

Benchmarking foundation models as feature extractors for weakly-supervised computational pathology

Peter Neidlinger (1,*), Omar S. M. El Nahhas (1, 2, *), Hannah Sophie Muti (1, 3, 4), Tim Lenz (1), Michael Hoffmeister (5), Hermann Brenner (5, 6, 7), Marko van Treeck (1), Rupert Langer (8), Bastian Dislich (9), Hans Michael Behrens (10), Christoph Röcken (10), Sebastian Foersch (11), Daniel Truhn (2, 12), Antonio Marra (13), Oliver Lester Saldanha (1), Jakob Nikolas Kather (1, 2, 14, 15, 16, +)

* equal contribution

+ Correspondence to jakob-nikolas.kather@alumni.dkfz.de

1. Else Kroener Fresenius Center for Digital Health, Technical University Dresden, Dresden, Germany
2. StratifAI GmbH, Dresden, Germany
3. Department for Visceral, Thoracic and Vascular Surgery, University Hospital and Faculty of Medicine Carl Gustav Carus, Technische Universität Dresden, Dresden, Germany
4. National Center for Tumor Diseases Dresden (NCT/UCC), a partnership between DKFZ, Faculty of Medicine and University Hospital Carl Gustav Carus, TUD Dresden University of Technology, and Helmholtz-Zentrum Dresden - Rossendorf (HZDR), Dresden, Germany
5. Division of Clinical Epidemiology and Aging Research, German Cancer Research Center (DKFZ), Heidelberg, Germany
6. Division of Preventive Oncology, German Cancer Research Center (DKFZ) and National Center for Tumor Diseases (NCT), Heidelberg, Germany
7. German Cancer Consortium (DKTK), German Cancer Research Center (DKFZ), Heidelberg, Germany
8. Institute of Pathology and Molecular Pathology, Kepler University Hospital, Johannes Kepler University Linz, Linz, Austria
9. Institute of Tissue Medicine and Pathology, University of Bern, Bern, Switzerland
10. Department of Pathology, University Hospital Schleswig-Holstein, Kiel, Germany
11. Institute of Pathology, University Medical Center Mainz, Mainz, Germany
12. Department of Diagnostic and Interventional Radiology, University Hospital Aachen
13. Division of New Drugs and Early Drug Development, European Institute of Oncology IRCCS, Milan, Italy
14. Medical Department 1, University Hospital and Faculty of Medicine Carl Gustav Carus, Technische Universität Dresden, Dresden, Germany
15. National Center for Tumor Diseases Dresden (NCT/UCC), a partnership between DKFZ, Faculty of Medicine and University Hospital Carl Gustav Carus, TUD Dresden University of Technology, and Helmholtz-Zentrum Dresden - Rossendorf (HZDR), Dresden, Germany
16. Medical Oncology, National Center for Tumor Diseases (NCT), University Hospital Heidelberg, Heidelberg, Germany

Abstract

Advancements in artificial intelligence have driven the development of numerous pathology foundation models capable of extracting clinically relevant information. However, there is currently limited literature independently evaluating these foundation models on truly external cohorts and clinically relevant tasks to uncover adjustments for future improvements. In this study, we benchmarked ten histopathology foundation models on 13 patient cohorts with 6,791 patients and 9,493 slides from lung, colorectal, gastric, and breast cancers. The models were evaluated on weakly-supervised tasks related to biomarkers, morphological properties, and prognostic outcomes. We show that a vision-language foundation model, CONCH, yielded the highest performance in 42% of tasks when compared to vision-only foundation models. The experiments reveal that foundation models trained on distinct cohorts learn complementary features to predict the same label, and can be fused to outperform the current state of the art. Creating an ensemble of complementary foundation models outperformed CONCH in 66% of tasks. Moreover, our findings suggest that data diversity outweighs data volume for foundation models. Our work highlights actionable adjustments to improve pathology foundation models.

1. Introduction

Artificial intelligence (AI) has revolutionized digital pathology (DP) by enabling biomarker prediction from cancer tissues using high-resolution whole slide images (WSIs)^{1–6}. Moreover, these algorithms can substantially enhance diagnostic accuracy, efficiency and consistency, significantly reducing the subjectivity associated with human interpretation^{7,8}. In particular, deep learning (DL) can perform tasks such as disease grading, cancer subclassification or prognostic prediction^{9–11}.

Recently, foundation models, which are trained on large-scale datasets, have been introduced to DP^{12,13}. These models use self-supervised Learning (SSL) techniques to learn meaningful representations of histology tissue, which are crucial for clinical pathology tasks. SSL techniques such as contrastive learning^{14,15} and masked image modeling (MIM)¹⁶ have shown improved performance, robustness and higher transferability compared to fully supervised learning. Another advantage lies in its ability to learn from vast amounts of unlabeled data, thereby significantly reducing the need for manual annotation¹⁷. The practical application of foundation models involves WSI tessellation into small, non-overlapping patches, after which image feature extraction is performed. These extracted features serve as inputs for training classification or regression models, such as ViTs¹⁸, tailored for specific tasks, like mutation prediction, survival analysis, disease grading, or cancer classification¹⁹.

The potential of foundation models has led to the development of many specialized models in a short amount of time. We refer to these models as “downstream” models because they are trained on top of a foundation model feature extractor which remains constant. However, many of these foundation models have only been benchmarked on a limited set of tasks and public datasets, with potential issues of data leakage and cherrypicking. The majority of these models have not systematically been tested on a variety of clinically relevant tasks. The limited availability and variable quality of public pathology data can hinder the performance of these models when applied to real-world clinical scenarios²⁰. Recent efforts have demonstrated the potential of large-scale foundation models in computational pathology. Unlike earlier models

that relied heavily on datasets like The Cancer Genome Atlas (TCGA), contemporary foundation models are now trained on much larger proprietary cohorts like Mass-100K (100k WSIs)²¹, Providence (171k WSIs)²² and Memorial Sloan Kettering Cancer Center (1488k WSIs)²³. These models have shown impressive capabilities in tasks such as mutation prediction, enhancing the scope of computational pathology beyond traditional applications.

In the present work, we introduce the most extensive benchmarking effort to date for histopathology foundation models. By including multiple proprietary cohorts from multiple countries, which were never part of any foundation model training, we effectively mitigate the risk of data leakage from pretraining datasets. Our benchmarking includes 10 foundation models and 31 clinically relevant evaluation tasks, 19 of which are the prediction of cancer biomarkers, using a total of 6,791 patients and 9,493 slides. This comprehensive evaluation bridges a notable gap in DP literature and will serve as an important reference point for the DP community helping to select the right foundation model for a specific DP task.

2 Results

2.1 Multimodal vision-language model outperforms vision-only models on imaging tasks in pathology

We benchmarked the performance of the foundation models, trained as vision-language or vision-only, on weakly-supervised prediction tasks related to morphology, biomarkers, and prognostication. Across the three task domains, the vision-language model CONCH consistently outperformed vision-only models as measured by the area under the receiver operating characteristic (AUROC) averaged over the tasks within the task domain. For the five morphology-related tasks, CONCH yielded the highest AUROC of 0.77, followed by Virchow and UNI with AUROCs of 0.75 and 0.74, respectively (Figure 2C). Across the 19 biomarker-related tasks, CONCH achieved the highest AUROC of 0.73, followed closely by UNI and Prov-GigaPath with an AUROC of 0.72 (Figure 2D). Finally, in the seven prognostic-related tasks, CONCH yielded the highest AUROC of 0.62, followed by H-optimus-0 and Virchow with AUROCs of 0.61 and 0.59, respectively (Figure 2E). Averaged across all 31 tasks, CONCH had the highest AUROC of 0.71, followed by UNI, H-optimus-0 and Prov-GigaPath with AUROCs of 0.69. Subsequent rankings included Virchow, Hibou and CTransPath with AUROCs of 0.67, and Kaiko and Phikon at 0.66. Moreover, CONCH achieved the highest average area under the precision recall characteristic (AUPRC), accuracy and F1 scores (Figure S1) and the highest average AUROC in each cancer type (Figure S2).

Among the 29 binary classification tasks, only a subset of foundation models were statistically significantly different from CONCH while yielding higher AUROCs. Specifically, this occurred for H-optimus-0 in two tasks, and in one task for UNI, Kaiko, Prov-GigaPath, Hibou and Virchow. Vice versa, CONCH yielded higher AUROCs with statistically significantly differences across a substantially larger amount of tasks: Phikon (12), Hibou (11), H-optimus-0 (10), CTransPath (9), Virchow (8), UNI (6), and Prov-GigaPath (3). Notably, CTransPath and Phikon were never significantly better than CONCH in any of the tasks ($p < 0.05$ across all comparisons) (Figure S3B).

Together, these data show that CONCH, the only vision-language model in this study, outperforms vision-only models on 13 of 31 tasks, achieving the highest performance metrics in the three highlighted domains of morphology, biomarkers and prognostication-based prediction tasks.

Task difficulty impacts pathology foundation model performance

2.2

Next, we categorized the 31 tasks into two groups to assess model effectiveness on tasks with more consistent weakly-supervised deep learning classification. A task was defined as “high-performance”, if at least one model achieved a mean AUROC of over 0.7 and a standard deviation (SD) below 0.05. The high-performance tasks identified included: NSCLC subtyping, EGFR, STK11, and TP53 mutations in NSCLC; Lauren and MSI status in Bern; Lauren, EBV status, and MSI status in Kiel; ESR1 and PGR expression in CPTAC-BRCA; MSI status and BRAF mutations in CPTAC-CRC; and CRC sidedness, MSI status, and BRAF mutations in DACHS.

The vision-language model CONCH excelled in five of 16 high-performance tasks, achieving an AUROC above 0.7 on all tasks. UNI and Prov-GigaPath reached this AUROC threshold in 15 tasks. Overall, CONCH, UNI, and H-optimus-0 yielded a 0.79 AUROC across these tasks, while Prov-GigaPath achieved 0.78 (**Figure S4B**). A larger disparity was observed among the 15 low-performance tasks. CONCH performed best on eight of 15 tasks and achieved an AUROC above 0.6 on twelve tasks, whereas H-optimus-0 reached this AUROC threshold on seven tasks, and Kaiko on two tasks. CONCH averaged 0.63 AUROC across the 15 low-performance tasks, while H-optimus-0, Prov-GigaPath, and UNI tied at a 0.59 AUROC (**Figure S4C**).

These findings indicate that the CONCH vision-language model consistently outperforms vision-only models in low-performance tasks with a more pronounced advantage, whereas performance differences are minimal in high-performance tasks. This highlights the robustness and efficacy of the vision-language model in addressing complex prediction scenarios (**Figure 2F, S4**).

Pathology foundation model data diversity outweighs data quantity in downstream classification performance

Next, we analyzed how pretraining dataset attributes affect foundation model performance. Smaller models (up to 100k WSIs) showed strong correlation between WSI count and AUROC (**Figure 3A**). This correlation weakened for larger models like Virchow and H-optimus-0, suggesting diminishing returns beyond a certain dataset size. For biomarker prediction, patient and tissue type diversity proved more crucial than WSI quantity. Diversity in anatomic tissue sites (**Figure S5A**) is reflected in various metrics (**Table S1**). A moderate correlation ($r = 0.41$, $p = 0.035$) exists between tissue-specific WSI count in pretraining and performance on related downstream tasks (**Figure 3B, S5B**).

Moreover, we evaluated the impact of constrained dataset volume on model performance. Models were trained on randomly selected cohorts of 299, 150, or 75 patients and tested on full external cohorts, using 29 of 31 tasks (excluding Lauren classification in STAD Bern and

Kiel due to insufficient data). With 299 patients, CONCH outperformed other models, leading in 13 tasks, followed by Prov-GigaPath (5 tasks), and UNI, H-optimus-0, and Virchow (3 each). With 150 patients, H-optimus-0 led in 8 tasks, while UNI and CONCH led in 6 each. Performance differences between 75 and 150 patients were minimal, with CONCH leading in 6 tasks (Figure 3A, 3B, S6). Overall, larger cohorts yielded higher AUROCs across all tasks (Figure 3C, 3D). However, the impact of training size varied substantially across tasks. For example, in MSI status prediction in CPTAC-CRC, the AUROC decreased from 0.85 with 299 patients (including 44 MSI-high cases) to 0.70 with 75 patients (including 13 MSI-high cases). Similarly, EBV status prediction in Kiel saw a sharp decline from 0.85 with 299 patients (24 EBV-positive cases) to 0.53 with 75 patients (five EBV-positive cases). Conversely, *BRAF* mutation prediction in CPTAC-CRC showed a slight AUROC increase from 0.66 with 299 patients (41 *BRAF*-mutated cases) to 0.67 with 75 patients (14 *BRAF*-mutated cases) (Figure 3C, S7).

Together, these results indicate that the patient count and anatomic site diversity prove more crucial than data quantity beyond 100k WSIs, particularly for biomarker prediction. The impact of finetuning dataset size varies significantly across tasks, highlighting the need for adequate training data in specific scenarios. All models show similar performance declines with reduced training sizes, underlining the weakness of current pathology foundation models in scarce data scenarios.

Pathology foundation models learn different tissue morphologies

To quantitatively measure prediction similarity across models, we calculated Cohen's kappa²⁴ for tasks where at least one model achieved a mean AUROC above 0.7, excluding BRAF in CPTAC due to universally low accuracy (Figure S1). For each task, labels were assigned using a majority vote across the cross-validation folds. Among the top 4 models, the combination of CONCH and UNI had the lowest kappa score (0.47), making them the ideal candidates for Ensemble approaches, as they provide complementary predictive power (Figure 4C). Notably, lower-performing models such as Hibou and Kaiko exhibited the least consensus, while top performers like Prov-GigaPath, CONCH, UNI, and Bioptimus showed higher agreement, reflecting their stronger predictive capabilities (Figure S8). Within individual model folds, CONCH achieved the highest average kappa (0.48), followed by H-optimus-0 (0.46) and UNI (0.31), with Hibou (0.17) and Phikon (0.16) ranking lowest, consistent with their AUROC performance (Figure 4D).

To elucidate the reasons behind the observed performance differences among the downstream models trained on top of the different foundation models, we investigated whether the models focus on different morphological properties for their predictions. We utilized attention heatmaps to compare model behavior when the models 1) consistently predicted the label correctly and 2) were in disagreement regarding the predicted label. In cases where all models were in agreement on the correct prediction, the validity of the classification would be supported by their focus on relevant tissue regions for diagnosis. For example, in the prediction of MSI status, models predominantly highlighted tumor regions, as expected. However, models such as UNI, Hibou, Virchow and Kaiko occasionally highlighted pen marks, which is an undesired behavior that suggests predictions are being made through some form of pattern association rather than understanding the underlying biology (Figure 4A, S10B). Models such as CONCH and Virchow focused on multiple small tissue areas, whereas Prov-

GigaPath appears less selective in its attention (**Figure 4A**). In NSCLC subtyping, models generally performed well, focussing mainly on tumor regions and ignoring healthy lung parenchyma (**Figure S9B**). In *ESR1* overexpression prediction, Prov-GigaPath and Kaiko highlighted the majority of the WSI area, whereas CONCH and Virchow focused on a few small tissue areas (**Figure S9C**). In contrast, when analyzing slides where models made inconsistent predictions, we found instances of model disagreement that led to errors. For instance, in the task of DACHS CRC sidedness, Virchow erroneously focused on pen marks (**Figure S10B**). However, no consistent pattern of errors emerged across the models to fully explain these discrepancies.

Together, these data indicate that foundation models vary in their focus on tissue regions and the morphological features they prioritize, which impacts their predictive performance. The differences in attention across models suggest that combining models with complementary strengths, as demonstrated by the lower Cohen's kappa scores between certain models, could enhance overall predictive accuracy in ensemble approaches.

Ensemble of pathology foundation models improve performance

2.5 Lastly, we tested the hypothesis that creating an ensemble of pathology foundation models improves prediction performance. We utilized two approaches for ensembling models, taking the average of the various downstream models' prediction scores trained on different foundation model backbones, and concatenating feature vectors from different foundation model backbones to create a single downstream model.

Experiments show that ensembling by taking the average of the models' prediction scores yielded a superior AUROC compared to either model used in isolation. The combination of the four top-performing models led to the highest improvement, achieving a mean AUROC 1.3% higher than CONCH (**Figure S11**), the leading individual model (**Figure 1B**). Therefore, these data show that ensembling the prediction scores of multiple high-performing models enhances performance on certain tasks beyond the capabilities of the best individual model.

Combining feature vectors from the top four performing models (CONCH, UNI, Prov-GigaPath, and H-optimus-0) resulted in a 3840-dimensional vector used for the downstream model, which achieved an average AUROC of 0.67, comparable to the individual mean AUROCs of UNI, Prov-GigaPath and H-optimus-0 (**Figure S11**). The concatenation of CONCH and Prov-GigaPath, resulting in a 1280-dimensional vector, produced a minimally higher average AUROC than CONCH alone (0.695 vs. 0.693) (**Figure 1B**). To quantify improvements, we conducted two-sided DeLong's tests comparing AUROC scores of CONCH with ensembles and other single-model baselines. For each model, we averaged prediction scores across five folds, and across up to 20 folds for ensembles. Bagging the five folds of the same foundation model increased AUROC scores, while integrating different models via stacking or concatenation yielded more pronounced improvements (**Figure S3A**). The CONCH and Prov-GigaPath ensemble showed statistically significant differences in performance with higher AUROCs than CONCH in eight of 29 tasks ($p < 0.05$), whereas the CONCH and H-optimus ensemble showed significant improvements in seven tasks (**Figure S3B**). When analyzing the Cohen's kappa of the concatenated vectors in comparison to the four individual models, the ensemble of CONCH and Prov-GigaPath was more similar to CONCH (0.72) than to Prov-GigaPath (0.62), indicating that the aggregator model relied more on the information given by

the CONCH vectors than the Prov-GigaPath vectors. This further suggests that CONCH extracts more relevant information from WSIs. In contrast, the ensemble of all four models focused similarly on all four models (scores between 0.54 and 0.58) (**Figure 4B**).

These results demonstrate that ensemble approaches for pathology foundation models, as well as their downstream models, lead to enhanced prediction performance. This suggests that merging multiple foundation models through ensemble techniques can be beneficial.

3. Discussion

Weakly-supervised computational pathology approaches, in which a deep learning system predicts a label directly from a whole slide image, have been massively successful in cancer research. They have been used to make the diagnosis of tumors, to predict biomarker status, and to predict clinical outcomes directly from image data. Over one hundred such tools are now approved for clinical use in the US and the European Union^{25,26}. Since 2022, foundation models have become an integral part of weakly supervised computational pathology pipelines and have improved performance and generalizability^{4,27}. However, the current internal evaluation strategy for foundation models in computational pathology for clinically relevant tasks is limited. When groups that publish pathology foundation models evaluate them on tasks of their own choosing, there is a high potential for bias. Moreover, concerns about data leakage arise when foundation models are tested on images from the same institutions where they were trained.

In this study, we conducted a comprehensive evaluation of pathology foundation models in weakly-supervised computational pathology on external datasets and a diverse set of biomarkers. Our results show that many of the existing foundation models achieve high performance on clinically relevant prediction tasks in cancer histopathology, highlighting their potential utility in this domain. Interestingly, the best performing model, CONCH, was trained with images and text, which insinuates that training on multimodal data boosts image-only embedding quality for pathology foundation models. In terms of prediction interpretability, our approach highlights that different foundation models focus on different areas in the tissue, while still having a high agreement on the predicted label. Given this property, we show that ensembling foundation models is beneficial for weakly-supervised computational pathology, combining several learned perspectives of the tissue morphology. However, it also indicates that there is not yet a single pathology foundation model that is both truly foundational and independent. Our ensembles of foundation models show potential for enhancing prediction performance in computational pathology.

A key insight of our study is that performance of foundation models does not scale well with increasing numbers of images in the training set used for self-supervised learning. Meaning, bigger is not always better. Rather, the diversity of the training set suggests to be a key factor, favoring various sources of data, races, and types of cancer. Our results will inform the future development of new foundation models. Specifically, using multimodal data to train models, even if the intention is just to apply them on unimodal data (i.e. on images alone), should be encouraged. For healthcare institutions, this means that data which is available at scale, even without clinical association with clinical endpoints, is a valuable resource to train such models.

Our study has limitations in that our evaluation tasks only contain certain tumor types. We were limited to pathology foundation models licenses which are accessible in a research setting. For example, this excludes RudolphV and PLUTO from our analysis. Our evaluation strategy is focused on a diverse set of biomarkers in cancer histopathology. Future work will expand upon the range of tumor types, biomarkers and patient cohorts to evaluate the robustness of foundation models in pathology.

Author contributions

PN, OSMEN, HSM and JNK designed the study. PN, TL and MVT developed the software. PN, MH, HB, HSM, RL, BD, HMB, CR, AM, OLS and JNK contributed to data collection and assembly. PN, OSMEN, TL, HSM, SF, DT and OLS interpreted and analyzed the data. All authors substantially contributed to writing and reviewing the report, approved the final version for submission, and have agreed to be personally accountable for the author's own contributions and to ensure that questions related to the accuracy or integrity of any part of the work, even ones in which the author was not personally involved, are appropriately investigated, resolved, and the resolution documented in the report.

Acknowledgements

No acknowledgments to declare.

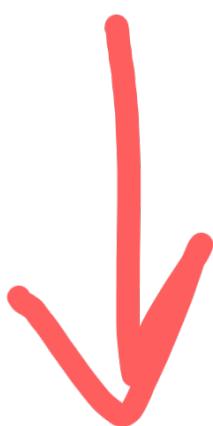
Disclosures

JNK declares consulting services for Bioptimus, France; Owkin, France; DoMore Diagnostics, Norway; Panakeia, UK; AstraZeneca, UK; Scailyte, Switzerland; Mindpeak, Germany; and MultiplexDx, Slovakia. Furthermore he holds shares in StratifAI GmbH, Germany, has received a research grant by GSK, and has received honoraria by AstraZeneca, Bayer, Daiichi Sankyo, Eisai, Janssen, MSD, BMS, Roche, Pfizer and Fresenius. DT received honoraria for lectures by Bayer and holds shares in StratifAI GmbH, Germany. SF has received honoraria from MSD and BMS. RL declares consulting services and honoraria from MSD, Janssen, Astra Zeneca, Astellas, Roche. AM has received honoraria as a consultant, advisor or speaker from Roche, Lilly and Menarini/Stemline, and has received support for accommodation and travel from AstraZeneca, all outside the submitted work. OSMEN holds shares in StratifAI GmbH, Germany. No other COIs are declared by any of the authors.

Funding

JNK is supported by the German Cancer Aid (DECADE, 70115166), the German Federal Ministry of Education and Research (PEARL, 01KD2104C; CAMINO, 01EO2101; SWAG, 01KD2215A; TRANSFORM LIVER, 031L0312A; TANGERINE, 01KT2302 through ERA-NET Transcan; Come2Data, 16DKZ2044A; DEEP-HCC, 031L0315A), the German Academic Exchange Service (SECAI, 57616814), the German Federal Joint Committee (TransplantKI, 01VSF21048) the European Union's Horizon Europe and innovation programme (ODELIA, 101057091; GENIAL, 101096312), the European Research Council (ERC; NADIR, 101114631), the National Institutes of Health (EPICO, R01 CA263318) and the National

Institute for Health and Care Research (NIHR, NIHR203331) Leeds Biomedical Research Centre. The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health and Social Care. This work was funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them. SF is supported by the German Federal Ministry of Education and Research (SWAG, 01KD2215A), the German Cancer Aid (DECADE, 70115166 and TargHet, 70115995) and the German Research Foundation (504101714). The DACHS study was supported by the German Research Council (BR 1704/6-1, BR 1704/6-3, BR 1704/6-4, CH 117/1-1, HO 5117/2-1, HO 5117/2-2, HE 5998/2-1, HE 5998/2-2, KL 2354/3-1, KL 2354 3-2, RO 2270/8-1, RO 2270/8-2, BR 1704/17-1, BR 1704/17-2); the Interdisciplinary Research Program of the National Center for Tumor Diseases (NCT), Germany; and the German Federal Ministry of Education and Research (01KH0404, 01ER0814, 01ER0815, 01ER1505A, and 01ER1505B). AM is supported by the European Society for Medical Oncology José Baselga Fellowship for Clinician Scientists founded by AstraZeneca (2023–2025).



Figures

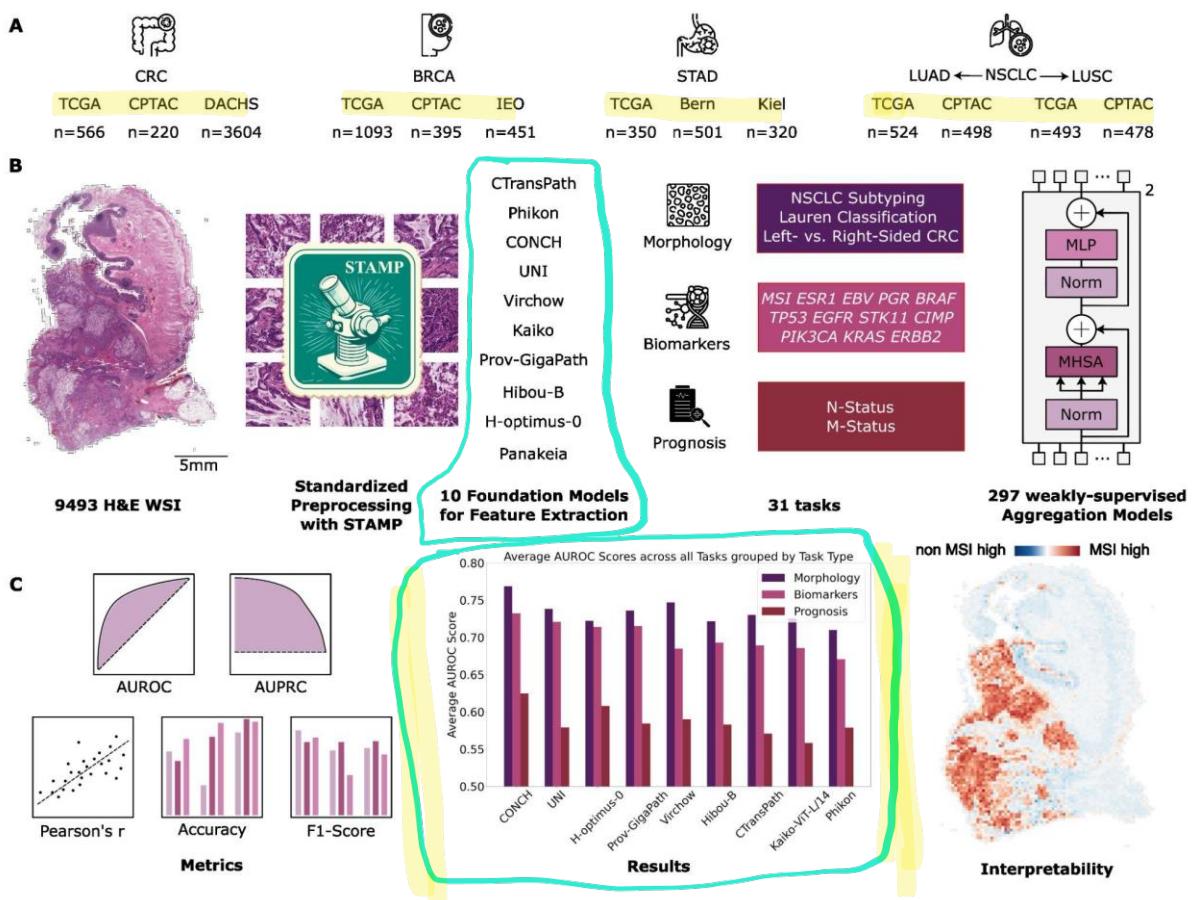
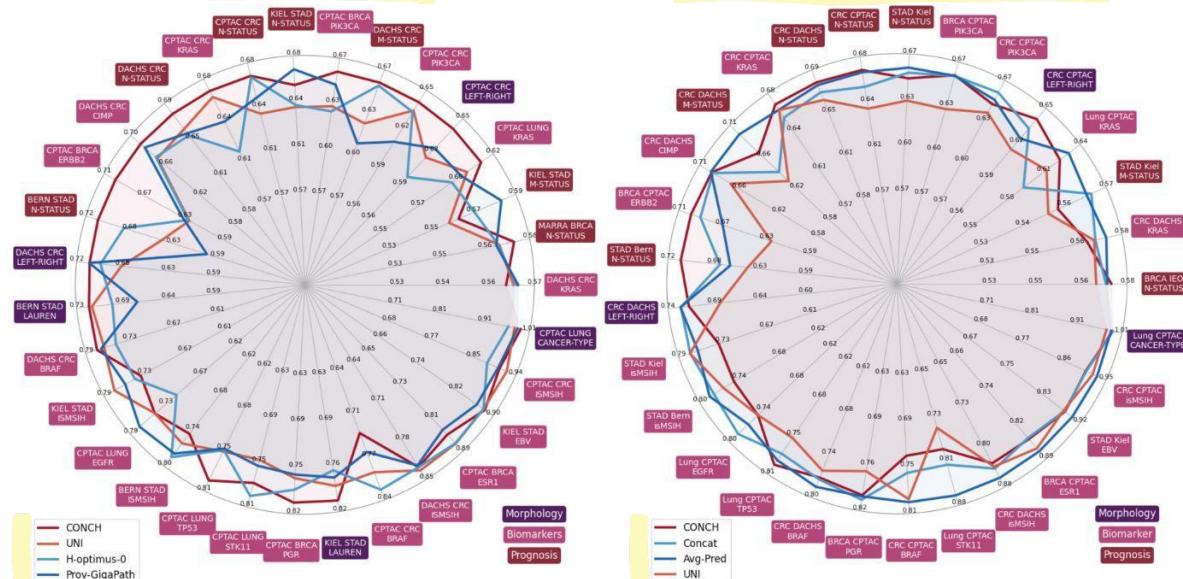


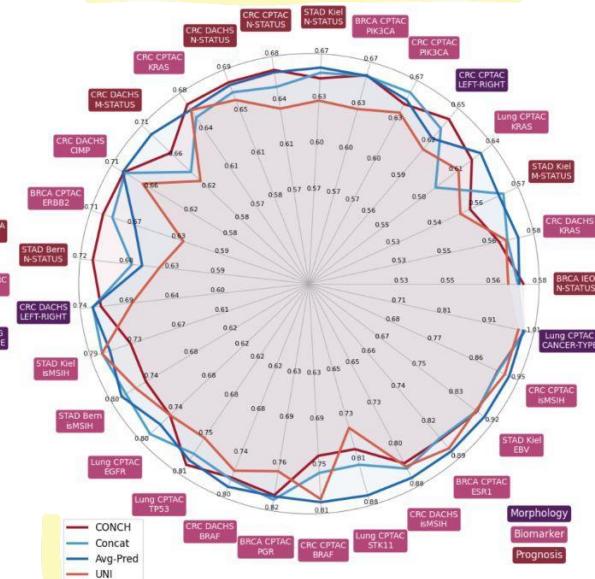
Fig. 1: Experimental design of the study. Benchmarking of ten histopathology foundation models using 13 cohorts and 31 tasks. **A**, number of slides used from each of the 13 cohorts including four cancer types. **B**, 9493 hematoxylin and eosin (H&E) stained whole slide images (WSIs) were preprocessed using the standardized STAMP¹⁹ pipeline. Feature extraction from the processed tiles was performed using ten foundation models analyzed in this study. The TCGA features were utilized for five-fold cross-validation with downstream Transformer models on 31 classification tasks using STAMP. All models were subsequently applied to external features from CPTAC, Bern, Kiel, DACHS, and IEO. **C**, All experiments were analyzed using AUROCs, supplemented by AUPRC, Pearson's correlation coefficient, accuracy and F1-score. CONCH generally outperforms other foundation models across all task categories, followed by UNI and H-optimus. Attention heatmaps were generated for some slides to interpret differences between foundation models.

A

Four best Foundation Models

**B**

Ensembles vs. two best Foundation Models

**D**

Biomarker Tasks

Task	CPTAC CRC MSI	CPTAC BRCA ESR1	KIEL STAD EBV	DACHS CRC MSI	CPTAC BRCA PGR	BERN STAD MSI	CPTAC CRC BRAF	DACHS CRC BRAF	KIEL STAD MSI	CPTAC LUNG TP53	CPTAC LUNG EGFR	CPTAC LUNG STK11	DACHS CRC CIMP	CPTAC BRCA PIK3CA	CPTAC CRC KRAS	CPTAC CRC PIK3CA	CPTAC LUNG KRAS	DACHS CRC KRAS	Average
	0.91	0.92	0.90	0.88	0.86	0.86	0.87	0.87	0.86	0.85	0.85								0.73
	0.84	0.87	0.83	0.87	0.87	0.86	0.86	0.86	0.84	0.78	0.78	0.89							0.72
	0.88	0.88	0.86	0.88	0.84	0.85	0.72	0.84	0.77										0.73
	0.83	0.83	0.82	0.83	0.75	0.82	0.79	0.85	0.77	0.83									0.73
	0.80	0.75	0.75	0.77	0.79	0.80	0.72	0.78	0.72	0.72	0.72	0.82							0.72
	0.73	0.75	0.77	0.78	0.72	0.68	0.75	0.74	0.70										0.71
	0.71	0.79	0.75	0.82	0.72	0.77	0.66	0.62	0.72	0.77									0.71
	0.77	0.75	0.76	0.73	0.73	0.69	0.70	0.73	0.71	0.72									0.72
	0.72	0.77	0.75	0.73	0.76	0.67	0.76	0.73	0.69										0.69
	0.79	0.73	0.72	0.73	0.70	0.71	0.73	0.72	0.70										0.70
	0.73	0.73	0.77	0.69	0.66	0.72	0.74	0.70	0.70	0.70									0.70
	0.77	0.73	0.74	0.79	0.70	0.61	0.73	0.55	0.61										0.61
	0.68	0.65	0.68	0.65	0.64	0.63	0.64	0.68	0.61	0.66									0.66
	0.65	0.60	0.63	0.59	0.56	0.65	0.59	0.61	0.60	0.64									0.64
	0.66	0.65	0.61	0.56	0.64	0.63	0.55	0.59	0.58	0.62									0.62
	0.63	0.62	0.56	0.61	0.58	0.50	0.57	0.63	0.62	0.58									0.58
	0.69	0.56	0.58	0.56	0.59	0.58	0.58	0.57	0.50	0.61	0.61	0.61							0.61
	0.60	0.58	0.57	0.55	0.54	0.55	0.51	0.46	0.57										0.54
	0.53	0.54	0.55	0.55	0.50	0.52	0.55	0.50	0.54	0.54	0.54	0.54							0.54
Average	0.73	0.72	0.72	0.71	0.69	0.69	0.69	0.68	0.67	0.71									0.71

C

Morphology Tasks

Task	CPTAC LUNG CANCER TYPE	KIEL STAD LAUREN	BERN STAD LAUREN	DACHS CRC LEFT RIGHT	CPTAC CRC LEFT RIGHT	Average
	0.99	0.99	0.98	0.99	0.98	0.97
	0.80	0.81	0.78	0.76	0.79	0.75
	0.71	0.69	0.71	0.64	0.67	0.71
	0.70	0.67	0.65	0.70	0.64	0.69
	0.63	0.57	0.57	0.59	0.57	0.58
Average	0.77	0.75	0.74	0.74	0.73	0.72

E

Prognosis Tasks

Task	DACHS CRC N STATUS	KIEL STAD N STATUS	CPTAC CRC N STATUS	DACHS CRC M STATUS	BERN STAD N STATUS	MARRA BRCA N STATUS	KIEL STAD M STATUS	Average
	0.67	0.63	0.63	0.63	0.59	0.64	0.62	0.60
	0.63	0.60	0.61	0.66	0.63	0.60	0.61	0.62
	0.66	0.66	0.59	0.62	0.58	0.60	0.57	0.57
	0.65	0.64	0.66	0.55	0.61	0.58	0.58	0.56
	0.70	0.66	0.60	0.52	0.57	0.60	0.54	0.55
	0.56	0.54	0.56	0.54	0.55	0.54	0.53	0.53
	0.51	0.53	0.47	0.57	0.55	0.58	0.51	0.49
Average	0.62	0.61	0.59	0.58	0.58	0.58	0.57	0.56

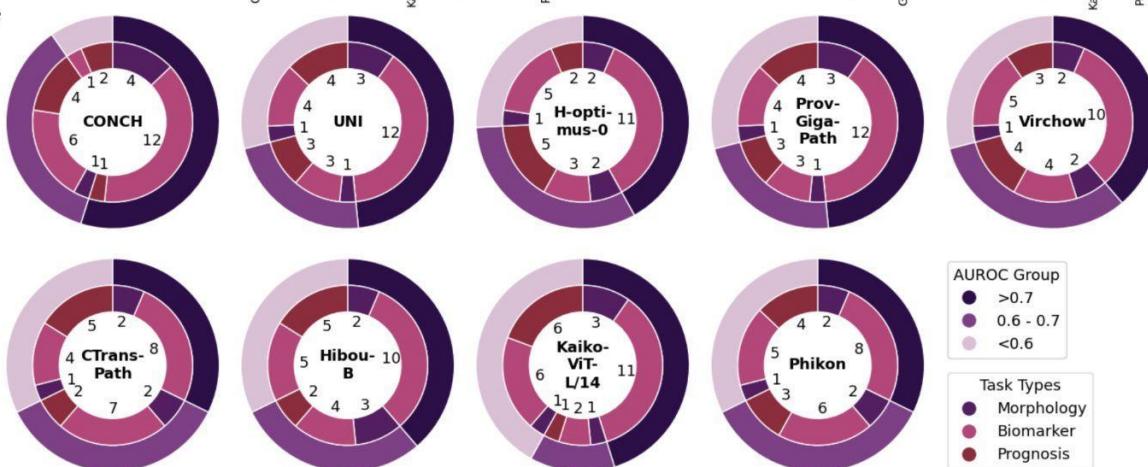
F

Fig. 2: Performance of ten pathology foundation models on 31 weakly-supervised prediction tasks. **A**, Area under the receiver operator characteristic (AUROC) scores of the four best foundation models, taskwise normalization. **B**, AUROC scores of the two best foundation models compared to the average prediction of the four best models (Avg-Pred) and the concatenated vectors of CONCH and Prov-GigaPath (Concat). **C-E**, Average AUROC scores of the five-folds of each foundation model on Morphology (**C**), Biomarker (**D**) and Prognosis (**E**) tasks. Taskwise normalization for better comparison of the foundation models. Tasks are sorted by their mean AUROC across all models, while models are sorted by their mean AUROC across all tasks. **F**, stacked pie charts showing the number of tasks where each model achieved an average AUROC of >0.7 , $0.6 - 0.7$, or <0.6 , grouped by task type.

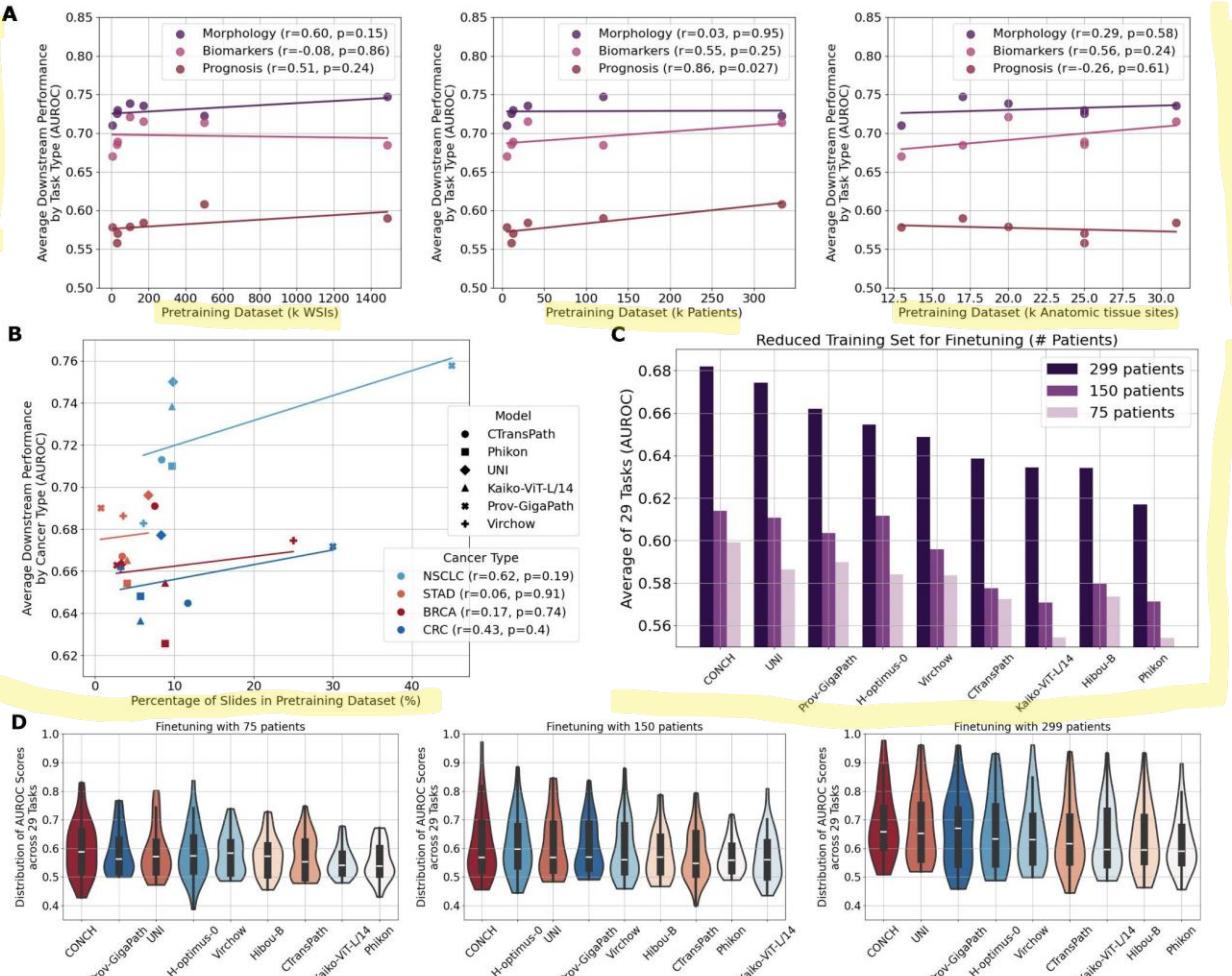


Fig. 3: The impact of data diversity and volume on downstream weakly-supervised classification performance. **A,B**, The impact of foundation model data diversity on downstream classification. Correlation between the number of WSIs, patients and anatomic tissue sites in the pretraining dataset and the average AUROC for each downstream task type for the models CTransPath, Phikon, Kaiko, UNI, Prov-GigaPath, H-optimus-0, and Virchow (A). Performance of the respective cancer types correlated with the proportion of the cancer type in the pretraining dataset (B). CONCH was excluded due to its unique pretraining approach, which involved first pretraining the image encoder with WSIs and then using image-caption pairs for vision-language pretraining. Hibou was also excluded as the performance of Hibou-L, which utilized all data, couldn't be tested in this study. All other information that was available is shown (Table S2, S3). **C-D**, Experiments with reduced downstream training sizes. Average AUROC scores across 29 tasks, trained with 75, 150, or 299 patients (C). Distribution of AUROC scores across all tasks for each model separately (D).



Fig. 4: Divergence in tissue focus and predictive similarity among foundation models and ensembles thereof. **A**, Attention Heatmap Analysis for MSI-H Classification in four different DACHS samples selected for correct predictions across all foundation models. Thumbnails of the original whole slide images (WSIs) and heatmaps of all foundation models. **B-C**, Objective measure of similarity of prediction scores using Cohen's Kappa and majority vote across the five folds to binarize the predictions. Kappa scores of the best foundation models tested in this study with the concatenated vectors of CONCH and Prov-GigaPath (CPG-Concat) and with the concatenated vectors of all four foundation models (Top-4-Concat) (**B**). CPG-Concat is more similar to CONCH than to Prov-GigaPath, suggesting that the prediction was primarily based on the CONCH features. Kappa Scores of the four best foundation models compared to each other shows the lowest similarity between UNI and CONCH, which therefore provide the highest complementary predictive power (**C**). Only tasks where at least one model achieves a mean AUROC over 0.7 are included, with "CPTAC CRC BRAF" excluded due to low majority vote accuracy. **D**, Cohen's Kappa between the five folds of each foundation model across the same tasks as in B and C.

4.

Material and Methods

Ethics statement

This study was carried out in accordance with the Declaration of Helsinki. The Clinical Proteomic Tumor Analysis Consortium (CPTAC) and TCGA did not require formal ethics approval for a retrospective study of anonymised samples. The analysis of the testing cohort DACHS (an epidemiological study which is led by the German Cancer Research Center, DKFZ, Heidelberg, Germany) was approved by the ethics committee of the Medical Faculty, University of Heidelberg under 310/2001²⁸⁻³⁰.

Datasets

The study utilized datasets from TCGA, CPTAC and proprietary cohorts. Specifically, cohorts from lung adenocarcinoma (LUAD), lung squamous cell carcinoma (LUSC), colorectal cancer (CRC), stomach adenocarcinoma (STAD), and breast cancer (BRCA) were included. TCGA datasets were used for training of the models and CPTAC, DACHS, Kiel, Bern and IEO were used for evaluation. This ensured that all testing was done on data that had neither been seen during training of the foundation models nor the fine-tuned models.

For external validation, CPTAC datasets for LUAD, LUSC, colorectal adenocarcinoma (COAD), and BRCA were used. No foundation models analyzed in this study were trained on CPTAC, ensuring its suitability as an independent test cohort. Additionally, for CRC, the DACHS cohort was utilized alongside CPTAC as another external test set. In STAD, proprietary datasets from Kiel and Bern served as external validation cohorts. For BRCA, the IEO dataset was used alongside CPTAC for external validation (**Figure 1A, S12**).

To ensure transparency and standardization in reporting the dataset and methods used, we adhered to the STARD (Standards for Reporting Diagnostic Accuracy Studies) checklist.

Experimental Design

DP involves several task categories, including morphological, biomarker and prognostic tasks, and foundation models should be capable of performing well across all of them. In this study, we assembled and benchmarked ten foundation models - CONCH, CTransPath, Hibou-B, H-optimus-0, Phikon, Prov-GigaPath, UNI, Kaiko (ViT-L/14), Virchow, and Panakeia - across a comprehensive set of tasks from all three categories. Each category was assessed across all cancer types, apart from morphological features in BRCA and prognostic features in NSCLC due to data unavailability. To enable both training and independent testing, each task required ground truth data to be available in TCGA (for training) and at least one test cohort. For each cohort, only tasks with at least 10 cases in each category were included.

First, we investigated morphological classification tasks related to cancer subgroups with distinct phenotypic characteristics. The aim was to assess foundation models by evaluating their ability to discern established phenotypic distinctions. In CRC, the morphological task involved predicting whether the slide originated from the left or right side of the colon, excluding colon transversum samples due to ambiguous classification. In STAD, the Lauren

classification³¹ was chosen as the morphological task, classifying slides as “intestinal”, “diffuse”, or “mixed”, given the unavailability of ground truth for newer classification systems^{32,33}. In lung cancer, the models were tasked with classifying samples into either adenocarcinoma or squamous cell carcinoma¹.

Biomarker prediction tasks focused mainly on clinically relevant targets with some type of morphological correlation as demonstrated by previous computational pathology models. For CRC, these included *BRAF*, *KRAS*, microsatellite instability (MSI) status, *PIK3CA*, and CpG island methylator phenotype (CIMP) status¹¹. For STAD, Epstein-Barr virus (EBV) presence and MSI status were selected³⁴. For LUAD, the targets were *EGFR*, *STK11*, *KRAS*, and *TP53*¹. For BRCA, the targets were the expression of HER2, ER, PR receptors, and *PIK3CA* mutations^{35,36}. MSI status and CIMP status were binarized into MSI-high versus not MSI-high and CIMP-high versus not CIMP-high, respectively. HER2, ER, and PR expression were binarized using the z-score of mRNA expression profiles, similar to a study by Wegscheider et al.³⁷. This approach was preferred over immunohistochemistry labels due to its objectivity and reduced variance error.

Prognostic tasks, which aim to predict clinical outcomes directly from whole-slide images (WSIs), were selected based on their prognostic relevance. The tasks included N status for CRC, STAD, and BRCA, where all stages except N0 were classified as N+ (excluding Nx cases). M-Status was analyzed in CRC and STAD, performing binary classification of M0 versus M+.

By focusing on tasks with clear therapeutic actionability or prognostic relevance, we aimed to evaluate the practical utility of these models in a clinical setting. This comprehensive benchmarking study included 31 tasks across 8 external test cohorts, encompassing a wide range of clinically relevant classification tasks (**Table S4**).

Image Processing and Deep Learning Techniques

4.3 The benchmarking was conducted using the STAMP pipeline version 1.1.0¹⁹. Each classification task followed a two-step procedure (**Figure 1B**). In the first step, feature vectors were extracted from WSIs utilizing the foundational models evaluated in this study. In the second step, these vectors were employed to train a slide-level aggregator on the downstream tasks described above.

WSIs were segmented into N tiles, with an edge length of 224x224px corresponding to 256 µm, resulting in an effective magnification of ~1.14 µm per pixel. Background tiles were excluded using Canny edge detection³⁸. Feature extraction was performed on each tile individually using the different foundational models. The embedding dimensions M varied across models, ranging from M=512 for CONCH to M=2560 for Virchow. Subsequently, each slide was transformed into a 2D matrix with dimension NxM. The extracted feature vectors were input into a Transformer-based aggregator model⁴. It utilizes multi-head attention, Gaussian Error Linear Unit (GELU) activation functions³⁹, layer normalization, and a multilayer perceptron (MLP) head to produce an output corresponding to the k possible classes for each task. A five-fold cross-validation approach was implemented, resulting in the creation of 1,395 models (9 foundation models, 31 tasks, 5 folds) trained exclusively on TCGA datasets. All

experiments were run on individual 40 GB NVIDIA RTX A6000 and L40 GPU (graphics processing unit) nodes.

Prov-GigaPath provides both a slide-level and a tile-level encoder and we tested both approaches²². The main results always include the slide-encoder version, as recommended by the authors. In the case of Virchow, Vorontsov et al. proposed concatenating the class token with the average pool of patch tokens for each tile embedding. To maintain consistency with other models that only use class tokens, two configurations were tested: one including and one excluding the averaged patch tokens. The main results show the version recommended by the authors. For CONCH, we used the output of the attentional pooler that corresponds to image-text alignment, with an embedding dimension of 512. For experiments involving combined feature vectors, vectors were concatenated, maintaining a single vector per tile. For instance, combining CONCH and Prov-GigaPath resulted in a combined embedding dimension M of M=1,280 (M=512 for CONCH + M=768 for Prov-GigaPath).

Explainability

4.4 To better interpret the output of the models, we generated whole-slide prediction heatmaps for selected tasks. These heatmaps illustrate the models' focus on specific tissue areas, by weighting the scores assigned to individual tiles using Gradient-weighted Class Activation Mapping (Grad-CAM)⁴⁰. It is important to note that a high number of positively contributing tiles do not automatically result in a high final score due to the non-linear aggregation process in neural networks⁴¹. The benchmarking effort involved 1395 models and 9493 slides, leading to a vast number of model-slide combinations. Thus, it was necessary to select a few informative examples methodically. Slides were selected by including cases where models showed strong disagreements and cases where all models performed well. The heatmaps were visually analyzed and compared to the underlying WSI. To further analyze the similarity between different models, Cohen's kappa²⁴ was measured between each pair of foundation models.

Statistical Analysis

4.5 The performance of the models was evaluated using the Area Under the Receiver Operating Characteristic curve (AUROC) employing five-fold cross-validation and deployment on external cohorts. Mean AUROC scores from the five cross-validation models deployed on external data were used for statistical and graphical evaluations. Predictions were made per patient, and all feature matrices belonging to one patient were concatenated for use in the model. In addition to AUROC, for completeness in the supplementary material, we also calculated the Area Under the Precision-Recall Curve (AUPRC), accuracy, and F1 score. The two-sided DeLong's test was used to test for statistically significant differences in AUROC scores. As the DeLong's test is only applicable when a single prediction score is available for each model and sample, the average prediction score across all five folds was employed. Due to its multi-class nature, we excluded Lauren classification tasks from this analysis. This differs from the main metrics, where the AUROC/AUPRC/F1/accuracy scores represent the mean across the five folds.

4.6 Data availability

The slides for TCGA are available at <https://portal.gdc.cancer.gov/>. The slides for CPTAC are available at <https://proteomics.cancer.gov/data-portal>. The molecular data for TCGA and CPTAC are available at <https://www.cbioportal.org/>. The slides and biomarker data for DACHS were generated for prior studies^{42–44} with restricted access. Biomarker data for DACHS are available by requesting Authorized Access to the phs001078 study [https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs001113.v1.p1]. Applications for access to DACHS biomarker data are reserved for Senior Investigators and NIH Investigators as defined in <https://dbgap.ncbi.nlm.nih.gov/aa/wga.cgi>, and upon successful application grants access to the data for 1 year with the option to renew access. The slides for DACHS can only be requested directly through the DACHS principal investigators. The contact details are listed at <http://dachs.dkfz.org/dachs/kontakt.html>. All other cohorts can be requested from the respective study investigators. The data generated in this study for the creation of the figures are provided in the Source Data file. Source data are provided with this paper.

Code availability

The open-source STAMP software for the implementation of the Benchmarking experiments is available on GitHub (<https://github.com/KatherLab/STAMP>).



References

1. 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).
2. Kather, J. N. *et al.* Deep learning can predict microsatellite instability directly from histology in gastrointestinal cancer. *Nat. Med.* **25**, 1054–1056 (2019).
3. Lu, M. Y. *et al.* Data-efficient and weakly supervised computational pathology on whole-slide images. *Nat Biomed Eng* **5**, 555–570 (2021).
4. Wagner, S. J. *et al.* Transformer-based biomarker prediction from colorectal cancer histology: A large-scale multicentric study. *Cancer Cell* **41**, 1650–1661.e4 (2023).
5. Loeffler, C. M. L. *et al.* Direct prediction of Homologous Recombination Deficiency from routine histology in ten different tumor types with attention-based Multiple Instance Learning: a development and validation study. *medRxiv* (2023)
doi:10.1101/2023.03.08.23286975.
6. Liu, Q. *et al.* Identification of lymph node metastasis in pre-operation cervical cancer patients by weakly supervised deep learning from histopathological whole-slide biopsy images. *Cancer Med.* **12**, 17952–17966 (2023).
7. da Silva, L. M. *et al.* Independent real-world application of a clinical-grade automated prostate cancer detection system. *J. Pathol.* **254**, 147–158 (2021).
8. Bagg, A. *et al.* Performance evaluation of a novel artificial intelligence-assisted digital microscopy system for the routine analysis of bone marrow aspirates. *Mod. Pathol.* **37**, 100542 (2024).
9. Yang, Z. *et al.* The devil is in the details: a small-lesion sensitive weakly supervised learning framework for prostate cancer detection and grading. *Virchows Arch.* **482**, 525–538 (2023).
10. El Nahhas, O. S. M. *et al.* Regression-based Deep-Learning predicts molecular biomarkers from pathology slides. *Nat. Commun.* **15**, 1253 (2024).
11. Niehues, J. M. *et al.* Generalizable biomarker prediction from cancer pathology slides

- with self-supervised deep learning: A retrospective multi-centric study. *Cell Rep Med* **4**, 100980 (2023).
12. Moor, M. *et al.* Foundation models for generalist medical artificial intelligence. *Nature* **616**, 259–265 (2023).
 13. Waqas, A. *et al.* Revolutionizing Digital Pathology With the Power of Generative Artificial Intelligence and Foundation Models. *Lab. Invest.* **103**, 100255 (2023).
 14. He, K., Fan, H., Wu, Y., Xie, S. & Girshick, R. B. Momentum Contrast for unsupervised visual representation learning. *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.* 9726–9735 (2019).
 15. Chen, T., Kornblith, S., Norouzi, M. & Hinton, G. A Simple Framework for Contrastive Learning of Visual Representations. in *Proceedings of the 37th International Conference on Machine Learning* (eds. lii, H. D. & Singh, A.) vol. 119 1597–1607 (PMLR, 13–18 Jul 2020).
 16. Filiot, A. *et al.* Scaling self-Supervised Learning for histopathology with Masked Image Modeling. *bioRxiv* (2023) doi:10.1101/2023.07.21.23292757.
 17. Wu, W., Gao, C., DiPalma, J., Vosoughi, S. & Hassanpour, S. Improving Representation Learning for Histopathologic Images with Cluster Constraints. *Proc. IEEE Int. Conf. Comput. Vis.* **2023**, 21347–21357 (2023).
 18. Dosovitskiy, A. *et al.* An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *arXiv [cs.CV]* (2020).
 19. El Nahhas, O. S. M. *et al.* From Whole-slide Image to Biomarker Prediction: A Protocol for End-to-End Deep Learning in Computational Pathology. *arXiv [cs.CV]* (2023).
 20. Schömig-Markiefka, B. *et al.* Quality control stress test for deep learning-based diagnostic model in digital pathology. *Mod. Pathol.* **34**, 2098–2108 (2021).
 21. Chen, R. J. *et al.* Towards a general-purpose foundation model for computational pathology. *Nat. Med.* **30**, 850–862 (2024).
 22. Xu, H. *et al.* A whole-slide foundation model for digital pathology from real-world data. *Nature* **630**, 181–188 (2024).

23. Vorontsov, E., Bozkurt, A., Casson, A. & Shaikovski, G. A foundation model for clinical-grade computational pathology and rare cancers detection. *Nat. Med.* (2024).
24. Cohen, J. *A Coefficient of Agreement for Nominal Scales*. (1960).
25. Geaney, A. *et al.* Translation of tissue-based artificial intelligence into clinical practice: from discovery to adoption. *Oncogene* **42**, 3545–3555 (2023).
26. Benjamins, S., Dhunnoor, P. & Meskó, B. The state of artificial intelligence-based FDA-approved medical devices and algorithms: an online database. *NPJ Digit Med* **3**, 118 (2020).
27. Wang, X. *et al.* Transformer-based unsupervised contrastive learning for histopathological image classification. *Med. Image Anal.* **81**, 102559 (2022).
28. Carr, P. R. *et al.* Estimation of Absolute Risk of Colorectal Cancer Based on Healthy Lifestyle, Genetic Risk, and Colonoscopy Status in a Population-Based Study. *Gastroenterology* **159**, 129–138.e9 (2020).
29. Hoffmeister, M. *et al.* Colonoscopy and Reduction of Colorectal Cancer Risk by Molecular Tumor Subtypes: A Population-Based Case-Control Study. *Am. J. Gastroenterol.* **115**, 2007–2016 (2020).
30. Brenner, H., Chang-Claude, J., Seiler, C. M., Stürmer, T. & Hoffmeister, M. Does a negative screening colonoscopy ever need to be repeated? *Gut* **55**, 1145–1150 (2006).
31. Lauren, P. THE TWO HISTOLOGICAL MAIN TYPES OF GASTRIC CARCINOMA: DIFFUSE AND SO-CALLED INTESTINAL-TYPE CARCINOMA. AN ATTEMPT AT A HISTO-CLINICAL CLASSIFICATION. *Acta Pathol. Microbiol. Scand.* **64**, 31–49 (1965).
32. Veldhuizen, G. P. *et al.* Deep learning-based subtyping of gastric cancer histology predicts clinical outcome: a multi-institutional retrospective study. *Gastric Cancer* **26**, 708–720 (2023).
33. Nagtegaal, I. D. *et al.* The 2019 WHO classification of tumours of the digestive system. *Histopathology* **76**, 182–188 (2020).
34. Muti, H. S. *et al.* Development and validation of deep learning classifiers to detect Epstein-Barr virus and microsatellite instability status in gastric cancer: a retrospective

- multicentre cohort study. *Lancet Digit Health* **3**, e654–e664 (2021).
35. Kather, J. N. *et al.* Pan-cancer image-based detection of clinically actionable genetic alterations. *Nat Cancer* **1**, 789–799 (2020).
 36. Mandair, D., Reis-Filho, J. S. & Ashworth, A. Biological insights and novel biomarker discovery through deep learning approaches in breast cancer histopathology. *NPJ Breast Cancer* **9**, 21 (2023).
 37. Wegscheider, A.-S. *et al.* Comprehensive and Accurate Molecular Profiling of Breast Cancer through mRNA Expression of ESR1, PGR, ERBB2, MKI67, and a Novel Proliferation Signature. *Diagnostics (Basel)* **14**, (2024).
 38. Canny, J. A computational approach to edge detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **8**, 679–698 (1986).
 39. Hendrycks, D. & Gimpel, K. Gaussian Error Linear Units (GELUs). *arXiv [cs.LG]* (2016).
 40. Selvaraju, R. R. *et al.* Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *Int. J. Comput. Vis.* **128**, 336–359 (2020).
 41. Chen, R. J. & Krishnan, R. G. Self-Supervised Vision Transformers Learn Visual Concepts in Histopathology. *arXiv [cs.CV]* (2022).
 42. Lilla, C. *et al.* Effect of NAT1 and NAT2 genetic polymorphisms on colorectal cancer risk associated with exposure to tobacco smoke and meat consumption. *Cancer Epidemiol. Biomarkers Prev.* **15**, 99–107 (2006).
 43. Brenner, H., Chang-Claude, J., Seiler, C. M. & Hoffmeister, M. Long-term risk of colorectal cancer after negative colonoscopy. *J. Clin. Oncol.* **29**, 3761–3767 (2011).
 44. Hoffmeister, M. *et al.* Statin use and survival after colorectal cancer: the importance of comprehensive confounder adjustment. *J. Natl. Cancer Inst.* **107**, djv045 (2015).
 45. Chen, X., Xie, S. & He, K. An empirical study of training self-supervised Vision Transformers. *ICCV* 9620–9629 (2021).
 46. Bao, H., Dong, L., Piao, S. & Wei, F. BEiT: BERT Pre-Training of Image Transformers. *arXiv [cs.CV]* (2021).
 47. Zhou, J. *et al.* iBOT: Image BERT Pre-Training with Online Tokenizer. *arXiv [cs.CV]*

- (2021).
48. Yu, J. *et al.* CoCa: Contrastive Captioners are Image-Text Foundation Models. *arXiv [cs.CV]* (2022).
 49. Lu, M. Y. *et al.* A visual-language foundation model for computational pathology. *Nat. Med.* **30**, 863–874 (2024).
 50. Oquab, M. *et al.* DINoV2: Learning Robust Visual Features without Supervision. *arXiv [cs.CV]* (2023).
 51. Caron, M., Touvron, H., Misra, I. & Jégou, H. Emerging properties in self-supervised vision transformers. *Proceedings of the* (2021).
 52. Ai, K. *et al.* Towards Large-Scale Training of Pathology Foundation Models. *arXiv [cs.CV]* (2024).
 53. Ding, J. *et al.* LongNet: Scaling Transformers to 1,000,000,000 Tokens. *arXiv [cs.CL]* (2023).
 54. Nechaev, D., Pchelnikov, A. & Ivanova, E. Hibou: A Family of Foundational Vision Transformers for Pathology. *arXiv [eess.IV]* (2024).
 55. Darzet, T., Oquab, M., Mairal, J. & Bojanowski, P. Vision Transformers Need Registers. *arXiv [cs.CV]* (2023).
 56. Saillard, C. *et al.* *H-Optimus-0*. (2024).
 57. Yang, Z. *et al.* A foundation model for generalizable cancer diagnosis and survival prediction from histopathological images. *bioRxiv* 2024.05.16.594499 (2024) doi:10.1101/2024.05.16.594499.
 58. Juyal, D. *et al.* PLUTO: Pathology-Universal Transformer. *arXiv [eess.IV]* (2024).
 59. Dippel, J. *et al.* RudolfV: A Foundation Model by Pathologists for Pathologists. *arXiv [eess.IV]* (2024).
 60. Hua, S., Yan, F., Shen, T. & Zhang, X. PathoDuet: Foundation Models for Pathological Slide Analysis of H&E and IHC Stains. *arXiv [cs.CV]* (2023).
 61. Campanella, G. *et al.* Computational Pathology at Health System Scale -- Self-Supervised Foundation Models from Three Billion Images. *arXiv [cs.CV]* (2023).

62. Campanella, G. *et al.* A Clinical Benchmark of Public Self-Supervised Pathology Foundation Models. *arXiv [eess.IV]* (2024).
63. Bellman, R. *Dynamic Programming*. (Princeton University Press, 1957).
64. Ainsworth, S. K., Hayase, J. & Srinivasa, S. Git Re-Basin: Merging Models modulo Permutation Symmetries. *arXiv [cs.LG]* (2022).



Supplementary Figures

Fig. S1: AUROCs, AUPRCs, Accuracy and F1-scores for all main experiments

A

Task	AUROC Scores by Model (Average) - Biomarker Tasks							
	CPTAC CRC MSI	CPTAC BRCA MSI	DACHS CRC MSI	CPTAC BRCA PGR	BERN STAD MSI	DACHS CRC BRAF	KIEL STAD MSI	CPTAC LUNG TP53
	0.91	0.92	0.90	0.88	0.86	0.86	0.87	0.86
	0.84	0.87	0.83	0.87	0.87	0.86	0.86	0.84
	0.88	0.88	0.86	0.88	0.84	0.85	0.72	0.84
	0.83	0.83	0.82	0.83	0.75	0.82	0.79	0.85
	0.80	0.75	0.77	0.77	0.79	0.80	0.72	0.78
	0.73	0.75	0.77	0.78	0.72	0.68	0.75	0.74
	0.77	0.75	0.76	0.73	0.73	0.69	0.70	0.73
	0.72	0.77	0.75	0.73	0.76	0.67	0.76	0.73
	0.71	0.79	0.75	0.82	0.72	0.77	0.66	0.62
	0.79	0.73	0.72	0.73	0.70	0.71	0.73	0.72
	0.73	0.73	0.77	0.69	0.66	0.72	0.74	0.70
	0.77	0.73	0.74	0.79	0.70	0.61	0.73	0.55
	0.65	0.65	0.68	0.65	0.64	0.63	0.64	0.68
	0.65	0.60	0.63	0.59	0.56	0.65	0.59	0.61
	0.66	0.65	0.61	0.56	0.64	0.63	0.55	0.58
	0.63	0.62	0.56	0.61	0.58	0.50	0.57	0.63
	0.69	0.56	0.58	0.56	0.59	0.58	0.58	0.57
	0.60	0.58	0.57	0.55	0.54	0.55	0.51	0.48
	0.53	0.54	0.55	0.55	0.50	0.52	0.55	0.50
Average	0.73	0.72	0.72	0.71	0.69	0.69	0.68	0.67
CONCH	UNI	Prov-GigaPath	H-opti-mus-0	Hibou-B	CTrans-Path	Kalgo-VIT-L14	Virchow	Phikon

AUROC Scores by Model (Average) - Morphology Tasks

Task	AUROC Scores by Model (Average) - Morphology Tasks													
	CPTAC LUNG CANCER TYPE	KIEL STAD LAUREN	BERN STAD LAUREN	DACHS CRC LEFT RIGHT	CPTAC CRC LEFT RIGHT	Average	CONCH	Virchow	UNI	Prov-GigaPath	CTrans-Path	Kalgo-VIT-L14	H-opti-mus-0	Hibou-B
	0.99	0.99	0.98	0.99	0.98	0.97	0.97	0.98	0.97	0.97	0.98	0.97	0.71	0.70
	0.80	0.81	0.78	0.76	0.79	0.73	0.75	0.75	0.75	0.75	0.76	0.75	0.62	0.62
	0.71	0.69	0.71	0.64	0.67	0.71	0.68	0.68	0.69	0.69	0.68	0.69	0.65	0.65
	0.70	0.67	0.65	0.70	0.64	0.64	0.63	0.64	0.63	0.63	0.64	0.63	0.62	0.62
	0.63	0.57	0.57	0.59	0.57	0.58	0.53	0.53	0.53	0.53	0.53	0.53	0.57	0.57
	0.77	0.75	0.74	0.74	0.73	0.73	0.72	0.72	0.72	0.72	0.72	0.72	0.21	0.21

B

Task	AUPRC Scores by Model (Average) - Biomarker Tasks							
	CPTAC CRC MSI	CPTAC BRCA MSI	DACHS CRC MSI	CPTAC BRCA PGR	BERN STAD MSI	DACHS CRC BRAF	KIEL STAD MSI	CPTAC LUNG TP53
	0.86	0.87	0.88	0.85	0.82	0.82	0.81	0.85
	0.81	0.85	0.81	0.85	0.84	0.83	0.83	0.76
	0.79	0.75	0.75	0.78	0.80	0.80	0.79	0.72
	0.71	0.73	0.77	0.68	0.66	0.71	0.69	0.70
	0.69	0.70	0.70	0.70	0.63	0.69	0.71	0.66
	0.66	0.71	0.67	0.72	0.67	0.70	0.62	0.59
	0.67	0.69	0.69	0.68	0.65	0.63	0.69	0.67
	0.73	0.69	0.69	0.68	0.69	0.65	0.71	0.56
	0.62	0.67	0.65	0.64	0.65	0.57	0.63	0.60
	0.67	0.66	0.62	0.60	0.65	0.65	0.60	0.56
	0.62	0.59	0.62	0.60	0.58	0.62	0.58	0.57
	0.63	0.58	0.60	0.58	0.62	0.59	0.60	0.58
	0.60	0.60	0.60	0.58	0.59	0.57	0.58	0.58
	0.63	0.60	0.56	0.60	0.58	0.52	0.59	0.56
	0.60	0.60	0.58	0.57	0.55	0.56	0.53	0.57
	0.59	0.54	0.56	0.54	0.55	0.54	0.55	0.53
	0.53	0.54	0.54	0.54	0.51	0.52	0.50	0.54
Average	0.68	0.67	0.67	0.67	0.65	0.65	0.64	0.63
CONCH	UNI	Prov-GigaPath	H-opti-mus-0	Hibou-B	CTrans-Path	Virchow	Kalgo-VIT-L14	Phikon

AUPRC Scores by Model (Average) - Morphology Tasks

Task	AUPRC Scores by Model (Average) - Morphology Tasks													
	CPTAC LUNG CANCER TYPE	KIEL STAD LAUREN	BERN STAD LAUREN	DACHS CRC LEFT RIGHT	CPTAC CRC LEFT RIGHT	Average	CONCH	Virchow	UNI	Prov-GigaPath	CTrans-Path	Kalgo-VIT-L14	H-opti-mus-0	Hibou-B
	0.99	0.99	0.98	0.99	0.98	0.97	0.97	0.98	0.97	0.97	0.98	0.97	0.71	0.70
	0.80	0.81	0.78	0.76	0.79	0.73	0.75	0.75	0.75	0.75	0.76	0.75	0.62	0.62
	0.71	0.69	0.71	0.64	0.67	0.71	0.68	0.68	0.69	0.69	0.68	0.69	0.65	0.65
	0.70	0.67	0.66	0.65	0.67	0.64	0.63	0.64	0.65	0.65	0.64	0.65	0.55	0.55
	0.63	0.57	0.57	0.59	0.57	0.54	0.53	0.54	0.55	0.55	0.54	0.55	0.50	0.50
	0.77	0.75	0.74	0.74	0.73	0.73	0.72	0.72	0.72	0.72	0.72	0.72	0.21	0.21

C

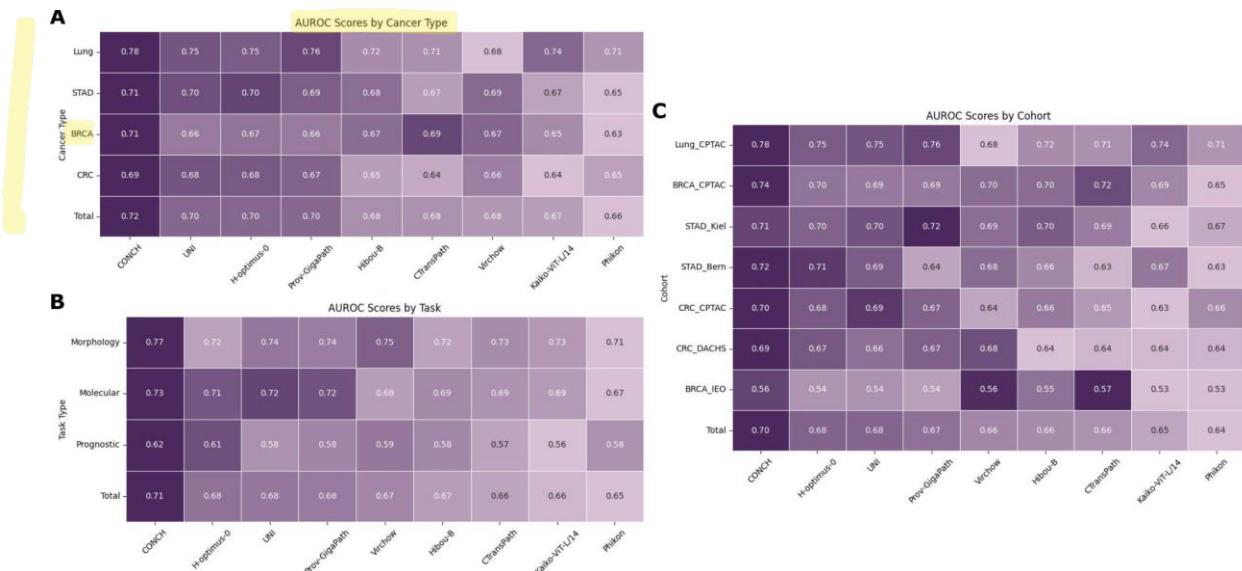
Task	ACC Scores by Model (Average) - Biomarker Tasks							
	KIEL STAD EBV	CPTAC BRCA MSI	DACHS CRC MSI	CPTAC BRCA PGR	BERN STAD MSI	DACHS CRC BRAF	KIEL STAD MSI	CPTAC LUNG TP53
	0.92	0.94	0.94	0.94	0.88	0.91	0.93	0.92
	0.83	0.88	0.89	0.90	0.89	0.86	0.88	0.86
	0.88	0.88	0.89	0.83	0.89	0.85	0.86	0.89
	0.84	0.87	0.86	0.87	0.82	0.85	0.81	0.82
	0.82	0.83	0.83	0.82	0.84	0.82	0.79	0.84
	0.81	0.83	0.83	0.84	0.82	0.84	0.74	0.84
	0.80	0.83	0.83	0.81	0.89	0.86	0.69	0.64
	0.74	0.71	0.63	0.81	0.39	0.66	0.64	0.64
	0.66	0.69	0.76	0.75	0.75	0.67	0.60	0.76
	0.66	0.61	0.73	0.70	0.61	0.75	0.72	0.72
	0.65	0.67	0.63	0.57	0.65	0.62	0.58	0.64
	0.63	0.68	0.63	0.63	0.63	0.63	0.62	0.63
	0.62	0.67	0.63	0.63	0.64	0.59	0.62	0.65
	0.62	0.57	0.50	0.57	0.54	0.57	0.47	0.55
	0.54	0.55	0.60	0.41	0.52	0.53	0.52	0.54
	0.57	0.56	0.47	0.48	0.55	0.51	0.47	0.48
	0.56	0.47	0.50	0.50	0.51	0.53	0.45	0.49
	0.54	0.47	0.50	0.50	0.53	0.53	0.49	0.52
	0.45	0.59	0.56	0.53	0.53	0.45	0.35	0.50
	0.49	0.49	0.54	0.49	0.47	0.37	0.50	0.47
	0.39	0.31	0.31	0.41	0.38	0.29	0.25	0.37
	0.32	0.18	0.31	0.22	0.22	0.19	0.24	0.25
	0.58	0.56	0.55	0.54	0.53	0.53	0.53	0.52
CONCH	UNI	Prov-GigaPath	H-opti-mus-0	CTrans-Path	Virchow	Kalgo-VIT-L14	Hibou-B	Phikon

ACC Scores by Model (Average) - Morphology Tasks

Task	ACC Scores by Model (Average) - Morphology Tasks														
	CPTAC LUNG CANCER TYPE	KIEL STAD LAUREN	BERN STAD LAUREN	DACHS CRC STATUS	CPTAC CRC STATUS	Average	CONCH	Virchow	UNI	Prov-GigaPath	CTrans-Path	Kalgo-VIT-L14	H-opti-mus-0	Hibou-B	Phikon
	0.91	0.91	0.95	0.91	0.89	0.94	0.86	0.89	0.91	0.91	0.92	0.91	0.71	0.70	
	0.77	0.72	0.59	0.73	0.57	0.63	0.62	0.62	0.61	0.66	0.62	0.61	0.61	0.61	0.61
	0.71	0.59</													

A-D, Average AUROC (**A**), AUPRC (**B**), accuracy (**C**) and F1 (**D**) scores of the five-folds of each foundation model on Morphology, Biomarker and Prognosis tasks.

Fig. S2: Average AUROCs sorted by task category, Cancer type and cohort



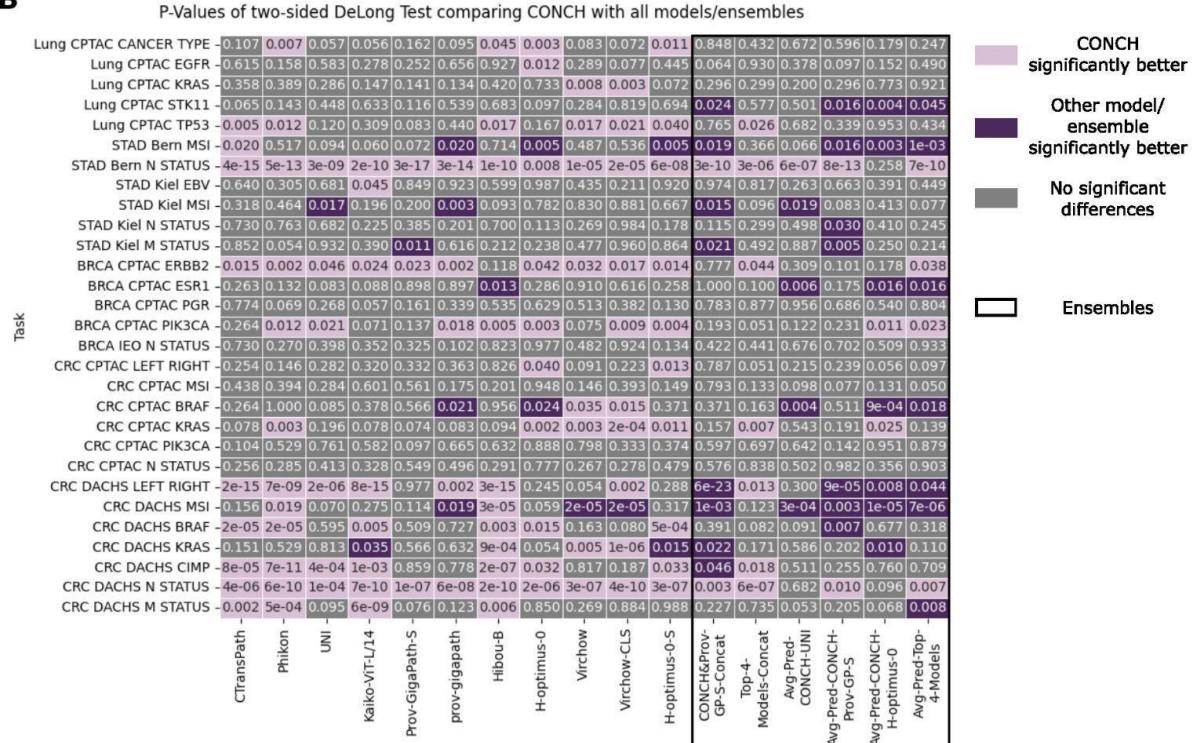
Mean AUROC across five folds for all models. **A**, The 31 tasks were grouped by cancer type (5 tasks for NSCLC, 5 tasks for BRCA, 8 tasks for STAD, 13 tasks for CRC). The final row presents the average of the mean values for each cancer type, thus weighing each cancer type equally. Cancer types are sorted by their mean AUROC across all models, while models are sorted by the Total row. **B**, The 31 tasks were grouped by type of task (5 morphological, 19 molecular, 7 prognostic tasks). The final row presents the average of the mean values for each task type, thus weighing each task type equally. Task types are sorted by their mean AUROC across all models, while models are sorted by the Total row. **C**, The 31 tasks were grouped by external test cohort (5 tasks for CPTAC-LUAD/LUSC, 3 tasks for Bern, 6 tasks for Kiel, 4 tasks for CPTAC-BRCA, 1 task for IEO-BRCA, 7 tasks for CPTAC-CRC, 9 tasks for DACHS-CRC). The final row presents the average of the mean values for each cohort, thus weighing each cohort equally. Cohorts are sorted by their mean AUROC across all models, while models are sorted by the Total row.

Fig. S3: Performance Comparison of Model Ensembles and Single-Model Baselines Using DeLong's Test

A

Task	AUROC Scores by Model (Average)																		
	Lung CPTAC CANCER TYPE	0.99	0.99	0.99	0.99	1.00	1.00	0.98	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.98	0.99
CRC CPTAC MSI	0.92	0.94	0.94	0.95	0.94	0.92	0.92	0.93	0.94	0.95	0.95	0.95	0.88	0.88	0.90	0.90	0.87	0.90	0.90
STAD Kiel EBV	0.89	0.92	0.93	0.92	0.90	0.89	0.89	0.88	0.91	0.90	0.89	0.87	0.86	0.87	0.73	0.89	0.85	0.82	
BRCA CPTAC ESR1	0.85	0.88	0.88	0.88	0.86	0.85	0.88	0.85	0.89	0.88	0.85	0.85	0.90	0.85	0.87	0.88	0.88	0.84	0.81
CRC DACHS MSI	0.86	0.86	0.86	0.87	0.86	0.84	0.86	0.85	0.86	0.85	0.86	0.85	0.77	0.88	0.85	0.82	0.85	0.89	0.80
Lung CPTAC STK11	0.88	0.88	0.82	0.89	0.88	0.80	0.90	0.90	0.73	0.82	0.75	0.82	0.74	0.69	0.83	0.77	0.81	0.67	
Lung CPTAC EGFR	0.83	0.79	0.81	0.81	0.83	0.80	0.74	0.83	0.81	0.80	0.81	0.81	0.77	0.82	0.78	0.78	0.75	0.76	
BRCA CPTAC PGR	0.81	0.82	0.82	0.81	0.81	0.82	0.80	0.78	0.79	0.81	0.78	0.79	0.74	0.80	0.81	0.76	0.76	0.79	0.75
CRC DACHS BRAF	0.79	0.79	0.81	0.81	0.81	0.80	0.76	0.81	0.79	0.77	0.80	0.75	0.78	0.71	0.74	0.74	0.77	0.74	
STAD Bern MSI	0.77	0.79	0.75	0.79	0.77	0.74	0.81	0.78	0.77	0.70	0.78	0.73	0.75	0.69	0.77	0.82	0.75	0.73	
Lung CPTAC TP53	0.80	0.80	0.80	0.79	0.79	0.80	0.75	0.74	0.76	0.76	0.78	0.73	0.73	0.73	0.77	0.73	0.73	0.72	
CRC CPTAC BRAF	0.75	0.82	0.79	0.81	0.75	0.73	0.92	0.77	0.82	0.78	0.82	0.74	0.62	0.79	0.68	0.69	0.58	0.73	
STAD Kiel MSI	0.78	0.75	0.76	0.76	0.76	0.74	0.74	0.77	0.78	0.69	0.79	0.78	0.74	0.71	0.77	0.75	0.73	0.71	
CRC DACHS LEFT RIGHT	0.76	0.73	0.72	0.73	0.73	0.72	0.71	0.72	0.68	0.70	0.69	0.65	0.71	0.66	0.65	0.71	0.69	0.68	
CRC DACHS M STATUS	0.69	0.71	0.70	0.72	0.70	0.70	0.70	0.67	0.67	0.70	0.68	0.67	0.71	0.66	0.61	0.70	0.70	0.65	
CRC DACHS CIMP	0.71	0.70	0.69	0.70	0.70	0.70	0.67	0.70	0.65	0.67	0.69	0.70	0.64	0.66	0.66	0.67	0.68	0.62	
STAD Kiel N STATUS	0.67	0.64	0.66	0.67	0.69	0.65	0.62	0.68	0.64	0.68	0.69	0.66	0.63	0.65	0.69	0.62	0.65	0.65	
CRC DACHS N STATUS	0.68	0.68	0.69	0.68	0.68	0.69	0.65	0.65	0.66	0.65	0.64	0.63	0.65	0.62	0.64	0.63	0.64	0.64	
CRC CPTAC KRAS	0.70	0.67	0.73	0.68	0.69	0.75	0.57	0.64	0.67	0.62	0.67	0.66	0.60	0.67	0.63	0.60	0.56	0.58	
BRCA CPTAC PIK3CA	0.67	0.66	0.67	0.65	0.68	0.69	0.59	0.65	0.62	0.64	0.61	0.58	0.64	0.66	0.57	0.61	0.59	0.62	
CRC CPTAC PIK3CA	0.67	0.65	0.66	0.65	0.62	0.65	0.66	0.57	0.67	0.63	0.62	0.62	0.66	0.64	0.61	0.67	0.68	0.62	
CRC CPTAC N STATUS	0.64	0.67	0.64	0.65	0.65	0.65	0.66	0.63	0.62	0.66	0.62	0.61	0.60	0.60	0.61	0.62	0.60	0.61	
STAD Bern N STATUS	0.65	0.73	0.69	0.64	0.64	0.74	0.68	0.53	0.62	0.63	0.55	0.57	0.65	0.57	0.59	0.60	0.65	0.56	
BRCAPCPTAC ERBB2	0.72	0.69	0.69	0.65	0.68	0.71	0.56	0.60	0.56	0.60	0.52	0.62	0.60	0.59	0.60	0.50	0.59	0.50	
Lung CPTAC KRAS	0.68	0.63	0.67	0.64	0.61	0.64	0.62	0.55	0.68	0.59	0.55	0.69	0.49	0.60	0.55	0.54	0.44	0.69	
CRC CPTAC LEFT RIGHT	0.63	0.60	0.62	0.59	0.62	0.64	0.53	0.60	0.58	0.56	0.60	0.56	0.57	0.58	0.59	0.51	0.59	0.58	
BRCA IEO N STATUS	0.57	0.57	0.57	0.57	0.57	0.57	0.57	0.55	0.56	0.56	0.56	0.54	0.57	0.58	0.57	0.55	0.54	0.57	
CRC DACHS KRAS	0.58	0.58	0.56	0.57	0.56	0.55	0.58	0.56	0.55	0.58	0.56	0.51	0.52	0.54	0.58	0.59	0.49	0.55	
STAD Kiel M STATUS	0.55	0.53	0.51	0.55	0.58	0.51	0.55	0.62	0.51	0.54	0.53	0.57	0.49	0.50	0.48	0.52	0.51	0.59	
Average		0.74	0.74	0.74	0.74	0.74	0.73	0.72	0.72	0.72	0.71	0.71	0.71	0.69	0.69	0.69	0.69	0.68	

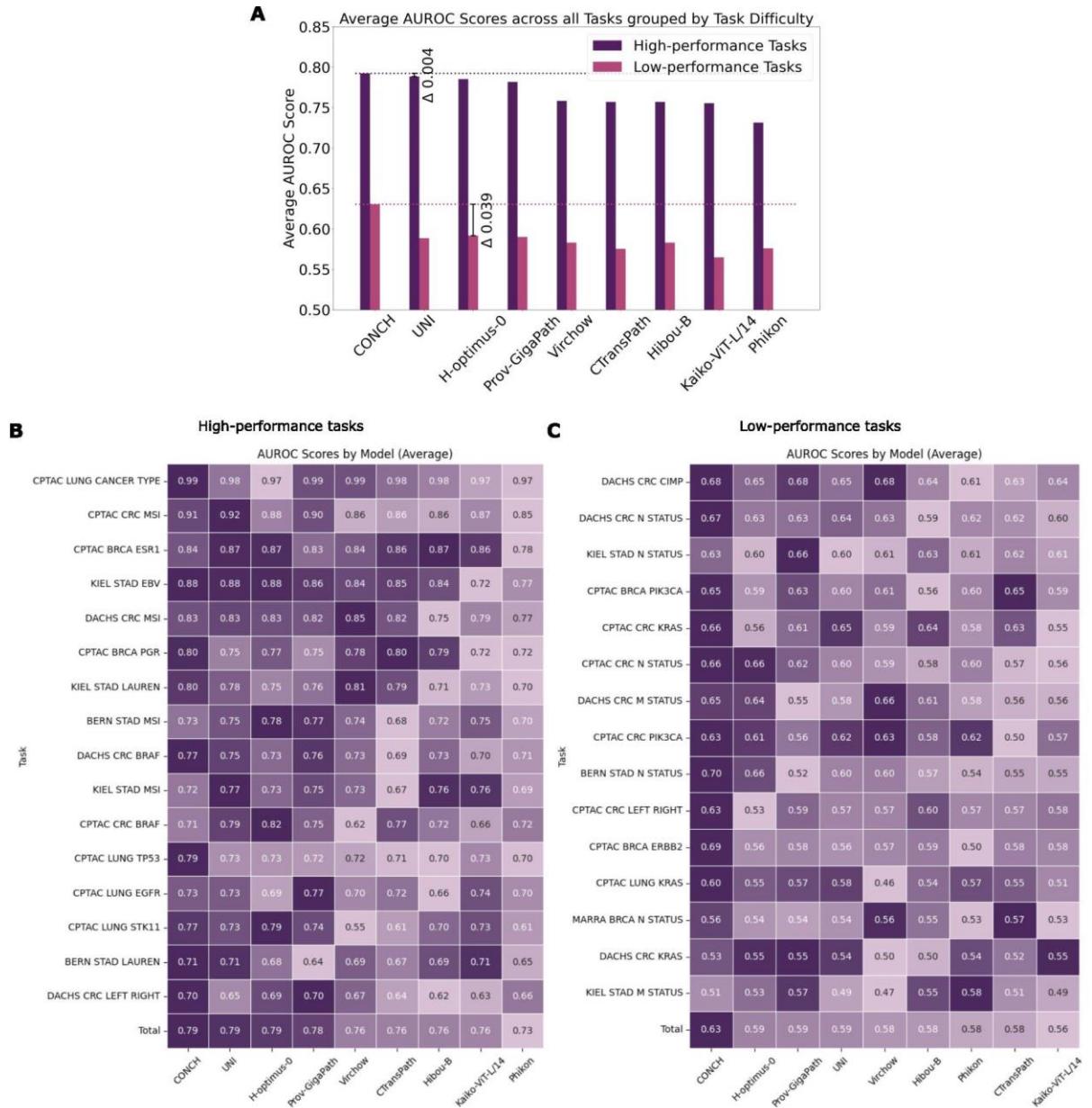
B



only tile embeddings, while Prov-GigaPath-S included both tile embeddings and the slide encoder. For H-optimus-0-S, the tile embeddings from H-optimus-0 were further processed using the GigaPath slide encoder. Virchow-CLS contained only class tokens, and Virchow represents the recommended version by the authors. Two ensembling approaches were utilized: taking the average prediction scores of downstream models trained on different foundation model backbones (prefix Avg-Pred) and concatenating feature vectors from different backbones to create a single downstream model (suffix Concat). The "Lauren" task was excluded as it's not a binary classification.

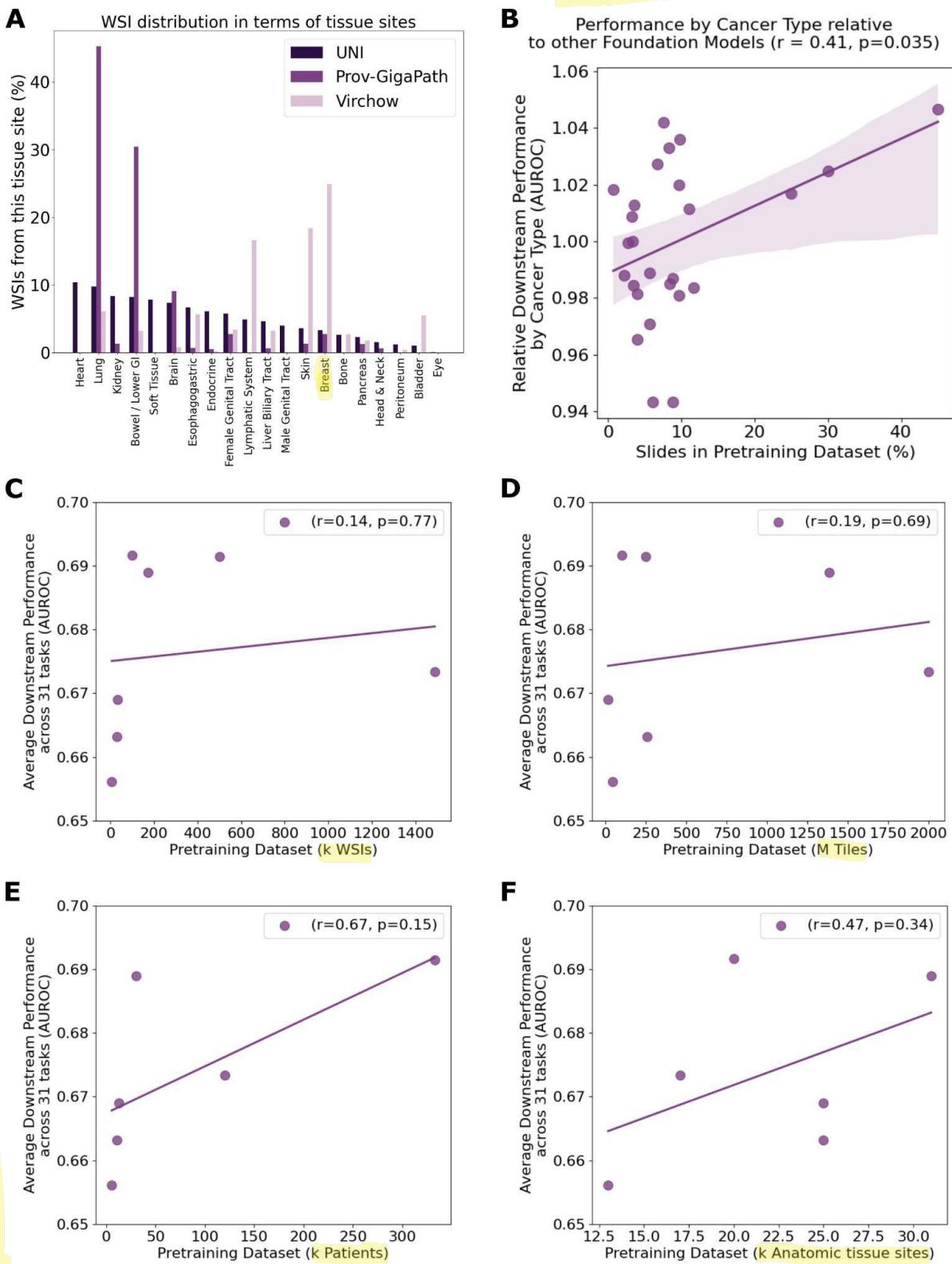
B, P-values from two-sided DeLong's tests comparing CONCH with other models and ensembles. Pink indicates CONCH performed significantly better, purple indicates the other model or ensemble performed significantly better, and grey shows no significant differences. No correction for multiple testing was applied; alpha was set to 0.05.

Fig. S4: High-performance vs. low-performance tasks



A, Average AUROC scores across 16 high-performance and 15 low-performance tasks. Tasks were selected by including only those where at least one foundation model achieved a mean AUROC over 0.70 with a SD less than 0.05 in high-performance and all others in low-performance tasks. **B-C**, The performance of each foundation model is listed. The final row presents the overall average AUROC for each model. Tasks are sorted by their mean AUROC across all models, while models are sorted by their mean AUROC across all tasks.

Fig. S5: Diversity of pretraining datasets



A. Relative number of slides per anatomic tissue site in the pretraining datasets of the three foundation models that provided this data. For Virchow, upper GI and stomach were combined into the esophagogastric category, and endometrium and ovary were combined into the

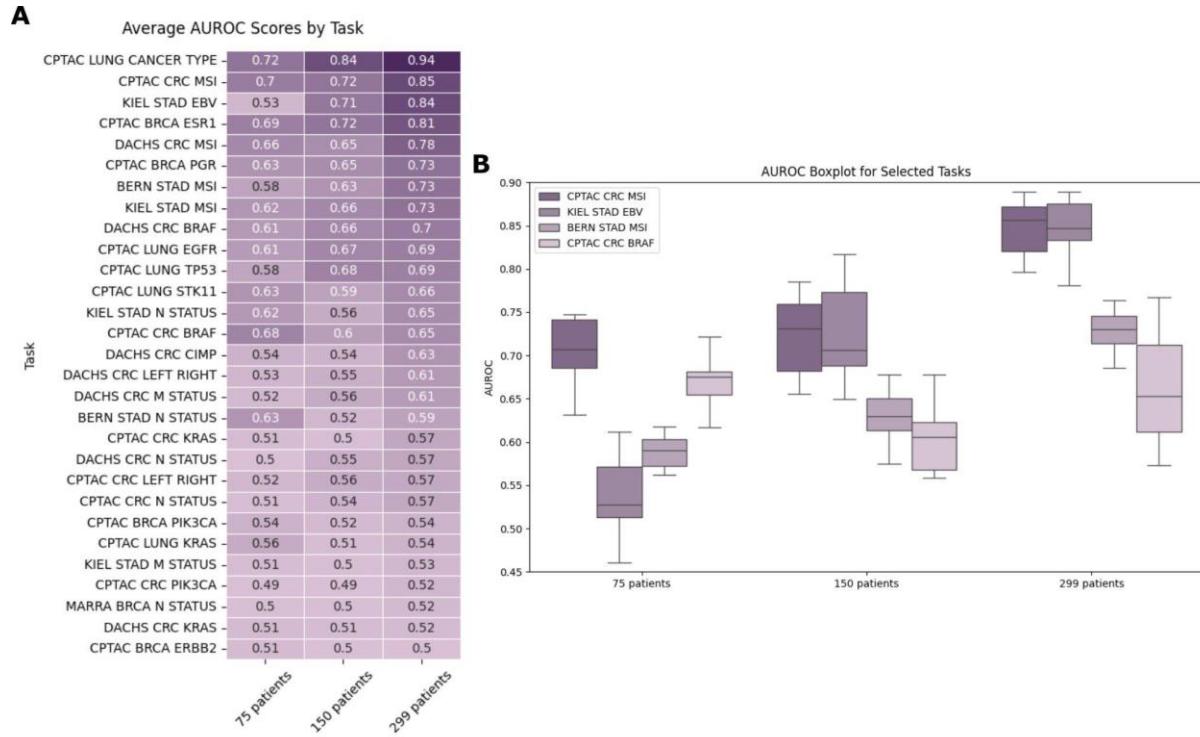
female genital tract category. For Prov-GigaPath, ovary/fallopian tube and uterus were combined into the female genital tract category, and liver and biliary tract were combined. These adjustments were made to match the tissue categories between the models, resulting in the 17 tissue types for Virchow being shown in 15 categories and 15 tissue types for Prov-GigaPath being shown in 13 categories. **B**, Proportion of a tissue type in the pretraining dataset with the relative downstream task performance of the model on tasks of the same tissue type, in comparison to all models on tasks of the same tissue type. Included models are CTransPath, Phikon, UNI, Kaiko, Prov-GigaPath, and Virchow. CONCH was excluded due to lack of information about the dataset composition in the vision-language pretraining, Hibou was excluded as the performance of Hibou-L, which utilized all data, couldn't be tested in this study and H-optimus-0 is excluded due to lack of information about the cancer types in the pretraining dataset. Refer to **Table S5** for detailed information on tissue type proportions. **C-F**, Correlation between the number of WSIs (**C**), Tiles (**D**), patients (**E**) and anatomic tissue sites (**F**) in the pretraining dataset and the average AUROC across all 31 tasks for the models CTransPath, Phikon, Kaiko, UNI, Prov-GigaPath, H-optimus-0, and Virchow.

Fig. S6: Model performance with reduced finetuning dataset

Task	Reduced Training Set for Finetuning (75 patients)										Reduced Training Set for Finetuning (150 patients)										Reduced Training Set for Finetuning (299 patients)									
	CONCH	Prov-GigaPath-S	H-optimus-0	Vrchow	Hibou-B	CtransPath	Kaliko-ViT-L14	Phikon	CONCH	H-optimus-0	Uni	Prov-GigaPath-S	Vrchow	Hibou-B	CtransPath	Phikon	CONCH	Uni	Prov-GigaPath-S	H-optimus-0	Vrchow	CtransPath	Kaliko-ViT-L14	Hibou-B	Phikon					
CPTAC LUNG CANCER TYPE	0.83	0.77	0.80	0.84	0.70	0.71	0.68	0.66	0.54	0.97	0.88	0.85	0.84	0.88	0.78	0.79	0.72	0.81	0.98	0.96	0.96	0.93	0.96	0.94	0.93	0.93	0.90			
CPTAC CRC MSI	0.71	0.74	0.74	0.74	0.69	0.68	0.75	0.60	0.63	0.82	0.76	0.82	0.77	0.69	0.79	0.66	0.59	0.59	0.89	0.89	0.86	0.86	0.85	0.86	0.81	0.82	0.80			
CPTAC BRCA ESR1	0.80	0.77	0.79	0.68	0.62	0.73	0.69	0.57	0.57	0.66	0.78	0.76	0.75	0.73	0.71	0.76	0.67	0.66	0.88	0.89	0.85	0.87	0.85	0.84	0.80	0.83	0.78			
CPTAC CRC BRAF	0.65	0.68	0.72	0.72	0.62	0.68	0.68	0.68	0.67	0.77	0.74	0.82	0.71	0.79	0.54	0.70	0.65	0.70	0.85	0.83	0.82	0.84	0.79	0.84	0.81	0.81	0.72			
DACHS CRC MSI	0.69	0.66	0.64	0.69	0.74	0.62	0.65	0.62	0.66	0.71	0.68	0.69	0.71	0.67	0.67	0.66	0.69	0.68	0.82	0.79	0.79	0.82	0.83	0.76	0.74	0.74	0.74			
CPTAC LUNG STK11	0.66	0.63	0.53	0.70	0.70	0.59	0.61	0.61	0.64	0.71	0.64	0.68	0.66	0.70	0.66	0.65	0.60	0.69	0.80	0.76	0.73	0.76	0.75	0.76	0.69	0.68	0.64			
CPTAC BRCA PGR	0.66	0.69	0.71	0.64	0.60	0.63	0.65	0.54	0.52	0.70	0.69	0.66	0.69	0.70	0.64	0.63	0.61	0.64	0.73	0.75	0.74	0.76	0.71	0.71	0.71	0.69	0.69			
BERN STAD N STATUS	0.72	0.60	0.60	0.62	0.64	0.60	0.61	0.59	0.66	0.69	0.70	0.69	0.66	0.68	0.63	0.65	0.63	0.62	0.74	0.74	0.73	0.75	0.72	0.71	0.73	0.73	0.69			
KIEL STAD MSI	0.62	0.62	0.62	0.64	0.65	0.59	0.59	0.65	0.62	0.81	0.68	0.71	0.69	0.63	0.65	0.66	0.55	0.48	0.76	0.73	0.79	0.82	0.83	0.76	0.74	0.74	0.74			
KIEL STAD N STATUS	0.66	0.63	0.59	0.61	0.66	0.59	0.63	0.58	0.60	0.67	0.72	0.66	0.62	0.70	0.60	0.61	0.61	0.64	0.66	0.72	0.71	0.73	0.64	0.66	0.70	0.71	0.58			
DACHS CRC BRAF	0.62	0.62	0.66	0.59	0.62	0.60	0.60	0.57	0.62	0.65	0.67	0.68	0.63	0.62	0.58	0.65	0.61	0.59	0.69	0.70	0.74	0.65	0.64	0.62	0.58	0.71	0.59			
CPTAC LUNG EGFR	0.63	0.68	0.57	0.60	0.58	0.64	0.62	0.56	0.57	0.57	0.61	0.62	0.60	0.57	0.63	0.56	0.63	0.62	0.67	0.65	0.67	0.63	0.64	0.66	0.65	0.66	0.66			
BERN STAD MSI	0.59	0.56	0.58	0.59	0.61	0.61	0.57	0.62	0.59	0.56	0.60	0.63	0.69	0.49	0.64	0.59	0.49	0.61	0.70	0.73	0.65	0.60	0.65	0.57	0.62	0.60	0.60			
CPTAC LUNG TP53	0.68	0.57	0.58	0.53	0.55	0.57	0.62	0.55	0.54	0.54	0.54	0.56	0.57	0.60	0.59	0.56	0.56	0.56	0.65	0.65	0.67	0.64	0.66	0.64	0.66	0.64	0.64			
CPTAC LUNG KRAS	0.58	0.54	0.57	0.56	0.52	0.61	0.55	0.53	0.59	0.59	0.59	0.53	0.57	0.56	0.53	0.55	0.58	0.55	0.65	0.63	0.65	0.68	0.61	0.62	0.59	0.59	0.61			
DACHS CRC CIMP	0.55	0.57	0.57	0.57	0.58	0.46	0.51	0.55	0.52	0.60	0.57	0.57	0.57	0.58	0.57	0.50	0.56	0.50	0.65	0.63	0.65	0.67	0.64	0.64	0.65	0.65	0.66			
CPTAC BRCA PIK3CA	0.58	0.51	0.54	0.58	0.62	0.55	0.49	0.50	0.49	0.58	0.56	0.56	0.55	0.56	0.55	0.55	0.52	0.53	0.61	0.62	0.62	0.63	0.62	0.57	0.58	0.59	0.59			
DACHS CRC LEFT RIGHT	0.52	0.54	0.52	0.57	0.51	0.53	0.52	0.53	0.52	0.54	0.61	0.55	0.56	0.52	0.53	0.51	0.51	0.56	0.60	0.72	0.73	0.65	0.60	0.65	0.57	0.62	0.60			
KIEL STAD EBV	0.61	0.51	0.58	0.39	0.58	0.56	0.53	0.52	0.46	0.56	0.57	0.53	0.54	0.54	0.52	0.51	0.54	0.55	0.60	0.58	0.59	0.58	0.57	0.53	0.53	0.57	0.57			
CPTAC CRC LEFT RIGHT	0.52	0.50	0.55	0.52	0.53	0.52	0.53	0.52	0.53	0.55	0.60	0.52	0.52	0.53	0.57	0.52	0.51	0.50	0.55	0.62	0.56	0.54	0.57	0.57	0.55	0.54	0.54			
DACHS M STATUS	0.58	0.55	0.50	0.53	0.53	0.50	0.49	0.48	0.50	0.53	0.54	0.50	0.53	0.53	0.53	0.54	0.57	0.57	0.61	0.58	0.58	0.61	0.57	0.60	0.59	0.56	0.56			
DACHS CRC KRAS	0.51	0.51	0.50	0.52	0.51	0.51	0.52	0.51	0.53	0.52	0.54	0.55	0.51	0.49	0.52	0.50	0.54	0.48	0.50	0.52	0.54	0.55	0.57	0.57	0.55	0.57	0.57			
CPTAC BRCA ERBB2	0.43	0.55	0.53	0.48	0.49	0.58	0.53	0.51	0.54	0.50	0.52	0.50	0.51	0.50	0.52	0.51	0.51	0.51	0.54	0.52	0.50	0.53	0.52	0.51	0.51	0.51	0.51			
CPTAC CRC KRAS	0.50	0.54	0.52	0.51	0.52	0.50	0.48	0.52	0.51	0.52	0.52	0.50	0.51	0.50	0.52	0.51	0.51	0.50	0.56	0.58	0.59	0.58	0.57	0.57	0.53	0.53	0.57			
KIEL STAD M STATUS	0.49	0.51	0.51	0.53	0.51	0.54	0.50	0.51	0.49	0.49	0.49	0.54	0.54	0.48	0.50	0.52	0.47	0.52	0.51	0.50	0.56	0.54	0.55	0.54	0.54	0.53				
CPTAC CRC N STATUS	0.48	0.53	0.54	0.49	0.51	0.49	0.49	0.49	0.49	0.48	0.46	0.49	0.57	0.49	0.51	0.48	0.54	0.56	0.44	0.55	0.56	0.58	0.57	0.55	0.58	0.55	0.55			
MARRA BRCA N STATUS	0.51	0.52	0.47	0.51	0.51	0.50	0.50	0.50	0.51	0.51	0.52	0.50	0.49	0.53	0.51	0.48	0.51	0.49	0.49	0.51	0.52	0.50	0.53	0.51	0.51	0.51	0.51			
DACHS CRC N STATUS	0.50	0.51	0.48	0.49	0.50	0.51	0.50	0.50	0.52	0.52	0.44	0.52	0.50	0.50	0.50	0.50	0.48	0.51	0.49	0.52	0.51	0.52	0.50	0.53	0.50	0.52				
CPTAC CRC PIK3CA	0.53	0.52	0.49	0.51	0.51	0.49	0.48	0.50	0.43	0.51	0.47	0.51	0.55	0.46	0.49	0.49	0.40	0.54	0.48	0.51	0.52	0.54	0.50	0.53	0.50	0.52				
Total	0.60	0.59	0.59	0.58	0.58	0.57	0.57	0.55	0.55	0.61	0.61	0.61	0.60	0.60	0.58	0.58	0.57	0.57	0.68	0.67	0.66	0.65	0.64	0.63	0.63	0.62				

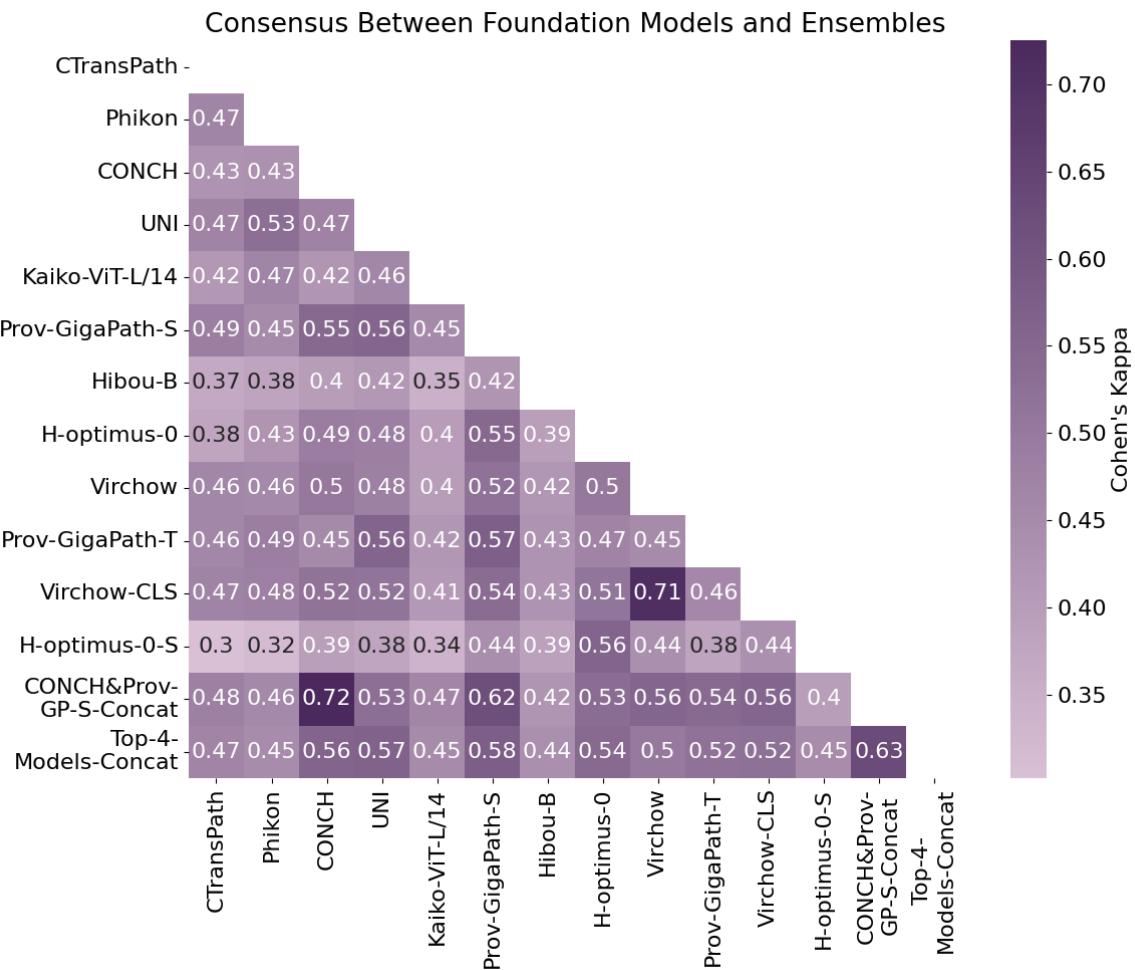
Mean AUROC across all five folds on 29 tasks for all foundation models trained with a reduced downstream dataset of 75 (A), 150 (B), or 299 patients (C). Patients were randomly selected from the TCGA cohorts, ensuring the ground truth was defined for all analyzed tasks. The tasks Lauren in Kiel and Bern were excluded due to insufficient patient numbers.

Fig. S7: Task performance across all models with reduced finetuning dataset



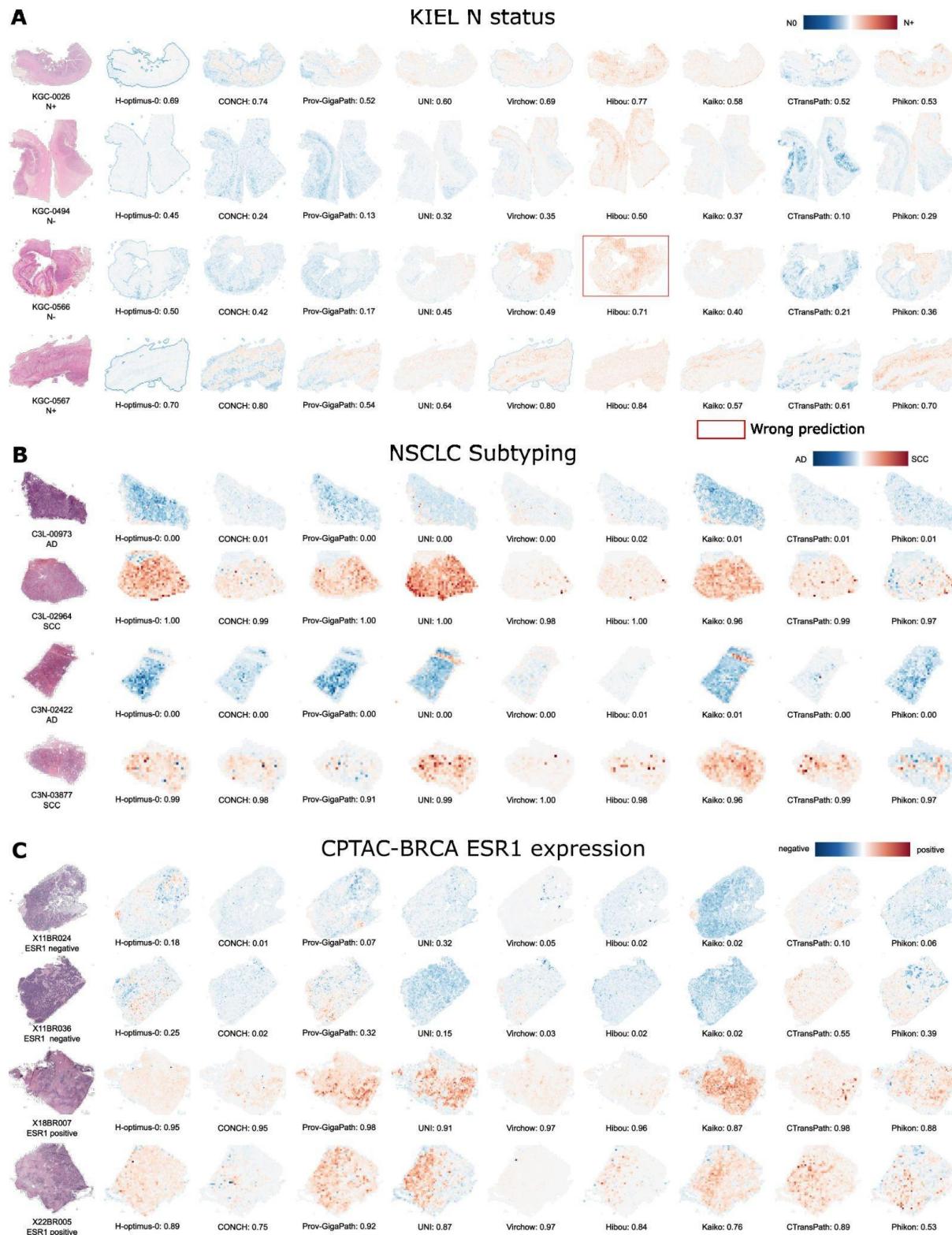
A, Mean AUROC for each task across all foundation models and five folds for models trained on 75, 150, or 299 patients. The mean AUROC values were calculated for each task to evaluate the impact of varying patient sample sizes on different tasks. **B**, Boxplots of four selected tasks showing the varying impact of reduced finetuning datasets on the performance. For EBV status prediction in Kiel for example, the performance is very similar to CPTAC-CRC MSI status prediction when finetuned with 299 patients, but drops to 0.53 when using 75 patients, in contrast to 0.70 for MSI status.

Fig. S8: Cohens kappa scores across all models and ensembles



Objective measure of similarity of prediction scores using Cohen's Kappa and majority vote across the five folds to binarize the predictions. Only tasks where at least one model achieves a mean AUROC over 0.7 are included, with "CPTAC CRC BRAF" excluded due to low majority vote accuracy. All tested versions of foundation models and concatenated features are shown.

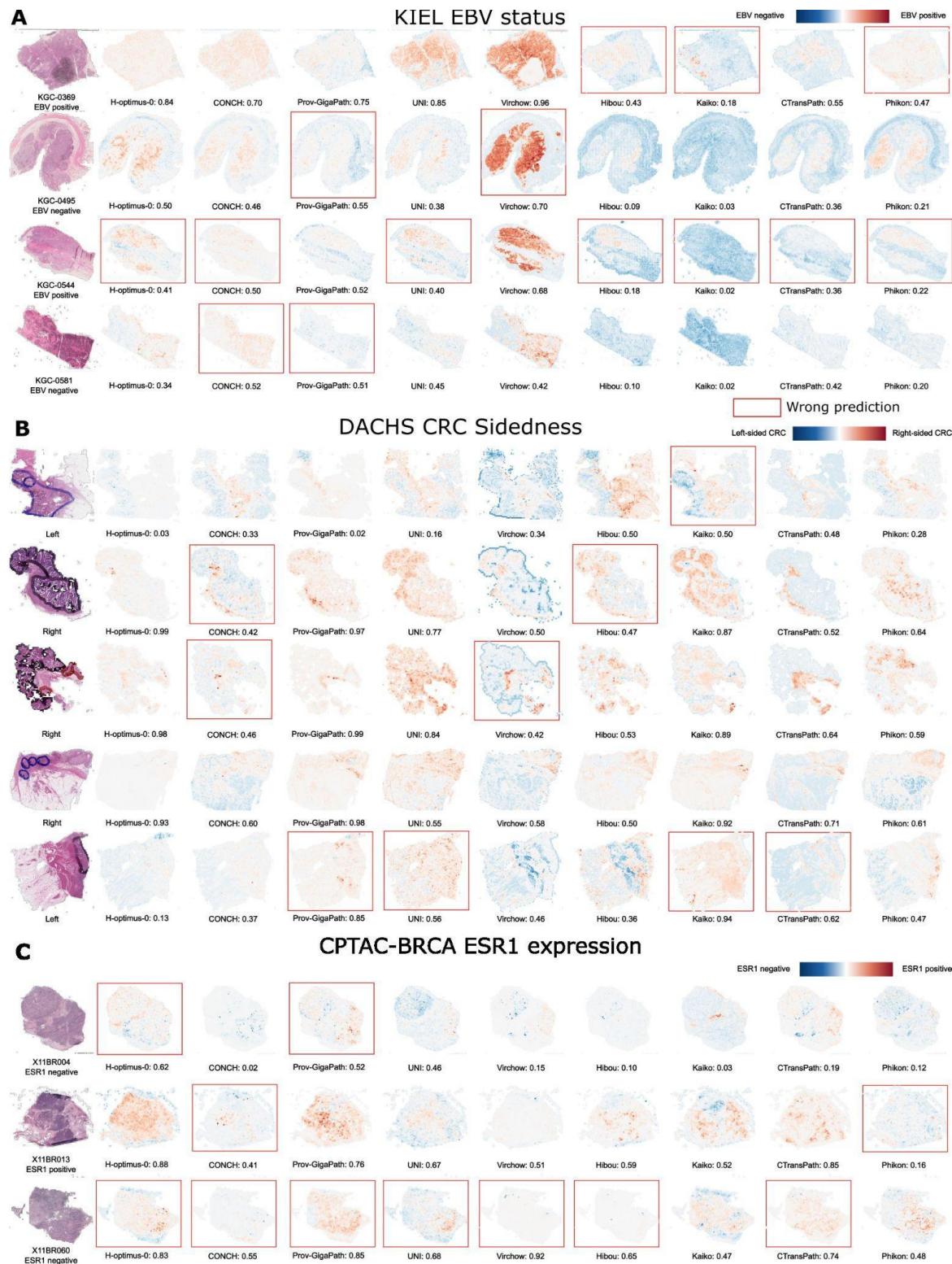
Fig. S9: Attention heatmaps of slides that all models predicted well



A-C, Attention Heatmap Analysis for Kiel N status (**A**), NSCLC subtyping (**B**) and CPTAC-BRCA ESR1 expression (**C**). Classification in four different samples per cohort selected for

correct predictions across almost all foundation models. Thumbnails of the original whole slide images (WSIs) and heatmaps of all foundation models are shown.

Fig. S10: Attention heatmaps of slides with large variations in prediction scores



A-C, Attention Heatmap Analysis for Kiel EBV status (**A**), DACHS CRC sidedness (**B**) and CPTAC-BRCA ESR1 expression (**C**). Classification in 12 different WSIs selected for diverse prediction scores across the foundation models. Thumbnails of the original whole slide images (WSIs) and heatmaps of all foundation models are shown.

Fig. S11: AUROC scores across all foundation models and ensembles

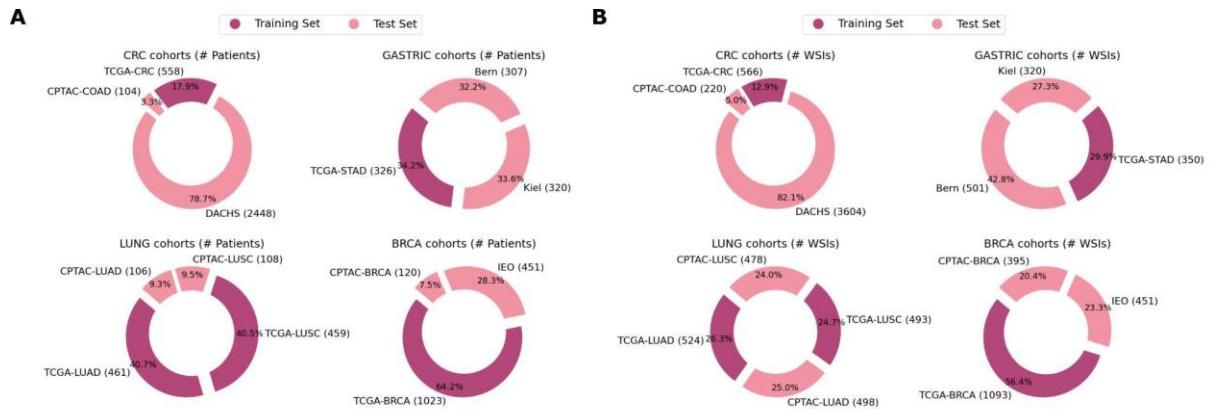
		AUROC Scores by Model (Average)																					
Task	Model	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.97	0.99	0.98	0.98	0.99	0.98	0.99	0.98	0.99	0.98	0.97	0.97			
	Lung CPTAC CANCER TYPE	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.97	0.99	0.98	0.98	0.99	0.98	0.99	0.98	0.99	0.98	0.97	0.97			
	CRC CPTAC MSI	0.93	0.93	0.93	0.92	0.93	0.90	0.91	0.91	0.90	0.88	0.90	0.92	0.92	0.93	0.86	0.86	0.86	0.87	0.83	0.87		
	STAD Kiel EBV	0.90	0.90	0.90	0.88	0.89	0.87	0.88	0.88	0.81	0.88	0.86	0.88	0.84	0.87	0.84	0.84	0.85	0.83	0.88	0.72	0.77	
	BRCA CPTAC ESR1	0.87	0.87	0.87	0.85	0.88	0.87	0.84	0.84	0.85	0.87	0.83	0.87	0.84	0.81	0.75	0.85	0.82	0.86	0.86	0.86	0.78	
	CRC DACHS MSI	0.86	0.85	0.85	0.84	0.86	0.86	0.84	0.83	0.84	0.83	0.82	0.83	0.84	0.81	0.75	0.85	0.82	0.86	0.82	0.79	0.77	
	BRCA CPTAC PGR	0.80	0.80	0.80	0.79	0.81	0.81	0.80	0.80	0.77	0.77	0.75	0.75	0.76	0.79	0.79	0.78	0.80	0.77	0.73	0.72	0.72	
	STAD Bern MSI	0.78	0.75	0.77	0.76	0.77	0.77	0.77	0.73	0.77	0.78	0.77	0.75	0.76	0.69	0.72	0.74	0.68	0.72	0.78	0.75	0.70	
	CRC DACHS BRAF	0.78	0.78	0.77	0.79	0.78	0.77	0.77	0.76	0.73	0.76	0.75	0.77	0.69	0.73	0.73	0.69	0.72	0.71	0.70	0.71	0.70	
	Lung CPTAC STK11	0.86	0.80	0.83	0.81	0.82	0.81	0.80	0.77	0.76	0.79	0.74	0.73	0.80	0.70	0.55	0.61	0.63	0.67	0.73	0.61		
	Lung CPTAC TPS3	0.78	0.78	0.78	0.78	0.76	0.75	0.76	0.79	0.73	0.73	0.72	0.73	0.75	0.73	0.70	0.72	0.71	0.71	0.70	0.73	0.70	
	STAD Kiel MSI	0.75	0.75	0.73	0.74	0.76	0.77	0.77	0.72	0.77	0.73	0.75	0.77	0.77	0.67	0.76	0.73	0.67	0.72	0.74	0.76	0.69	
	Lung CPTAC EGFR	0.76	0.75	0.72	0.77	0.77	0.77	0.78	0.73	0.75	0.69	0.77	0.73	0.76	0.75	0.66	0.70	0.72	0.72	0.70	0.74	0.70	
	BRCA CPTAC BRAF	0.79	0.77	0.78	0.73	0.75	0.78	0.74	0.71	0.77	0.82	0.75	0.79	0.81	0.67	0.72	0.62	0.77	0.58	0.67	0.66	0.72	
	CRC DACHS LEFT RIGHT	0.72	0.71	0.71	0.72	0.71	0.71	0.72	0.70	0.70	0.69	0.70	0.65	0.67	0.68	0.62	0.67	0.64	0.65	0.69	0.63	0.66	
	CRC DACHS CIMP	0.69	0.68	0.68	0.69	0.68	0.69	0.68	0.68	0.68	0.65	0.68	0.65	0.68	0.65	0.64	0.62	0.60	0.59	0.63	0.62	0.62	0.61
	CRC DACHS N STATUS	0.66	0.67	0.66	0.66	0.65	0.65	0.65	0.67	0.64	0.63	0.63	0.64	0.62	0.60	0.59	0.63	0.62	0.62	0.62	0.60	0.62	
	CRC DACHS M STATUS	0.69	0.67	0.68	0.65	0.69	0.68	0.60	0.65	0.68	0.64	0.64	0.55	0.58	0.61	0.62	0.61	0.66	0.56	0.65	0.61	0.56	0.58
	STAD Kiel N STATUS	0.65	0.63	0.63	0.67	0.65	0.65	0.64	0.63	0.63	0.60	0.66	0.60	0.64	0.65	0.63	0.61	0.62	0.63	0.59	0.61		
	BRCA CPTAC PIK3CA	0.65	0.65	0.64	0.65	0.64	0.65	0.65	0.65	0.65	0.59	0.63	0.60	0.59	0.63	0.56	0.61	0.65	0.59	0.59	0.59	0.60	
	CRC CPTAC KRAS	0.65	0.68	0.63	0.65	0.65	0.63	0.64	0.66	0.59	0.56	0.61	0.65	0.63	0.58	0.64	0.59	0.63	0.57	0.59	0.55	0.58	
	CRC CPTAC N STATUS	0.65	0.65	0.67	0.65	0.63	0.62	0.63	0.66	0.61	0.66	0.62	0.60	0.61	0.61	0.58	0.59	0.57	0.59	0.57	0.56	0.60	
	CRC CPTAC PIK3CA	0.64	0.65	0.63	0.62	0.64	0.64	0.65	0.63	0.65	0.61	0.61	0.56	0.60	0.62	0.60	0.58	0.63	0.50	0.61	0.59	0.57	0.62
	STAD Bern N STATUS	0.63	0.67	0.70	0.62	0.60	0.61	0.64	0.70	0.61	0.66	0.52	0.60	0.54	0.62	0.57	0.60	0.55	0.59	0.58	0.55		
	BRCA CPTAC ERBB2	0.64	0.66	0.67	0.67	0.61	0.63	0.68	0.69	0.62	0.56	0.58	0.56	0.51	0.57	0.59	0.57	0.58	0.57	0.51	0.58	0.50	
	CRC CPTAC LEFT RIGHT	0.59	0.61	0.60	0.62	0.59	0.58	0.61	0.63	0.59	0.53	0.59	0.57	0.59	0.57	0.60	0.57	0.57	0.57	0.51	0.58	0.57	
	Lung CPTAC KRAS	0.62	0.61	0.59	0.61	0.58	0.59	0.54	0.60	0.59	0.55	0.57	0.58	0.54	0.55	0.54	0.46	0.55	0.45	0.50	0.51	0.57	
	BRCA IEO N STATUS	0.55	0.56	0.56	0.56	0.56	0.55	0.55	0.56	0.56	0.54	0.54	0.54	0.54	0.55	0.55	0.56	0.57	0.55	0.53	0.53	0.53	
	CRC DACHS KRAS	0.56	0.55	0.55	0.55	0.56	0.55	0.54	0.53	0.55	0.55	0.55	0.54	0.54	0.54	0.50	0.50	0.52	0.49	0.57	0.55	0.54	
	STAD Kiel M STATUS	0.55	0.51	0.53	0.56	0.53	0.56	0.55	0.51	0.55	0.53	0.57	0.49	0.52	0.53	0.55	0.47	0.51	0.50	0.52	0.49	0.58	
	Total	0.72	0.72	0.72	0.72	0.71	0.71	0.71	0.71	0.70	0.69	0.69	0.69	0.69	0.68	0.67	0.67	0.66	0.66	0.66	0.65		

AUROC scores for all foundation models, foundation model variations, and multiple ensemble approaches. For Prov-GigaPath-T, only tile embeddings were used, while Prov-GigaPath-S included both tile embeddings and the slide encoder. For H-optimus-0-S, the H-optimus-0 tile embeddings were further processed in the GigaPath slide encoder. Virchow-CLS contained only class tokens, with Virchow representing the recommended version by the authors.

Two experiments were conducted using concatenated feature vectors: one combining features from CONCH and Prov-GigaPath-S, and another combining features from the top four models (CONCH, H-optimus-0, UNI, and Prov-GigaPath-S).

For each task and slide, mean prediction scores were calculated for various model combinations, including: all models, the model with the prediction score furthest from 0.5 (“most confident model”), the three most confident models, CONCH and UNI, CONCH and Prov-GigaPath-S, CONCH and H-optimus-0, and the combination of CONCH, UNI, Prov-GigaPath-S, and H-optimus-0. These scores were used to evaluate the performance of combined predictions. The combination of the four best foundation models achieved the highest average AUROC of 0.702. As this approach is limited to binary classification, the task “Lauren” was excluded.

Fig. S12: Datasets overview



Composition of all cohorts used in this study and the size comparison of training and test sets for each cancer type. Number of patients (A) and number of WSIs (B) in each cohort. Training was conducted using the TCGA-CRC, TCGA-STAD, TCGA-LUAD, TCGA-LUSC, and TCGA-BRCA cohorts, with TCGA-BRCA being the largest training cohort. Testing was performed using the CPTAC-COAD, CPTAC-LUAD, CPTAC-LUSC, CPTAC-BRCA, DACHS, Bern, Kiel, and IEO cohorts, with DACHS being the largest testing cohort. In total, 6,791 patients and 9,493 slides were used in this study.



Supplementary Tables

Table S1: Analysis of Diversity Metrics in the pretraining datasets

Model	Shannon Entropy	Simpson's Diversity Index	Evenness
UNI	3.4264	0.857	0.707
Prov-GigaPath	1.8714	0.58	0.386
Virchow	2.0481	0.528	0.423

Shannon Entropy: A measure of the uncertainty or randomness in the data distribution. Higher values indicate more diversity.

Simpson's Diversity Index: Measures the probability that two individuals randomly selected from a sample will belong to different categories. Higher values indicate more diversity.

Evenness: Indicates how evenly the data are distributed across categories. It is derived from the Shannon Entropy and ranges from 0 (low evenness) to 1 (high evenness).

These analyses indicate that the UNI model has a more diversified dataset compared to Prov-GigaPath and Virchow, which might have implications for the robustness and generalizability of models trained on these datasets.

Table S2: Models' architecture overview

Name	Released	SSL	Architecture	Pret training Tile size (px)	Patch token size (px)	Magnification	Embed dim	Dataset	Special attributes
CTransPath	Dec 2021	SRC-L	CNN + Swin-Transformer	1024	4	20x	768	TCGA, PAIP	Mean of all tokens as embedding
Phikon	Jul 2023	iBOT	ViT-Base	224	16	20x	768	TCGA	
CONCH	Jul 2023	iBOT + CoCa	ViT-Base	256	16	20x	512	MGH	Vision language model pretraining using 1.2M image-caption pairs
UNI	Aug 2023	DINO v2	ViT-Large	256 & 512	16	20x	1024	BWH, MGH, GTEx	
Virchow	Sep 2023	DINO v2	ViT-Huge	224	14	20x	2560	MSKCC	Mean patch tokens added to the tile embeddings
Kaiko	Mar 2024	DINO v2	ViT-Large	256	14	5x, 10x, 20x, 40x	1024	TCGA	
Prov-Giga Path	May 2024	DINO v2	ViT-Giant	256	14	20x	1536/768	Providence	LongNet slide encoder to further preprocess tile embeddings
Hibou-B	Jun 2024	DINO v2	ViT-Base	?	16	?	768	proprietary	Proprietary Hibou-L model with 1.2B tiles and ViT-L architecture
H-optimus-0	Jul 2024	DINO v2/iBOT	ViT-Giant	224	14	20x	1536	proprietary	
Panakeia	-	?	ViT-Small	224	16	?	384	proprietary	Specific cancer models only for BRCA and CRC

Table S3: Models' pretraining dataset composition

Name	WSI s (K)	Tiles (M)	Patients (K)	Cancer subtypes	Anatomic sites/Organs	malignant WSIs
CTransPath	32	16	~13	32	25	100%
Phikon	6	43	5.6	16	13	100%
CONCH	21	16	?	350	?	?
UNI	100	100	?	?	20	?
Virchow	1488	2000	120	?	17	38%
Kaiko	29	256	11	32	25	100%
Prov-GigaPath	171	1385	30	?	31	?
Hibou-B	1139	512	306	?	?	?
H-optimus-0	500	"several hundreds of millions"	~333	?	?	?
Panakeia-BRCA	5	13	?	?	1	?
Panakeia-CRC	1	4.5	?	?	1	?

Table S4: Patient numbers for individual experiments

CRC				
Marker	Value	Dataset	Cohort	Count
LEFT_RIGHT	left	train	TCGA	242
LEFT_RIGHT	right	train	TCGA	173
isMSIH	nonMSIH	train	TCGA	381
isMSIH	MSIH	train	TCGA	64
BRAF	WT	train	TCGA	458
BRAF	MUT	train	TCGA	61
KRAS	WT	train	TCGA	304
KRAS	MUT	train	TCGA	215
CIMP	nonCIMPH	train	TCGA	390
CIMP	CIMPH	train	TCGA	57
NRAS	WT	train	TCGA	486
NRAS	MUT	train	TCGA	93
PIK3CA	WT	train	TCGA	377
PIK3CA	MUT	train	TCGA	202
N_STATUS	N0	train	TCGA	282
N_STATUS	N+	train	TCGA	221
M_STATUS	M0	train	TCGA	370
M_STATUS	M+	train	TCGA	70
LEFT_RIGHT	left	test	Dachs	1607
LEFT_RIGHT	right	test	Dachs	819
isMSIH	nonMSIH	test	Dachs	1836
isMSIH	MSIH	test	Dachs	210
BRAF	WT	test	Dachs	1930
BRAF	MUT	test	Dachs	151
KRAS	WT	test	Dachs	1397
KRAS	MUT	test	Dachs	677
CIMP	nonCIMPH	test	Dachs	1878
CIMP	CIMPH	test	Dachs	362
N_STATUS	N0	test	Dachs	1295
N_STATUS	N+	test	Dachs	1085
M_STATUS	M0	test	Dachs	1459
M_STATUS	M+	test	Dachs	337
LEFT_RIGHT	right	test	CPTAC	55
LEFT_RIGHT	left	test	CPTAC	47
isMSIH	nonMSIH	test	CPTAC	80
isMSIH	MSIH	test	CPTAC	24
BRAF	WT	test	CPTAC	89
BRAF	MUT	test	CPTAC	15
KRAS	WT	test	CPTAC	70
KRAS	MUT	test	CPTAC	34

PIK3CA	WT	test	CPTAC	82
PIK3CA	MUT	test	CPTAC	22
N_STATUS	N0	test	CPTAC	56
N_STATUS	N+	test	CPTAC	48
Marker	Value	Dataset	Cohort	Count
LAUREN	intestinal	train	TCGA	148
LAUREN	diffuse	train	TCGA	61
LAUREN	mixed	train	TCGA	10
EBV	negative	train	TCGA	301
EBV	positive	train	TCGA	26
isMSIH	nonMSIH	train	TCGA	270
isMSIH	MSIH	train	TCGA	57
N_STATUS	N+	train	TCGA	225
N_STATUS	N0	train	TCGA	98
M_STATUS	M0	train	TCGA	290
M_STATUS	M+	train	TCGA	21
LAUREN	intestinal	test	Bern	172
LAUREN	diffuse	test	Bern	78
LAUREN	mixed	test	Bern	54
isMSIH	nonMSIH	test	Bern	261
isMSIH	MSIH	test	Bern	43
N_STATUS	N+	test	Bern	205
N_STATUS	N0	test	Bern	99
LAUREN	intestinal	test	Kiel	187
LAUREN	diffuse	test	Kiel	75
LAUREN	mixed	test	Kiel	20
EBV	negative	test	Kiel	302
EBV	positive	test	Kiel	18
isMSIH	nonMSIH	test	Kiel	293
isMSIH	MSIH	test	Kiel	27
N_STATUS	N+	test	Kiel	222
N_STATUS	N0	test	Kiel	98
M_STATUS	M0	test	Kiel	259
M_STATUS	M+	test	Kiel	61
Marker	Value	Dataset	Cohort	Count
EGFR	WT	train	TCGA	420
EGFR	MUT	train	TCGA	51
KRAS	WT	train	TCGA	325
KRAS	MUT	train	TCGA	146
STK11	WT	train	TCGA	405
STK11	MUT	train	TCGA	66
TP53	MUT	train	TCGA	245

TP53	WT	train	TCGA	226
EGFR	WT	test	CPTAC	72
EGFR	MUT	test	CPTAC	34
KRAS	WT	test	CPTAC	74
KRAS	MUT	test	CPTAC	32
STK11	WT	test	CPTAC	88
STK11	MUT	test	CPTAC	18
TP53	MUT	test	CPTAC	55
TP53	WT	test	CPTAC	51
NSCLC				
Marker	Value	Dataset	Cohort	Count
CANCER_TYPE	AC	train	TCGA	471
CANCER_TYPE	SCC	train	TCGA	459
CANCER_TYPE	AC	test	CPTAC	106
CANCER_TYPE	SCC	test	CPTAC	108
BRCA				
Marker	Value	Dataset	Cohort	Count
ERBB2	negative	train	TCGA	903
ERBB2	positive	train	TCGA	127
ESR1	positive	train	TCGA	773
ESR1	negative	train	TCGA	257
PGR	positive	train	TCGA	708
PGR	negative	train	TCGA	322
PIK3CA	WT	train	TCGA	699
PIK3CA	MUT	train	TCGA	331
N_STATUS	N+	train	TCGA	548
N_STATUS	N0	train	TCGA	463
ERBB2	negative	test	CPTAC	106
ERBB2	positive	test	CPTAC	14
ESR1	positive	test	CPTAC	79
ESR1	negative	test	CPTAC	41
PGR	positive	test	CPTAC	70
PGR	negative	test	CPTAC	50
PIK3CA	WT	test	CPTAC	81
PIK3CA	MUT	test	CPTAC	39
N_STATUS	N+	test	IEO	247
N_STATUS	N0	test	IEO	208

Table S5: Proportion of analyzed tissue types in the pretraining data

Foundation Model	Number represents	Lung	Breast	Stomach	Colon
Prov-GigaPath	Tissue slides	45%	2.7%	0.7%	30%
CONCH	Image-text pairs	103k (9.5%)	65k (5.6%)	121k (10.4%)	
UNI	Tissue slides	9846 (9.8%)	3364 (3.3%)	6705 (6.7%)	8303 (8.3%)
Hibou	Tissue slides in total 112.5k estimated	2.5k (2.2%)	12k (11%)	35k (31%)	
Virchow	Tissue slides	6.1%	25%	3.5%	3.2%
Phikon, Kaiko (TCGA)	Patients	1089 (9.7%)	979 (8.8%)	443 (4.0%)	633 (5.7%)
CTransPath	Patients	1089 (8.4%)	979 (7.5%)	443 (3.4%)	1533 (11.7%)

No information available for H-optimus-0.



5. Supplementary methods

5.1 Description of foundation models

CTransPath, introduced by Wang et al. in December 2021, is the pioneering Transformer-based unsupervised feature extractor for histopathological images. It integrates a convolutional neural network with a multi-scale Swin Transformer architecture, trained on 15 million patches from 32 cancer subtypes using semantically-relevant contrastive learning, a framework introduced in the same paper based on MoCo v3⁴⁵. The model was evaluated across multiple tasks, including patch retrieval, patch classification, whole-slide image (WSI) classification, mitosis detection, and colorectal adenocarcinoma gland segmentation²⁷.

Phikon, published by Filion et al. in July 2023, employs the iBOT framework (image BERT pre-training with Online Tokenizer) for SSL, using MIM and self-distillation^{46,47}. The architecture is a ViT-Base model with 80 million parameters, trained on 6,093 WSIs from 16 cancer types, comprising 43 million patches. Phikon was assessed on tile-level and slide-level tasks for subtype, genomic alteration, and overall survival prediction¹⁶.

CONCH, released by Lu et al. in July 2023, is a vision-language model based on CoCa⁴⁸, which pretrains an image-text encoder-decoder model using contrastive and captioning losses. For this analysis, only the image encoder was considered. It was pretrained on 16 million image tiles from 21,442 WSIs covering over 350 cancer subtypes. Unlike CTransPath and Phikon, CONCH was trained on proprietary datasets rather than public ones like TCGA and PAIP. The vision-language model is trained by seeking to align image and text modalities in the model's representation space and by predicting the caption corresponding to an image. For this vision-language pretraining, over 1.1 million image-text pairs were used mainly taken from publicly available research articles. Its performance was evaluated on various subtyping, tissue classification, and grading tasks⁴⁹.

UNI, introduced by Chen et al. in August 2023, is notable for being the first model trained on over 100,000 slides. It utilizes DINOv2 for pretraining, incorporating MIM and self-distillation⁵⁰. The training dataset, Mass-100K, was collected from Massachusetts General Hospital and Brigham and Women's Hospital and consists of over 100 million tissue patches from 20 major tissue types. UNI was tested on the challenging 43-class OncoTree cancer type classification and 108-class OncoTree code classification tasks²¹.

Virchow, introduced by Vorontsov et al. in September 2023, stands out as the model trained on the largest dataset to date, with 1.5 million slides from Memorial Sloan Kettering Cancer Center. The model employs DINOv2 for pretraining and features a ViT-Huge architecture with 632 million parameters. Unique to Virchow, the final tile embedding is created using both class tokens and mean patch tokens, doubling the embedding dimension to 2560. It was evaluated on tissue classification and biomarker prediction tasks²³.

Kaiko.ai released a series of pretrained foundation model based on ViT and DINO/DINOv2⁵¹. Among these, the ViT-L14 model, trained with DINOv2, was tested in this study. Interestingly, in their own tests, the ViT-B8 trained on DINO performed better on some of the test sets despite using the older SSL method and the smaller ViT-Base architecture. This was attributed

to the reduced patch size of eight in comparison to 14 of the ViT-Large. Unlike other recently published foundation models, Kaiko.ai's models were trained on a relatively modest dataset of 29,000 WSIs from TCGA. The performance of these models was assessed on tissue classification tasks using five different datasets⁵².

Prov-GigaPath, published by Xu et al. in May 2024, employs a two-stage pretraining approach. Initially, the tile encoder, a ViT-Giant model, is pretrained using DINOv2. Subsequently, a slide encoder contextualizes each tile on the WSI using a LongNet model⁵³. Prov-GigaPath was trained on 1.3 billion patches from 170,000 WSIs, sourced from 28 cancer centers within the Providence health network. These WSIs represent over 30,000 patients and encompass 31 major tissue types. Prov-GigaPath was evaluated on mutation prediction using 18 pan-cancer biomarkers and on cancer subtyping tasks, utilizing both TCGA and Providence datasets²².

Hibou, developed by Nechaev et al. and released in June 2024, includes two versions: Hibou-B and Hibou-L. The public version, Hibou-B, utilizes a ViT-Base architecture with 86 million parameters and was trained on 510 million tiles. The proprietary version, Hibou-L, employs a ViT-L architecture trained on 1.2 billion tiles. The training data includes 936,441 H&E and 202,464 non-H&E stained slides from 306,400 individual cases including veterinary biopsies and cytology slides. Due to the unavailability of Hibou-L weights, only Hibou-B was included in this study. The performance of Hibou-B was assessed using six datasets for patch-level benchmarks, including tasks such as tissue classification, detection of tumor-infiltrating lymphocytes, and mutation prediction. Additionally, three datasets were used for slide-level benchmarks, focusing on tissue classification⁵⁴.

lowest?

The final model, **H-optimus-0**, was released on GitHub in July 2024 by Saillard et al. It was trained on a proprietary dataset comprising over 500,000 WSIs, from which hundreds of millions of tiles were extracted. Notably, H-optimus-0 features the highest number of WSIs per patient, with an average of 1.5 slides per patient, compared to other models in this study. For instance, Hibou has 3.7 slides per patient, Prov-GigaPath has 5.7 slides per patient, and Virchow has 12.4 slides per patient. H-optimus-0 employs a ViT-Giant architecture with a patch size of 14 and four registers⁵⁵. The model was evaluated on tile-level tissue classification tasks and slide-level biomarker or metastasis prediction tasks⁵⁶.

The collection of histopathology foundation models analyzed in this study is not exhaustive. Additional models published in the past year include BEPH⁵⁷, PLUTO⁵⁸, RudolfV⁵⁹, PathoDuet⁶⁰, and models by Campanella et al.⁶¹. However, these models are either not publicly accessible (PLUTO, RudolfV, Campanella et al.) or have been trained exclusively on TCGA data using the ViT-Base architecture, which renders them less competitive compared to the more recent models evaluated in this study. Furthermore, the publications associated with these models do not offer comparisons with the latest foundation models.

Comparison of foundation models

5.2

In the case of Prov-GigaPath, Xu et al. introduced a slide encoder aimed at analyzing global patterns in WSIs. Previous benchmarking efforts and comparisons by the authors of other foundation models did not evaluate Prov-GigaPath using both tile and slide encoders^{54,56,62}. Consequently, we deemed it beneficial to include both versions in this benchmarking study.

The results indicate that incorporating the slide encoder does not enhance the performance of the Prov-GigaPath model within a pipeline like STAMP¹⁹. This is likely because the aggregator model is capable of comprehending slide-level patterns as effectively as the LongNet model. Thus, it appears feasible to use only the tile encoder in a setup like ours. However, using the slide encoder remains beneficial for reducing the dimensionality of feature vectors from 1536 to 768 without compromising performance, suggesting that further compressing vectors of other foundation models may be possible without losing relevant information for the classification tasks analyzed. When testing the GigaPath LongNet model on tile embeddings generated by H-optimus-0, a slight performance reduction was observed, which is unsurprising given the LongNet model's pretraining context. Nonetheless, its performance was still comparable to that of Hibou-B and Kaiko ViT-L14. For Virchow, Vorontsov et al. recommended utilizing both the class token and the mean patch token to create the final tile embedding²³. This approach doubles the memory requirements for the feature vectors but does not improve performance in our setup. Therefore, it is sufficient to use only the class tokens for Virchow, as is standard practice with the other foundation models in this study (**Add. Figure 1**).

CONCH demonstrates exceptional performance despite its relatively modest size as a ViT-Base model. Notably, the vision encoder underwent pretraining on 21,000 WSIs, followed by training the foundation model on over 1.1 million image-text pairs, thus leveraging extensive high-quality training data. Given that the WSIs used for both CONCH and UNI originate from the same source, it is unlikely that there is a qualitative difference between them. The fact that CONCH outperforms UNI, despite being trained on fewer WSIs and utilizing a smaller architecture, underscores the effectiveness of the vision-language approach. These findings suggest that future advancements in histopathological feature extraction may benefit more from the strategic combination of modalities and the utilization of high-quality data rather than merely scaling up model size and data quantity.

We observed a marked difference in performance among the pure vision models, with UNI achieving the highest performance, closely followed by H-optimus-0 and Prov-GigaPath. Surprisingly, Virchow, despite being trained on the largest dataset in terms of WSIs and tiles, underperformed relative to the other models in Biomarkers tasks. A detailed analysis of the pretraining dataset attributes and their correlation with model performance revealed that the number of patients and the diversity of anatomic sites were more influential than the sheer quantity of data (**Figure 3A**). This suggests that data diversity is a critical factor for enhancing model performance, overshadowing the benefits of data volume alone. This trend is particularly evident when comparing the UNI and Virchow models. UNI, with its training dataset comprising 100k WSIs and a smaller ViT-Large architecture, surpassed Virchow, which utilized a much larger dataset of 1488k WSIs and a ViT-Giant architecture (Tables S1 & S2). Although Virchow included a greater number of patients than UNI (120k patients, UNI only included 100k WSIs, the patient number is unknown), UNI's dataset is more diverse in terms of tissue type distribution (**Figure S5**). The differences in data diversity are further underscored by various diversity metrics assessed for these models (Table S1). Unfortunately, comparable detailed information for the remaining models was unavailable, preventing a comprehensive comparison.

Model Ensembles

Combining the prediction scores of different models and concatenating their feature vectors yielded modest performance improvements compared to the best individual models. Heatmap analyses revealed that different models focus on distinct tissue regions and interpret WSIs differently, suggesting potential benefits in leveraging the strengths of multiple foundation models. The approaches explored in this study, however, were relatively rudimentary. Concatenating feature vectors likely introduced redundancy and created excessively long feature vectors, thereby increasing the risk of overfitting, as per Bellman's curse of dimensionality⁶³. This might explain why combining the four best feature vectors resulted in inferior performance compared to CONCH alone. Therefore, it would be interesting to find ways of merging the models without increasing the feature space. To enhance model combination strategies without expanding the feature space, future research could explore dimensionality reduction techniques or the integration of foundation models into a unified framework using merging strategies⁶⁴.

Ablation studies

The challenge of overfitting in machine learning models can often be mitigated by increasing the size of the training cohort. Exclusively for this experiment, we leveraged the DACHS dataset for downstream training, which enabled us to utilize up to 1700 patients across six different tasks. We conducted a 5-fold cross-validation on DACHS and evaluated the models on 11 tasks derived from the TCGA and CPTAC datasets. It is worth noting that models such as CTransPath, Phikon, and Kaiko might possess an inherent advantage in this experimental setup due to their pretraining on TCGA data. Our experiments involved varying the training cohort sizes (100, 200, 400, 850, and 1700 patients) to investigate whether larger embedding vectors yield improved performance with larger training cohorts. We correlated the embedding dimension of each foundation model with the mean AUROC across all tasks and five folds. Contrary to our initial hypothesis, the size of the model's embedding vectors did not consistently influence performance in a straightforward manner. Specifically, smaller virchow-class vectors (1280 dimensions) underperformed compared to virchow vectors (2560 dimensions) with a smaller number of patients. However, this performance disparity diminished when the models were trained on cohorts of 850 or 1700 patients. Similarly, Prov-GigaPath-Slide vectors (768 dimensions) performed worse with smaller patient counts compared to Prov-GigaPath vectors (1536 dimensions), but their performance converged when the full patient cohort was used. A noteworthy observation is that the CONCH model exhibited superior performance with limited training data compared to other models. However, this advantage dissipated as the size of the training cohort increased (Add. Figure 2).

Data diversity in foundation models

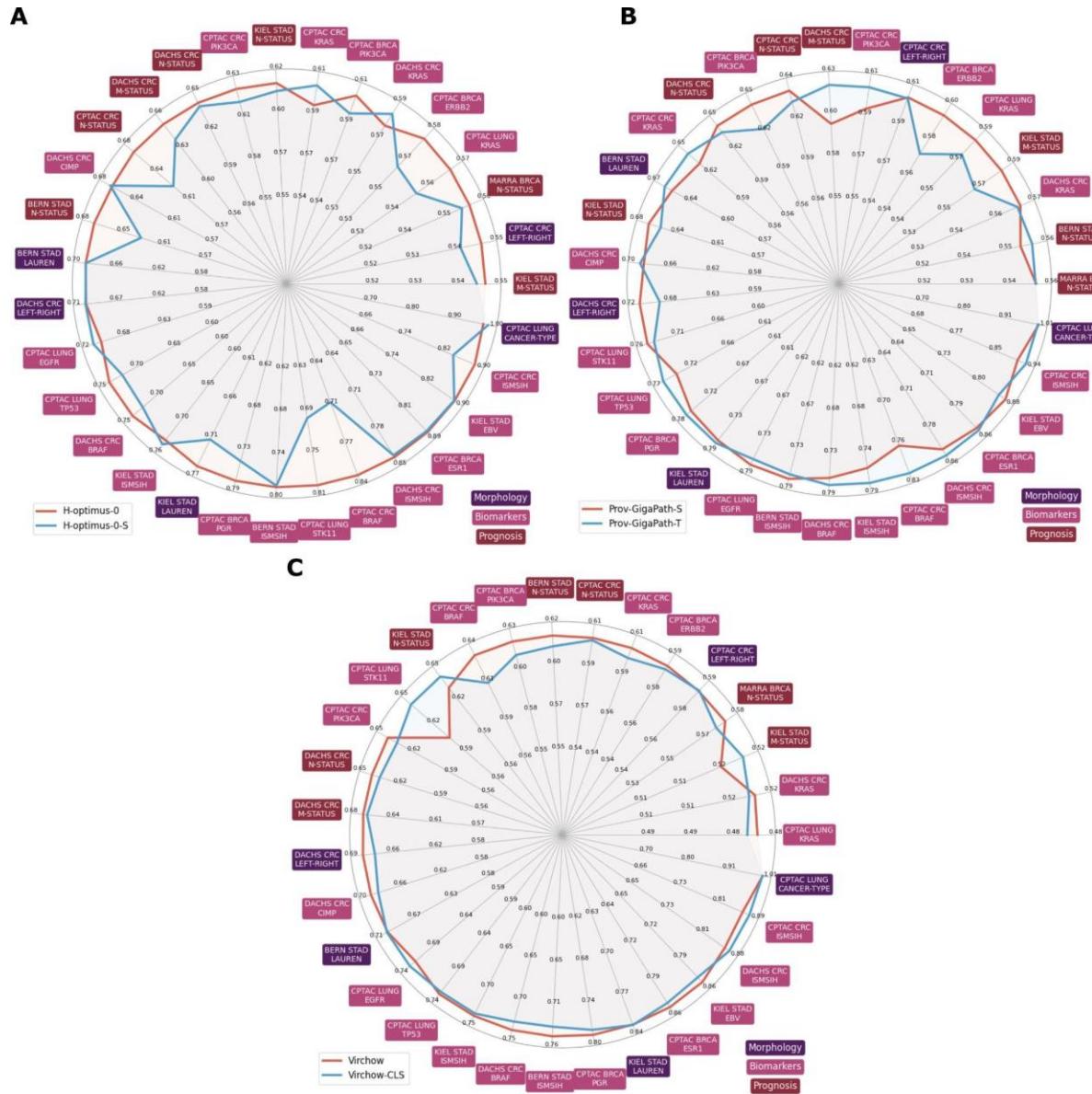
An obvious difference between the foundation models lies in the composition of their training data (Table S5). For instance, Virchow's training data consisted of 25% breast tissue, 18.4% skin, and only 6.1% lung. In contrast, UNI's training data predominantly comprised heart and lung tissues, with less than half as many skin and breast cases. Prov-GigaPath's dataset was

45% lung tissue slides, 30% bowel tissue, and only 2.76% breast tissue. Lastly, CONCH placed greater emphasis on GI and lung tissues, with approximately half as much weight given to breast tissue. Given this variability, Virchow's underperformance in lung cases compared to UNI and Prov-GigaPath may reflect the relatively small proportion of lung cases in its training data. However, the fact that CTransPath outperformed Phikon only in BRCA tasks, despite both being trained on the same BRCA cases from TCGA (as PAIP lacks breast tissue), is contrary to this logic. Overall, there is a moderate correlation ($r = 0.41$) between the number of WSIs of a specific tissue type in the pretraining data and relative performance in downstream tasks involving the same tissue type compared to the average performance of all models in the same tasks (Fig S5). While this correlation is not strong, it underscores the importance of considering tissue type diversity when benchmarking foundation models for histopathology.

Due to relying on the STAMP protocol, it was not feasible to include regression or tissue segmentation tasks, which were often part of the original studies for the foundation models. Additionally, expanding the analysis to include more cancer types would be beneficial, as there are noticeable differences in model performance across various cancers. However, given that CONCH consistently performs best across all analyzed cancer types, it is likely a strong candidate for other tasks as well.

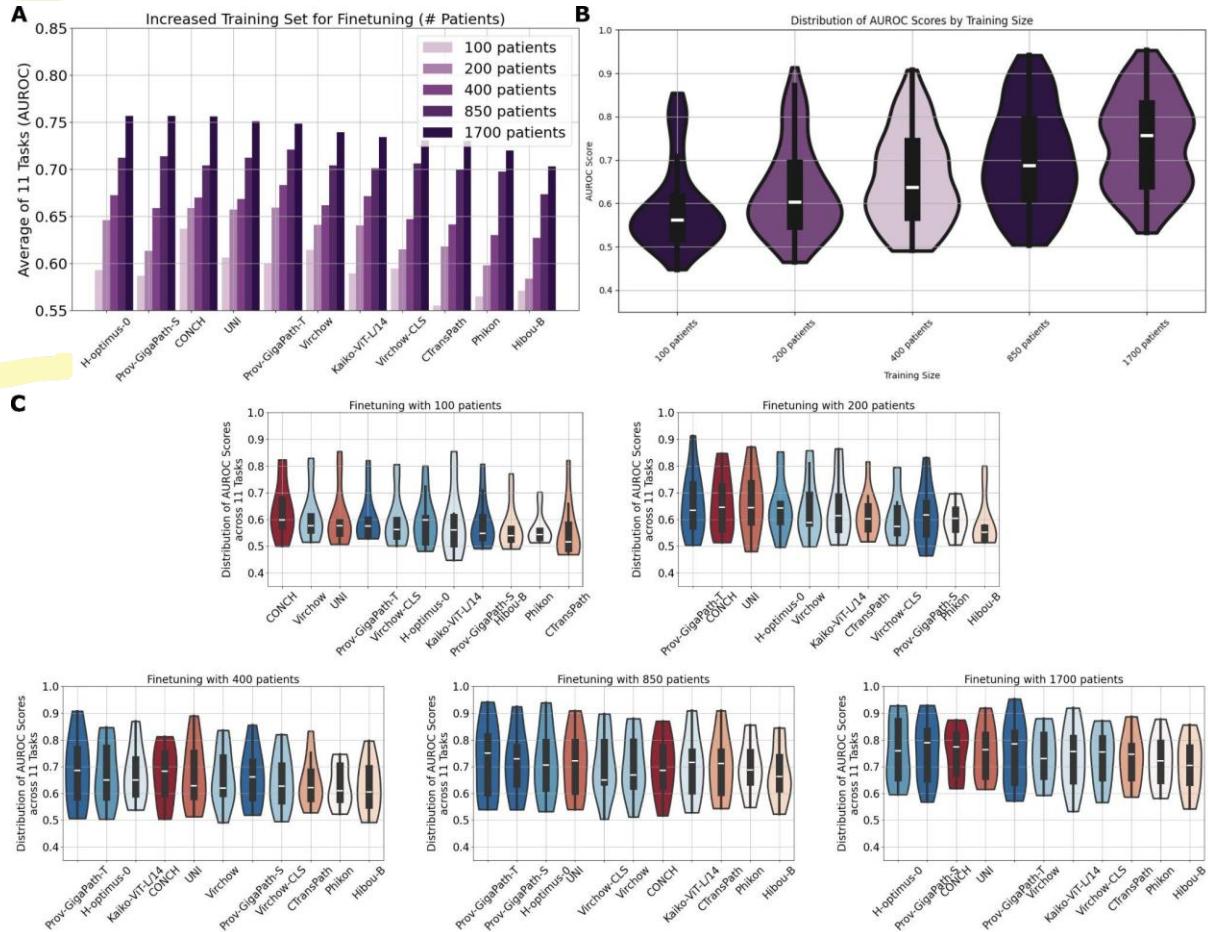
Figures for supplementary methods

Add. Fig. 1: Comparison of alternative foundation model versions for Prov-GigaPath, Virchow and H-optimus-0



A, AUROC scores of H-optimus-0 (red) vs H-optimus-0 vectors further processed in the Prov-GigaPath Slide Encoder (blue). **B**, AUROC scores of Prov-GigaPath with Slide Encoder (red) vs without Slide Encoder (blue). **C**, AUROC scores of Virchow with CLS and mean patch tokens (red) vs only CLS tokens. Red always shows the recommended version. All plots are normalized taskwise.

Add. Fig. 2: Experiments with increased finetuning dataset sizes using DACHS as training cohort



A, Average AUROC across five folds on 11 tasks for models trained on a finetuning dataset with 100, 200, 400, 850, or 1700 patients. **B**, Distribution of AUROC scores from all foundation models grouped by finetuning dataset size. **C**, Distribution of AUROC scores for each foundation model individually. For Prov-GigaPath-T, the tile embeddings were used, for Prov-GigaPath-S, the slide encoder was also included. Virchow-CLS only contained the class tokens, Virchow is the version recommended by the authors. Patients were randomly selected from the DACHS cohort, ensuring ground truth was defined for all analyzed tasks. The task "M-Status," was excluded due to insufficient patient numbers. The models were deployed using the CPTAC-CRC cohort and, unlike other experiments, also included the TCGA-CRC cohort. Consequently, Kaiko, CTransPath, and Phikon models might have an advantage as they had prior exposure to TCGA data during pretraining.