nature medicine

Article

https://doi.org/10.1038/s41591-023-02504-3

A visual–language foundation model for pathology image analysis using medical Twitter

Received: 26 March 2023

Accepted: 18 July 2023

Published online: 17 August 2023

Check for updates

Zhi Huang[1,2,4], Federico Bianchi[3,4], Mert Yuksekgonul[3], Thomas J. Montine[2] & James Zou[1,3]

The lack of annotated publicly available medical images is a major barrier for computational research and education innovations. At the same time, many de-identified images and much knowledge are shared by clinicians on public forums such as medical Twitter. Here we harness these crowd platforms to curate OpenPath, a large dataset of 208,414 pathology images paired with natural language descriptions. We demonstrate the value of this resource by developing pathology language–image pretraining (PLIP), a multimodal artificial intelligence with both image and text understanding, which is trained on OpenPath. PLIP achieves state-of-the-art performances for classifying new pathology images across four external datasets: for zero-shot classification, PLIP achieves F1 scores of 0.565–0.832 compared to F1 scores of 0.030–0.481 for previous contrastive language–image pretrained model. Training a simple supervised classifier on top of PLIP embeddings also achieves 2.5% improvement in F1 scores compared to using other supervised model embeddings. Moreover, PLIP enables users to retrieve similar cases by either image or natural language search, greatly facilitating knowledge sharing. Our approach demonstrates that publicly shared medical information is a tremendous resource that can be harnessed to develop medical artificial intelligence for enhancing diagnosis, knowledge sharing and education.

Recent advances in artificial intelligence (AI) algorithms in computational pathology can help to distinguish cell or tissue types, generate diagnoses and retrieve relevant images from routinely stained hematoxylin and eosin (H&E) images[1–5]. Despite the availability of several high-quality datasets for task-specific machine learning, such as PanNuke[6], Lizard[7] and NuCLS[8], progress in computational pathology has been constrained by the need for more diversified datasets that include well-annotated labels in natural language. This data limitation is particularly noticeable when considering that there are more than 8,000 diseases[9] and their pathological classification is constantly evolving as knowledge of the molecular and cellular bases of disease advances[10]. While few-shot learning approaches with fine-tuning may alleviate this limitation[11,12], it is crucial to develop a general pathology AI system capable of serving multiple purposes.

At the same time, many de-identified pathology images are shared on the Internet, especially on social media[13], where clinicians discuss de-identified medical images with their colleagues[14–17]. For example, in Schaumberg et al.[13], researchers curated 13,626 images from Twitter in addition to PubMed article images, and developed machine learning models for multiple tasks. These public data and discussions hold

[1]Department of Biomedical Data Science, Stanford University School of Medicine, Stanford, CA, USA. [2]Department of Pathology, Stanford University School of Medicine, Stanford, CA, USA. [3]Department of Computer Science, Stanford University, Stanford, CA, USA. [4]These authors contributed equally: Zhi Huang, Federico Bianchi. e-mail: jamesz@stanford.edu

substantial value to the pathology community, enabling knowledge sharing and serving educational purposes. In particular, there is a comprehensive collection of pathology subspecialty-specific hashtags in the Twitter community, which were user-generated and systematically arranged by the 2016 United States and Canadian Academy for Pathology (USCAP) meeting[18,19]. This set of data includes both common and rare pathology cases, which were routinely stained with H&E dyes, and sometimes using immunohistochemistry. These public discourses represent an underutilized source of data and knowledge for medical AI[13,20,21].

In this study, we used the popular pathology Twitter hashtags to curate 243,375 public pathology images. We expanded this collection to include pathology data from other sites on the Internet (collected from the Large-scale Artificial Intelligence Open Network (LAION[22])), followed by strict data quality filtering, finally creating a collection of 208,414 pathology image–text pairs called OpenPath. To our knowledge, OpenPath is so far the largest publicly available pathology image collection that is annotated with text descriptions. We then leveraged this large-scale structured set of pathology image–text pairs to develop a versatile image and language AI foundation model for pathology. Unlike previous works, our approach integrates comprehensive natural language annotations into the learning process. By doing so, the model gains the capacity to understand image-based semantic knowledge, thereby empowering it to perform a wide range of downstream tasks.

This study first presents a complete description of the collected OpenPath dataset and proposes the PLIP (pathology language–image pretraining) model, which was trained on paired images and captions from OpenPath via contrastive learning. Next, a comprehensive evaluation of our proposed model was conducted to assess its ability to adapt to new captions through zero-shot learning[23]. Moreover, PLIP can function as a general purpose image encoder to capture a better image representation for pathology, making it possible to train and classify new sets of data via linear probing, leading to improved performances across diverse tissue types and learning tasks. This general purpose image encoder can be particularly helpful on clinical tasks with limited annotated data. Finally, PLIP enables a flexible search engine for pathology images, which can serve as a powerful educational and information sharing tool for clinicians and pathology trainees. We conducted a systematic evaluation of image retrieval[4,5] to demonstrate its ability to retrieve relevant pathology images by text or image inputs, which holds tremendous knowledge sharing potential.

Results

Creating OpenPath from Twitter and other public sources

The USCAP and the Pathology Hashtag Ontology projects[24] recommends 32 Twitter pathology subspecialty-specific hashtags[18,19]. We used these 32 hashtags to retrieve relevant tweets from 21 March 2006 (the date of the first Twitter post) to 15 November 2022 (Fig. 1a) to establish so far the largest public pathology dataset with natural language descriptions for each image: OpenPath. The detailed definition of each hashtag is presented in Extended Data Table 1. We followed the usage policy and guidelines from Twitter and other entities in retrieving the data. To ensure data quality, OpenPath followed rigorous protocols for cohort inclusion and exclusion, including the removal of retweets, sensitive tweets and non-pathology images, as well as additional text cleaning (Fig. 1a, Extended Data Fig. 1 and Methods). The final OpenPath dataset (Fig. 1b) consists of: (1) tweets: 116,504 image–text pairs from Twitter posts (tweets) across 32 pathology subspecialty-specific hashtags (Fig. 1c); (2) replies: 59,869 image–text pairs from the associated replies that received the highest number of likes in the tweet, if applicable (Fig. 1c); and (3) PathLAION: 32,041 additional image–text pairs scraped from the Internet and the LAION dataset. The captions in OpenPath used a median number of 17 words (Fig. 1d and Supplementary Table 1) to describe the medical conditions in the corresponding images. The detailed dataset extraction and description are elaborated further in the Methods, and the complete dataset of the inclusion–exclusion procedure is demonstrated in Extended Data Fig. 1.

Training a visual–language AI using OpenPath

Unlike other supervised learning and segmentation pathology models that were trained solely on categorical labels, texts in natural language are enriched with semantic and interrelated knowledge, which can further enhance the understanding of the images and facilitate multiple downstream applications. In this study, we fine-tuned a pretrained contrastive language–image pretraining (CLIP) model[25] on OpenPath using contrastive learning. To accomplish this, a pathology image preprocessing pipeline was integrated, including image downsampling, random cropping and data augmentations (Methods). During the training phase, the PLIP model generates two embedding vectors from both the text and image encoders (Fig. 1e). These vectors were then optimized to be similar for each of the paired image and text vectors and dissimilar for non-paired images and texts via contrastive learning (Fig. 1f and Methods).

PLIP can handle multiple types of inferences across a broad spectrum of medical applications, which does not require explicit training. In subsequent sections, we demonstrate how PLIP can be used to perform different downstream tasks.

PLIP can classify new images without further training

In this study, we conducted a systematic evaluation of PLIP's zero-shot capability[23], which enables learning new classes at scale without the need for retraining (Fig. 2a). The evaluation was performed across four external validation datasets: (1) the Kather colon dataset[26] with nine different tissue types; (2) the PanNuke dataset[6] (benign and malignant); (3) the DigestPath dataset[27] (benign and malignant); and (4) the WSSS4LUAD dataset[28] (tumor and normal) (Fig. 2b and Supplementary Fig. 1). PLIP was evaluated on those datasets by converting labels to sentences, for example, 'tumor' to 'an H&E image of tumor'. As a natural comparison, we compared PLIP with the original CLIP model, which has been frequently used for other medical image tasks[29–31] and has already been trained from other medical images. By evaluating the weighted F1 score, which is a combined measure of precision and recall while accounting for class imbalances, our analysis showed that PLIP consistently outperformed the baseline CLIP model and the results from predicting the majority class (or Majority) (Fig. 2c). For instance, PLIP achieved F1 = 0.565 (95% confidence interval (CI) = 0.559–0.572) on the Kather colon dataset (nine classes). In the PanNuke dataset (benign versus malignant), PLIP achieved F1 = 0.656 (95% CI = 0.639–0.667). In the DigestPath dataset (benign versus malignant), PLIP achieved F1 = 0.832 (95% CI = 0.829–0.834). In WSSS4LUAD (tumor versus normal), PLIP achieved F1 = 0.734 (95% CI = 0.723–0.745). All these performances are substantially higher than CLIP at predicting the majority class (Supplementary Tables 3 and 4). The Matthews correlation coefficient (MCC) was also calculated and showed that PLIP was the best model across all the datasets (Supplementary Table 4).

In addition, the confusion matrix with ground truth and predicted annotations for the Kather colon dataset is presented in Fig. 2d. Compared to other models (Extended Data Fig. 2), PLIP exhibited reasonable zero-shot learning capabilities for this sophisticated task, which can accurately distinguish several key tissue types, including adipose tissue (ADI), background (BACK), colorectal carcinoma epithelium and lymphocytes (LYMs). However, it may face challenges when dealing with other tissue types, such as mucus (MUC) and debris (DEB). This observation is likely due to MUC, DEB and smooth muscle (MUS) sometimes being considered cancer-associated stroma (STR).

Moreover, we conducted a more in-depth investigation into the zero-shot performances to determine benign versus malignant among samples within each of 19 different organs in the PanNuke dataset. Compared to the baseline CLIP in Fig. 2e, we found that PLIP achieved superior weighted F1 score on 14 out of 19 organs (Supplementary

https://doi.org/10.1038/s41591-023-02504-3

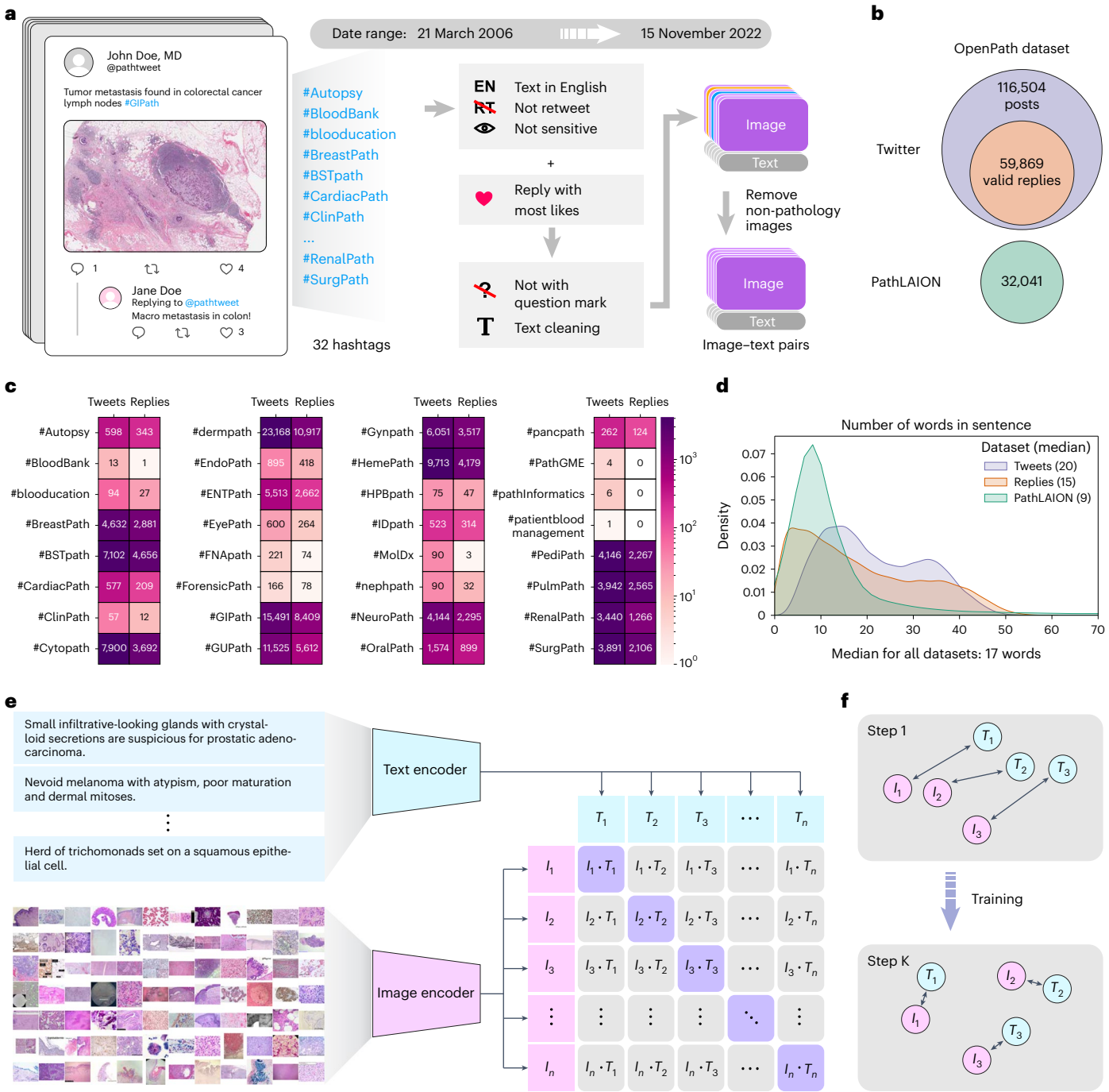


Fig. 1 | Overview of the study. a, Flowchart of data acquisition from medical Twitter. **b**, Overview of the OpenPath dataset. **c**, Total number of available image–text pairs from tweets and replies within each Twitter hashtag (sorted in alphabetical order). Replies are those that received the highest number of likes in Twitter posts, if applicable. **d**, Density plot of the number of words per sentence in the OpenPath dataset. **e**, The process of training the PLIP model with paired image–text dataset via contrastive learning. **f**, Graphical demonstration of the contrastive learning training process.

Table 3). Among them, seven subspecialties (adrenal gland, esophagus, liver, ovarian, stomach, testis and uterus) achieved reasonably high F1 scores (>0.8), whereas the baseline CLIP performed only at F1 = 0.3–0.6. However, imbalanced classes were observed among samples from some organs, such as kidney, lung and prostate, and this may have contributed to inferior performances by PLIP.

PLIP improves image representations for training models

To gain a deeper understanding of the capabilities of the PLIP image encoder, we used four different testing datasets (Kather colon[26], PanNuke[6], DigestPath[27] and WSSS4LUAD[28]) to evaluate the ability of image representation. Image embeddings were first calculated by the PLIP image encoder; then dimensionality reduction was applied via uniform manifold approximation and projection (Fig. 3a–d). Without training on these datasets, we found that PLIP could still effectively distinguish between various tissue subtypes in the Kather colon dataset (Fig. 3a). Compared to the performance of other baseline models (for example, the CLIP model in Extended Data Fig. 3b), PLIP effectively differentiated between normal colon mucosa (NORM) and colorectal adenocarcinoma epithelium (TUM), which both shared similar morphological and textural patterns.

In the PanNuke dataset, PLIP revealed intriguing organ-specific separations, particularly highlighting breast and colon subsets. Indeed, the colon set formed two relatively clean subclusters, with one being enriched with the malignant tissue images (Fig. 3b and Extended Data Fig. 4). Furthermore, the PLIP model separated normal and benign from tumoral and malignant image patches from the DigestPath and WSSS4LUAD datasets, respectively (Fig. 3c,d and Extended Data Figs. 5 and 6). For DigestPath, we also noticed a clear separation between different image downsampling rates and staining variations (Extended Data Fig. 7).

Encouraged by these findings, we hypothesized that the PLIP image encoder may serve as a preferred pretrained backbone for several pathology image classification tasks. To this end, we trained a simple linear classifier on top of the image embedding vectors (Fig. 3e) from the training splits of four different datasets (Kather colon, PanNuke, DigestPath and WSSS4LUAD; Fig. 2b and Supplementary Fig. 1a), and compared the classification performances on the testing splits of four datasets with two baseline models: image encoder from the original CLIP model and the multitask pretraining of deep neural networks[32] (MuDiPath).

By taking text description in natural language into account, our PLIP model sets itself apart from what others have done (for example, MuDiPath), and demonstrated superior performances across all four testing datasets (average macro F1: PLIP = 0.891, CLIP = 0.813, MuDiPath = 0.866; Fig. 3f). In the Kather colon dataset with nine-class classification, PLIP achieved F1 = 0.877, surpassing the second-best model MuDiPath (F1 = 0.825) with a 6.30% improvement ($P = 9.4 \times 10^{-12}$). In the PanNuke dataset with binary classification, PLIP achieved F1 = 0.902, which was the highest score among the compared models. In the DigestPath dataset, PLIP also improved by 3.5% in terms of F1 = 0.856 when compared to the second-best model ($P = 6.2 \times 10^{-4}$). Finally, PLIP achieved the highest F1 = 0.927 in the WSSS4LUAD datasets. Furthermore, an ablation study was conducted with different combinations of datasets and the PLIP model performed best when all the OpenPath data (Tweets + Replies + PathLAION) were used in training (Supplementary Table 5).

To gain further insight into the benefits of PLIP, we compared PLIP with the end-to-end deep learning model ViT-B/32 by fine-tuning them on four external validation datasets. ViT-B/32 is a state-of-the-art model and has the same architecture as the PLIP image encoder, thus facilitating a direct comparison of the benefit of PLIP's contrastive learning. We assessed the fine-tuning performance in terms of data efficiency by using different proportions of training data (1%, 5%, 10%, 50% and 100%). From Extended Data Fig. 8, our findings indicate that PLIP achieves better performances; the improvement over end-to-end supervised learning is especially large when the training set size is small.

The results suggest that PLIP can achieve comparable or higher performances compared to traditional deep learning models trained solely with categorical labels. This is probably due to the rich content of the text annotations, which may enable the model to exploit higher-level semantic visual–linguistic relationships and provide a more comprehensive understanding of the images, including both visual and subvisual intercellular patterns.

PLIP enhances pathology image retrieval from text inputs

PLIP can identify and retrieve the most relevant images based on a given text input, also known as text-to-image retrieval[33] (Fig. 4a). To evaluate this ability, we collected four sets of images with captions: (1) Twitter validation dataset (Twitter); (2) PathPedia images (PathPedia); (3) PubMed pathology images[34] (PubMed); and (4) pathology book collections[34] (Books) (Fig. 4b). The Twitter validation dataset contained 2,023 paired image–text from 16 November 2022 to 15 January 2023 (Fig. 4c and Extended Data Fig. 1b), and was expected to have a similar image–text distribution to what the PLIP model had trained on. In contrast, PathPedia (number of candidates for image retrieval = 210), PubMed (1,419 image–text pairs) and Books (558 image–text pairs) consisted of relatively concise texts (Fig. 4d).

We evaluated image retrieval performance using the Recall@10 and Recall@50 metrics on the Twitter validation dataset (Methods). Finding the exact image associated with a given text is challenging because of many similar images that could match one description. Nonetheless, we found that PLIP greatly improved image retrieval performances with Recall@10 = 0.271 (4.5× higher than CLIP) and Recall@50 = 0.527 (4.1× higher than CLIP) (Fig. 4e). With a large pool of candidates ($n = 2{,}023$) and different tissue types, the 52.7% chance of retrieving the target image within the top 50 demonstrates a challenging yet achievable task.

In addition, PLIP demonstrated a substantial improvement in performance compared to both the baseline CLIP and random performances across different datasets (Fig. 4e). In the PathPedia collection, PLIP achieved Recall@10 = 0.409 and Recall@50 = 0.752. In the PubMed pathology image collection, PLIP achieved Recall@10 = 0.069 and Recall@50 = 0.206. In the Books pathology image collection, PLIP achieved Recall@10 = 0.265 and Recall@50 = 0.659. On average, these performances are 2–5 times higher than the baseline CLIP model. PLIP demonstrated the largest advantage over baseline methods on the Twitter validation dataset (fold change = 55.3 and 21.4 for Recall@10 and Recall@50 compared to random retrieval). The PathPedia image collection showed the least improvement, probably due to the curated PathPedia collection not covering all of the nuances and variations in the text for the pathology images.

We further conducted the text-to-image retrieval tasks for the Twitter validation dataset within each of the top ten hashtags that had more than 100 available candidates (Fig. 4f,g and Extended Data Fig. 9a,b). When measured with Recall@10, we found that the gynecological pathology (#Gynpath) may benefit most from the PLIP model (Recall@10 = 0.557) (Fig. 4f). When measured with Recall@50, head and neck pathology (#ENTPath) benefited most from the PLIP model (Recall@50 = 0.925) (Extended Data Fig. 9a). Furthermore, the Spearman correlation analysis suggested that the performance improvements of the PLIP model were significantly correlated with the number of candidate images. For Recall@10, $\rho = 0.88$ ($P = 8.14 \times 10^{-4}$) for PLIP versus random, and $\rho = 0.64$ ($P = 4.79 \times 10^{-2}$) for PLIP versus CLIP (Fig. 4g). For Recall@50, $\rho = 0.98$ ($P = 1.47 \times 10^{-6}$) for PLIP versus random, and $\rho = 0.85$ ($P = 1.64 \times 10^{-3}$) for PLIP versus CLIP (Extended Data Fig. 9b).

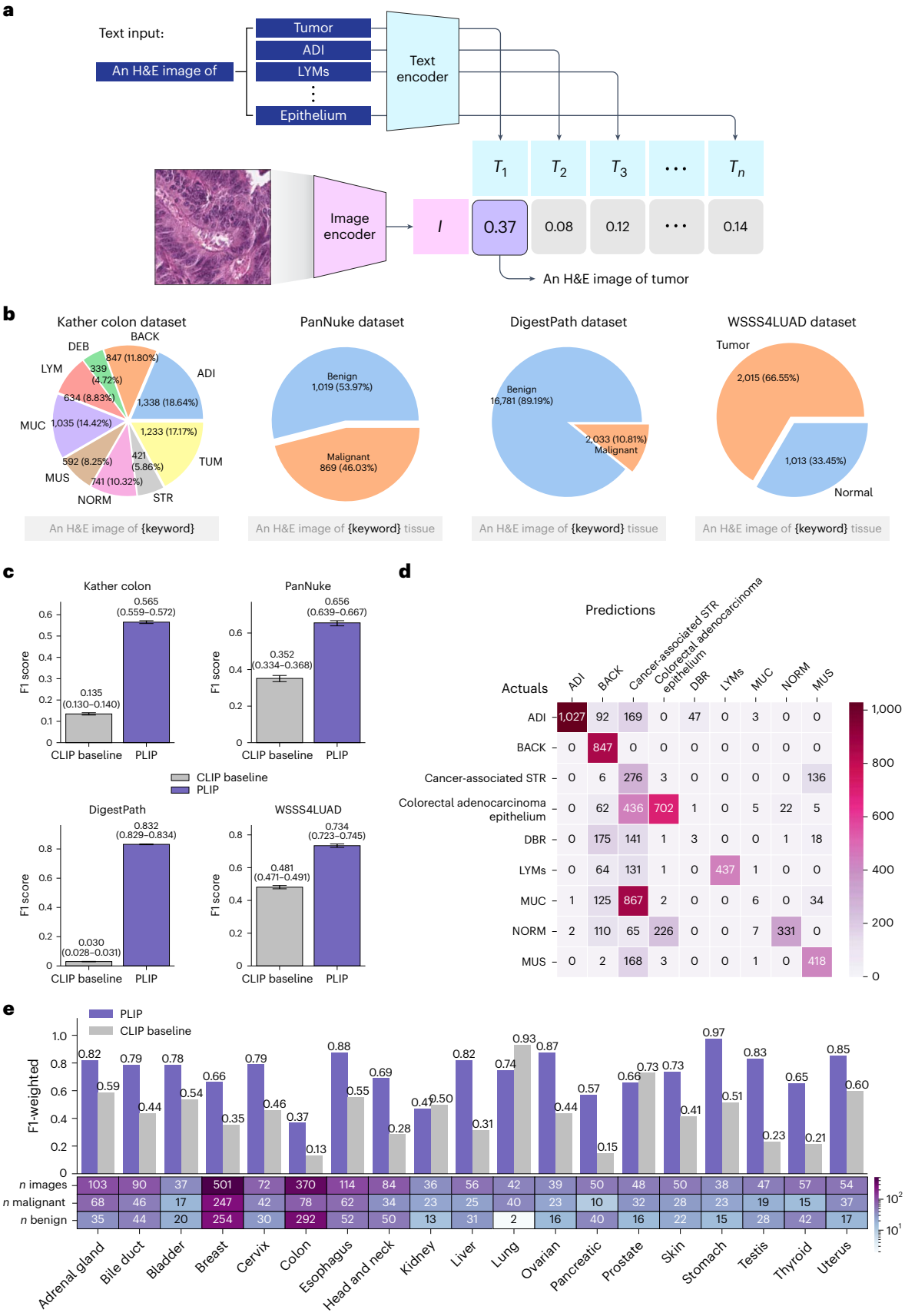
PLIP enhances pathology image retrieval from image inputs

We performed image-to-image retrieval[4] by calculating the similarity between the image embedding of the target image and the image embeddings of the candidate images (Fig. 5a). Evaluations were initially carried out from the Twitter validation dataset (Fig. 5b). Out of 2,023 image–text pairs from 925 tweets in the Twitter validation dataset, 525 tweets had at least two images, contributing to a total of 1,623 images to

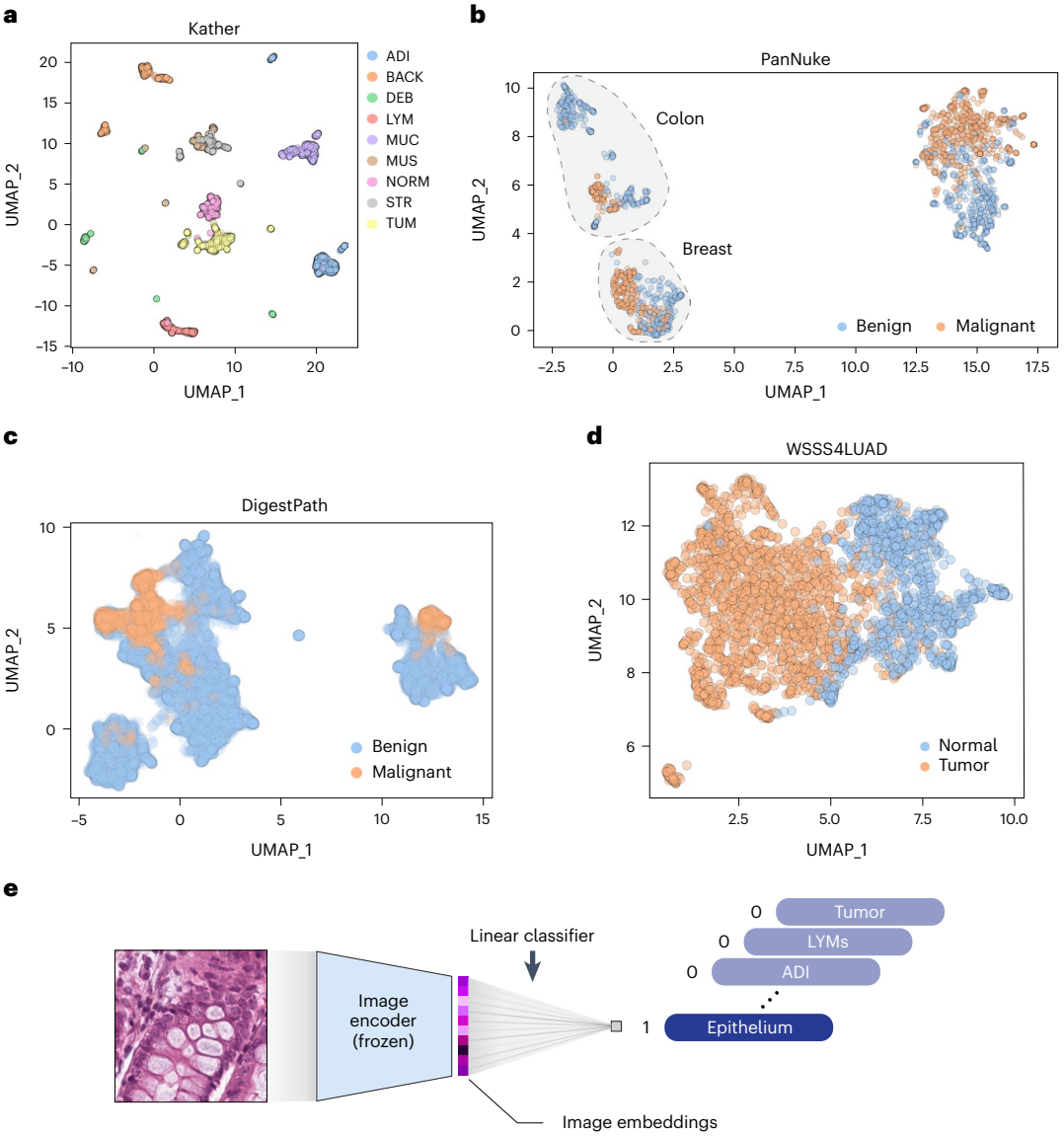
Fig. 2 | PLIP predicts new classes via zero-shot transfer learning. a, Graphical illustration of zero-shot classification. The classification output is determined by selecting the candidate text with the highest cosine similarity to the input image. **b**, Four external validation datasets: Kather colon dataset with nine tissue types; PanNuke dataset (benign and malignant tissues); DigestPath dataset (benign and malignant tissues); and WSSS4LUAD dataset (tumor and normal tissues). **c**, Zero-shot performances with weighted F1 scores across the four datasets. Note that the performances in the Kather colon dataset are based on a nine-class zero-shot learning evaluation, while the performances for other datasets are based on binary zero-shot learning evaluation. Within each box plot, the center line represents the mean and the error bar indicates the 95% CI. Number of test samples for each dataset: Kather colon ($n = 7{,}180$); PanNuke ($n = 1{,}888$); DigestPath ($n = 18{,}814$); and WSSS4LUAD ($n = 3{,}028$). **d**, Confusion matrix of the Kather colon dataset. The actual and predicted labels are displayed in rows and columns, respectively. **e**, Zero-shot evaluation of the PanNuke dataset within each organ type.

https://doi.org/10.1038/s41591-023-02504-3

be searched. We compared each target image with all other images via the Recall@10 and Recall@50, which measures the number of images in the top 10 and top 50 retrieved that originate from the same Twitter post. The benchmark comparison was conducted by comparing PLIP with three baseline models: CLIP; MuDiPath; and SISH[5]. The results presented in Fig. 5c suggested that all four models were capable of retrieving relevant images, while the PLIP model achieved the best performance with Recall@10 = 0.646 (compared to CLIP at 0.353,



https://doi.org/10.1038/s41591-023-02504-3



f

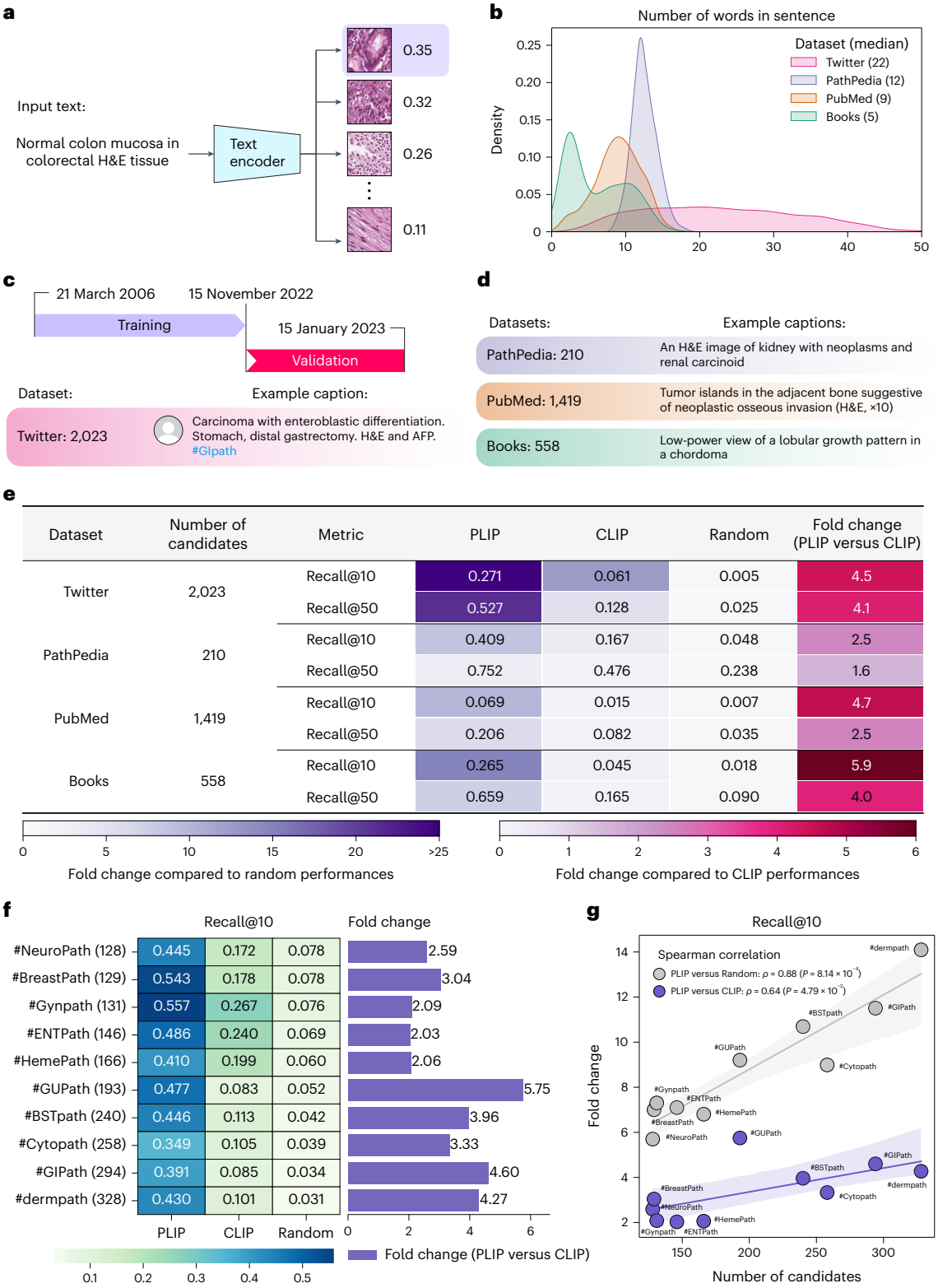
	Kather colon	PanNuke	DigestPath	WSSS4LUAD	Average
CLIP	0.797 (±0.006)	0.852 (±0.002)	0.753 (±0.009)	0.850 (±0.022)	0.813 (±0.043)
MuDiPath	0.825 (±0.001)	0.896 (±0.001)	0.827 (±0.007)	0.917 (±0.003)	0.866 (±0.041)
PLIP	0.877 (±0.001)	0.902 (±0.010)	0.856 (±0.008)	0.927 (±0.007)	0.891 (±0.028)
PLIP versus CLIP	2.9×10^{-9}	9.4×10^{-6}	1.5×10^{-7}	1.5×10^{-4}	—
PLIP versus MuDiPath	9.4×10^{-12}	0.249	6.2×10^{-4}	3.0×10^{-2}	—

Fig. 3 | Image embedding analysis and linear probing results. a, Image embeddings generated from the PLIP model in the Kather colon dataset. **b**, Image embeddings generated from the PLIP model in the PanNuke dataset. **c**, Image embeddings generated from the PLIP model in the DigestPath dataset. **d**, Image embeddings generated from the PLIP model in the WSSS4LUAD dataset. **e**, Graphical illustration of linear probing transfer learning. 'Frozen' means that the loss from the linear classifier will not be used to update the parameters of the image encoder. **f**, F1 score in testing sets with the mean (± s.d.) from five repeated experiments with different random seeds. The 'Average' column shows the averaged performances across the four datasets. *P* values were calculated using a two-sided Student's *t*-test and are presented in the bottom two rows.

MuDiPath at 0.336 and SISH at 0.356). Similarly, PLIP achieved the best performance with Recall@50 = 0.814 (compared to CLIP at 0.513, MuDiPath at 0.485 and SISH at 0.474).

Additional evaluations were conducted on three external validation datasets, each with a distinct study focus (Supplementary Fig. 2): (1) tissue types: the Kather colon dataset (nine colon tissue types); (2) organ types: the PanNuke dataset (19 organs); and (3) staining textures: the KIMIA Path24C dataset (24 staining textures)[35]. By evaluating the class retrieval accuracy at the top *K*, we determined the purity of the retrieved *K* images from the same given class. From the results, we

https://doi.org/10.1038/s41591-023-02504-3



Dataset	Number of candidates	Metric	PLIP	CLIP	Random	Fold change (PLIP versus CLIP)
Twitter	2,023	Recall@10	0.271	0.061	0.005	4.5
		Recall@50	0.527	0.128	0.025	4.1
PathPedia	210	Recall@10	0.409	0.167	0.048	2.5
		Recall@50	0.752	0.476	0.238	1.6
PubMed	1,419	Recall@10	0.069	0.015	0.007	4.7
		Recall@50	0.206	0.082	0.035	2.5
Books	558	Recall@10	0.265	0.045	0.018	5.9
		Recall@50	0.659	0.165	0.090	4.0

Fig. 4 | Text-to-image retrieval for pathology images. a, Graphical illustration of pathology image retrieval from text input. **b**, Density plot of the number of words per sentence across the four validation datasets. **c**, Description of the Twitter validation dataset and an example text caption. **d**, Descriptions of the PathPedia, PubMed and Books datasets and example text captions. **e**, Image retrieval performances across the validation datasets. **f**, Text-to-image retrieval performances for Recall@10 within each of the pathology subspecialty-specific hashtags. **g**, Spearman correlations between the number of candidates and fold changes for Recall@10 when comparing the PLIP model with CLIP and random, respectively. Regression estimates are displayed with the 95% CIs in gray or purple.

https://doi.org/10.1038/s41591-023-02504-3

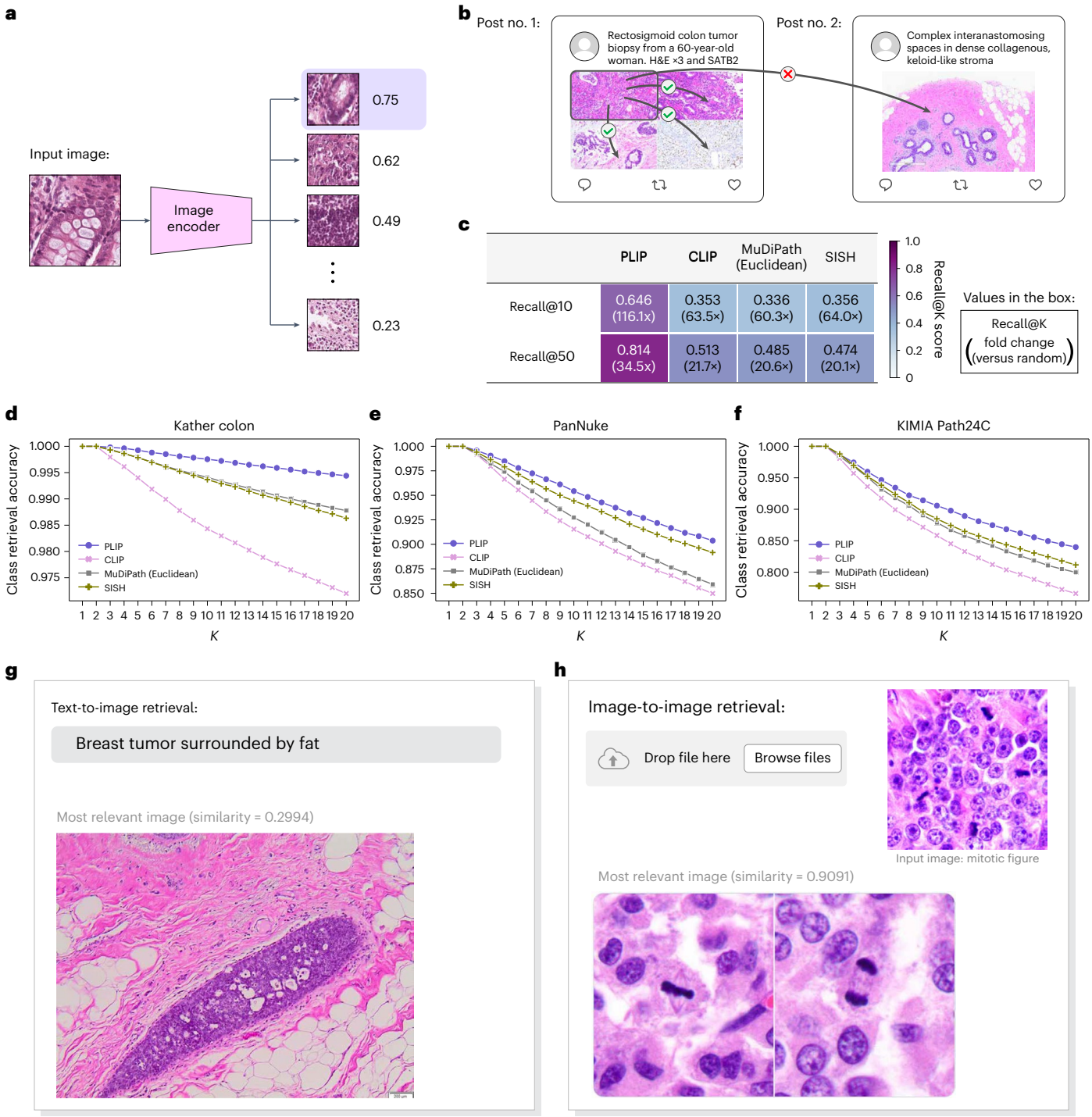


Fig. 5 | Image-to-image retrieval for pathology images. a, Graphical illustration of image-to-image retrieval. **b**, Illustration of image-to-image retrieval analysis on the Twitter validation dataset. **c**, Image-to-image retrieval performances on the Twitter validation dataset. The values in the boxes represent the Recall@10 and Recall@50 scores and the fold changes compared to random performances. **d**, Image-to-image retrieval performances on the Kather colon dataset. **e**, Image-to-image retrieval performances on the PanNuke dataset. **f**, Image-to-image retrieval performances on the KIMIA Path24C dataset. **g**, Examples of text-to-image retrieval. **h**, Examples of image-to-image retrieval (featuring the mitotic figure).

found that PLIP consistently outperformed other models across all three datasets (Fig. 5d–f and Supplementary Table 6). For example, in the Kather colon dataset, we found that PLIP achieved 0.998 when $K = 10$ (meanwhile, CLIP = 0.984, MuDiPath = 0.994, SISH = 0.993). In the PanNuke dataset, we found that PLIP achieved 0.954 when $K = 10$ (meanwhile, CLIP = 0.915, MuDiPath = 0.927, SISH = 0.944). In the KIMIA Path24C dataset, we found that PLIP achieved 0.906 when $K = 10$ (meanwhile, CLIP = 0.858, MuDiPath = 0.879, SISH = 0.885). These results under several testing scenarios, including tissue types, organ types and staining textures, suggested that PLIP is a preferred model to be used as an image-to-image retrieval system in pathology.

Finally, the text-to-image and image-to-image retrieval systems can function as image search engines, enabling users to match images from multiple queries and retrieve the most relevant image based

on a sentence description or an input image. As demonstrated and presented on our website (https://tinyurl.com/webplip), this generic system can understand semantic and interrelated knowledge, such as 'breast tumor surrounded by fat' (Fig. 5g). This capability provides a powerful tool for exploring and retrieving large pathology datasets, allowing users to efficiently and accurately identify relevant images that meet specific criteria. Additionally, image-to-image retrieval can be used to retrieve relevant pathology images similar to the target image input, for example, images that contain mitotic figures, demonstrating its ability to understand the key concepts from the input image (Fig. 5h).

Discussion

The rapid advance of machine learning for computer vision and natural language processing has relied on annotated data. Unlike other fields, annotating pathology images is extremely expensive and laborious, and requires a high level of domain expertise and years of professional education[36,37]. This obstructs the advancement of AI to understand and interpret histopathological features, decipher anatomy and disease heterogeneity, and identify diverse disease subtypes to enable precision medicine[38].

The explosion of data shared on social media presents a valuable and underused opportunity for medical AI. Twitter, in particular, has become an active community for pathologists[21]. By curating this publicly shared knowledge, the OpenPath dataset of image–text pairs can help AI to understand both global and local pathological features. In this study, we leveraged OpenPath to develop PLIP by fine-tuning the state-of-the-art model in visual–language representation and learning[25].

Unlike canonical machine learning approaches in digital pathology that learn from a fixed set of labels, our PLIP model is a general solution that can be applied to a wide range of tasks, including adaptation to new data and provide zero-shot predictions given any image inputs. This capability of handling a varying number of classes is particularly valuable in scenarios where the learning objectives change after the model has been trained. Furthermore, this zero-shot ability also aligns with the constantly evolving criteria for diagnosis in pathology[10]. The improved image representation ability was then quantitatively demonstrated by linear probing and fine-tuning analysis. Comparing the fine-tuning results from the PLIP image encoder to task-specific deep learning models, PLIP exhibited improved performance across four validation datasets. This was particularly notable when a smaller amount of training data was used for training, highlighting the representation learning benefits of PLIP.

This study has several limitations. First, Twitter data can be noisy. Our stringent filtering pipeline improved data quality, as confirmed by our human evaluation. Moreover, contrastive learning is also tolerant to some level of data noise. Second, although several image preprocessing and transformation algorithms were applied to the image encoder upfront, identifying and accounting for the various magnifications of pathology image patches and different staining styles still remained a challenging problem. By training over diverse data, PLIP demonstrates potential to recognize and adapt to images at various magnifications and staining protocols. Third, it is worth noting that zero-shot classification using prompts may be unstable because variations in the prompt could alter the results[39]. We expect that continued endeavors to optimize prompts may lead to further improvements in zero-shot performance. Moreover, we demonstrated PLIP's capability via zero-shot learning, and further expanded our evaluation via linear probing, fine-tuning, text-to-image retrieval and image-to-image retrieval. In the future, these results can serve as a starting point for more advanced diagnostic tasks, such as disease subtyping or grading. While the text-to-image retrieval results presented promising performances and large improvements over the baseline models, the retrieval datasets used were limited in size owing to the limited literature available on this topic; current results do not suggest that search can be replaced entirely by PLIP. As a challenging yet meaningful technique, the image retrieval task requires larger datasets of images and captions to better understand the effectiveness of the model. Fourth, while PLIP demonstrated its strong performance across multiple tasks and datasets, it is important to acknowledge that specialized models optimized for individual tasks can outperform general purpose models like PLIP in those specific domains. For example, a VGG19 model trained specifically to predict nine tissue types on the Kather colon dataset achieves an accuracy of 0.943 (ref. 26), which is higher than PLIP's linear probing accuracy of 0.913. Lastly, the training method was established with the standard CLIP model. While the CLIP model can be trained efficiently, future improvements on the algorithmic side are necessary to further enhance the model capabilities. Limited by computational ability, all input images were resized to 224 × 224 pixels according to the original CLIP, which might lose some visual and subvisual patterns in pathology images.

Advances in the PLIP model across diverse learning tasks were made possible due to the curation of the largest publicly available dataset, OpenPath, containing paired pathology images and text descriptions. We anticipate both the open-source PLIP and OpenPath to benefit the medical AI community by enabling further advances in pathology AI, building on this foundation model and facilitating medical knowledge sharing through the PLIP search engine.

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/s41591-023-02504-3.

References

1. 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).
2. Campanella, G. et al. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nat. Med.* **25**, 1301–1309 (2019).
3. Dawood, M., Branson, K., Rajpoot, N. M. & Ul Amir Afsar Minhas, F. ALBRT: cellular composition prediction in routine histology images. In *Proc. IEEE/CVF International Conference on Computer Vision Workshops* 664–673 (IEEE, 2021).
4. Hegde, N. et al. Similar image search for histopathology: SMILY. *NPJ Digit. Med.* **2**, 56 (2019).
5. Chen, C. et al. Fast and scalable search of whole-slide images via self-supervised deep learning. *Nat. Biomed. Eng.* **6**, 1420–1434 (2022).
6. Gamper, J., Alemi Koohbanani, N., Benet, K., Khuram, A. & Rajpoot, N. PanNuke: An Open Pan-Cancer Histology Dataset for Nuclei Instance Segmentation and Classification. In *Digital Pathology* (eds Reyes-Aldasoro, C. et al.) 11–19 (Springer International Publishing, 2019).
7. Graham, S. et al. Lizard: a large-scale dataset for colonic nuclear instance segmentation and classification. In *Proc. IEEE/CVF International Conference on Computer Vision Workshops* 684–693 (IEEE, 2021).
8. Amgad, M. et al. Structured crowdsourcing enables convolutional segmentation of histology images. *Bioinformatics* **35**, 3461–3467 (2019).
9. Singh, H. & Graber, M. L. Improving diagnosis in health care—the next imperative for patient safety. *N. Engl. J. Med.* **373**, 2493–2495 (2015).

10. Erickson, L. A., Mete, O., Juhlin, C. C., Perren, A. & Gill, A. J. Overview of the 2022 WHO classification of parathyroid tumors. *Endocr. Pathol.* **33**, 64–89 (2022).
11. van Rijthoven, M. et al. Few-shot weakly supervised detection and retrieval in histopathology whole-slide images. *Medical Imaging 2021: Digital Pathology* (eds Tomaszewski, J. E. & Ward, A. D.) 137–143 (Society of Photographic Instrumentation Engineers, 2021).
12. Chen, J., Jiao, J., He, S., Han, G. & Qin, J. Few-shot breast cancer metastases classification via unsupervised cell ranking. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **18**, 1914–1923 (2021).
13. 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).
14. Schukow, C. P., Booth, A. L., Mirza, K. M. & Jajosky, R. P. #PathTwitter: a positive platform where medical students can engage the pathology community. *Arch. Pathol. Lab. Med.* **147**, 135–136 (2023).
15. Crane, G. M. & Gardner, J. M. Pathology image-sharing on social media: recommendations for protecting privacy while motivating education. *AMA J. Ethics* **18**, 817–825 (2016).
16. El Hussein, S. et al. Next-generation scholarship: rebranding hematopathology using twitter: the MD Anderson experience. *Mod. Pathol.* **34**, 854–861 (2021).
17. Mukhopadhyay, S. et al. The network that never sleeps. *Lab. Med.* **52**, e83–e103 (2021).
18. Allen, T. C. Social media: pathologists' force multiplier. *Arch. Pathol. Lab. Med.* **138**, 1000–1001 (2014).
19. Misialek, M. J. & Allen, T. C. You're on social media! So now what? *Arch. Pathol. Lab. Med.* **140**, 393 (2016).
20. Katz, M. S. et al. Disease-specific hashtags for online communication about cancer care. *JAMA Oncol.* **2**, 392–394 (2016).
21. Oltulu, P., Mannan, A. A. S. R. & Gardner, J. M. Effective use of Twitter and Facebook in pathology practice. *Hum. Pathol.* **73**, 128–143 (2018).
22. Schuhmann, C. et al. LAION-5B: An open large-scale dataset for training next generation image-text models. In *Proc. 35th International Conference on Neural Information Processing Systems* (eds Koyejo, S. et al.) 25278–25294 (2022).
23. Palatucci, M., Pomerleau, D., Hinton, G. & Mitchell, T. M. Zero-shot learning with semantic output codes. In *Proc. 22nd International Conference on Neural Information Processing Systems* (eds Bengio, Y. et al.) 1410–1418 (Curran Associates, 2009).
24. Pathology Tag Ontology. *Symplur* https://www.symplur.com/healthcare-hashtags/ontology/pathology/ (2023).
25. Radford, A. et al. Learning transferable visual models from natural language supervision. In *Proc. 38th International Conference on Machine Learning* (eds Meila, M. & Zhang, T.) 8748–8763 (PMLR, 2021).
26. Kather, J. N. et al. Predicting survival from colorectal cancer histology slides using deep learning: a retrospective multicenter study. *PLoS Med.* **16**, e1002730 (2019).
27. 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).
28. 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).
29. 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).
30. Eslami, S., Meinel, C. & de Melo, G. PubMedCLIP: How much does CLIP benefit visual question answering in the medical domain? In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics* (eds Vlachos, A. & Augenstein, I.) 1181–1193 (EACL, 2023).
31. Wang, Z., Wu, Z., Agarwal, D. & Sun, J. MedCLIP: contrastive learning from unpaired medical images and text. Preprint at https://doi.org/10.48550/arXiv.2210.10163 (2022).
32. Mormont, R., Geurts, P. & Maree, R. Multi-task pre-training of deep neural networks for digital pathology. *IEEE J. Biomed. Health Inform.* **25**, 412–421 (2021).
33. Kherfi, M. L., Ziou, D. & Bernardi, A. Image retrieval from the world wide web: issues, techniques, and systems. *ACM Comput. Surv.* **36**, 35–67 (2004).
34. Gamper, J. & Rajpoot, N. Multiple instance captioning: learning representations from histopathology textbooks and articles. In *Proc. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 16549–16559 (IEEE, 2021).
35. Shafiei, S., Babaie, M., Kalra, S. & Tizhoosh, H. R. Colored Kimia Path24 dataset: configurations and benchmarks with deep embeddings. Preprint at https://doi.org/10.48550/arXiv.2102.07611 (2021).
36. Madabhushi, A. & Lee, G. Image analysis and machine learning in digital pathology: challenges and opportunities. *Med. Image Anal.* **33**, 170–175 (2016).
37. Srinidhi, C. L., Kim, S. W., Chen, F.-D. & Martel, A. L. Self-supervised driven consistency training for annotation efficient histopathology image analysis. *Med. Image Anal.* **75**, 102256 (2022).
38. Tizhoosh, H. R. & Pantanowitz, L. Artificial intelligence and digital pathology: challenges and opportunities. *J. Pathol. Inform.* **9**, 38 (2018).
39. Zhou, C., He, J., Ma, X., Berg-Kirkpatrick, T. & Neubig, G. Prompt consistency for zero-shot task generalization. In *Findings of the Association for Computational Linguistics* (eds Goldberg, Y. et al.) 2613–2626 (EMNLP, 2022).

Methods

Description of the OpenPath dataset

Release policy. In accordance with the policy and regulation of Twitter and other entities including LAION, all information provided in the datasets is linked to the original source of the data. Specifically, data that have been collected from Twitter are released in the form of Tweet IDs; data that have been collected from LAION are released in the form of URLs to images. Interested users will need to refer to the original sources to understand if their usage is compliant with the policies and regulations.

Twitter collection. All Twitter posts (tweets) with English captions under 32 pathology subspecialty-specific hashtags were included according to the recommendation from the 2016 USCAP meeting and the Pathology Hashtag Ontology projects[24]. The complete list of hashtags and descriptions is shown in Extended Data Table 1. Tweets were collected from 21 March 2006 (the date of the first Twitter post) to 15 November 2022. Conversations including replies were collected for each of the tweets from 21 March 2006 to 22 November 2022 (7 days after the last tweet was collected on 15 November 2022). In total, we collected 232,067 tweets and 243,375 image–text pairs. Among those tweets, we further collected 88,250 replies for which (1) the associated tweets had replies, (2) sentences contained at least one keyword from the International Classification of Diseases, 11th Revision codebook (February 2022 version) and (3) received the highest number of likes among all replies.

Several cohort exclusion criteria were applied to our raw dataset (Extended Data Fig. 1a), which included: (1) tweets that were marked as possibly sensitive by Twitter; (2) duplicate image–text pairs; (3) images that cannot be downloaded or opened; (4) non-pathology images by the pathology image classification model; and (5) texts that had question marks because questions often do not contain information about the image. After our stringent data inclusion and exclusion criteria, we ended up with 116,504 unique image–text pairs from tweets and 59,869 unique image–text pairs from the top replies. As part of the image retrieval tasks, the Twitter validation dataset was further collected from 16 November 2022 to 15 January 2023, with the exact same cohort inclusion and exclusion criteria (Extended Data Fig. 1b), and resulted in 2,023 unique image–text pairs.

PathLAION collection. We established the PathLAION collection, which contains the pathology image–text data from sources beyond Twitter on the Internet. PathLAION is a subset of pathology images from the LAION-5B dataset, which originally contained 5.85 billion image–text pairs from the Internet[22]. This PathLAION collection was obtained by feeding a pathology image from the Twitter dataset to LAION-5B to retrieve the 5,000 most similar images, measured by the cosine similarity of CLIP image embeddings. The subsampling process stopped once all images retrieved were duplicated. By repeating this step with 1,000 images sampled from the Twitter dataset, we obtained 32,041 unique pathology images that were not present in the Twitter collection.

Quality control of the training dataset. An additional quality control pipeline was applied to the established OpenPath dataset. A pathology image classifier was developed to exclude non-pathology images from both the Twitter and LAION datasets. Because not all downloaded images were microscopic pathology images, this step was necessary to ensure the quality and relevance of the dataset.

To ensure high-quality images, an additional manual inspection was conducted by randomly sampling 1,000 images from the OpenPath dataset (Supplementary Table 2). This evaluation showed that 98.5% of the subsampled images were pathology images and 94.5% of the sampled images were high-quality regions of interest from whole-slide images, indicating that OpenPath contains data of reasonably good quality.

To ensure high-quality text descriptions of the associated images, the texts of the tweets and replies were cleaned using the following pipeline: (1) remove '@username' in the sentence; (2) remove hashtags '#'; (3) remove HTML tags for italic and bold formatting (for example, replace '*keyword*' or '**keyword**' with 'keyword'); (4) remove emojis; (5) remove the symbols of newline (\n) and carriage return (\r); (6) remove extra white spaces; and (7) remove links starting with 'http://' and 'https://'. Furthermore, tweets containing question marks were removed because they are often used by practitioners to inquire about a pathology image rather than provide informative descriptions. For PathLAION, all image–text pairs with non-English captions were removed using langdetect (https://pypi.org/project/langdetect/). The complete statistics of text lengths in the final training dataset are shown in Supplementary Table 1.

External validation datasets

To evaluate the performance of the proposed machine learning algorithm, a set of publicly available datasets were gathered.

For zero-shot and linear probing analysis, four datasets were collected: (1) the Kather colon dataset[26], which consisted of 100,000 training image patches and 7,180 validation image patches; (2) the PanNuke dataset[6], which contained 19 organ types with five different types of nuclei (neoplastic, inflammatory, connective tissue, epithelial and dead) and had a total of 7,558 image patches that contained at least one cell. Among them, images were deemed to be 'malignant' if both the total number of neoplastic cells was ten or more and occupied over 30% of the total cells. Alternatively, images were deemed to be 'benign' if no neoplastic cells existed. This resulted in 2,866 malignant images and 3,368 benign images; (3) DigestPath[27], initially from low-magnification whole-slide images, which were then cropped into multiple patches with several downsampling rates (2×, 4×, 8×, 16× and 32×) with pixel sizes = 224 × 224 and 10% patch overlapping. Background patches were excluded if more than 50% of the pixels contained no tissue (RGB-based color threshold = 200). With pixel-level annotation of malignancy, images were deemed to be malignant if the region of malignancy occupied more than 50% of the tissue area within an image patch. Thus, 6,690 malignant images and 56,023 benign images were obtained; (4) for WSSS4LUAD[28], all images were binarized to either tumor or normal, based on the existence of tumors on an image. This resulted in 6,579 tumor images and 3,512 normal images. All image patches were then resized to 224 × 224 pixels before feeding into the image encoders. Except for the Kather colon dataset which had preexisting training and validation splits, all other datasets (PanNuke, DigestPath and WSSS4LUAD) were randomly divided into training and validation sets with a ratio of 70% to 30%. To prevent data leakage on DigestPath, training and validation splits were separated according to their unique sample ID. To ensure that the benchmarks were comparable, both zero-shot and linear probing analyses were evaluated on the validation split.

For the text-to-image retrieval analysis, we collected three datasets in addition to the Twitter validation dataset. These datasets included PathPedia (http://www.pathpedia.com/) and the pathology collections of PubMed[34] and Books[34]. For PathPedia, we subset the original set of data by dropping duplicated captions, which resulted in 210 unique image–text pairs. In PathPedia, captions were further curated with 'an image of category with subclass', where category refers to the organ type (for example, kidney) and the subclass refers to a specific disease subtype (for example, neoplasm). For PubMed and Books, duplicated captions were removed. Captions containing more than 100 characters or fewer than three characters were considered outliers and excluded from our subsequent analysis. In addition to this, looking at the captions we realized that often most of the information is contained in the first sentence. Thus, we removed all the text after the first period and finally checked again if the caption was between three and 100 characters.

For the image-to-image retrieval analysis, we used the Kather colon dataset with nine different tissue types and the PanNuke dataset with 19 different organ types. In addition, we included the KIMIA Path24C dataset[35] with 24 different pathology image staining textures to evaluate the model's ability to retrieve images with the same textural pattern.

Model training and tuning

All experiments were run in Python v.3.9. Detailed software versions are: pytorch v.1.13; CUDA v.11.7; CUDNN v.8.5.0; scipy v.1.9.3; torchvision v.0.14.0; pillow v.9.1.0; scikit-learn v.1.1.2; scikit-image v.0.19.2; pandas v.1.4.2; numpy v.1.23.5; multiprocess v.0.70.13; langdetect v.1.0.9; and Twitter API v.2.0 with Python v.3.9.

Model training. We established our PLIP model with the same architecture described in Radford et al.[25]. This architecture is based on a vision transformer as the image encoder (ViT-B/32, which can take in input images of size = 224 × 224 pixels), and a text transformer as the text encoder (with a maximum sequence length = 76 tokens)[40]. Images were initially resized to a maximum dimension of 512 pixels; then, they were randomly cropped to 224 × 224 pixels before feeding into the image encoder. Both the image and text encoders output vectors of 512 dimensions and were optimized by minimizing the contrastive loss on a given batch[41]. Contrastive learning imposes a higher cosine similarity in paired image and text, forcing the model to learn the correct relationships between images and texts.

To find the optimal set of hyperparameters, different combinations of training datasets and learning rates were searched on the linear probing task. With the model being evaluated on every quarter of an epoch (1 step = 1/4 epochs) given a total of 12 epochs, we found that the best-performing model was trained with an optimal learning rate = 1×10^{-5}, steps = 10 and using all training datasets (tweets + replies + PathLAION). An ablation study was conducted with different combinations of datasets and the same hyperparameters; the results shown in Supplementary Table 5 indicate that the combination of all data (tweets + replies + PathLAION) allowed us to get the best-performing PLIP model.

Zero-shot classification. Benefiting from its understanding of the text, the proposed PLIP model can classify and recognize new labels for unseen data. This capability, commonly referred to as zero-shot learning[23], enables learning new classes at scale without the need for retraining. In zero-shot learning, the classification result is based on identifying the candidate texts with the highest representational similarity to the input image.

Because images and texts are encoded into the same vector space, the proposed model has the ability to learn the similarity between a target image with text candidates; text prediction is the one with the highest similarity. This ability to transfer a new unseen dataset of images and texts requires zero-training data. In this study, we adapted four datasets (Kather colon, PanNuke, DigestPath and WSSS4LUAD) with their validation splits to evaluate the ability of zero-shot classification performances. The text candidates were generated according to their tissue type. In the Kather colon dataset, we generated the sentences with 'an H&E image of {keyword}', where the keyword can be (1) adipose tissue, (2) background, (3) debris, (4) lymphocytes, (5) mucus, (6) smooth muscle, (7) normal colon mucosa, (8) cancer-associated stroma and (9) colorectal adenocarcinoma epithelium. In the PanNuke and DigestPath datasets, we generated the sentences with 'an H&E image of {keyword} tissue', where the keywords can be (1) benign and (2) malignant. In the WSSS4LUAD dataset, we generated the sentences with 'an H&E image of {keyword} tissue', where the keywords can be (1) normal and (2) tumor. To determine the CI, zero-shot predictions were bootstrapped 100 times, each with 70% of the data used in the calculation.

Linear probing. Linear probing is a commonly used technique for assessing the quality of the features extracted by a model[42,43]. In this study, feature embeddings were extracted from the images and then a linear classifier was trained on top of these embeddings for a given classification task. All feature embeddings were L2-normalized before being fed into a linear classifier. To accomplish this, we used a logistic regression classifier from the stochastic gradient descent classifier (SGDClassifier) module in the sklearn Python package[44]. For benchmark comparison, we compared our PLIP model backbone to the baseline CLIP model backbone, as well as the multitask pretraining deep neural network model[32] (MuDiPath) with DenseNet121 architecture[45], which was trained on a collection of 22 classification tasks with approximately 900,000 images.

We evaluated the performance of the linear classifier using L2 regularization with several regularization multipliers (α = 0.1, 0.01, 0.001 and 0.0001) on the validation splits for all models. The best-performing linear classifier was selected based on the average macro F1 performance from the results generated by five different random seeds (seed = 1, 2, 3, 4, 5). The linear probing analysis was also used to determine the most suitable PLIP model for the other experiments.

Comparison to task-specific supervised models. Fine-tuning was conducted to compare the PLIP image encoder with the end-to-end deep learning model ViT-B/32 for the image classification tasks across four external validation datasets. The ViT-B/32 model was pretrained on the ImageNet dataset[46]. For the PLIP image encoder, the last layer of the model was concatenated with a linear classifier. Batch size was set to 128 for both the PLIP image encoder and the ViT-B/32 model. The Adam with decoupled weight decay optimizer[47] was adopted for fine-tuning, with a weight decay of 0.1 and a total of ten training epochs. A hyperparameter search was conducted on the validation split to determine the optimal learning rate. The learning rate was selected from a set of values: $\{1 \times 10^{-6}, 1 \times 10^{-5}, 1 \times 10^{-4}, 1 \times 10^{-3}, 1 \times 10^{-2}\}$. The selection of the learning rate was based on the highest weighted F1 score achieved on the validation split, which was created by taking 30% of the original training split. Once the optimal hyperparameter was determined, the models were trained using all available training data. Their performances were evaluated on the testing split to assess the model's ability on a new pathology image dataset. For the Kather dataset, we used 10% of its original training data as a test split.

Image retrieval. Similar to zero-shot learning, which can identify the closest text from a pool of candidates given an image, image retrieval is a technique that can identify the closest image from a pool of candidates given a text[33] or image[4]. This was accomplished by directly calculating the cosine similarity of each paired image–text or image–image under the same embedding space.

For text-to-image retrieval, we generated a vector representation for a searching query (in natural language) and identified the image that best matched the query. The performance of text-to-image retrieval was evaluated by two metrics: Recall@10 and Recall@50, which refer to the precision of the target image being among the top ten and top 50 retrieved images, respectively[48]. For example, given a set of 500 image–text pairs, each text was fed into the image retrieval task to find the corresponding image. If this process is repeated for all 500 pairs and 60% (or 300 images) of the true positive images can be found under the top ten closest matches from the pool of 500, then Recall@10 becomes 0.60.

The performance of the image-to-image retrieval was evaluated by class retrieval accuracy across models, which is also known as the mean average precision at K (MAP@K)[5]. We considered the retrieved image 'relevant' if the class of the retrieved image was the same as the target input and the precision was the fraction of relevant

images among the top K retrieved images. The average precision at K (AP@K) is defined as:

$$\text{AP@K} = \frac{1}{K}\sum_{i=1}^{K} P_i \times R_i$$

where P_i is precision at i and R_i is the relevant indicator at i ($R_i = 1$ if the i^{th} item is relevant, and $R_i = 0$ if not). The final class retrieval accuracy was calculated by:

$$\text{MAP@K} = \frac{1}{n}\text{AP@K}$$

where n is the total number of samples. It is worth noting that the score may decrease as K increases, as more images may be irrelevant in the top retrieval results. In our image-to-image benchmark comparison, cosine similarity was adopted to compare the similarity between image embeddings for the PLIP and CLIP models, while the MuDiPath and SISH models[5] calculated the similarity between image embeddings by Euclidean and Hamming distances, respectively. According to the guideline for the SISH model, we used a large vector quantized variational autoencoder based on their pretrained weights.

Evaluation metrics and statistical analysis

The F1 score was used to evaluate the performances of zero-shot and linear probing. Ranging from 0 to 1, the F1 score is calculated from the harmonic mean of precision and recall score:

$$\text{F1} = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}} = \frac{2 \times TP}{2 \times TP + FP + FN}$$

where TP stands for the number of true positives, FP stands for the number of false positives, and FN stands for the number of false negatives. While the higher the better, the F1 score represents the overall performance of a model on a classification task. The weighted F1 score was calculated by averaging the F1 score of each class, with the score of each class weighted by the number of occurrences in the class. In addition, the MCC was calculated and compared with the baseline CLIP model for zero-shot analysis. MCC ranges from −1 to 1, where MCC = 1 indicates a perfect prediction, MCC = 0 indicates that the prediction is no better than random chance and MCC = −1 suggests complete disagreement between prediction and ground truth.

To evaluate the performance of the image retrieval task,

$$\text{Precision} = \frac{TP}{TP + FP}$$

was adopted to quantitatively measure the fraction of the target image that exists among the top K retrieved images. In our image retrieval experiments, we evaluated precision with $K = 10$ and $K = 50$.

A two-sided Student's t-test was used to evaluate the significance between model performances. Spearman correlation with a two-sided P value was used to evaluate the correlation between the number of candidates and fold changes for the image retrieval task.

Reporting summary

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

Data availability

All data in OpenPath are publicly available from Twitter and LAION-5B (https://laion.ai/blog/laion-5b/). The Twitter IDs used for training and validation can be accessed at https://tinyurl.com/openpathdata. The validation datasets are publicly available and can be accessed from the following: Kather colon dataset (https://zenodo.org/record/1214456); PanNuke (https://warwick.ac.uk/fac/cross_fac/tia/data/pannuke); DigestPath (https://digestpath2019.grand-challenge.org/); WSSS4LUAD (https://wsss4luad.grand-challenge.org/); PathPedia (https://www.pathpedia.com/Education/eAtlas/Default.aspx); PubMed and Books pathology collection (https://warwick.ac.uk/fac/cross_fac/tia/data/arch); KIMIA Path24C (https://kimialab.uwaterloo.ca/kimia/index.php/pathology-images-kimia-path24/). The ImageNet dataset (https://www.image-net.org/) was adopted for the pretrained ViT-B/32 model. The trained model, source codes and interactive results can also be accessed at https://tinyurl.com/webplip.

Code availability

The trained model and source codes can be accessed at https://tinyurl.com/webplip.

References

40. Radford, A. et al. Language models are unsupervised multitask learners. *OpenAI blog* **1**, 9 (2019).
41. van den Oord, A., Li, Y. & Vinyals, O. Representation learning with contrastive predictive coding. Preprint at https://doi.org/10.48550/arXiv.1807.03748 (2018).
42. Alain, G. & Bengio, Y. Understanding intermediate layers using linear classifier probes. In *5th International Conference on Learning Representations Workshop* (2017).
43. Liang, Y., Zhu, L., Wang, X. & Yang, Y. A simple episodic linear probe improves visual recognition in the wild. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 9559–9569 (IEEE, 2022).
44. Pedregosa, F. et al. Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
45. Huang, G., Liu, Z., Maaten, L. van der & Weinberger, K. Q. Densely connected convolutional networks. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 4700–4708 (IEEE, 2017).
46. Deng, J. et al. ImageNet: a large-scale hierarchical image database. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition* 248–255 (IEEE, 2009).
47. Loshchilov, I. & Hutter, F. Decoupled weight decay regularization. Preprint at https://doi.org/10.48550/arXiv.1711.05101 (2017).
48. Zhang, S., Yang, M., Cour, T., Yu, K. & Metaxas, D. N. Query specific fusion for image retrieval. In *Proc. European Conference on Computer Vision 2012* (eds Fitzgibbon, A. et al.) 660–673 (ECCV, 2012).

Acknowledgements

F.B. is supported by the Hoffman-Yee Research Grant Program and the Stanford Institute for Human-Centered Artificial Intelligence. J.Z. is supported by the Chan Zuckerberg Biohub.

Author contributions

Z.H., F.B. and J.Z. designed the study. Z.H. and F.B. carried out the data collection, data analysis, model construction, model validation and manuscript writing. M.Y. carried out the data analysis, model construction, model validation and manuscript writing. T.J.M. provided knowledge support, interpreted the findings and helped with manuscript writing. J.Z. provided knowledge support, interpreted the findings, helped with manuscript writing and supervised the study. All authors contributed to writing the manuscript and reviewed and approved the final version.

Competing interests
The authors declare no competing interests.

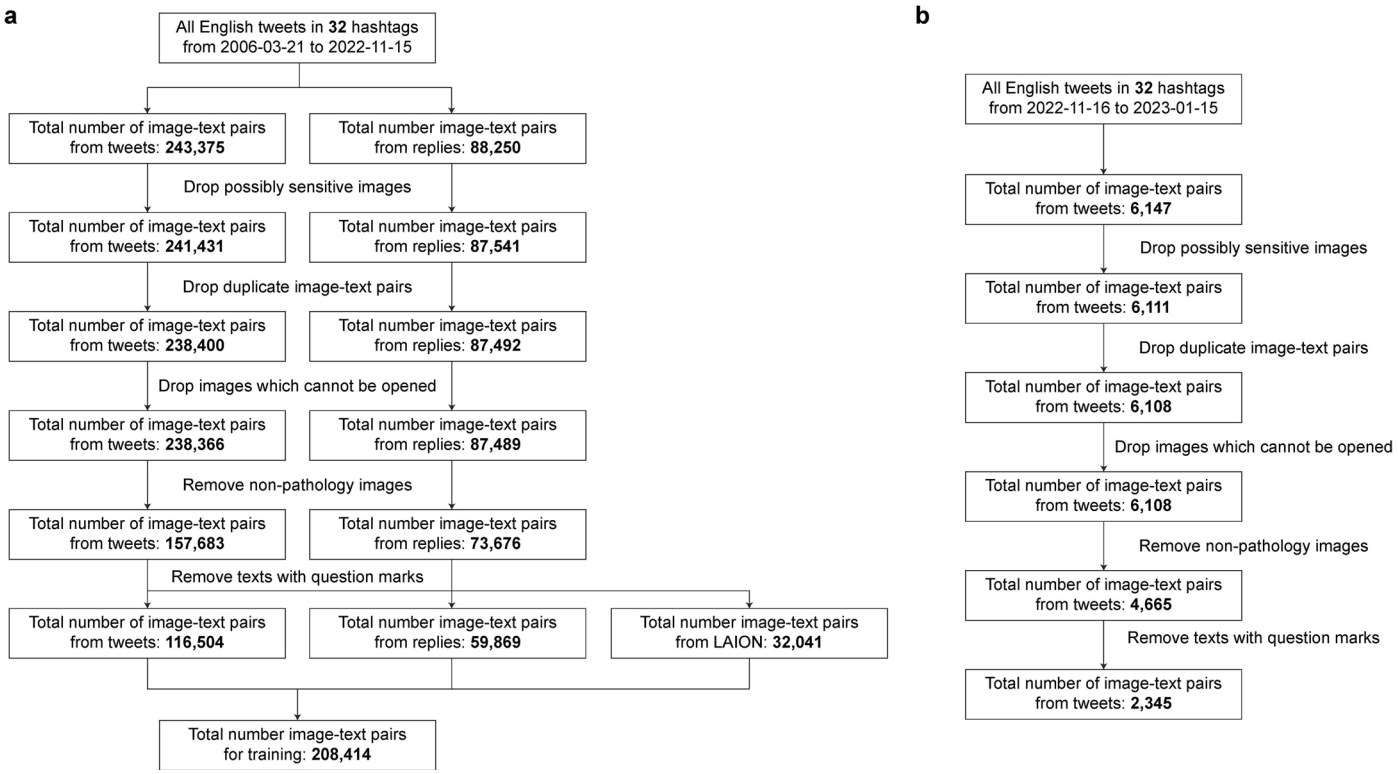
Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41591-023-02504-3.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41591-023-02504-3.

Correspondence and requests for materials should be addressed to James Zou.

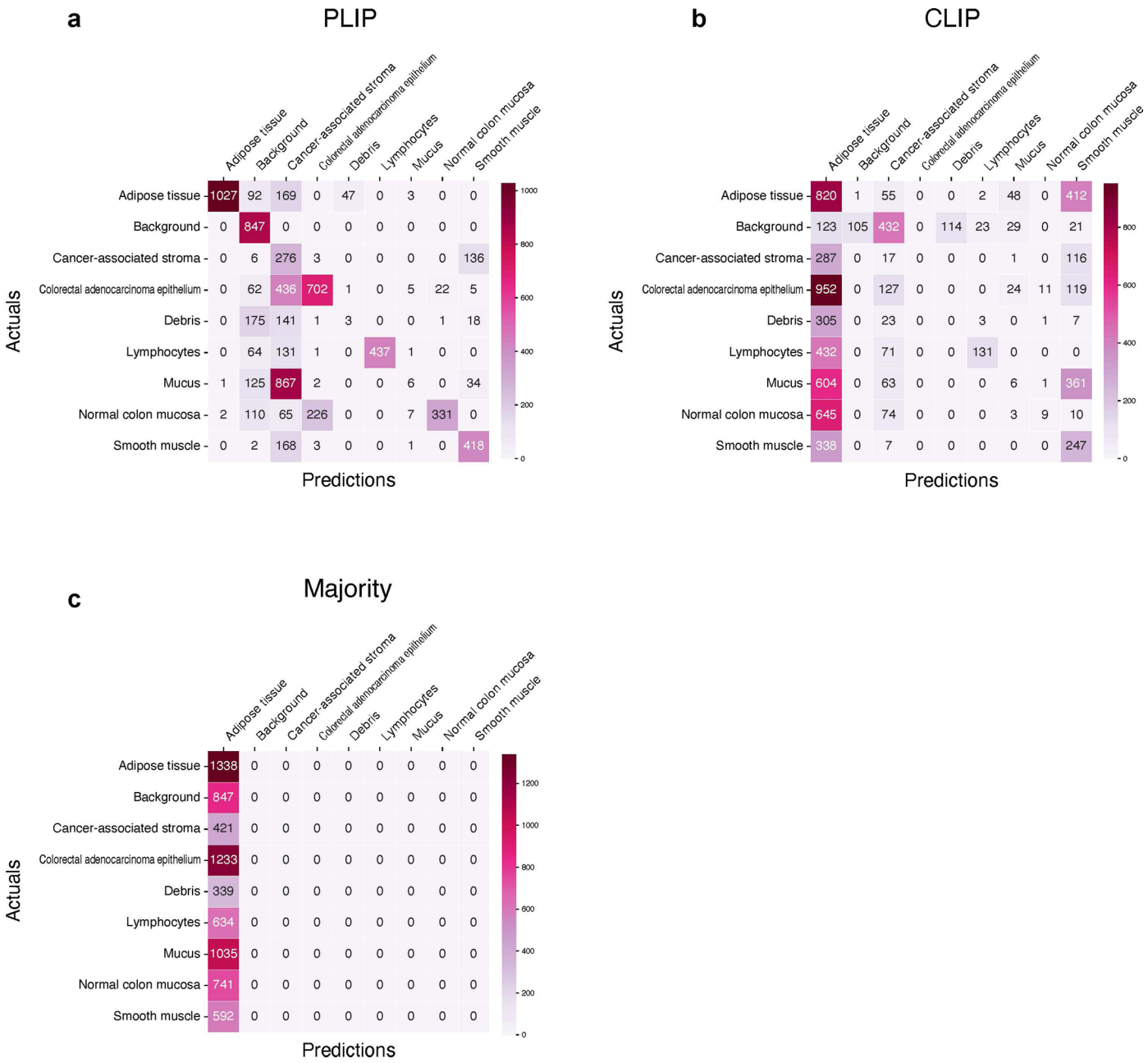
Peer review information *Nature Medicine* thanks Geert Litjens, Lee Cooper and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Primary Handling Editor: Lorenzo Righetto, in collaboration with the *Nature Medicine* team.

https://doi.org/10.1038/s41591-023-02504-3



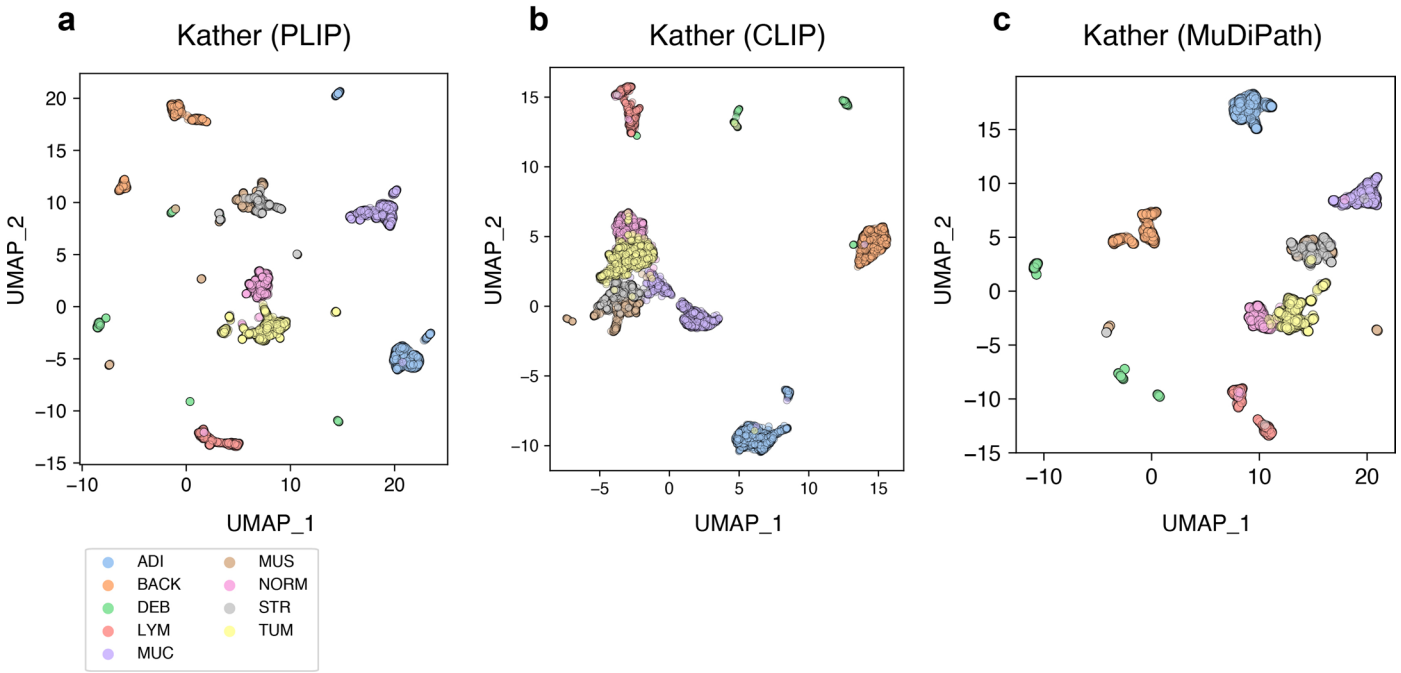
Extended Data Fig. 1 | Cohort inclusion and exclusion criteria and flowcharts. a, Twitter training dataset from 2006-03-21 to 2022-11-15. **b**, Twitter validation dataset for image retrieval from 2022-11-16 to 2023-01-15.

https://doi.org/10.1038/s41591-023-02504-3



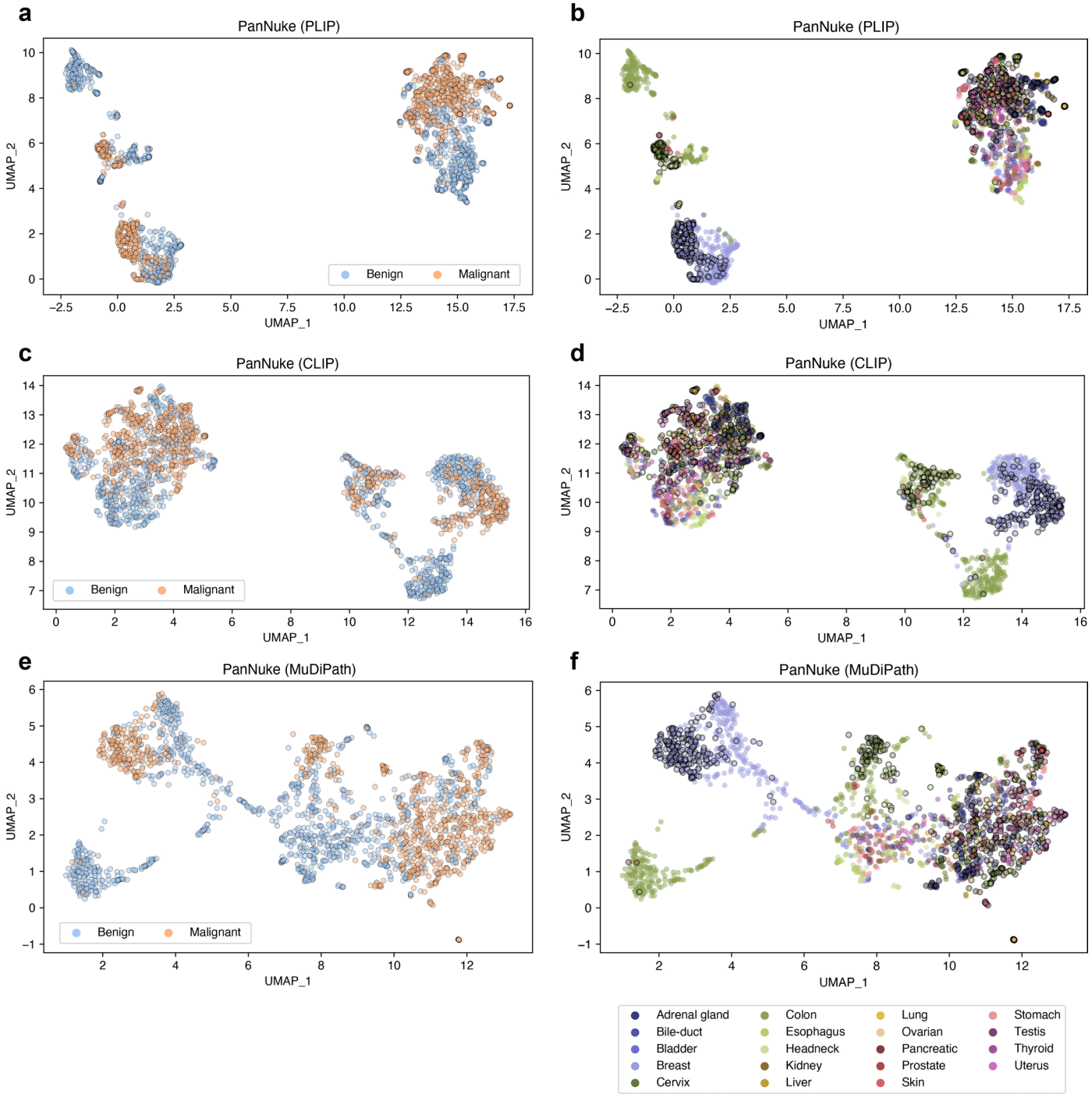
Extended Data Fig. 2 | Confusion matrix from zero-shot learning in the Kather colon dataset. a, Confusion matrix of PLIP model; **b**, Confusion matrix of CLIP model; **c**, Confusion matrix of the results from predicting the majority class (or Majority in short).

https://doi.org/10.1038/s41591-023-02504-3



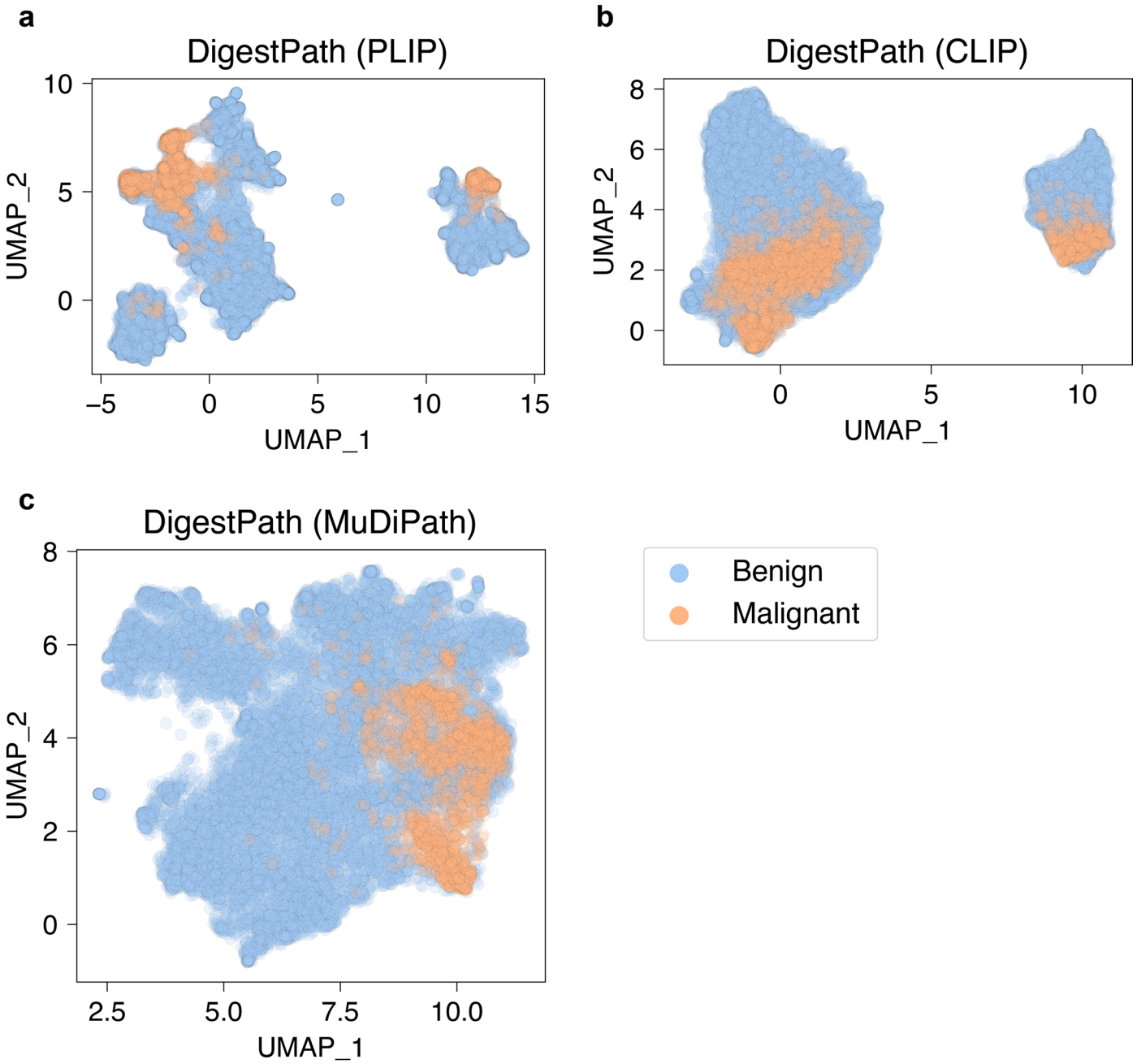
Extended Data Fig. 3 | Comparison of image embeddings derived from models in the Kather colon dataset. a, image embeddings derived from PLIP, **b**, image embeddings derived from baseline CLIP. **c**, image embeddings derived from MuDiPath. ADI: Adipose tissue, BACK: background, DEB: debris, LYM: lymphocytes, MUC: mucus, MUS: smooth muscle, NORM: normal colon mucosa, STR: cancer-associated stroma, TUM: colorectal adenocarcinoma epithelium.

https://doi.org/10.1038/s41591-023-02504-3



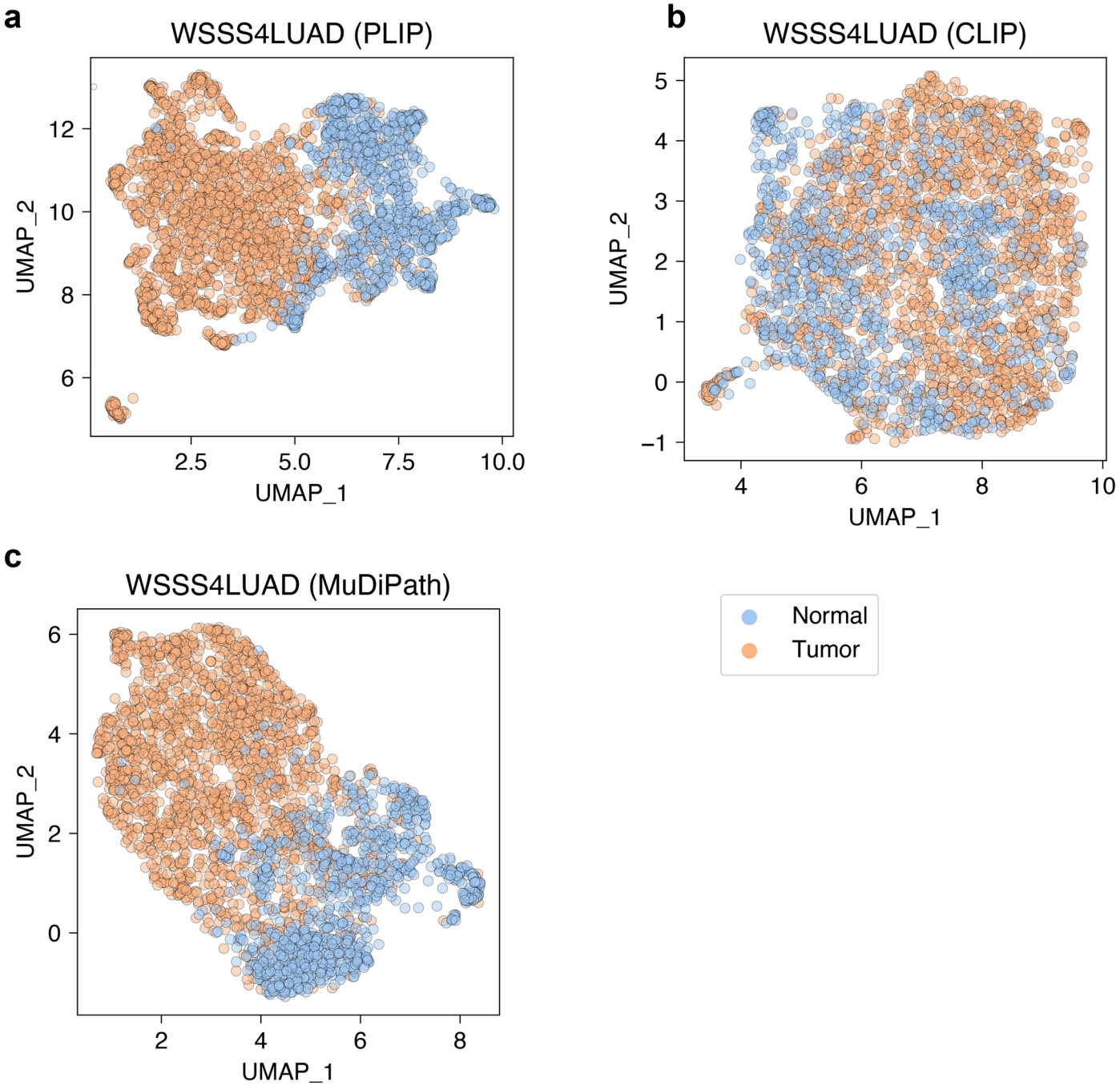
Extended Data Fig. 4 | Comparison of image embeddings between PLIP, CLIP, and MuDiPath models for the PanNuke dataset. a, Image embeddings generated by the PLIP model, colored by benign and malignant. **b**, Image embeddings generated by the PLIP model, colored by 19 pathology subspecialties. Scatters with black edges indicate malignant images. **c**, Image embeddings generated by the CLIP model, colored by benign and malignant. **d**, Image embeddings generated by the CLIP model, colored by 19 pathology subspecialties. Scatters with black edges indicate malignant images. **e**, Image embeddings generated by the MuDiPath model, colored by benign and malignant. **f**, Image embeddings generated by the MuDiPath model, colored by 19 organs. Scatters with black edges indicate malignant images.

https://doi.org/10.1038/s41591-023-02504-3



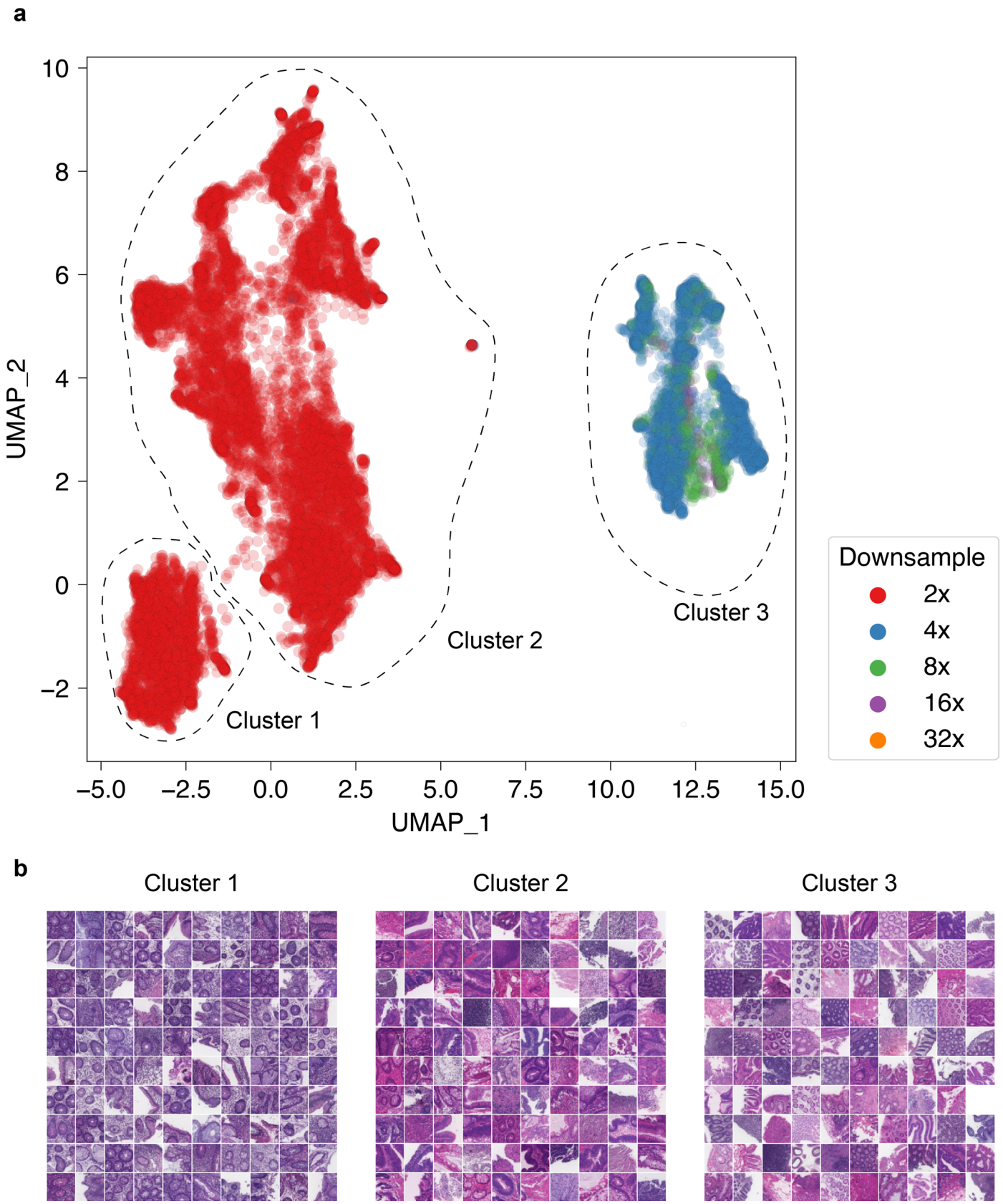
Extended Data Fig. 5 | Comparison of image embeddings derived from models in the DigestPath dataset. a, image embeddings derived from PLIP, **b**, image embeddings derived from baseline CLIP. **c**, image embeddings derived from MuDiPath.

https://doi.org/10.1038/s41591-023-02504-3



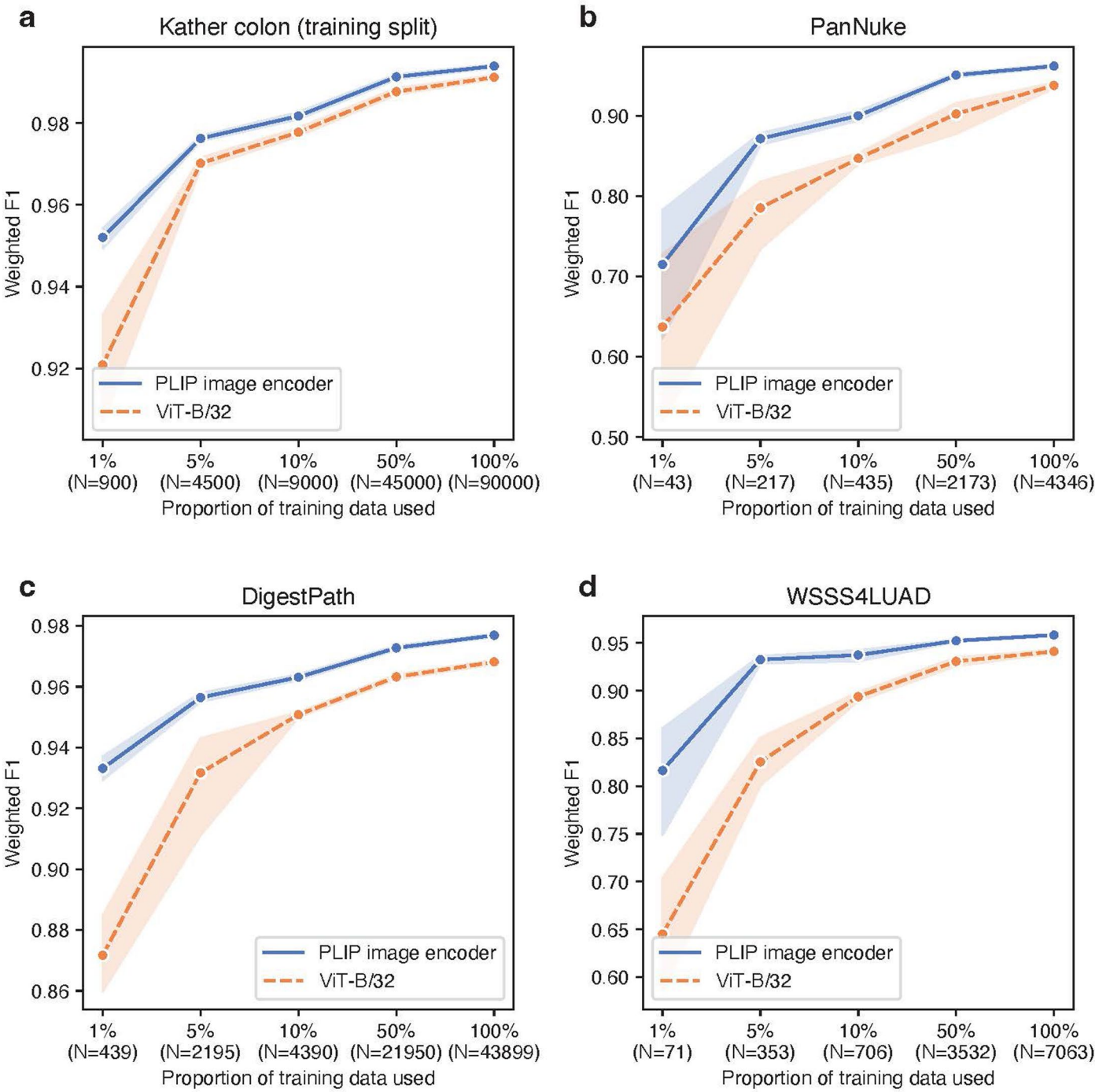
Extended Data Fig. 6 | Comparison of image embeddings derived from models in the WSSS4LUAD dataset. a, image embeddings derived from PLIP, **b**, image embeddings derived from baseline CLIP. **c**, image embeddings derived from MuDiPath.

https://doi.org/10.1038/s41591-023-02504-3



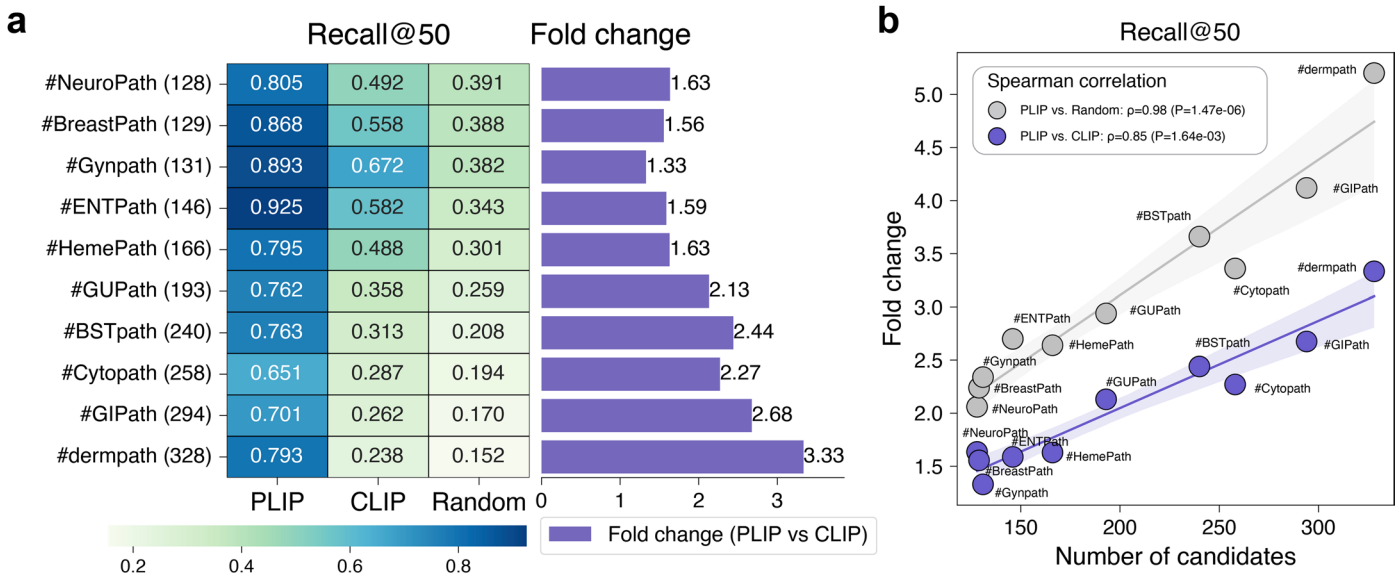
Extended Data Fig. 7 | Cluster visualization of images in DigestPath dataset. a, Image patches in low-dimensional space colored by different downsampling rates. **b**, Visualization of image patches on different clusters.

https://doi.org/10.1038/s41591-023-02504-3



Extended Data Fig. 8 | Comparison to supervised deep learning models. The fine-tuning was conducted on **a**, Kather colon dataset training split, **b**, PanNuke dataset, **c**, DigestPath dataset, and **d**, WSSS4LUAD dataset, by comparing the PLIP image encoder to ViT-B/32 (pre-trained on ImageNet). In the line plots, mean values and 95% confidence intervals are presented by using 10 different random seeds for subsetting the data and running the models. The improvements for PLIP are particularly large for smaller datasets. For instance, when comparing the weighted F1 scores across the four datasets using only 1% of the training data: (i) for Kather training split, the PLIP image encoder achieved F1 = 0.952, while ViT-B/32 achieved F1 = 0.921; (ii) for PanNuke dataset, the PLIP image encoder achieved F1 = 0.715, while ViT-B/32 achieved F1 = 0.637; (iii) for DigestPath dataset, the PLIP image encoder achieved F1 = 0.933, while ViT-B/32 achieved F1 = 0.872; (iv) for WSSS4LUAD dataset, the PLIP image encoder achieved F1 = 0.816, while ViT-B/32 achieved F1 = 0.645. When comparing the weighted F1 scores across the four datasets using all of the training data: (i) for Kather training split, the PLIP image encoder achieved F1 = 0.994, while ViT-B/32 achieved F1 = 0.991; (ii) for PanNuke dataset, the PLIP image encoder achieved F1 = 0.962, while ViT-B/32 achieved F1 = 0.938; (iii) for DigestPath dataset, the PLIP image encoder achieved F1 = 0.977, while ViT-B/32 achieved F1 = 0.968; (iv) for WSSS4LUAD dataset, the PLIP image encoder achieved F1 = 0.958, while ViT-B/32 achieved F1 = 0.941.

https://doi.org/10.1038/s41591-023-02504-3



Extended Data Fig. 9 | Text-to-image retrieval performances for Recall@50. **a**, Image retrieval performances for Recall@50 within each of the pathology subspecialty-specific hashtags. **b**, Two-sided Spearman correlations between the number of candidates and fold changes for Recall@50 when comparing the PLIP model with random and CLIP, respectively. In regression plots, the regression estimates are displayed with 95% confidence intervals in grey or purple colors.

https://doi.org/10.1038/s41591-023-02504-3

Extended Data Table 1 | List of pathology hashtags on Twitter used in this study

	Hashtag	Text	# tweets	# tweets with images	# images
1	Autopsy	Autopsy Pathology	33,804	7,287	8,444
2	BloodBank	Blood Banking & Transfusion Medicine	16,030	6,064	7,996
3	blooducation	Blood Banking & Transfusion Medicine	10,574	2,968	3,264
4	BreastPath	Breast Pathology	5,999	3,883	8,436
5	BSTpath	Bone And Soft Tissue Pathology	8,361	5,680	13,839
6	CardiacPath	cardiovascular pathology	1,788	1,104	1,788
7	ClinPath	Clinical Pathology	1,083	487	599
8	Cytopath	Cytopathology	13,036	9,005	16,531
9	dermpath	Dermatopathology	30,612	21,307	42,434
10	EndoPath	Endocrine Pathology	1,889	1,004	1,997
11	ENTPath	Head And Neck Pathology	6,342	4,506	9,541
12	EyePath	Ophthalmic Pathology	908	621	1,135
13	FNApath	Fine Needle Aspirate (FNA) Cytopathology	361	295	565
14	ForensicPath	Forensic Pathology & Forensics	1,042	464	719
15	GIPath	Gastrointestinal and Liver Pathology	17,214	11,985	27,284
16	GUPath	Genitourinary Pathology	15,216	9,500	19,819
17	Gynpath	Gynecologic Pathology	7,543	5,378	11,597
18	HemePath	Hematopathology	18,395	11,088	20,671
19	HPBpath	Hepatobiliary pathology	128	59	122
20	IDpath	Infectious Disease Pathology	733	556	1,063
21	MolDx	Molecular Pathology	1,502	788	959
22	nephpath	Nephropathology	207	130	240
23	NeuroPath	Neuropathology	9,978	4,424	9,052
24	OralPath	Oral Pathology	2,083	1,403	2,813
25	pancpath	Pancreatic pathology	362	187	429
26	PathGME	Pathology Graduate Medical Education	19	11	14
27	pathInformatics	Pathology Informatics	158	92	134
28	patientbloodmanagement	Blood Banking & Transfusion Medicine	1,752	504	561
29	PediPath	Pediatric Pathology	8,028	4,424	9,072
30	PulmPath	Pulmonary And Pleural Pathology	6,611	3,891	8,430
31	RenalPath	Renal and Medical Kidney Pathology	5,933	3,738	6,563
32	SurgPath	Surgical Pathology	4,376	3,405	7,264
		Total	232,067	126,238	243,375

The hashtags used from 21 March 2006 (the date of the first Twitter post) to 15 November 2022 and the original number of data before data filtering are presented.

nature portfolio

Corresponding author(s): James Zou

Last updated by author(s): Jul 11, 2023

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	Data was collected from Twitter and LAION using Twitter API v2.0 with python v3.9.
Data analysis	All experiments were run in python 3.9. Detailed software versions are: pytorch v1.13, CUDA v11.7, CUDNN v8.5.0, scipy v1.9.3, torchvision v0.14.0, pillow v9.1.0, scikit-learn v1.1.2, scikit-image v0.19.2, pandas v1.4.2, numpy v1.23.5, and multiprocess v0.70.13. The trained model and source codes can be accessed at https://tinyurl.com/webplip.

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

All data in OpenPath is publicly available from Twitter and LAION-5B (https://laion.ai/blog/laion-5b/). Twitter IDs used for training and validation can be accessed at https://tinyurl.com/openpathdata. Validation datasets are all publicly available and can be accessed from: Kather colon dataset (https://zenodo.org/

record/1214456), PanNuke (https://warwick.ac.uk/fac/cross_fac/tia/data/pannuke), DigestPath (https://digestpath2019.grand-challenge.org/), WSSS4LUAD (https://wsss4luad.grand-challenge.org/), PathPedia (https://www.pathpedia.com/Education/eAtlas/Default.aspx), PubMed and Books pathology collection (https://warwick.ac.uk/fac/cross_fac/tia/data/arch), KIMIA Path24C (https://kimialab.uwaterloo.ca/kimia/index.php/pathology-images-kimia-path24/). ImageNet dataset (https://www.image-net.org/) was adopted for the pre-trained ViT-B/32 model. The trained model, source codes, and interactive results can also be accessed at https://tinyurl.com/webplip.

Human research participants

Policy information about studies involving human research participants and Sex and Gender in Research.

Reporting on sex and gender	No sex- or gender-based analysis in this study as all the data consisted of de-identified pathology image patches.
Population characteristics	No population characteristics applied to this study as all the data consisted of de-identified pathology image patches.
Recruitment	No participants were recruited in this study as all the data consisted of de-identified pathology image patches.
Ethics oversight	We confirm that there is no human subject involved in this study as all the data consisted of de-identified pathology image patches.

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	208,414 pathology images paired with natural language were used in this study to ensure an adequate representation of the digital pathology images under investigation. No statistical methods were used to predetermine sample size. Sample size was then determined by data acquisition followed with a data exclusion and data filtering pipeline.
Data exclusions	A rigorous protocol was applied for data exclusion, including the removal of retweets, sensitive tweets, and non-pathology images. The exclusion criteria is established in this manuscript.
Replication	We confirm that all experimental findings can be reproduce with our source code provided.
Randomization	For Twitter dataset, the training / validation groups were divided based on a datetime criterion. For external validation datasets, samples were randomly allocated to either training or validation groups using a 70%:30% ratio.
Blinding	Blinding was not applicable to our study as all the data was de-identified.

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

Methods

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