM [13]	0.523 ± 0.022
	GIGAPATH [6]	0.423 ± 0.021
	CHIEF [11]	0.393 ± 0.018
	THREADS	<u>0.578 ± 0.024</u>
Supervised	ABMIL	0.512 ± 0.019
	GIGAPATH [6]	0.340 ± 0.018
	CHIEF [11]	0.478 ± 0.022
	THREADS Random Init	**0.598 ± 0.020**
	THREADS	0.576 ± 0.022

Extended Data Table 23: **Performance comparison between THREADS and baselines on BRACS[28] tasks.** Best in **bold**, second best is underlined. **Fine Subtype**: This is a slide-level classification task evaluated using balanced accuracy, mean and standard error reported over 5-fold cross-validation. **Coarse Subtype**: This is a slide-level classification task evaluated using balanced accuracy, mean and standard error reported over 5-fold cross-validation.

	Model	Tasks	
		Fine Subtype (↑) (n=547)	Coarse Subtype (↑) (n=547)
Linear Probe	VIRCHOW [7] MEAN POOLING	0.348 ± 0.015	0.618 ± 0.028
	GIGAPATH [6] MEAN POOLING	0.318 ± 0.016	0.569 ± 0.031
	CHIEF [11] MEAN POOLING	0.350 ± 0.029	0.621 ± 0.030
	CONCHv1.5 MEAN POOLING	0.378 ± 0.036	0.620 ± 0.036
	PRISM [13]	0.419 ± 0.014	0.659 ± 0.019
	GIGAPATH [6]	0.342 ± 0.019	0.606 ± 0.037
	CHIEF [11]	0.452 ± 0.015	0.704 ± 0.023
	THREADS	<u>0.481 ± 0.015</u>	0.716 ± 0.019
Supervised	ABMIL	0.478 ± 0.038	**0.740 ± 0.023**
	GIGAPATH [6]	0.444 ± 0.020	0.678 ± 0.018
	CHIEF [11]	0.228 ± 0.007	0.495 ± 0.039
	THREADS Random Init	0.456 ± 0.020	0.715 ± 0.020
	THREADS	**0.492 ± 0.013**	<u>0.731 ± 0.012</u>

Extended Data Table 24: **Performance comparison between THREADS and baselines on EBRAINS[27] tasks.** Best in **bold**, second best is underlined. **Fine Subtype**: This is a slide-level classification task evaluated using balanced accuracy, mean and 95% CI reported over 100 bootstraps of a single fold. **Coarse Subtype**: This is a slide-level classification task evaluated using balanced accuracy, mean and 95% CI reported over 100 bootstraps of a single fold.

	Model	Tasks	
		Fine Subtype (↑) (n=2319)	Coarse Subtype (↑) (n=2319)
Linear Probe	VIRCHOW [7] MEAN POOLING	0.708 (0.666-0.756)	0.873 (0.841-0.913)
	GIGAPATH [6] MEAN POOLING	0.740 (0.700-0.776)	0.903 (0.871-0.941)
	CHIEF [11] MEAN POOLING	0.637 (0.594-0.671)	0.785 (0.730-0.832)
	CONCHv1.5 MEAN POOLING	0.693 (0.646-0.733)	0.870 (0.836-0.910)
	PRISM [13]	0.721 (0.686-0.766)	0.871 (0.823-0.908)
	GIGAPATH [6]	0.722 (0.679-0.757)	0.887 (0.849-0.928)
	CHIEF [11]	0.660 (0.621-0.704)	0.840 (0.793-0.893)
	THREADS	0.742 (0.707-0.775)	**0.916 (0.873-0.951)**
Supervised	ABMIL	0.720 (0.684-0.758)	0.900 (0.865-0.929)
	GIGAPATH [6]	**0.751 (0.708-0.787)**	0.890 (0.856-0.923)
	CHIEF [11]	0.033 (0.033-0.033)	0.156 (0.150-0.163)
	THREADS Random Init	0.737 (0.697-0.774)	0.898 (0.858-0.933)
	THREADS	0.734 (0.692-0.769)	0.881 (0.838-0.920)

Extended Data Table 25: **Performance comparison between THREADS and baselines on OV-Bevacizumab[41] tasks.** Best in **bold**, second best is underlined. **Response**: This is a patient-level classification task evaluated using AUROC, mean and standard error reported over 50-fold Monte Carlo.

	Model	Tasks
		Response (↑) (n=36)
Linear Probe	VIRCHOW [7] MEAN POOLING	0.607 ± 0.033
	GIGAPATH [6] MEAN POOLING	0.523 ± 0.034
	CHIEF [11] MEAN POOLING	0.667 ± 0.035
	CONCHv1.5 MEAN POOLING	0.623 ± 0.030
	PRISM [13]	0.433 ± 0.034
	GIGAPATH [6]	0.613 ± 0.040
	CHIEF [11]	0.620 ± 0.041
	THREADS	**0.863 ± 0.021**
Supervised	ABMIL	0.553 ± 0.034
	GIGAPATH [6]	0.593 ± 0.043
	CHIEF [11]	0.453 ± 0.042
	THREADS Random Init	0.757 ± 0.027
	THREADS	0.793 ± 0.029

Extended Data Table 26: **Performance comparison between THREADS and baselines on NADT Prostate[42] tasks.** Best in **bold**, second best is <u>underlined</u>. **Response**: This is a patient-level classification task evaluated using AUROC, mean and standard error reported over 50-fold Monte Carlo.

	Model	Tasks
		Response (↑) (n=36)
Linear Probe	VIRCHOW [7] MEAN POOLING	0.612 ± 0.027
	GIGAPATH [6] MEAN POOLING	0.585 ± 0.028
	CHIEF [11] MEAN POOLING	0.643 ± 0.024
	CONCHV1.5 MEAN POOLING	0.620 ± 0.025
	PRISM [13]	<u>0.723 ± 0.022</u>
	GIGAPATH [6]	0.552 ± 0.024
	CHIEF [11]	0.685 ± 0.033
	THREADS	**0.730 ± 0.022**
Supervised	ABMIL	0.577 ± 0.029
	GIGAPATH [6]	0.450 ± 0.036
	CHIEF [11]	0.540 ± 0.031
	THREADS Random Init	0.702 ± 0.022
	THREADS	0.695 ± 0.023

Extended Data Table 27: **Performance comparison between THREADS and baselines on GBM-Treatment tasks.** Best in **bold**, second best is <u>underlined</u>. **Response**: This is a patient-level classification task evaluated using AUROC, mean and standard error reported over 50-fold Monte Carlo.

	Model	Tasks
		Response (↑) (n=93)
Linear Probe	VIRCHOW [7] MEAN POOLING	0.580 ± 0.018
	GIGAPATH [6] MEAN POOLING	0.605 ± 0.017
	CHIEF [11] MEAN POOLING	0.570 ± 0.017
	CONCHV1.5 MEAN POOLING	0.703 ± 0.018
	PRISM [13]	0.622 ± 0.022
	GIGAPATH [6]	0.642 ± 0.014
	CHIEF [11]	0.575 ± 0.019
	THREADS	**0.741 ± 0.016**
Supervised	ABMIL	0.677 ± 0.016
	GIGAPATH [6]	0.680 ± 0.018
	CHIEF [11]	0.550 ± 0.021
	THREADS Random Init	0.655 ± 0.017
	THREADS	<u>0.705 ± 0.015</u>

Extended Data Table 28: **Performance comparison between THREADS and baselines on Post-NAT-BRCA tasks.** Best in **bold**, second best is <u>underlined</u>. **Lymphovascular Invasion**: This is a slide-level classification task evaluated using AUROC, mean and standard error reported over 50-fold Monte Carlo.

	Model	Tasks
		Lymphovascular Invasion (↑) (n=53)
Linear Probe	VIRCHOW [7] MEAN POOLING	0.612 ± 0.028
	GIGAPATH [6] MEAN POOLING	0.603 ± 0.025
	CHIEF [11] MEAN POOLING	0.497 ± 0.023
	CONCHV1.5 MEAN POOLING	0.654 ± 0.022
	PRISM [13]	0.540 ± 0.029
	GIGAPATH [6]	0.582 ± 0.024
	CHIEF [11]	0.542 ± 0.030
	THREADS	**0.701 ± 0.027**
Supervised	ABMIL	0.530 ± 0.029
	GIGAPATH [6]	0.607 ± 0.023
	CHIEF [11]	0.495 ± 0.028
	THREADS Random Init	0.600 ± 0.025
	THREADS	<u>0.662 ± 0.024</u>

Extended Data Table 29: **Performance comparison between THREADS and baselines on SURGEN tasks.** Best in **bold**, second best is <u>underlined</u>. **BRAF Mutation**: This is a patient-level classification task evaluated using AUROC, mean and standard error reported over 5-fold cross-validation. **RAS Mutation**: This is a patient-level classification task evaluated using AUROC, mean and standard error reported over 5-fold cross-validation. **MMR Loss**: This is a patient-level classification task evaluated using AUROC, mean and standard error reported over 5-fold cross-validation. **Death in 5 Years**: This is a patient-level classification task evaluated using AUROC, mean and standard error reported over 5-fold cross-validation.

	Model	Tasks			
		BRAF Mutation (↑) (n=388)	RAS Mutation (↑) (n=389)	MMR Loss (↑) (n=389)	Death in 5 Years (↑) (n=387)
Linear Probe	VIRCHOW [7] MEAN POOLING	0.693 ± 0.037	0.586 ± 0.030	0.880 ± 0.017	0.642 ± 0.021
	GIGAPATH [6] MEAN POOLING	0.694 ± 0.033	0.605 ± 0.015	0.877 ± 0.017	0.663 ± 0.016
	CHIEF [11] MEAN POOLING	0.670 ± 0.033	0.627 ± 0.036	0.840 ± 0.020	0.671 ± 0.012
	CONCHv1.5 MEAN POOLING	0.650 ± 0.025	0.629 ± 0.028	0.796 ± 0.037	0.662 ± 0.023
	PRISM [13]	<u>0.740 ± 0.034</u>	0.632 ± 0.014	0.883 ± 0.010	0.635 ± 0.023
	GIGAPATH [6]	0.680 ± 0.026	0.622 ± 0.022	0.860 ± 0.015	0.632 ± 0.013
	CHIEF [11]	0.736 ± 0.043	<u>0.652 ± 0.018</u>	0.830 ± 0.031	0.683 ± 0.020
	THREADS	0.727 ± 0.042	0.633 ± 0.027	**0.910 ± 0.023**	0.685 ± 0.017
Supervised	ABMIL	0.692 ± 0.039	0.629 ± 0.021	0.893 ± 0.028	<u>0.698 ± 0.024</u>
	GIGAPATH [6]	0.665 ± 0.020	0.634 ± 0.015	0.779 ± 0.027	0.668 ± 0.022
	CHIEF [11]	0.725 ± 0.013	0.540 ± 0.024	0.749 ± 0.050	0.612 ± 0.013
	THREADS Random Init	0.697 ± 0.023	0.629 ± 0.022	<u>0.896 ± 0.016</u>	0.687 ± 0.028
	THREADS	**0.754 ± 0.037**	**0.676 ± 0.026**	0.885 ± 0.032	**0.715 ± 0.024**

Extended Data Table 30: **Performance comparison between THREADS and baselines on MBC tasks.** Best in **bold**, second best is <u>underlined</u>. **Recist**: This is a patient-level classification task evaluated using quadratic weighted kappa, mean and standard error reported over 50-fold Monte Carlo.

	Model	Tasks
		Recist (↑) (n=76)
Linear Probe	VIRCHOW [7] MEAN POOLING	0.051 ± 0.028
	GIGAPATH [6] MEAN POOLING	0.131 ± 0.034
	CHIEF [11] MEAN POOLING	0.204 ± 0.038
	CONCHv1.5 MEAN POOLING	0.253 ± 0.038
	PRISM [13]	0.206 ± 0.034
	GIGAPATH [6]	0.016 ± 0.032
	CHIEF [11]	0.230 ± 0.033
	THREADS	0.258 ± 0.037
Supervised	ABMIL	<u>0.268 ± 0.036</u>
	GIGAPATH [6]	0.111 ± 0.032
	CHIEF [11]	0.015 ± 0.027
	THREADS Random Init	0.209 ± 0.027
	THREADS	**0.326 ± 0.030**

Extended Data Table 31: **Survival prediction on CPTAC-PDAC[38].** Best in **bold**, second best is underlined. α is the cost parameter for CoxNet, set empirically to allow for convergence. **Overall Survival**: This is a patient-level survival task evaluated using C-index, mean and standard error reported over 5-fold cross-validation.

	Model	α	Tasks
			Overall Survival (↑) (n=97)
CoxNet	VIRCHOW [7] MEAN POOLING	0.07	0.502 ± 0.051
	GIGAPATH [6] MEAN POOLING	0.07	0.510 ± 0.043
	CHIEF [11] MEAN POOLING	0.07	0.508 ± 0.042
	CONCHv1.5 MEAN POOLING	0.07	0.569 ± 0.018
	PRISM [13]	0.07	0.554 ± 0.027
	GIGAPATH [6]	0.07	0.426 ± 0.032
	CHIEF [11]	0.07	0.517 ± 0.032
	THREADS	0.07	**0.616 ± 0.031**
Supervised	ABMIL	–	0.611 ± 0.042
	GIGAPATH [6]	–	0.410 ± 0.013
	CHIEF [11]	–	0.489 ± 0.046
	THREADS Random Init	–	0.582 ± 0.037
	THREADS	–	0.577 ± 0.043

Extended Data Table 32: **Survival prediction on CPTAC-LUAD[38].** Best in **bold**, second best is underlined. α is the cost parameter for CoxNet, set empirically to allow for convergence. **Overall Survival**: This is a patient-level survival task evaluated using C-index, mean and standard error reported over 5-fold cross-validation.

	Model	α	Tasks
			Overall Survival (↑) (n=105)
CoxNet	VIRCHOW [7] MEAN POOLING	0.07	0.443 ± 0.072
	GIGAPATH [6] MEAN POOLING	0.07	0.363 ± 0.041
	CHIEF [11] MEAN POOLING	0.07	0.483 ± 0.009
	CONCHv1.5 MEAN POOLING	0.07	0.579 ± 0.060
	PRISM [13]	0.07	**0.614 ± 0.032**
	GIGAPATH [6]	0.07	0.469 ± 0.010
	CHIEF [11]	0.07	0.462 ± 0.030
	THREADS	0.07	0.613 ± 0.054
Supervised	ABMIL	–	0.524 ± 0.069
	GIGAPATH [6]	–	0.462 ± 0.062
	CHIEF [11]	–	0.518 ± 0.050
	THREADS Random Init	–	0.576 ± 0.046
	THREADS	–	0.535 ± 0.074

Extended Data Table 33: **Survival prediction on CPTAC-CCRCC[38].** Best in **bold**, second best is underlined. α is the cost parameter for CoxNet, set empirically to allow for convergence. **Overall Survival**: This is a patient-level survival task evaluated using C-index, mean and standard error reported over 5-fold cross-validation.

	Model	α	Tasks
			Overall Survival (↑) (n=94)
CoxNet	VIRCHOW [7] MEAN POOLING	0.07	0.554 ± 0.093
	GIGAPATH [6] MEAN POOLING	0.07	0.675 ± 0.063
	CHIEF [11] MEAN POOLING	0.07	0.463 ± 0.034
	CONCHv1.5 MEAN POOLING	0.07	0.555 ± 0.091
	PRISM [13]	0.07	0.567 ± 0.048
	GIGAPATH [6]	0.07	0.550 ± 0.048
	CHIEF [11]	0.01	0.626 ± 0.076
	THREADS	0.07	0.673 ± 0.075
Supervised	ABMIL	–	**0.693 ± 0.043**
	GIGAPATH [6]	–	0.527 ± 0.079
	CHIEF [11]	–	0.369 ± 0.039
	THREADS Random Init	–	0.529 ± 0.101
	THREADS	–	0.495 ± 0.090

Extended Data Table 34: **Survival prediction on CPTAC-HNSC[38].** Best in **bold**, second best is underlined. α is the cost parameter for CoxNet, set empirically to allow for convergence. **Overall Survival**: This is a patient-level survival task evaluated using C-index, mean and standard error reported over 5-fold cross-validation.

| | Model | α | Tasks |
			Overall Survival (↑) (n=102)
CoxNet	VIRCHOW [7] MEAN POOLING	0.07	0.569 ± 0.049
	GIGAPATH [6] MEAN POOLING	0.07	0.590 ± 0.056
	CHIEF [11] MEAN POOLING	0.07	0.480 ± 0.021
	CONCHV1.5 MEAN POOLING	0.07	0.568 ± 0.053
	PRISM [13]	0.07	0.587 ± 0.066
	GIGAPATH [6]	0.07	0.562 ± 0.063
	CHIEF [11]	0.07	0.474 ± 0.011
	THREADS	0.07	**0.631 ± 0.076**
Supervised	ABMIL	–	0.495 ± 0.024
	GIGAPATH [6]	–	0.531 ± 0.064
	CHIEF [11]	–	0.489 ± 0.047
	THREADS Random Init	–	0.550 ± 0.014
	THREADS	–	0.514 ± 0.011

Extended Data Table 35: **Survival prediction on SURGEN.** Best in **bold**, second best is underlined. α is the cost parameter for CoxNet, set empirically to allow for convergence. **Overall Survival**: This is a patient-level survival task evaluated using C-index, mean and standard error reported over 5-fold cross-validation.

| | Model | α | Tasks |
			Overall Survival (↑) (n=144)
CoxNet	VIRCHOW [7] MEAN POOLING	0.07	0.593 ± 0.025
	GIGAPATH [6] MEAN POOLING	0.07	0.606 ± 0.027
	CHIEF [11] MEAN POOLING	0.07	0.606 ± 0.023
	CONCHV1.5 MEAN POOLING	0.07	0.625 ± 0.022
	PRISM [13]	0.07	0.578 ± 0.026
	GIGAPATH [6]	0.07	0.611 ± 0.025
	CHIEF [11]	0.07	0.600 ± 0.026
	THREADS	0.07	**0.638 ± 0.014**
Supervised	ABMIL	–	0.587 ± 0.026
	GIGAPATH [6]	–	0.612 ± 0.022
	CHIEF [11]	–	0.531 ± 0.037
	THREADS Random Init	–	0.632 ± 0.022
	THREADS	–	0.613 ± 0.034

Extended Data Table 36: **Survival prediction on MBC.** Best in **bold**, second best is underlined. α is the cost parameter for CoxNet, set empirically to allow for convergence. **Overall Survival**: This is a patient-level survival task evaluated using C-index, mean and standard error reported over 5-fold cross-validation.

| | Model | α | Tasks |
			Overall Survival (↑) (n=75)
CoxNet	VIRCHOW [7] MEAN POOLING	0.07	0.517 ± 0.031
	GIGAPATH [6] MEAN POOLING	0.07	0.472 ± 0.024
	CHIEF [11] MEAN POOLING	0.07	0.441 ± 0.040
	CONCHV1.5 MEAN POOLING	0.07	0.510 ± 0.052
	PRISM [13]	0.07	0.511 ± 0.038
	GIGAPATH [6]	0.07	0.440 ± 0.030
	CHIEF [11]	0.07	0.460 ± 0.046
	THREADS	0.07	0.550 ± 0.027
Supervised	ABMIL	–	0.519 ± 0.043
	GIGAPATH [6]	–	0.433 ± 0.014
	CHIEF [11]	–	0.529 ± 0.064
	THREADS Random Init	–	0.512 ± 0.034
	THREADS	–	**0.608 ± 0.030**

Extended Data Table 37: **Survival prediction on BOEHMK.** Best in **bold**, second best is <u>underlined</u>. α is the cost parameter for CoxNet, set empirically to allow for convergence. GIGAPATH-supervised could not be evaluated due to a bug in PyTorch autocast (https://github.com/pytorch/pytorch/issues/81876). **Progression Free Survival**: This is a patient-level survival task evaluated using C-index, mean and standard error reported over 5-fold cross-validation.

	Model	α	Tasks
			Progression Free Survival ($\uparrow$) (n=183)
CoxNet	VIRCHOW [7] MEAN POOLING	0.01	0.513 ± 0.017
	GIGAPATH [6] MEAN POOLING	0.01	0.487 ± 0.016
	CHIEF [11] MEAN POOLING	0.01	0.480 ± 0.031
	CONCHv1.5 MEAN POOLING	0.01	0.536 ± 0.019
	PRISM [13]	0.02	0.500 ± 0.019
	GIGAPATH [6]	0.01	0.536 ± 0.043
	CHIEF [11]	0.01	0.520 ± 0.031
	THREADS	0.01	0.541 ± 0.013
Supervised	ABMIL	–	0.523 ± 0.036
	GIGAPATH [6]	–	—
	CHIEF [11]	–	0.517 ± 0.033
	THREADS Random Init	–	<u>0.553 ± 0.058</u>
	THREADS	–	**0.575 ± 0.049**

Extended Data Table 38: **Effect of increasing linear probe cost on our benchmark.** Lower cost corresponds to stronger regularization. Adaptive cost[64] is computed as $\frac{embedding\_dim \times num\_classes}{100}$.

Model	Cost	Benchmark (without survival)
VIRCHOW [7] MEAN POOLING		0.653
GIGAPATH [6] MEAN POOLING		0.663
CHIEF [11] MEAN POOLING		0.592
CONCHV1.5 MEAN POOLING		0.690
PRISM [13]	0.001	0.719
GIGAPATH [6]		0.651
CHIEF [11]		0.648
THREADS		**0.758**
VIRCHOW [7] MEAN POOLING		0.682
GIGAPATH [6] MEAN POOLING		0.691
CHIEF [11] MEAN POOLING		0.616
CONCHV1.5 MEAN POOLING		0.716
PRISM [13]	0.01	0.729
GIGAPATH [6]		0.679
CHIEF [11]		0.668
THREADS		**0.772**
VIRCHOW [7] MEAN POOLING		0.690
GIGAPATH [6] MEAN POOLING		0.691
CHIEF [11] MEAN POOLING		0.652
CONCHV1.5 MEAN POOLING		0.717
PRISM [13]	0.1	0.720
GIGAPATH [6]		0.685
CHIEF [11]		0.698
THREADS		**0.779**
VIRCHOW [7] MEAN POOLING		0.681
GIGAPATH [6] MEAN POOLING		0.683
CHIEF [11] MEAN POOLING		0.670
CONCHV1.5 MEAN POOLING		0.708
PRISM [13]	0.5	0.709
GIGAPATH [6]		0.675
CHIEF [11]		0.710
THREADS		**0.774**
VIRCHOW [7] MEAN POOLING		0.676
GIGAPATH [6] MEAN POOLING		0.679
CHIEF [11] MEAN POOLING		0.671
CONCHV1.5 MEAN POOLING		0.703
PRISM [13]	1.0	0.704
GIGAPATH [6]		0.669
CHIEF [11]		0.711
THREADS		**0.769**
VIRCHOW [7] MEAN POOLING		0.665
GIGAPATH [6] MEAN POOLING		0.670
CHIEF [11] MEAN POOLING		0.663
CONCHV1.5 MEAN POOLING		0.691
PRISM [13]	10.0	0.694
GIGAPATH [6]		0.659
CHIEF [11]		0.694
THREADS		**0.751**
VIRCHOW [7] MEAN POOLING		0.662
GIGAPATH [6] MEAN POOLING		0.668
CHIEF [11] MEAN POOLING		0.659
CONCHV1.5 MEAN POOLING		0.688
PRISM [13]	Adaptive	0.691
GIGAPATH [6]		0.657
CHIEF [11]		0.690
THREADS		**0.746**

Extended Data Table 39: **Generalizability experiments using linear probe (for classification) and CoxNet (for survival).** All samples from the train dataset are used for training, and all samples from the test dataset are used for testing. All tasks are binary classification. CI: 95% confidence interval over 100 bootstraps.

Train Dataset	Test Dataset	Task	Method	Mean AUC (CI)
CPTAC-CCRCC[38]	MUT-HET-RCC	BAP1 Mutation	PRISM	0.776 (0.745-0.809)
			GIGAPATH	0.793 (0.765-0.824)
			CHIEF	0.813 (0.788-0.840)
			THREADS	**0.840 (0.809-0.866)**
		PBRM1 Mutation	PRISM	0.665 (0.636-0.695)
			GIGAPATH	0.577 (0.548-0.603)
			CHIEF	0.681 (0.654-0.710)
			THREADS	**0.701 (0.676-0.729)**
TCGA-BRCA	BCNB[29]	ER	PRISM	0.627 (0.590-0.663)
			GIGAPATH	0.731 (0.698-0.770)
			CHIEF	0.835 (0.804-0.864)
			THREADS	**0.885 (0.859-0.906)**
		PR	PRISM	0.677 (0.638-0.708)
			GIGAPATH	0.688 (0.654-0.716)
			CHIEF	0.787 (0.760-0.813)
			THREADS	**0.794 (0.767-0.822)**
TCGA-GBMLGG	EBRAINS[27]	IDH Status	PRISM	0.931 (0.909-0.947)
			GIGAPATH	0.904 (0.882-0.926)
			CHIEF	0.935 (0.913-0.954)
			THREADS	**0.961 (0.947-0.975)**
TCGA-NSCLC	MGB Lung	Subtype	PRISM	0.969 (0.957-0.981)
			GIGAPATH	0.890 (0.871-0.905)
			CHIEF	0.951 (0.937-0.962)
			THREADS	**0.984 (0.978-0.989)**
TCGA-BRCA	MGB Breast	Subtype	PRISM	0.963 (0.950-0.973)
			GIGAPATH	0.919 (0.905-0.934)
			CHIEF	**0.970 (0.962-0.978)**
			THREADS	0.965 (0.956-0.973)
TCGA-LUAD	CPTAC-LUAD[38]	Overall Survival	PRISM	0.545 (0.407-0.681)
			GIGAPATH	0.561 (0.413-0.685)
			CHIEF	0.495 (0.369-0.607)
			THREADS	**0.654 (0.511-0.748)**
TCGA-PDAC	CPTAC-PDAC[38]	Overall Survival	PRISM	0.505 (0.432-0.587)
			GIGAPATH	0.571 (0.493-0.642)
			CHIEF	0.505 (0.419-0.592)
			THREADS	**0.613 (0.546-0.696)**

Extended Data Table 40: **Performance comparison between THREADS and baselines on BCNB[29] tasks in a few-shot setting.** Best in **bold**, second best is underlined. **ER**: This is a patient-level classification task evaluated using AUROC, mean and standard error reported over 5-fold cross-validation. **PR**: This is a patient-level classification task evaluated using AUROC, mean and standard error reported over 5-fold cross-validation. **HER2**: This is a patient-level classification task evaluated using AUROC, mean and standard error reported over 5-fold cross-validation.

Model	# Shots	Tasks		
		ER (n=1058)	PR (n=1058)	HER2 (n=1058)
VIRCHOW[7] MEAN POOLING	1	0.520 ± 0.025	0.555 ± 0.012	0.540 ± 0.011
GIGAPATH[6] MEAN POOLING	1	0.505 ± 0.014	0.576 ± 0.018	0.542 ± 0.012
CHIEF[11] MEAN POOLING	1	0.485 ± 0.016	0.569 ± 0.010	0.493 ± 0.017
CONCHv1.5 MEAN POOLING	1	0.520 ± 0.082	<u>0.627 ± 0.035</u>	<u>0.568 ± 0.031</u>
PRISM[13]	1	<u>0.574 ± 0.055</u>	0.560 ± 0.051	**0.624 ± 0.023**
GIGAPATH[6]	1	0.514 ± 0.024	0.574 ± 0.021	0.523 ± 0.010
CHIEF[11]	1	0.566 ± 0.031	0.603 ± 0.018	0.488 ± 0.020
THREADS	1	**0.671 ± 0.054**	**0.678 ± 0.091**	0.543 ± 0.062
VIRCHOW[7] MEAN POOLING	2	0.512 ± 0.035	0.546 ± 0.023	0.541 ± 0.011
GIGAPATH[6] MEAN POOLING	2	0.583 ± 0.046	0.594 ± 0.031	0.531 ± 0.019
CHIEF[11] MEAN POOLING	2	0.560 ± 0.038	0.557 ± 0.026	0.487 ± 0.046
CONCHv1.5 MEAN POOLING	2	0.602 ± 0.073	<u>0.675 ± 0.020</u>	<u>0.574 ± 0.020</u>
PRISM[13]	2	<u>0.687 ± 0.043</u>	0.627 ± 0.050	**0.580 ± 0.039**
GIGAPATH[6]	2	0.573 ± 0.042	0.572 ± 0.032	0.525 ± 0.027
CHIEF[11]	2	0.657 ± 0.032	0.614 ± 0.016	0.497 ± 0.030
THREADS	2	**0.783 ± 0.055**	**0.770 ± 0.030**	**0.580 ± 0.037**
VIRCHOW[7] MEAN POOLING	4	0.568 ± 0.044	0.587 ± 0.035	0.523 ± 0.012
GIGAPATH[6] MEAN POOLING	4	0.668 ± 0.051	0.629 ± 0.039	0.545 ± 0.031
CHIEF[11] MEAN POOLING	4	0.585 ± 0.036	0.599 ± 0.039	0.524 ± 0.023
CONCHv1.5 MEAN POOLING	4	0.696 ± 0.068	<u>0.713 ± 0.040</u>	0.565 ± 0.034
PRISM[13]	4	<u>0.734 ± 0.039</u>	0.657 ± 0.040	<u>0.643 ± 0.019</u>
GIGAPATH[6]	4	0.638 ± 0.050	0.611 ± 0.042	0.526 ± 0.019
CHIEF[11]	4	0.724 ± 0.036	0.666 ± 0.028	0.578 ± 0.017
THREADS	4	**0.838 ± 0.021**	**0.725 ± 0.061**	**0.651 ± 0.024**
VIRCHOW[7] MEAN POOLING	8	0.601 ± 0.036	0.592 ± 0.024	0.555 ± 0.022
GIGAPATH[6] MEAN POOLING	8	0.686 ± 0.046	0.651 ± 0.023	0.594 ± 0.032
CHIEF[11] MEAN POOLING	8	0.623 ± 0.034	0.618 ± 0.023	0.540 ± 0.023
CONCHv1.5 MEAN POOLING	8	0.716 ± 0.060	<u>0.721 ± 0.037</u>	0.608 ± 0.033
PRISM[13]	8	<u>0.764 ± 0.027</u>	0.695 ± 0.027	<u>0.655 ± 0.017</u>
GIGAPATH[6]	8	0.663 ± 0.042	0.619 ± 0.028	0.567 ± 0.027
CHIEF[11]	8	0.756 ± 0.027	0.687 ± 0.028	0.606 ± 0.026
THREADS	8	**0.836 ± 0.028**	**0.756 ± 0.034**	**0.666 ± 0.012**
VIRCHOW[7] MEAN POOLING	16	0.683 ± 0.030	0.641 ± 0.019	0.591 ± 0.024
GIGAPATH[6] MEAN POOLING	16	0.762 ± 0.021	0.698 ± 0.015	0.605 ± 0.029
CHIEF[11] MEAN POOLING	16	0.678 ± 0.021	0.635 ± 0.019	0.561 ± 0.026
CONCHv1.5 MEAN POOLING	16	0.786 ± 0.032	0.723 ± 0.025	0.626 ± 0.030
PRISM[13]	16	<u>0.828 ± 0.017</u>	<u>0.744 ± 0.019</u>	<u>0.656 ± 0.030</u>
GIGAPATH[6]	16	0.731 ± 0.034	0.643 ± 0.014	0.599 ± 0.026
CHIEF[11]	16	0.801 ± 0.015	0.713 ± 0.020	0.645 ± 0.026
THREADS	16	**0.874 ± 0.018**	**0.786 ± 0.019**	**0.684 ± 0.023**
VIRCHOW[7] MEAN POOLING	32	0.757 ± 0.031	0.688 ± 0.023	0.614 ± 0.020
GIGAPATH[6] MEAN POOLING	32	0.787 ± 0.021	0.728 ± 0.025	0.612 ± 0.024
CHIEF[11] MEAN POOLING	32	0.720 ± 0.023	0.667 ± 0.024	0.598 ± 0.029
CONCHv1.5 MEAN POOLING	32	0.810 ± 0.024	0.744 ± 0.021	0.632 ± 0.019
PRISM[13]	32	<u>0.822 ± 0.023</u>	<u>0.752 ± 0.025</u>	<u>0.654 ± 0.017</u>
GIGAPATH[6]	32	0.769 ± 0.022	0.692 ± 0.020	0.619 ± 0.023
CHIEF[11]	32	0.809 ± 0.016	0.734 ± 0.022	0.646 ± 0.026
THREADS	32	**0.882 ± 0.014**	**0.798 ± 0.026**	**0.694 ± 0.020**

Extended Data Table 41: **Performance comparison between THREADS and baselines on BRACS[28] tasks in a few-shot setting.** Best in **bold**, second best is <u>underlined</u>. **Fine Subtype**: This is a slide-level classification task evaluated using balanced accuracy, mean and standard error reported over 5-fold cross-validation. **Coarse Subtype**: This is a slide-level classification task evaluated using balanced accuracy, mean and standard error reported over 5-fold cross-validation.

Model	# Shots	Tasks	
		Fine Subtype (n=547)	Coarse Subtype (n=547)
VIRCHOW [7] MEAN POOLING	1	0.229 ± 0.024	0.335 ± 0.020
GIGAPATH [6] MEAN POOLING	1	0.201 ± 0.014	0.328 ± 0.021
CHIEF [11] MEAN POOLING	1	0.216 ± 0.021	0.333 ± 0.024
CONCHv1.5 MEAN POOLING	1	0.241 ± 0.024	0.308 ± 0.024
PRISM [13]	1	**0.334 ± 0.057**	**0.462 ± 0.055**
GIGAPATH [6]	1	0.198 ± 0.021	0.309 ± 0.017
CHIEF [11]	1	<u>0.291 ± 0.040</u>	0.373 ± 0.019
THREADS	1	<u>0.291 ± 0.032</u>	<u>0.402 ± 0.050</u>
VIRCHOW [7] MEAN POOLING	2	0.204 ± 0.023	0.369 ± 0.017
GIGAPATH [6] MEAN POOLING	2	0.227 ± 0.022	0.369 ± 0.015
CHIEF [11] MEAN POOLING	2	0.254 ± 0.023	0.350 ± 0.017
CONCHv1.5 MEAN POOLING	2	0.231 ± 0.015	0.398 ± 0.022
PRISM [13]	2	**0.361 ± 0.041**	**0.549 ± 0.031**
GIGAPATH [6]	2	0.196 ± 0.016	0.369 ± 0.014
CHIEF [11]	2	0.319 ± 0.029	0.482 ± 0.045
THREADS	2	<u>0.346 ± 0.033</u>	<u>0.521 ± 0.031</u>
VIRCHOW [7] MEAN POOLING	4	0.258 ± 0.023	0.430 ± 0.021
GIGAPATH [6] MEAN POOLING	4	0.234 ± 0.018	0.413 ± 0.023
CHIEF [11] MEAN POOLING	4	0.255 ± 0.027	0.410 ± 0.022
CONCHv1.5 MEAN POOLING	4	0.259 ± 0.026	0.492 ± 0.018
PRISM [13]	4	<u>0.363 ± 0.019</u>	<u>0.551 ± 0.027</u>
GIGAPATH [6]	4	0.237 ± 0.015	0.402 ± 0.028
CHIEF [11]	4	0.351 ± 0.041	0.547 ± 0.041
THREADS	4	**0.387 ± 0.022**	**0.588 ± 0.025**
VIRCHOW [7] MEAN POOLING	8	0.315 ± 0.033	0.459 ± 0.015
GIGAPATH [6] MEAN POOLING	8	0.285 ± 0.020	0.473 ± 0.024
CHIEF [11] MEAN POOLING	8	0.276 ± 0.021	0.450 ± 0.022
CONCHv1.5 MEAN POOLING	8	0.290 ± 0.022	0.523 ± 0.051
PRISM [13]	8	**0.416 ± 0.026**	0.551 ± 0.033
GIGAPATH [6]	8	0.281 ± 0.015	0.461 ± 0.028
CHIEF [11]	8	<u>0.411 ± 0.027</u>	<u>0.567 ± 0.031</u>
THREADS	8	0.394 ± 0.014	**0.602 ± 0.038**
VIRCHOW [7] MEAN POOLING	16	0.368 ± 0.036	0.480 ± 0.017
GIGAPATH [6] MEAN POOLING	16	0.314 ± 0.033	0.494 ± 0.038
CHIEF [11] MEAN POOLING	16	0.343 ± 0.033	0.526 ± 0.011
CONCHv1.5 MEAN POOLING	16	0.318 ± 0.041	0.539 ± 0.038
PRISM [13]	16	0.418 ± 0.014	<u>0.635 ± 0.015</u>
GIGAPATH [6]	16	0.314 ± 0.033	0.512 ± 0.036
CHIEF [11]	16	**0.434 ± 0.024**	0.618 ± 0.017
THREADS	16	<u>0.425 ± 0.015</u>	**0.701 ± 0.027**
VIRCHOW [7] MEAN POOLING	32	0.349 ± 0.020	0.527 ± 0.009
GIGAPATH [6] MEAN POOLING	32	0.353 ± 0.025	0.505 ± 0.032
CHIEF [11] MEAN POOLING	32	0.335 ± 0.041	0.544 ± 0.019
CONCHv1.5 MEAN POOLING	32	0.361 ± 0.042	0.607 ± 0.045
PRISM [13]	32	<u>0.435 ± 0.030</u>	0.632 ± 0.023
GIGAPATH [6]	32	0.362 ± 0.031	0.535 ± 0.037
CHIEF [11]	32	0.417 ± 0.021	<u>0.649 ± 0.023</u>
THREADS	32	**0.502 ± 0.028**	**0.726 ± 0.021**

Extended Data Table 42: **Performance comparison between THREADS and baselines on EBRAINS[27] tasks in a few-shot setting.** Best in **bold**, second best is <u>underlined</u>. **Fine Subtype**: This is a slide-level classification task evaluated using balanced accuracy, mean and 95% CI reported over 5 bootstraps of the official train split. **Coarse Subtype**: This is a slide-level classification task evaluated using balanced accuracy, mean and 95% CI reported over 5 bootstraps of the official train split.

| Model | # Shots | Tasks | |
		Fine Subtype (n=2319)	Coarse Subtype (n=2319)
VIRCHOW [7] MEAN POOLING	1	0.303 ± 0.017	0.368 ± 0.013
GIGAPATH [6] MEAN POOLING	1	0.304 ± 0.016	0.347 ± 0.015
CHIEF [11] MEAN POOLING	1	0.215 ± 0.006	0.273 ± 0.008
CONCHv1.5 MEAN POOLING	1	0.361 ± 0.012	0.414 ± 0.021
PRISM [13]	1	<u>0.406 ± 0.017</u>	<u>0.463 ± 0.018</u>
GIGAPATH [6]	1	0.293 ± 0.012	0.331 ± 0.015
CHIEF [11]	1	0.234 ± 0.016	0.262 ± 0.020
THREADS	1	**0.503 ± 0.008**	**0.595 ± 0.028**
VIRCHOW [7] MEAN POOLING	2	0.380 ± 0.010	0.472 ± 0.016
GIGAPATH [6] MEAN POOLING	2	0.413 ± 0.005	0.490 ± 0.020
CHIEF [11] MEAN POOLING	2	0.282 ± 0.007	0.373 ± 0.015
CONCHv1.5 MEAN POOLING	2	0.449 ± 0.015	0.537 ± 0.008
PRISM [13]	2	<u>0.505 ± 0.008</u>	<u>0.614 ± 0.018</u>
GIGAPATH [6]	2	0.395 ± 0.006	0.460 ± 0.013
CHIEF [11]	2	0.311 ± 0.020	0.369 ± 0.030
THREADS	2	**0.570 ± 0.004**	**0.736 ± 0.022**
VIRCHOW [7] MEAN POOLING	4	0.489 ± 0.006	0.612 ± 0.009
GIGAPATH [6] MEAN POOLING	4	0.522 ± 0.009	0.625 ± 0.009
CHIEF [11] MEAN POOLING	4	0.371 ± 0.009	0.450 ± 0.014
CONCHv1.5 MEAN POOLING	4	0.525 ± 0.009	0.631 ± 0.020
PRISM [13]	4	<u>0.565 ± 0.009</u>	<u>0.693 ± 0.010</u>
GIGAPATH [6]	4	0.490 ± 0.008	0.590 ± 0.010
CHIEF [11]	4	0.402 ± 0.012	0.456 ± 0.012
THREADS	4	**0.635 ± 0.007**	**0.802 ± 0.012**
VIRCHOW [7] MEAN POOLING	8	0.593 ± 0.013	0.703 ± 0.010
GIGAPATH [6] MEAN POOLING	8	0.598 ± 0.008	0.719 ± 0.010
CHIEF [11] MEAN POOLING	8	0.457 ± 0.006	0.544 ± 0.014
CONCHv1.5 MEAN POOLING	8	0.592 ± 0.004	0.748 ± 0.010
PRISM [13]	8	<u>0.603 ± 0.006</u>	<u>0.782 ± 0.004</u>
GIGAPATH [6]	8	0.592 ± 0.007	0.689 ± 0.012
CHIEF [11]	8	0.494 ± 0.007	0.590 ± 0.012
THREADS	8	**0.683 ± 0.008**	**0.859 ± 0.004**
VIRCHOW [7] MEAN POOLING	16	0.644 ± 0.005	0.793 ± 0.005
GIGAPATH [6] MEAN POOLING	16	<u>0.674 ± 0.007</u>	0.800 ± 0.006
CHIEF [11] MEAN POOLING	16	0.541 ± 0.007	0.618 ± 0.007
CONCHv1.5 MEAN POOLING	16	0.641 ± 0.006	0.801 ± 0.004
PRISM [13]	16	0.643 ± 0.008	<u>0.809 ± 0.005</u>
GIGAPATH [6]	16	0.661 ± 0.007	0.777 ± 0.007
CHIEF [11]	16	0.566 ± 0.006	0.692 ± 0.010
THREADS	16	**0.712 ± 0.004**	**0.882 ± 0.002**
VIRCHOW [7] MEAN POOLING	32	—	0.841 ± 0.005
GIGAPATH [6] MEAN POOLING	32	—	<u>0.866 ± 0.005</u>
CHIEF [11] MEAN POOLING	32	—	0.700 ± 0.006
CONCHv1.5 MEAN POOLING	32	—	0.832 ± 0.004
PRISM [13]	32	—	0.826 ± 0.005
GIGAPATH [6]	32	—	0.847 ± 0.007
CHIEF [11]	32	—	0.751 ± 0.003
THREADS	32	—	**0.905 ± 0.002**

Extended Data Table 43: **Performance comparison between THREADS and baselines on GBM-Treatment tasks in a few-shot setting.** Best in **bold**, second best is <u>underlined</u>. **Response**: This is a patient-level classification task evaluated using AUROC, mean and standard error reported over 50-fold Monte Carlo.

Model	# Shots	Tasks
		Response (n=93)
VIRCHOW [7] MEAN POOLING	1	**0.583 ± 0.021**
GIGAPATH [6] MEAN POOLING	1	0.519 ± 0.020
CHIEF [11] MEAN POOLING	1	0.526 ± 0.022
CONCHv1.5 MEAN POOLING	1	<u>0.578 ± 0.021</u>
PRISM [13]	1	0.573 ± 0.023
GIGAPATH [6]	1	0.550 ± 0.022
CHIEF [11]	1	0.525 ± 0.019
THREADS	1	0.551 ± 0.021
VIRCHOW [7] MEAN POOLING	2	<u>0.555 ± 0.026</u>
GIGAPATH [6] MEAN POOLING	2	0.530 ± 0.023
CHIEF [11] MEAN POOLING	2	0.542 ± 0.022
CONCHv1.5 MEAN POOLING	2	0.554 ± 0.021
PRISM [13]	2	**0.558 ± 0.022**
GIGAPATH [6]	2	0.551 ± 0.025
CHIEF [11]	2	0.509 ± 0.019
THREADS	2	0.548 ± 0.020
VIRCHOW [7] MEAN POOLING	4	0.556 ± 0.020
GIGAPATH [6] MEAN POOLING	4	0.539 ± 0.022
CHIEF [11] MEAN POOLING	4	0.528 ± 0.019
CONCHv1.5 MEAN POOLING	4	**0.606 ± 0.019**
PRISM [13]	4	0.554 ± 0.023
GIGAPATH [6]	4	0.553 ± 0.025
CHIEF [11]	4	0.489 ± 0.023
THREADS	4	<u>0.566 ± 0.020</u>
VIRCHOW [7] MEAN POOLING	8	0.591 ± 0.020
GIGAPATH [6] MEAN POOLING	8	0.581 ± 0.023
CHIEF [11] MEAN POOLING	8	0.534 ± 0.022
CONCHv1.5 MEAN POOLING	8	**0.634 ± 0.019**
PRISM [13]	8	0.574 ± 0.026
GIGAPATH [6]	8	0.613 ± 0.024
CHIEF [11]	8	0.511 ± 0.023
THREADS	8	<u>0.618 ± 0.022</u>
VIRCHOW [7] MEAN POOLING	16	0.572 ± 0.017
GIGAPATH [6] MEAN POOLING	16	0.572 ± 0.019
CHIEF [11] MEAN POOLING	16	0.527 ± 0.019
CONCHv1.5 MEAN POOLING	16	<u>0.638 ± 0.020</u>
PRISM [13]	16	0.630 ± 0.021
GIGAPATH [6]	16	0.608 ± 0.017
CHIEF [11]	16	0.503 ± 0.021
THREADS	16	**0.676 ± 0.017**

Extended Data Table 44: **Retrieval performance on EBRAINS fine (30 classes) and coarse (12 classes) subtyping.** Best in **bold**, second best is underlined. **mAP@K**: mean average precision using top-k retrieved examples. CI: 95% confidence interval.

Model	Subtyping Task	mAP@1 (CI)	mAP@5 (CI)	mAP@10 (CI)
VIRCHOW [7] MEAN POOLING	Fine	0.593 (0.555-0.627)	0.453 (0.425-0.480)	0.393 (0.366-0.417)
GIGAPATH [6] MEAN POOLING	Fine	0.581 (0.536-0.626)	0.456 (0.427-0.490)	0.395 (0.367-0.425)
CHIEF [11] MEAN POOLING	Fine	0.463 (0.425-0.504)	0.317 (0.288-0.340)	0.260 (0.235-0.279)
CONCHv1.5 MEAN POOLING	Fine	0.645 (0.610-0.675)	0.514 (0.488-0.540)	0.463 (0.438-0.487)
PRISM [13]	Fine	0.648 (0.615-0.686)	0.534 (0.506-0.560)	0.488 (0.463-0.514)
GIGAPATH [6]	Fine	0.607 (0.572-0.640)	0.454 (0.426-0.482)	0.383 (0.355-0.411)
CHIEF [11]	Fine	0.502 (0.469-0.541)	0.359 (0.333-0.387)	0.297 (0.272-0.320)
THREADS	Fine	**0.706 (0.667-0.739)**	**0.606 (0.573-0.633)**	**0.568 (0.538-0.597)**
VIRCHOW [7] MEAN POOLING	Coarse	0.797 (0.759-0.830)	0.702 (0.670-0.726)	0.648 (0.618-0.670)
GIGAPATH [6] MEAN POOLING	Coarse	0.782 (0.741-0.820)	0.698 (0.665-0.724)	0.642 (0.611-0.670)
CHIEF [11] MEAN POOLING	Coarse	0.659 (0.619-0.696)	0.534 (0.502-0.563)	0.483 (0.455-0.511)
CONCHv1.5 MEAN POOLING	Coarse	0.857 (0.829-0.880)	0.774 (0.746-0.796)	0.731 (0.704-0.753)
PRISM [13]	Coarse	0.845 (0.818-0.874)	0.772 (0.746-0.793)	0.738 (0.713-0.760)
GIGAPATH [6]	Coarse	0.798 (0.763-0.828)	0.696 (0.667-0.720)	0.635 (0.608-0.659)
CHIEF [11]	Coarse	0.697 (0.661-0.740)	0.582 (0.551-0.613)	0.526 (0.497-0.555)
THREADS	Coarse	**0.904 (0.883-0.928)**	**0.854 (0.834-0.873)**	**0.831 (0.809-0.849)**

Extended Data Table 45: **Tissue type retrieval (10 classes) performance using CPTAC.** Best in **bold**, second best is underlined. **mAP@K**: mean average precision using top-k retrieved examples. CI: 95% confidence interval.

Model	mAP@1 (CI)	mAP@5 (CI)	mAP@10 (CI)
VIRCHOW MEAN POOLING	0.935 ± 0.004	0.818 ± 0.003	0.754 ± 0.004
GIGAPATH MEAN POOLING	0.943 ± 0.005	0.831 ± 0.007	0.753 ± 0.005
CHIEF MEAN POOLING	0.910 ± 0.003	0.778 ± 0.002	0.698 ± 0.003
CONCHv1.5 MEAN POOLING	0.955 ± 0.002	0.890 ± 0.004	0.848 ± 0.004
PRISM	0.944 ± 0.003	0.883 ± 0.004	0.850 ± 0.005
GIGAPATH	0.943 ± 0.006	0.830 ± 0.006	0.749 ± 0.005
CHIEF	0.916 ± 0.002	0.801 ± 0.006	0.734 ± 0.007
THREADS	**0.967 ± 0.003**	**0.900 ± 0.004**	**0.861 ± 0.006**

Extended Data Table 46: **Prompting experiments.** All samples from the train dataset are used for training, and all samples from the test dataset are used for testing. Best in **bold**, second best is underlined. CI: 95% confidence interval.

Train Dataset	Test Dataset	Task	Method	Mean AUC (CI)
TCGA-BRCA	MGB Breast	ER	CONCHv1.5 MEAN POOLING	0.702 (0.673-0.744)
			THREADS Molecular	**0.779 (0.747-0.810)**
		HER2	CONCHv1.5 MEAN POOLING	0.580 (0.522-0.623)
			THREADS Molecular	**0.713 (0.668-0.764)**
		PR	CONCHv1.5 MEAN POOLING	0.662 (0.620-0.694)
			THREADS Molecular	**0.736 (0.698-0.771)**
		Subtype	CONCHv1.5 MEAN POOLING	**0.916 (0.901-0.934)**
			THREADS Molecular	0.882 (0.858-0.905)
TCGA-GBMLGG	EBRAINS[27]	IDH Status	CONCHv1.5 MEAN POOLING	0.910 (0.892-0.927)
			THREADS Molecular	**0.960 (0.947-0.972)**
TCGA-NSCLC	MGB Lung	Subtype	CONCHv1.5 MEAN POOLING	0.889 (0.870-0.905)
			THREADS Molecular	**0.984 (0.974-0.990)**
TCGA-CCRCC	CPTAC-CCRCC[38]	Overall Survival	CONCHv1.5 MEAN POOLING	0.581 (0.497-0.653)
			THREADS Molecular	**0.687 (0.591-0.762)**
TCGA-PDAC	CPTAC-PDAC[38]	Overall Survival	CONCHv1.5 MEAN POOLING	0.531 (0.481-0.584)
			THREADS Molecular	**0.589 (0.543-0.635)**

Extended Data Table 47: **Impact of pretraining size on performance.** We train THREADS on pretraining datasets of increasing size and report the average linear probe performance of each family of tasks (Table 2, 3, 4, 5, 6, 7). Full benchmark refers to our proposed pan-tissue benchmark.

% of pretraining data (number of WSIs)	Clinical subtyping and grading	IHC prediction	Mutation prediction	Treatment response and survival prediction	Full benchmark performance
CONCHv1.5 MEAN POOLING	0.779	0.770	0.680	0.574	0.689
1 (470)	0.806	0.791	0.716	0.583	0.714
5 (2356)	0.829	0.804	0.727	0.585	0.725
25 (11791)	**0.836**	0.802	0.727	0.613	0.732
50 (23584)	**0.836**	0.814	0.737	0.614	0.739
75 (35377)	0.828	0.815	0.752	0.629	0.747
100 (47171)	0.832	**0.825**	**0.755**	**0.635**	**0.753**

Extended Data Table 48: **Impact of model size on performance.** Number of pretraining heads in THREADS (1, 2, 4 heads) compared against PRISM and GIGAPATH. Evaluations done using linear probing on Full benchmark. RESNET50-IN MEAN POOLING is ResNet50 model pretrained on ImageNet (IN)

Model name	Number of parameters in WSI encoder (million)	Full benchmark performance
CONCHv1.5 MEAN POOLING	N/A	0.689
VIRCHOW [7] MEAN POOLING	N/A	0.662
GIGAPATH [6] MEAN POOLING	N/A	0.663
CHIEF [11] (Patch) MEAN POOLING	N/A	0.651
RESNET50-IN [65] MEAN POOLING	N/A	0.560
THREADS (1 head)	5.0	0.734
THREADS (2 heads)	11.3	**0.753**
THREADS (4 heads)	19.7	<u>0.743</u>
THREADS (6 heads)	28.1	<u>0.743</u>
THREADS (ViT)	16.1	0.690
PRISM [13]	45.0	0.690
GIGAPATH [6]	85.1	0.654
CHIEF [11] (Slide)	1.2	0.686

Extended Data Table 49: **CONCHv1.5 hyperparameters**. CONCHv1.5 was initialized with UNI (Vision Transformer Large). *Batch size* refers to the total batch size across GPUs. Effective batch size used for optimization is *batch size × gradient accumulation steps*. Learning rate is increased from zero linearly to the *peak learning rate* over the course of *warmup steps* and decays back to zero following the *learning rate scheduler*. The model was trained for 20 epochs with 1.26 million image / caption pairs where the maximum sequence length for captions is set to 128. Non-squared images are first padded to square and then resized to 448×448.

Hyperparameter	Values
Image size	448×448
Automatic mixed precision	FP16
Batch size	256
Gradient accumulation steps	3
Learning rate scheduler	Cosine
Warmup steps	250
Peak learning rate	1e-4
AdamW β	(0.9, 0.999)
AdamW ϵ	1e-8
Weight decay	0.2
Softmax temperature	Learned
Epochs	20

Extended Data Table 50: **CLIP hyperparameters in THREADS pretraining.**

Hyperparameter	Value
GPU	$4\times$ 80GB A100
Batch size per GPU	300
Patches sampled during training	512
AdamW β	(0.9, 0.999)
WSI embedder learning rate	0.00005
RNA embedder learning rate	0.00005
DNA embedder learning rate	0.00005
WSI embedder weight decay	0.0001
RNA embedder weight decay	0.0001
DNA embedder weight decay	0.0001
Learning rate schedule	Cosine
Learning rate (start)	0
Learning rate (post warmup)	1e-5
Learning rate (final)	1e-8
Warmup epochs	5
Max epochs	101
INFONCE Temperature	0.07
Automatic mixed precision	bfloaft16
Distributed Data Parallel Backend	GLOO
Early stopping criteria	SmoothRank [56]

Extended Data Table 51: **THREADS architectural hyperparameters.** WSI: wsi encoder; RNA: scGPT RNA encoder; DNA: DNA encoder

	Hyperparameter	Value
WSI	Pre-attention hidden dimension	1024
	Pre-attention hidden layers	2
	Pre-attention droput	0.1
	Attention heads	2
	Head activation	GeLU
	Head dropout	0.1
	Patch embedding dimension	768
RNA	Encoder	scGPT [18]
	Normalization	$\log_2(\text{Transcripts per million})$
	scGPT data binning	None
	scGPT pretrained weights	Pancancer
	scGPT number of genes to sample	1199
	scGPT hidden dim	512
DNA	Encoder	MLP
	Hidden layers	2
	Hidden dimension	1024
	Activation	ReLU
	Dropout	0.2

Extended Data Table 52: **Label breakdown for the MGB-BRCA morphological subtyping task.** Each sample corresponds to a WSI.

Grade	# Samples
Invasive ductal carcinoma (IDC)	981
Invasive lobular carcinoma (ILC)	283

Extended Data Table 53: **Label breakdown for the MGB-Lung morphological subtyping task.** Each sample corresponds to a WSI.

Grade	# Samples
Lung adenocarcinoma (LUAD)	1616
Lung squamous cell carcinoma (LUSC)	325

Extended Data Table 54: **Label breakdown for the IMP tumor grading task.** Each sample corresponds to a WSI.

Grade	# Samples
0 (non-neoplastic)	847
1 (low-grade)	2847
2 (high-grade)	1639

Extended Data Table 55: **Label breakdown for the PANDA tumor grading task.** Each sample corresponds to a WSI.

ISUP Grade	# Samples
0	2603
1	2399
2	1209
4	1124
3	1118
5	1102

Extended Data Table 56: **Label breakdown for the BRACS coarse-grained morphological subtyping task.** Each sample corresponds to a WSI.

Label	# Samples
Benign tumor	265
Malignant tumor	193
Atypical tumor	89

Extended Data Table 57: **Label breakdown for the BRACS fine-grained morphological subtyping task.**
Each sample corresponds to a WSI.

Label	# Samples
Pathological benign	147
Invasive carcinoma	132
Usual ductal hyperplasia	74
Ductal carcinoma *in situ*	61
Atypical ductal hyperplasia	48
Normal	44
Flat epithelial atypia	41

Extended Data Table 58: **Label breakdown for the EBRAINS coarse-grained diagnosis task.** Each sample
corresponds to a WSI.

Label	# Samples
Adult-type diffuse gliomas	837
Meningiomas	430
Mesenchymal, non-meningothelial tumours involving the CNS	190
Tumours of the sellar region	184
Circumscribed astrocytic gliomas	173
Ependymal Tumours	96
Haematolymphoid tumours involving the CNS	91
Glioneuronal and neuronal tumours	88
Cranial and paraspinal nerve tumours	81
Paediatric-type diffuse low-grade gliomas	70
Metastatic tumours	47
Embryonal Tumors	32

Extended Data Table 59: **Label breakdown for the EBRAINS fine-grained diagnosis task.** Each sample corresponds to a WSI.

Label	# Samples
Glioblastoma, IDH-wildtype	474
Pilocytic astrocytoma	173
Meningothelial meningioma	104
Pituitary adenoma	99
Anaplastic oligodendroglioma, IDH-mutant and 1p/19q codeleted	91
Ganglioglioma	88
Haemangioblastoma	88
Adamantinomatous craniopharyngioma	85
Oligodendroglioma, IDH-mutant and 1p/19q codeleted	85
Atypical meningioma	83
Schwannoma	81
Diffuse astrocytoma, IDH-mutant	70
Transitional meningioma	68
Diffuse large B-cell lymphoma of the CNS	59
Gliosarcoma	59
Fibrous meningioma	57
Anaplastic ependymoma	50
Anaplastic astrocytoma, IDH-wildtype	47
Metastatic tumours	47
Anaplastic astrocytoma, IDH-mutant	47
Ependymoma	46
Anaplastic meningioma	46
Secretory meningioma	41
Lipoma	38
Haemangiopericytoma	34
Glioblastoma, IDH-mutant	34
Medulloblastoma, non-WNT/non-SHH	32
Langerhans cell histiocytosis	32
Angiomatous meningioma	31
Haemangioma	30

References

1. Song, A. H. *et al.* Artificial intelligence for digital and computational pathology. *Nature Reviews Bioengineering* (2023). URL `https://doi.org/10.1038/s44222-023-00096-8`.

2. Van der Laak, J., Litjens, G. & Ciompi, F. Deep learning in histopathology: the path to the clinic. *Nature Medicine* **27**, 775–784 (2021).

3. Bulten, W. *et al.* Artificial intelligence for diagnosis and gleason grading of prostate cancer: the panda challenge. *Nature medicine* **28**, 154–163 (2022).

4. Bejnordi, B. E. *et al.* Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *JAMA* **318**, 2199–2210 (2017).

5. Chen, R. J. *et al.* Towards a general-purpose foundation model for computational pathology. *Nature Medicine* (2024).

6. Xu, H. *et al.* A whole-slide foundation model for digital pathology from real-world data. *Nature* 1–8 (2024).

7. Vorontsov, E. *et al.* A foundation model for clinical-grade computational pathology and rare cancers detection. *Nature Medicine* 1–12 (2024).

8. Chen, R. J. *et al.* Scaling vision transformers to gigapixel images via hierarchical self-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2022).

9. Jaume, G. *et al.* Transcriptomics-guided slide representation learning in computational pathology. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9632–9644 (2024).

10. Jaume, G. *et al.* Multistain pretraining for slide representation learning in pathology. In *Computer Vision – European Conference on Computer Vision 2024*, 19–37 (Springer Nature Switzerland, Cham, 2025).

11. Wang, X. *et al.* A pathology foundation model for cancer diagnosis and prognosis prediction. *Nature* 1–9 (2024).

12. Consortium, G. *et al.* The genotype-tissue expression (gtex) pilot analysis: multitissue gene regulation in humans. *Science* **348**, 648–660 (2015).

13. Shaikovski, G. *et al.* Prism: A multi-modal generative foundation model for slide-level histopathology. *arXiv preprint arXiv:2405.10254* (2024).

14. Van der Maaten, L. & Hinton, G. Visualizing data using t-sne. *Journal of machine learning research* **9** (2008).

15. Lu, M. *et al.* Towards a visual-language foundation model for computational pathology. *Nature Medicine* (2024).

16. Dosovitskiy, A. *et al.* An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations* (2021).

17. Vaswani, A. *et al.* Attention is all you need. *Advances in neural information processing systems* **30** (2017).

18. Cui, H. *et al.* scGPT: Toward building a foundation model for single-cell multi-omics using generative ai. *Nature Methods* 1–11 (2024).

19. Chen, R. J. *et al.* Pathomic fusion: an integrated framework for fusing histopathology and genomic features for cancer diagnosis and prognosis. *IEEE Transactions on Medical Imaging* (2020).

20. Jaegle, A. *et al.* Perceiver: General perception with iterative attention. In *International conference on machine learning*, 4651–4664 (PMLR, 2021).

21. Oquab, M. *et al.* DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research* (2024).

22. Ding, J. *et al.* Longnet: Scaling transformers to 1,000,000,000 tokens. *arXiv preprint arXiv:2307.02486* (2023).

23. Wang, X. *et al.* Transpath: Transformer-based self-supervised learning for histopathological image classification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* (2021).

24. Wang, X. *et al.* Transformer-based unsupervised contrastive learning for histopathological image classification. *Medical Image Analysis* **81**, 102559 (2022).

25. Ilse, M., Tomczak, J. & Welling, M. Attention-based deep multiple instance learning. In *Proceedings of the 35th International Conference on Machine Learning*, 2132–2141 (2018).

26. Oliveira, S. P. *et al.* A CAD system for automatic dysplasia grading on H&E cervical whole-slide images. *Sci. Rep.* **13**, 1–12 (2023).

27. Roetzer-Pejrimovsky, T. *et al.* The Digital Brain Tumour Atlas, an open histopathology resource. *Scientific Data* **9**, 1–6 (2022).

28. Brancati, N. *et al.* BRACS: A Dataset for BReAst Carcinoma Subtyping in H&E Histology Images. *Database* **2022** (2022).

29. Xu, F. *et al.* Predicting axillary lymph node metastasis in early breast cancer using deep learning on primary tumor biopsy slides. *Frontiers in oncology* **11**, 759007 (2021).

30. Radford, A. *et al.* Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020* (2021).

31. Zheng, Z. *et al.* Anchored multiplex pcr for targeted next-generation sequencing. *Nature medicine* **20**, 1479–1484 (2014).

32. Garcia, E. P. *et al.* Validation of oncopanel: a targeted next-generation sequencing assay for the detection of somatic variants in cancer. *Archives of Pathology and Laboratory Medicine* **141**, 751–758 (2017).

33. Jaume, G. *et al.* Modeling dense multimodal interactions between biological pathways and histology for survival prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2024).

34. Vaidya, A. *et al.* Demographic bias in misdiagnosis by computational pathology models. *Nature Medicine* **30**, 1174–1190 (2024).

35. Acosta, P. H. *et al.* Intratumoral resolution of driver gene mutation heterogeneity in renal cancer using deep learning. *Cancer research* **82**, 2792–2806 (2022).

36. Neto, P. C. *et al.* An interpretable machine learning system for colorectal cancer diagnosis from pathology slides. *NPJ precision oncology* **8**, 56 (2024).

37. Pati, P. *et al.* Hierarchical graph representations in digital pathology. *Medical Image Analysis* **75**, 102264 (2022).

38. Edwards, N. J. *et al.* The CPTAC Data Portal: A Resource for Cancer Proteomics Research. *Journal of Proteome Research* **14**, 2707–2713 (2015).

39. Thangudu, R. R. *et al.* Abstract LB-242: Proteomic Data Commons: A resource for proteogenomic analysis. *Cancer Research* **80** (2020).

40. Liao, Y. *et al.* A proteogenomics data-driven knowledge base of human cancer. *Cell Systems* **14**, 777–787 (2023).

41. Wang, C.-W. *et al.* Weakly supervised deep learning for prediction of treatment effectiveness on ovarian cancer from histopathology images. *Computerized Medical Imaging and Graphics* **99**, 102093 (2022).

42. Wilkinson, S. *et al.* Nascent Prostate Cancer Heterogeneity Drives Evolution and Resistance to Intense Hormonal Therapy. *European Urology* **80**, 746–757 (2021).

43. Touat, M. *et al.* Mechanisms and therapeutic implications of hypermutation in gliomas. *Nature* **580**, 517–523 (2020).

44. Ayoub, G. *et al.* Path-50. ai-powered automated tissue segmentation improves outcome stratification in glioblastoma. *Neuro-Oncology* **26**, viii190–viii190 (2024).

45. Domingo-Musibay, E. & Galanis, E. What next for newly diagnosed glioblastoma? *Future Oncology* **11**, 3273–3283 (2015).

46. Martel, A., Nofech-Mozes, S., Salama, S., Akbar, S. & Peikari, M. Assessment of residual breast cancer cellularity after neoadjuvant chemotherapy using digital pathology [data set]. *The Cancer Imaging Archive* (2019).

47. Myles, C., Um, I. H., Harrison, D. J. & Harris-Birtill, D. Leveraging foundation models for enhanced detection of colorectal cancer biomarkers in small datasets. In *Annual Conference on Medical Image Understanding and Analysis*, 329–343 (Springer, 2024).

48. Galland, L. *et al.* Efficacy of platinum-based chemotherapy in metastatic breast cancer and HRD biomarkers: utility of exome sequencing. *npj Breast Cancer* **8**, 1–12 (2022).

49. Bergstrom, E. N. *et al.* Deep learning artificial intelligence predicts homologous recombination deficiency and platinum response from histologic slides. *Journal of Clinical Oncology* **42**, 3550–3560 (2024).

50. Boehm, K. M. *et al.* Multimodal data integration using machine learning improves risk stratification of high-grade serous ovarian cancer. *Nat. Cancer* **3**, 723–733 (2022).

51. Lin, T.-Y. *et al.* Feature pyramid networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2017).

52. Yu, J. *et al.* CoCa: Contrastive Captioners are Image-Text Foundation Models. *Transactions on Machine Learning Research* (2022).

53. Lu, M. Y. *et al.* A multimodal generative AI copilot for human pathology. *Nature* 1–3 (2024).

54. Lu, M. Y. *et al.* Data efficient and weakly supervised computational pathology on whole slide images. *Nature Biomedical Engineering* (2021).

55. Wang, H. *et al.* Path-gptomic: A balanced multi-modal learning framework for survival outcome prediction. *arXiv preprint arXiv:2403.11375* (2024).

56. Garrido, Q., Balestriero, R., Najman, L. & Lecun, Y. Rankme: Assessing the downstream performance of pretrained self-supervised representations by their rank. In *International conference on machine learning*, 10929–10974 (PMLR, 2023).

57. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778 (2016).

58. Deng, J. *et al.* Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255 (Ieee, 2009).

59. Pölsterl, S. scikit-survival: A library for time-to-event analysis built on top of scikit-learn. *Journal of Machine Learning Research* **21**, 1–6 (2020). URL `http://jmlr.org/papers/v21/20-729.html`.

60. Dunn, O. J. Multiple comparisons among means. *Journal of the American statistical association* **56**, 52–64 (1961).

61. Searle, S. R., Speed, F. M. & Milliken, G. A. Population marginal means in the linear model: an alternative to least squares means. *The American Statistician* **34**, 216–221 (1980).

62. Robertson, H. *et al.* Decoding the hallmarks of allograft dysfunction with a comprehensive pan-organ transcriptomic atlas. *Nature Medicine* 1–10 (2024).

63. Pedregosa, F. *et al.* Scikit-learn: Machine learning in python. *the Journal of machine Learning research* **12**, 2825–2830 (2011).

64. Kolesnikov, A., Zhai, X. & Beyer, L. Revisiting self-supervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2019).

65. Russakovsky, O. *et al.* Imagenet large scale visual recognition challenge. *International journal of computer vision* **115**, 211–252 (2015).