Article

AI-based pathology predicts origins for cancers of unknown primary

https://doi.org/10.1038/s41586-021-03512-4

Received: 27 June 2020

Accepted: 1 April 2021

Published online: 5 May 2021

Check for updates

Ming Y. Lu[1,2,3], Tiffany Y. Chen[1,2,5], Drew F. K. Williamson[1,2,5], Melissa Zhao[1], Maha Shady[1,2,3,4], Jana Lipkova[1,2,3] & Faisal Mahmood[1,2,3] ✉

Cancer of unknown primary (CUP) origin is an enigmatic group of diagnoses in which the primary anatomical site of tumour origin cannot be determined[1,2]. This poses a considerable challenge, as modern therapeutics are predominantly specific to the primary tumour[3]. Recent research has focused on using genomics and transcriptomics to identify the origin of a tumour[4–9]. However, genomic testing is not always performed and lacks clinical penetration in low-resource settings. Here, to overcome these challenges, we present a deep-learning-based algorithm—Tumour Origin Assessment via Deep Learning (TOAD)—that can provide a differential diagnosis for the origin of the primary tumour using routinely acquired histology slides. We used whole-slide images of tumours with known primary origins to train a model that simultaneously identifies the tumour as primary or metastatic and predicts its site of origin. On our held-out test set of tumours with known primary origins, the model achieved a top-1 accuracy of 0.83 and a top-3 accuracy of 0.96, whereas on our external test set it achieved top-1 and top-3 accuracies of 0.80 and 0.93, respectively. We further curated a dataset of 317 cases of CUP for which a differential diagnosis was assigned. Our model predictions resulted in concordance for 61% of cases and a top-3 agreement of 82%. TOAD can be used as an assistive tool to assign a differential diagnosis to complicated cases of metastatic tumours and CUPs and could be used in conjunction with or in lieu of ancillary tests and extensive diagnostic work-ups to reduce the occurrence of CUP.

The site of a primary tumour has an important role in guiding the clinical care of patients with metastatic tumours and can typically be determined through the histopathological examination of tissue and through a clinical and radiological assessment of the patient. Despite improvements through sophisticated imaging modalities, specific and sensitive testing using immunohistochemistry (IHC), a concrete determination of the site of origin of the primary tumour can still be a diagnostic challenge. In fact, 1–2% of cancers are often categorized as CUPs, for which the anatomic site of primary origin cannot be assigned despite extensive diagnostic investigation and clinical correlation[1,2]. Patients with CUP often undergo comprehensive diagnostic work-ups including pathology, radiology, endoscopic and laboratory examinations to determine the occult primary site[2,3], as most cases of CUP for which a putative primary origin cannot be assigned are treated with empirical combination chemotherapy and have a poor prognosis (median overall survival of 2.7–16 months)[1,2]. Recent studies have proposed using genomics and transcriptomics to identify the primary origin[4–10]. However, molecular profiling is not routinely performed for every patient, especially in low-resource settings. In addition, uncertainty in classifying a tumour as primary or metastatic and misdiagnosing a relapse of an antecedent malignancy have also been reported in the literature[11,12]. Advances in deep learning[13] have demonstrated accurate, reliable and reproducable performance on a variety of different diagnostic tasks in medicine, including the ability to readily identify characteristics that are not typically recognized by human experts[14–24].

AI-based assessment of tumour origins

Here we present TOAD, a high-throughput, interpretable deep-learning-based solution that uses scanned haematoxylin and eosin whole-slide images (WSIs)—which are routinely used for clinical diagnosis—for the identification of the site of primary origin for tumour specimens. Our approach can be used to simultaneously predict whether a tumour is metastatic and assign a differential diagnosis for the origin of the primary tumour. TOAD can act as an assistive tool for pathologists for assessing complicated cases of metastatic and unknown primary tumours for which a large number of clinical and ancillary tests are required to narrow a differential diagnosis.

Our study uses 32,537 gigapixel WSIs from large public data repositories and the Brigham and Women's Hospital, spanning 18 common origins of primary cancer (see Methods, Supplementary Tables 1, 2). We trained our model using 22,833 gigapixel WSIs using a weakly supervised multitask training paradigm. We assess the performance of TOAD by first testing on a held-out test set of 6,499 WSIs with known primary tumour origins, and then analysing the subset of increasingly difficult

[1]Department of Pathology, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA. [2]Cancer Program, Broad Institute of Harvard and MIT, Cambridge, MA, USA. [3]Cancer Data Science Program, Dana-Farber Cancer Institute, Boston, MA, USA. [4]Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA. [5]These authors contributed equally: Tiffany Y. Chen, Drew F. K. Williamson. ✉e-mail: faisalmahmood@bwh.harvard.edu

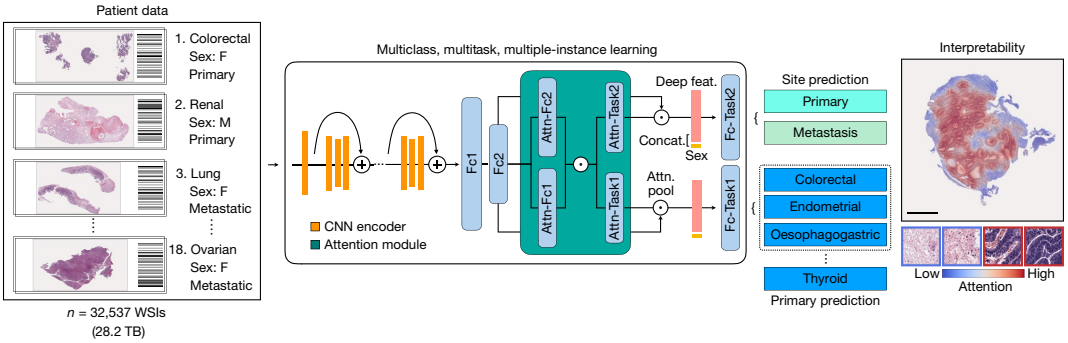


Fig. 1 | TOAD workflow. Patient data in the form of digitized high-resolution histology slides serve as the input into the main network. For each WSI, the tissue content is automatically segmented and divided into an average of thousands to tens of thousands of regions as small image patches. These images are processed by a convolutional neural network (CNN) with fixed pretrained parameters, which serves as an encoder to extract a compact, descriptive feature vector from each patch. Using an attention-based multiple-instance learning algorithm, TOAD learns to rank all of the tissue regions in the slide using their feature vectors and aggregate their information across the whole slide based on their relative importance, assigning greater weights to regions perceived to have high diagnostic relevance. As an additional covariate, the sex of the patient can be fused with the aggregated histology features to further guide classification. By using a multi-branched network architecture and a multitask objective, TOAD can predict both the tumour origin and whether the cancer is primary or metastatic. Additionally, the attention scores that the network assigns to each region can be used to interpret the prediction of the model. Further details of the model architecture are described in the Methods. Scale bar, 0.5 mm. Attn, attention; Concat, concatenation; Fc1, Fc2, fully connected layers; feat, features; F, female; M, male.

cases of metastatic tumours to determine the capability of TOAD for assigning a differential diagnosis (Supplementary Table 3). To further assess the adaptability of our model, we evaluated the model using an external test set of 682 samples submitted from more than 200 medical centres (Supplementary Table 4). We also assessed our model using an additional test dataset of 317 cases of CUP that were assigned a primary differential based on ancillary tests, radiology, patient history, clinical correlation or at autopsy (Supplementary Table 5). An overview of the study design is provided in Extended Data Fig. 1.

We combine transfer learning and weakly supervised multitask learning to enable a single, unified predictive model to be efficiently trained on tens of thousands of gigapixel WSIs. Using attention-based learning[15,25], our approach automatically locates regions in the slide that are of high diagnostic relevance and aggregates their information to make the final predictions. By visualizing the attention scores, the relative importance of each region examined by the model can be displayed for human interpretability and validation. We incorporate the sex of the patient into the prediction of the model by fusing it with slide-level features before the final classification layers, which predict both the origin of the cancer and whether the tumour is primary or metastatic. The two classification problems are learned jointly during training by using a multitask objective and sharing the model parameters of intermediate layers (Fig. 1). Separate attention layer weights, however, are learned for each task to increase the expressivity of the model, allowing it to attend to different sets of information-rich regions of the slide depending on the task. Further details of the model architecture are described in the Methods.

Evaluation of model performance

On the held-out test set ($n = 6{,}499$), which was not seen by the model during training, our model achieved an overall accuracy of 83.4%. When the model is evaluated using top-k differential diagnosis accuracy—that is, how often the ground truth label is found in the k highest confidence predictions of the model—TOAD achieved a top-3 accuracy of 95.5% and top-5 accuracy of 98.1% (Fig. 2e). The top predictions can be useful for complicated cases of metastatic tumours and CUP for which narrowing down potential primary tumour origins can assist with the diagnostic workflow and reduce the number of ancillary tests that are required to find the primary tumour. Performance for each individual primary site is shown in Fig. 2a and a summary table of classification performance metrics is included in Supplementary Table 6. Furthermore, detailed performance on metastatic tumours is shown in Extended Data Fig. 2, 3, 6a and Supplementary Table 7. Performance stratified by detailed cancer subtype is shown in Supplementary Table 8 and per-case assessments can be found in Supplementary Table 9. Ablation studies, which analyse the benefit of using the sex of the patient and multitask learning for the prediction of cancer origins, as well as the effect of adding the tissue sampling or biopsy site as an input covariate are discussed in Extended Data Fig. 4 and Supplementary Figs. 1, 2.

The high top-k accuracy suggests that we can potentially use the top predictions of TOAD for a given slide to narrow down the origin of the tumour to a handful of possible origins (Fig. 2e). Predictions made by the model with a high confidence were observed to be generally reliable (Fig. 2b, g). In addition, TOAD was able to predict whether the tumour specimen is a metastasis with an accuracy of 85.0% and an area under the receiver operator characteristic curve (AUC ROC) of 0.942 (Fig. 2f). We further analyse the performance on this binary task for a number of tissue sampling sites at which both primary and metastatic tumours are frequently found as well as for cases stratified by the same primary origin (Extended Data Fig. 5).

Generalization to external test cohort

To assess the adaptability of our model across different healthcare systems with different staining protocols and patient populations, we also validated TOAD on an additional test set of 682 external patients submitted from more than 200 US and international medical centres (Supplementary Fig. 3 and Supplementary Table 4). Without tuning or domain adaptation, our trained model produced an accuracy of 79.9% and a top-3 accuracy of 93.4% (Fig. 2c, e). Similarly, on the second task of distinguishing between a metastasis and a primary tumour, the model had an AUC of 0.919 (Fig. 2d, f). The performance indicates that our model is capable of generalization to diverse data sources and staining protocols that were not encountered during training. Individual case assessments are available in Supplementary Table 10.

Evaluation on difficult cases of metastatic tumours

It is challenging to objectively evaluate the ability of the model to predict the origin of the tumour for cases of CUP because of limited ground truth labels and the inherent associated uncertainty even when

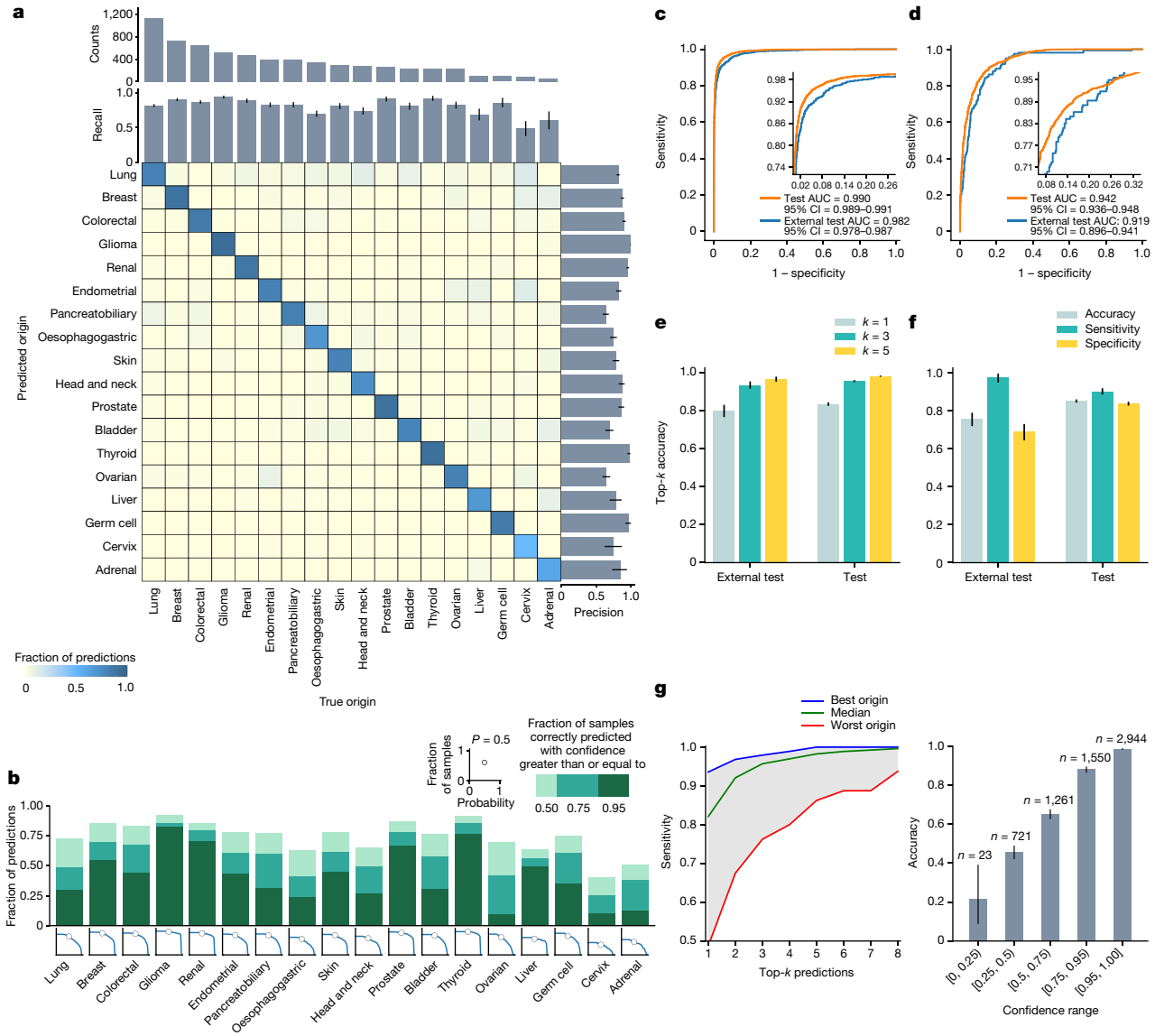


Fig. 2 | Model performance of TOAD. a, Slide-level performance for the prediction of the tumour origin on the overall held-out test set ($n = 6{,}499$) for 18 primary origins ($n = 5{,}091$ primary and $n = 1{,}408$ metastatic). Analysis on only metastatic tumours is shown in Extended Data Figs. 2, 3. Per origin count, precision and recall are plotted next to the confusion matrix. Columns of the confusion matrix represent the true origin of the tumour and rows represent the origins predicted by the model. **b**, Top, for each origin, the fraction of samples that was correctly classified with a confidence (probability) score above certain thresholds is shown. Bottom, the fraction of samples (*y* axis) that was correctly classified at or above a certain confidence threshold (*x* axis, computed over increments of 0.025 in probability score) is shown. **c**, Micro-averaged one-versus-rest ROC curves for the classification of the primary origin, evaluated on the test set ($n = 6{,}499$) and an independent test set of external cases only ($n = 682$). **d**, ROC curves for the auxiliary binary task of predicting primary versus metastasis in the test set ($n = 6{,}499$) and external test set ($n = 682$). **e**. Top-*k* model accuracies for tumour origin classification on the test set ($n = 6{,}499$) and external test set ($n = 682$) for $k \in \{1, 3, 5\}$. **f**, Accuracy, sensitivity and specificity of the model for predicting primary versus metastasis in the test set ($n = 6{,}499$) and external test set ($n = 682$). Without loss of generality, metastasis is defined as the 'positive' class. Performance for tumours at common metastatic sites and for tumours grouped by primary origin are further analysed in Extended Data Fig. 5. **g**, Further analysis of the model predictions on the test set ($n = 6{,}499$). Left, tumour origin prediction based on the top-*k* predictions of the model for $k \in \{1, 2, \ldots, 8\}$. The band bounded between the best and worst performing origin for each value of *k* is shaded in grey. Right, accuracy of the predictions for different bins of prediction confidence. **a**, **e**–**g**, Error bars indicate 95% confidence intervals (CIs); the centre is always the computed value of each classification performance metric (specified by its respective axis labels).

such labels do exist. Therefore, we first analysed the performance of TOAD on metastatic tumours in our test set for which a diagnosis is assigned and available (Extended Data Figs. 2, 3b). For these metastatic tumours ($n = 1{,}408$), TOAD achieved an accuracy of 82.8% and a top-3 accuracy of 94.9% (Extended Data Fig. 6a). We used the number of diagnostic IHC stains performed as an indirect measure to identify difficult-to-diagnose cases[26] and examined the performance of the model across different levels of IHC usage. As expected, in patients who were diagnosed without requiring IHC ($n = 603$), TOAD scored the highest accuracy of 87.4% and a top-3 accuracy of 96.7%. However, even in more difficult cases ($n = 449$) who required three (75th percentile) or more IHC tests, TOAD still achieved an accuracy of 75.7%, and a top-3 accuracy of 92.0%. Additionally, we identified and analysed the performance on two other subsets of challenging cases, including 88 patients whose tumours could not be diagnosed with IHC analysis and required further clinical correlation (top-1 accuracy, 79.5%; top-3 accuracy, 95.5%) and 215 cases whose tumours were characterized as poorly differentiated in the pathology reports (top-1 accuracy, 77.2%; top-3

accuracy, 92.1%). It is worth noting that the model was able to achieve the reported performance without having access to additional clinical variables, or IHC results, as it makes its predictions solely on the basis of the digitized haematoxylin and eosin slide and the sex of the patient. We also calculated Cohen's κ score[27], which measures the inter-observer agreement between the model and the assigned diagnosis, while taking into account agreement by chance. The κ scores fell in the range of substantial agreement for cases of metastatic tumours (Extended Data Fig. 6a) as well as on our overall test ($\kappa = 0.819$) and external test ($\kappa = 0.761$) sets. Overall, the performance of the model based on routine histology is largely comparable to the performance reported by several recent origin prediction studies based on genomics[4,5] (Supplementary Table 11).

Evaluation on patients with CUP

We further curated a dataset of 743 patients from 152 medical centres that were assigned a diagnosis of CUP at some point during the course of diagnosis and treatment (Supplementary Fig. 3). These patients could not be assigned a primary diagnosis using the histology slide alone and required thorough clinical work-ups and ancillary tests. After analysing all electronic medical records (EMRs), we identified a subset of 317 patients that were assigned a primary differential (Supplementary Table 5). These differential diagnoses for cases of CUP involve elements of uncertainty and conjecture and should be distinguished from confident, ground truth labels, which cannot be realistically obtained for cases of CUP.

We observed that the top prediction of the model directly agreed with the origin site indicated by the primary differential assigned in 192 of the 317 patients (60.6%; $\kappa = 0.520$) and the agreement reached 82.0% and 92.1%, respectively, when considering the top-3 and top-5 predictions of the model (Extended Data Fig. 6b and Supplementary Table 12). These results are encouraging as our model was able to assign concordant differential diagnoses on the basis of the routine histology image, whereas differential diagnoses for cases of CUP are commonly assigned after extensive investigative diagnostic work-ups. We also noted higher agreement in high-confidence predictions (Extended Data Fig. 6b).

We further categorized the cases of CUP into high-certainty ($n = 193$) and low-certainty ($n = 124$) diagnoses on the basis of the strength of the evidence used to make the determination and language used in the EMRs. As expected, agreement is low for cases in the low-certainty bin ($\kappa = 0.372$) compared to high-certainty diagnoses ($\kappa = 0.611$). In Extended Data Fig. 8 and Supplementary Fig. 4, we demonstrate how the top predictions of the model could be used in conjunction with IHC testing to assist in assigning a primary differential.

Owing to the difficulty of assessing cases of CUP, pathologists sometimes assign multiple possible primary origins in pathology reports. We identified a subset of 73 patients who were assigned multiple primary origins in the pathology reports (median, 2; minimum, 2; maximum, 5). We found that in 62 out of 73 cases (84.9%), 50% or more of the possible primary origins that were assigned in the report were also predicted by the model. Similar to a previous origin prediction study based on genomics[5], we also analysed the confidence of the model in cases for which there was no diagnosis assigned or EMRs were missing. The average confidence of the model for these 426 cases of CUP is 0.611 (s.d., 0.209) and 269 out of 426 cases (63.1%) are predicted with a confidence of 0.5 or higher.

Model interpretability

For each slide, the attention of the model can be visualized for human interpretability and validation (see Methods). Examples of attention heat maps for metastatic tumours are shown in Extended Data Figs. 7, 8 and Supplementary Fig. 4. High-resolution heat maps for cases of all primary sites can be accessed through our interactive demo website (http://toad.mahmoodlab.org). We also quantitatively analysed the relative proportions of cell types captured by the high-attention regions that the model identified in metastatic tumours in our test set (Extended Data Fig. 9 and Supplementary Table 13), which shows that the attention of the model mainly captures regions of tumour cells for all considered primary origins.

Discussion

We present TOAD, a deep learning algorithm developed to predict the primary origin of a tumour based on routine histopathology slides. It can be challenging to identify the origins of metastatic tumours using minimal clinical information and especially when evaluating the tumour on the basis of histology alone[28]. We show that using routine H&E histology and the sex of the patient as input, our model, which was trained using weakly supervised learning on tens of thousands of samples, can make reasonably accurate predictions, particularly in assigning top-3 or top-5 differential diagnoses for cases of metastatic tumours that required ancillary testing, radiological imaging or clinical correlations to diagnose. Similarly, for patients with CUPs for whom a primary differential was assigned, often after extensive clinical work-ups, the model was able to make predictions that are in agreement to a meaningful degree with the assigned primary differential.

We demonstrate that our single network trained using multitask learning can additionally predict whether a tumour is metastatic and can distinguish between primary and metastatic tumours found at the same tissue site, which poses occasional diagnostic challanges[11,12,29]. As an example, when asked to predict whether tumours found in the central nervous system and liver were primary or metastatic, the model achieved an AUC ROC of 0.971 and 0.952, respectively (see Extended Data Fig. 5 for additional examples). From additional experiments, we verified that weakly supervised AI models can also be developed to predict the primary origin for subsets of tumours that share the same morphological appearance or have metastasized to a common site (Extended Data Fig. 10).

In resource-constrained settings in which ancillary and clinical testing, advanced imaging and necessary pathology expertise may not be available, the top-1 origin prediction from TOAD can potentially be used to assign a primary differential. We also envision that TOAD can be used as an assistive tool by pathologists. Pertaining to the high top-k accuracy of the model, TOAD can be used to narrow the potential primary origins using the top predicted differential diagnoses, followed by more informed ancillary testing or other clinical correlations to obtain a final diagnosis (Extended Data Fig. 8 and Supplementary Fig. 4). We envision the TOAD-assisted workflow to present potential benefits such as fewer clinical and ancillary tests, reduced tissue sampling, and more accurate and standardized predictions. In both settings, we believe that TOAD can help to reduce the turnaround time and resources for diagnosing complicated metastatic tumours and cases for which it may be unclear whether the tumour is primary or metastatic. Similarly, when the diagnosis of a patient is initially ambiguous, as a secondary reader, TOAD can help to prompt re-evaluation when the model produces a high-confidence prediction that disagrees with the assessment of the expert or encourage exploration of alternative hypotheses that were previously overlooked. Several recent studies[15,22] have shown that weakly supervised deep learning can be used to train accurate cancer-subtyping models without detailed human annotations. Therefore, if required, additional classifier models may be trained for each origin to predict the tumour subtype and guide further treatment decisions once the TOAD model has predicted the primary origin. Incidences of CUP have been decreasing in part because of the advances in radiological and molecular assessments coupled with sensitive ancillary tests[30]. Our approach has the potential to further contribute to this trend by assisting with the fast diagnosis of

complicated cases of metastatic tumours and CUP. Overall, our study serves as a proof-of-concept for developing large-scale, weakly supervised AI models for origin prediction and paves the way for prospective studies and clinical trials to further assess the efficacy of AI-based origin prediction using conventional histology.

Online content

Any methods, additional references, Nature Research 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/s41586-021-03512-4.

1. Rassy, E. & Pavlidis, N. Progress in refining the clinical management of cancer of unknown primary in the molecular era. *Nat. Rev. Clin. Oncol*. **17**, 541–554 (2020).
2. Varadhachary, G. R. & Raber, M. N. Cancer of unknown primary site. *N. Engl. J. Med*. **371**, 757–765 (2014).
3. Massard, C., Loriot, Y. & Fizazi, K. Carcinomas of an unknown primary origin—diagnosis and treatment. *Nat. Rev. Clin. Oncol*. **8**, 701–710 (2011).
4. Jiao, W. et al. A deep learning system accurately classifies primary and metastatic cancers using passenger mutation patterns. *Nat. Commun*. **11**, 728 (2020).
5. Penson, A. et al. Development of genome-derived tumor type prediction to inform clinical cancer care. *JAMA Oncol*. **6**, 84–91 (2020).
6. Grewal, J. K. et al. Application of a neural network whole transcriptome-based pan-cancer method for diagnosis of primary and metastatic cancers. *JAMA Netw. Open* **2**, e192597 (2019).
7. Zhao, Y. et al. CUP-AI-Dx: a tool for inferring cancer tissue of origin and molecular subtype using RNA gene-expression data and artificial intelligence. *EBioMedicine* **61**, 103030 (2020).
8. Shen, Y. et al. TOD-CUP: a gene expression rank-based majority vote algorithm for tissue origin diagnosis of cancers of unknown primary. *Brief. Bioinformatics* **22**, 2106–2118 (2020).
9. Kerr, S. E. et al. Multisite validation study to determine performance characteristics of a 92-gene molecular cancer classifier. *Clin. Cancer Res*. **18**, 3952–3960 (2012).
10. Hayashi, H. et al. Site-specific and targeted therapy based on molecular profiling by next-generation sequencing for cancer of unknown primary site: a nonrandomized phase 2 clinical trial. *JAMA Oncol*. **6**, 1931–1938 (2020).
11. Nass, D. et al. MiR-92b and miR-9/9* are specifically expressed in brain primary tumors and can be used to differentiate primary from metastatic brain tumors. *Brain Pathol*. **19**, 375–383 (2009).
12. Estrella, J. S., Wu, T. T., Rashid, A. & Abraham, S. C. Mucosal colonization by metastatic carcinoma in the gastrointestinal tract: a potential mimic of primary neoplasia. *Am. J. Surg. Pathol*. **35**, 563–572 (2011).
13. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).
14. Liu, Y. et al. A deep learning system for differential diagnosis of skin diseases. *Nat. Med*. **26**, 900–908 (2020).
15. Lu, M. Y. et al. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nat. Biomed. Eng*. https://doi.org/10.1038/s41551-020-00682-w (2021).
16. Campanella, G. et al. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nat. Med*. **25**, 1301–1309 (2019).
17. Chen, P. C. et al. An augmented reality microscope with real-time artificial intelligence integration for cancer diagnosis. *Nat. Med*. **25**, 1453–1457 (2019).
18. Ouyang, D. et al. Video-based AI for beat-to-beat assessment of cardiac function. *Nature* **580**, 252–256 (2020).
19. Hollon, T. C. et al. Near real-time intraoperative brain tumor diagnosis using stimulated Raman histology and deep neural networks. *Nat. Med*. **26**, 52–58 (2020).
20. Esteva, A. et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **542**, 115–118 (2017).
21. Kalra, S. et al. Pan-cancer diagnostic consensus through searching archival histopathology images using artificial intelligence. *NPJ Digit. Med*. **3**, 31 (2020).
22. 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).
23. Kather, J. N. et al. Deep learning can predict microsatellite instability directly from histology in gastrointestinal cancer. *Nat. Med*. **25**, 1054–1056 (2019).
24. Poplin, R. et al. Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. *Nat. Biomed. Eng*. **2**, 158–164 (2018).
25. Ilse, M., Tomczak, J. M. & Welling, M. Attention-based deep multiple instance learning. In *International Conference on Machine Learning* 2132–2141 (2018).
26. Handorf, C. R. et al. A multicenter study directly comparing the diagnostic accuracy of gene expression profiling and immunohistochemistry for primary site identification in metastatic tumors. *Am. J. Surg. Pathol*. **37**, 1067–1075 (2013).
27. McHugh, M. L. Interrater reliability: the kappa statistic. *Biochem. Med*. **22**, 276–282 (2012).
28. Sheahan, K. et al. Metastatic adenocarcinoma of an unknown primary site. A comparison of the relative contributions of morphology, minimal essential clinical data and CEA immunostaining status. *Am. J. Clin. Pathol*. **99**, 729–735 (1993).
29. Jurmeister, P. et al. Machine learning analysis of DNA methylation profiles distinguishes primary lung squamous cell carcinomas from head and neck metastases. *Sci. Transl. Med*. **11**, eaaw8513 (2019).
30. Rassy, E. & Pavlidis, N. The currently declining incidence of cancer of unknown primary. *Cancer Epidemiol*. **61**, 139–141 (2019).

Methods

Data reporting

No statistical methods were used to determine sample size. We used all available data from large public repositories and in house data (2004–2020). For rare origins and subtypes the sample size was limited by the number of cases available in public repositories and in-house or off-site pathology archives. For disease models with abundant slides deep learning models were trained while increasing the number of input slides until asymptotic improvement of the model performance was achieved to suggest an appropriate sample size was obtained. The data were split randomly into train, validation and held-out test sets; no other form of randomization was used. The study involved retrospective analysis of pathology slides and patients were not directly involved or recruited for the study. No form of blinding was used or required for this retrospective study. The Mass General Brigham institutional review board approved the retrospective analysis of pathology slides and corresponding pathology reports. Informed consent was waived for analysing archival pathology slides retrospectively. All pathology slides were de-identified before scanning and digitization. All digital data, including whole slide images, pathology reports and EMRs were de-identified before computational analysis and model development.

Dataset description

Our overall dataset, which includes publicly available and in-house data, was composed of 32,537 digitized haematoxylin and eosin (H&E) slides (25,419 primary and 7,118 metastatic WSIs from 29,107 patient; 52.8% women, 47.2% men), spanning 18 groups of common primary cancer origins; each origin encompasses both common and rare tumour subtypes (Supplementary Tables 1, 2). This roughly amounted to 28 terabytes of raw data. All WSIs were processed and analysed at the 20× equivalent magnification. The dataset is randomly partitioned and is stratified by class, into a training set (70% of cases), a validation set (10% of cases) and a test set (20% of cases), as shown in Supplementary Table 3. The partitioning was performed at the patient level and therefore all slides from the same patient are always placed into the same set. As some patients in the The Cancer Genome Atlas (TCGA) and Clinical Proteomic Tumour Analysis Consortium (CPTAC) datasets have multiple slides per patient, 10.6% (692 out of 6,499) of slides in the test set come from patients with more than one slide per patient.

Publicly available data. We queried TCGA Data Commons and CPTAC Pathology Data Portal for WSIs containing tumour tissue corresponding to the 18 grouped classes of primary origins. Among slides obtained from the TCGA, only representative formalin-fixed, paraffin-embedded diagnostic slides of primary and metastatic tumours were considered, and slides that did not contain tumours, lacked lower magnification downsamples or had missing sex information were excluded. Similarly, non-tumour-containing slides from CPTAC and cases without sex information were excluded before analysis. In total, we gathered 10,406 WSIs from 8,794 patients across 25 TCGA studies and 2,969 WSIs from 1,151 patients across 7 CPTAC studies.

In-house data. For model development and evaluation, we additionally curated a dataset of 19,162 WSIs (12,215 primary, 6,947 metastatic) from internal patients at the Brigham and Women's Hospital. These slides were collected between 2004 and 2020 and either scanned at 20× using an Aperio GT450 scanner or 40× using a Hamamatsu S210 scanner. Unless indicated otherwise, each WSI corresponds to a unique patient. For in-house patients, the assigned diagnosis is always based on review by multiple pathologists. Many challenging cases of metastatic tumours were difficult to diagnose with certainty based on H&E histology alone and were diagnosed based on concurrence by multiple clinical experts often in conjunction with clinical correlation and ancillary tests.

External test set. As an independent test cohort, we used 682 external cases received at the Brigham & Women's Hospital from 203 medical centres across 34 states in USA and 20 international medical centres from 9 other countries (see Supplementary Fig. 3 for geographical diversity and Supplementary Table 4 for per-origin breakdown and more detailed summary). Slides for these patients were prepared at their respective institutions using a variety of different tissue preparation, processing and staining protocols. These submitted cases were reviewed by in-house expert pathologists and received a final diagnosis similar to the cases in our in-house dataset, while taking into consideration other submitted materials including IHC analyses (in addition to the H&E sections) and previous clinical history where applicable. Additional IHC analyses were sometimes ordered at the discretion of the experts reviewing the case.

CUP test set. To further validate our model, we also identified 743 patients that were assigned a diagnosis of CUP. These patients were received from 146 medical centres across 21 US states and 6 international centres from 3 other countries (see Supplementary Fig. 3 for geographical diversity). For each case, we reviewed electronic medical records (EMRs) including the pathology report in combination with laboratory results, patient history, oncology, radiology, endoscopy and autopsy reports, treatment and follow-up history where applicable and if available; we determined a subset of 317 patients with a primary differential. These differential diagnoses were assigned during the course of diagnosis or treatment. It was verified that none of these cases could be diagnosed using histology alone and required extensive ancillary and clinical testing. Although it is not possible to obtain an absolute ground truth origin for cases of CUP, the assigned primary differential diagnosis was used to assess the value of our model in assigning appropriate differential diagnoses to cases of CUP using histology alone. For further analysis, we additionally split these 317 cases (Supplementary Table 5) into high-certainty (n = 193) and low-certainty (n = 124) based on the language and evidence used in the EMRs. High-certainty diagnoses were defined as being compatible on the basis of stronger evidence supported by IHC findings or clinical, radiological or molecular correlation, whereas low-certainty diagnoses include cases that may not suggest a specific primary or lacked definitive supporting evidence for the putative differential assigned.

TOAD model architecture

We used deep learning to simultaneously predict the origin of the tumour in each WSI and whether it was the primary tumour or a metastasis. Owing to the enormous size of gigapixel WSIs as well as the large variation in the shape of tissue content captured by the image, it is generally considered inefficient or intractable to use deep learning algorithms based on convolutional neural networks directly on top of the entire WSI for training or inference. Although it is possible to use smaller regions of interests for training, the drawback of this approach is that as the slide-level diagnosis (for example, lung adenocarcinoma) is only manifested in a fraction of the tissue content in the WSI—unless human expertise and manual labour is involved to ensure these smaller regions are representative of the diagnosis made for the entire slide—naively associating them with the slide-level diagnosis will lead to noisy and erroneous labels. To overcome this limitation, we used a form of weakly supervised machine learning known as multiple instance learning. By considering each WSI as a collection (known as a bag) of smaller image regions (known as instances), we trained a multitask neural network model directly with slide-level labels without the need for manual extraction of regions of interests, while taking into account information from the entire slide. For computational efficiency, after tissue segmentation and patching (see 'WSI processing'), we first performed dimensionality reduction on the raw image data by encoding each 256 × 256 RGB image patch into a descriptive 1,024-dimensional

feature vector using a ResNet50-based[31] convolutional neural network with fixed parameters pretrained on ImageNet[32]. Learning subsequently occurs on the fixed, low-dimensional feature representations instead of the pixel space. In the feature space, the information from all of the tissue regions in each slide is aggregated by extending attention-based pooling[15,25] to multiple tasks, based on which the classification layers of the network outputs the final slide-level predictions. Specifically, two stacked fully connected layers Fc1 and Fc2, parameterized by $W_1 \in \mathbb{R}^{512\times 1,024}$, $\mathbf{b}_1 \in \mathbb{R}^{512}$ and $W_2 \in \mathbb{R}^{512\times 512}$, $\mathbf{b}_2 \in \mathbb{R}^{512}$ in the base of the network, each followed by rectified linear unit (ReLU) activation, allow the model to learn histology-specific feature representations by tuning deep features extracted through transfer learning, mapping the set of patch feature embeddings $\{\mathbf{z}_k\}$, $\mathbf{z}_k \in \mathbb{R}^{1,024}$ in a given WSI to 512-dimensional vectors:

$$\mathbf{h}_k = \mathrm{ReLU}(W_2(\mathrm{ReLU}(W_1\mathbf{z}_k + \mathbf{b}_1)) + \mathbf{b}_2). \quad (1)$$

Multitask attention pooling. In the proposed multitask learning framework, the multilayered attention module consists of layers Attn-Fc1 and Attn-Fc2 with weight parameters $V_a \in \mathbb{R}^{384\times 512}$ and $U_a \in \mathbb{R}^{384\times 512}$ (shared across all tasks), and one independent set of weights $W_{a,t} \in \mathbb{R}^{1\times 384}$ for each task t. This network module is trained to assign an attention score $a_{k,t}$ (equation (2)) to each patch, for which—after softmax activation—a high score (near 1) indicates that a region is highly informative for determining the slide-level classification task and a low score (near 0) indicates that the region has no diagnostic value (for simplicity the bias parameters are not shown in the equation; $\odot$, element-wise product; sigm, sigmoid activation function; N, total number of patch embeddings in a particular slide):

$$a_{k,t} = \frac{\exp\{W_{a,t}(\tanh(V_a\mathbf{h}_k)\odot\mathrm{sigm}(U_a\mathbf{h}_k))\}}{\sum_{j=1}^{N}\exp\{W_{a,t}(\tanh(V_a\mathbf{h}_j)\odot\mathrm{sigm}(U_a\mathbf{h}_j))\}}. \quad (2)$$

Attention pooling then simply averages the feature representations $\{\mathbf{h}_k\}$ of all patches in the slide, weighted by their respective predicted attention scores $\{a_{k,t}\}$ and the resulting feature vector $\mathbf{h}_{\mathrm{slide},t} \in \mathbb{R}^{512}$ is treated as the histology deep features representing the entire slide for task t. This intuitive, trainable aggregation function allowed the network to learn to automatically identify the subset of informative regions in the slide to predict the primary tumour without requiring detailed annotations that outlined the precise regions of tumour.

Late-stage fusion and classification. We adopt a simple fusion mechanism to incorporate the biological sex of each patient into the prediction of the model by treating the sex s as an additional covariate encoded by binary values, and concatenating it to the deep features extracted from the histology slide. The concatenation results in a 513-dimensional feature vector that is fed into the final classification (cls) layer $W_{\mathrm{cls},t}$ (with bias parameters $\mathbf{b}_{\mathrm{cls},t}$) for task t to obtain the slide-level probability prediction scores (softmax, softmax activation; concat, concatenation):

$$\mathbf{p}_t = \mathrm{softmax}(W_{\mathrm{cls},t}\mathrm{concat}([\mathbf{h}_{\mathrm{slide},t}, s]) + \mathbf{b}_{\mathrm{cls},t}). \quad (3)$$

In our study, the first task of predicting the origin site of tumour is a 18-class classification problem and the second task of predicting whether a tumour is primary or metastatic is a binary problem. Accordingly, the task-specific classification layers are parameterized by $W_{\mathrm{cls},1} \in \mathbb{R}^{18\times 513}$ and $W_{\mathrm{cls},2} \in \mathbb{R}^{2\times 513}$, respectively.

Training details. We randomly sampled slides using a mini-batch size of 1 WSI and used a multitask objective to supervise the neural network during training. The sampling frequency of each slide is set based on the inverse relative proportion of metastatic to primary slides in the training set. For each slide, the total loss is a weighted sum of loss incurred from the first task of predicting the tumour origin and the loss from the second task of predicting primary versus metastasis:

$$\mathcal{L}_{\mathrm{total}} = c_1\mathcal{L}_{\mathrm{cls},1} + c_2\mathcal{L}_{\mathrm{cls},2}. \quad (4)$$

The standard cross-entropy loss function was used for both tasks and to give higher importance to the main task of tumour origin prediction, we used $c_1 = 0.75$ and $c_2 = 0.25$. Values of $c_1 \in \{0.5, 0.25\}$ were also tested (without loss of generality, we let $c_2 = 1 - c_1$), with $c_1 = 0.75$ resulting in the highest validation accuracy on origin prediction whereas the validation AUC ROC on the binary task of distinguishing primary versus metastatic tumours was largely insensitive to the hyperparameter choice of c_2 (Supplementary Fig. 5). After each mini-batch, the model parameters are updated via the Adam optimizer with an ℓ_2 weight decay of 1×10^{-5} and a learning rate of 2×10^{-4}. $\beta_1 = 0.9$ and $\beta_2 = 0.999$ (default coefficient values) were used for computing the running averages of the first and second moment of the gradient and by default, the epsilon term (added to the denominator for numerical stability) was set to 1×10^{-8}. To curb the model from potential overfitting, we also used dropout layers with $P = 0.25$ after every hidden layer. Learning curves for training and validation for different model configurations are shown in Supplementary Fig. 6.

Model selection. During training, the performance of the model on the validation set was monitored each epoch. Beyond epoch 50, if the validation loss on the tumour origin prediction task had not decreased for 20 consecutive epochs, early stopping was triggered and the best model with the lowest validation loss was used for reporting the performance on the held-out test set.

Evaluation

Primary versus metastatic tumours. For the binary classification task of distinguishing primary versus metastatic tumours, without loss of generality, metastatic tumours are defined as the 'positive' class for computing the sensitivity and specificity, and the ROC curve and its associated AUC. The operating point was adjusted on the basis of the Youden's J statistic[33]. We additionally assessed the performance on this binary task for a number of tissue sites in which both primary and metastatic tumours are frequently found (sites with at least 10 metastatic and 10 primary tumours in the test set are considered in the analysis; sites along the female reproductive tract, including the ovary, uterus and cervix were grouped into 'Müllerian'), as well as for tumours of the same primary origin (Extended Data Fig. 5). Nonparametric bootstrapping with 1,000 samples was used to compute 95% confidence intervals.

Origin prediction. For the multiclass classification task of origin prediction, the prediction of the network is the argmax of the class probabilities predicted by the model (that is, the class with the highest predicted probability score). We assessed the prediction of the model using a wide range of classification metrics including the one-versus-rest precision, recall, F_1-score, mean average precision and AUC ROC calculated for each primary origin and for an aggregation across all classes through micro-averaging, macro-averaging and weighted averaging (Supplementary Tables 6, 7). Additionally, we reported the top-k accuracy of the model for $k \in 1, 3, 5$, which measures how often the ground truth label is found in the k highest confidence predictions of the model, as well as Cohen's κ, which measures interobserver agreement while taking into account agreement by chance. For the assessment of model performance on cases of CUP, oesophagogastric and colorectal origins are grouped under gastrointestinal origin and similarly the ovary, uterus and cervix were grouped under the female reproductive tract as Müllerian, to be consistent with the convention used by the primary differential

assigned. Owing to uncertainties associated with the ground truth labels for cases of CUP, we use the term top-*k* agreement instead of accuracy to denote the concordance between the model and the assigned differential diagnosis. Nonparametric bootstrapping with 1,000 samples was used to compute 95% confidence intervals.

Additional experiments and analysis

Classification of adenocarcinoma and squamous cell carcinoma. Often pathologists can readily distinguish between adenocarcinoma and squamous cell carcinoma based on the morphological and architectural appearance of the tumour cells that are present in the tissue. However, within the respective family of adenocarcinoma and squamous cell carcinoma subtypes, determining the origin of the tumour can remain a challenging task. Therefore, to analyse the feasibility of using weakly supervised deep learning to distinguish between origins when a sample is known to be an adenocarcinoma (which represent 50% of cases of CUP[1]) or squamous cell carcinoma, we trained two additional networks. The adenocarcinoma model was developed and validated using a subset of 14,653 adenocarcinoma WSIs that fall under 6 of the 18 tumour origin classes considered by the main network: lung (3,737), breast (3,605), colorectal (2,979), pancreatobiliary (1,698), oesophagogastric (1,380) and prostate (1,254). Similarly, the squamous cell carcinoma network was developed using a subset of 3,093 squamous cell carcinoma WSIs from 4 origins: lung (1,334), head and neck (1,284), cervix (286), and oesophagogastric (189). For all experiments, the cases were partitioned into 70:10:20 splits for training:validation:testing. The model architecture (except for the number of classes, which is adjusted accordingly from the original 18 classes), learning schedule and hyperparameters used were the same as for the main network. These networks can act as additional readers in addition to our main 18-class TOAD network. Results for this analysis are presented in Extended Data Fig. 10.

Classification of tumours metastasized to the liver and lymph nodes. One additional way to pose the origin prediction problem is to train individual networks for common metastatic sites. To explore this further, we trained two networks by grouping common origins that metastasize to lymph nodes and the liver. The lymph-node-specific model was developed using a subset of 1,649 WSIs of metastatic tumours from seven primary origins including: lung (402), skin (275), head and neck (267), breast (262), thyroid (233), oesophagogastric (113) and bladder (97). The liver-specific network was developed using a subset of 1,137 WSIs of metastatic tumours from four primary origins including: colorectal (380), pancreatobiliary (379), breast (247) and lung (131). For all experiments, the cases were partitioned into 70:10:20 splits for training:validation:testing. The model architecture (in addition to the number of output classes), learning schedule and hyperparameters used were the same as for the main network except the multitask attention branch for predicting primary versus metastatic was disabled as all cases were metastatic. Results for this analysis are presented in Extended Data Fig. 10.

Analysing IHC usage for cases of metastasis. We used the number of IHC stains performed during diagnosis as an indirect measure of analysing difficult-to-diagnose cases. This is in line with a previous study that established correlations between the number of IHC analyses used and case difficulty[26]. To analyse the performance of the model for cases of metastases that required different levels of IHC usage for diagnosis, we used pathology reports from the in-house database to extract the number of IHC stains used for each case. For the purpose of the analysis, we excluded non-diagnostic stains from the IHC counts, such as ER, PR, HER2, ALK, ROS and PD-L1). We were able to extract the number of diagnostic IHC tests performed in 1,286 reports (mean, 2.1; minimum, 0; maximum, 21). Note that pathologists often rely on clinical correlation and the knowledge of the known primary tumour of a patient to confirm or guide their diagnosis for metastatic tumours, which the deep learning model does not have access to. It is therefore reasonable to assume that the number of IHC analyses required will be higher if pathologists are blinded to clinical variables. Results for this analysis are presented in Extended Data Fig. 6.

Assessing model performance for cases of CUP with multiple assigned diagnoses. Owing to the difficulty of assessing cases of CUP, pathologists sometimes assign multiple diagnoses in the pathology report. As these diagnoses were not necessarily arranged in terms of certainty or physician confidence, to assess model agreement with all differential diagnoses assigned (without considering their ordering), for a case with *k* differential diagnoses, we computed its intersection with the set of top-*k* predictions of the model.

Ablation studies

We conducted ablation experiments to study the effect of: (1) including sex; (2) including tissue sampling site; (3) multitask learning; and (4) including cases of primary tumours in the study. To assess the benefit of multitask learning and using the sex of the patient as an input in addition to histology slides for the task of origin prediction, we evaluated the performance of the model trained using histology slides as the only input, and the model trained using both histology slides and the sex of the patient as input (no multitask learning). Results show that using the sex of the patient as input and multitask learning leads to a substantial improvement in the top-1 accuracy of the model especially on metastatic tumours (Extended Data Fig. 4a, b and Supplementary Fig. 1). Next, we assessed the value of training on primary tumour slides by removing them from the training set and instead training the model on only metastatic tumour slides (multitask training is disabled when training on only metastatic tumours). We observed a substantial decrease in performance testing on metastatic tumours when the model is trained on only metastatic tumours, which suggests that the ability of the model to recognize metastatic tumours indeed benefits from learning the morphology of primary tumours (Extended Data Fig. 4c). Lastly, we additionally experimented with providing the tissue sampling or biopsy site to the model as an additional input. Biopsy sites with a frequency of fewer than 100 slides in the dataset are grouped together into 'other'. During training and inference, the biopsy site is one-hot encoded and concatenated with attention pooled feature vector and patient sex. Using the biopsy site and training on metastatic tumours slightly outperforms the model trained on metastatic tumours without using biopsy site. However, when trained on both primary and metastatic tumours, the model that does not use biopsy site achieves a much higher top-1 accuracy than the model that uses biopsy site as an additional input. This is not surprising as the biopsy site can provide a direct shortcut to the ground truth label for primary tumours and therefore the model no longer has to rely on learning from the morphology of the primary tumours, which has been shown to be beneficial for the ability of the model to recognize metastatic tumours. Therefore, we did not include the tissue sampling site as an additional input into our main network. All results of the ablation experiments are shown in Extended Data Fig. 4. All models use the same hyperparameters and learning schedule as what is reported throughout the study and are trained and tested using the same training:validation:test splits.

Computational hardware and software

We processed all WSIs on Intel Xeon W-2123 multi-core CPUs (central processing units) and a total of 16 NVIDIA 2080Ti GPUs (graphics processing units) using our custom, publicly available CLAM package[15] whole-slide processing pipeline implemented in Python (version 3.7.7). CLAM uses OpenSlide (version 4.3.1) and openslide-python (version 1.1.1) for reading WSIs and pillow (version 7.0.0) and opencv-python (version 4.1.1) for image processing. Each deep learning model was trained on multiple GPUs using the Pytorch deep learning library

(version 1.5.1). Unless otherwise specified, plots were generated in Python using matplotlib (version 3.1.1) and numpy (version 1.18.1) was used for vectorized numerical computation. The geographical diversity maps were generated using additional Python packages including pyshp (version 2.1.0), basemap (version 1.1.0) and geopy version (version 1.22.0). Other Python libraries used to support data analysis include pandas (version 1.0.3), scipy (version 1.3.1), tensorflow (version 1.14.0), tensorboardX (version 1.9), torchvision (version 0.6), timm (version 0.3.4) and ptflops (version 0.6.4). The scientific computing library scikit-learn (version 0.22.1) was used to compute various classification metrics and estimate the AUC ROC. The confusion matrix plot was plotted in R (version 4.0.2) using ComplexHeatmap (version 2.4.3). The interactive demo website was developed using OpenSeadragon (version 2.4.2) and jQuery (version 3.6.0). Using a single 2080Ti GPU, the average run times per slide in seconds for 100 slides randomly sampled from the test set are 17.1 s (s.d., 11.17) and 171.6 s (s.d., 108) for inference (includes tissue segmentation, patching, feature extraction and prediction) and heat map generation (computed for non-overlapping patches and displayed at 10×), respectively.

WSI processing

Segmentation. Tissue segmentation of WSIs was performed automatically using the CLAM[15] library at a downsampled magnification of each slide. A binary mask for the tissue regions was computed by applying binary thresholding to the saturation channel of the image downsample after conversion from RGB to the HSV colour space. Median blurring and morphological closing were also performed to smooth the detected tissue contours and suppress artefacts such as small gaps and holes. The approximate contours of the detected tissue as well as tissue cavities were then filtered based on their area to produce the final segmentation mask.

Patching. We exhaustively cropped segmented tissue contours into 256 × 256 patches (without overlap) at 20× magnification (if the 20× downsample is not found in the image pyramid, 512 × 512 patches were instead cropped from the 40× downsample and downscaled to 256 × 256). We refer to the collective set of all patches (known as instances) extracted from a particular WSI as a bag.

Feature extraction. Given the enormous bag sizes (number of patches in each WSI) in our dataset, we first used a convolutional neural network based on the ResNet50 architecture to encode each patch into a compact low-dimensional feature vector. Specifically, a ResNet50 model pretrained on Imagenet[32] was truncated after the third residual block and was followed by an adaptive mean-spatial pooling layer to reduce the spatial feature map obtained from each 256 × 256 × 3 RGB image patch into a descriptive, one-dimensional feature representation of length of 1,024. The choice of this particular feature encoder is described in the next section. To perform the feature extraction step efficiently, we used up to 16 GPUs in parallel with a batch-size of 128 per GPU.

Choice of feature encoder. The motivation behind the choice of using a truncated ResNet50 model as the feature encoder is that it strikes the balance between the compactness of the smaller ResNet34 256-dimensional feature space, which risks insufficient expressive capabilities, and the more-expressive 2,048-dimensional feature space of the full ResNet50, which would lead to much higher cost during data processing and model training. In an attempt to assess the trade-off between using the 1,024-dimensional features from the third residual block and the full 2,048-dimensional features from the full ResNet50 model, we performed bench-marking as well as evaluation using the binary classification problem of distinguishing primary lung adenocarcinoma and lung squamous cell carcinoma using the available TCGA data, which is a common problem analysed in recent computational pathology literature[15,22]. Although arguably limited in scope, our results (Supplementary Table 14) suggest that despite enabling faster data processing and training and requiring lower disk storage space, the use of 1,024-dimensional features from an earlier convolution layer of the pretrained ResNet50 does not harm the predictive performance of the downstream classification model. Notably, feature extraction with the full ResNet50 is expected to take more than 12% longer because of the increased number of compute operations in the forward pass, as well as the increased time required to serialize the 2,048-dimensional feature vectors to a disk (which also substantially increases the storage space required to perform a study at scale). Additionally, although the classification network, which is developed on top of extracted features, is relatively compact and the increase in real compute time for performing forward/backward pass in such a network is small, we note that the bottleneck of training/inference currently lies with deserialization of the saved feature vectors from disk to fast CPU/GPU memory, which takes around 80% longer for the 2,048-dimensional features. Therefore, using 2,048-dimensional features will markedly increase the training time in practice. Empirically, training the lung adenocarcinoma versus lung squamous cell carcinoma classifier slides took 43 s per epoch with 1,024-dimensional features, and nearly twice as long (82 s per epoch) with 2,048-dimensional features. Therefore, unless we can achieve considerably faster deserialization, using 2,048-dimensional features will substantially increase the time to perform our model training and evaluation, especially for large-scale studies. We found that the increased feature dimension does not benefit classification performance in the context of lung adenocarcinoma versus lung squamous cell carcinoma. This phenomenon can likely be explained by the observation that later layers of a deep neural network learn filters that are increasingly specific to features that are relevant to the source task and data (in this case, natural image classification with ImageNet), whereas features from earlier layers are more general and applicable to different datasets[34]. Therefore, because of the vast difference between the domain of histopathology and natural images, the finding that using penultimate layer features of ResNet50 does not improve performance over earlier features is not necessarily unexpected.

Model interpretability

Interpreting model prediction using attention heat maps. To visually interpret the importance of each region in a WSI to classification predictions of the model, we first computed the reference distribution of attention scores by tiling the WSI into 256 × 256 patches without overlap and computing the attention score for each patch for the task of primary origin prediction. To generate more fine-grained heat maps, we subsequently repeated the tiling but with an overlap of up to 95% and converted the attention scores computed from overlapping crops to normalized percentile scores between 0.0 (low attention) and 1.0 (high attention) based on the initial reference distribution. The normalized scores were then registered onto the original WSI corresponding to the spatial location of each patch and scores in overlapped regions were averaged. Finally, a colour map was applied to the attention scores and the heat map was displayed as an overlay layer with a transparency value of 0.5. These attention maps are shown in Extended Data Figs. 7, 8 and Supplementary Fig. 4 and can also be visualized in our interactive demo website (http://toad.mahmoodlab.org).

Quantitative analysis of high attention regions. We further analysed the attention of the model by quantifying the cell populations that are localized within the high-attention regions proposed by the model for all metastatic tumours in the test set (n = 1,408). Specifically, the top-10 high-attention patches from each slide were extracted at the 20× equivalent magnification and a HoverNet[35] model was trained for multi-organ nucleus segmentation and classification was used to identify different cellular populations including tumour cells, lymphocytes, connective tissue, dead cells and non-neoplastic epithelial cells. The relative fraction of each cell type is plotted using box plots (Extended Data Fig. 9).

Reporting summary

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

Data availability

The TCGA diagnostic whole-slide data and corresponding labels are available from NIH genomic data commons (https://portal.gdc.cancer.gov/). The CPTAC histology data and corresponding labels are available from the TCIA CPTAC Pathology Portal (https://cancerimagingarchive.net/datascope/cptac/). Processed data that are included in the figures presented in the paper are available as source data. Restrictions apply to the availability of the raw in-house and external data, which were used with institutional permission through IRB approval for the current study, and are thus not publicly available. Please email all requests for academic use of raw and processed data to the corresponding author (and also include M.Y.L. (mlu16@bwh.harvard.edu)). All requests will be evaluated based on institutional and departmental policies to determine whether the data requested is subject to intellectual property or patient privacy obligations. Data can only be shared for non-commercial academic purposes and will require a formal material transfer agreement. Source data are provided with this paper.

Code availability

All code was implemented in Python using PyTorch as the primary deep learning package. All code and scripts to reproduce the experiments of this paper are available at https://github.com/mahmoodlab/TOAD.

31. He, K. et al. Deep residual learning for image recognition. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition* 770–778 (IEEE, 2016).
32. Russakovsky, O. et al. Imagenet large scale visual recognition challenge. *Int. J. Comput. Vis.* **115**, 211–252 (2015).
33. Youden, W. J. Index for rating diagnostic tests. *Cancer* **3**, 32–35 (1950).
34. Yosinski, J., Clune, J. Bengio, Y. & Lipson, H. How transferable are features in deep neural networks? In *Proc. 27th International Conference on Neural Information Processing Systems* Vol. 2, 3320–3328 (2014).
35. Graham, S. et al. Hover-net: simultaneous segmentation and classification of nuclei in multi-tissue histology images. *Med. Image Anal.* **58**, 101563 (2019).

Acknowledgements We thank A. Bruce for scanning internal cohorts of histology slides of patients at BWH; J. Wang, M. Barbieri, K. Bronstein, L. Cirelli and E. Askeland for querying the BWH slide database and retrieving archival slides; C. Li for assistance with EMRs and Research Patient Data Registry (RPDR); M. Bragg, T. Mellen, T. A. Mages and S. Zimmet for administrative support; Z. Noor for developing the interactive demo website; and K. Tung of Boston Children's Hospital for anatomical illustrations. This work was supported in part by internal funds from BWH Pathology, NIH NIGMS R35GM138216 (F.M.), Google Cloud Research Grant and Nvidia GPU Grant Program. M.S. was additionally supported by the NIH Biomedical Informatics and Data Science Research Training Program, NIH NLM T15LM007092. The content is solely the responsibility of the authors and does not reflect the official views of the National Institute of Health, National Institute of General Medical Sciences or the National Library of Medicine.

Author contributions F.M. and M.Y.L. conceived the study and designed the experiments. M.Y.L. performed the experimental analysis. D.F.K.W., T.Y.C. and M.Z. curated the in-house, external and CUP datasets. M.Y.L., F.M., D.F.K.W., T.Y.C. and M.S. analysed the results. M.Y.L., M.S., J.L. and F.M. developed the data-visualization tools. M.Y.L. and F.M. prepared the manuscript with input from all co-authors. F.M. supervised the research.

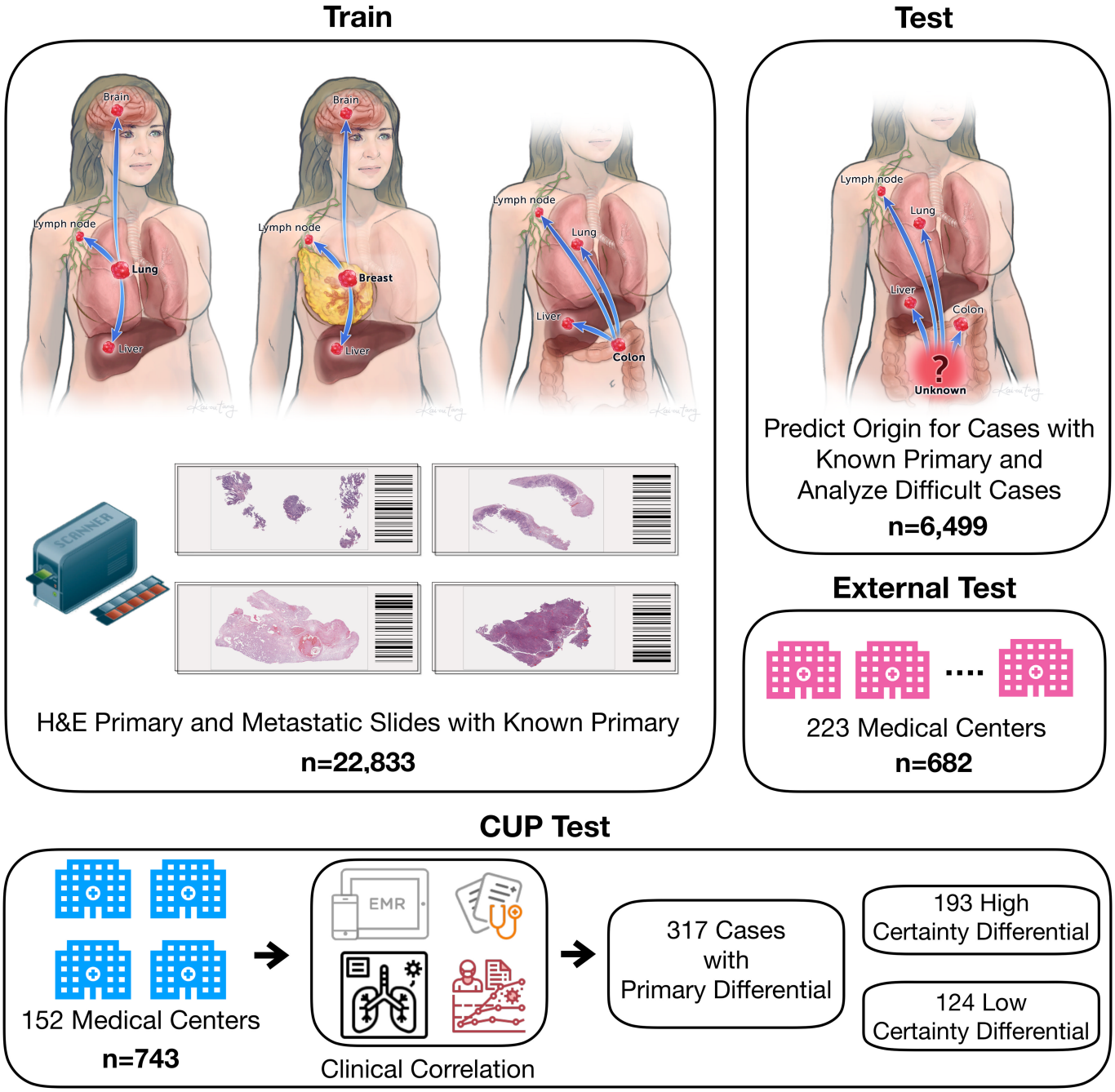
Competing interests The authors declare no competing interests.

Additional information

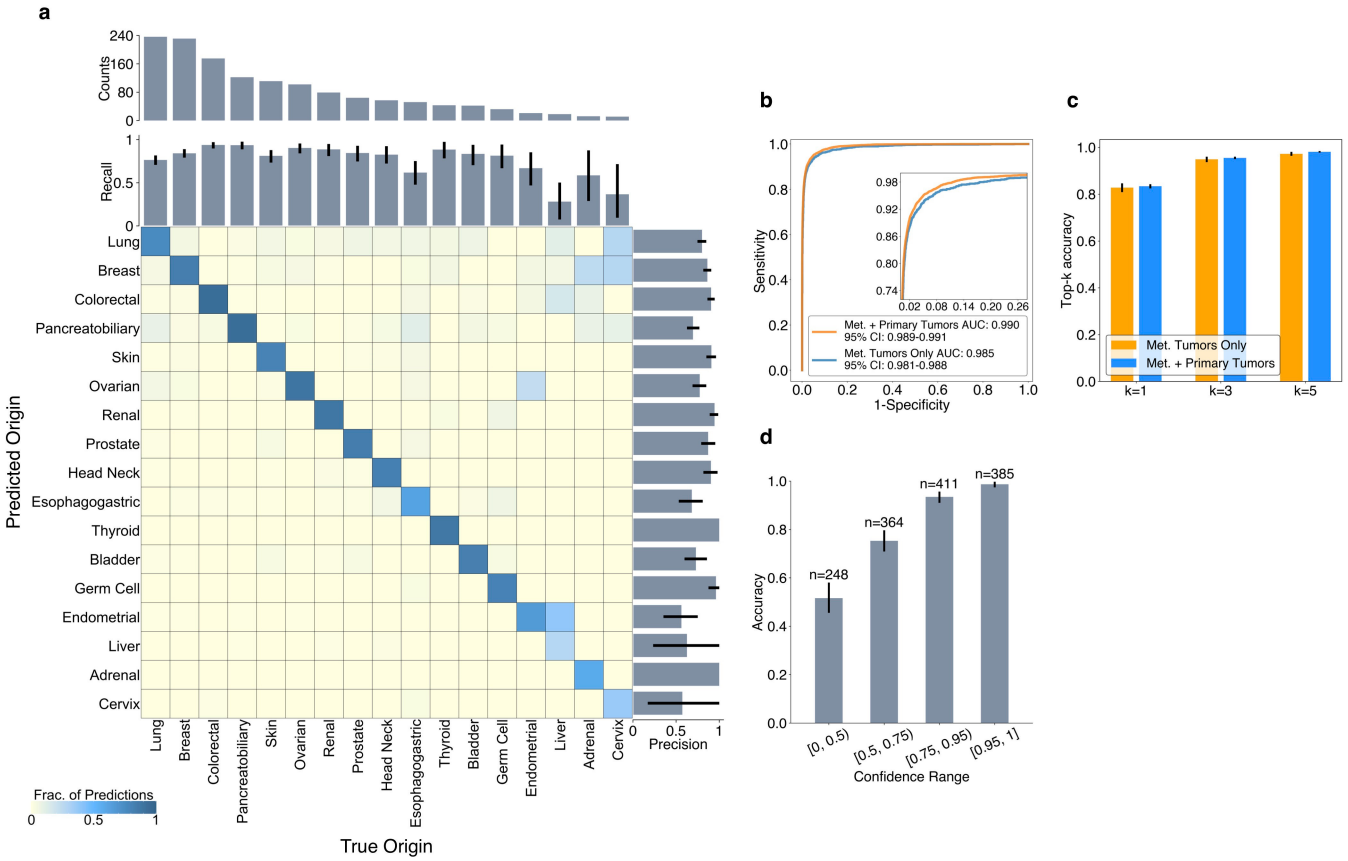
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41586-021-03512-4.

Correspondence and requests for materials should be addressed to F.M.

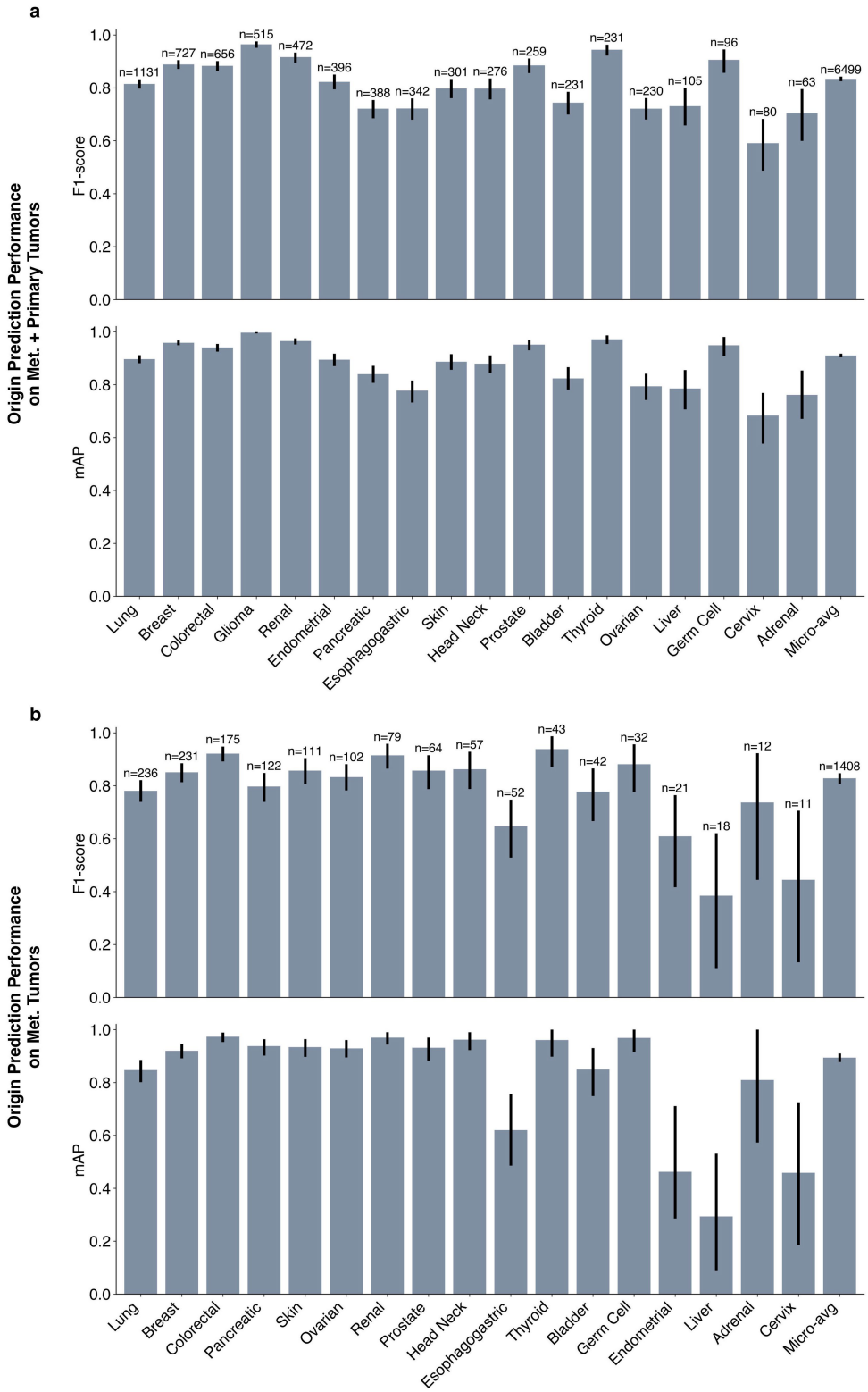
Peer review information *Nature* thanks Beatrice Knudsen and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.



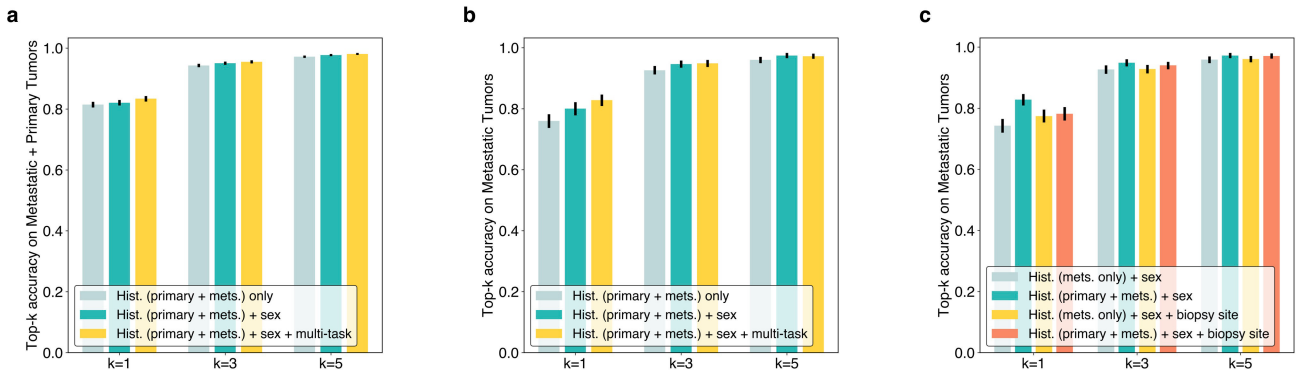
Extended Data Fig. 1 | Overall study design. The model was first trained and tested on tumours of known primary origins. For model development and testing, we collected, in total, 32,537 H&E digitized diagnostic slides (from 29,107 patients) with confirmed diagnosis and randomly sampled 70% of cases (22,833 slides) to train the model and 20% of cases (6,499 slides) were held-out for evaluation. The remaining 10% of cases (3,205 slides) was used for validation during training to select the best performing model. To further assess the ability of the model to generalize on data from sources and staining protocols that it did not encounter during training, we also evaluated the model on an external test cohort of 682 cases, submitted from more than 200 US and international medical centres. The model was then assesed on increasingly difficult cases of metastatic tumours. Lastly, to assess the ability of the model to inform meaningful predictions for origins of cancers that cannot be readily diagnosed by human experts using H&E histology alone, we curated an additional diverse dataset of 743 cases of CUP sourced from institutions across the country and outside the USA. Although the primary cancer could not be initially assigned for all of these cases based on H&E histology alone, using EMRs and evidence from clinical and ancillary tests, we identified a subset of 317 cases for which a primary differential was eventually assigned over the course of the patient's history (see Methods). We validated our model against the recorded primary differential for agreement, showcasing the applicability of the model to cases without clear morphological indication for a particular primary cancer.



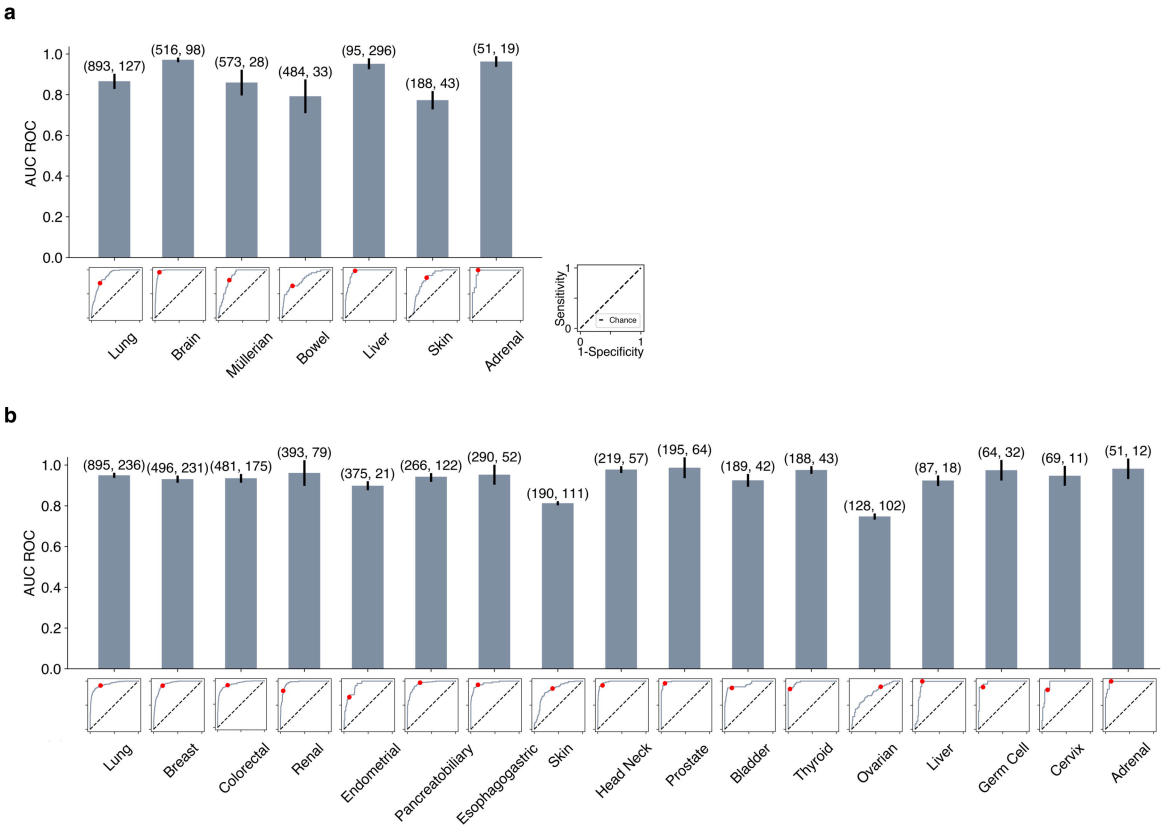
Extended Data Fig. 2 | Classification performance for the prediction of cancer origins on metastatic tumours. a, The confusion matrix, along with the precision and recall of each class and its count is plotted for metastatic tumours in the test set ($n = 1{,}408$). Glioma was excluded as there were no metastatic glioma specimens in the test set and it was verified that no case of metastasis was predicted as glioma by the model. **b**. The micro-averaged, one-versus-rest AUC ROC. **c**, Top-k accuracies of the model on only metastatic tumours ($n = 1{,}408$), and on the combined set of metastatic and primary tumours ($n = 6{,}499$). **d**, Accuracy of the model on metastatic tumours binned into different levels of prediction confidence. **a**, **c**, **d**, Error bars indicate 95% confidence intervals, the centre is always the computed value of each classification performance metric (specified by its respective axis labels).



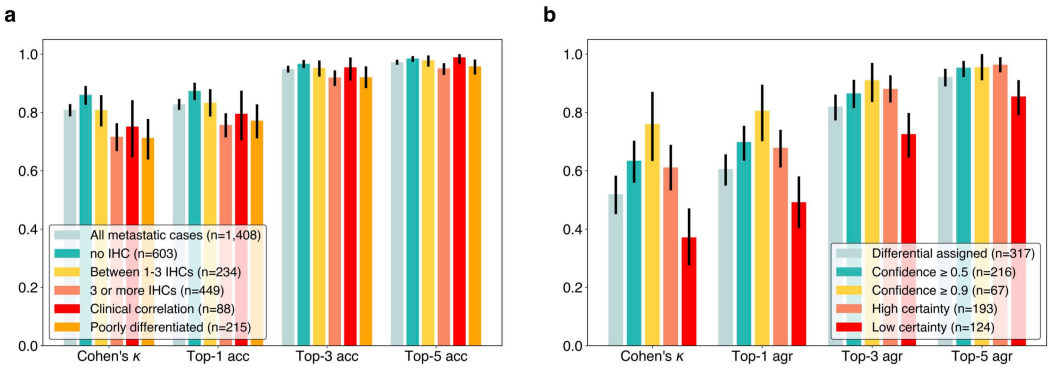
Extended Data Fig. 3 | Performance for the prediction of cancer origins on metastatic and primary tumours. a, **b**, Additional metrics including per-class and micro-averaged F_1-score and mean average precision score are computed for the combined set of primary and metastatic tumours (**a**; $n = 6,499$) and only metastatic tumours (**b**; $n = 1,408$) in the test set. **a**, **b**, Error bars indicate 95% confidence intervals, plotted around the computed value of each classification performance metric (specified by its respective axis labels). Note that the micro-averaged F_1-score is the same as the overall accuracy. See Supplementary Table 3 for the number of metastatic and primary tumours for each origin in the test set.



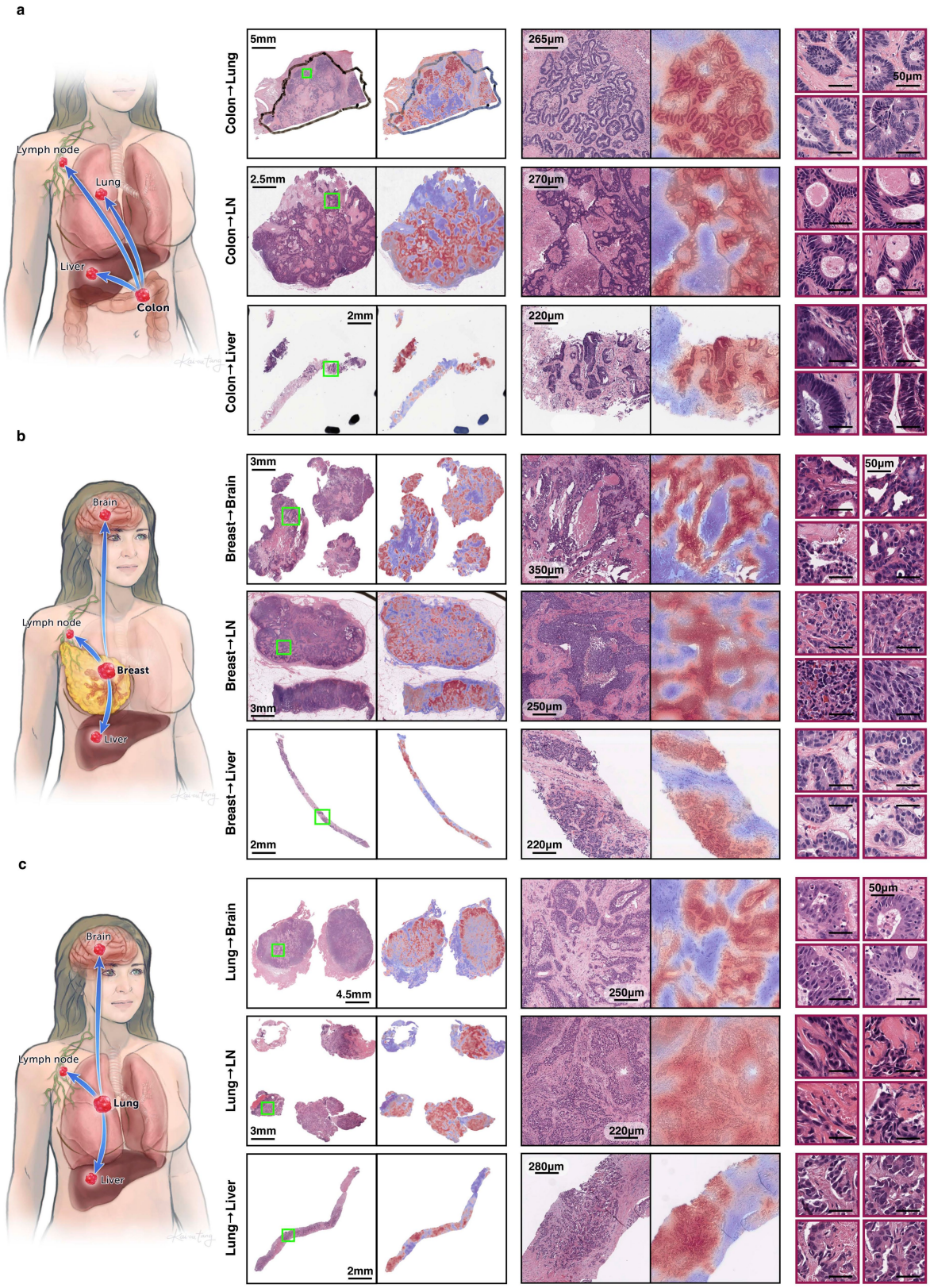
Extended Data Fig. 4 | Ablation studies. a, **b**, Ablation experiments were performed to assess the benefit of multitask learning and including patient sex as an input in addition to histology on the performance for the prediction of cancer origins (see Methods, 'Ablation studies'). Top-*k* accuracies for testing on both primary and metastatic tumours (**a**; $n = 6{,}499$) in the held-out test set and testing on only metastatic tumours (**b**; $n = 1{,}408$). The multitask model with access to patient sex scored nearly 2.0% higher in top-1 accuracy compared to the baseline, single-task model using histology only when testing on the entire test set, and is 6.8% higher when testing on only the metastatic tumours. **c**, Additional experiments are performed to assess the importance of including primary tumour slides during training and the effect of adding the tissue sampling or biopsy site as another input covariate (in addition to sex) on model performance on metastatic tumours ($n = 1{,}408$). The accuracy of the model decreased by 8.5% when trained on only metastatic tumours in the training set, showing that the ability of the model to recognize metastatic tumours benefits substantially from also learning from primary tumours. We additionally experimented with providing the tissue sampling or biopsy site to the model. Multitask training is used when training on both primary and metastatic tumours. A decrease of 4.6% in model accuracy is observed when the biopsy site information is incorporated. This is probably because the biopsy site can provide a direct shortcut to the ground truth label for primary tumour slides and therefore discourages the model from learning from the morphology of primary tumours, which we have found to be beneficial for the ability of the model to recognize metastatic tumours. **a**–**c**, Error bars indicate 95% confidence intervals, plotted around the computed value of each classification performance metric (specified by its respective axis labels).



Extended Data Fig. 5 | Model performance on the binary problem of distinguishing between primary and metastatic tumours. a, Performance for tumours at common metastatic sites. The AUC ROCs (*y* axis) with associated 95% confidence intervals and ROC curves are shown for organ sites (*x* axis) with at least 10 metastatic and 10 primary tumours in the test set. The ovary, uterus and cervix were grouped into upper female reproductive tract ('Müllerian'). The number of primary tumours (first element) and metastatic tumours (second element) at each site are indicated as a tuple above each bar. **b**, Performance for tumours of different primary origins. The AUC ROCs (*y* axis) with associated 95% confidence intervals and ROC curves are shown for tumours from each origin site (*x* axis) except for glioma, for which no metastatic tumours were present in our test set. The number of primary tumours (first element) and metastatic tumours (second element) for each origin are indicated as a tuple above each bar. **a**, **b**, Without the loss of generality, metastatic tumours are designated as the 'positive' class, and primary tumours as the 'negative' class for computing sensitivities and specificities. The operating point of the model is indicated by a red dot on each ROC curve, and is based on maximizing Youden's *J* index.

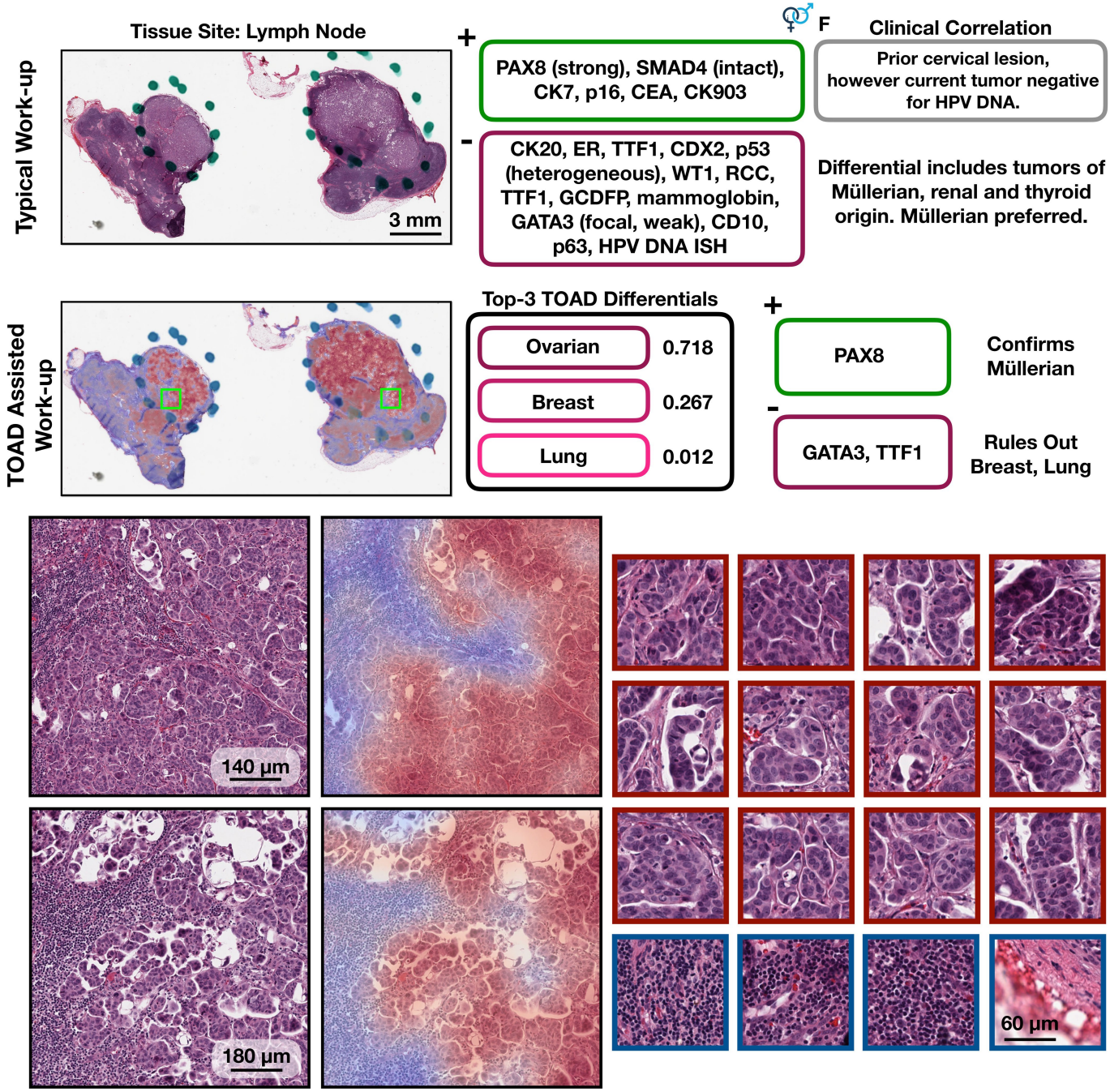


Extended Data Fig. 6 | Model performance on difficult metastatic and unknown primary tumours. a, The performance of the model for the prediction of cancer origins is evaluated in terms of top-*k* accuracies (acc) and Cohen's κ score for patients with metastatic tumours in the held-out test set (n = 1,408). Performance is additionally reported for subsets of patients with metastatic tumours depending on the number of diagnostic IHC stains used, whether recommendation for clinical or radiological correlation was given and whether the tumour was categorized as poorly differentiated. **b**, For the held-out test set of cases of CUP with assigned primary differential diagnosis (n = 317), the model performance is assessed using agreement (agr) with the assigned differential. Performance is additionally reported for high-confidence model predictions (for example, model confidence ≥ 0.5) as well as for cases with a high versus low degree of diagnostic certainty associated with the assigned differential. For cases of CUP, based on the strength of evidence used to support the differential diagnosis and language used in EMRs, we define high-certainty diagnoses as being compatible with morphological evidence or supported by IHC findings or clinical, radiological or molecular correlation, whereas low-certainty diagnoses may not suggest a single specific primary origin or lacked definitive supporting evidence for the assigned primary differential. **a**, **b**, Error bars indicate 95% confidence intervals, plotted around the computed value of each classification performance metric (specified by its respective axis labels).

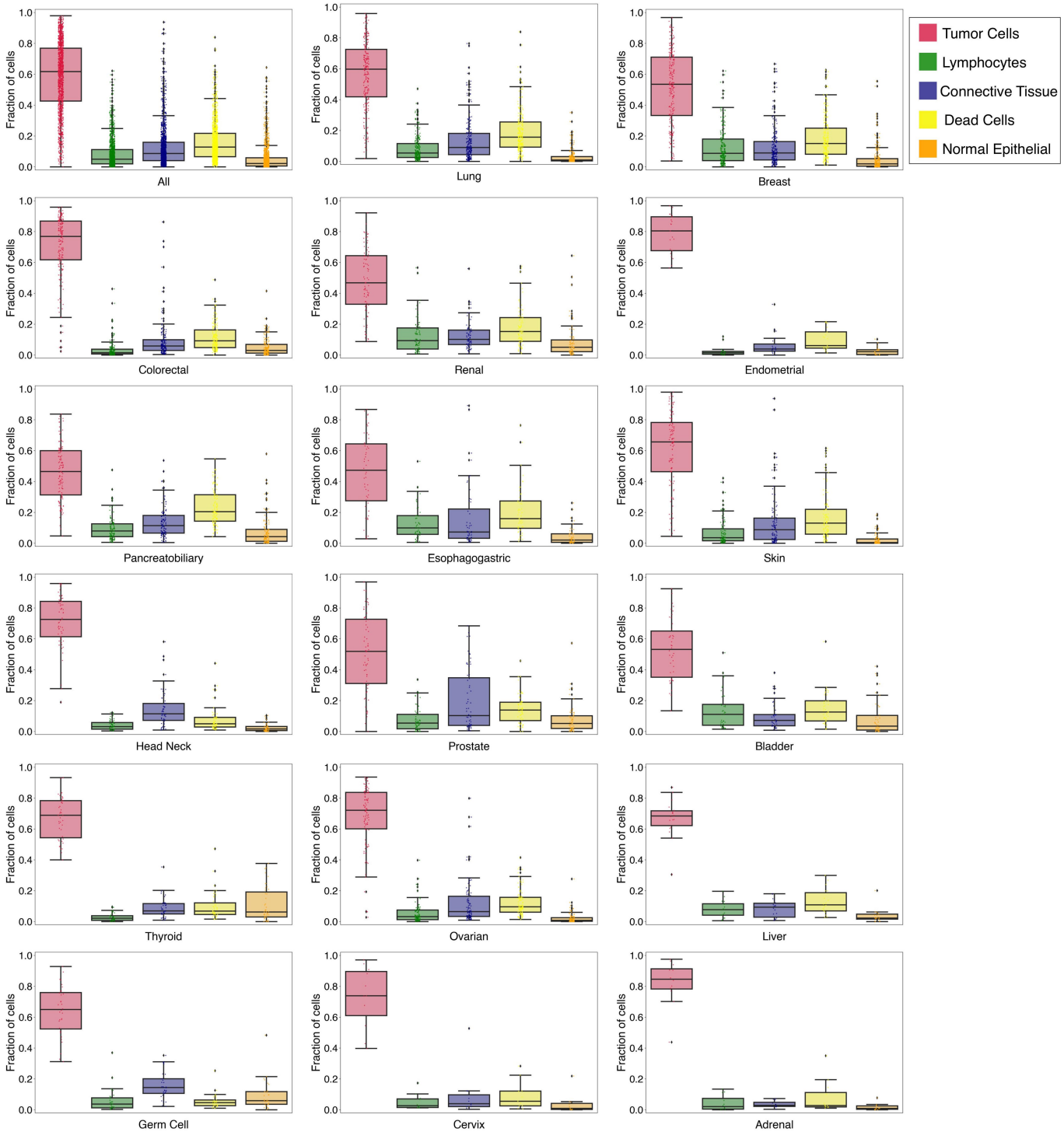


Extended Data Fig. 7 | See next page for caption.

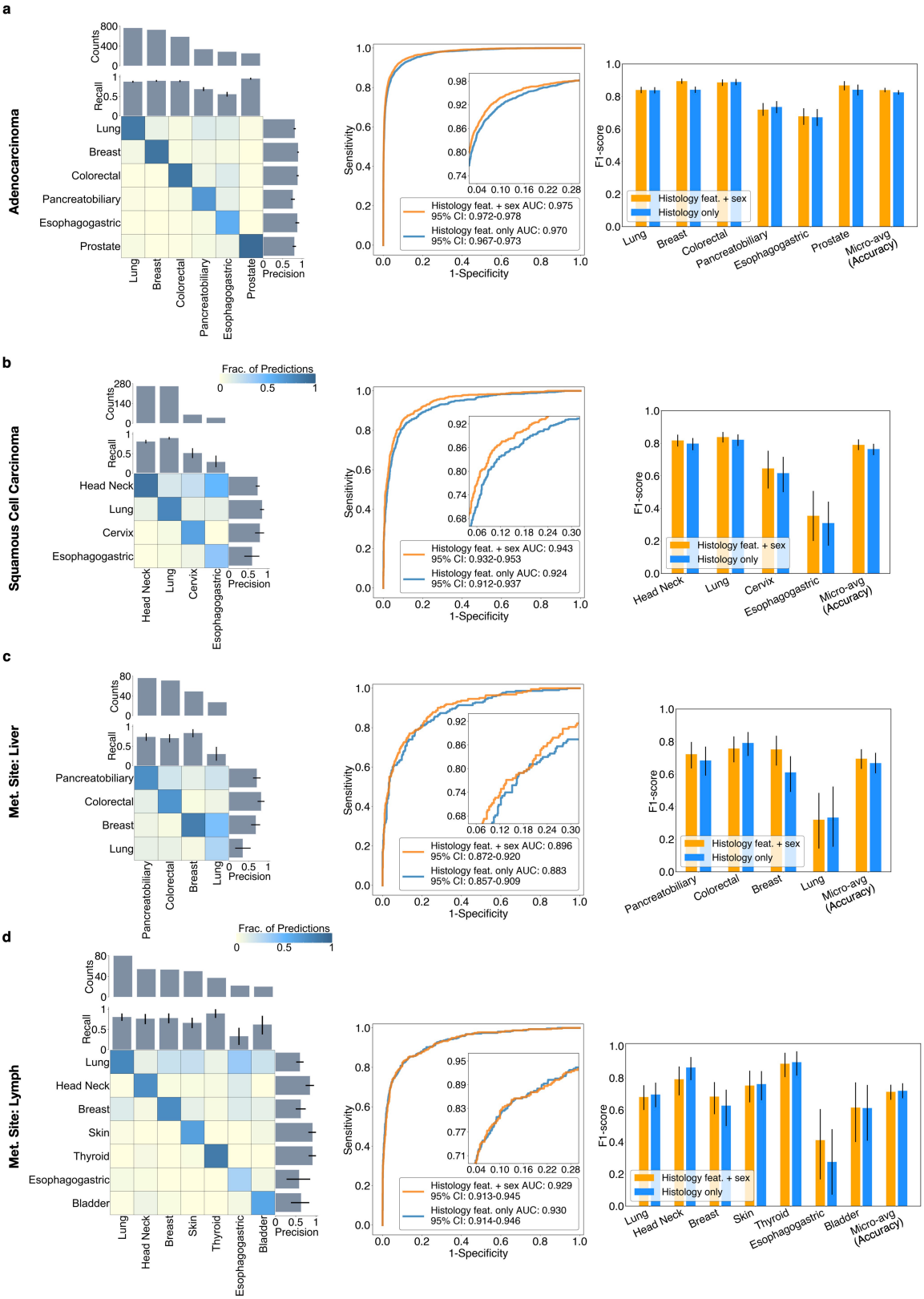
Extended Data Fig. 7 | Examples of metastases from colorectal, breast and lung primary tumours with attention heat maps. a–**c**, Example metastases from colorectal (**a**), breast (**b**) and lung (**c**) primary tumours are shown. For each case, the attention heat map of the model is displayed on top of the original H&E WSI as a semi-transparent overlay in which the overlaid regions range from crimson (high attention, high diagnostic relevance) to navy (low attention, low diagnostic relevance). Left, sites of metastasis are shown, including the lung, lymph node (LN), liver and brain. Right, H&E images show, from left to right, low magnification with corresponding attention map, medium magnification with corresponding attention map, and high-magnification patches. **a**, Medium- and high-magnification views demonstrate so-called 'dirty necrosis' and variably sized glands with densely packed, hyperchromatic nuclei that are characteristic of colorectal adenocarcinoma. **b**, Medium- and high-magnification views demonstrate sheets of cells as well as small tubules and glands—morphologies that are consistent with metastatic breast carcinomas. **c**, Medium- and high-magnification views demonstrate sheets of cells, variably sized glands and cells in infiltrative single files. The cells have large, hyperchromatic nuclei and high nuclear:cytoplasmic ratios, which are consistent with metastatic lung carcinomas. **a**–**c**, The attention heat maps allow the predictions of the model for each case to be visually interpretable for human experts, revealing the morphological features used by the model for the determination of the classification. High-resolution heat maps for cases from all primary sites can be accessed through our interactive demo website (http://toad.mahmoodlab.org).



Extended Data Fig. 8 | TOAD-assisted CUP work-up: example 1. Top, a representative case that underwent a standard CUP work-up involving extensive IHC staining and clinical correlation. Strong PAX8 staining suggested a Müllerian origin and multiple IHC tests were used to rule out other primary tumours. Retrospectively, we analysed the case with TOAD and found that the top-3 determinations were ovarian, breast and lung, and, after this determination, that only three IHC stains (PAX8, GATA3 and TTF1) needed to be used to confirm a Müllerian origin and rule out breast carcinoma and lung adenocarcinoma. This workflow demonstrates how TOAD can be used as an assistive diagnostic tool. Bottom, medium magnification and corresponding heat maps of representative areas of tumour, with high-magnification, high-attention patches on the right outlined in crimson and low-attention patches outlined in navy.



Extended Data Fig. 9 | Analysis of high-attention regions for metastatic tumours. Relative counts of different cell types localized within the high-attention regions proposed by the model were quantified. Specifically, the top-10 high-attention patches from each slide were extracted at the 20× equivalent magnification and a HoverNet[35] model trained for multi-organ nucleus segmentation and classification was used to detect different cellular populations including tumour cells (red), lymphocytes (green), connective tissue (blue), dead cells (yellow) and non-neoplastic epithelial cells (orange). The fraction of cells for each cell type is plotted using box plots for all metastatic slides in the test set (n = 1,408) and is stratified by each primary origin site: lung (n = 236), breast (n = 231), colorectal (n = 175), pancreatobiliary (n = 122), skin (n = 111), ovarian (n = 102), renal (n = 79), prostate (n = 64), head and neck (n = 57), oesophagogastric (n = 52), thyroid (n = 43), bladder (n = 42), germ cell (n = 32), endometrial (n = 21), liver (n = 18), adrenal (n = 12) and cervix (n = 11). Boxes indicate quartile values and whiskers extend to data points within 1.5× the interquartile range. This analysis demonstrates in addition to the attention heat maps, that the model attends strongly to regions of tumour presence for its predictions.



Extended Data Fig. 10 | See next page for caption.

Extended Data Fig. 10 | Classification performance of adenocarcinoma network, squamous cell carcinoma network and site-specific networks for tumour metastasized to the liver and lymph node. a, b, Often pathologists can readily distinguish between adenocarcinoma and squamous cell carcinoma based on the morphological and architectural appearance of the tumour cells that are present in the tissue. However, within the respective family of adenocarcinoma and squamous cell carcinoma subtypes, determining the origin of the tumour can remain a challenging task. Therefore, we hypothesized that we can develop models to specifically predict the origin of tumours for top primary sites of adenocarcinoma (**a**) and squamous cell carcinoma (**b**). Cases from six primary sites (breast, lung, colorectal, pancreatobiliary, prostate and oesophagogastric) and four primary sites (head and neck, lung, cervix and oesophagogastric) were chosen for the development of the adenocarcinoma and squamous cell carcinoma classifiers, respectively, based on their frequency in the database. We also explored the additional scenarios of predicting the primary origins of metastatic tumours grouped by a common metastatic site, including the liver (**c**) and lymph node (**d**). Cases of metastasis from the top-four and top-seven primary origins for liver and lymph nodes, respectively, were chosen on the basis of their frequency in our database. See Methods, 'Additional experiments and analysis' for details. **a**–**d**, Left, the confusion matrix, along with the precision and recall of each class and its count is plotted for the adenocarcinoma model test set (**a**; $n = 2{,}920$) and squamous cell carcinoma model test set (**b**; $n = 621$), the liver metastasis (met.) site test set (**c**; $n = 223$) and lymph node metastasis site test set (**d**; $n = 318$), respectively. Consistent with the model developed using examples of all 18 primary sites, the adenocarcinoma-, squamous-cell-carcinoma- and site-specific models were trained by including the sex of the patient. Performance for models trained with and without the sex of the patient in terms of the micro-averaged, one-versus-rest AUC ROC (middle) and F_1-scores for each primary site and overall model accuracy (micro-averaged F_1-score) (right) are shown. All error bars indicate 95% confidence intervals, plotted around the computed value of each classification performance metric (specified by its respective axis labels).

nature research

Corresponding author(s): Faisal Mahmood

Last updated by author(s): Mar 20, 2021

Reporting Summary

Nature Research 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 Research 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	The code used for this study has been made publicly available here: http://github.com/mahmoodlab/TOAD
Data analysis	Openslide (3.4.1) was used for reading whole slide images. Analysis code was written in Python (3.7.7). These Python libraries were used: h5py (2.10.0), matplotlib (3.1.1), numpy (1.18.1), opencv-python (4.1.1), openslide-python (1.1.1), pandas (1.0.3), pillow (7.0.0), PyTorch (1.5.1), scikit-learn (0.22.1), scipy (1.3.1), tensorflow (1.14.0) , tensorboardX (1.9), torchvision (0.6), pyshp (2.1.0), basemap (1.1.0), geopy (1.22.0), ptflops (0.6.4), and timm (0.3.4). Additionally, R (version 4.0.2) and R library ComplexHeatmap (2.4.3) were used. The interactive demo website was developed using OpenSeadragon (version 2.4.2) and jQuery (version 3.6.0).

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 Research 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 list of figures that have associated raw data
- A description of any restrictions on data availability

The TCGA diagnostic whole slide data and corresponding labels are available from NIH genomic data commons (https://portal.gdc.cancer.gov/). The CPTAC histology data and corresponding labels are available from the TCIA CPTAC Pathology Portal (https://cancerimagingarchive.net/datascope/cptac/). Processed data corresponding figures presented in the paper is available as source data. Restrictions apply to the availability of the raw in-house and external data, which were used with institutional permission via IRB approval for the current study, and thus are not publicly available due to patient privacy obligations. Please email all

requests for academic use of raw and processed data to the corresponding author (and copy to mlu16@bwh.harvard.edu). All requests will be evaluated based on institutional and departmental policies to determine if the data requested is subject to intellectual property or patient privacy obligations. Data can only be shared for non-commercial academic purposes and will require a formal material transfer agreement.

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. We used all available data from public and in house repositories (2004-2020) grouped into 18 predominant primary origins (32,537 whole slide images from 29,107 patients) for model development and held out testing. All available external (682) and CUP cases with assigned differentials (317) were used for additional testing. For rare origins and subtypes the sample size was limited by the number of cases available, for disease models with abundant slides deep learning models were trained while increasing the number of input slides until asymptotic improvement of model performance was achieved to suggest an appropriate sample size was obtained. Further details regarding all datasets are available in the methods and Supplementary tables 2, 3.
Data exclusions	Pre-established exclusion criteria included: 1. Slides that were corrupted or did not have a lower magnification downsample for segmenting, processing the tissue image. 2. Slides that did not have any tumor content. 3. Slides corresponding patients with missing sex information. No other slides were excluded.
Replication	Code for our training and evaluation protocols are publicly available for reproducibility and may be accessed here: http://github.com/mahmoodlab/TOAD. Attempts at replication were successful for all results reported in the study.
Randomization	We randomly split our dataset of 32,537 whole slide images from 29,107 patients into 70% train, 10% validation and 20% test splits, stratified by class, on a patient level i.e. ensuring that the same patient was never represented in multiple splits.
Blinding	Blinding was not necessary because our experiments involved retrospective analysis of digitized histology slides.

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
☐	☒ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

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

Human research participants

Policy information about studies involving human research participants

Population characteristics	Public Training Data: TCGA and CPTAC contain data from a diverse population representing multiple hospitals. In-house BWH Data: Patient demographics are consistent with demographics of all patients who undergo pathology diagnosis at the hospital. Population characteristics including sex and diagnosis are summarized in the Dataset Description section, Supplementary Table 2 presents distribution w.r.t the primary cancer origin, Supplementary Table 3 presents distribution wrt disease subtype. Additionally, covariates corresponding each slide used for model assessment is included in Supplementary Data Tables 9, 10 and 12.

Recruitment	Patients were not directly involved or recruited for the study. This study involved retrospective analysis of pathology slides from patients obtained during standard clinical care.
Ethics oversight	Massachusetts General Brigham IRB committee approved the study. Only retrospective data was used for research, without any active involvement of patients.

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