nature communications

Article

https://doi.org/10.1038/s41467-025-58796-1

A clinical benchmark of public self-supervised pathology foundation models

Received: 8 July 2024

Accepted: 2 April 2025

Published online: 17 April 2025

Check for updates

Gabriele Campanella[1,2] ✉, Shengjia Chen[1,2], Manbir Singh[1,2], Ruchika Verma[1,2], Silke Muehlstedt[1,2], Jennifer Zeng[3], Aryeh Stock[3], Matt Croken[3], Brandon Veremis[3], Abdulkadir Elmas[4], Ivan Shujski[5,6], Noora Neittaanmäki[5,6], Kuan-lin Huang[4], Ricky Kwan[3], Jane Houldsworth[3], Adam J. Schoenfeld[7] & Chad Vanderbilt[8] ✉

The use of self-supervised learning to train pathology foundation models has increased substantially in the past few years. Notably, several models trained on large quantities of clinical data have been made publicly available in recent months. This will significantly enhance scientific research in computational pathology and help bridge the gap between research and clinical deployment. With the increase in availability of public foundation models of different sizes, trained using different algorithms on different datasets, it becomes important to establish a benchmark to compare the performance of such models on a variety of clinically relevant tasks spanning multiple organs and diseases. In this work, we present a collection of pathology datasets comprising clinical slides associated with clinically relevant endpoints including cancer diagnoses and a variety of biomarkers generated during standard hospital operation from three medical centers. We leverage these datasets to systematically assess the performance of public pathology foundation models and provide insights into best practices for training foundation models and selecting appropriate pretrained models. To enable the community to evaluate their models on our clinical datasets, we make available an automated benchmarking pipeline for external use.

Artificial Intelligence (AI) is revolutionizing the medical field. The introduction of deep learning[1] has greatly accelerated the development of predictive models for high-dimensional data modalities such as images and text that are not readily amenable to classical machine learning algorithms. Convolutional neural networks (CNNs) and vision transformers[2] (ViTs) have been used to solve numerous problems using supervised learning and have enabled the training of predictive models for a variety of tasks with high performance[3–9]. Recently, the development of self-supervised learning (SSL) algorithms has marked a paradigm shift by enabling the training of deep neural networks on very large unlabeled datasets, yielding results on par with supervised learning strategies[10,11]. Large neural networks trained this way can be described as foundation models that can be used for a wide variety of downstream tasks with little to no fine-tuning. Despite the great successes in the computer vision and natural language fields, SSL algorithms and foundation models are still

[1]Windreich Department of AI and Human Health, Icahn School of Medicine at Mount Sinai, New York 10029 NY, USA. [2]Hasso Plattner Institute at Mount Sinai, Icahn School of Medicine at Mount Sinai, New York 10029 NY, USA. [3]Department of Pathology, Icahn School of Medicine at Mount Sinai, New York 10029 NY, USA. [4]Department of Genetics and Genomics, Icahn School of Medicine at Mount Sinai, New York 10029 NY, USA. [5]Department of Clinical Pathology, Sahlgrenska University Hospital, Gothenburg, Sweden. [6]Department of Laboratory Medicine, University of Gothenburg, Gothenburg, Sweden. [7]Department of Medicine, Memorial Sloan Kettering Cancer Center, New York 10065 NY, USA. [8]Department of Pathology, Memorial Sloan Kettering Cancer Center, New York 10065 NY, USA. ✉e-mail: gabriele.campanella@mssm.edu; vanderbc@mskcc.org

in their infancy in the medical domain. One of the main reasons is the lack of medical datasets and the necessary computing infrastructure, which makes large-scale SSL experiments only possible at large well-funded institutions.

In pathology, the lack of data is even more acute due to the still low adoption of digital pathology. Additionally, digital whole slide images (WSI) are orders of magnitude larger than other image modalities, with resolutions of tens to hundreds of thousands of pixels in each dimension. This poses challenges in terms of the methods used to analyze the images and the hardware requirements to effectively perform experiments. A common strategy to analyze these images is to divide the slide into small tiles or patches and encode them using a deep neural network, expressing the slide as a list of feature vectors and thus reducing the dimensionality of the slide by multiple orders of magnitude[2,12]. In a second step, the feature vectors are aggregated using a neural network to obtain a slide-level representation[12,13]. The first step is by far the most computationally expensive, while the second step requires much fewer resources. This is why most studies in computational pathology rely on already existing pretrained encoders, usually trained on natural images and not WSIs[12,14–17]. There is a need for strategies that enable training of encoders directly on pathology images, and SSL lends itself well for this task as it does not require any sort of labels and thus enables the training of pathology foundation models on large unannotated datasets. SSL for pathology has recently received lots of attention, and there are many academic and non-academic efforts to build a general-purpose pathology foundation model (Table 1).

Wang et al.[18] proposed SRCL, an SSL method based on MoCo v3[10], along with CTransPath, a model architecture that combines convolutional layers with the Swin Transformer[19] model. They trained their model on 15.6 million tiles from 32,220 slides from the TCGA[20] and PAIP datasets spanning 25 anatomic sites and over 32 cancer subtypes. The downstream performance was assessed on patch retrieval, supervised patch classification, weakly supervised WSI classification, mitosis detection, and colorectal adenocarcinoma gland segmentation. Methodological advances include the introduction of a strategy to sample positive examples for the contrastive learning approach, and the hybrid convolutional-transformer model architecture.

Filiot et al.[21] analyzed the performance of iBOT[22], an SSL framework that combines masked image modeling and contrastive learning, on histology data. They trained several ViT models on a dataset consisting of up to 43.3 million tiles from 6093 TCGA slides of 13 anatomic sites. They assessed the performance of learned features on 17 downstream tasks across seven cancer indications, including tile-level and slide-level tasks for subtype, genomic alteration, and overall survival prediction. Ultimately, they publicly released Phikon[21], a ViT-base model.

Chen et al.[23] introduced UNI, a ViT-large model trained on 100,000 proprietary slides using the DINOv2[24] SSL algorithm. The pretraining dataset they used included 100 million tiles from 20 major tissue types. They evaluated the downstream performance across 33 tasks, which included tile-level tasks such as classification, segmentation, retrieval, as well as slide-level classification tasks.

Vorontsov et al.[25] introduced Virchow, a ViT-huge model trained on 2 billion tiles from almost 1.5 million proprietary slides with DINOv2[24]. Slides were included from 17 tissue types and the performance on downstream tasks was evaluated using tile-level and slide-level benchmarks, encompassing tissue classification and biomarker prediction.

Campanella et al.[26] compared the performance of masked autoencoders[27] (MAE) and DINO[28] using over 3 billion tiles sourced from more than 423,000 pathology slides. The models were evaluated on six clinical tasks spanning three anatomical sites and two institutions. Their results showed the superiority of the DINO algorithm for pathology foundation model pretraining.

Dippel et al.[29] introduced RudolfV, a model that integrates pathologist expertise, semi-automated data curation, and a diverse dataset from over 15 laboratories. Their dataset comprised 134,000 slides from 34,000 cases, representing a broad spectrum of histological samples with various fixation, staining, and scanning protocols from laboratories across the EU and US. Additionally, semantically similar slides and tissue patches were grouped to optimize data sampling for training, and stain-specific data augmentation was applied.

Xu et al.[30] introduced Prov-GigaPath, that was created by tile-level pretraining using DINOv2[24], followed by slide-level pretraining using a masked autoencoder[27] and LongNet[31]. This model was pretrained on 1.3 billion tiles derived from 171,189 WSIs comprising H&E-stained and immunohistochemistry (IHC) slides from Providence Health and Services. These WSIs originated from over 30,000 patients encompassing 31 tissue types. Prov-GigaPath was evaluated on 17 genomic prediction tasks and 9 cancer subtyping tasks using both Providence and TCGA[32] data.

Zimmermann et al.[33] introduced two models, Virchow2 (ViT-huge) and Virchow2G (ViT-giant), trained on 1.7 billion and 1.9 billion tiles, respectively, from 3.1 million histopathology whole slide images with DINOv2. Virchow2 examines the impact of increased data scale and diversity across multiple magnifications, while Virchow2G focuses on scaling model size. Slides from 225,401 patients, with

Table 1 | A summary of recently published pathology foundation models

Model	Param. (M)	Algorithm	Training Data Source	Tiles (M)	Slides (K)	Training Resolution
CTransPath[18]	28	SRCL	TCGA, PAIP	16	32	20x
Phikon[21]	86	iBOT	TCGA	43	6	20x
UNI[23]	303	DINOv2	MGB	100	100	20x
Virchow[25]	631	DINOv2	MSKCC	2000	1488	20x
Ref. 26	22	DINO	MSHS	1600	423	20x
Ref. 26	303	MAE	MSHS	3200	423	20x
Rudolf-V[29]	304	DINOv2	Multicenter	1200	134	20x
Prov-GigaPath[30]	1135	DINOv2	PHS	1300	171	20x
Virchow2[33]	631	DINOv2	MSKCC	1700	3100	5,10,20,40 ×
H-optimus-0[34]	1135	DINOv2	Proprietary	>100	>500	20x
Phikon-v2[35]	307	DINOv2	Multicenter	456	58	20x

MGB Mass General Brigham, *MSKCC* Memorial Sloan Kettering Cancer Center, *MSHS* Mount Sinai Health System, *PHS* Providence Health and Services.

magnifications of 5x, 10x, 20x, and 40x, were included from both H&E and IHC stains across nearly 200 tissue types. Their approach achieved state-of-the-art performance on 12 tile-level tasks, surpassing the top-performing competing models.

Saillard et al.[34] introduced H-optimus-0, a ViT-giant model trained on hundreds of millions of tiles derived from over 500,000 proprietary H&E stained whole-slide histology images. The dataset included slides from a wide range of tissue types, and the model's performance was assessed on tile-level and slide-level benchmarks, covering tasks such as tissue classification, mutation prediction, and survival analysis.

Phikon-v2[35], a self-supervised vision transformer trained using DINOv2[24], demonstrates superior performance compared to its predecessor, Phikon-v1[21], and achieves results comparable to other leading histopathology foundation models. The model was pretrained on a diverse dataset of 460 million pathology tiles derived from over 100 publicly available cohorts, encompassing more than 50,000 histopathology slides across 30 cancer types. Benchmark evaluations, covering eight slide-level tasks with external validation cohorts, highlight its robust performance and generalizability.

It is becoming abundantly clear that using SSL to train image encoders on unlabeled pathology data is superior to relying on models pretrained on other domains such as natural images[26,36,37]. While SSL-trained pathology models hold immense potential, there are still challenges that need to be overcome before pathology foundation models can be used reliably in clinical workflows. One consideration is that datasets used to train pathology models are still relatively small compared to other domains, in particular natural images, especially when considering the number of slides or cases. Since each pathology slide can contain tens of thousands of tiles, it is possible to generate large number of tiles from a small number of slides. Thus, it is essential to consider not only the number of tiles or slides used, but also other metrics of tissue heterogeneity such as anatomic sites and organ inclusion. Given the evidence from the natural language and vision domains that larger datasets and higher capacity models will produce better performance especially in the SSL setting[38–40], training on larger pathology datasets should be a priority. Recent works show progress in this respect as the digitization of pathology data becomes more prevalent[25,30,33]. Most importantly, the downstream performance of SSL models for pathology should be assessed on clinically derived data, preferably from multiple institutions, for clinically relevant tasks such as diagnostic assessment, biomarker prediction, and outcome prediction. This effect is compounded by the use of curated public datasets which may not be suited for assessing generalization to real world data. It should be noted that progress in this regard is being made and a trend towards the use of more clinical data in recent publications can be observed. Yet, there is still a lack of a systematic comparison of current models on a wide variety of clinical tasks.

In the present work we overcome this limitation by introducing a clinical benchmark dataset which is used to systematically compare public pathology foundation models. In contrast to previous efforts[36,41], the dataset consists of clinical data generated during standard hospital operations from three health systems. It includes two broad task types (disease detection and biomarker prediction), and a wide range of disease indications and anatomic sites. Considering the rate of progress in computational pathology, as new foundation models are published and additional datasets are added to our benchmarks, we will regularly update our findings to provide the community with a comprehensive view of the state of foundation models in computational pathology. The live benchmark can be found in the official GitHub repository. In addition, we provide an automated benchmarking mechanism for external users who wish to take advantage of our clinical cohorts. Instruction are provided on GitHub.

Results

Disease Detection Tasks

For disease detection, all models show consistent performance across all tasks with AUCs above 0.9 in all cases (Fig. 1). The ImageNet pretrained encoder is consistently under-performing the pathology trained encoders. This behavior is statistically significant across tasks with the exception of the Thyroid cohort (see Supplementary Fig. 7). Among the pathology trained encoders, CTransPath consistently shows inferior performance. This behavior is statistically significant across tasks with the exception of the Colorectal, Oral, and Thyroid cohorts (see Supplementary Fig. 7). CTransPath was trained on a relatively small dataset and used a contrastive learning algorithm, which may explain the difference in performance. The other foundation models tested were trained with either iBOT, DINO, or DINOv2. In general, they all achieve similar performances with largely non statistically significant differences despite the diversity in pretraining datasets and model architectures (see Supplementary Fig. 7). Ranking the models based on the average AUC across tasks (Supplementary Fig. 9), the top 3 ranked models are H-optimus-0, Prov-GigaPath, and SP85M. Overall, for detection tasks, all the DINO and DINOv2 trained models achieve comparable performance and the choice of model may depend on other considerations, such as inference cost.

Computational Biomarker Prediction Tasks

Biomarker prediction tasks are more challenging than disease detection tasks since it is generally unknown whether measurable morphological changes in H&E stained slides even exist for the biomarker of interest. For some biomarkers, prediction from H&E may not be feasible. As expected, the biomarker prediction tasks show a higher degree of variability in performance than the detection tasks (Fig. 2). The gap in performance of the ImageNet pretrained model is more evident here than in the detection tasks. Pairwise comparisons with TRes50 are statistically significant except for the NSCLC IO task (see Supplementary Fig. 8). As before, CTransPath tends to perform worse than the DINO and DINOv2 models. This difference is for the most part statistically significant with the exception of the NSCL IO, breast HRD, and melanoma BRAF tasks (see Supplementary Fig. 8). On the other end of the spectrum, H-optimus-0 and Prov-GigaPath tend to be significantly better than other models in the majority of tasks (see Supplementary Fig. 8). The main exceptions include the Breast HRD, Melanoma, and NSCLC IO tasks. For the other models, it is more difficult to make general statements. Ranking the models based on the average AUC across tasks (Supplementary Fig. 10), H-optimus-0, Prov-GigaPath, and UNI are the top 3 ranked.

Stratifying the analysis by biomarker panels, we can make some further observations. For the breast cancer IHC/FISH biomarkers, the observations made before are largely accurate with H-optimus-0, Prov-GigaPath, and UNI performing-generally better. Here Virchow and Virchow2 also compare positively to some of the other models. For the somatic mutation panel in melanoma, differences in performance between the various models are less obvious. For the somatic NGS panel in LUAD, again H-optimus-0, Prov-GigaPath, and UNI tend to be significantly better than the other models. Interestingly, we noticed that the prevalence of lung tissue in the pretraining cohort explain in part the lung biomarker results. For UNI, lung is the second most common tissue in their dataset, with around 10% of the slides or about ten thousand WSIs. For Prov-GigaPath, lung is the most common tissue, comprising over 45% of the slides, or about 77 thousand WSIs. This points to the hypothesis that while for detection tasks, dataset composition seems not an important factor, it may play a significant role for biomarker prediction.

Finally, for the task of predicting ICI response in NSCLC, all models obtained equally poor results with AUCs barely above chance. UNI, with and average AUC of 0.6 performed performed significantly

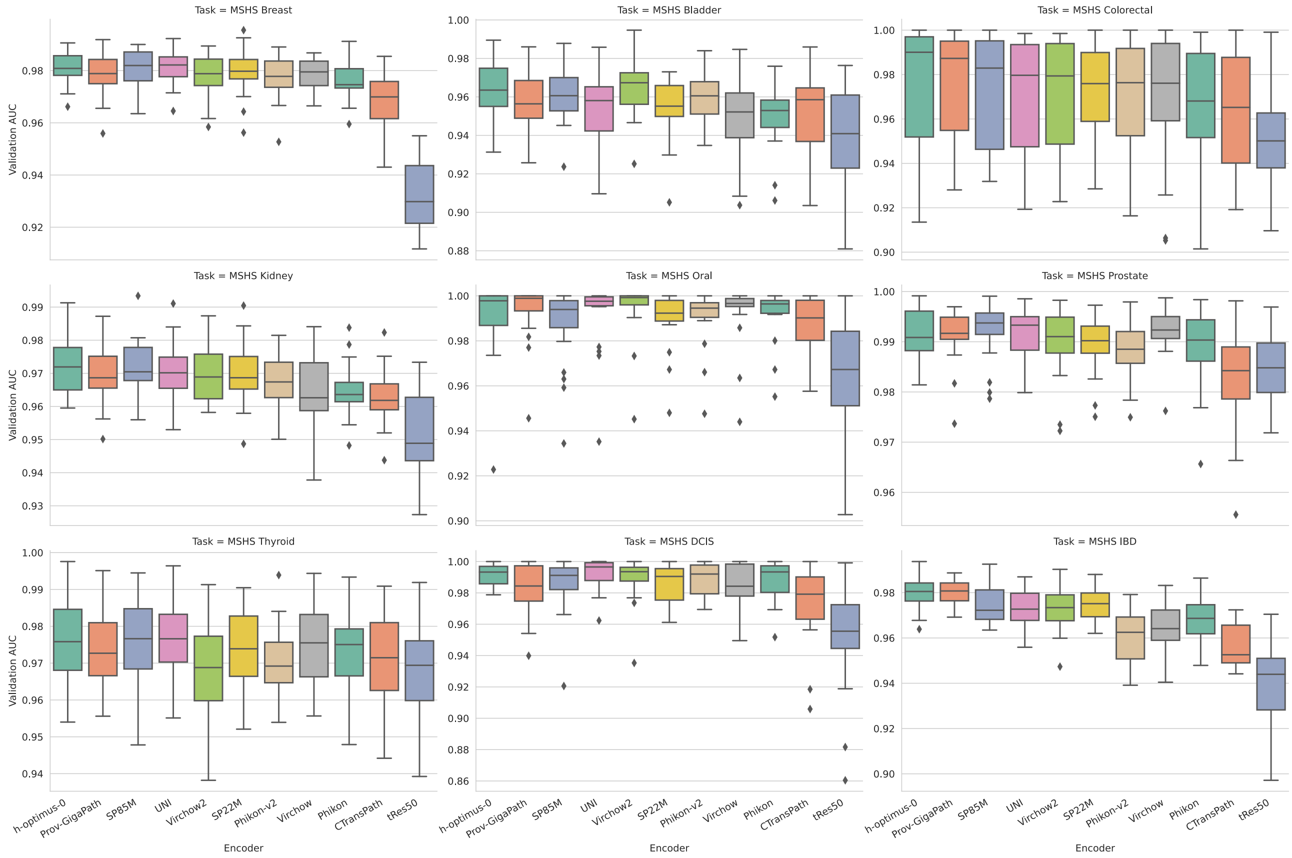

Fig. 1 | Benchmarking Results: Detection Tasks. Each box plots summarizes the distribution of validation performance across 20 MCCV splits ($N = 20$). Boxes show the quartiles of each distribution, while the whiskers extend 1.5 times the interquartile range. Source data are provided as a Source Data file.

better than the other models. ICI response prediction from H&E slides is a challenging task, yet there is evidence that descriptors of local cellular networks[42], that better model the tumor microenvironment (TME) can achieve AUCs of around 0.7, on-par with PD-L1 IHC, the current clinical gold standard. It is reasonable to hypothesize that SSL-trained foundation models should be able to capture local cellular information and reach similar performance. One potential explanation is that the pretraining data may be skewed in terms of cancer presence, cancer subtype, and cancer stage. Given that foundation models are trained on large data collections with minimal to no data curation, the magnitude of these biases with this level of detail is generally not measurable. Yet, this result suggests that the composition of the pretraining dataset may be crucial, especially for challenging response prediction tasks.

Foundation Model Size

One important aspect of foundation models is their representational capacity which can be roughly estimated by the model's parameter count. Here we investigate how model size correlates with downstream performance to assess whether scaling laws observed in other domains, such as natural language processing are occurring for pathology data. For this analysis we excluded tRes50 and CTransPath to restrict the analysis to vision transformers trained with iBOT, DINO, or DINOv2 (UNI, Virchow, Prov-GigaPath, SP22M, SP85M, Virchow2, h-optimus-0, Phikon-v2, Phikon). Model sizes range from 22 million (SP22M) to 1.1 billion (Prov-GigaPath) parameters (see Table 1).

Figure 3 shows how the downstream performance of detection and biomarker prediction tasks correlate with encoder model sizes. For detection tasks, our results suggest a tendency of downstream performance scaling with model size, but this effect is rather minor (Pearson statistic: 0.055, *p* value: 2.59e-2). As we showed previously, on average a 22 million parameter model is comparable to a 1.1 billion parameter model for these tasks. For biomarker prediction, an overall tendency of higher performance with larger models is observed to a more significant extent compared to the detection tasks (Pearson statistic: 0.091, *p* value: 9.52e-6). While overall biomarkers for biomarkers the model size correlates with performance, we observed that this effect may be task-dependent. Focusing on several breast biomarkers, there is no benefit from larger models, whereas for the NGS lung tasks there seems to be a larger benefit. As noted earlier, this may be due to the pretraining dataset composition and not to the larger model capacity.

Pretraining Dataset Size

Next, we investigated the effect of pretraining dataset size on the downstream performance in terms of number of slides and number of tiles. The models included in the analysis were trained on datasets with a wide range of number of slides, from six thousand (Phikon) to over three million (Virchow2). Focusing on slides used for pretraining, we observed no evidence that larger pretraining datasets are correlated with better performance. This was true for both detection and biomarker tasks (see Supplementary Fig. 2; detection tasks: $r = 0.008$, *p* val = 7.58e-01; biomarker tasks: $r = -0.033$, *p* val = 1.07e-01). Similarly, the number of tiles used to pretrain the foundation models varied widely, from 43 million (Phikon) to 1.7 billion (Virchow2). For detection tasks we observed a slight trend of increased performance, while for biomarker tasks we observed a slight trend of decreased performance (see Supplementary Fig. 3; detection tasks: $r = 0.050$, *p* value = 7.79e-02; biomarker tasks: $r = -0.048$, *p* value = 4.03e-02). Overall, the dataset size does not correlate strongly with downstream performance.

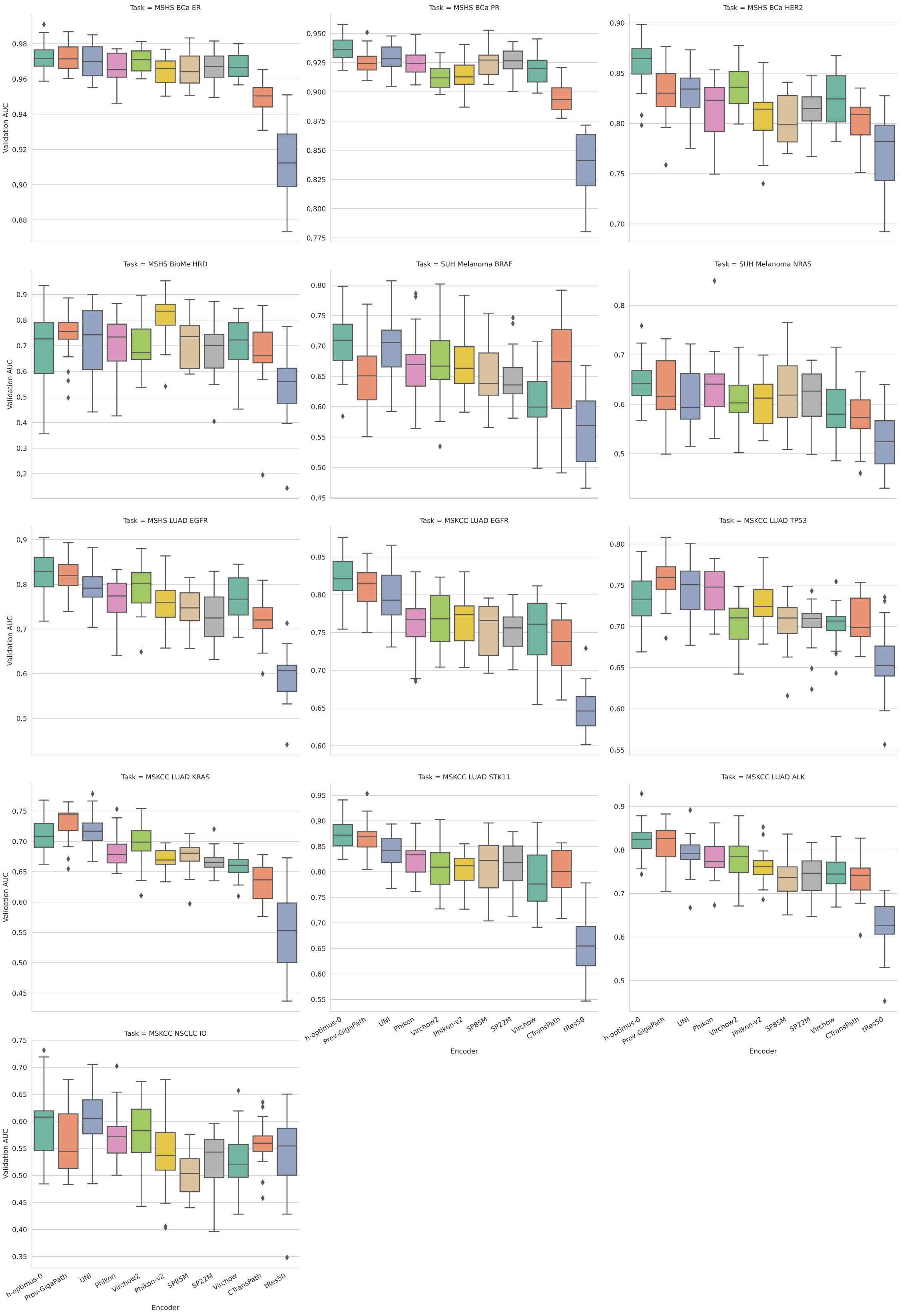

Fig. 2 | Benchmarking Results: Biomarker Prediction Tasks. Each box plots summarizes the distribution of validation performance across 20 MCCV splits ($N = 20$). Boxes show the quartiles of each distribution, while the whiskers extend 1.5 times the interquartile range. Source data are provided as a Source Data file.

Computational Resources

Previously, we analyzed model size and dataset size independently. To assess their joint effect, we studied the extent to which the overall pretraining computational resources explain downstream performance. Computational resources needed to train a model depend on the model size, dataset size, and also the pretraining algorithm. For example, SP22M with 22 million parameters was trained with DINO using full precision on 40GB GPUs with a batch size per GPU of 90 tiles. In

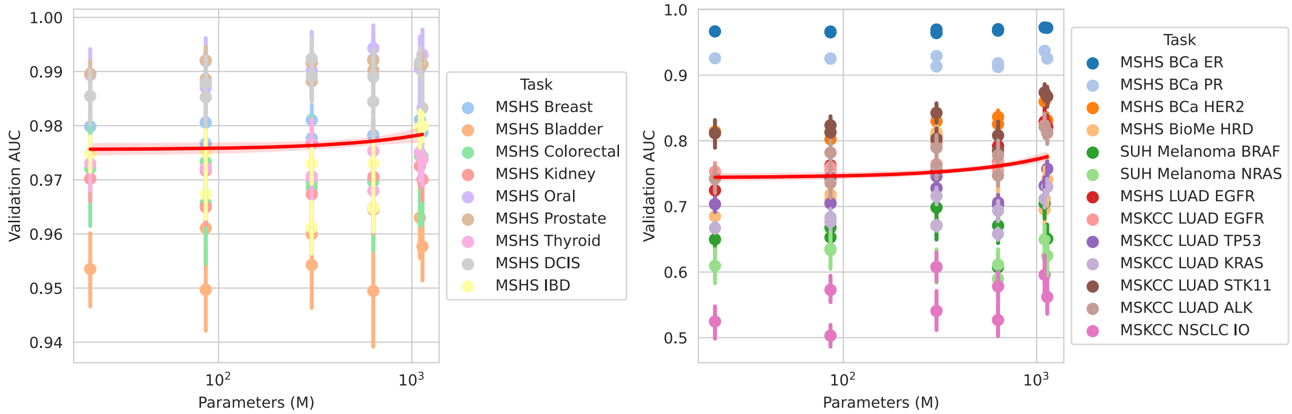

Fig. 3 | Scaling Laws: downstream performance vs foundation model size. Scatter plots summarize the validation performance across 20 MCCV runs for each task (*N* = 20). The error bars show the 95% confidence interval calculated via bootstrapping with 1000 iterations. Linear tendency line and 95% bootstrapped confidence interval is shown in red. The line summarizes the performance across all tasks at once (detection tasks *N* = 1620; biomarker tasks *N* = 2340). Left: detection tasks, Pearson correlation coefficient: 0.055, two-sided *p*-value: 2.59e-2. Right: biomarker tasks, Pearson correlation coefficient: 0.091, two-sided *p* value: 9.52e-6. *P* values not adjusted for multiple hypothesis testing. Source data are provided as a Source Data file.

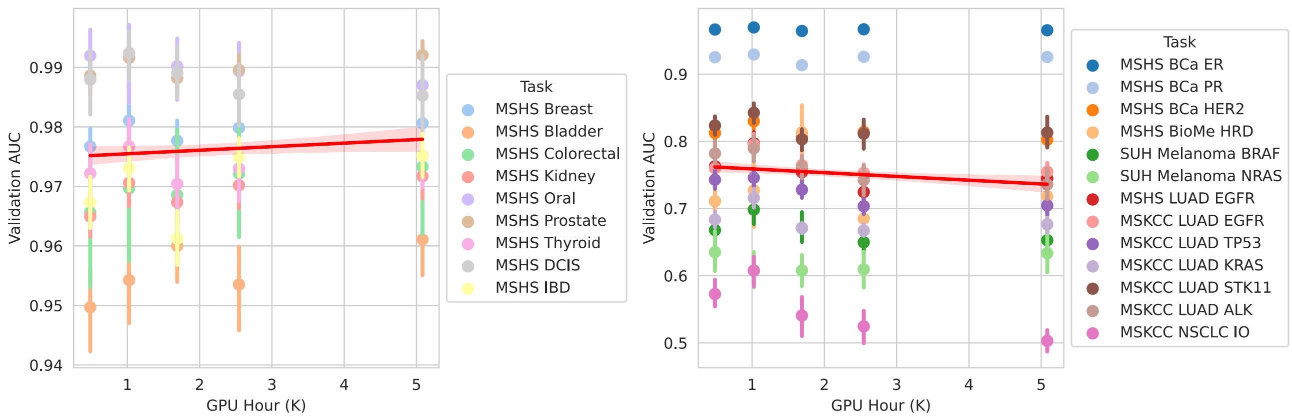

Fig. 4 | Scaling Laws: downstream performance vs computational resources used for pretraining the foundation models. Scatter plots summarize the validation performance across 20 MCCV runs for each task (*N* = 20). The error bars show the 95% confidence interval calculated via bootstrapping with 1000 iterations. Linear tendency line and 95% bootstrapped confidence interval is shown in red. The line summarizes the performance across all tasks at once (detection tasks *N* = 900; biomarker tasks *N* = 1300). Left: detection tasks, Pearson correlation coefficient: 0.055, two-sided *p* value: 1.01e-1. Right: biomarker tasks, Pearson correlation coefficient: -0.074, two-sided *p* value: 7.51e-3. *P* values not adjusted for multiple hypothesis testing. Source data are provided as a Source Data file.

comparison, Prov-GigaPath with 1.1 billion parameters was trained with DINOv2 using half-precision on 80GB GPUs with a batch size of 12 tiles per GPU. To harmonize the various training runs, we measured overall computational resources using GPU-hours normalized to a hypothetical 80GB GPU card. We assume that, for models trained on a 40GB card, the computation time would be halved by using an 80GB card. GPU usage and training times were obtained from each respective paper or model cards in public repositories and are summarized in Supplementary Table 6. Foundation models included in this analysis are: SP22M, SP85M, UNI, Phikon, Phikon-v2. Others had to be excluded due to lack of data.

Figure 4 shows how the downstream performance of detection and biomarker prediction tasks correlate with computational resources used for training. For detection tasks, our results show no evidence of improved performance associated with higher computational costs (Pearson statistic: 0.055, *p* value: 1.01e-1). The same conclusion can be made for biomarker prediction tasks where the linear tendency even had a slight negative slope (Pearson statistic: −0.074, *p* value: 7.51e-3). It is important to note the lack of data for Prov-GigaPath and H-optimus-0 in this analysis. Based on these results, we highlight how UNI, while trained with a comparatively modest resource budget, achieves competitive performance, especially in our biomarker tasks. While this could be explained again by the prevalence of lung tasks, it may also point to the importance of pretraining dataset composition.

Pretraining Dataset Composition

We have previously hinted that the composition of the pretraining dataset may be a crucial aspect for explaining downstream performance. We explore this hypothesis more in detail by correlating the performance of tissue-specific tasks with the percentage of the pretraining dataset devoted to that tissue in terms of slides. To perform this analysis, we collected tissue percentages for each model from their respective papers when available. The data collected is presented in Supplementary Table 7. The foundation models included in this analysis are: SP22M, SP85M, UNI, Prov-GigaPath, and Virchow. To investigate the effect of the tissue prevalence, we combined tasks by tissue and analyzed the AUC performance as a function of tissue percentage. We focused on four tissue/organs that had the most complete data: lung, breast, colon/rectum, and prostate.

For lung, our benchmark contains seven biomarker tasks. Tissue prevalence ranged from 0.36% (SP22M, SP85M) to 45.29% (Prov-GigaPath). Correlating the performance on lung tasks with tissue prevalence yielded the strongest effect with a Pearson statistic of 0.229 (*p*-value: 9.50e-10, Supplementary Fig. 4a). For breast, we have two detection tasks and four biomarker tasks. Tissue prevalence ranged from 2.76% (Prov-GigaPath) to 24.90% (Virchow). We observed no significant correlation between performance and tissue prevalence for the breast tasks considered with a Pearson statistic of -0.044 (*p* value: 2.83e-01, Supplementary Fig. 4b). For colon and rectum, we have two detection tasks. Tissue prevalence ranged from 3.20% (Virchow) to 30.43% (Prov-GigaPath). We didn't observe a significant correlation between performance and tissue prevalence for the breast tasks considered with a Pearson statistic of 0.122 (*p* value: 8.59e-2, Supplementary Fig. 4c). Finally, for prostate, only one task was considered. Tissue prevalence ranged from 0% (UNI) to 10.61% (SP22M, SP85M). No significant correlation was observed with a Pearson statistic of -0.068 (*p*-value: 4.99e-1, Supplementary Fig. 4d). Overall, the importance of the tissue prevalence seems to be tissue-specific. While there was an indication of positive correlation for the lung tasks, the same was not the case for the other organs tested.

Foundation Model Inference

Compared to pretraining, model inference covers a fraction of the computational expense, yet it is an important consideration for model deployment. On comparable hardware, inference largely depends on model architecture. To assess the inference performance of foundation models, we measured the minimal GPU memory required and the maximum throughput in tiles per second (TPS). These analyses were conducted using synthetic data. Each condition was repeated 20 times to assess variability.

First, we considered the minimal GPU memory required in gigabytes (GB) as the memory necessary to run a forward pass through a model with a single image (batch size of 1). Supplementary Fig. 5a shows the memory requirements for each foundation model we considered. Requirements range from 0.127GB for SP22M, to 4.643GB for H-optimus-0 (see Supplementary Table 8). We related the memory requirements with the average performance for detection and biomarker tasks for each model. For detection tasks (Supplementary Fig. 5b), we identified SP85M as the best trade-off between memory and performance with an average AUC of 0.978 across all detection tasks and a minimal memory requirement of 0.387GB. It compares favorably against the best-performing H-optimus-0 which requires 4643GB for a 0.2% increase in performance on average over SP85M. For biomarker tasks (Supplementary Fig. 5c), we identified UNI as offering the best trade-off with its average AUC of 0.773 across all biomarker tasks with a 1.259GB memory requirement. The best-performing H-optimus-0 required 4.643GB to achieve an average AUC of 0.785, a 1.3% increase in performance over UNI.

We also analyzed the inference performance in terms of maximal throughput. For this, we considered the average number of tiles that can be processed by a foundation model every second assuming no data loading bottleneck. We run this analysis on a single H100 80GB GPU. To maximize throughput, it is important to utilize the GPU close to its full memory capacity. To this end, we first identified for each foundation model the largest batch size (that is a multiple of 8) that allows to run a forward pass and used that for the analysis. Supplementary Fig. 6a shows the TPS throughput for all models considered. These ranged from 4417.4 TPS by the truncated ResNet50 (followed by 2569.4 TPS by SP22M) to 75.3 TPS by H-optimus-0 (see also Supplementary Table 9). We analyzed the relation between TPS and average AUC performance for detection and biomarker tasks. The results mostly coincided with our previous analysis of minimum memory requirements. For detection tasks (see Supplementary Fig. 6b), SP85M seems to strike the best trade-off between TPS and AUC. Compared to the best performer, it shows a 0.2% decrease in average performance (0.980 for H-optimus-0 vs 0.978 for SP85M) with a 14-fold increase in throughput (75.3 TPS for H-optimus-0 vs 1063.6 TPS for SP85M). For biomarker tasks (see Supplementary Fig. 6c), these results highlight UNI and Phikon. Compared to the best performer, UNI shows a 1.5% drop in average AUC (0.773 for UNI vs 0.785 for H-optimus-0) with a 4.8-fold increase in throughput. Instead, Phikon shows a 3.5% decrease in AUC compared to H-optimus-0 and a 2.1% decrease compared to UNI, but achieves a throughput 14.2 and 3 times higher, respectively.

Discussion

Self-supervised learning and foundation models have the potential to revolutionize medical research. Training foundation models for computational pathology is showing a clear benefit over traditional supervised approaches in terms of performance and generalizability. Notably, recent models trained by both academic and private institutions are being released in public repositories, empowering researchers with the tools to develop the next generation of predictive models. While there is still much work to be done towards democratizing computational pathology-based decision support systems and making them available to the research community, the emergence of foundation models likely will play a significant role.

As more and more foundation models are trained, an independent benchmark of clinically relevant tasks becomes essential for both researchers training foundation models and looking to apply these pretrained foundation models on downstream tasks. Training new foundation models is expensive and it is important to learn from previous efforts. A benchmark can provide insights for improving pretraining and yield better models in the future. For downstream clinical applications, a benchmark can guide the decision to use one model over another, considering a variety of factors, from performance on various tasks to computational resource constraints. In this work, we presented a benchmark of publicly available pathology foundation models focusing on 22 clinically relevant slide-level tasks across a variety of tissues and disease indications. Importantly, all the data was generated during clinical operations without further curation, representing the variability, both biological and technical, that can be observed under real-world conditions. Further, most foundation models were trained on a combination of public and private datasets without any overlap with the cohorts in this study. As an exception, we must note that since the Virchow and Virchow2 models were trained on a large sample of slides from MSKCC, we can't ensure that there is no overlap between their pretraining cohort and the clinical tasks based on MSKCC data. In addition, the SP22M and SP85M models were trained on MSH data but we ensured no overlap with the clinical tasks.

We made a deliberate decision to not release the test data used for these benchmarks. Efforts to scrape all publicly available data for pretraining foundation models may lead to data contamination and negatively impact the relevance of such benchmark results. Instead, we will regularly update the benchmark results with the latest models as they become publicly available. In addition, we are exposing an automated pipeline allowing external users to benchmark their foundation models on our clinical cohorts. Instructions are provided in the GitHub repository.

In summary, the ImageNet pretrained encoder, and CTransPath to a lesser degree, consistently underperformed compared to newer models. For the disease detection tasks, in general all DINO and DINOv2-trained models performed comparably. H-optimus-0 and Prov-GigaPath performed significantly better in a few tasks, yet the improvement is minimal. For biomarker tasks, there was a larger spread of performances, here H-optimus-0, Prov-GigaPath, and UNI compared favorably to the other models in most tasks. We also found that model size and pretraining dataset composition influenced performance, particularly for biomarker prediction, but not necessarily for disease detection. Additionally, inference cost and computational efficiency varied, with models like SP85M and UNI offering a good balance between performance and resource usage. The findings underscore the importance of pretraining dataset composition and model architecture in optimizing performance for specific tasks.

From our analyses, we can make the following observations: Strong evidence does not yet exist supporting that scaling laws observed in pretraining SSL models for natural language and images are applicable for tile encoders in pathology. Performance does not scale with model size and dataset size as clearly as in other domains given current training algorithms. Smaller models perform on par with much larger models on most tasks and are only marginally worse in others, particularly for detection tasks. Similarly, the dataset size and overall computational expense does not appear to lead to significantly better models. It is likely that dataset composition may be a crucial aspect in the downstream performance, and more efforts in the curation of the pretraining data is likely to be beneficial. While general-purpose foundation models may be desirable, tissue-specific foundation models may be a viable alternative. Recent efforts to benchmark pathology foundation models by Neidlinger et al.[41] have come to similar conclusions. Moving forward, we expect small incremental performance gains with current SSL algorithms. It is possible that we are saturating the capabilities of current SSL strategies in pathology and great leaps forward may not occur without innovations from the algorithmic side or integration of SSL with other forms of supervision. Finally, we hypothesize that for challenging tasks like ICI response, tile-level encoders alone, especially with current tile sizes in the order of 224 pixels, are not enough to fully describe all the relevant features and slide-level aggregators are likely to play an important role. More research in this direction will be needed as there is currently a lack of strategies that are capable of fully leveraging the global topology of the tissue in a slide[43].

In this work we have focused on gathering a large set of clinically relevant downstream tasks. We were able to include a variety of disease indications and task types. Yet, in the current version of the benchmark, the technical variability (tissue fixation, staining, scanning, etc.) is limited to three institutions. Additionally, here we focused on comparing the expressivity of tile-level features generated by pathology foundation models. To that end we leveraged GMA aggregator with a linear classifier. We considered the exploration of more complex aggregators as out of scope and it has been addressed in other studies[43].

There are several aspects of pretraining pathology foundation models that we could not address at this time due to lack of evidence. The majority of foundation models have been trained at 20x magnification as it allows to use the largest possible cohort of data. One question is whether a higher resolution may be beneficial especially for tasks where cellular features may be important. Some works have started to appear where several magnification levels are used jointly. Whether mixing magnifications or training magnification specific models is of an advantage is largely unanswered. Similarly, a majority of efforts have focused on using H&E stained slides and ignoring IHC ones. H&E slides are the basis of diagnostic work and are the fastest and cheapest to produce. Meanwhile, IHC slides provide supporting information but are slower and more costly to generate and are not routine except for in very small subset of pathologists' workflow. Further, the technical complexity of IHC (e.g. differences in tissue processing, antigen extraction, unique antibodies for same protein target, unique automation platforms) make the inter-institutional variability much greater than H&E. As a result, it has yet to be proven that foundation models can be useful for IHC-based computational pathology models. Furthermore, it might be possible that the inclusion of IHC slides could be beneficial for H&E based tasks. Future work will be needed to address these questions. Finally, gathering large collections of pathology slides for pretraining is a daunting task within the constraints of single institutions. While, collecting multi-institutional pretraining data might improve the robustness and generalizability of foundation models, there are several important obstacles in the way, and it has yet to be proven beneficial or necessary.

Table 2 | Summary of detection downstream tasks currently included

Origin	Disease	Slides (Positive)	Scanner
MSHS	Breast Cancer	1998 (999)	Philips Ultrafast
MSHS	Oral Cancer	279 (145)	Philips Ultrafast
MSHS	Bladder Cancer	448 (272)	Philips Ultrafast
MSHS	Kidney Cancer	1000 (562)	Philips Ultrafast
MSHS	Thyroid Cancer	710 (390)	Philips Ultrafast
MSHS	DCIS	233 (135)	Philips Ultrafast
MSHS	Prostate Cancer	1000 (547)	Philips Ultrafast
MSHS	Colo-rectal Cancer	413 (257)	Philips Ultrafast
MSHS	IBD	1448 (717)	Philips Ultrafast

MSHS Mount Sinai Health System.

As the development of foundation models in pathology progresses, we will continue providing the community with a leader board of publicly available foundation models as well as external models automatically benchmarked by external users. At the same time we will expand the scope of the tasks included in terms of technical variability and prediction endpoints. We will include data from partner institutions, national and international. We will increase our focus on tasks related to biomarker prediction and treatment response as well as survival analysis tasks. Based on the accumulated evidence, we will update our recommendations on how to train foundation models in computational pathology. As the field develops, future work will also focus on assessing the performance of slide-level foundation models.

Methods

This research study was approved by the respective Institutional Review Boards at the Icahn School of Medicine at Mount Sinai (Protocol 19-00951) and Memorial Sloan Kettering Cancer Center (Protocol 18-013). Informed consent was waived as per the IRB protocols. Participants were not compensated. Sex and/or gender was not considered in the study design as cohorts were generated as random samples of the patient population.

Downstream Tasks

To assess the representation power of pathology foundation models, we collected a series of clinical datasets spanning clinically relevant tasks from three institutions and scanned with a variety of scanners. For analysis, data was extracted always at 20x magnification (0.5 microns per pixel). The tasks are described below and summarized in Tables 2 and 3 for the detection and the biomarker tasks respectively. Additional demographic information for each cohort are provided in Supplementary Table 1, Supplementary Table 2, and Supplementary Table 3.

Disease Detection

MSHS Breast Cancer Detection Cohort. Breast cancer blocks and normal breast blocks were obtained from the pathology LIS. A total of 1998 slides were sampled, with 999 positive and 999 negative. The positive slides were selected from blocks that received the routine biomarker panel for cancer cases (estrogen receptor ER, progesterone receptor PR, HER2, and Ki67), while negative slides were selected from breast cases that did not have an order for the routine panel. Additionally, negative cases were selected if they were not a mastectomy case, did not have a synoptic report associated with the case, and had no mention of cancer or carcinoma in the report.

MSHS Oral Cancers Detection Cohort. Tumor (positive) and normal (negative) block information were extracted from structured synoptic

Table 3 | Summary of downstream tasks currently included for computational biomarker prediction

Origin	Biomarker	Specimen	Slides (Positive)	Scanner
MSHS	IHC ER	Breast Cancer	2000 (1000)	Philips Ultrafast
MSHS	IHC PR	Breast Cancer	1986 (953)	Philips Ultrafast
MSHS	IHC/ FISH HER2	Breast Cancer	2018 (760)	Philips Ultrafast
MSHS	BioMe HRD	Breast	563 (188)	Philips Ultrafast
SUH	NGS *BRAF*	Skin	283 (113)	Nanozoomer S210
SUH	NGS *NRAS*	Skin	283 (94)	Nanozoomer S210
MSHS	NGS *EGFR*	LUAD	294 (103)	Philips Ultrafast
MSKCC	NGS *EGFR*	LUAD	1000 (307)	Aperio AT2
MSKCC	NGS *ALK*	LUAD	999 (144)	Aperio AT2
MSKCC	NGS *STK11*	LUAD	998 (122)	Aperio AT2
MSKCC	NGS *KRAS*	LUAD	998 (325)	Aperio AT2
MSKCC	NGS *TP53*	LUAD	998 (430)	Aperio AT2
MSKCC	ICI Response	NSCLC	454 (86)	Aperio AT2

MSHS Mount Sinai Health System, *SUH* Sahlgrenska University Hospital, *MSKCC* Memorial Sloan Kettering Cancer Center.

reports obtained from the LIS. Synoptic reports for “Lip and Oral Cavity” were included. The positive samples included a variety of cancer diagnoses: squamous cell carcinoma, adenoid cystic carcinoma, mucoepidermoid carcinoma, and others.

MSHS Bladder Cancers Detection Cohort. Tumor (positive) and normal (negative) block information were extracted from structured synoptic reports obtained from the LIS. Synoptic reports for “Cystectomy, Anterior Exenteration” and “Transurethral Resection of Bladder Tumor” were included. The positive samples included a variety of cancer diagnoses: urothelial carcinoma, small cell neuroendocrine carcinoma, adenocarcinoma, squamous cell carcinoma, and others.

MSHS Kidney Cancers Detection Cohort. Tumor (positive) and normal (negative) block information were extracted from structured synoptic reports obtained from the LIS. Synoptic reports for “Nephrectomy” were included. The positive samples included a variety of cancer diagnoses: clear cell renal cell carcinoma, chromophobe renal cell carcinoma, papillary renal cell carcinoma, Xp11 translocation renal cell carcinoma, clear cell sarcoma, and others.

MSHS Thyroid Cancers Detection Cohort. Tumor (positive) and normal (negative) block information were extracted from structured synoptic reports obtained from the LIS. Synoptic reports for “Thyroid Gland” were included. The positive samples included a variety of cancer diagnoses: papillary carcinoma, follicular carcinoma, Hurthle cell carcinoma, and others.

MSHS Prostate Cancer Detection Cohort. Tumor (positive) and normal (negative) block information was extracted from structured synoptic reports obtained from the LIS. Synoptic reports for “Radical Prostatectomy” and “Transurethral Prostatic Resection” were included. The positive samples included acinar and ductal prostate adenocarcinomas.

MSHS Colo-rectal Cancers Detection Cohort. Tumor (positive) and normal (negative) block information was extracted from structured synoptic reports obtained from the LIS. Synoptic reports for “Resection”, “Transanal Disk Excision of Rectal Neoplasms”, “Excisional Biopsy (Polypectomy)”, and “Neuroendocrine Tumor” were included. The positive samples included a variety of cancer diagnoses: adenocarcinoma, signet-ring cell carcinoma, micropapillary carcinoma, and others.

MSHS DCIS Detection Cohort. Tumor (positive) and normal (negative) block information was extracted from structured synoptic reports obtained from the LIS. The synoptic report “DCIS of the Breast” was used for this cohort.

MSHS IBD Detection Cohort. Normal mucosa samples were obtained from patients undergoing screening and routine surveillance lower endoscopy from 2018 to 2022. Inflammatory bowel disease (IBD) cases, including first diagnoses and follow ups, were included. Active IBD samples were scored using the Mount Sinai histologic disease criteria and found to have Histologic Activity Score (HAI) > = 1. A total of 1441 slides were sampled, 717 with active inflammation and 724 with normal mucosa.

MSHS Breast Cancer IHC/FISH Panel. Breast cancer cases with orders for Estrogen Receptor (ER), Progesterone (PR), and HER2 were queried from the LIS. The test results were automatically extracted from the respective pathology report.

MSHS Breast Cancer ER Prediction Cohort. ER IHC orders were included and a total of 2000 slides were sampled, 1000 positive, 1000 negative.

MSHS Breast Cancer PR Prediction Cohort. PR IHC orders were included and a total of 1986 slides were sampled, 953 positive, 1033 negative.

MSHS Breast Cancer HER2 Prediction Cohort. Orders for HER2 IHC and FISH were included and a total of 2018 slides were sampled, 760 positive, 1258 negative.

MSHS Breast HRD Prediction Cohort. Mount Sinai BioMe is a whole-exome sequencing cohort of 30k individuals, where carriers of pathogenic and protein-truncating variants affecting Homologous Repair Deficiency (HRD) genes (i.e., *BRCA1*, *BRCA2*, *BRIP1*, *PALB2*, *RAD51*, *RAD51C*, *RAD51D*, *ATM*, *ATR*, *CHEK1*, and *CHEK2*), where included as positives. A subset of the BioMe dataset of patients with available breast pathology slides were included. Slides containing solely normal breast tissue and slides with breast cancer were both included.

SUH Melanoma Somatic Mutation Panel. A total of 283 melanoma cases were retrospectively collected from the archives of the Departments of Pathology at Sahlgrenska University Hospital (SUH), Södra Älvsborg hospital and Norra Älvsborgs hospital in the Region Västra Götaland, Sweden. *BRAF* and *NRAS* mutation status was verified by NGS or IHC. The dataset included both primary and metastatic samples.

SUH *BRAF* Mutation Prediction in Melanoma Cohort. Of the 283 samples, 113 had verified V600E/K *BRAF* mutations and were considered positive. The rest had no clinically relevant *BRAF* mutations and were considered negative.

SUH *NRAS* Mutation Prediction in Melanoma Cohort. Of the 283 samples, 94 were detected to have an *NRAS* mutation and were considered positive.

MSHS *EGFR* mutation detection in Lung Adenocarcinoma. Lung adenocarcinoma (LUAD) patients that underwent next generation sequencing (NGS) profiling for their cancer were identified. A total of 294 slides were obtained from MSHS’s clinical slide database, 103 positive and 191 negative. Mutations outside of the *EGFR* kinase domain (exons 18-24) are not considered oncogenic and are considered negative in this analysis.

MSKCC Lung Adenocarcinoma Somatic Mutation Panel. LUAD patients at Memorial Sloan Kettering Cancer Center with respective molecular analysis from the MSK-IMPACT assay[44,45] and corresponding digitized slides where identified. MSK-IMPACT is an NGS assay that can detect variants in up to 505 unique cancer genes, including *EGFR*, *TP53*, *KRAS*, *STK11*, and *ALK*.

MSKCC *EGFR* Mutation Prediction in LUAD. LUAD samples with an oncogenic *EGFR* mutation detected by MSK-IMPACT were included. Mutations outside of the *EGFR* kinase domain (exons 18-24) are not considered oncogenic and are considered negative in this analysis. This is a sample of the dataset described in Campanella et al.[46] from which 1000 slides were sampled at random, 307 positive and 693 negative.

MSKCC *TP53* Mutation Prediction in LUAD. MSK-IMPACT derived *TP53* mutational status. A total of 998 slides were sampled, 430 positive and 568 negative.

MSKCC *KRAS* Mutation Prediction in LUAD. MSK-IMPACT derived *KRAS* mutational status. A total of 998 slides were sampled, 325 positive and 673 negative.

MSKCC *STK11* Mutation Prediction in LUAD. MSK-IMPACT derived *STK11* mutational status. A total of 998 slides were sampled, 122 positive and 876 negative.

MSKCC *ALK* Mutation Prediction in LUAD. MSK-IMPACT derived *ALK* mutational status. A total of 999 slides were sampled, 144 positive and 855 negative.

MSKCC ICI Therapy Response Prediction in NSCLC. Non-small cell lung cancer (NSCLC) patients who received PD-L1 blockade-based immune checkpoint inhibitor (ICI) therapy between 2013 and 2019 at MSKCC were considered. Cytology specimens were excluded. The objective overall response was determined by RECIST and performed by a blinded thoracic radiologist. A total of 454 slides were obtained, 86 positive and 368 negative.

Downstream Task Training

In the SSL literature, the performance of downstream tasks is frequently assessed by training a linear classifier (linear probing) on top of features extracted by a frozen encoder, or via zero-shot approaches such as k-NN. For pathology slides, there is no direct way to translate these approaches without having tile-level annotations. Instead, it is common practice to train a slide-level aggregator. For this purpose we chose the popular Gated MIL Attention (GMA) model[47] with a linear classifier on top. Since GMA does not consider the spatial distribution of tiles over the slide in its prediction, it is a simple method to test the expressiveness of the feature space generated by the SSL pretraining. In fact GMA is widely used to assess the performance of pathology foundation models[23,25,26]. Further, despite being a simple strategy, it is highly performant even compared to more recent aggregators in computational pathology[43].

For each slide, tissue tiles were extracted at 20x magnification (0.5 microns per pixel, MPP) and embedded into a feature representation using a specific foundation model. This magnification is appropriate for all foundation models considered. Each slide is then converted to a 2D matrix where every row corresponds to a tile in the slide and the columns contain the features. The vectorized slide is the input to the GMA model, which combines the tile representations into a slide-level representation, which is then linearly projected to class scores.

To estimate generalization performance, we employed a Monte Carlo Cross-Validation (MCCV) strategy. For each MCCV split, 80% of the samples were assigned to the training set and the remaining 20% were assigned to the validation set. For each benchmark task, the 20 MCCV folds were randomly sampled and kept fixed for all experiments. Each MCCV split was run twice to assess stochastic fluctuations during training and the results were averaged across the two replicas. All models were trained using a single GPU for 50 epochs using the AdamW[48] optimizer. A cosine decay with warm-up schedule was used for the learning rate with a peak learning rate of 0.0001. The exact parameters used for training can be found in the GitHub repository and in Supplementary Tables 4 and 5. For each task and foundation model, the distribution of validation AUCs across the 20 MCCVs are used to assess the trained model performance.

Foundation Models

In this work we focus on benchmarking publicly available vision foundation models trained on large pathology corpora. More specifically, we consider only tile-level vision encoder models. Pre-trained aggregation models are not considered and vision-language models are also not included. Given that the number of publicly available foundation models has been increasing steadily, it is beyond the scope of this manuscript to exhaustively benchmark all models available. We strived to chose a wide selection of models of different sizes from academia and private companies using public and private pretraining datasets, including CTransPath[18], UNI[23], Virchow[25], Prov-GigaPath[30], Virchow2[33], h-optimus-0[34], and Phikon-v2[35]. We also included a truncated ResNet50 (tRes50) pretrained on ImageNet as a baseline due to its popularity in the computational pathology community. We must note that for Virchow and Virchow2, since they were trained on slides from MSKCC, we can't ensure that there is no overlap between their pretraining cohort and the clinical tasks based on MSKCC data. In addition, only the tile-level encoder of Prov-GigaPath was considered in this work. For each foundation model, we followed the embedding instructions provided by the authors in each respective repository.

For comparison, we further include two in-house trained foundation models: a ViT-small (21.7M parameters) and a ViT-base (85.8M parameters) trained with DINO[28]. These models were pretrained on a clinical dataset compiled at MSHS during normal hospital operation. The pretraining dataset consisted of 423,563 H&E stained slides from 88,035 cases and 76,794 patients. These include slides from 42 organs across all pathology specialities. We ensured that no overlap exists between this pretraining dataset and the clinical benchmarking dataset. All slides were scanned on a Philips Ultrafast scanner at 40x magnification (0.25 MPP), de-identified and converted to tiff format. The total storage required for the raw tiff files was around 600TB. As a preprocessing step, tissue tiles were extracted from each slide at 0.5 MPP resolution, yielding approximately 3.2 billion tiles. The ViT-small (SP22M) was trained on 12 Nvidia A100 40GB GPUs with a batch size of 90 per GPU for 17 days and 16 hours. The ViT-base (SP85M) was trained on 8 Nvidia H100 80GB GPUs with a batch size of 100 per GPU for 26 days and 11 hours. Both models were trained on approximately 1.6 billion tiles. The models are publicly available on HuggingFace: SP22M, SP85M.

We can observe that older foundation models are trained with variants of contrastive learning. After the introduction of DINO, and later DINOv2, recent foundation models have used the latter as go-to pretraining algorithm. While evidence emerged that DINO tends to outperform contrastive learning and masked image modeling approaches for pathology pretraining[26,36], there is no direct comparison of DINO and DINOv2. To summarize, current pathology foundation models are trained in a very homogeneous manner, leveraging for the most part DINOv2 and a small number of ViT architectures. The main differences between the various efforts mainly lay in the composition of the pretraining dataset.

Automated External Benchmarking

To facilitate the benchmarking of external models, we have developed an automated pipeline that leverages the Azure cloud infrastructure for interfacing with users and on-premise computing to minimize costs. Interested users will need to fill out a Microsoft Form to express their interest in benchmarking their model. This form collects essential details such as the user's email address and model information. The form submission triggers a Power Automate process in the back-end which generates a OneDrive folder accessible to the external user. The process also generates an email containing the OneDrive link and instructions, which is sent to the external user. The instructions prompt the user to upload two files to OneDrive: i) a docker (or singularity) container that includes the model weights, and ii) an inference script that returns the model's output. We provide a template of the inference script that users can customize and detailed instructions on GitHub. Currently, files up to 250GB can be uploaded. Once uploaded, these files will trigger the benchmarking pipeline. Files are copied to the local cluster and are used to generate tile embeddings which are stored as binary files. These are then used to train a GMA aggregator as described previously. Results of the benchmark are then returned to the user via email. If the user opted to release to the leaderboard, the results will be posted on our GitHub page. An overview of the workflow is presented in Supplementary Fig. 1.

Reporting summary

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

Data availability

Digital pathology benchmark data will not be made available due to legal, privacy, and data contamination considerations. If benchmark data would be made available, it would likely be scraped for pretraining foundation models, negating the benefits of an independent benchmark. We provide a mechanism to evaluate user provided foundation models on our benchmarking tasks. Instructions can be found in the GitHub repository. Source data for each figure are provided as a Source Data file. Source data are provided with this paper.

Code availability

Code used for pretraining SP22M and SP85M was taken from the official DINO repository. The code associated with this work is available in this GitHub repository with a MIT License[49].

References

1. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).
2. Dosovitskiy, A. et al. An image is worth 16 × 16 words: Transformers for image recognition at scale. Preprint at *arXiv:2010.11929* (2020).
3. Gao, Z. et al. Instance-based vision transformer for subtyping of papillary renal cell carcinoma in histopathological image. In *Proc. Part VIII 24, Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference* 299–308 (Strasbourg, 2021).
4. Akinyelu, A. A., Zaccagna, F., Grist, J. T., Castelli, M. & Rundo, L. Brain tumor diagnosis using machine learning, convolutional neural networks, capsule neural networks and vision transformers, applied to mri: a survey. *J. Imaging* **8**, 205 (2022).
5. Kumar, N. et al. Convolutional neural networks for prostate cancer recurrence prediction. *Med. Imaging 2017: Digi. Pathol.* **10140**, 106–117 (2017).
6. Coudray, N. et al. Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning. *Nat. Med.* **24**, 1559–1567 (2018).
7. Verma, R. et al. Sexually dimorphic computational histopathological signatures prognostic of overall survival in high-grade gliomas via deep learning. *Sci. Adv.* **10**, eadi0302 (2024).
8. Shen, Y. et al. Explainable survival analysis with convolution-involved vision transformer. *Proc. AAAI Conf. Artif. Intell.* **36**, 2207–2215 (2022).
9. Mobadersany, P. et al. Predicting cancer outcomes from histology and genomics using convolutional networks. *Proc. Natl Acad. Sci.* **115**, E2970–E2979 (2018).
10. Chen, X., Xie, S. & He, K. An empirical study of training self-supervised vision transformers. In *Proc. IEEE/CVF International Conference on Computer Vision* 9640–9649 (IEEE, 2021).
11. Khan, A. et al. A survey of the self supervised learning mechanisms for vision transformers. Preprint at *arXiv:2408.17059* (2024).
12. Lu, M. Y. et al. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nat. Biomed. Eng.* **5**, 555–570 (2021).
13. Ilse, M., Tomczak, J. & Welling, M. Attention-based deep multiple instance learning. *Int. Confer. Mach. Learn.* **80**, 2127–2136 (2018).
14. Shao, Z. et al. Transmil: Transformer based correlated multiple instance learning for whole slide image classification. *Adv. neural Inf. Process. Syst.* **34**, 2136–2147 (2021).
15. Mormont, R., Geurts, P. & Marée, R. Comparison of deep transfer learning strategies for digital pathology. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition Workshops* 2262–2271 (IEEE, 2018).
16. Tabibu, S., Vinod, P. & Jawahar, C. Pan-renal cell carcinoma classification and survival prediction from histopathology images using deep learning. *Sci. Rep.* **9**, 10509 (2019).
17. Carmichael, I. et al. Incorporating intratumoral heterogeneity into weakly-supervised deep learning models via variance pooling. In *Proc. International Conference on Medical Image Computing and Computer-Assisted Intervention* 387–397 (2022).
18. Wang, X. et al. Transformer-based unsupervised contrastive learning for histopathological image classification. *Med. Image Anal.* **81**, 102559. https://linkinghub.elsevier.com/retrieve/pii/S1361841522002043 (2022).
19. Liu, Z. et al. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proc. IEEE/CVF International Conference on Computer Vision* 10012–10022 (IEEE, 2021).
20. Weinstein, J. N. et al. The cancer genome atlas pan-cancer analysis project. *Nat. Genet.* **45**, 1113–1120 (2013).
21. Filiot, A. et al. Scaling self-supervised learning for histopathology with masked image modeling. *medRxiv* 2023–07 (2023).
22. Zhou, J. et al. ibot: Image bert pre-training with online tokenizer. Preprint at *arXiv:2111.07832* (2021).
23. Chen, R. J. et al. Towards a general-purpose foundation model for computational pathology. *Nat. Med.* **30**, 850–862 (2024).
24. Oquab, M. et al. Dinov2: Learning robust visual features without supervision. Preprint at *arXiv:2304.07193* (2023).
25. Vorontsov, E. et al. A foundation model for clinical-grade computational pathology and rare cancers detection. *Nat. Med.* **30**, 2924–2935 (2024).
26. Campanella, G. et al. Computational pathology at health system scale–self-supervised foundation models from three billion images. Preprint at *Xiv:2310.07033* (2023).
27. He, K. et al. Masked autoencoders are scalable vision learners. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition* 16000–16009 (IEEE, 2022).
28. Caron, M. et al. Emerging properties in self-supervised vision transformers. In *Proc. IEEE/CVF International Conference on Computer Vision* 9650–9660 (IEEE, 2021).
29. Dippel, J. et al. Rudolfv: a foundation model by pathologists for pathologists. Preprint at *arXiv:2401.04079* (2024).
30. Xu, H. et al. A whole-slide foundation model for digital pathology from real-world data. *Nature* **630**, 181–188 (2024).
31. Ding, J. et al. Longnet: Scaling transformers to 1,000,000,000 tokens. Preprint at *arXiv:2307.02486* (2023).

32. Network, C. G. A. R. et al. Comprehensive molecular profiling of lung adenocarcinoma. *Nature* **511**, 543 (2014).
33. Zimmermann, E. et al. Virchow 2: Scaling self-supervised mixed magnification models in pathology. Preprint at *arXiv:2408.00738* https://arxiv.org/abs/2408.00738 (2024).
34. Saillard, C. et al. H-optimus-0. https://github.com/bioptimus/releases/tree/main/models/h-optimus/v0 (2024).
35. Filiot, A., Jacob, P., Mac Kain, A. & Saillard, C. Phikon-v2, a large and public feature extractor for biomarker prediction. Preprint at *arXiv:2409.09173* (2024).
36. Kang, M., Song, H., Park, S., Yoo, D. & Pereira, S. Benchmarking self-supervised learning on diverse pathology datasets. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition* 3344–3354 (IEEE, 2023).
37. Campanella, G. et al. A clinical benchmark of public self-supervised pathology foundation models. Preprint at *arXiv:2407.06508* (2024).
38. Goyal, P. et al. Vision models are more robust and fair when pre-trained on uncurated images without supervision. Preprint at *arXiv:2202.08360*. https://arxiv.org/abs/2202.08360 (2022).
39. Bhattacharyya, P., Huang, C. & Czarnecki, K. Ssl-lanes: Self-supervised learning for motion forecasting in autonomous driving. *Confer. Robot Learn*. 1793–1805 (2023).
40. Radford, A. et al. Learning transferable visual models from natural language supervision. *Int. Confer. Mach. Learning* 8748–8763 (2021).
41. Neidlinger, P. et al. Benchmarking foundation models as feature extractors for weakly-supervised computational pathology. Preprint at *arXiv:2408.15823* (2024).
42. Xie, C. et al. Computational biomarker predicts lung ici response via deep learning-driven hierarchical spatial modelling from h&e. https://www.researchsquare.com/article/rs-1251762/v1 (2022).
43. Chen, S. et al. Benchmarking embedding aggregation methods in computational pathology: A clinical data perspective. In *Proceedings of the MICCAI Workshop on Computational Pathology*, (eds Ciompi, F. et al.) Vol. 254, 38–50 (PMLR, 2024). https://proceedings.mlr.press/v254/chen24a.html.
44. Cheng, D. T. et al. Comprehensive detection of germline variants by msk-impact, a clinical diagnostic platform for solid tumor molecular oncology and concurrent cancer predisposition testing. *BMC Med. Genom*. **10**, 1–9 (2017).
45. Zehir, A. et al. Mutational landscape of metastatic cancer revealed from prospective clinical sequencing of 10,000 patients. *Nat. Med.* **23**, 703–713 (2017).
46. Campanella, G. et al. H&e-based computational biomarker enables universal egfr screening for lung adenocarcinoma. Preprint at *arXiv:2206.10573* (2022).
47. Ilse, M., Tomczak, J. & Welling, M. Attention-based deep multiple instance learning. *Int Confer. Mach. Learning*. 2127–2136 https://arxiv.org/abs/1802.04712 (2018).
48. Loshchilov, I. & Hutter, F. Decoupled weight decay regularization. Preprint at *arXiv:1711.05101* https://arxiv.org/abs/1711.05101 (2017).
49. Campanella, G. sinai-computational-pathology/SSL_tile_benchmarks: First release. https://doi.org/10.5281/zenodo.15110130.

Acknowledgements

We acknowledge Sahar Alawieh for curating the SUH cohorts. This work is supported in part through the use of the AI-Ready Mount Sinai (AIR.MS) research platform and the expertise provided by the team at the Hasso Plattner Institute for Digital Health at Mount Sinai (HPI.MS). We also utilized computational resources and expertise from Scientific Computing and Data at the Icahn School of Medicine, supported by the Clinical and Translational Science Awards (CTSA) grant UL1TR004419. Additionally, research funding was provided in part by a grant from from the National Library of Medicine (NLM) (R01 LM013766, G.C.), by a Cancer Center Support Grant from the NIH/NCI (P30CA008748), and by a grant from the Warren Alpert Foundation (C.V.) through the Warren Alpert Center for Digital and Computational Pathology at Memorial Sloan Kettering Cancer Center.

Author contributions

G.C., C.V. conceived the study. G.C., S.C., R.V. performed the experiments and analyzed the results. R.K., J.Z., A.S., B.V. curated the MSHS detection cohorts. R.K., J.Z. curated the MSHS breast biomarker cohorts. M.C., J.H. curated the MSHS NGS cohort. A.E., K.H. curated the MSHS HRD cohort. I.S., N.N. curated the SUH NGS cohorts. C.V. provided the MSKCC NGS cohorts. C.V., A.J.S. provided the MSKCC IO cohort. S.M. facilitated the collaborative research agreements between institutions. G.C., M.S. developed the automatic benchmarking pipeline.

Competing interests

C.V. reports intellectual property rights and equity interest in Paige.AI, Inc. A.J.S. reports consulting/advising role to J&J, Bayer, KSQ therapeutics, BMS, Merck, Astrazeneca, Synthekine, cTRL therapeutics, Regeneron, Enara Bio, Perceptive Advisors, Oppenheimer and Co, Umoja Biopharma, Legend Biotech, Iovance Biotherapeutics, Obsidian Therapeutics, Prelude Therapeutics, Immunocore, Lyell Immunopharma, Amgen and Heat Biologics. A.J.S. receives research funding from GSK (Inst), Obsidian (Inst), Lilly (Inst), PACT pharma (Inst), Iovance Biotherapeutics (Inst), Achilles therapeutics (Inst), Merck (Inst), Synthekine (Inst), BMS (Inst), Harpoon Therapeutics (Inst), AffiniT therapeutics (Inst), Legend Therapeutics (Inst), Synthekine (Inst) and Amgen (Inst). G.C., S.C., M.S., R.V., S.M., J.Z., A.S., M.C., B.V., A.E., I.S., N.N., K.H., R.K., J.H. declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-025-58796-1.

Correspondence and requests for materials should be addressed to Gabriele Campanella or Chad Vanderbilt.

Peer review information *Nature Communications* thanks Frauke Wilm, and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. A peer review file is available.

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

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

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

© The Author(s) 2025