nature medicine

Article

https://doi.org/10.1038/s41591-024-02856-4

A visual-language foundation model for computational pathology

Received: 2 August 2023

Accepted: 5 February 2024

Published online: 19 March 2024

Check for updates

Ming Y. Lu[1,2,3,4,5,11], Bowen Chen[1,2,11], Drew F. K. Williamson[1,2,3,11], Richard J. Chen[1,2,3,4,6], Ivy Liang[1,7], Tong Ding[1,7], Guillaume Jaume[1,2,3,4], Igor Odintsov[1], Long Phi Le[2], Georg Gerber[1], Anil V. Parwani[8], Andrew Zhang[1,2,3,4,9] & Faisal Mahmood[1,2,3,4,10] ✉

The accelerated adoption of digital pathology and advances in deep learning have enabled the development of robust models for various pathology tasks across a diverse array of diseases and patient cohorts. However, model training is often difficult due to label scarcity in the medical domain, and a model's usage is limited by the specific task and disease for which it is trained. Additionally, most models in histopathology leverage only image data, a stark contrast to how humans teach each other and reason about histopathologic entities. We introduce CONtrastive learning from Captions for Histopathology (CONCH), a visual-language foundation model developed using diverse sources of histopathology images, biomedical text and, notably, over 1.17 million image–caption pairs through task-agnostic pretraining. Evaluated on a suite of 14 diverse benchmarks, CONCH can be transferred to a wide range of downstream tasks involving histopathology images and/or text, achieving state-of-the-art performance on histology image classification, segmentation, captioning, and text-to-image and image-to-text retrieval. CONCH represents a substantial leap over concurrent visual-language pretrained systems for histopathology, with the potential to directly facilitate a wide array of machine learning-based workflows requiring minimal or no further supervised fine-tuning.

The gold standard for the diagnosis of many diseases remains the examination of tissue by a pathologist. The recent rise of computational pathology[1–4], which leverages artificial intelligence (AI) to solve problems in pathology, has demonstrated considerable advances across many tasks, including metastasis detection[5], cancer subtyping[6,7], survival prediction[8–10], unknown primary origin site prediction[11,12], image search[13–16] and prediction of molecular alterations[17,18], among other tasks[19]. Additionally, current strides in the field are made under the paradigm of developing models targeting specific tasks using large cohorts of labeled training examples, such as in lymph node metastasis detection[20] and prostate cancer grading[21,22]. However, the process of data collection and annotation of whole-slide images (WSIs) is labor intensive and is not scalable to open-set recognition problems or rare diseases, both of which are common to the practice of pathology. With thousands of possible diagnoses and many other tasks, training separate models for every step of the pathology workflow is untenable.

[1]Department of Pathology, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA. [2]Department of Pathology, Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA. [3]Cancer Program, Broad Institute of Harvard and MIT, Cambridge, MA, USA. [4]Cancer Data Science Program, Dana-Farber Cancer Institute, Boston, MA, USA. [5]Electrical Engineering and Computer Science, Massachusetts Institute of Technology (MIT), Cambridge, MA, USA. [6]Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA. [7]Harvard John A. Paulson School of Engineering and Applied Sciences, Harvard University, Cambridge, MA, USA. [8]Department of Pathology, Wexner Medical Center, Ohio State University, Columbus, OH, USA. [9]Health Sciences and Technology, Harvard-MIT, Cambridge, MA, USA. [10]Harvard Data Science Initiative, Harvard University, Cambridge, MA, USA. [11]These authors contributed equally: Ming Y. Lu, Bowen Chen, Drew F. K. Williamson. ✉e-mail: faisalmahmood@bwh.harvard.edu

Additionally, as diverse as these tasks are, they are all analyses of visual data or include other structured information such as 'omics' (refs. 23–26) and other multimodal data sources[27–29]. However, the practice of pathology and the communication of pathological findings make extensive use of natural language, be it in the form of the report that the pathologist prepares for the patient and their treating clinician, the journal article that details a new histopathologic entity or the textbook chapter that teaches residents how to practice pathology.

The general machine learning community has made immense strides in foundation models that use both visual and language information. Representative tools such as CLIP[30], ALIGN[31] and CoCa[32], among others[33–38], use large-scale image–caption pairs[39] to pretrain visual-language foundation models—task-agnostic pretrained models that demonstrate robust performance in downstream vision and visual-language tasks. In the broader biomedical imaging domain, visual-language data have been leveraged for a variety of tasks, including X-ray report generation[40,41], zero-shot classification[42–45] and retrieval[45–48], among others[49–53]. However, the number of studies integrating vision and language data for representation learning in computational pathology is small, with recent studies[44,54–58] demonstrating the potential of using paired image–caption data to learn meaningful visual representations and to develop foundation models for histopathology that can be transferred to multiple downstream tasks in a zero-shot setting, that is, using no task-specific training data. However, these studies[44,54,56] were limited in the scale of histopathology-specific pretraining data due to the lack of readily available image–caption pairs in this domain, leading to limited practical utility from relatively poor performance. Additionally, the broader capabilities of these models remain underexplored.

Given the diversity of tasks, the difficulty in acquiring large datasets of rare diseases or combinations of findings, and the central nature of language to the practice of pathology, there is a need for (1) high-performing visual-language foundation models that leverage large-scale pretraining and generalize well across tasks; and (2) extensive studies on the wide range of potential applications of these models to understand their utility and limitations. We introduce CONtrastive learning from Captions for Histopathology (CONCH), a visual-language foundation model developed using diverse sources of histopathology images, biomedical text and over 1.17 million image–caption pairs (Fig. 1a–b and Extended Data Fig. 1) through task-agnostic pretraining to address these unfilled needs. Based on CoCa[32], a state-of-the-art visual-language foundation pretraining framework, CONCH uses an image encoder, a text encoder and a multimodal fusion decoder, and it is trained using a combination of contrastive alignment objectives that seek to align the image and text modalities in the model's representation space and a captioning objective that learns to predict the caption corresponding to an image (Fig. 1c). We investigate the capabilities of CONCH on a wide array of tasks, including classification of image tiles and gigapixel WSIs, cross-modal image-to-text and text-to-image retrieval, image segmentation and image captioning, using a total of 14 diverse benchmarks. We demonstrate that our model achieves state-of-the-art performance across all benchmarks relative to other visual-language foundation models (Fig. 1d), including PLIP[54], BiomedCLIP[44] and OpenAICLIP[30], and it outperforms concurrent baselines, often by a large margin (Figs. 2–5).

Results

Zero-shot classification of diverse tissues and diseases

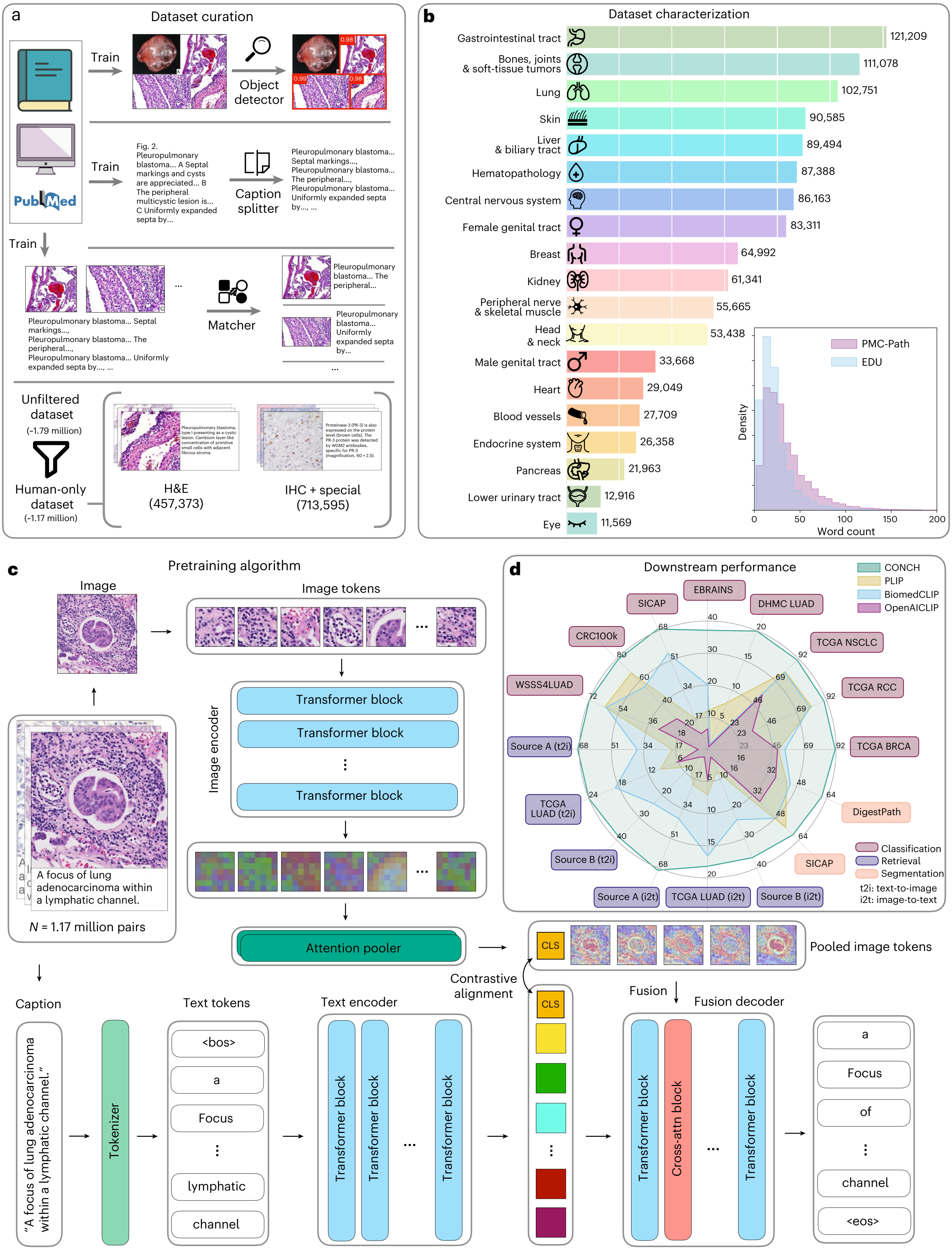
Contrastively aligned visual-language pretraining allows the model to be directly applied to downstream classification tasks without requiring further labeled examples for supervised learning or fine-tuning. This zero-shot transfer capability allows a single pretrained foundation model to be applied off the shelf to different downstream datasets with an arbitrary number of classes compared with the current paradigm of training a new model for every new task. While we do not expect zero-shot classification to currently be sufficiently accurate for most clinical use cases, in some tasks, we found CONCH to perform surprisingly well, and it may serve as a strong baseline for conventional supervised learning, especially when training labels are scarce.

Given a task, we first represented the set of class or category names using a set of predetermined text prompts, where each prompt corresponded to a class. An image was then classified by matching it with the most similar text prompt in the model's shared image–text representation space (Fig. 2a; see Methods for details). In practice, there are often multiple ways to phrase the same concept in text (for example, 'invasive lobular carcinoma (ILC) of the breast' and 'breast ILC'); therefore, we created an ensemble of multiple text prompts for each class during prediction, which was found to generally boost predictive performance compared to using a single text prompt (Extended Data Fig. 2). Additionally, while previous studies[44,54] primarily focused on classification tasks at the region-of-interest (ROI) level, we also investigated the zero-shot capability of our model on gigapixel WSIs by leveraging MI-Zero[56], which divides a WSI into smaller tiles and subsequently aggregates individual tile-level scores into a slide-level prediction (Fig. 2b).

In total, we evaluated CONCH on four slide-level classification tasks: The Cancer Genome Atlas (TCGA) BRCA (invasive breast carcinoma subtyping), TCGA NSCLC (non-small-cell lung cancer subtyping), TCGA RCC (renal cell carcinoma subtyping) and Dartmouth Hitchcock Medical Center (DHMC) LUAD (lung adenocarcinoma histologic pattern classification) and three ROI-level tasks: CRC100k (colorectal cancer tissue classification), WSSS4LUAD (LUAD tissue classification) and SICAP (Gleason pattern classification). We used balanced accuracy as the primary evaluation metric for TCGA NSCLC, TCGA RCC, TCGA LUAD, CRC100k and WSSS4LUAD, which accounted for class imbalance by weighing the accuracy score of each class equally. Following the community standard, we used Cohen's κ and quadratic weighted Cohen's κ as primary metrics for LUAD pattern classification and Gleason pattern classification, respectively, as they are regarded as more subjective tasks, which typically translates to higher inter-rater variability. We refer readers to Supplementary Tables 1–14 for more detailed reporting of model performance and Methods for detailed descriptions of evaluation datasets.

Fig. 1 | Data curation and model schematic. a, Automated data cleaning pipeline. Educational sources (EDU) and parts of the PubMed Central Open Access Dataset (PMC OA) were manually cleaned and used to train an object detector to detect histopathology images, a language model to split captions referring to multiple images and a matching model to match detected images to their corresponding captions. The cleaning process yielded a dataset of 1.79 million image–text pairs, and we then filtered out pairs referring to nonhumans to create our CONCH (human-only) pretraining dataset of 1.17 million (see Methods for details on data cleaning and Discussion on ablation experiments investigating data filtering). **b**, Estimated distribution of image–text pairs in the human-only pretraining dataset by topic. Note that pretraining data cover a diverse range of pathology topics. Inset, comparison of the distribution of caption lengths between PMC-Path and EDU (see Extended Data Fig. 1 for wordclouds of captions from each category). **c**, Visual-language pretraining setup. CONCH consists of an image encoder, a text encoder and a multimodal text decoder. The pretraining process uses both contrastive and captioning objectives. The contrastive objectives align the image and text encoders by maximizing the cosine-similarity scores between paired image and text embeddings, while the captioning objective maximizes the likelihood of generating the correct text conditioned on the image and previously generated text (see Methods for details). <bos>, beginning of sentence; attn, attention; <eos>, end of sentence. **d**, Radar plot comparing the performance of CONCH and baselines on various downstream tasks. CONCH outperforms baselines by a significant margin on a diverse set of tasks spanning zero-shot classification, retrieval and zero-shot segmentation (see Results for detailed descriptions of each task and metric).

On slide-level benchmarks, CONCH outperformed state-of-the-art visual-language foundation models (PLIP, BiomedCLIP and OpenAI-CLIP) on all tasks, often by a wide margin (Fig. 2c). For instance, for NSCLC subtyping and RCC subtyping, CONCH achieved a zero-shot accuracy of 90.7% and 90.2%, respectively, and it outperformed the next-best-performing model, PLIP, by 12.0% and 9.8% on each



 https://doi.org/10.1038/s41591-024-02856-4

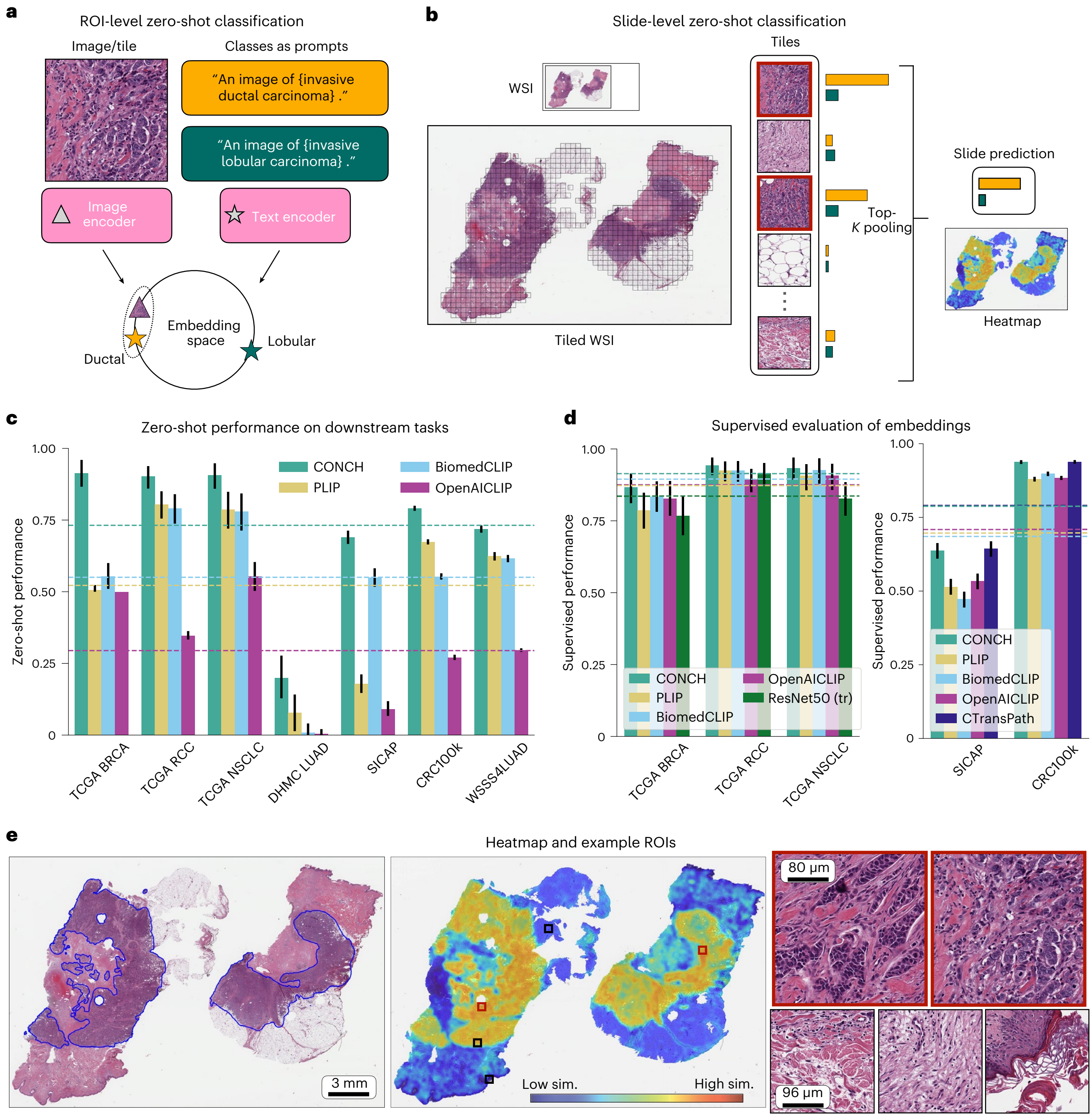


Fig. 2 | Zero-shot and supervised classification. **a**, Schematic of zero-shot classification using contrastively aligned image and text encoders. A prompt is constructed for each class, and the image is classified according to the prompt whose embedding is closest to that of the image in the shared embedding space. **b**, Zero-shot classification of WSIs. Each WSI is divided into tiles and processed as in **a**. The similarity scores for tiles are aggregated using top-*K* pooling to form slide-level similarity scores, the highest of which corresponds to the slide-level prediction. In **c**,**d**, dashed lines represent the average over tasks. Error bars represent 95% confidence intervals, and the centers correspond to computed values of each metric, as specified below. **c**, Zero-shot performance on downstream subtyping (TCGA BRCA, $n = 150$; TCGA RCC, $n = 225$; TCGA NSCLC, $n = 150$; DHMC LUAD, $n = 143$; CRC100k, $n = 7,180$; WSSS4LUAD, $n = 4,693$) and grading (SICAP, $n = 2,122$) tasks. Cohen's κ is reported for DHMC LUAD and quadratically weighted Cohen's κ is reported for SICAP, while balanced accuracy is reported for all other tasks. Additional metrics are reported in Supplementary Tables 1–7. **d**, Supervised evaluation of embeddings of each model. Linear probing is used for ROI-level tasks (CRC100k and SICAP), while ABMIL is used for slide-level tasks, with the same metrics reported as in **c** (see Supplementary Tables 15–19 for more detailed results). **e**, From left to right: pathologist-annotated IDC, corresponding heatmap and selected tiles at higher power. The heatmap is colored on the basis of the cosine-similarity score between each tile within the slide and the text prompt corresponding to the predicted class label. We find excellent agreement between the annotated image and high-similarity regions, with the tiles demonstrating classic IDC morphology within the high-similarity (high sim.) regions and stroma or other normal constituents of the breast in the low-similarity (low sim.) regions.

task with $P < 0.01$ according to a two-sided paired permutation test (Methods, 'Statistical analysis'). On the more difficult BRCA subtyping task, CONCH achieved a zero-shot accuracy of 91.3%, while other models performed at near-random chance, with accuracies ranging from 50.7% (PLIP) to 55.3% (BiomedCLIP), nearly 35% ($P < 0.01$) lower than CONCH. Lastly, on the LUAD pattern classification task, CONCH achieved a κ score of 0.200, which was 0.12 higher than that for the next-best-performing model, PLIP, although no significance was noted ($P = 0.055$). On ROI-level benchmarks, we observed similar findings, where CONCH achieved a zero-shot quadratic κ of 0.690 on SICAP (outperforming BiomedCLIP by 0.140, $P < 0.01$), a zero-shot accuracy of 79.1% on CRC100k (outperforming PLIP by 11.7%, $P < 0.01$) and a zero-shot accuracy of 71.9% on WSSS4LUAD (outperforming PLIP by 9.5%, $P < 0.01$). These results demonstrate that, in addition to achieving more accurate predictions on relatively easy tasks, CONCH was still able to achieve meaningful predictions on some more challenging tasks where other models may especially struggle.

When classifying a WSI using zero-shot transfer, in addition to computing an aggregated, slide-level prediction, we can create a heatmap to visualize the cosine-similarity score between each tile in the slide and the text prompt corresponding to the predicted class label. Regions with high similarity scores are deemed by the model to be close matches with the diagnosis (for example, invasive ductal carcinoma (IDC)), while regions with low similarity scores do not match the diagnosis (Fig. 2e). In an example of a breast IDC slide, we found that regions highlighted in the heatmap closely resembled the tumor regions as delineated by pathologist annotation (Fig. 2e, left and middle). Because the slide-level prediction score is a simple average of the similarity scores of the top-K tiles for a given class, the heatmap enables human interpretability by directly highlighting regions involved in the model's decision-making process, which can be displayed in high resolution to the human user for inspection (Fig. 2e, right). Additional examples are visualized in Extended Data Figs. 3–5. These findings suggest the possibility of using the zero-shot recognition ability of our model for coarse-grained tissue segmentation on WSIs, which we quantitatively evaluated in Results ('Zero-shot segmentation').

Few-shot classification with task-specific supervised learning

The zero-shot recognition capability of contrastive pretrained visual-language models for histopathology enables efficient and expedited application of a single foundation model to a potentially wide range of tasks without going through the laborious processes of training data collection, annotation and supervised model training for each new task. Sometimes, however, it may still be desirable to specialize the model with labeled training examples to maximize performance for a given task, ideally using as few labels as possible. In this section, we investigate the label efficiency when using the pretrained representation of the image encoder backbone of the visual-language foundation models for task-specific supervised classification. For each benchmark using supervised training, we used either the official training set (if provided) or the remaining cases from the dataset after holding out the set of cases used for zero-shot evaluation (Methods, 'Downstream evaluation datasets'). For slide-level tasks, we trained weakly supervised classification models using slide-level labels based on the widely used attention-based multiple-instance learning (ABMIL) algorithm[59]. For ROI-level tasks, we used logistic regression on top of the global (for example, classification (<CLS>) token) representation of each encoder, a practice commonly known as linear probing. In addition to PLIP, BiomedCLIP and OpenAICLIP encoders, we introduced supplementary baselines for comparison: for slide-level tasks, given its popularity, we used ResNet50 (ref. 60) (truncated after the third residual block) pretrained on ImageNet[61], while, for ROI-level tasks, we included CTransPath[62]—a state-of-the-art self-supervised pretrained histopathology image encoder (see Methods for details).

On the slide-level tasks (Fig. 2d, left), CONCH achieved a balanced accuracy score of 86.7%, 94.2% and 93.3% on BRCA subtyping, RCC subtyping and NSCLC subtyping, respectively, outperforming the commonly used ResNet50 ImageNet baseline by 10.0%, 2.6% and 10.7%, respectively ($P < 0.01$, $P = 0.223$ and $P = 0.033$). Overall, CONCH obtained an average accuracy of 91.4% across the three tasks, whereas PLIP and BiomedCLIP had an average accuracy of 87.3% and 89.4%, respectively, but no statistical significance was detected other than for BRCA subtyping in the comparison with PLIP ($P = 0.04$). In the ROI-level tasks (Fig. 2d, right), CONCH performed nearly identically to the state-of-the-art CTransPath encoder (93.8% versus 93.8% balanced accuracy on CRC100k and 0.833 versus 0.835 quadratically weighted κ on SICAP), while outperforming PLIP, BiomedCLIP and OpenAICLIP by 4.0–5.8% in balanced accuracy on CRC100k and by 0.071–0.128 in quadratically weighted κ on SICAP ($P < 0.01$ for all comparisons). These results demonstrated that, overall, CONCH provides a strong image encoder that performed either comparably to or better than all visual encoders tested, including a strong, vision-only self-supervised baseline (see Supplementary Tables 15–19 for detailed reporting of model performance).

Next, we investigated the label efficiency of different visual-language pretrained encoders in the few-shot setting, where we varied the number of training labels per class (n_c), for $n_c = 1, 2, 4, 8$, up to 512 per class or until we reached the maximum number of available labels in the training set. In the few-shot setting, for each experiment, we sampled five different sets of training examples and showed their individual performance by boxplot to account for the high variance in model performance when performing supervised learning with very few training examples (Fig. 3 and Extended Data Fig. 6). We first observed that CONCH achieved better performance (in terms of the median accuracy of five runs) than other encoders for all sizes of training set and for all tasks, which translated to requiring fewer labels to achieve the same performance. For instance, in BRCA subtyping, using the CONCH encoder and 8 training labels per class outperformed using PLIP, BiomedCLIP or OpenAICLIP with 64 labels per class, representing a nontrivial reduction in training set size—a trend we also observed for most tasks tested. Additionally, we noted that the zero-shot performance of CONCH was highly competitive when compared to few-shot supervised learning. Aside from relatively easy tasks such as RCC subtyping and CRC tissue classification, CONCH zero-shot outperformed PLIP-based and BiomedCLIP-based supervised learning in BRCA subtyping (up to 64 labels per class), NSCLC subtyping (up to 128 labels per class) and Gleason grading (up to 8 labels per class for PLIP and 64 labels per class for BiomedCLIP). These findings suggest that the zero-shot capability of a good visual-language foundation model should not be trivialized and, in fact, can serve as a very good baseline when evaluating the performance of task-specific diagnostic models trained with supervised learning. On the other hand, we found that the zero-shot capability of previous visual-language foundation models (that is, PLIP and BiomedCLIP) could be relatively easily surpassed by using supervised learning on top of the CONCH vision encoder with just a few labeled examples.

Application to classification of rare diseases

While previous investigations have focused on evaluating zero-shot and few-shot performance of visual-language pretrained models on relatively narrow tasks corresponding to a small set of possible classes (2–5 classes), to our best knowledge, the effectiveness of such models in large-scale, potentially fine-grained disease classification involving rare diseases has yet to be studied. Here, we investigated the utility of CONCH in recognizing up to 30 categories of brain tumors, all of which are classified as rare cancers following the definition of the RARECARE project[63] as having an annual crude incidence rate smaller than 6 per 100,000, the definition adopted by the National Cancer Institute's Surveillance, Epidemiology and End Results (SEER) program. We constructed a large-scale subtyping benchmark using the EBRAINS dataset

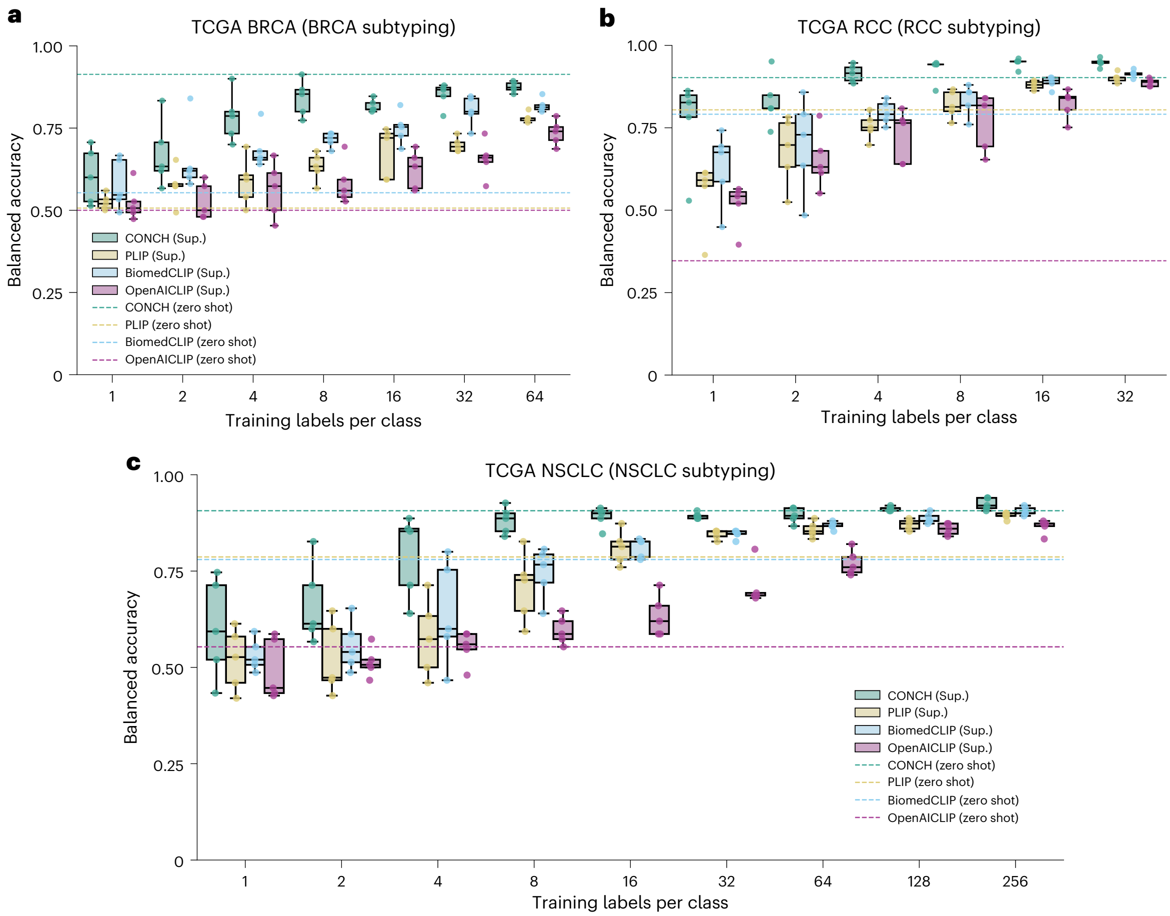


Fig. 3 | Slide-level few-shot classification experiments. a–c, We investigated the label efficiency of different visual-language pretrained encoders in the few-shot setting where we varied the number of training labels per class (n_c), for $n_c = 1, 2, 4, 8, 16$… until we reached the maximum number of available labels in the training set. For each n_c, we sampled five different sets of training examples and trained a weakly supervised ABMIL model on each training set using slide-level labels (see Methods, 'Supervised and weakly supervised classification experiments' for details). We show their individual model performance for BRCA subtyping (**a**), RCC subtyping (**b**) and NSCLC subtyping (**c**) by boxplot ($n = 5$ for each box) to study the variance in model performance when performing supervised learning with very few training examples. Boxes indicate quartile values and whiskers extend to data points within 1.5× the interquartile range. For reference, the zero-shot performance of each model is shown as a dashed line on the same plot. In terms of few-shot supervised learning, CONCH achieves better performance (in terms of the median accuracy of five runs) than other encoders for different sizes of training set and for all tasks. Additionally, the zero-shot performance of CONCH is surprisingly competitive, exceeding the few-shot performance of PLIP, BiomedCLIP and OpenAICLIP with up to 64 labels per class in the case of BRCA and NSCLC subtyping. Sup., supervised learning.

and evaluated the effectiveness of both zero-shot and supervised learning of various models.

In zero-shot classification, CONCH achieved a balanced accuracy score of 37.1% on the 30-class subtyping problem (Extended Data Fig. 7 and Supplementary Table 20), far surpassing the random chance baseline of 3.3%, as well as the second-best-performing visual-language pretrained zero-shot classifier, BiomedCLIP (+17.0%, $P < 0.01$). However, the generally low zero-shot performance of these models suggests that the current generation of visual-language foundation models may not yet be capable of directly performing 'in the wild', that is, open-set recognition of diverse diseases in pathology, and they are likely to achieve limited performance when evaluated on more challenging benchmarks involving many classes and rare entities.

Next, we studied the quality of pretrained representations of our vision encoder for training weakly supervised ABMIL classification models. Similar to the previous section, we also included additional baselines for pretrained vision encoders, including CTransPath, KimiaNet[64] and truncated ResNet50 (ImageNet initialized weights). We found that, while the zero-shot performance of CONCH was limited due to the challenging nature of the task, image embeddings of the frozen CONCH encoder could be used to develop strong-performing classification models when combined with weakly supervised learning. Specifically, CONCH combined with ABMIL achieved a balanced accuracy of 68.2% (Extended Data Fig. 7a and Supplementary Table 21), surpassing the vision-only self-supervised learning (SSL) pretrained CTransPath model (+6.8%, $P < 0.01$), as well as all other visual-language pretrained models tested by a substantial margin (+10.7%, $P < 0.01$ for PLIP, +14.4%, $P < 0.01$ for BiomedCLIP and +17.8%, $P < 0.01$ for OpenAICLIP). These results demonstrate the potential utility of a strong pretrained visual-language model as an effective image-only encoder for standard weakly supervised learning of computational pathology workflows, even when the task predominantly involves rare diseases. Lastly, we also investigated the few-shot learning performance of various models, motivated by the need for high label efficiency when training diagnostic models for rare diseases due to limited data availability. We observed a similar trend of superior label efficiency for CONCH compared to all other models tested, with other models generally requiring around four times as many labels to achieve comparable performance (Extended Data Fig. 7b).

Zero-shot cross-modal retrieval

By learning an aligned latent space for visual and language embeddings, our model is capable of cross-modal retrieval in a zero-shot setting, that

https://doi.org/10.1038/s41591-024-02856-4

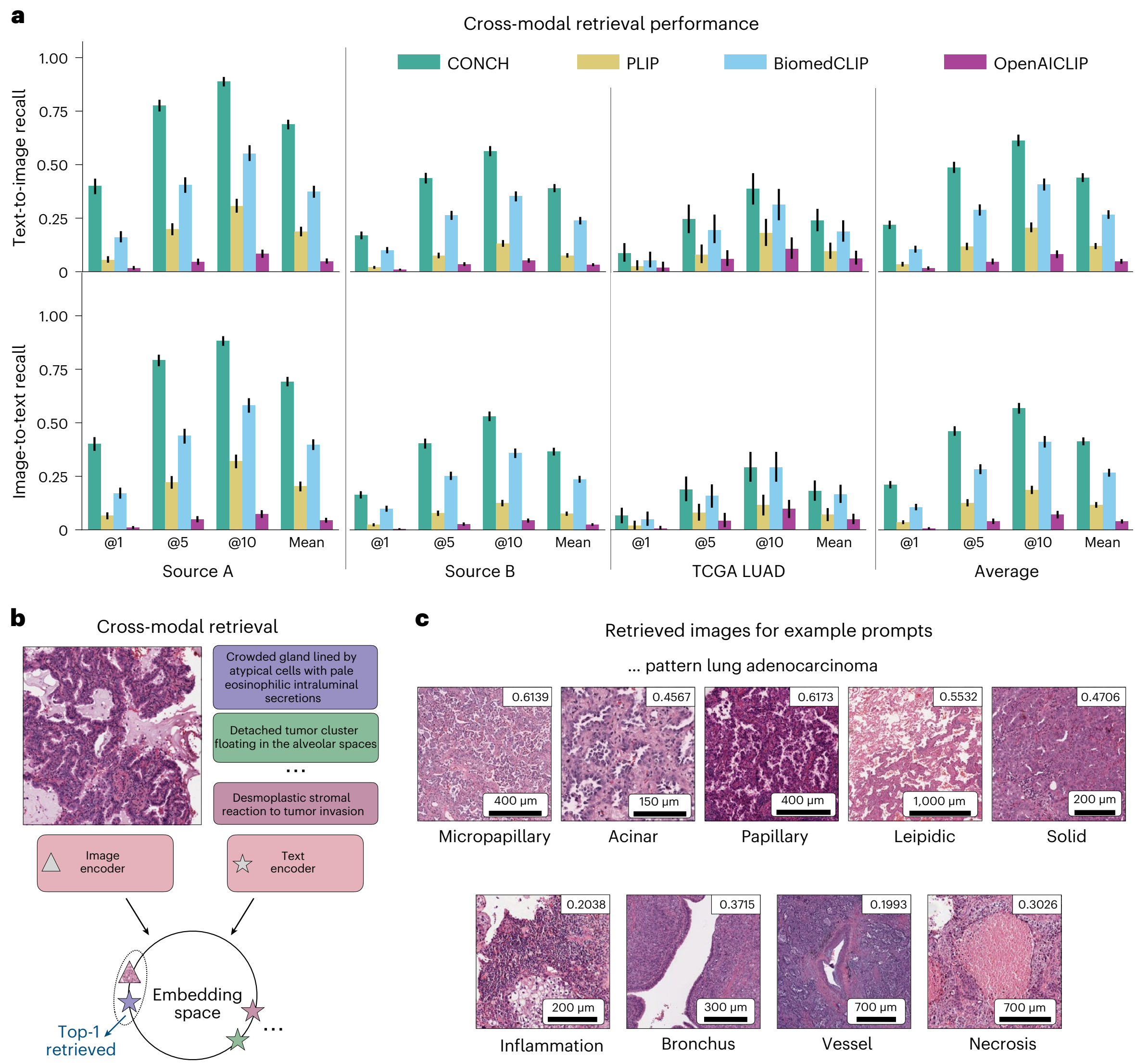


Fig. 4 | Zero-shot cross-modal retrieval. a, Model performance in cross-modal retrieval was evaluated on three datasets of image–text pairs (source A, $n = 797$; source B, $n = 1{,}755$; TCGA LUAD, $n = 165$). Similarity in the embedding space was computed between the query image and all text samples in the database. The top-K most similar texts were retrieved. We report Recall@K for $K \in \{1, 5, 10\}$ and the mean recall, which averages over K. We show both text-to-image (top row) and image-to-text (bottom row) retrieval for each retrieval task (columns). The rightmost column reports the average across tasks for each metric. CONCH outperforms other baselines on all retrieval tasks. Error bars indicate 95% confidence intervals. **b**, Schematic for zero-shot image-to-text retrieval (the text-to-image direction is analogous). **c**, Examples of images in the top five retrieved results from TCGA LUAD using LUAD-relevant queries with cosine-similarity scores shown in the top-right corner. Examples of other datasets using more diverse queries are shown in Extended Data Fig. 7. In general, we found that the images retrieved by the model matched what was described in the text prompt.

is, retrieving the corresponding text entry on the basis of an image query (image-to-text, abbreviated as 'i2t') or vice versa (text-to-image, abbreviated as 't2i'). This task naturally lends itself to image search applications, which are useful in the biomedical domain for applications such as identifying cases for inclusion in research cohorts or clinical trials, assistance with rare disease presentations or morphologies, and collecting cases for or helping to create educational resources. To perform text-to-image retrieval (the image-to-text direction was analogous), we used the text encoder to embed a text input that served as a query. We then used the query text embedding to retrieve similar images in the latent space (Fig. 4b).

We evaluated our model on three image–caption datasets, source A and source B (both are held-out sources from model pretraining that cover a diverse range of general pathology concepts) and TCGA LUAD (a much more specific dataset of tiles extracted from LUAD slides in TCGA and annotated with captions in house). Following previous studies[31,44,54], we used Recall@K as the metric for cross-modal retrieval (see Methods for more detailed descriptions of retrieval datasets).

On average, over the three datasets, CONCH significantly outperformed baselines by a large margin, achieving mean recall for text-to-image retrieval of 44.0%, and it outperformed the next-best model, BiomedCLIP, by 17.3% with $P < 0.01$ according to a two-sided paired permutation test (Fig. 4a). For source A and source B, CONCH achieved mean recall for text-to-image retrieval of 68.8% and 39.0%, respectively, outperforming the second-best model, BiomedCLIP, by 31.5% and 15.1% ($P < 0.01$ for both). For TCGA LUAD, CONCH achieved text-to-image mean recall of 24.0%, outperforming the next-best model, BiomedCLIP, by 5.3% but with no statistical significance ($P = 0.22$). However, CONCH significantly outperformed PLIP and OpenAICLIP ($P < 0.01$). Image-to-text retrieval for all three datasets followed the same trend as text-to-image retrieval in terms of performance and statistical significance, except for TCGA LUAD where the gap for CONCH and BiomedCLIP was slightly smaller (1.6%). We refer readers to Supplementary Tables 22–27 for more detailed reporting of model performance. On the basis of these results, CONCH was able to perform more accurate cross-modal retrieval than baselines.

In addition to using the paired captions as queries, we show examples of retrieved results using CONCH with simple text prompts of concepts related to LUAD (for example, 'solid-pattern LUAD') on the TCGA LUAD dataset (Fig. 4c). To provide examples from more complex text queries, such as 'cribriform prostatic adenocarcinoma', we used a highly diverse dataset of 321,261 tiles sampled from 1,620 cases held out during pretraining, spanning 108 OncoTree[65] codes (Extended Data Fig. 8). However, as this dataset did not have paired text data, we were not able to quantify the retrieval performance. The presented examples were confirmed by a pathologist to represent the text query closely.

Zero-shot segmentation

While WSIs can be gigapixels in size, they are generally heterogeneous, with diverse cell types, morphologies and tissue architectures represented, each often making up a small share of the slide. Consequently, segmentation on the slide level is a difficult and useful task to identify distinct regions of a WSI on the basis of the characteristics of interest, and it can reduce the number of tiles needed for downstream applications. However, because annotated data at the sub-slide level are expensive and laborious to collect, a general model capable of performing slide-level segmentation in a zero-shot setting is valuable. In this work, we explored the possibility of performing coarse-grained tissue segmentation on WSIs without labeled examples, instead directly using the demonstrated zero-shot retrieval and classification capabilities of our model.

Given a WSI, we divided the tissue regions into smaller image tiles and posed a given segmentation task as classifying each tile using zero-shot classification and assigning the predicted class label to all pixels in the tile, performed for all tiles (Fig. 5a). To minimize sharp transition in predicted values for pixels at the boundary of neighboring tiles, we tiled the WSIs with a 75% overlap and averaged the prediction scores in overlapped regions to achieve a smoother appearance in the predicted segmentation map. We evaluated our model on SICAP for prostate tumor versus normal tissue segmentation and on DigestPath for malignant versus benign tissue segmentation in CRC specimens. We report the widely used Dice score, in addition to precision and recall, for each task against ground-truth pixel-level annotations, with scores macro-averaged over all images in each dataset (see Methods for more details). We refer the reader to Supplementary Tables 28 and 29 for more detailed results of model performance.

CONCH outperformed other models in both tasks (Fig. 5b,c). In SICAP, CONCH achieved an average Dice score of 0.601 (0.549, $P = 0.08$ for PLIP and 0.484, $P < 0.01$ for BiomedCLIP), an average recall score of 0.751 (0.644, $P < 0.01$ for PLIP and 0.557, $P < 0.01$ for BiomedCLIP) and an average precision core of 0.672 (0.605, $P = 0.024$ for PLIP and 0.536, $P < 0.01$ for BiomedCLIP). In DigestPath, CONCH achieved an average Dice score of 0.615 (0.426, $P < 0.01$ for PLIP and 0.446, $P < 0.01$ for BiomedCLIP), an average recall score of 0.709 (0.541, $P < 0.01$ for PLIP and 0.601, $P < 0.01$ for BiomedCLIP) and an average precision core of 0.663 (0.526, $P = 0.024$ for PLIP and 0.581, $P < 0.01$ for BiomedCLIP). Additionally, we found that, despite the coarse-grained and zero-shot nature of the approach, the model was able to produce reasonably accurate pixel-level segmentation masks in some instances, as visualized in Fig. 5d,e.

Discussion

Most previous tools in computational pathology have attempted to extract meaningful patterns and discriminative signals from image data and/or structured patient data such as genomics and have ignored the textual aspect of pathology. However, these approaches leave on the table a huge amount of information present in descriptions of images, information that allows pathology trainees to generalize from a few exemplar images of an entity to images in the real world that are often substantially more diverse. While several recent studies[44,54] attempted to leverage image and caption data from social media or biomedical research articles to build visual-language foundation models applicable to the domain of histopathology, we found that, across a number of tasks, both their zero-shot and their supervised classification performance remain limited, hindering their practical value as general-purpose recognition or retrieval systems for histopathology. Additionally, beyond working on small ROIs, the models' abilities to perform in more complex settings (for example, classification of rare diseases or tumor segmentation on heterogeneous gigapixel WSIs) remain underexplored.

In this study, we demonstrated that, by using the currently largest histopathology-specific, paired image–text dataset of over 1.17 million examples for task-agnostic pretraining, we could build a high-performance visual-language foundation model that could then demonstrate utility in a wide range of clinically relevant downstream tasks such as classification, retrieval and tissue segmentation. Our model is equipped with strong zero-shot recognition capabilities out of the box, which can potentially relieve the burden of annotating training examples for many specific classification tasks, and we demonstrated that its zero-shot performance often rivaled or even outperformed conventional supervised learning baselines in these tasks under few-shot settings. Additionally, the much-improved zero-shot image-to-text and text-to-image retrieval capabilities of our model will potentially empower trainees, physicians and researchers to more accurately and flexibly retrieve relevant patient cases or educational examples based on image or natural language queries once it can be efficiently implemented into healthcare systems or databases. Equipped with a multimodal decoder, our visual-language foundation model also provides the flexibility to be further fine-tuned in downstream tasks that involve language generation (for example, image captioning; see Methods, 'Captioning with fine-tuning' for details and Extended Data Fig. 9 and Supplementary Table 30 for exploratory results) and/or multimodal reasoning based on both visual and textual inputs. However, beyond promising results in select tasks, we also found and noted that current visual-language pretrained models, including CONCH, still perform poorly on challenging zero-shot problems (relative to their supervised learning counterparts) that involve a large number of classes and rare diseases. These observations suggest that we still potentially have a long way to go before achieving the goal of building a foundation model capable of truly universal zero-shot recognition or retrieval for histopathology.

We additionally performed ablation experiments to investigate the effect of data filtering, different pretraining algorithms and unimodal pretraining on the performance of our model. Most notably, we found that performing unimodal pretraining (especially vision encoder SSL pretraining) could improve model performance in zero-shot classification and retrieval across most tasks (see Extended Data Fig. 10 for more details).

Another relatively underexplored aspect is the compatibility of visual-language pretrained foundation models with conventional end-to-end supervised learning aimed at targeting specific tasks. For some widely studied, single-disease model tasks such as prostate adenocarcinoma Gleason grading, there have been substantial efforts by various groups around the world to build large and diverse datasets with detailed ROI or pixel-level annotations suitable for end-to-end supervised machine learning. A natural question is, given the abundance of annotated data, does pretraining a foundation model on images and captions from diverse tissue types and diseases still lead to tangible benefits for these specific tasks? We attempted to provide some insight into this question by assembling a large and diverse dataset of more than 200,000 labeled ROIs for the task of prostate cancer Gleason grading from multiple publicly available sources, before

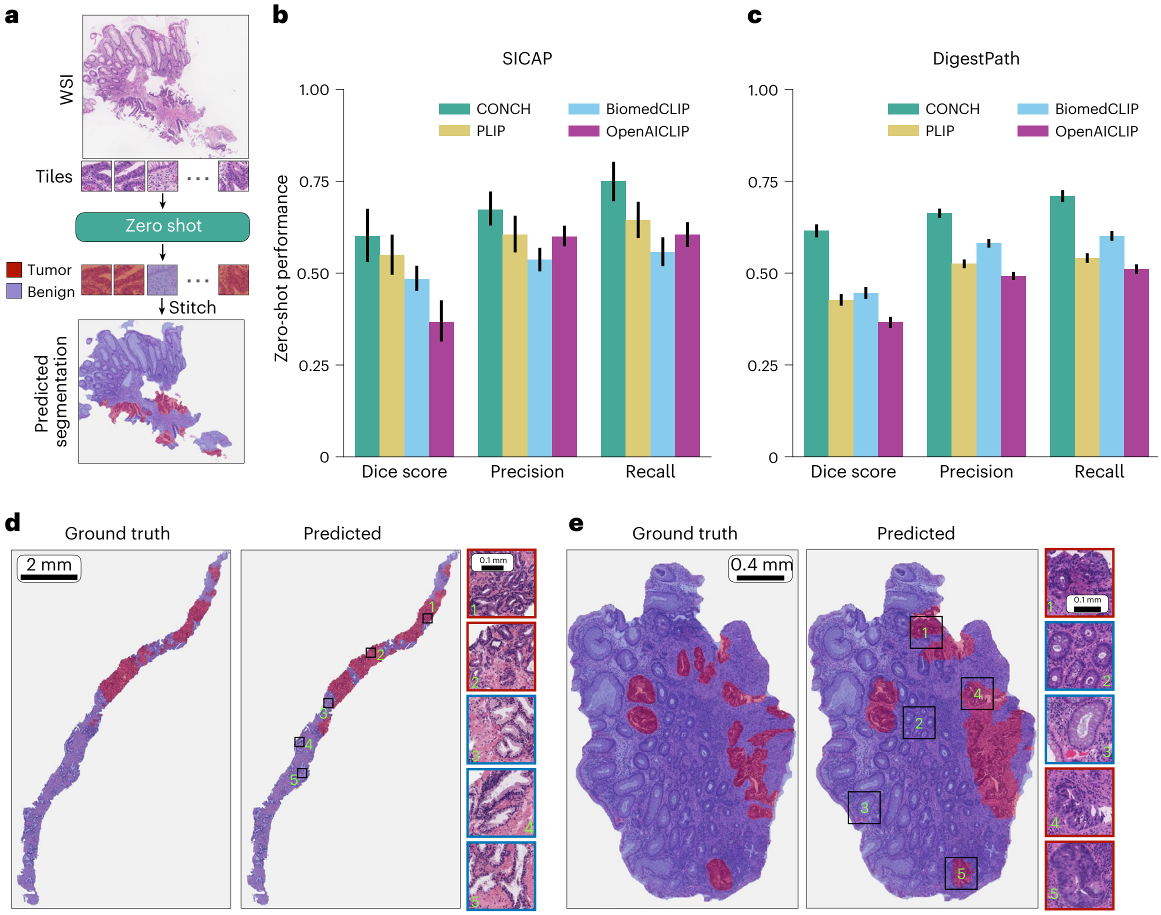


Fig. 5 | Zero-shot segmentation. a, Schematic illustrating zero-shot segmentation on WSIs (or large tissue sections). To perform segmentation, we divided each WSI into tiles and used zero-shot classification to predict the label of each tile. The tile-level predictions were stitched together to form the predicted segmentation mask. **b**,**c**, Zero-shot segmentation performance of CONCH and baselines on SICAP ($n = 31$) (**b**) and DigestPath ($n = 250$) (**c**) datasets. The macro-averaged Dice score, precision and recall are reported. Error bars represent 95% confidence intervals. **d**,**e**, Examples of CONCH segmentation prediction on WSIs for SICAP (**d**) and DigestPath (**e**). The left panel shows the ground truth, and the right panel shows the predicted segmentation mask, with example regions enlarged. Red and blue indicate tumor and normal tissue, respectively. In general, in these examples, CONCH displays excellent sensitivity to tumor regions with slightly lower specificity, although most of the regions that CONCH segments as tumor that are in fact nontumor are adjacent to cancerous glands or contain cancer-associated stroma for both SICAP and DigestPath.

performing end-to-end fine-tuning of our vision encoder, as well as a handful of other pretrained standard convolutional neural network (CNN)-based and vision transformer (ViT)-based models including domain-specific encoders such as KimiaNet[64] and CTransPath[62]. In our experiments, we found that, even with hundreds of thousands of labeled ROIs paired with transfer learning from ImageNet weights or SSL pretraining, a fine-tuned CONCH model can still provide a sizeable improvement, even when compared to a much larger ViT-Large model (Supplementary Table 31).

While a recent investigation found that current visual-language pretrained foundational models may perform worse than smaller encoders in the specific scenario of WSI-to-WSI matching using one specific algorithm[66], our experiments in both rare disease few-shot and weakly supervised classification, as well as end-to-end fine-tuning, showed that CONCH can serve as a state-of-the-art visual encoder for histopathology images, in addition to providing a shared image–text latent space that unlocks additional multimodal capabilities. Nevertheless, these findings highlight the importance of continuous research and evaluation to better understand the strengths and limitations of foundational models for computational pathology.

A key limitation of our study is the scale of data pretraining, which still pales in comparison to billion-scale datasets used in developing large-scale visual-language foundation models in the general machine learning community; therefore, we are likely to see further potential improvement in zero-shot recognition capabilities, representation quality and robustness by increasing both the quantity and the quality of histopathology image–caption datasets. However, given the increasing data scale used in pretraining, the potential for unintentional data overlap between pretraining data and downstream test data becomes increasingly high, a limitation also shared by previous vision-language pretraining approaches in the biomedical domain[44,54]. Detecting and removing duplicates and near-duplicates typically relies on a combination of heuristics and manual assessment, and this has not been sufficiently explored in the biomedical domain, serving as an open research question for future work. In this study, we minimized the potential for data overlap by ensuring that no publicly available test dataset was directly derived from any training sources and by only holding out data at the source level. Another limitation of the study is that we did not investigate the robustness of zero-shot classification (for both image ROIs and WSIs) across different data cohorts with potentially different staining variations, tissue preparation protocols and scanner-specific imaging profiles, compared to using conventional supervised learning or parameter-efficient fine-tuning techniques[67,68]. Additionally, while we showed that simply ensembling a small number of templates and class names written by a pathologist can already work well for several tasks, we did not attempt to explicitly engineer the prompts on the basis of the model's performance (for example, by using a validation set). We note that doing an explicit search for 'good' prompts on a small validation set (if it is available) may be much more effective in practice while still retaining the benefit of not needing to fine-tune the model, although it would no longer be strictly considered zero-shot transfer[69,70]. Moreover, as a zero-shot classification algorithm for WSIs, MI-Zero is only best suited for tasks where the defining morphological patterns of each class are mutually exclusive, and it may not work on tasks with specific assumptions or guidelines. This includes tasks such as Gleason scoring where both the primary and the secondary pattern may need to be considered to inform the classification or tumor versus normal classification, in which a slide may be appropriately labeled as 'positive' as soon as a single tumor-containing region is identified. We note that, for these types of tasks, the pooling function of MI-Zero can be adjusted to better suit the nature of the task, and we leave its implementation and evaluation to future studies. Lastly, while the current landscape of visual-language foundation models for histopathology focuses primarily on image-level tasks, the ability of these models to recognize fine-grained visual concepts at the region level (that is, cellular or even subcellular level) has not yet been studied, meaning that other important tasks such as mitosis detection, fine-grained tissue segmentation or cell counting currently remain outside the scope of their downstream capabilities.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41591-024-02856-4.

References

1. Song, A. H. et al. Artificial intelligence for digital and computational pathology. *Nat. Rev. Bioeng.* **1**, 930–949 (2023).
2. Bera, K., Schalper, K. A., Rimm, D. L., Velcheti, V. & Madabhushi, A. Artificial intelligence in digital pathology—new tools for diagnosis and precision oncology. *Nat. Rev. Clin. Oncol.* **16**, 703–715 (2019).
3. Shmatko, A., Ghaffari Laleh, N., Gerstung, M. & Kather, J. N. Artificial intelligence in histopathology: enhancing cancer research and clinical oncology. *Nat. Cancer* **3**, 1026–1038 (2022).
4. Lipkova, J. et al. Artificial intelligence for multimodal data integration in oncology. *Cancer Cell* **40**, 1095–1110 (2022).
5. Bejnordi, B. E. et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *JAMA* **318**, 2199–2210 (2017).
6. Coudray, N. et al. Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning. *Nat. Med.* **24**, 1559–1567 (2018).
7. Lu, M. Y. et al. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nat. Biomed. Eng.* **5**, 555–570 (2021).
8. Skrede, O.-J. et al. Deep learning for prediction of colorectal cancer outcome: a discovery and validation study. *Lancet* **395**, 350–360 (2020).
9. Chen, R. J. et al. Pan-cancer integrative histology–genomic analysis via multimodal deep learning. *Cancer Cell* **40**, 865–878 (2022).
10. Courtiol, P. et al. Deep learning-based classification of mesothelioma improves prediction of patient outcome. *Nat. Med.* **25**, 1519–1525 (2019).
11. Lu, M. Y. et al. AI-based pathology predicts origins for cancers of unknown primary. *Nature* **594**, 106–110 (2021).
12. Zhu, L. et al. An accurate prediction of the origin for bone metastatic cancer using deep learning on digital pathological images. *EBioMedicine* **87**, 104426 (2023).
13. Kalra, S. et al. Yottixel—an image search engine for large archives of histopathology whole slide images. *Med. Image Anal.* **65**, 101757 (2020).
14. Hegde, N. et al. Similar image search for histopathology: SMILY. *NPJ Digit. Med.* **2**, 56 (2019).
15. Wang, X. et al. RetCCL: clustering-guided contrastive learning for whole-slide image retrieval. *Med. Image Anal.* **83**, 102645 (2023).
16. Chen, C. et al. Fast and scalable search of whole-slide images via self-supervised deep learning. *Nat. Biomed. Eng.* **6**, 1420–1434 (2022).
17. Kather, J. N. et al. Pan-cancer image-based detection of clinically actionable genetic alterations. *Nat. Cancer* **1**, 789–799 (2020).
18. Saldanha, O. L. et al. Self-supervised attention-based deep learning for pan-cancer mutation prediction from histopathology. *NPJ Precis. Oncol.* **7**, 35 (2023).
19. Graham, S. et al. Hover-Net: simultaneous segmentation and classification of nuclei in multi-tissue histology images. *Med. Image Anal.* **58**, 101563 (2019).
20. Campanella, G. et al. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nat. Med.* **25**, 1301–1309 (2019).
21. Bulten, W. et al. Automated deep-learning system for Gleason grading of prostate cancer using biopsies: a diagnostic study. *Lancet Oncol.* **21**, 233–241 (2020).
22. Nagpal, K. et al. Development and validation of a deep learning algorithm for improving Gleason scoring of prostate cancer. *NPJ Digit. Med.* **2**, 48 (2019).
23. Mobadersany, P. et al. Predicting cancer outcomes from histology and genomics using convolutional networks. *Proc. Natl Acad. Sci. USA* **115**, E2970–E2979 (2018).
24. Chen, R. J. et al. Multimodal co-attention transformer for survival prediction in gigapixel whole slide images. In *Proc. IEEE/CVF International Conference on Computer Vision* 4015–4025 (IEEE, 2021).
25. Fu, Y. et al. Pan-cancer computational histopathology reveals mutations, tumor composition and prognosis. *Nat. Cancer* **1**, 800–810 (2020).
26. Sammut, S.-J. et al. Multi-omic machine learning predictor of breast cancer therapy response. *Nature* **601**, 623–629 (2022).
27. Huang, Z. et al. Artificial intelligence reveals features associated with breast cancer neoadjuvant chemotherapy responses from multi-stain histopathologic images. *NPJ Precis. Oncol.* **7**, 14 (2023).
28. Foersch, S. et al. Multistain deep learning for prediction of prognosis and therapy response in colorectal cancer. *Nat. Med.* **29**, 430–439 (2023).
29. Vanguri, R. S. et al. Multimodal integration of radiology, pathology and genomics for prediction of response to PD-(L)1 blockade in patients with non-small cell lung cancer. *Nat. Cancer* **3**, 1151–1164 (2022).
30. Radford, A. et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning* (eds Meila, M. & Zhang, T.) 8748–8763 (PMLR, 2021).
31. Jia, C. et al. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning* (eds Meila, M. & Zhang, T.) 4904–4916 (PMLR, 2021).
32. Yu, J. et al. CoCa: contrastive captioners are image–text foundation models. *Trans. Mach. Learn. Artif. Intell.* https://openreview.net/forum?id=Ee277P3AYC (2022).
33. Li, J., Li, D., Xiong, C. & Hoi, S. BLIP: bootstrapping language–image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning* (eds Chaudhur, K. et al.) 12888–12900 (PMLR, 2022).

34. Singh, A. et al. FLAVA: a foundational language and vision alignment model. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition* 15638–15650 (IEEE, 2022).
35. Li, H. et al. Uni-Perceiver v2: a generalist model for large-scale vision and vision-language tasks. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition* 2691–2700 (IEEE, 2023).
36. Alayrac, J.-B. et al. Flamingo: a visual language model for few- shot learning. *Adv. Neural Inf. Process. Syst.* **35**, 23716–23736 (2022).
37. Li, Y., Fan, H., Hu, R., Feichtenhofer, C. & He, K. Scaling language–image pre-training via masking. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition* 23390–23400 (IEEE, 2023).
38. Wang, W. et al. Image as a foreign language: BEiT pretraining for vision and vision-language tasks. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition* 19175–19186 (IEEE, 2023).
39. Schuhmann, C. et al. LAION-5B: an open large-scale dataset for training next generation image-text models. *Adv. Neural Inf. Process. Syst.* **35**, 25278–25294 (2022).
40. Chen, Z., Song, Y., Chang, T.-H. & Wan, X. Generating radiology reports via memory-driven transformer. In *Proc. 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (eds Webber, B. et al.) 1439–1449 (Association for Computational Linguistics, 2020); https://aclanthology.org/2020.emnlp-main.112
41. Liu, G. et al. Clinically accurate chest X-ray report generation. In *Proc. 4th Machine Learning for Healthcare Conference* (eds Doshi-Velez, F. et al.), Vol. 106, 249–269 (PMLR, 2019).
42. Tiu, E. et al. Expert-level detection of pathologies from unannotated chest X-ray images via self-supervised learning. *Nat. Biomed. Eng.* **6**, 1399–1406 (2022).
43. Huang, S.-C., Shen, L., Lungren, M. P. & Yeung, S. GLoRIA: a multimodal global–local representation learning framework for label-efficient medical image recognition. In *Proc. IEEE/CVF International Conference on Computer Vision* 3942–3951 (IEEE, 2021).
44. Zhang, S. et al. BiomedCLIP: a multimodal biomedical foundation model pretrained from fifteen million scientific image–text pairs. Preprint at https://doi.org/10.48550/arXiv.2303.00915 (2023).
45. Wang, Z., Wu, Z., Agarwal, D. & Sun, J. MedCLIP: contrastive learning from unpaired medical images and text. In *Proc. 2022 Conference on Empirical Methods in Natural Language Processing* (eds Che, W. & Shutova, E.) 3876–3887 (Association for Computational Linguistics, 2022).
46. Schaumberg, A. J. et al. Interpretable multimodal deep learning for real-time pan-tissue pan-disease pathology search on social media. *Mod. Pathol.* **33**, 2169–2185 (2020).
47. Maleki, D. & Tizhoosh, H. R. LILE: look in-depth before looking elsewhere—a dual attention network using transformers for cross-modal information retrieval in histopathology archives. In *International Conference on Medical Imaging with Deep Learning* (eds Konukoglu, E. et al.) 879–894 (PMLR, 2022).
48. Zhang, Y., Jiang, H., Miura, Y., Manning, C. D. & Langlotz, C. P. Contrastive learning of medical visual representations from paired images and text. In *Machine Learning for Healthcare Conference* (eds Lipton, Z. et al.) 2–25 (PMLR, 2022).
49. Zhang, H. et al. PathNarratives: data annotation for pathological human–AI collaborative diagnosis. *Front. Med.* **9**, 1070072 (2023).
50. Tsuneki, M. & Kanavati, F. Inference of captions from histopathological patches. In *International Conference on Medical Imaging with Deep Learning* (eds Konukoglu, E. et al.) 1235–1250 (PMLR, 2022).
51. Zhang, R., Weber, C., Grossman, R. & Khan, A. A. Evaluating and interpreting caption prediction for histopathology images. In *Machine Learning for Healthcare Conference* (eds Doshi-Velez, F. et al.) 418–435 (PMLR, 2020).
52. Naseem, U., Khushi, M. & Kim, J. Vision-language transformer for interpretable pathology visual question answering. *IEEE J. Biomed. Health Inform.* **27**, 1681–1690 (2022).
53. He, X. Towards visual question answering on pathology images. In *Proc. 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)* (eds Zong, C. et al.) 708–718 (Association for Computational Linguistics, 2021).
54. Huang, Z., Bianchi, F., Yuksekgonul, M., Montine, T. J. & Zou, J. A visual-language foundation model for pathology image analysis using medical Twitter. *Nat. Med.* **29**, 2307–2316 (2023).
55. Gamper, J. & Rajpoot, N. Multiple instance captioning: learning representations from histopathology textbooks and articles. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition* 16549–16559 (IEEE, 2021).
56. Lu, M. Y. et al. Visual language pretrained multiple instance zero-shot transfer for histopathology images. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* 19764–19775 (IEEE, 2023).
57. Lin, W. et al. PMC-CLIP: contrastive language–image pre-training using biomedical documents. In *Medical Image Computing and Computer Assisted Intervention—MICCAI 2023* (ed. Greenspan, H. et al.) 525–536 (Springer Nature, 2023).
58. Ikezogwo, W. O. et al. Quilt-1M: one million image–text pairs for histopathology. In *Advances in Neural Information Processing Systems* (eds Oh, A. et al.) 37995–38017 (Curran Associates, Inc., 2023).
59. Ilse, M., Tomczak, J. & Welling, M. Attention-based deep multiple instance learning. In *International Conference on Machine Learning* (eds Dy, J. & Krause, A.) 2127–2136 (PMLR, 2018).
60. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition* 770–778 (IEEE, 2016).
61. Deng, J. et al. ImageNet: a large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition* 248–255 (IEEE, 2009).
62. Wang, X. et al. Transformer-based unsupervised contrastive learning for histopathological image classification. *Med. Image Anal.* **81**, 102559 (2022).
63. Gatta, G. et al. Burden and centralised treatment in Europe of rare tumours: results of RARECAREnet—a population-based study. *Lancet Oncol.* **18**, 1022–1039 (2017).
64. Riasatian, A. et al. Fine-tuning and training of densenet for histopathology image representation using TCGA diagnostic slides. *Med. Image Anal.* **70**, 102032 (2021).
65. Kundra, R. et al. OncoTree: a cancer classification system for precision oncology. *JCO Clin. Cancer Inform.* **5**, 221–230 (2021).
66. Alfasly, S. et al. When is a foundation model a foundation model. Preprint at https://doi.org/10.48550/arXiv.2309.11510 (2023).
67. Zhou, K., Yang, J., Loy, C. C. & Liu, Z. Learning to prompt for vision-language models. *Int. J. Comput. Vis.* **130**, 2337–2348 (2022).
68. Gao, P. et al. CLIP-Adapter: better vision-language models with feature adapters. *Int. J. Comput. Vis.* **132**, 581–595 (2024).
69. Perez, E., Kiela, D. & Cho, K. True few-shot learning with language models. *Adv. Neural Inf. Process. Syst.* **34**, 11054–11070 (2021).
70. Sanh, V. et al. Multitask prompted training enables zero-shot task generalization. In *10th International Conference on Learning Representations* https://openreview.net/forum?id=9Vrb9D0WI4 (OpenReview.net 2021).

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

https://doi.org/10.1038/s41591-024-02856-4

Methods

Dataset curation

Most data used for this study were obtained from publicly available research articles. For internal data, the Mass General Brigham institutional review board approved the retrospective analysis of internal pathology images, corresponding reports and electronic records. All internal digital data, including WSIs, pathology reports and EMRs were deidentified before computational analysis and model development. Patients were not directly involved or recruited for the study. Informed consent was waived for analyzing archival pathology slides retrospectively. We used publicly available articles from PubMed to curate the largest-to-date dataset of histopathology image–caption pairs. We used deep learning to automate data cleaning iteratively. For curation, we divided the data sources into two categories: EDU, which consists of data extracted from educational notes, and PMC OA, which consists of data downloaded from the PubMed Central Open Access Dataset (https://ncbi.nlm.nih.gov/pmc/tools/openftlist/).

The data curation process poses two main challenges: filtering for histopathology data and handling image panels. The first challenge is that the raw downloaded data comprised both histopathology and non-histopathology examples. The second challenge is that a substantial portion of the data were in the form of figure panels, where the images consisted of multiple subimages arranged in a panel with parts of the caption addressing all or some of the subimages. In light of these challenges, manually cleaning the data was infeasible. We cleaned the data in three steps: (1) detecting histopathology images (as single images or subimages); (2) splitting captions that referred to image panels into separate captions into subcaptions; and (3) aligning subimages with subcaptions within each image panel.

To detect histopathology images, we used an object detection model (YOLOv5)[71] to generate bounding boxes for extracting detected images. To avoid the laborious task of manually labeling ground-truth bounding boxes, we generated synthetic data by randomly selecting single-panel images and arranging them in an image panel. We iteratively refined the detection model by validating it on a small subset (<0.5%) of PMC OA and adding incorrectly labeled samples to the training set.

For caption splitting, we collected a dataset of original and split captions (while cleaning the EDU dataset) to fine-tune a generative pretrained transformer (GPT)-style model pretrained on PubMed and other medical text[72]. We posed the problem of splitting captions as causal language modeling, where we fine-tuned the language model to take the original full caption as input and predicted the subcaptions separated by the keyword 'next caption'. We used the fine-tuned model to perform caption splitting.

To align the detected histopathology images with split captions, we first trained a CLIP model[30] on the cleaned EDU dataset, along with PMC OA single figures that did not require splitting and alignment. Using the trained model, given a set of m detected images and n split captions from an image panel, we computed the image embeddings $\{\mathbf{u}_0, \mathbf{u}_1, \ldots, \mathbf{u}_m\}$ and text embeddings $\{\mathbf{v}_0, \mathbf{v}_1, \ldots, \mathbf{v}_n\}$ in the aligned latent space. For each image embedding $\mathbf{u}_i$, we computed the cosine-similarity score with each text embedding $\mathbf{v}_j$. We retrieved the text with the highest cosine-similarity score $s_{i,j} := \mathbf{u}_i^T \mathbf{v}_j$ and considered $\{\mathbf{u}_i, \mathbf{v}_j\}$ to be an image–caption pair for our cleaned dataset.

By applying the three steps above to PMC OA, we created PMC-Path, a pathology-specific image–caption dataset derived from PubMed figures. We then combined it with EDU to form our full, unfiltered pretraining dataset of 1,786,362 image–caption pairs. However, PMC-Path also contained a substantial number of pairs referring to animal histopathology, as well as non-hematoxylin and eosin (H&E) stains (immunohistochemistry (IHC), Masson's trichrome, Congo red, etc.). Because our downstream evaluation concerned only human histopathology and H&E tasks, we wanted to assess how the animal and special staining data would affect performance. We first parsed the captions to exclude samples referencing nonhuman animals, forming a dataset of 1,170,647 human pairs. Additionally, we trained a classifier that identified H&E stains to further filter the human-only dataset and create a dataset of 457,372 pairs. We found that CONCH pretrained on the human-only dataset performed the best on downstream tasks in general (Extended Data Fig. 10a).

Visual-language pretraining

For visual-language pretraining, we used an equal-weighted combination of the image–text contrastive loss and the captioning loss following CoCa[32], a state-of-the-art visual-language foundation model pretrained on general-domain image–caption pairs. The model consisted of an image encoder, $f(\cdot\,;\theta)$, a text encoder, $g(\cdot\,;\phi)$, and a multimodal text decoder, $h(\cdot\,;\psi)$. The image encoder included the backbone and two attentional pooler modules, parameterized by θ_{backbone}, θ_{contrast} and θ_{caption}, respectively. The backbone was a ViT[73] following the standard ViT-base architecture with 12 transformer layers, 12 attention heads, an embedding dimension of 768 and a hidden dimension of 3,072. The token size was 16 × 16, and learned absolute positional embeddings were added to each token. The backbone transformed images in the form of raw red–green–blue (RGB) pixel values to dense feature maps in a more semantically rich representation space learned from data. Each attentional pooler was responsible for computing a fixed number (denoted by n) of image tokens from the last layer representation of the ViT backbone using multiheaded attention and n learned queries. For enabling cross-modal retrieval through contrastive learning, the first attentional pooler $f_{\text{contrast}}(\cdot\,;\theta_{\text{contrast}})$ used a single query ($n_{\text{contrast}} = 1$) to compute a single image token designed to capture the global representation of the image. The second attentional pooler $f_{\text{caption}}(\cdot\,;\theta_{\text{caption}})$ used $n_{\text{caption}} = 256$ queries to generate a set of 256 image tokens designed to capture more local and fine-grained details of the image, which are typically required for captioning. The text encoder and multimodal decoder were both GPT-style models that used causal attention masks for left-to-right autoregressive language modeling. Similar to the image encoder, the text encoder and multimodal decoder consisted of 12 transformer layers with an embedding dimension of 768 and a hidden dimension of 3,072. The text encoder included an embedding table for mapping discrete word tokens to continuous embeddings and a set of learned absolute positional embeddings. Additionally, the text encoder appended a learned <CLS> token to each tokenized caption, which had access to the full context during transformer attention to extract a global representation of a given caption. The multimodal decoder inserted a cross-attention layer after each multiheaded self-attention layer to incorporate information from image tokens and included a final language modeling head for predicting the distribution of the next token over the supported vocabulary.

During visual-language pretraining, a mini-batch consisted of M image–caption pairs $(\mathbf{x}_i, \mathbf{w}_i)_{i=1}^{M}$, where $\mathbf{w}_i = (\texttt{<BOS>}, \mathrm{w}_{i,1}, \ldots, \mathrm{w}_{i,T}, \texttt{<EOS>})$ is a sequence of T word tokens representing the ith caption. For a given pair $(\mathbf{x}_i, \mathbf{w}_i)$, we let $(\mathbf{u}_i, \mathbf{v}_i)$ be the output of $f_{\text{contrast}}(\cdot\,;\theta_{\text{contrast}})$ and the output of $g(\cdot\,;\phi)$ at the position corresponding to the <CLS> token after ℓ_2-normalization. The complete objective is given by:

$$\mathcal{L} = -\frac{1}{2M}\sum_{i=1}^{M}\log\frac{\exp(\tau\mathbf{u}_i^T\mathbf{v}_i)}{\sum_{j=1}^{M}\exp(\tau\mathbf{u}_i^T\mathbf{v}_j)} - \frac{1}{2M}\sum_{j=1}^{M}\log\frac{\exp(\tau\,\mathbf{v}_j^T\mathbf{u}_j)}{\sum_{i=1}^{M}\exp(\tau\,\mathbf{v}_j^T\mathbf{u}_i)}$$

$$-\frac{1}{M}\sum_{i=1}^{M}\sum_{t=1}^{T+1}\log p\left(\mathrm{w}_{i,t}|\mathrm{w}_{i,0:t-1}, \mathbf{x}_i; \theta, \phi, \psi\right)$$

The first and second terms represent image-to-text and text-to-image contrastive loss, respectively, to maximize the cosine-similarity scores between paired image and text embeddings relative to remaining negative pairings in the mini-batch. The last term seeks to maximize the log-likelihood of each observed token under the multimodal autoregressive language model (jointly parameterized by the image encoder, text encoder and multimodal decoder), conditioned on previous tokens in the caption, as well as the corresponding image. Each visual-language pretraining experiment was trained for 40 epochs, distributed across eight NVIDIA A100 80-GB graphics processing units (GPUs) with a local batch size of 48 per GPU, and gradient accumulation was used to achieve an effective global batch size of 1,536. We set the image size to 448 × 448 pixels, where larger images were first resized along the shorter edge and center-cropped, and smaller images were zero-padded as needed. For all optimization hyperparameters, refer to Supplementary Table 32.

Pretraining unimodal encoders

Prior work[56] showed that performing self-supervised pretraining of unimodal modules using unpaired data before joint visual-language pretraining using paired image–caption data can substantially improve downstream zero-shot transfer performance. We pretrained our image encoder using iBOT[74], a state-of-the-art, self-supervised pretraining algorithm for unlabeled image data. An in-house dataset of 16 million 256 × 256-sized image tiles were sampled and extracted at ×20-equivalent magnification from the tissue regions of 21,442 WSIs spanning over 350 cancer subtypes under the OncoTree classification system[65]. Detailed hyperparameters for image-only pretraining are provided in Supplementary Table 33. For pretraining the language model, we built a diverse corpus of pathology-relevant texts ranging from pathology educational texts to final diagnosis sections of over 550,000 surgical pathology reports from Massachusetts General Hospital and over 400,000 select histopathology-relevant PubMed abstracts. We used regex to deidentify in-house diagnostic reports, notably replacing patient and physician names, specimen identifiers, medical record numbers and dates with a corresponding special token in the vocabulary. We pretrained a 24-layer GPT-style autoregressive model using the next-word prediction loss. Specifically, given a sequence of word tokens $\mathbf{w} = (\texttt{<BOS>}, \mathrm{w}_1, \ldots, \mathrm{w}_T, \texttt{<EOS>})$, we maximized the log-likelihood of each token under an autoregressive generative model parameterized by ξ:

$$\mathcal{L}_{\mathrm{clm}}(\xi) = -\sum_{t=1}^{T+1} \log p\left(\mathrm{w}_t | \mathrm{w}_{0:t-1}; \xi\right)$$

Detailed hyperparameters for text-only pretraining are provided in Supplementary Table 34. After pretraining, the first 12 layers of the transformer-based language models and the embedding table were used to initialize the unimodal text encoder, while the last 12 layers and the language modeling classifier head were used to initialize the corresponding parameters in the multimodal decoder.

We assessed the benefit of unimodal pretraining by comparing downstream performance between the unimodal domain-specific pretraining scheme above versus CONCH with the image encoder pretrained on ImageNet versus CONCH with the language model randomly initialized (Extended Data Fig. 10). We found that CONCH with domain-specific pretraining outperformed CONCH with ImageNet pretraining on both zero-shot transfer and retrieval tasks. CONCH with the pretrained language model performed similarly to CONCH with a randomly initialized language model on classification and grading tasks but outperformed it in retrieval tasks.

Zero-shot transfer on ROIs and tiles

For zero-shot transfer, we used the method described in CLIP[30]. Each class was associated with a text prompt consisting of a class name (for example, '`adenocarcinoma`') and a template (for example, '`this is {}.`'; see Supplementary Table 35 for templates used across all tasks). For a prompt associated with class $j \in \{1, 2, \ldots, C\}$, we computed the ℓ_2-normalized embedding $\mathbf{v}_j$ using a text encoder trained on our paired dataset to form the linear classifier weights. Because model performance can vary considerably depending on the choice of prompts, we measured the performance spread by sampling subsets from a pathologist-curated set of prompts and reporting the median. Alternatively, we could also ensemble all the prompts within a class by using the mean embedding over the prompts as the text embedding associated with that class (see Extended Data Fig. 2 for a comparison with and without ensembling). Analogously, for each image, we computed the ℓ_2-normalized embedding $\mathbf{u}_i$. We then computed cosine-similarity scores between the image and each text embedding, and the predicted class was consequently the class with the highest similarity score:

$$\hat{y}_i = \underset{j}{\mathrm{argmax}}\ \mathbf{u}_i^T \mathbf{v}_j$$

Because some evaluation sets were imbalanced, we report the balanced accuracy (that is, the macro average over the accuracy obtained on each class) and the average F_1 score weighted by the support of each class. For SICAP, we also report the quadratic Cohen's κ score, which is often used for prostate Gleason grading[75], where errors between adjacent grading classes are penalized less.

Similarly, for cross-modal retrieval, we used the same method as zero-shot classification above to retrieve the top-K images that were closest in the aligned latent space to a specific text query (text-to-image retrieval). Image-to-text retrieval was performed analogously. To evaluate retrieval, we followed ALIGN[31] and used Recall@K, that is, for what percentage of the test set is the correct result in the top-K retrieved samples. We chose $K \in \{1, 5, 10\}$, and we also report mean recall by averaging the scores over the three Recall@K values.

Unless otherwise specified, we enforced the maximum image size to be 448 × 448 for CONCH through image resizing and center cropping, similar to its pretraining configuration. For all models that were not ours, we used their provided processor function and default configuration for image and text processing in downstream evaluation.

Extending zero-shot transfer to WSIs

To extend zero-shot transfer to gigapixel images, we followed the method introduced by MI-Zero[56]. Specifically, for classification over C classes, the WSI was first divided into N tiles, and the ℓ_2-normalized embeddings were computed independently using the image encoder. For each tile embedding, we computed similarity scores with each text embedding following the method for tiles described above, obtaining a set of C similarity scores for each tile. To aggregate similarity scores across tiles, we used the top-K pooling operator by averaging over the highest K similarity scores for each class to obtain the slide-level similarity score. Consequently, the class with the highest slide-level score was the predicted class. We chose $K \in \{1, 5, 10, 50, 100\}$, and we report metrics for the K value with the highest balanced accuracy for classification tasks and Cohen's κ for DHMC LUAD. Similarly to the classification of tiles, we report the slide-level balanced accuracy and weighted F_1 score for classification tasks. For DHMC LUAD, because the task of LUAD subtyping can be subjective, we report Cohen's κ score.

We performed zero-shot slide-level segmentation using a similar approach to that used for classification. We divided the WSI into tiles and computed similarity scores for each tile independently. However, instead of aggregating the scores across tiles into a single slide-level prediction, we mapped the tile-level scores to their corresponding spatial locations in the WSI, averaging the scores in overlapped regions. Finally, for each pixel, we assigned the class with the highest score as the prediction, producing a pixel-level segmentation mask. We computed the Dice score[76] to quantify the quality of the predicted segmentation mask relative to the ground truth.

Details of WSI preprocessing for both classification and segmentation tasks are described in Methods, 'WSI processing'.

Supervised and weakly supervised classification experiments

We performed supervised classification experiments on all tasks with a labeled set of training examples available, including TCGA BRCA for BRCA subtyping, TCGA NSCLC for NSCLC subtyping, TCGA RCC for

RCC subtyping, CRC100k for CRC tissue classification and SICAP for Gleason grading. For each dataset, we used the official training and testing split if it was available or we used the remaining labeled cases for training after holding out the cases used for zero-shot classification evaluation (see Methods, 'Downstream evaluation datasets' for a more detailed breakdown). For slide-level experiments, we considered four visual-language pretrained image encoders, namely, CONCH, PLIP, BiomedCLIP and OpenAICLIP. All four encoders followed the ViT-base architecture with a patch size of 16 except PLIP, which used a patch size of 32. For slide-level tasks, we additionally considered a ResNet50 encoder truncated after the third residual block, with weights initialized from supervised classification on ImageNet, as it has been a common choice in the weakly supervised classification of WSIs. For ROI-level tasks, we added CTransPath[62] as a baseline, which is a state-of-the-art general-purpose vision encoder trained with SSL on a large dataset of unlabeled histopathology images. We did not use CTransPath for TCGA slide-level tasks because TCGA slides (including those used in our test sets) made up a large portion of the data used to train CTransPath; therefore, this could have resulted in information leakage that unfairly inflated the performance of CTransPath on TCGA benchmarks.

For all experiments, we standardized the image input size to 224 × 224. We used each image encoder to extract a low-dimensional feature embedding from each image (tiles in the case of WSIs). For CONCH, we used the output of the attentional pooler that corresponded to image–text alignment, with an embedding dimension of 512. For CLIP-based models, including PLIP, BiomedCLIP and OpenAICLIP, we used the <CLS> token, which was also used for image–text alignment during pretraining and similarly had a dimension of 512. For ResNet50, we used global average pooling after the third residual block to obtain a 1,024-dimensional embedding. For CTransPath, we also used the <CLS> token representation, which had an embedding dimension of 768.

For WSI classification, we used the same preprocessing setup as zero-shot classification with MI-Zero. We used the widely used ABMIL[59] for weakly supervised classification of WSIs using slide-level labels. The ABMIL model architecture consists of a fully connected layer and a rectified linear unit (ReLU) nonlinearity that first maps the inputs to an embedding dimension of 512, followed by a two-layer, gated variant (as described in the original paper) of the attention network, with a hidden dimension of 384. Lastly, a fully connected classifier head maps the attention-pooled slide-level representation to logits, which are interpreted as class probabilities after softmax normalization. We used dropout with $P = 0.25$ after each intermediate layer in the network for regularization. We trained each model for 20 epochs on the training set, using an AdamW optimizer, a cosine learning rate scheduler and a learning rate of 1×10^{-4}. We used a weighted data sampler that increased the sampling probability of slides from minority classes such that, on average, the model saw the same number of slides from each class each epoch. The full set of hyperparameters is summarized in Supplementary Table 36.

For ROI-level classification, we conducted linear probing by training a logistic regression model on top of the pretrained image embeddings of each encoder. We followed a practice recommended by the large-scale self-supervised representation learning community[77] and set the ℓ_2 regularization coefficient λ to $\frac{100}{MC}$, where M is the embedding dimension and C is the number of classes. We used the limited-memory Broyden–Fletcher–Goldfarb–Shanno (L-BFGS) solver and set the maximum number of iterations to 800.

For few-shot classification, we kept the test set the same, and we varied the number of labeled examples per class for training (known as 'shot') from $n_c = 1, 2, 4, 8, 16, 32$, up to either $n_c = 512$ or the maximum number of labeled examples available for a given class. Otherwise, the hyperparameters and training setup remained the same as described above.

End-to-end fine-tuning for classification experiments

We evaluated the utility of CONCH in image ROI classification using standard end-to-end fine-tuning on a four-class Gleason grading benchmark with a total of 228,482 (training, 189,484; validation, 9,959; testing, 29,039) image ROIs individually labeled as NC, G3, G4 or G5 (see Methods, 'Downstream evaluation datasets' for more details). We compared its performance against that of five other models covering a variety of model architectures, pretraining strategies and sizes, including ViT-B/16 (ViT of the same architecture as the CONCH vision encoder backbone), ViT-L/16 (larger ViT with ~3.5 times the number of parameters as ViT-B), ResNet50 (popular, widely used standard CNN architecture), CTransPath (a histopathology-specific image encoder based on the Swin transformer architecture, pretrained using large-scale vision SSL, which has achieved state-of-the-art performance on many computational pathology tasks) and KimiaNet[64] (a lightweight CNN based on the DenseNet121 architecture, pretrained on a histopathology image classification task using supervised learning). For ViT-B/16, ViT-L/16 and ResNet50, we initialized the models using weights pretrained on ImageNet; for CTransPath and KimiaNet, we used the pretrained weights provided by their respective authors. We also investigated the label efficiency of each model by further subsampling 10% and 1% of labels from the full training set (189,484 ROIs from 4,622 slides) at the slide level, corresponding to 19,304 ROIs from 462 slides and 1,864 ROIs from 46 slides, respectively. The results are summarized in Supplementary Table 31.

We used eight 80-GB NVDIA A100 GPUs for each experiment using a batch size per GPU of 32 for ViT-L/16 (due to GPU memory constraints) and a batch size of 128 for all other models. All images were resized to 448 × 448 for both training and inference. We warmed up the learning rate over 250 steps and used the AdamW optimizer with $\beta = (0.9, 0.999)$ with fp16 automatic mixed precision training. For each model, we swept the learning rate over $\{1 \times 10^{-6}, 1 \times 10^{-5}, 1 \times 10^{-4}, 1 \times 10^{-3}, 1 \times 10^{-2}\}$ using the validation set. We trained for a maximum of 20 epochs and monitored the validation performance for early stopping with a patience of five epochs, using the best-performing model on the validation set for evaluation on the test set. We increased the maximum number of epochs to 40 and 80 for training with 10% labels and 1% labels, respectively, to account for the fewer training iterations per epoch, and we similarly increased the early-stopping patience to 10 and 20 epochs, respectively. We used standard data augmentation techniques during training, including random horizontal and vertical flips, discrete angle rotation ($\theta_{rot} \in \{0, 90, 180, 270\}$) and color jittering (brightness, 16/255; contrast, 0.125; saturation, 0.075; hue, 0.01).

Captioning with fine-tuning

Image captioning has been a widely explored task in the general visual-language domain[36,78,79]. In addition to distilling a top-level diagnosis of the image, image captioning can potentially provide morphological and contextual details, as well as additional interpretability, offering a much richer set of information than discrete labels. While prior studies[44,54,56] in visual-language pretraining showed applications in classification and retrieval, they are not equipped with generative capabilities. By adding a generative loss along with alignment and a text encoder module using the CoCa framework, our model is augmented with the ability to generate text conditioned on image inputs. We explored the captioning capabilities of CONCH on image–caption pairs extracted from a held-out source, source A, where a board-certified pathologist manually reviewed and condensed each caption such that it retained only information that could be inferred from the image, including the top-level diagnosis and detailed morphological descriptions. Given that our pretraining data were far from the scale of high-quality zero-shot captioning, we performed fine-tuning on the dataset. We partitioned the dataset into training, validation and testing splits and fine-tuned CONCH and baselines. Because PLIP and BiomedCLIP are not readily adaptable to captioning tasks, we compared the results against

https://doi.org/10.1038/s41591-024-02856-4

GenerativeImage2Text (GIT)[78], a widely used family of open-source visual-language pretrained models for image captioning.

We fine-tuned the entire model on a small training set of image–caption pairs. When fine-tuning CONCH, we simply set the contrastive loss to zero and kept only the captioning loss in the training objective. To evaluate performance, we report the commonly used metrics METEOR (metric for evaluation of translation with explicit ordering)[80] and ROUGE (recall-oriented understudy for gisting evaluation)[81]. For each model, we trained for a maximum of 40 epochs and selected the checkpoint with the highest METEOR on the validation set using an early-stopping patience of 10 epochs. At inference time, we generated captions using top-K sampling[82] as the decoding strategy with $K = 50$, where, at each timestep, the K most likely tokens were filtered and the probability mass was redistributed before sampling. Similar to zero-shot classification and retrieval, we set the maximum image size to 448 × 448. The full set of hyperparameters used to fine-tune captioning is presented in Supplementary Table 37.

Evaluation metrics

For classification tasks, we report balanced accuracy, weighted F_1 score and the area under the receiver operating characteristic curve (AUROC). Balanced accuracy is defined as the macro average of the recall of each class. Weighted F_1 score is computed by taking the average of the F_1 score (the harmonic mean of precision and recall) of each class, weighted by the support of each class. In the binary case, the AUROC is calculated from a plot of the true positive rate against the false positive rate as the classification threshold is varied. The AUROC is generalized to the multiclass case by averaging over the AUROC of all pairwise combinations of classes. For retrieval, we used the metric Recall@K, which is the proportion of the data correctly retrieved among the top-K retrieved samples. Following ALIGN[31], we chose $K \in \{1, 5, 10\}$, and we also computed the mean recall, which averages over the Recall@K values. For segmentation, we report the Dice score, which is the same as the F_1 score, and the precision and recall score, macro-averaged across all images and classes. For captioning, we report METEOR and ROUGE for comparing the predicted caption with the ground-truth caption. METEOR[80] is a metric based on unigram matching that considers both precision and recall between the original and ground truth and takes into account synonyms and word forms. ROUGE[81] computes the overlap of n-grams between the predicted caption and ground truth. We used ROUGE-1, which considers unigrams.

Downstream evaluation datasets

Source A was a dataset of image–caption pairs extracted from a held-out source. We split multipanel figures and matched them with captions manually. Because we also used this dataset for captioning, and because the captions were generally noisy and often contained information not present in the images, a board-certified pathologist cleaned the text, and we used the cleaned version for all downstream tasks. After filtering and cleaning, we obtained 797 images with an average width of 570 pixels and an average height of 428 pixels. We used this dataset in its entirety for cross-modal retrieval. We also used this dataset for captioning after performing a 70–10–20 split for training, validation and testing. To avoid information leakage, the dataset split was performed at the figure level (taking into account multifigure panels that were separated).

Source B was a dataset of image–caption pairs extracted from a held-out source. Similar to source A, we split multipanel figures and matched them with captions manually. After filtering and cleaning, we obtained 1,755 images with an average width of 512 pixels and an average height of 410 pixels. Because the dataset was much bigger than source A, we did not perform manual cleaning of the captions. We used this dataset for cross-modal retrieval.

TCGA LUAD consisted of 165 image–caption pairs extracted from 49 LUAD H&E histopathology slides from TCGA (https://portal.gdc.cancer.gov/).

For each slide, a board-certified pathologist chose up to five tiles of interest from each slide and provided captions describing the tissue pattern and any notable morphological features. This process yielded a set of 165 image tiles with an average width of 656 pixels and an average height of 642 pixels. We used this set of image tiles for cross-modal retrieval.

TCGA BRCA consisted of BRCA H&E formalin-fixed paraffin-embedded (FFPE) diagnostic histopathology WSIs from TCGA. This dataset consisted of cases for primary IDC and ILC. After removing slides with missing metadata, we collected a total of 1,048 slides (837 IDC and 211 ILC). The zero-shot test set was a sampled subset of the full TCGA RCC dataset consisting of 150 WSIs (75 for each class). For the supervised learning experiments, we held out the zero-shot test set as the test set and used the remaining slides as the supervised training set after excluding slides from patients who appeared in the test set. This process yielded a training set of 881 slides (754 IDC and 127 ILC; see Supplementary Table 38 for prompts used for each class in zero-shot classification).

TCGA NSCLC consisted of NSCLC H&E FFPE diagnostic histopathology WSIs from TCGA. This dataset consisted of cases of primary LUAD and lung squamous cell carcinoma (LUSC). After removing slides with missing or incorrect metadata, we collected a total of 1,041 slides (529 LUAD and 512 LUSC). The zero-shot test set was a sampled subset of the full TCGA RCC dataset consisting of 150 WSIs (75 for each class). For the supervised learning experiments, we held out the zero-shot test set as the test set and used the remaining slides as the supervised training set after excluding slides from patients who appeared in the test set. This process yielded a training set of 846 slides (432 LUAD and 414 LUSC; see Supplementary Table 38 for prompts used for each class in zero-shot classification).

TCGA RCC consisted of RCC H&E FFPE diagnostic histopathology WSIs from TCGA. This dataset consisted of cases of primary clear cell RCC (CCRCC), papillary RCC (PRCC) and chromophobe RCC (CHRCC). After removing slides missing low-resolution downsamples, we collected a total of 922 WSIs (519 CCRCC, 294 PRCC and 109 CHRCC). The zero-shot test set was a sampled subset of the full TCGA RCC dataset consisting of 225 WSIs (75 for each of the three classes). For the supervised learning experiments, we held out the zero-shot test set as the test set and used the remaining slides as the supervised training set after excluding slides from patients who appeared in the test set. This process yielded a training set of 693 slides (444 CCRCC, 215 PRCC and 34 ChRCC; see Supplementary Table 38 for prompts used for each class in zero-shot classification).

DHMC LUAD[83] consisted of 143 H&E LUAD slides, each labeled with the primary histologic growth pattern (59 acinar, 51 solid, 19 lepidic, 9 micropapillary and 5 papillary). We only used this dataset for zero-shot classification (see Supplementary Table 39 for prompts used for each class in zero-shot classification).

CRC100k[84] consisted of 224 × 224 pixel image tiles at 0.5 µm per pixel (mpp) extracted from 50 patients with colorectal adenocarcinoma. Each image belonged to one of nine classes: adipose, background, debris, lymphocytes, mucus, smooth muscle, normal colon mucosa, cancer-associated stroma or colorectal adenocarcinoma epithelium. For the supervised dataset, we used the officially provided splits of 100,000 images in the training set and 7,180 images in the test set. For the zero-shot test set, we used only the official test set (see Supplementary Table 40 for prompts used for each class in zero-shot classification).

WSSS4LUAD[85] consisted of LUAD image tiles of around 200–500 pixels in dimension, each labeled as tumor, tumor-associated stroma and/or normal. For our evaluation, we filtered for the samples with only one ground-truth label. We were left with 4,693 images from the official training split (see Supplementary Table 41 for prompts used for each class in zero-shot classification).

SICAP[75] consisted of 512 × 512 pixel images extracted from 155 WSIs of core-needle biopsies of prostate cancer, digitized at ×10 magnification. The official training and testing split partitioned the dataset into 9,959 images from 124 WSIs for training and 2,122 images from 31 WSIs for testing. Each tile was labeled with the primary Gleason pattern (G3, G4 or G5) or as noncancerous (NC). For zero-shot classification, we used only the official test set for evaluation, while, for supervised classification, we used the official splits for training and testing. For zero-shot segmentation (tumor versus benign), we used the slides from the official test split and corresponding pixel-level segmentation mask for evaluation (combining Gleason patterns G3, G4 and G5 as the tumor class; see Supplementary Table 41 for prompts used for each class in zero-shot classification and segmentation).

DigestPath[86] consisted of 660 colonoscopy H&E tissue section images from 324 patients, acquired at ×20-equivalent magnification. We used the subset of 250 images from 93 patients for which pixel-level lesion annotation for colorectal cancer tissue was provided, and we performed zero-shot segmentation evaluation (see Supplementary Table 41 for prompts used for each class in zero-shot segmentation).

EBRAINS[87,88] consisted of H&E histopathology WSIs of brain tissue from the EBRAINS Digital Tumor Atlas. We used a subset of 2,319 slides corresponding to a 30-way fine-grained brain tumor subtyping task, where only classes with at least 30 slides were kept to ensure that a reasonable number of slides were available for both model training and evaluation. For the supervised dataset, we performed a 50–25–25 split for training (1,151 slides), validation (595 slides) and testing (573 slides). For the zero-shot test set, we used the testing split of 573 slides (see Supplementary Tables 42–44 for prompts used for each class in zero-shot classification). The WSI counts for each class in the dataset were as follows: (1) *IDH1*-wild-type glioblastoma (474 slides); (2) pilocytic astrocytoma (173 slides); (3) meningothelial meningioma (104 slides); (4) pituitary adenoma (99 slides); (5) *IDH1*-mutant and 1p/19q codeleted anaplastic oligodendroglioma (91 slides); (6) ganglioglioma (88 slides); (7) hemangioblastoma (88 slides); (8) adamantinomatous craniopharyngioma (85 slides); (9) *IDH1*-mutant and 1p/19q codeleted oligodendroglioma (85 slides); (10) atypical meningioma (83 slides); (11) schwannoma (81 slides); (12) *IDH1*-mutant diffuse astrocytoma (70 slides); (13) transitional meningioma (68 slides); (14) diffuse large B cell lymphoma of the central nervous system (59 slides); (15) gliosarcoma (59 slides); (16) fibrous meningioma (57 slides); (17) anaplastic ependymoma (50 slides); (18) *IDH1*-wild-type anaplastic astrocytoma (47 slides); (19) metastatic tumors (47 slides); (20) *IDH1*-mutant anaplastic astrocytoma (47 slides); (21) ependymoma (46 slides); (22) anaplastic meningioma (46 slides); (23) secretory meningioma (41 slides); (24) lipoma (38 slides); (25) hemangiopericytoma (34 slides); (26) *IDH1*-mutant glioblastoma (34 slides); (27) non-Wingless-related integration (Wnt)/non-Sonic hedgehog (Shh) medulloblastoma (32 slides); (28) Langerhans cell histiocytosis (32 slides); (29) angiomatous meningioma (31 slides); and (30) hemangioma (30 slides).

Prostate Gleason Grading consisted of 228,482 image ROIs of H&E-stained prostate tissue curated from three publicly available datasets: AGGC[89], PANDA[90] and SICAP[75]. In the case of PANDA and AGGC, each ROI was extracted at ×10-equivalent magnification with dimensions 512 × 512 pixels and was labeled as NC, G3, G4 or G5, assigned using the pixel-level annotation masks provided by the respective dataset. We used this dataset to compare end-to-end fine-tuning performance between our model and other vision encoders commonly used in computational pathology. We partitioned the dataset at the slide level and split the dataset into training (189,000 ROIs from 4,622 slides in PANDA and the AGGC official training set), validation (10,000 ROIs from 124 slides in the SICAP official training set), and testing (29,000 ROIs from 92 slides in the official test sets of AGGC and SICAP).

WSI processing

For slide-level tasks, the processing pipeline for WSIs consisted of tissue segmentation, tiling and feature extraction. We used the CLAM library[7] for tissue segmentation, which computes a binary mask for tissue using binary thresholding along the saturation channel after converting a downsample of the slide from the RGB to hue–saturation–value (HSV) color space. Median blurring and morphological closing were used to smooth tissue contours and remove artifacts. The contours were filtered by area to yield the segmentation mask. For zero-shot and supervised classification, we followed previous conventions[7,62] and divided the segmented tissue regions into contiguous 256 × 256 pixel tiles at ×10-equivalent magnification. For segmentation, we extracted tiles using a smaller tile size (224 × 224 pixels) with 75% overlap at the highest magnification possible (that is, ×10 for SICAP and ×20 for DigestPath) to achieve more fine-grained predictions. After tiling, for feature extraction, we resized all tiles to 224 × 224 pixels and computed embeddings for each tile independently using a frozen pretrained image encoder, before caching them for downstream evaluation.

Pretraining dataset characterization

We estimated the distribution of topics covered by our pretraining captions. We first created a list of 19 topics that covered major anatomical sites relevant to the study of pathology. For each topic, a board-certified pathologist then curated a list of keywords associated with the topic. We then mapped a caption to a topic if it contained a specific word. Because it was impractical to curate an exhaustive set of keywords to cover all captions, we used k-nearest neighbors (kNN) with $k = 5$ to categorize the remaining captions. The distribution of captions on the topics is shown in Fig. 1b. Within each topic (as well as the overall dataset), we qualitatively visualized the contents of the captions using wordclouds (Extended Data Fig. 1).

Statistical analysis

Nonparametric bootstrapping with 1,000 samples was used to construct 95% confidence intervals for model performance. For each evaluation metric, observed differences in model performance were tested for statistical significance using a two-sided paired permutation test with 1,000 permutations. In each permutation, independent predictions of two models were randomly swapped to obtain a new difference in model performance. The P value was the proportion of differences in model performance greater than the observed difference in terms of absolute value. The null hypothesis was that there was no difference in model performance for the given test set and evaluation metric.

Computing hardware and software

We used Python (version 3.8.13) for all experiments and analyses in the study, which can be replicated using open-source libraries as outlined below. For task-agnostic pretraining, we used eight 80-GB NVIDIA A100 GPUs configured for multi-GPU training using DistributedDataParallel (DDP) as implemented by the popular open-source deep learning framework PyTorch (version 2.0.0, CUDA 11.7) (https://pytorch.org). All downstream experiments were conducted on single 24-GB NVIDIA 3090 GPUs. For unimodal pretraining of our visual encoder using iBOT, we modified the ViT implementation maintained by the open-source Timm library (version 0.9.2) from Hugging Face (https://huggingface.co) for the encoder backbone and used the original iBOT implementation (https://github.com/bytedance/ibot) for training. For natural language processing (NLP) workflows, we used open-source libraries provided by Hugging Face. Notably, we used Transformers (version 4.27.3) and Accelerate (version 0.15.0) for tokenization of text data and unimodal pretraining of our language model, and we used Evaluate (version 0.4.0) for accessing common machine translation and image captioning metrics including ROUGE (from rouge-score version 0.1.2) and METEOR (from nltk version 3.6.7). We integrated our pretrained unimodal visual encoder and language model into the open clip library (version 2.14.0) for visual-language pretraining using the CoCa framework. All WSI processing was supported by OpenSlide (version 4.3.1) and openslide-python (version 1.2.0). We used Scikit-learn (version 1.2.1) for its implementation of common machine learning model evaluation metrics for image classification and to train logistic regression models for linear probe experiments. Numpy (version 1.20.3) and Pandas (version 1.5.3) were used data collection and preparation. Implementations of other visual-language models benchmarked in the study were found on the Hugging Face model hub (https://huggingface.co/models): PLIP (https://huggingface.co/vinid/plip), BiomedCLIP (https://huggingface.co/microsoft/BiomedCLIP-PubMedBERT_256-vit_base_patch16_224), OpenAICLIP (https://huggingface.co/openai/clip-vit-base-patch16), GIT-base (https://huggingface.co/microsoft/git-base) and GIT-large (https://huggingface.co/microsoft/git-large). Pillow (version 9.3.0) and Opencv-python were used to perform basic image processing tasks. Matplotlib (version 3.7.1) and Seaborn (version 0.12.2) were used to create plots and figures. Usage of other miscellaneous Python libraries is listed in the Nature Portfolio Reporting Summary.

Reporting summary

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

Data availability

TCGA whole-slide data and labels are available from the NIH genomic data commons (http://portal.gdc.cancer.gov). DHMC LUAD whole-slide data and labels can be accessed through the Dartmouth Biomedical Informatics Research and Data Science website (http://bmirds.github.io/LungCancer/). SICAP whole-slide and tile data with corresponding labels can be accessed through the data portal at http://data.mendeley.com/datasets/9xxm58dvs3/1. CRC100k tile data and labels can be found at http://zenodo.org/record/1214456. WSSS4LUAD image tiles and labels can be found at http://wsss4luad.grand-challenge.org/. Pretraining data were curated from image–caption pairs in educational resources and PubMed. EBRAINS WSIs can be found at http://search.kg.ebrains.eu/instances/Dataset/8fc108ab-e2b4-406-8999-60269dc1f994. AGGC and PANDA WSIs can be accessed through their respective Grand Challenge portals (http://aggc22.grand-challenge.org/data/ and http://panda.grand-challenge.org/data/). The unprocessed PubMed Central Open Access dataset is available from the NIH PubMed Central website (http://ncbi.nlm.nih.gov/pmc/tools/openftlist/). Restrictions apply to the availability of anonymized patient data that were used retrospectively for this project with institutional permission and are, thus, not publicly available. All requests for processed or raw data collected or curated in house should be made to the corresponding author and will be evaluated according to institutional and departmental policies to determine whether the data requested are subject to intellectual property or patient privacy obligations.

Code availability

Model weights for CONCH can be assessed for academic research purposes at http://huggingface.co/MahmoodLab/conch. Code for using the pretrained model is provided at http://github.com/mahmoodlab/CONCH. We have documented all technical deep learning methods and software libraries used in the study while ensuring the paper is accessible to the broader clinical and scientific audience.

References

71. Redmon, J., Divvala, S., Girshick, R. & Farhadi, A. You Only Look Once: unified, real-time object detection. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition* 779–788 (IEEE, 2016).
72. Luo, R. et al. BioGPT: generative pre-trained transformer for biomedical text generation and mining. *Brief. Bioinform.* **23**, bbac409 (2022).
73. Dosovitskiy, A. et al. An image is worth 16×16 words: transformers for image recognition at scale. In *9th International Conference on Learning Representations* https://openreview.net/forum?id=YicbFdNTTy (OpenReview.net, 2021).
74. Zhou, J. et al. Image BERT pre-training with online tokenizer. In *10th International Conference on Learning Representations* https://openreview.net/forum?id=ydopy-e6Dg (OpenReview.net, 2022).
75. Silva-Rodriguez, J., Colomer, A., Dolz, J. & Naranjo, V. Self-learning for weakly supervised Gleason grading of local patterns. *IEEE J. Biomed. Health Inform.* **25**, 3094–3104 (2021).
76. Dice, L. R. Measures of the amount of ecologic association between species. *Ecology* **26**, 297–302 (1945).
77. Kolesnikov, A., Zhai, X. & Beyer, L. Revisiting self-supervised visual representation learning. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition* 1920–1929 (IEEE, 2019).
78. Wang, J. et al. GIT: a generative image-to-text transformer for vision and language. *Trans. Mach. Learn. Res.* https://openreview.net/forum?id=b4tMhpN0JC (2022).
79. Li, J., Li, D., Savarese, S. & Hoi, S. BLIP-2: bootstrapping language–image pre-training with frozen image encoders and large language models. In *Proc. 40th International Conference on Machine Learning* (eds Krause, A. et al.) 19730–19742 (PMLR, 2023).
80. Banerjee, S. & Lavie, A. METEOR: an automatic metric for MT evaluation with improved correlation with human judgments. In *Proc. ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization* 65–72 (Association for Computational Linguistics, 2005).
81. Lin, C.-Y. ROUGE: a package for automatic evaluation of summaries. In *Text Summarization Branches Out* 74–81 (Association for Computational Linguistics, 2004).
82. Lewis, M., Dauphin, Y. & Fan, A. Hierarchical neural story generation. In *Proc. 56th Annual Meeting of the Association for Computational Linguistics* (eds Gurevych, I. & Miyao, Y.) 889–898 (Association for Computational Linguistics, 2018).
83. Wei, J. W. et al. Pathologist-level classification of histologic patterns on resected lung adenocarcinoma slides with deep neural networks. *Sci. Rep.* **9**, 3358 (2019).
84. Kather, J. N. et al. Predicting survival from colorectal cancer histology slides using deep learning: a retrospective multicenter study. *PLoS Med.* **16**, e1002730 (2019).
85. Han, C. et al. WSSS4LUAD: Grand Challenge on weakly-supervised tissue semantic segmentation for lung adenocarcinoma. Preprint at https://doi.org/10.48550/arXiv.2204.06455 (2022).
86. Da, Q. et al. DigestPath: a benchmark dataset with challenge review for the pathological detection and segmentation of digestive-system. *Med. Image Anal.* **80**, 102485 (2022).
87. Roetzer-Pejrimovsky, T. et al. The Digital Brain Tumour Atlas, an open histopathology resource. *Sci. Data* **9**, 55 (2022).
88. Roetzer-Pejrimovsky, T. et al. The Digital Brain Tumour Atlas, an open histopathology resource [Data set]. *EBRAINS* https://doi.org/10.25493/WQ48-ZGX (2022).
89. Huo, X. et al. Comprehensive AI model development for Gleason grading: from scanning, cloud-based annotation to pathologist–AI interaction. Preprint at *SSRN* https://doi.org/10.2139/ssrn.4172090 (2022).
90. Bulten, W. et al. Artificial intelligence for diagnosis and Gleason grading of prostate cancer: the PANDA challenge. *Nat. Med.* **28**, 154–163 (2022).

Acknowledgements

We thank Jinghao Zhou for providing insights into the training dynamics for iBOT. We thank A. Song for his feedback. This work was supported in part by the BWH president's fund, BWH and MGH Pathology, and NIH NIGMS R35GM138216 (F.M.). M.Y.L. was also supported by the Siebel Scholars program. D.F.K.W. was also funded by the NIH NCI Ruth L. Kirschstein National Service Award, T32CA251062. R.J.C. was also supported by the NSF Graduate Fellowship. T.D. was also supported by the Harvard SEAS Fellowship. G.G. was supported by the BWH president's scholar award, NIGMS R35GM149270, NIDDK P30DK034854 and the Massachusetts Life Sciences Center. We thank T. Janicki, R. Kenny and the system administration staff at the MGB Enterprise Research Infrastructure and Services (ERIS) research computing core for maintaining the GPU computing resources that were instrumental in this study. We also thank T. Mages and T. Ramsey for their administrative support. The content is solely the responsibility of the authors and does not reflect the official views of the National Institutes of Health or the National Science Foundation.

Author contributions

F.M., M.Y.L., B.C. and D.F.K.W. conceptualized the study and designed the experiments. M.Y.L., B.C., R.J.C., T.D., I.L., D.F.K.W., I.O. and L.P.L. performed collection and cleaning of data used for unimodal and visual-language pretraining. M.Y.L., B.C. and R.J.C. performed model development. M.Y.L., B.C., D.F.K.W. and G.J. performed experimental analysis. M.Y.L., B.C., D.F.K.W., A.Z., R.J.C., I.L., T.D., G.J., F.M., G.G., L.P.L. and A.V.P. interpreted experimental results and provided feedback on the study. M.Y.L., B.C., D.F.K.W. and F.M. prepared the paper with input from all co-authors. F.M. supervised the research.

Competing interests

M.Y.L., B.C., R.J.C. and F.M. are inventors on a provisional US patent (application number 63/610,645) filed corresponding to the methodological aspects of this work.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41591-024-02856-4.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41591-024-02856-4.

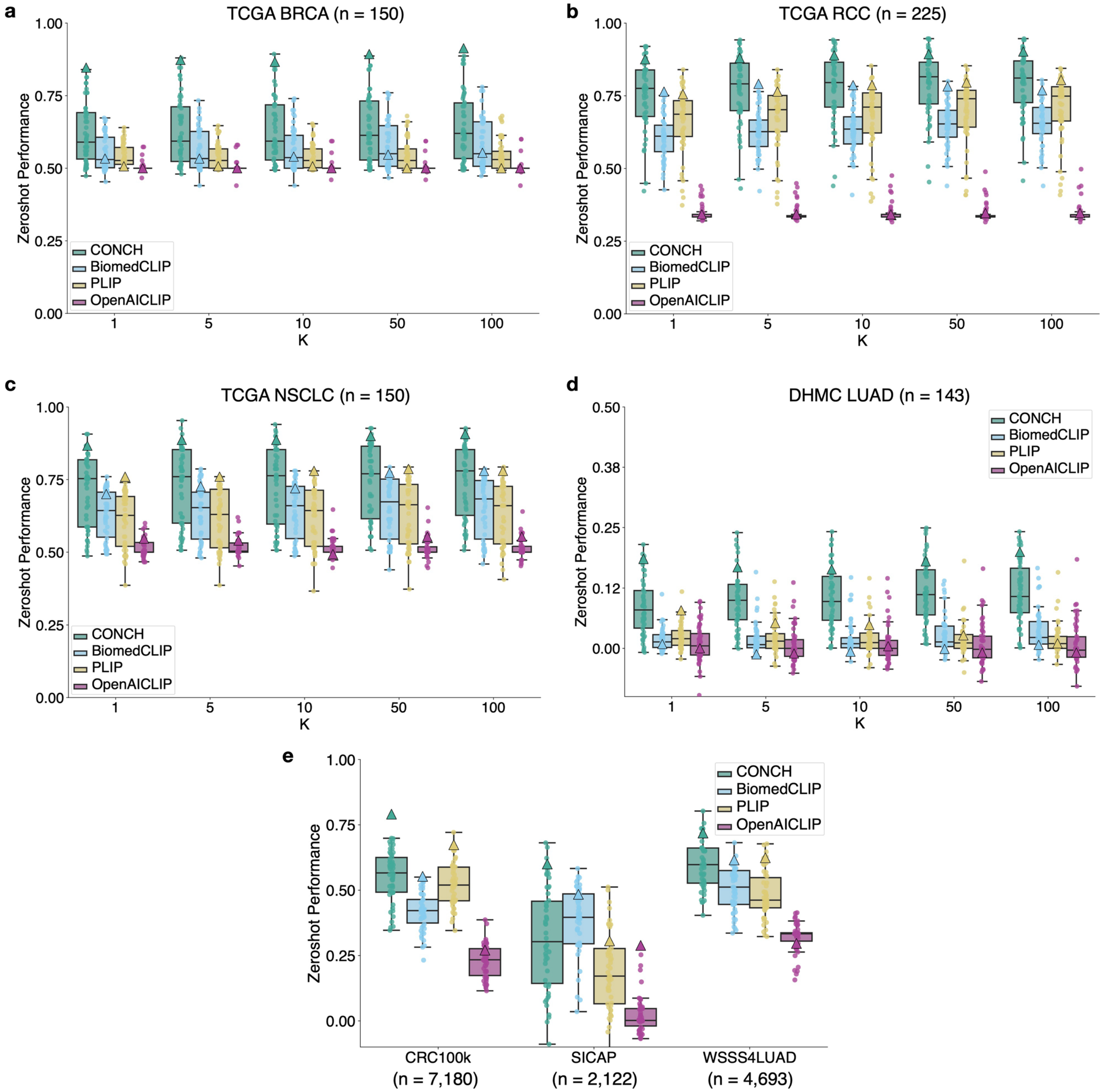
Correspondence and requests for materials should be addressed to Faisal Mahmood.

Peer review information *Nature Medicine* thanks Andrew Beck, Lee Cooper and Geert Litjens for their contribution to the peer review of this work. Primary Handling Editor: Lorenzo Righetto, in collaboration with the *Nature Medicine* team.

https://doi.org/10.1038/s41591-024-02856-4

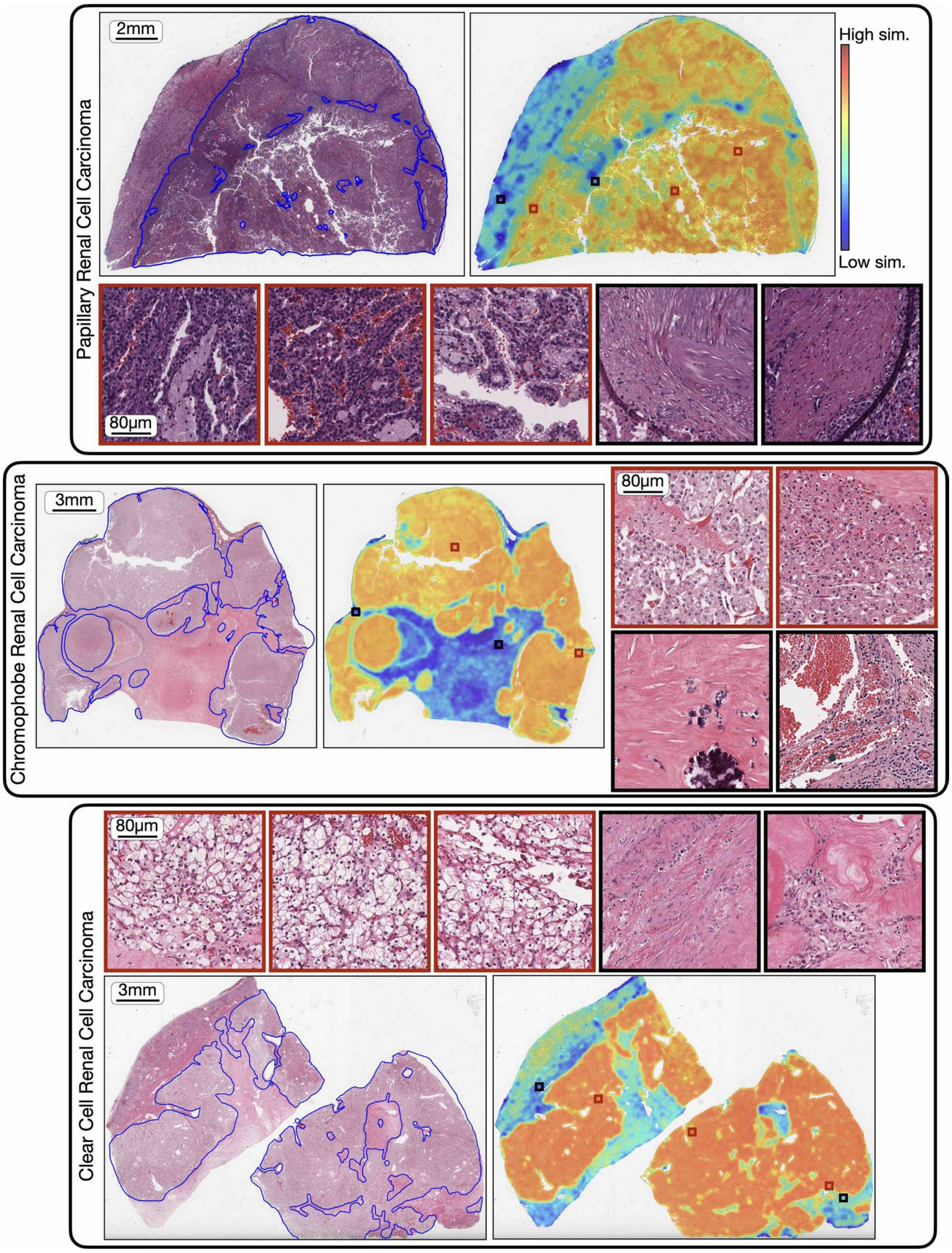
Extended Data Fig. 1 | Caption content of pre-training dataset. Wordclouds of captions to qualitatively visualize the caption content of each category in the pre-training dataset. Larger words are more represented in the captions. Common articles, nouns, and verbs are ignored.

https://doi.org/10.1038/s41591-024-02856-4



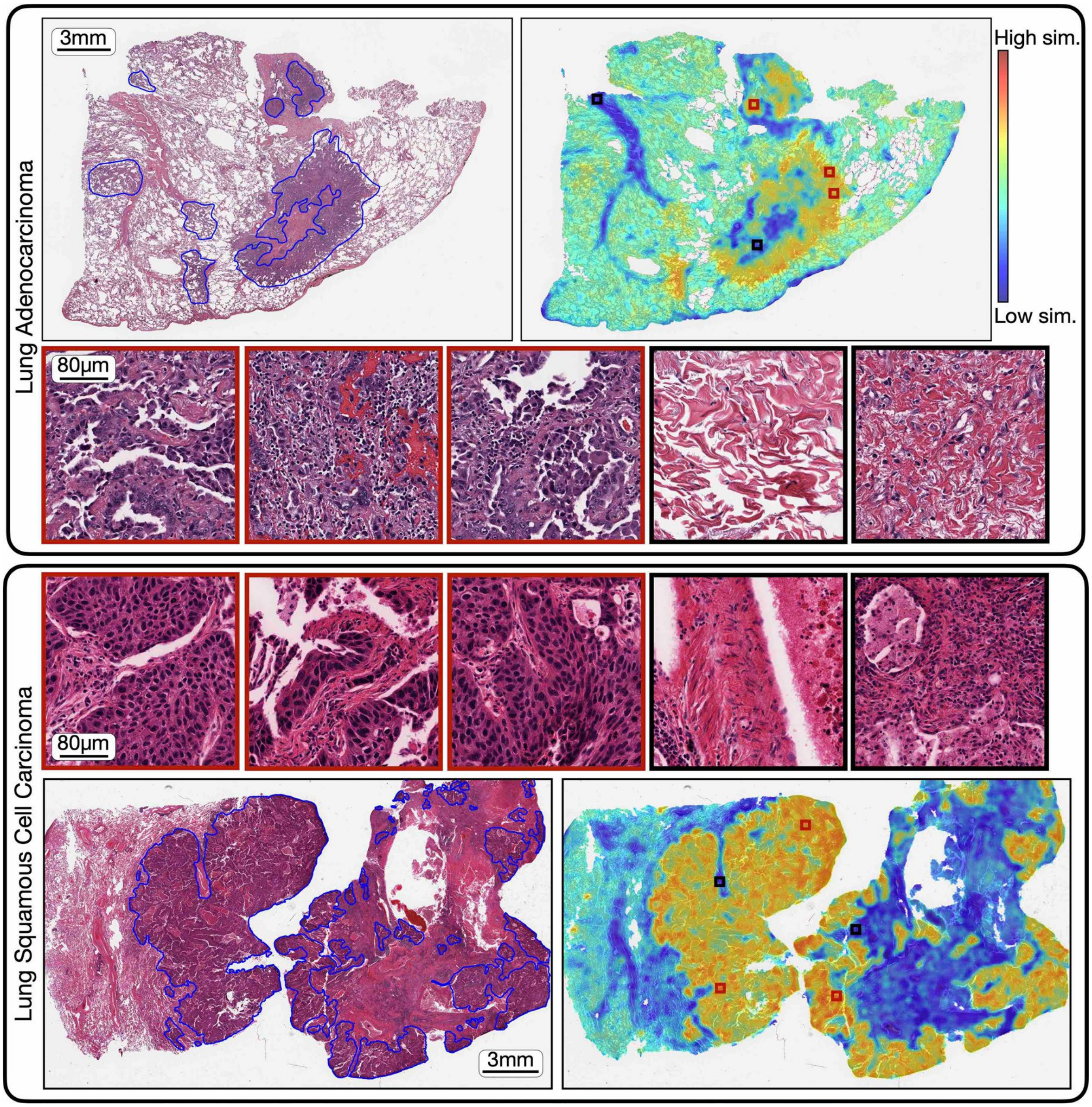
Extended Data Fig. 2 | Zero-shot classification: single prompt vs. ensembling. **a-d**, slide-level tasks. **e**, ROI-level tasks. We compare using a single text prompt per class *vs.* ensembling over multiple class names and templates. Since zero-shot performance of a visual-language pretrained model can be sensitive to the prompts used[52] when using a single prompt per class, for each class, we independently randomly sample a prompt from the pool of candidate templates and class names (see Supplementary Data Tables 34-38 for the prompt pools). We randomly sample 50 sets of prompts for each task, and plot the resulting distribution of zero-shot performance for each model using boxplot. Each dot corresponds to a single set of prompts ($n = 50$ for each box). Boxes indicate quartile values, and whiskers extend to data points within 1.5 × the interquartile range. Triangles indicate the performance of prompt ensembling. For slide-level tasks, we show performance for all *K*s used in top-*K* pooling. We observe prompt ensembling can substantially boost performance (relative to the median performance of randomly sampled single prompts) for most models in most tasks, except when the median performance is near random chance, such as for OpenAICLIP on most tasks and PLIP on TCGA BRCA. The poor median performance in these scenarios indicates that the model fails to perform under the majority of prompts sampled and therefore it is unsurprising that the ensembled prompt performs equally bad or worse. See Supplementary Data Tables 1-14 for more results.

 https://doi.org/10.1038/s41591-024-02856-4



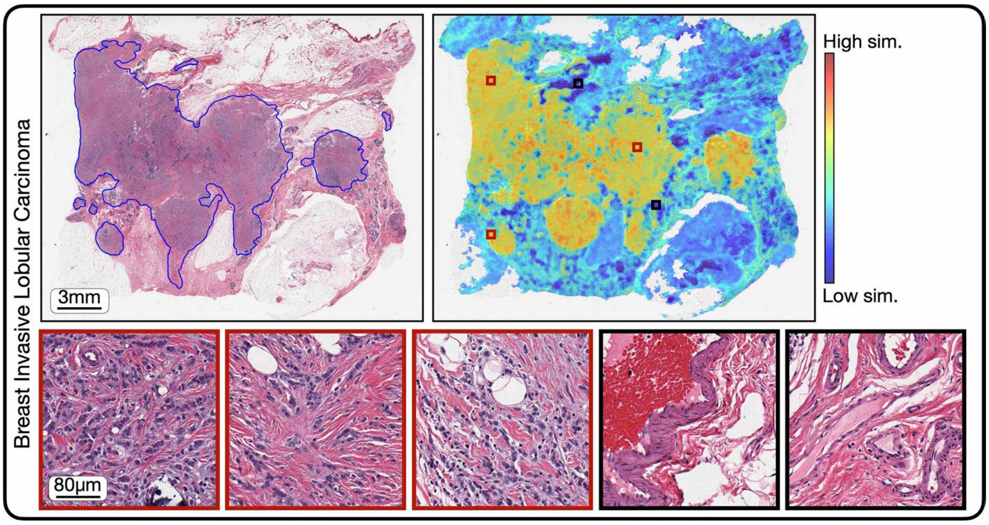
Extended Data Fig. 3 | CONCH heatmaps, renal cell carcinomas. Pathologist-annotated H&E images, corresponding cosine-similarity heatmaps of, from top to bottom, papillary, chromophobe, and clear cell renal cell carcinomas. Tiles of high similarity (red border) and low similarity (black border) with the predicted class label are randomly sampled and displayed next to each heatmap. We find excellent agreement between the annotated image and the regions of the slide with high similarity, with the tiles demonstrating stereotypical morphology of the tumors within the high-similarity regions and stroma or other normal constituents of the kidney in the low similarity regions.

 https://doi.org/10.1038/s41591-024-02856-4



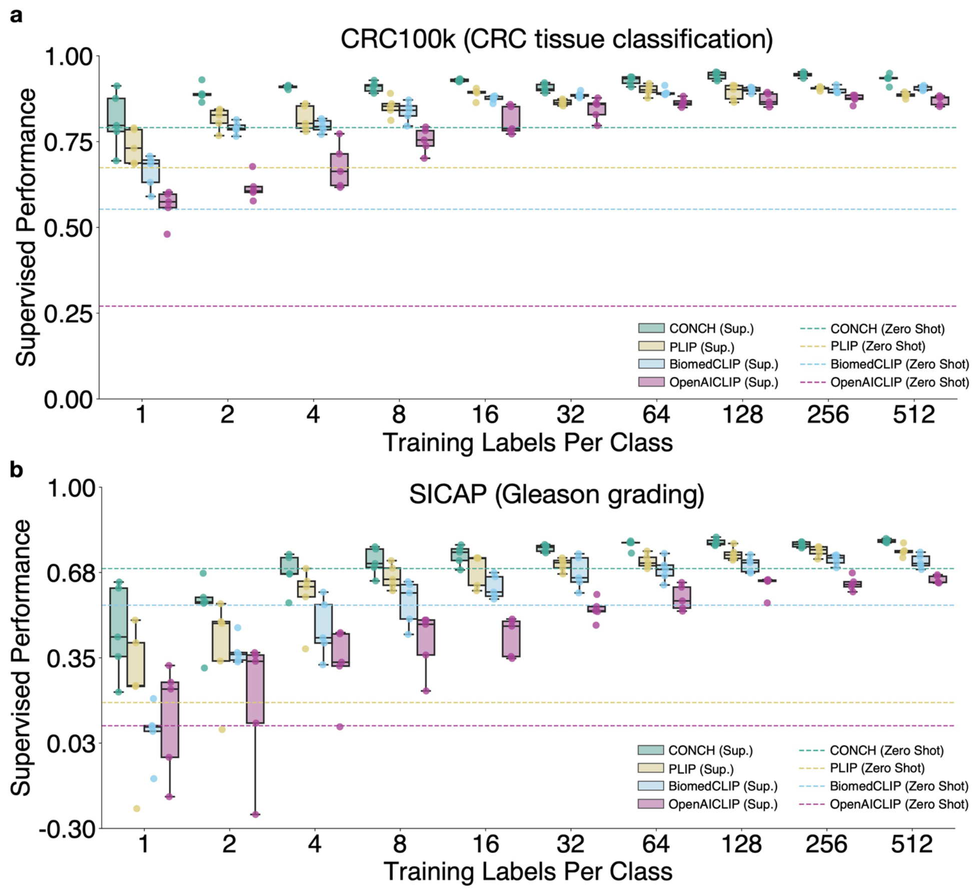
Extended Data Fig. 4 | CONCH heatmaps, non-small cell lung carcinomas. Pathologist-annotated H&E images, corresponding cosine-similarity heatmaps of adenocarcinoma (top) and squamous cell carcinoma (bottom) of the lung. Tiles of high similarity (red border) and low similarity (black border) with the predicted class label are randomly sampled and displayed next to each heatmap. We find excellent agreement between the annotated image and the regions of the slide with high similarity, with the tiles demonstrating stereotypical morphology of the tumors within the high-similarity regions and stroma or other normal constituents of the lung in the low similarity regions.

https://doi.org/10.1038/s41591-024-02856-4



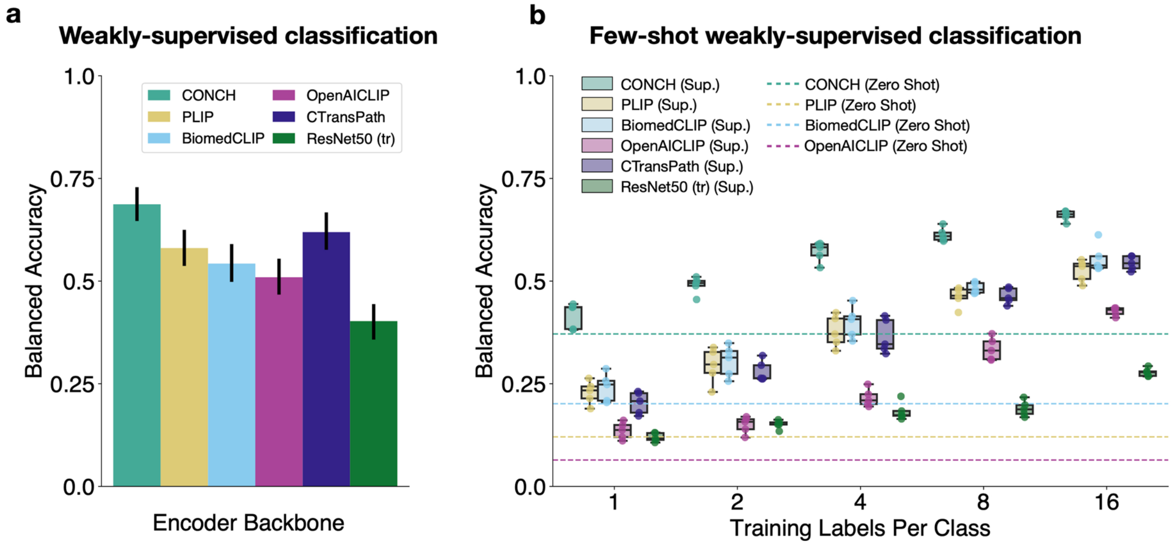
Extended Data Fig. 5 | CONCH heatmap, lobular carcinoma of the breast. Pathologist-annotated H&E image, corresponding cosine-similarity heatmap of lobular carcinoma of the breast. Tiles of high similarity (red border) and low similarity (black border) with the predicted class label are randomly sampled and displayed next to the heatmap. As with the ductal carcinoma heatmap in Fig. 2e, we find excellent agreement between the annotated image and the regions of the slide with high similarity, with the tiles demonstrating stereotypical morphology of lobular caricnoma within the high-similarity regions and stroma or other normal constituents of the breast in the low similarity regions.

https://doi.org/10.1038/s41591-024-02856-4



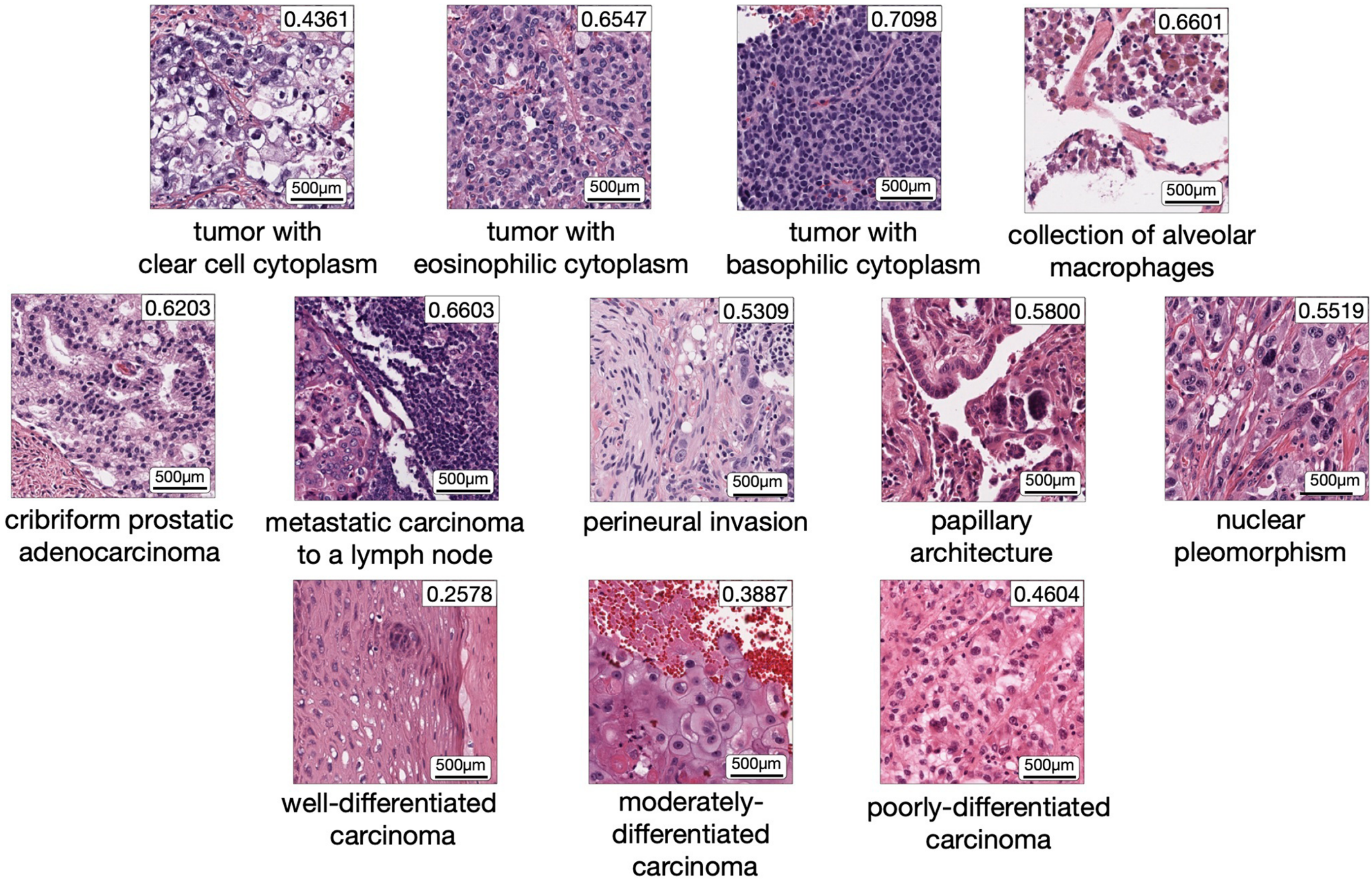
Extended Data Fig. 6 | ROI-level few-shot classification experiments. a, b. We investigate the label efficiency of different visual-language pretrained encoders in the few-shot setting where we vary the number of training labels per class (n_c), for n_c=1,2,4,8,16,… up to 512. For each n_c, we sample 5 different sets of training examples and perform linear probing on each training set using associated image labels (see **Supervised classification experiments** for details). We show their individual model performance via boxplot (*i.e.*, n = 5 for each box) to study the variance in model performance when performing supervised learning with very few training examples. Boxes indicate quartile values and whiskers extend to data points within 1.5 × the interquartile range. For reference, the zero-shot performance of each model is shown as a dotted line on the same plot. In terms of few-shot supervised learning, CONCH achieves better performance (*i.e.* in terms of the median accuracy of 5 runs) than other encoders for different sizes of training set and for all tasks. Additionally, in SICAP, we find CONCH zero-shot performance to be competitive with PLIP and BiomedCLIP few-shot up to 64 labels per class.

https://doi.org/10.1038/s41591-024-02856-4



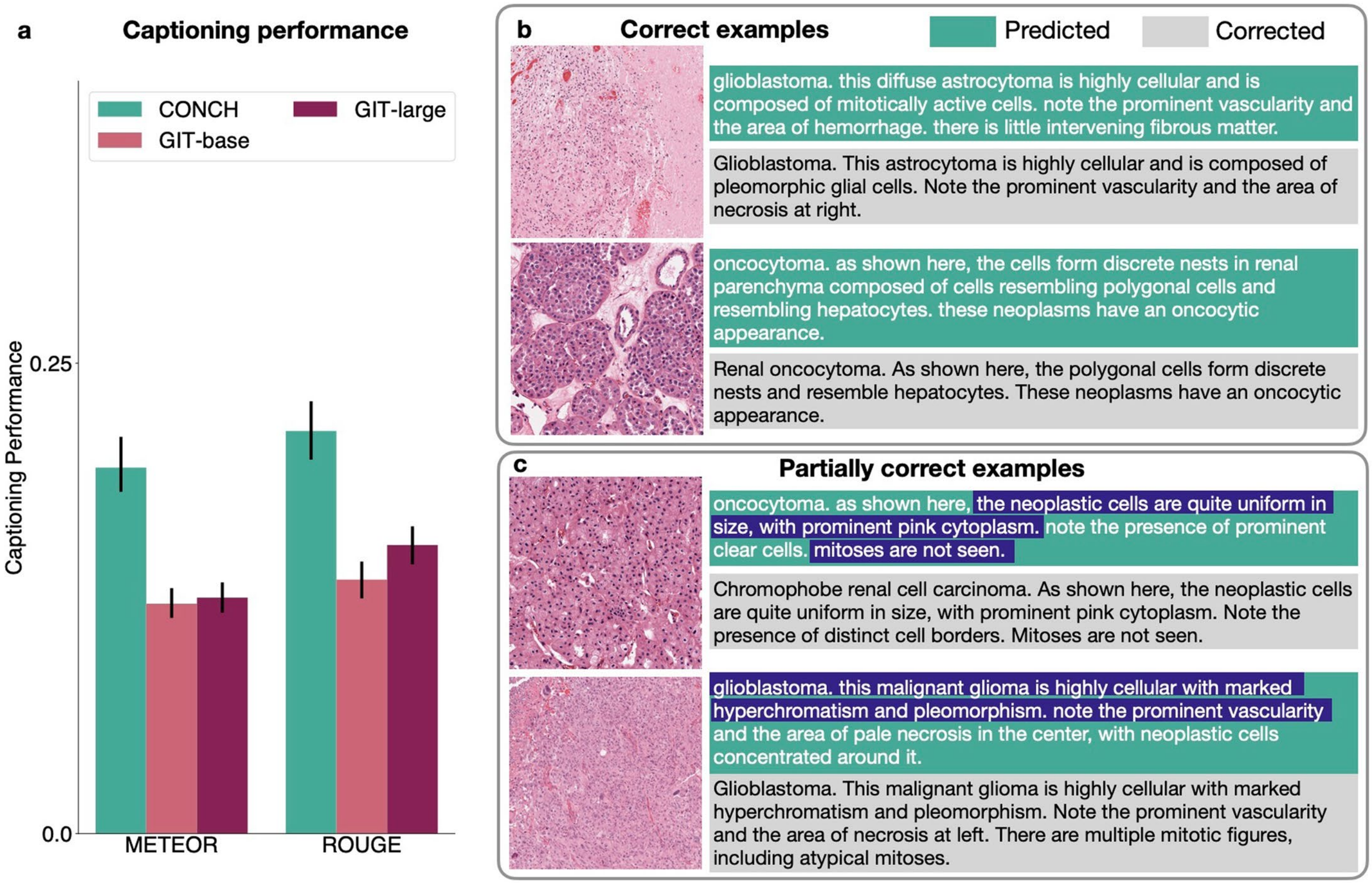
Extended Data Fig. 7 | Rare disease classification results on EBRAINS. **a**. Weakly-supervised ABMIL performance for CONCH and other pretrained encoder models on the EBRAINS 30-class brain tumor subtyping task ($n = 573$). Error bars represent 95% confidence intervals; the center is the computed value of balanced accuracy. **b**. We investigate the label efficiency of different pretrained encoders in the few-shot setting where we vary the number of training labels per class (n_c), for $n_c \epsilon \{1, 2, 4, 8, 16\}$. For each n_c, we sample 5 different sets of training examples and follow the experimental protocol in **a** to train an ABMIL model on each training set using associated slide labels (see **Supervised classification experiments** for details). We show their individual model performance via boxplot (*i.e.*, $n = 5$ for each box) to study the variance in model performance when performing supervised learning with very few training examples. Boxes indicate quartile values and whiskers extend to data points within 1.5 × the interquartile range. For reference, the zero-shot performance of each model is shown as a dotted line on the same plot. Additional metrics are reported in Supplementary Data Table 20 - 21. We find that CONCH consistently outperform all other visual language pretrained models in zeroshot classification and all pretrained encoders in weakly-supervised learning in terms of both performance and label efficiency.

 https://doi.org/10.1038/s41591-024-02856-4



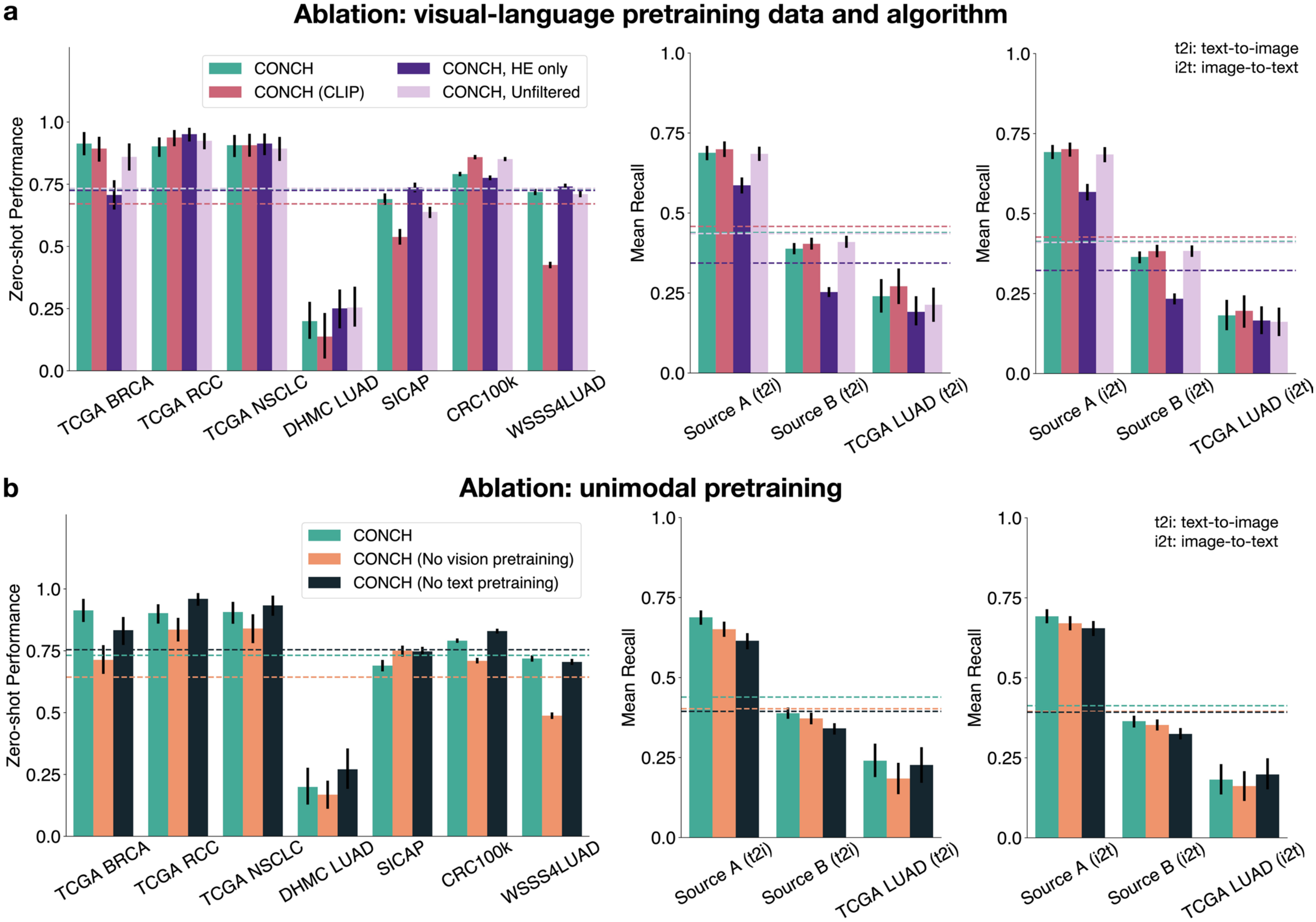
Extended Data Fig. 8 | Additional Retrieval Examples. Retrieved examples (among top 10) using complex prompts with detailed morphological information. Images are from an in-house dataset of tiles sampled from 1,620 cases held-out during pretraining, spanning 108 OncoTree codes (5 for each code). Similarity scores between each image and prompt are shown in the top-right corner of each image.

https://doi.org/10.1038/s41591-024-02856-4



Extended Data Fig. 9 | Image captioning results. a. Captioning performance of CONCH and baselines fine-tuned on Source A (train n=558, validation n=77, test n=162). The METEOR and ROUGE metrics are both calculated to evaluate the quality of generated captions. Captions were generated using top-K sampling with K=50 as the decoding strategy. Error bars representing 95% confidence intervals; the center is the computed value of each metric indicated by the x-axis label. CONCH outperforms both GIT baselines with $p < 0.01$. Although our absolute performance on these metrics is not ideal, image captioning is a considerably more difficult task than classification and retrieval, and we show that our pretraining data and approach can significantly improve performance over general visual-language models. **b**. Examples of captions generated by CONCH considered by a pathologist to be high quality. The green text boxes show generated captions and gray text boxes show captions corrected by a pathologist. **c**. Examples of partially correct captions generated by CONCH. Reasonably correct portions of the generated caption are highlighted in blue. In general, we noticed that some of the generated captions are regurgitated verbatim from the training dataset, likely due to the limited scale of fine-tuning (training split: n=558). Given that our current pretraining scale is still relatively small compared to works in the general visual-language domain, we expect the fine-tuned captioning performance to potentially improve substantially with more high-quality training data.

 https://doi.org/10.1038/s41591-024-02856-4



Extended Data Fig. 10 | CONCH pretraining ablations. In **a**, **b**, error bars represent 95% confidence intervals and the centres correspond to computed values of each metric as specified by the legend (**left**) or the y-axis label (**middle, right**). **a**. Comparison between CONCH pretrained on human-only data (n = 1,170,647) using CoCa *vs*. human-only data using CLIP *vs*. H&E only data (n = 457,372) *vs*. the full unfiltered dataset (n = 1,786,362). **Left**. Zero-shot performance on downstream subtyping (TCGA BRCA, n = 150; TCGA RCC, n = 225; TCGA NSCLC, n = 150; DHMC LUAD, n = 143; CRC100k, n = 7, 180; WSSS4LUAD, n = 4, 693) and grading (SICAP, n = 2, 122) tasks. Following pre-established conventions, quadratically weighted Cohen's κ is reported for SICAP and Cohen's κ is reported for DHMC LUAD, while balanced accuracy is reported for all other tasks. CONCH performs the best on average. **Middle and right**. Model performance in cross-modal retrieval on 3 datasets of image-text pairs (Source A, n = 797; Source B, n = 1,755; TCGA LUAD, n = 165). CONCH (CLIP) performs the best on average. **b**. Comparison between CONCH and no domain-specific unimodal pretraining. CONCH (No vision pretraining) replaces the image encoder pretrained on histopathology image patches with an analogous encoder pretrained on ImageNet. CONCH (No language pretraining) initializes the text encoder randomly instead of pretraining on pathology-related text. **Left**. Zero-shot performance on subtyping and grading tasks. **Middle and right**. Cross-modal retrieval performance.

Corresponding author(s): Faisal Mahmood

Last updated by author(s): Mar 3, 2024

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
☒	☐	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☒	☐	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. F, t, r) with confidence intervals, effect sizes, degrees of freedom and P value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's d, Pearson's r), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Scripts for data collection and preparation was written in Python (version 3.8.13) using the libraries numpy (version 1.20.3), pandas (version 1.5.3), and Pillow (version 9.3.0).
Data analysis	All WSI processing was supported by OpenSlide (version 4.3.1) and openslide-python (version 1.2.0). The CLAM package was used for feature extraction (http://github.com/mahmoodlab/CLAM). Code for data analysis was written in Python (version 3.8.13). The following Python libraries were used for analysis: HuggingFace Evaluate (version 0.4.0), nltk (version 3.6.7), rouge-score (version 0.1.2), scikit-learn (version 1.2.1), Pillow (version 9.3.0), opencv-python (version 4.7.0.72), matplotlib (version 3.7.1), seaborn (version 0.12.2), numpy (1.20.3), pandas (1.5.3), PyTorch (2.0.0, CUDA 11.7).

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

TCGA whole slide data and labels are available from the NIH genomic data commons (https://portal.gdc.cancer.gov). DHMC LUAD whole slide data and labels can be accessed through the Dartmouth Biomedical Informatics Research and Data Science website (https://bmirds.github.io/LungCancer/). SICAP whole slide and tile data with corresponding labels can be accessed through a data portal at https://data.mendeley.com/datasets/9xxm58dvs3/1. CRC100k tile data and labels can be found at (https://zenodo.org/record/1214456). WSSS4LUAD image tiles and labels can be found at (https://wsss4luad.grand-challenge.org/). EBRAINS whole slide images can be found at (https://search.kg.ebrains.eu/instances/Dataset/8fc108ab-e2b4-406f-8999-60269dc1f994). Pretraining data was curated from image-caption pairs in educational resources and PubMed. AGGC and PANDA whole slide images can be accessed through their respective Grand Challenge portal (https://aggc22.grand-challenge.org/Data/, https://panda.grand-challenge.org/data/). Educational resources are subject to copyright terms of publishers and will not be made available. The unprocessed PubMed Central Open Access dataset are available from the NIH PubMed Central website (https://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/). In house data are not publicly available. All requests for data collected or curated in-house will be evaluated based on institutional and departmental policies to determine whether the data requested is subject to intellectual property or patient privacy obligations.

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	Gender or sex was not included as a covariate in our experimental analysis at any stage of the study. Data pertaining to sex and gender may have been collected for data used in downstream tasks, which were curated by their original investigators, we refer readers to their original source for more detailed descriptions. For pretraining data sourced from the web and educational resources, no gender or sex data are made available in a structured format. For in-house BWH data used for pretraining, we provide the aggregate distribution of self-reported sex as follows: 11855 Female, 9575 Male, 12 Unspecified. For in-house BWH data used for in retrieval analysis (ED figure 8), the distribution is as follows: 908 Female, 711 Male, 1 Unspecified.
Reporting on race, ethnicity, or other socially relevant groupings	No covariates regarding race, ethnicity, and other social groupings were collected, used or analyzed in the study.
Population characteristics	No covariates relating to population characteristics were collected, used or analyzed in the study.
Recruitment	No patient recruitment was necessary for using histology whole slide images retrospectively.
Ethics oversight	Brigham and Women's Hospital IRB committee approved the retrospective analysis of pathology data.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

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

Sample size	No sample size calculations were performed. For pretraining data, we internally curated as much paired histology data as possible from educational resources and the PubMed Central Open Access data portal. After cleaning and curation, we yielded an unfiltered dataset of 1,786,362 cases, a human-only subset of 1,170,647 cases (on which we base most of analysis), and a human-only subset of 457,372 cases. For downstream datasets, see the Downstream evaluation datasets subsection of the Methods section in the manuscript for more details.
Data exclusions	For pretraining data, no additional data exclusions were performed after curation. For WSSS4LUAD, we excluded data that had more than one ground truth label. For EBRAINS, we excluded classes that had fewer than 30 total WSIs.
Replication	Replication was successful for all experiments in this study.

Randomization	For downstream evaluation that required creating train, validation, test splits, we created them randomly. We created 70% train, 10% validation, 20% test splits for captioning on the figure level, i.e., ensuring that extracted images from the same figure are only in the same split. For few-shot and supervised analyses, all splits were created stratified by class and on a patient level, i.e., ensuring that slides from the same patient are only in the same split.
Blinding	Blinding was not necessary for our study because our experiments were based on digitized histology slides or region-level images.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

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