nature machine intelligence

Article

https://doi.org/10.1038/s42256-024-00868-w

A transformer-based weakly supervised computational pathology method for clinical-grade diagnosis and molecular marker discovery of gliomas

Received: 19 January 2024

Accepted: 17 June 2024

Published online: 18 July 2024

Check for updates

Rui Jiang[1,8], Xiaoxu Yin[1,8], Pengshuai Yang[1,2,8], Lingchao Cheng[3,8], Juan Hu[4], Jiao Yang[5], Ying Wang[3], Xiaodan Fu[3], Li Shang[3], Liling Li[3], Wei Lin[3], Huan Zhou[6], Fufeng Chen[7], Xuegong Zhang[1] ✉, Zhongliang Hu[3,4] ✉ & Hairong Lv[1] ✉

The complex diagnostic criteria for gliomas pose great challenges for making accurate diagnoses with computational pathology methods. There are no in-depth analyses of the accuracy, reliability and auxiliary capability of present approaches from a clinical perspective. Previous studies have overlooked the exploration of molecular and morphological correlations. To overcome these limitations, we propose ROAM, a multiple-instance learning model based on large regions of interest and a pyramid transformer. ROAM enlarges regions of interest to facilitate the consideration of tissue contexts. It utilizes the pyramid transformer to model both intrascale and interscale correlations of morphological features and leverages class-specific multiple-instance learning based on attention to extract slide-level visual representations that can be used to diagnose gliomas. Through comprehensive experiments on both in-house and external glioma datasets, we demonstrate that ROAM can automatically capture key morphological features consistent with the experience of pathologists and thus provide accurate, reliable and adaptable clinical-grade diagnoses of gliomas. Moreover, ROAM has clinical value for auxiliary diagnoses and could pave the way for the study of molecular and morphological correlations.

Gliomas are the most common primary intracranial tumour and have an extremely low 5 year relative survival[1]. The histopathological classification of gliomas is intricate. They are typically categorized into three subtypes, astrocytomas, oligodendrogliomas and ependymomas[2], each of which can be further divided into several grades. Consequently, accurate classification and grading are critical to prognostic assessments of and treatment plans for gliomas[3]. The fifth edition of the *WHO Classification of Tumors of the Central Nervous System*, released in 2021, has led to a growing interest in the molecular pathogenesis of brain tumours[4]. It is now believed that molecular features, such as the mutation of isocitrate dehydrogenase (IDH) and the methylation of the O-6-methylguanine-DNA methyltransferase (MGMT) promoter, contribute to the discrimination of gliomas[4,5], making a comprehensive and precise diagnosis even more complicated. The diagnosis of gliomas is usually accomplished by experienced pathologists, who rely on their personal expertise, by observing tissue sections. However, the scarcity of experienced pathologists, the subjective nature of diagnosis[6] and the lengthy diagnostic process[7] all contribute to the inadequacy of manual

A full list of affiliations appears at the end of the paper. ✉e-mail: zhangxg@mail.tsinghua.edu.cn; huzhongliang@csu.edu.cn; lvhairong@tsinghua.edu.cn

https://doi.org/10.1038/s42256-024-00868-w

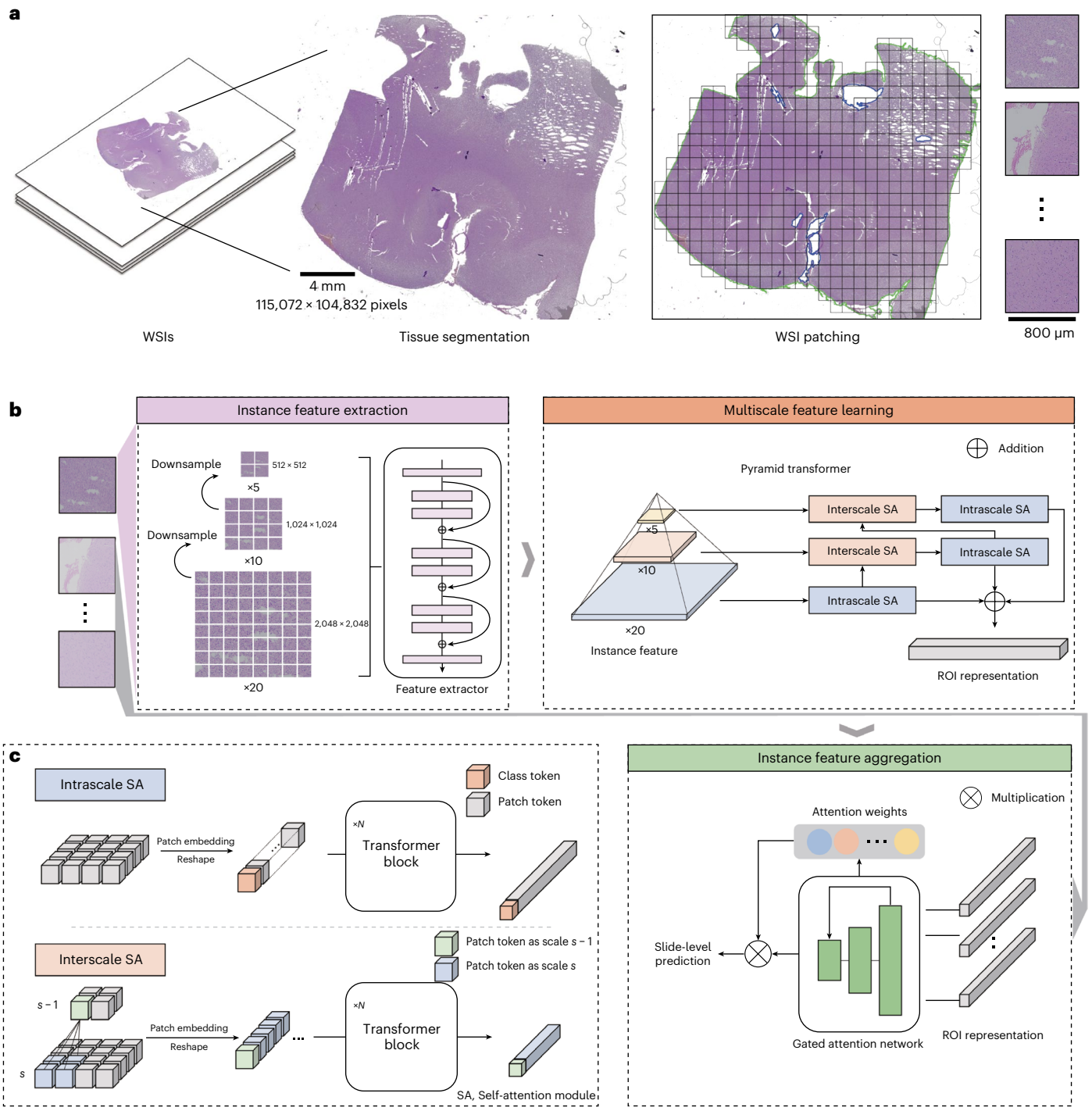


Fig. 1 | Overview of the basic framework and architecture of ROAM. a, WSIs are segmented to obtain tissue regions (left) and, subsequently, large image patches, referred to as ROIs, are extracted from these regions. **b**, For each ROI, two consecutive downsamplings are performed, resulting in images at three different magnification levels. Each image is then segmented into small image patches, which are encoded using a pretrained convolutional neural network to extract their visual representations, which are used as inputs for the MIL model (left). Multiscale SA modules (right) and an attention network (lower right) are used to generate the instance-level representation and aggregate this information into a slide-level representation. **c**, Two distinct types of SA module progressively fuse visual features from high to low magnification levels to generate a visual representation of the ROI. Intrascale SA modules and interscale SA modules learn the intrascale and interscale correlation characteristics of ROIs, respectively. Both types of module comprise several multi-head SA layers and feedforward layers.

diagnosis, which can hardly meet the current diagnostic demands of gliomas. Therefore, there is an urgent need for an effective artificial intelligence (AI) system that could assist doctors in diagnosing gliomas.

Recent advances in digital pathology and machine learning have enabled the digitization of histology slides to gigapixel whole-slide images (WSIs), which encompass extensive contextual data. Such slides have a huge potential for the diagnosis, prognosis and analysis of molecular features that may lead to precision oncology[8–10]. Researchers have also begun to employ deep-learning methods to facilitate the diagnosis and prognosis of gliomas[11–14]. However, these methods

analyse only regions of interest (ROIs) in pathology images that have already been selected by pathologists and so fail to achieve an automated analysis of an entire slide. To address this limitation, slide-level labels have been directly assigned to patches within a WSI to enable supervised learning with the labelled patches[15]. Nonetheless, this intuitive strategy, though remaining the primary approach at present[12,16], inevitably generates a large number of patches with noisy labels, and thus, it is not suited for datasets with limited tumour contents, as many normal patches may be incorrectly labelled as a tumour.

Multiple-instance learning (MIL) has been leveraged to alleviate this problem. With the assumption that a WSI is diagnosed as a tumour only if there are regions of lesioned tissue present within the image, researchers have modelled patches extracted from a slide as instances and obtained the feature representation of the slide through various aggregation methods[17–19]. Given the limited capacity of small image patches to capture large-scale image features, some studies have sought to learn visual representations of WSIs at various magnification levels[20–23]. Additionally, graph-based methods have been employed to extend the receptive fields of small image patches, thereby enabling the learning of large-scale features in WSIs[24–26]. Although these methods have made commendable attempts in learning multiscale features and enlarging the receptive fields, they still have limitations in learning visual representations of WSIs, especially due to coupling effects of multiscale features. Moreover, present approaches primarily concentrate on enhancing performance metrics, which may overlook their practical value, such as the potential to assist in clinical diagnosis.

In light of the above considerations, we propose ROAM, a multiscale self-attention (SA) MIL method based on large ROIs, which comprehensively addresses the above issues in computational pathology in the context of glioma diagnosis. ROAM effectively extracts rich multiscale information from pathological images, thereby achieving state-of-the-art performance across a variety of glioma classification tasks, including tumour detection, subtyping, grading and molecular feature prediction. Furthermore, ROAM can generate interpretable heat maps at both slide and patch levels with the attention mechanism and relevance propagation[27], which greatly assists pathologists in a wide spectrum of clinical applications, such as the identification of morphological features and the exploration of molecular features relevant to tumour diagnosis, thereby gaining insights into not only the clinical diagnosis but also the medical discovery of gliomas.

Results

Overview of the ROAM model

ROAM is designed on the premise that a large image patch extends the field of view and, thus, can effectively capture contextual information within a WSI, whereas the integration of visual features across different spatial positions and magnification levels of an image patch leads to a proper representation of the corresponding ROI. As illustrated in Fig. 1, ROAM first segments a WSI to identify tissue regions and subsequently extracts large image patches (2,048 × 2,048 pixels) to obtain ROIs. These regions are then fed to an instance feature extraction module, which applies two consecutive downsampling processes to obtain region images at three different magnification levels (2,048 × 2,048, 1,024 × 1,024 and 512 × 512). It segments each region image into small image patches of size 256 × 256 and uses a pretrained convolutional neural network to obtain visual representations of these patches. The resulting instance features are then processed by a multiscale feature-learning module, which relies on a pyramid transformer framework (Extended Data Fig. 1) to capture correlations between the visual representations of different patches within the same magnification level of an ROI by an intrascale SA mechanism and those among different magnification levels of an ROI by an interscale SA mechanism. The resulting multiscale features produced by the pyramid transformer are then fused from high to low magnification levels, thereby generating a visual representation of the ROI. Finally, an instance feature aggregation module takes ROIs as instances, follows the class-specific attention-weighted aggregation paradigm and produces slide-level predictions. Further details of the ROAM model can be found in Methods.

ROAM enables accurate diagnosis of glioma on in-house data

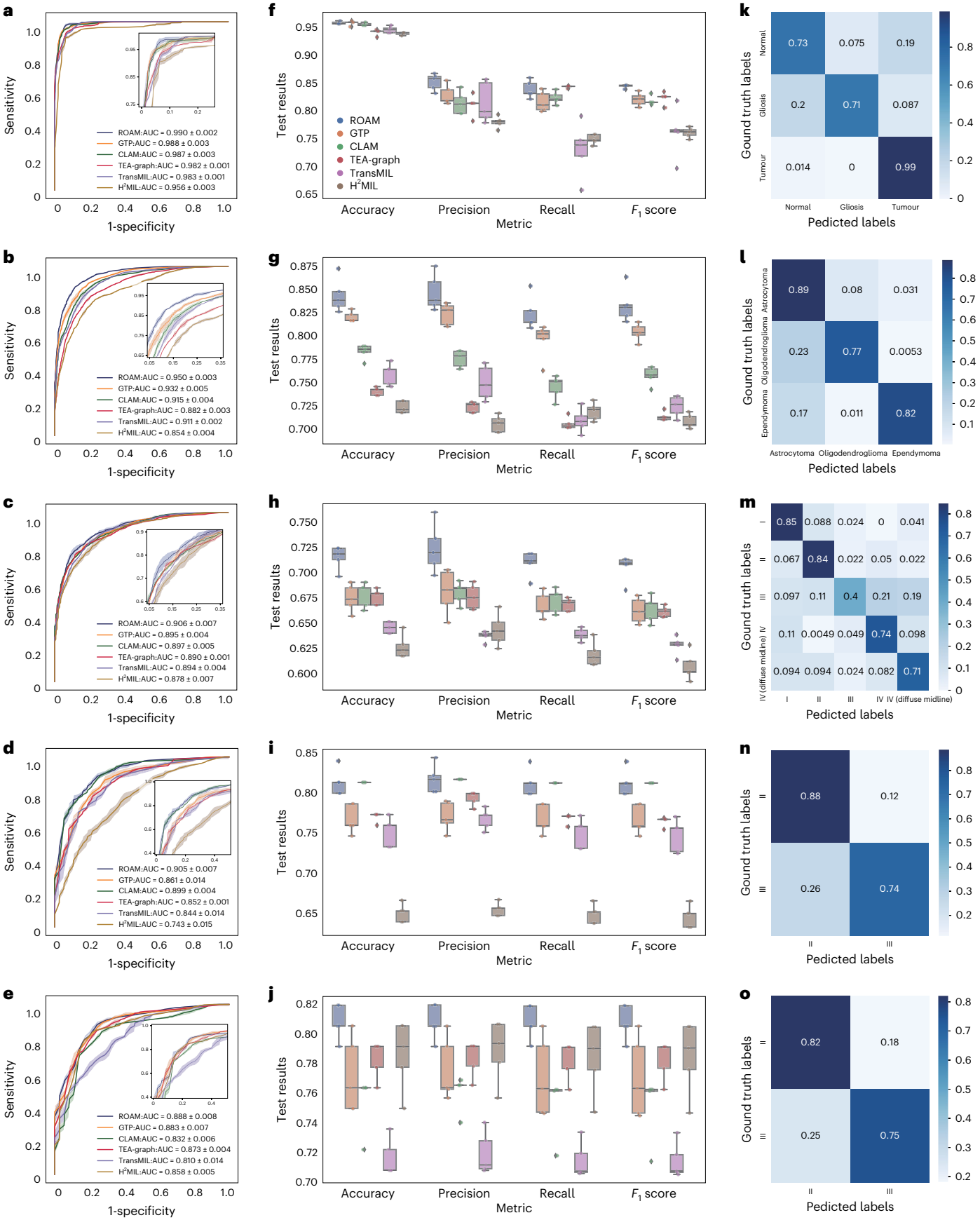
We simulated the scenario of real clinical applications and evaluated our method on independent test data. To achieve this objective, the Xiangya dataset was randomly split in a ratio of 2:1 to form an in-house training dataset (two-thirds of the data) and an independent in-house test dataset (one-third of the data). The glioma diagnosis process was divided into five diagnostic subtasks, namely glioma detection and classification and the grading of three subtypes (Extended Data Fig. 2). For each diagnosis task, we trained five ensemble models using the in-house training dataset and evaluated these models using the in-house test dataset. The results in Fig. 2 demonstrate the superior performance of our method in comparison to the ensemble versions of five baseline approaches—CLAM[18], TransMIL[19], GTP[25], TEA-graph[26] and H^2MIL (ref. 24)—as detailed in Methods. In the three-class glioma detection of normal, gliosis and tumour, ROAM achieved a macro-averaged one-versus-rest area under the receiver operating characteristic (ROC) curve of AUC = 0.990 ± 0.002 (Fig. 2a). For the three-class glioma subtyping of astrocytoma, oligodendroglioma and ependymoma, ROAM achieves an AUC = 0.950 ± 0.003 (Fig. 2b). In all these tasks for glioma diagnosis, ROAM achieved the highest AUC and outperformed all the baseline methods, demonstrating the effectiveness and high performance of our model in glioma diagnosis. Per-class AUCs are reported in Supplementary Figs. 1 and 2. We also evaluated the performance of ROAM for glioma diagnosis through fivefold cross-validation using the Xiangya dataset. The results clearly show the superior performance of our method over the baseline approaches (Extended Data Fig. 3). For a more detailed account of the results, see Supplementary Figs. 3–6 and Supplementary Note 1.

We further computed the confusion matrices and four other metrics (accuracy, macro precision, macro recall and macro F_1 score) in a comprehensive evaluation of our method. The results again confirm the superior performance of ROAM compared with the baseline approaches (Fig. 2f–o). Notice, however, that ROAM is not good at discriminating astrocytoma grade III (Fig. 2m). This is reasonable because making an accurate diagnosis of high-grade astrocytoma in clinical practice still needs the evidence of molecular features. We then removed data for astrocytoma grade IV (diffuse midline gliomas),

Fig. 2 | Performance of ROAM and baseline methods in glioma diagnosis on in-house data. a–**o**, Tasks in glioma diagnosis include glioma detection (**a**,**f**,**k**, *n* = 373), glioma subtyping (**b**,**g**,**l**, *n* = 324), astrocytoma grading (**c**,**h**,**m**, *n* = 178), oligodendroglioma grading (**d**,**i**,**n**, *n* = 75) and ependymoma grading (**e**,**j**,**o**, *n* = 72). Results were derived from five ensemble models trained using the in-house training dataset and tested on the in-house test dataset, both split from the Xiangya dataset in the ratio 2:1. **a**–**e**, ROC curves and the corresponding AUC ± s.d. for glioma detection (**a**), glioma subtyping (**b**), astrocytoma grading (**c**), oligodendroglioma grading (**d**) and ependymoma grading (**e**). The confidence bound shows ±1 s.d. for a curve. Insets, enlarged views of the curves. **f**–**j**, Accuracy, macro precision, macro recall and macro F_1 score for glioma detection (**f**), glioma subtyping (**g**), astrocytoma grading (**h**), oligodendroglioma grading (**i**) and ependymoma grading (**j**). The metrics are plotted as box plots with each box ranging from the upper to the lower quartile. The median is the horizontal line. Whiskers extend to 1.5 times the interquartile range. Diamonds represent outliers. All colours are consistent with those in **f**. **k**–**o**, Mean normalized confusion matrices for ROAM for glioma detection (**k**), glioma subtyping (**l**), astrocytoma grading (**m**), oligodendroglioma grading (**n**) and ependymoma grading (**o**).

https://doi.org/10.1038/s42256-024-00868-w

which relies heavily on molecular evidence, and found that our model had a significant improvement in predicting the other categories (Supplementary Fig. 7). Additionally, ROAM demonstrated superior data efficiency compared to the baseline approaches, maintaining good stability and classification performance even with limited training data (Supplementary Fig. 8).



https://doi.org/10.1038/s42256-024-00868-w

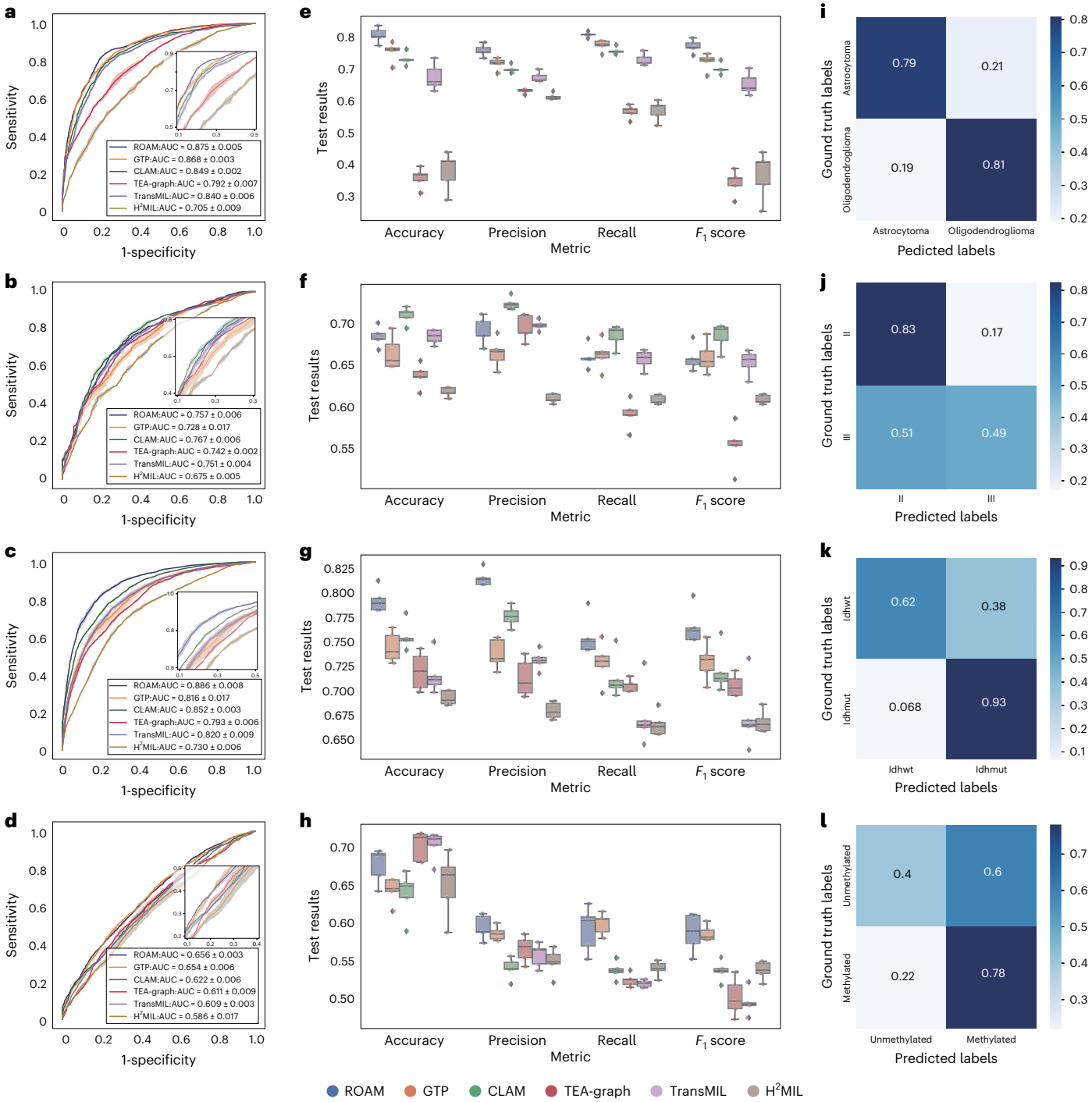


Fig. 3 | Performance of ROAM and baseline methods when generalized to independent external data. a–l, Results from five ensemble models trained using the entire Xiangya dataset and tested on the independent TCGA dataset. **a–d**, ROC curves and the corresponding AUCs for the four diagnosis tasks: glioma subtyping (**a**, $n = 621$), oligodendroglioma grading (**b**, $n = 156$), prediction of IDH mutation (**c**, $n = 621$) and prediction of MGMT promoter methylation (**d**, $n = 621$). Insets, enlarged views of the curves. **e–h**, Average accuracy, macro precision, macro recall and macro F_1 score for glioma subtyping (**e**), oligodendroglioma grading (**f**), prediction of IDH mutation (**g**) and prediction of MGMT promoter methylation (**h**). The metrics are plotted as box plots with each box ranging from the upper to the lower quartile. The median is the horizontal line. Whiskers extend to 1.5 times the interquartile range. Diamonds represent outliers. **i–l**, Mean normalized confusion matrices for ROAM for glioma subtyping (**i**), oligodendroglioma grading (**j**), prediction of IDH mutation (**k**) and prediction of MGMT promoter methylation (**l**). Idhmut, IDH mutation; Idhwt, IDH wild type.

We further evaluated the performance of our method in predicting two molecular features closely related to glioma diagnosis: IDH mutation and MGMT promoter methylation. Briefly, ROAM achieved an AUC of 0.918 ± 0.007 for the former and 0.762 ± 0.003 for the latter (Extended Data Fig. 4), clearly outperforming all the baseline methods. These results suggest the feasibility of determining molecular status by summarizing the characteristics of pathological images and, more importantly, indicate that ROAM has, indeed, captured some highly effective morphology features in images that could be used to predict molecular features, which is the crucial knowledge that we aimed

https://doi.org/10.1038/s42256-024-00868-w

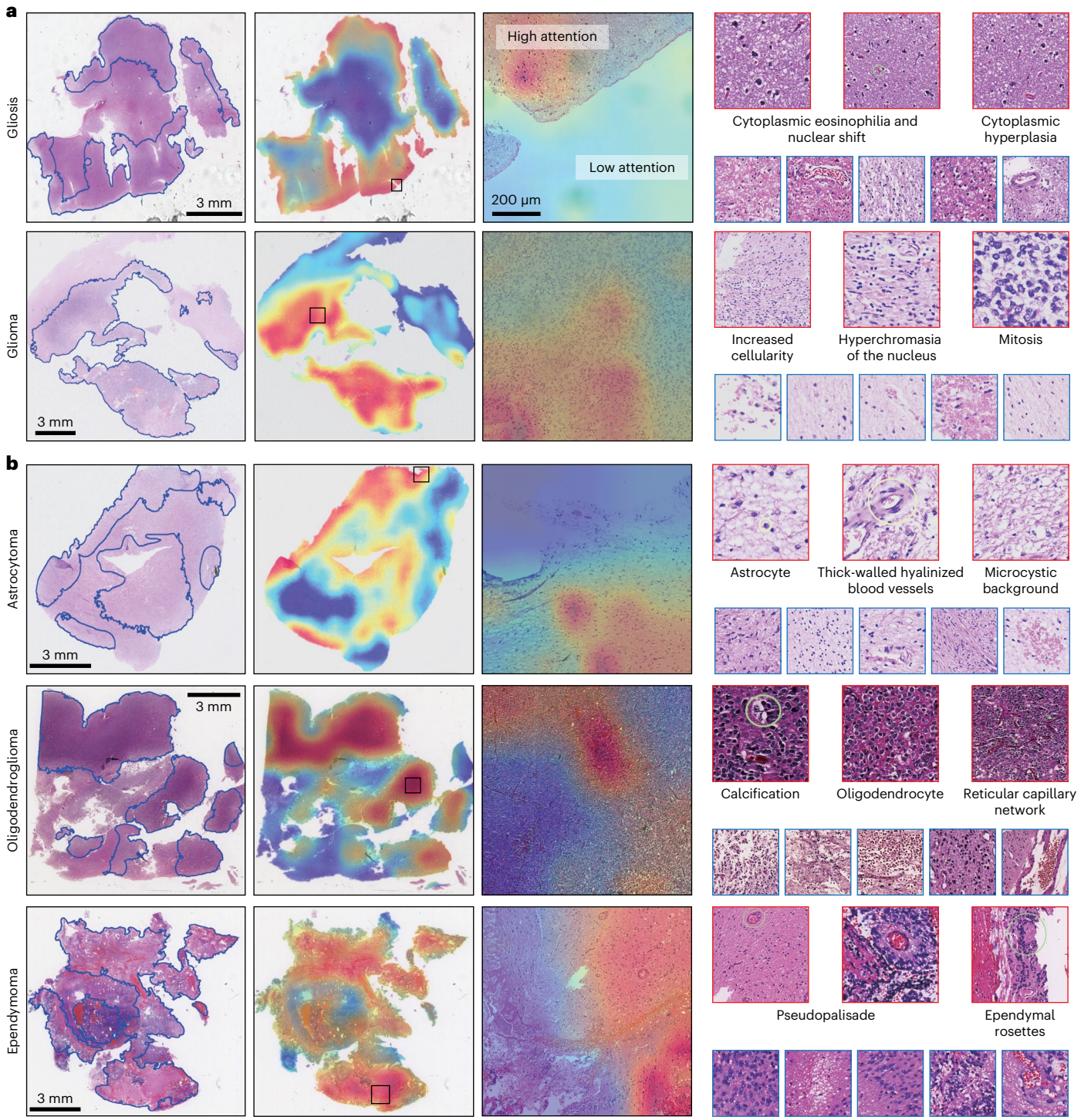


Fig. 4 | Visualization for glioma detection and subtyping. a,**b**, A representative slide from each category was selected to generate corresponding visualization results (first column). Each slide was annotated by a pathologist. **a**, Glioma detection with a slide showing gliosis and a slide showing glioma. **b**, Glioma subtyping with slides showing astrocytoma, oligodendroglioma and ependymoma. A whole-slide attention heat map was generated for each slide by calculating the attention scores for the predicted class over ROIs patched with an overlap of 90% (second column). The slide-level visualization of each ROI is depicted in a single colour due to having only one attention score. Therefore, the fine-grained ROI-level visualization heat map was generated by computing attention scores for each small image block within the ROI with respect to the predicted class based on the SA matrix (third column). Patches cropped from ROIs with high attention scores (red borders in the fourth column) have cellular or tissue features highly consistent with the corresponding class, which cannot be observed on patches with low attention scores (blue borders). Green circles highlight specific morphology features.

to uncover. We provide information on various diagnostic subtasks in the glioma diagnosis process and the two molecular features in Supplementary Note 2.

ROAM can be generalized to independent external data

Due to differences in slide processing and imaging equipment, WSIs from different institutions usually have considerable differences in

appearance. With this consideration, we demonstrated the excellent generalizability of our method to an independent external dataset containing 618 WSIs curated from The Cancer Genome Atlas (TCGA).

Considering the data distribution in the TCGA dataset, we evaluated the generalizability of our method in four diagnostic tasks: glioma subtyping, oligodendroglioma grading and predicting two molecular features. For each diagnosis task, we constructed five ensemble models using the Xiangya dataset, and we evaluated these models using the independent TCGA dataset. The results clearly show the superior generalizability of our method compared to the baseline approaches. which were also trained using the same ensemble strategy (Fig. 3). Specifically, ROAM achieved an AUC of 0.875 ± 0.005 for the two-class glioma subtyping of astrocytoma and oligodendroglioma (Fig. 3a) and 0.757 ± 0.006 for the grading of oligodendroglioma (Fig. 3b and Supplementary Fig. 9). Besides, ROAM achieved an AUC of 0.886 ± 0.008 for predicting IDH mutations and 0.656 ± 0.003 for predicting MGMT promoter methylation (Fig. 3c,d). Although the diagnostic performance of ROAM decreased slightly with the independent external test dataset, two out of the four tasks (glioma subtyping and IDH mutation prediction) still achieved AUCs greater than 0.870, suggesting the effectiveness of the generalization of our method to independent external data. Overall, ROAM outperformed all baseline methods in three out of the four diagnostic tasks and ranked second for oligodendroglioma grading in terms of the AUC (Fig. 3a–d), confirming the superior generalizability of our method over the baseline approaches. We also computed the confusion matrices and the four metrics for each diagnostic task in a comprehensive evaluation of our method with the consideration of data imbalance (Fig. 3e–l). The results again suggest the excellent generalizability of our method to independent external data. We further confirmed the generalizability of ROAM through multi-centre validation experiments. The results and analysis are detailed in Supplementary Fig. 10 and Supplementary Note 3.

ROAM enhances the visualization and interpretation of diagnosis

A clear and readable interpretation of the results of a diagnostic model is crucial for allowing pathologists to validate the reliability of the diagnostic basis of the model and extract valuable information to form new knowledge for future diagnoses. ROAM first learns class-specific attention scores for each ROI to obtain the contribution of each tissue region to the diagnosis of the WSI. It then aggregates the visual representation of each ROI based on attention scores to provide the final slide-level prediction. Therefore, the morphological features in the regions with high attention scores serve as the primary basis for the model's diagnostic predictions. To visualize the contributions of these important ROIs, we used a colour map to convert attention scores into RGB colours and applied them to the corresponding spatial locations in the WSI. ROIs with a 90% overlap were cropped, followed by averaging operations across these regions to generate fine-grained attention heat maps. In addition, to further explore the contributions of different regions within each high-attention ROI, we generated more refined ROI-level heat map visualizations based on the SA mechanism. In conjunction with slide-level and ROI-level visualization results, we can distinctly observe the morphological structures within the pathological images that substantially contribute to the predictions made by the model.

An essential finding is that the morphological features in high-attention regions are generally consistent with the established diagnostic criteria recognized by pathologists and widely used in clinical practice. For example, the presence of a pseudopalisade and ependymal rosettes serves as critical criteria for diagnosing ependymoma clinically. These two morphological features are also the primary focus of the model trained for glioma subtyping when diagnosing ependymoma, as it appears frequently within the predicted high-attention regions (Fig. 4b). The model's focus on astrocytes for diagnosing astrocytoma and on oligodendrocytes for diagnosing oligodendroglioma is also an important criterion for identifying these two subtypes in clinical practice (Fig. 4b). In addition, for the glioma grading, the model pays more attention to the unique morphological features of each grade rather than the common features of the subtype, which are not beneficial to grading (Supplementary Figs. 11 and 12). In clinical practice, diffuse midline gliomas exhibit similar morphological features to other high-grade astrocytomas (Supplementary Fig. 13). Moreover, the high-attention regions of a slide, which were predicted by our method to be at different stages of the cascade diagnosis of glioma, show remarkable differences and strong relevance to the specific diagnostic task (Supplementary Fig. 14). This observation suggests that the model focuses on task-relevant and informative features, rather than relying on generic features with weak discriminative power. In addition, fine-grained heat maps demonstrate the great potential of our weakly supervised method in finer-grained pathological image tasks, such as segmentation or detection. Furthermore, these heat maps can be used not only for verification but also as valuable references to aid pathologists in diagnosing gliomas in clinical practice.

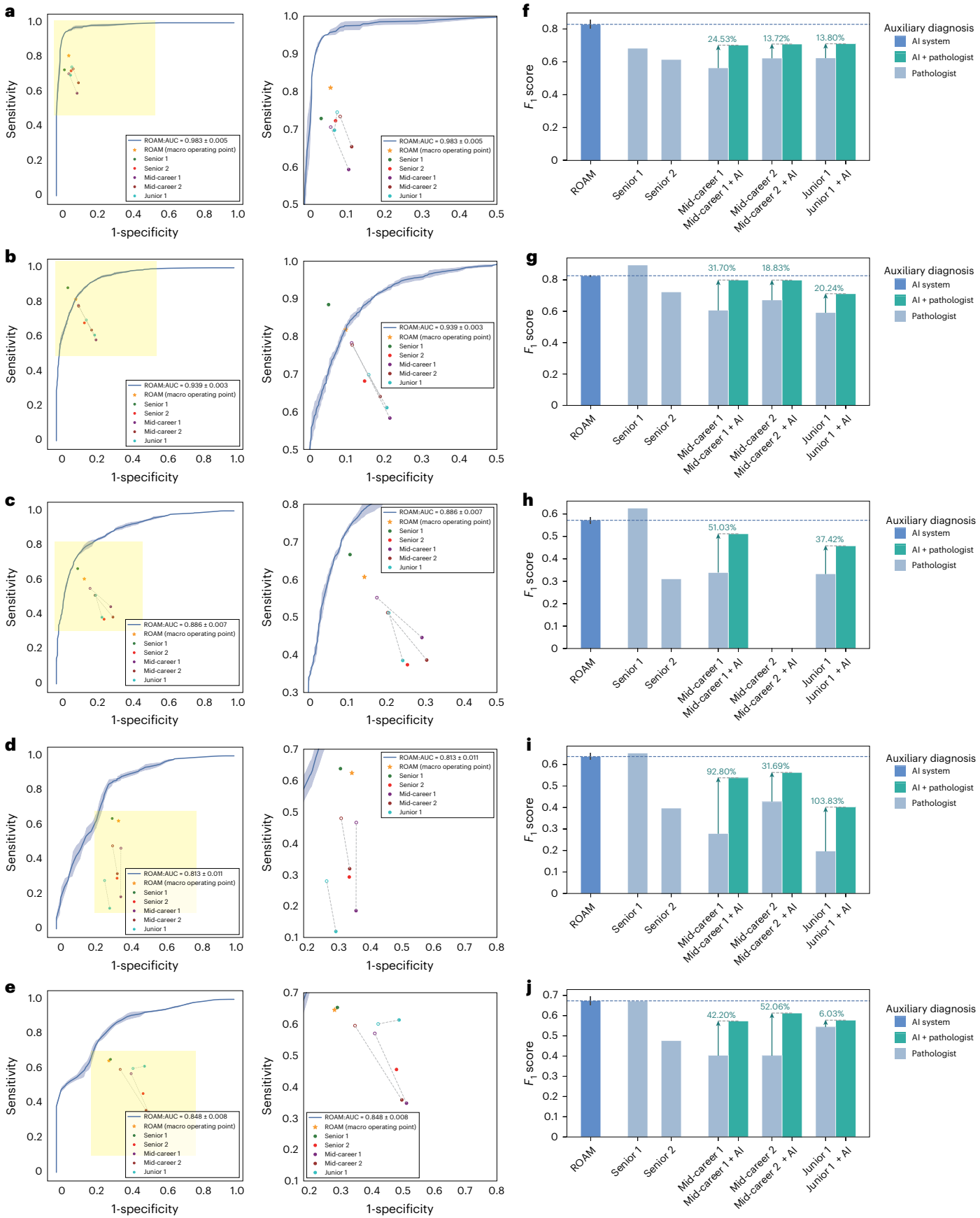
ROAM promotes auxiliary diagnosis

We compared the diagnostic predictions of our method with those of practising pathologists to assess the real ability of ROAM in clinical settings. Diagnosing a glioma in clinical practice is a progressive process. Given a WSI, a pathologist first assesses whether it contains tumour tissues. If the slide does, the pathologist will carefully determine the glioma subtype and further infer the degree of glioma grading to give the final diagnosis. Following this procedure, we constructed a cascade diagnostic system based on the ensemble models previously trained using the in-house training data to evaluate the value of ROAM in real diagnostic scenarios (Extended Data Fig. 2d). In this system, the classification result at the previous level determines which classification model the current slide is assigned to for classification at the next level. For example, WSIs input to the astrocytoma grading model have the same upper-level label, astrocytoma. WSIs that are incorrectly predicted will then be put into incorrect models for further prediction at the next level, leading to an accumulation of errors as the cascade deepens. We invited five pathologists in three groups to participate in the study, including one pathologist in the junior group with less than 5 years of clinical experience, two pathologists in the mid-level group with 5 to 15 years of clinical experience and two pathologists in the senior group with more than 15 years of clinical experience. Each pathologist independently diagnosed WSIs in the in-house test dataset.

We evaluated the results predicted by our cascade diagnostic system in comparison to those of the pathologists (Supplementary Note 4). As shown in Fig. 5 and Supplementary Fig. 15, our system

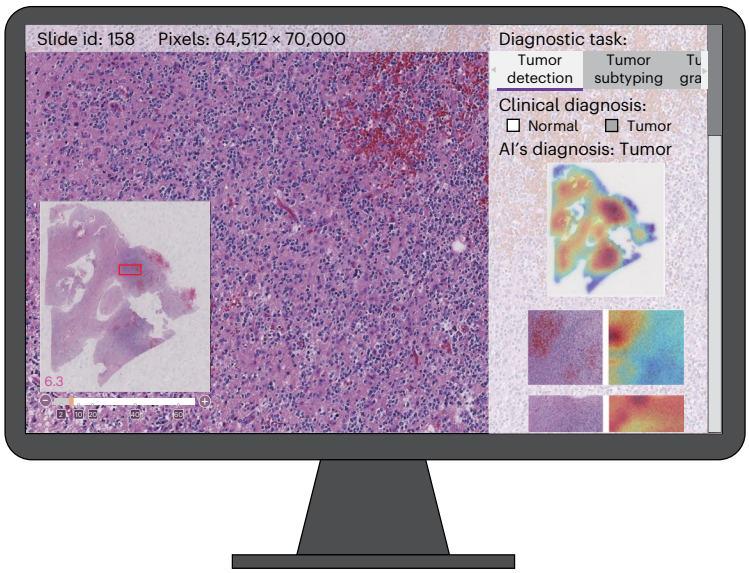
Fig. 5 | Performance of the cascade diagnostic system and auxiliary diagnosis. a–e, Performance of the cascade diagnostic system and the five pathologists. Macro-averaged one-versus-rest ROC curves, mean test AUC ± s.d. (first column) and their locally enlarged views (second column) are plotted for the subtask at each step: glioma detection (**a**, $n = 373$), glioma subtyping (**b**, $n = 324$), astrocytoma grading (**c**, $n = 178$), oligodendroglioma grading (**d**, $n = 75$) and ependymoma grading (**e**, $n = 72$). The confidence bound shows ±1 s.d. for each curve. Orange stars denote the performance of the cascade diagnostic system based on ROAM (average of the five ensemble models trained previously), filled dots that of the pathologists and hollow dots that of the junior and mid-level groups with the assistance of ROAM. A dashed line links the paired performance of a junior or mid-level pathologist. **f–j**, Macro-averaged F_1 score of ROAM (mean F_1 score ± s.d.) and pathologists (F_1 score) plotted for glioma detection (**f**), glioma subtyping (**g**), astrocytoma grading (**h**), oligodendroglioma grading (**i**) and ependymoma grading (**j**). Percentage increases in diagnostic metrics for junior and mid-level pathologists after AI assistance are marked above green arrows. The F_1 score for mid-career 1 in **h** is absent, indicating that this pathologist was unsuccessful in this diagnostic task.

https://doi.org/10.1038/s42256-024-00868-w



https://doi.org/10.1038/s42256-024-00868-w

a



b

Molecular feature	Associated morphological features
IDH mutated	(1) Prominent eosinophilia and homogeneous cytoplasm
	(2) Presence of hyperchromatic nucleus
MGMT methylated	No significant morphological features identified

c

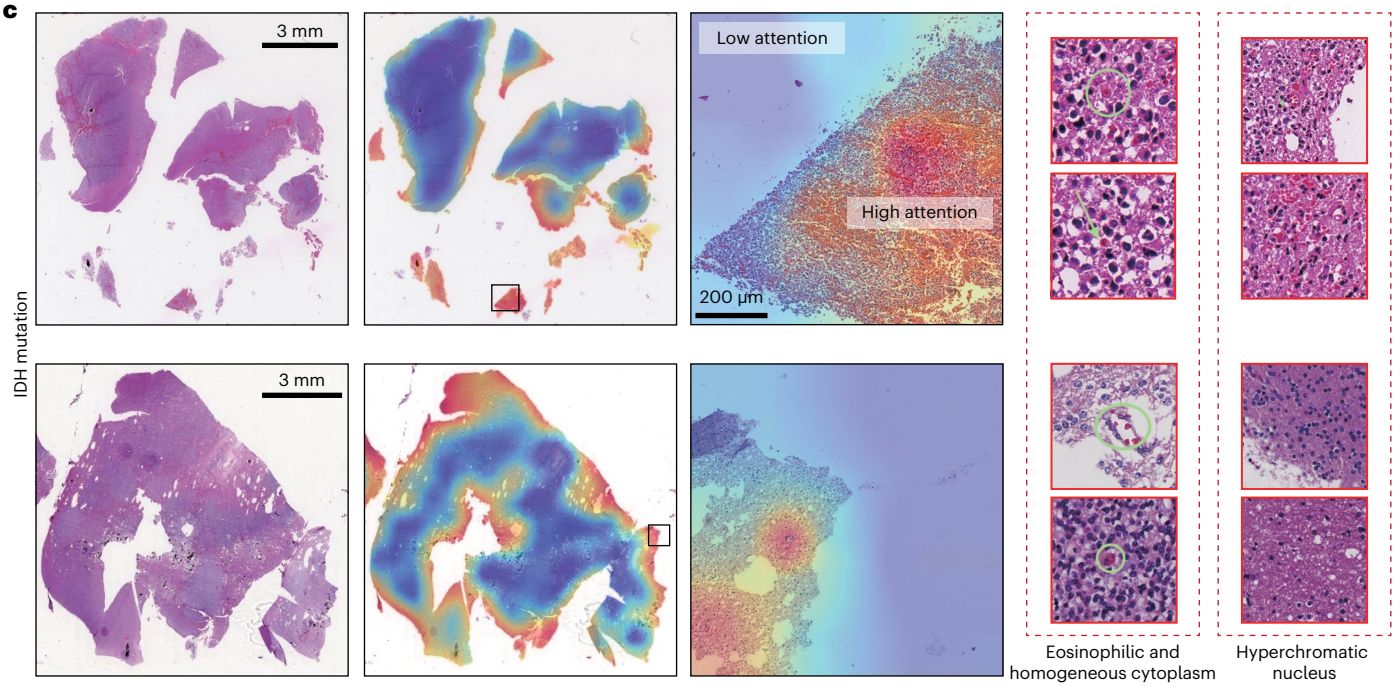


Fig. 6 | Molecular and morphological biomarker discovery through human–AI collaboration. a, Illustration of the online platform used by pathologists to browse pathology images and make diagnoses. Predictions made by ROAM are uploaded to the platform to assist pathologists in making more accurate diagnoses and uncovering valuable morphological features. **b**, Discoveries regarding molecular and morphological markers of glioma. **c**, First column: two slides showing IDH mutation. Second column: whole-slide attention heat maps. Third column: ROI-level visualization of regions, marked by boxes in the first column, with high attention scores. Fourth column: patches extracted from high-attention ROIs (red solid box) exhibit distinct features highly correlated with IDH mutation (red dashed boxes). Refer to the green arrows or ellipses in the figure for specific morphological features.

demonstrated excellent performance on the five tasks of glioma cascade diagnosis, outperforming four out of five pathologists and achieving comparable results to the best-performing senior pathologist (Senior 1). Specifically, for the glioma detection, our system achieved an average macro F_1 score of 0.828, significantly outperforming all the pathologists, including the best-performing pathologist, by 21.30% (Fig. 5f). These results demonstrate the significant potential of our system in the practical clinical diagnosis of gliomas.

We then asked the three junior and mid-level pathologists to make diagnoses with the assistance of ROAM so that we could investigate whether their diagnostic performance could be improved. The experiments were conducted 2 months after their initial diagnoses to avoid a potential memorization bias. Specifically, for each slide in the in-house test dataset, we provided the prediction and corresponding visualization results, including the heat maps of the entire slide and the coordinates of the top five tissue regions with the highest attention scores at each stage during the cascade diagnostic process used by our system. This allowed the pathologists to combine their own observations and expertise with the morphological features of the tissue regions that our system focused on to make a more accurate diagnosis. The entire auxiliary diagnostic process was performed through an online platform (Fig. 6a). With the help of ROAM, the diagnostic

accuracy of the three pathologists increased by an average of 7.27% (junior 1), 12.87% (mid-level 1) and 9.96% (mid-level 2) across all tasks (Fig. 5). The auxiliary diagnostic results of the overall glioma diagnostic task, as reported in Extended Data Fig. 5, clearly demonstrate that the assistance of ROAM significantly improves the diagnostic accuracy of less experienced pathologists, thereby contributing to the elevation of healthcare standards in economically disadvantaged regions that lack skilled pathologists and demonstrating the immense value of our method in clinical practice.

ROAM boosts the discovery of molecular and morphological biomarkers

Its outstanding performance in predicting molecular features suggests that ROAM may have identified some crucial morphological features that are helpful in discriminating the molecular characteristics of gliomas. To uncover these important potential molecular and morphological biomarkers for IDH mutation, we conducted the following analysis in collaboration with the two senior pathologists, each of whom had over 15 years of clinical experience. Briefly, ROAM provided visualizations of each pathological slide with IDH mutation, including whole-slide attention heat maps and visualizations of high-attention ROIs through the same online platform used for the auxiliary diagnoses (Fig. 6a). The pathologists observed these high-attention ROIs and summarized their common characteristics (Fig. 6b). Based on the visualization of such ROIs in diffuse astrocytoma and oligodendroglioma, distinct features in gliomas with IDH mutation were uncovered (Fig. 6c). These features, summarized as (1) prominent eosinophilic and homogeneous cytoplasm and (2) the presence of a hyperchromatic nucleus, were confirmed from an analysis of frozen intraoperative glioma sections and imprints. Using these features as molecular and morphological biomarkers, the IDH status in glioma could be determined without DNA sequencing, thereby providing valuable knowledge for subsequent diagnostic work that may improve patient prognosis. More importantly, these biomarkers could be used to optimize and refine clinical diagnostic standards and, thus, promote more accurate diagnoses in the absence of molecular sequencing data. We also conducted similar analysis for pathological slides with MGMT promoter methylation but did not find distinct morphological features. We conjecture that the reason may be related to the relatively weak performance of ROAM when predicting MGMT promoter methylation.

Discussion

This paper describes ROAM, a transformer-based, weakly supervised, computational pathology framework that relies on MIL. Our aim is to provide a comprehensive, general and effective solution for the pathological diagnosis of gliomas in clinical practice. We performed a rigorous evaluation of ROAM through detailed experiments that confirmed its effectiveness and robustness. Using only slide-level annotations for training, ROAM achieved diagnostic performance comparable to or exceeding that of experienced senior pathologists and exhibited strong generalizability with an independent external test cohort. ROAM could serve as a supporting diagnostic tool by offering credible predictions and diagnostic references and also aid pathologists in making decision by providing strong interpretable evidence through fine-grained attention heat maps. Furthermore, using a human–AI collaboration, we were successful in summarizing a set of highly reliable morphological features that are associated with molecular characteristics of gliomas. This knowledge could assist pathologists in making effective assessments of molecular features when only pathological images are available, thereby enabling comprehensive clinical diagnoses.

The following aspects of ROAM could be extended to overcome its current limitations. First, large ROIs compromise the precision of heat maps, hamper the observation of subcellular structures and are not suitable for diagnosing tumours with small tissue areas. The exploration of a flexible means of extracting image information at variable scales is, therefore, desirable. Second, although it is believed that adult and paediatric gliomas have different morphological characteristics, this was not clearly observed in our work, probably due to the limited amount of data. Therefore, it is necessary to integrate data from other modalities, such as radiology images, electronic health records and molecular biomarkers, which would allow the precise discrimination of tumours with only subtle morphological differences in fine-grained diagnostic tasks. Third, the success of the cascade diagnostic system based on ROAM suggests the importance of using prior clinical knowledge to guide the diagnostic process of an AI model. Although our efforts in this direction are just beginning, we believe that such a knowledge-guided and data-driven approach will greatly improve interpretability and have important clinical applicability. Finally, we hold the opinion that close human–AI collaboration is indispensable for achieving scientific discoveries in future research into computational pathology. We hope that our work will provide researchers with new insights that will allow them to progressively build a reliable, comprehensive, interpretable and general clinical diagnostic system and extract valuable knowledge from WSIs, thereby promoting the research and clinical applications of computational pathology.

Methods

The ROAM model

ROAM constructs a MIL framework based on large-scale ROIs to tackle the intricate and multilevel clinical diagnostic tasks relating to glioma using only slide-level annotations. The overall pipeline of ROAM includes the following four steps: (1) extraction of ROIs and patch features, (2) learning the representation of instances by transformer-based multiscale SA modules, (3) class-specific, attention-weighted, instance aggregation and (4) multilevel supervision.

Extraction of ROIs and patch features. Tissue regions are segmented from an WSI. ROIs of size 2,048 × 2,048 are segmented from the resulting tissue regions at magnification ×20 (0.5 µm per pixel) and treated as instances of the slide. A single WSI can be divided into dozens to hundreds of ROIs without overlapping. To extract information at different magnification levels, ×2 and ×4 downsampling are applied to each ROI, generating region images of size 1,024 × 1,024 and 512 × 512, which can be regarded as images captured at ×10 and ×5 magnification levels with the same field of view as the original ROI. The ith ROI of the slide and its downsampled images at ×10 and ×5 magnification levels are denoted as R_i^0, R_i^1 and R_i^2, respectively. Subsequently, patches of size 256 × 256 are segmented from these images, namely R_i^0, R_i^1 and R_i^2, without overlapping and then put into a pretrained convolutional neural network to generate corresponding features. The visual features at ×20, ×10 and ×5 are denoted as $X_i^0 \in \mathbb{R}^{64\times c}$, $X_i^1 \in \mathbb{R}^{16\times c}$ and $X_i^2 \in \mathbb{R}^{4\times c}$, where 64, 16 and 4 are the numbers of patches within the ROI of different magnification levels, and c the dimension of the features. Therefore, the initial multiscale visual features of the ith ROI of the slide are represented as $X_i = \left[X_i^0, X_i^1, X_i^2\right] \in \mathbb{R}^{84\times d}$. Following the MIL formulation, a WSI can be viewed as a bag b composed of several instances (ROIs of the slide), $b = \{X_1, X_2, \ldots, X_M\}$, where M is the number of ROIs in the WSI.

Transformer-based multiscale SA learning. ROAM constructs a multiscale SA module to fully leverage contextual information about ROIs. As illustrated in Extended Data Fig. 1, two distinct types of module, intrascale SA modules and interscale SA modules, model correlations between patches at disparate positions and between different scales of histological morphological features within ROIs. The intrascale SA module is implemented at each scale as in ViT[28]. Given an ROI at scale s, the initial visual features, $X^s = \{\mathbf{x}_1^s, \mathbf{x}_2^s, \ldots, \mathbf{x}_{n_s}^s\}$, where $\mathbf{x}_i^s \in \mathbb{R}^d$ is the ith patch token and n_s the number of patches, can be

acquired through the aforementioned process. A class token $\mathbf{x}^s_{\text{cls}}$ and a learnable position encoding E_{pos} are introduced to compute the embedded token Z^s_0:

$$Z^s_0 = [\mathbf{x}^s_{\text{cls}}; \mathbf{x}^s_1 E; \mathbf{x}^s_2 E; \dots; \mathbf{x}^s_{n_s} E] + E_{\text{pos}}, \tag{1}$$

where E is the patch embedding projector and $Z^s_0, E_{\text{pos}} \in \mathbb{R}^{(n_s+1)\times d}$. Subsequently, the transformer encoder[29], which contains L layers of multi-head SA modules (Supplementary Note 5), and multilayer perceptron blocks are applied to learn the spatial correlation of ROI (intrascale SA), as given by

$$Z^{s'}_l = \text{MSA}(\text{LayerNorm}(Z^s_{l-1})) + Z^s_{l-1}, \quad l = 1, 2, \dots, L, \tag{2}$$

$$Z^s_l = \text{MLP}(\text{LayerNorm}(Z^{s'}_l)) + Z^{s'}_l, \quad l = 1, 2, \dots, L. \tag{3}$$

Relative position bias in the multi-head SA was introduced to better model the positional relationships between tokens[30]. A multilayer perceptron contains two fully connected layers and a Gaussian error linear unit with a nonlinear activation function. Accordingly, the visual representation of the ROI at scale s, denoted $\mathbf{h}^s_{\text{cls}}$, is given by $\mathbf{h}^s_{\text{cls}} = Z^{s(0)}_L$, where $Z^{s(0)}_L$ is the first token of Z^s_L, the class token. Patch tokens that incorporate spatial correlations are given by $X^s = [\mathbf{z}^s_{L,1}; \mathbf{z}^s_{L,2}; \dots; \mathbf{z}^s_{L,n_s}]$.

With the representation of patches at scale s, the interscale SA is constructed to integrate finer-scale image information into the learning of larger-scale visual features. The scaling factor between adjacent scales of ROIs is ×2. Each patch of size 256 × 256 within the ROI at scale $s+1$ originates from a single patch of size 512 × 512 at the same location at scale s, which is subdivided into four 256 × 256 patches for individual learning. This implies that for each patch within the ROI at scale $s+1$, four corresponding patches at scale s that represent the same tissue region can be identified, with the only difference being the scale of magnification. Patch tokens representing the same tissue region are assigned the same number in Extended Data Fig. 1. By concatenating these five patches together, the interscale SA module transfers information from the lower scale (s) to the higher scale ($s+1$), thereby learning a patch token that integrates multiscale features. Therefore, the concatenated multiscale patch tokens of the ith patch of ROI at scale $s+1$ that will be put into interscale SA module for learning, denoted $Z^{s\to(s+1)}_{0,i} \in \mathbb{R}^{(1+4)\times d}$, are given by

$$Z^{s\to(s+1)}_{0,i} = [\mathbf{x}^{s+1}_i E; \mathbf{z}^s_{L,i_1}; \mathbf{z}^s_{L,i_2}; \mathbf{z}^s_{L,i_3}; \mathbf{z}^s_{L,i_4}], \tag{4}$$

where i_1, i_2, i_3 and i_4 are the indices of four corresponding patches at scale s. Then, multiscale patch tokens are put into the transformer encoder with L layers, which is the interscale SA module. Patch tokens integrating finer-scale image information, denoted $\mathbf{z}^{s+1}_i$, can be obtained by extracting the first token of $Z^{s\to(s+1)}_{0,i}$. Subsequently, $[\mathbf{z}^{s+1}_0, \mathbf{z}^{s+1}_1, \dots, \mathbf{z}^{s+1}_{n_{s+1}}]$ is put into another intrascale SA to learn visual representations of ROI at scale $s+1$. This iterative process continues until the model learns the visual representations of ROIs across all scales. The final representation of the ith ROI of the slide, denoted $\mathbf{h}_i$, is given by

$$\mathbf{h}_i = \sum_{s=0}^{2} w_s \mathbf{h}^s_{i,\text{cls}}, \tag{5}$$

where $\mathbf{h}^s_{i,\text{cls}}$ represents the visual representation of the ROI at scale s and w_s is the corresponding weight coefficient. The influence of the size of the ROI and multiscale feature fusion on the performance of the model is illustrated in Supplementary Fig. 16. The full process for learning the representation of an ROI (instance) is summarized in Algorithm 1.

Algorithm 1: Multiscale SA learning

function Multiscale_SA_Learning(X^0, X^1, X^2)
  for s ← 0,1,2 **do**
    $Z^s = [\mathbf{x}^s_{\text{cls}}; X^s E] + E_{\text{pos}}$
  end for
  for s ← 0,1,2 **do**
    $\mathbf{h}^s_{\text{cls}}, Z^s \leftarrow$ **IntrascaleSA**(Z^s) # Intrascale SA learning
    $Z^{s\to(s+1)} \leftarrow$ **Concat**($[Z^{s+1}; Z^s]$) #Concatenatefeaturesbetween two scales
    $Z^{s+1}, \_ \leftarrow$ **InterscaleSA**($Z^{s\to(s+1)}$) # Interscale SA learning
  end for
  $\mathbf{h} \leftarrow \sum_{s=0}^{2} w_s \mathbf{h}^s_{\text{cls}}$
end function

Class-specific, attention-weighted, instance-aggregation function. ROAM adopts the same instance-aggregation function as CLAM[18]. A gated attention mechanism[31] is constructed to learn the attention score of each ROI for class k, denoted $a_{k,i}$, as given by equation (6), where $k \in \{1, 2, \dots, K\}$ and K is the number of classes. The attention score $a_{k,i}$ represents the contribution of the ith ROI to the predicted class k for the slide. The class-specific representation for class k of the slide, denoted $\mathbf{h}_{\text{slide},k}$, is given by equation (7):

$$a_{k,i} = \frac{\exp\{W_{a,k}(\tanh(V_a \mathbf{h}_i) \odot \text{sigmoid}(U_a \mathbf{h}_i))\}}{\sum_{j=1}^{N} \exp\{W_{a,k}(\tanh(V_a \mathbf{h}_j) \odot \text{sigmoid}(U_a \mathbf{h}_j))\}}, \tag{6}$$

$$\mathbf{h}_{\text{slide},k} = \sum_{i=1}^{N} a_{k,i} \mathbf{h}_i, \tag{7}$$

where $V_a, U_a \in \mathbb{R}^{\frac{d}{2}\times d}$ are two layers of the attention network, and $W_{a,k} \in \mathbb{R}^{1\times d}$ is the fully connected layer that projects the representation vector into attention scores. The prediction score for class k can then be given by the corresponding classifier $W_{c,k} \in \mathbb{R}^{1\times d}$ by $s_{\text{slide},k} = W_{c,k} \mathbf{h}_{\text{slide},k}$. The probability distribution over each class is eventually obtained by applying a softmax function to the prediction score of the slide $\mathbf{s}_{\text{slide}}$. ROAM does not use a transformer for instance aggregation because the correlations between instances have already been sufficiently considered in the previous instance feature-learning stage. Besides, simpler structures facilitate model training. The results in Supplementary Table 1 prove our point.

Multilevel supervision. The multilevel supervision mechanism, which consists of bag-level supervision and instance-level supervision, is employed to enhance the efficiency of the model in utilizing data and encourage the representation learning of instances. With K class-specific representations of the slide obtained by equation (7), the same number of corresponding class-specific classifiers are constructed to individually predict probabilities of the slide being classified into a class. Accordingly, the slide-level prediction of probabilities for K-class classification, denoted $\mathbf{p}$, is given by equation (8). Given a batch of WSIs with ground-truth labels assigned as $\{\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_M\}$, where $\mathbf{y}_i \in \mathbb{R}^K$ is one-hot encoding of the ground-truth class label for the ith slide, the cross-entropy loss is adopted as the loss function, as given by equation (9).

$$\mathbf{p} = \text{softmax}([\mathbf{h}_{\text{slide},1}, \mathbf{h}_{\text{slide},2}, \dots, \mathbf{h}_{\text{slide},K}]), \tag{8}$$

$$\mathcal{L}_{\text{bag}} = -\frac{1}{M} \sum_{i=1}^{M} \sum_{j=1}^{K} \mathrm{y}_{i,j} \log \mathbf{p}_{i,j}. \tag{9}$$

In addition to bag-level supervision, instance-level supervision with pseudo-labels is introduced, as in CLAM[18]. Considering that instance-level supervision does not have ground-truth labels

for training, attention scores of ROIs for each class are used to generate pseudo-labels. Given a WSI with ground-truth label y, the attention scores of each ROI in the slide for class y can be obtained from the output of the attention network. Although the true labels of ROIs are unknown, it can be reasonably inferred that ROIs with high attention scores for class y have a higher probability of being assigned label y. This is because these ROIs contribute critically to the correct prediction of label y for the slide, as indicated by their attention scores. Consequently, all ROIs are sorted according to their attention scores for label y, $a_{y,i}$, and the label of the slide y is assigned to the top B ROIs with the highest scores as pseudo-labels. The representations of these ROIs are put into an instance-level K-class classifier $W_{\text{inst}} \in \mathbb{R}^{K\times d}$ to predict class labels. The instance-level loss is calculated by a cross-entropy function. The loss $\mathcal{L}_{\text{inst}}$ is given by equations (10) and (11):

$$p_{\text{inst},i} = W_{\text{inst}}\mathbf{h}_i, \tag{10}$$

$$\mathcal{L}_{\text{inst}} = -\frac{1}{B}\sum_{i=1}^{B}\log p_{\text{inst},i}^{(y)}, \tag{11}$$

where $p_{\text{inst},i}^{(y)}$ represents the probability that the ith ROI of the top B ROIs is predicted as class y.

In contrast to CLAM, negative samples with pseudo-labels based on attention scores are not included in the analysis because the pseudo-labels for negative samples are noisy. For certain ROIs, such pseudo-labels exhibit low attention scores for class y or high attention scores for other classes, and thus, the contributions of such pseudo-labels to the classification of the slide as class y are indeed ambiguous. Therefore, directly assigning pseudo-labels to negative samples for instance-level supervision will introduce considerable interference in model training (Supplementary Tables 2 and 3). Accordingly, the total loss for training $\mathcal{L}$ is

$$\mathcal{L} = \mathcal{L}_{\text{bag}} + \alpha\mathcal{L}_{\text{inst}}, \tag{12}$$

where α is the weighting coefficient of $\mathcal{L}_{\text{inst}}$.

Model training. During the preprocessing stage, visual features of patches in ROIs with different magnifications were extracted using ResNet50 (ref. 32), which is pretrained on ImageNet[33] by default. The sampling probability of a slide was inversely proportional to the number of slides in its class in the training set to ensure that the class distribution was balanced during training. The batch size was set to four by default. An ROI dropout mechanism that randomly discarded 20% of the ROIs when the number of ROIs in the slide was more than 16 was adopted to enhance the generalizability of the model during training. Class tokens were initialized with a truncated normal distribution using a standard deviation of 0.02. The number of layers in both the intrascale SA and interscale SA modules was configured as two. The weights assigned to the visual representations of ROIs at ×20, ×10 and ×5 magnifications were 1/3, 1/3 and 1/3, which indicates the use of average pooling. ROIs within the same slide may exhibit considerable variation in the information contained, leading to only ordinary performance when attempting to learn multiscale weights (Supplementary Fig. 17). For instance-level supervision, the top four ROIs with the highest attention scores per slide were selected to compute $\mathcal{L}_{\text{inst}}$. The weighting coefficient α in equation (12) was set to 1 based on the results of parameter tuning (Supplementary Fig. 18). The model parameters were updated by the Adam optimizer with a learning rate of 2×10^{-4} and l_2 weight decay of 1×10^{-5}. The model underwent five-epoch warm-up training at the initial stage of training to ensure stable learning and prevent overfitting. A unified set of parameters was employed to measure the robustness of the model across all tasks in glioma diagnosis. Early stopping was adopted to avoid overfitting. An internal validation set containing 20% of the training data was selected at random to monitor the generalizability. A model was trained using the remaining 80% of the training data for up to 200 epochs, and the training stopped early if the loss on the validation set did not decrease for 20 consecutive epochs. To further reduce random effects in parameter initialization, the training procedure was repeated with different random seeds to obtain a model with high generalizability on the validation set. The number of repeats was set to one by default when the training data were from a single centre and, thus, had high internal consistency, and five otherwise. The influence of hyperparameters on the performance of the model is illustrated in Supplementary Fig. 18.

Ensemble strategy. An ensemble learning strategy was adopted to improve the performance and stability of our method. To construct an ensemble model, five base models, as described above, were trained using the same training data according to the aforementioned procedure. Each model allocated 20% of the training data at random for internal validation, which ensured that there was no overlapping of the validation data among the five models. The prediction results of the five models were then averaged to construct an ensemble model.

Data collection

Two large-scale datasets of digital histopathology WSIs of gliomas were collected.

Xiangya glioma WSI dataset. A dataset of WSIs of gliomas was collected from Xiangya Hospital Central South University. As illustrated in Extended Data Fig. 2b, this dataset consists of 1,109 WSIs consistently magnified at ×40 for the same number of distinct cases. It encompasses diagnostic tasks of glioma detection, subtyping, grading and molecular feature prediction. This dataset has only slide-level annotations, which indicate the subtype and grading of glioma for 530 astrocytoma cases, 224 oligodendroglioma cases and 213 ependymoma cases. Furthermore, molecular testing was performed on 634 IDH mutation cases and 641 MGMT promoter methylation cases. The annotations in this dataset were collectively determined by two subspecialist neuropathologists with over 15 years of clinical experience after careful examination of the pathology images and molecular test results, following the criteria outlined in the 2016 edition of the diagnostic guidelines[34]. To avoid any interference with the experimental results, these two pathologists refrained from participating in any doctor-related experiments mentioned in this article. The exclusion criteria for this dataset are provided in Supplementary Fig. 19. This dataset was partitioned at random into an in-house training dataset containing 736 WSIs for model training and an in-house test dataset containing 373 WSIs for model evaluation and doctor-related experiments. The class proportions in both datasets were the same as those in the whole dataset.

TCGA glioma WSI dataset. Another dataset of histopathology WSIs for glioma was collected from the Brain Lower Grade Glioma and Glioblastoma Multiforme projects. Although there were 860 Glioblastoma Multiforme slides from 389 cases and 844 Brain Lower Grade Glioma slides from 491 cases, the diagnostic criteria were different from those used for the Xiangya dataset. We, therefore, retained only the slide-level annotations in the data and invited the two pathologists who were involved in annotating the Xiangya dataset to review and revise the diagnostic results of these slides based on the 2016 edition of the diagnostic guidelines. As shown in Extended Data Fig. 2c, the final dataset resulting from the review consists of 618 WSIs magnified at ×40 and ×20, encompassing four tasks that align with the Xiangya dataset. This dataset served as the external test dataset for glioma subtyping, grading and molecular feature prediction. The exclusion criteria for this dataset are also presented in Supplementary Fig. 19.

Data processing

The following procedures were employed to deal with WSIs in both the Xiangya and the TCGA datasets.

Segmentation. Tissue regions were segmented from a WSI following the approach used by CLAM. Briefly, ×32 downsampling was applied to a WSI to obtain a low-resolution thumbnail that was easier to process. The thumbnail was then converted from the RGB colour space to the HSV colour space, followed by the application of median blurring on the saturation channel to smooth the image edges. Subsequently, a thresholding operation was applied to the saturation channel of the smoothed image to obtain a binary mask for tissue regions. Morphological closing was performed to fill small holes and gaps in the binary mask. The final mask for tissue regions was obtained by discarding connected components with small areas in the binary mask.

Patching. With the mask of tissue regions obtained, non-overlapping patching was applied to extract regions within the tissue area from a WSI. The hallmark of our method is the segmentation of large patches, namely, ROIs. Specifically, patches of size 4,096 × 4,096 and 2,048 × 2,048 were extracted at magnification levels ×40 and ×20, respectively, to ensure that ROIs had the same resolution at different magnification levels. The coordinates and metadata for WSIs were saved in hdf5 format to enable rapid retrieval of ROIs when extracting visual representations of patches. A WSI can usually be divided into dozens to hundreds of ROIs, depending on the size of the slide.

Feature extraction. A WSI was downsampled with ×2 and ×4 magnifications, resulting in ROIs of size 1,024 × 1,024 and 512 × 512, respectively. Patches of size 256 × 256 were cropped from ROIs at three different magnifications without overlapping, resulting in a total of 64, 16 and four patches at each respective magnification. These patches were then processed through a ResNet50 model pretrained on ImageNet for feature extraction. As a result, multiscale patch features were obtained with a shape of $84 \times d$ for each WSI, where d represents the dimension of the extracted features. Note that using ResNet50 pretrained on ImageNet for comparisons between ROAM and the baseline methods may not be optimal in the context of pathology images, because ImageNet mainly contains natural images, and thus, the pretrained ResNet50 model may not be suitable for extracting visual features of pathology images. We took this into consideration and explored self-supervised feature extraction models specifically designed for pathology images[35–38]. The results, as shown in Supplementary Fig. 20, indicate that these models are capable of extracting valuable features compared to models pretrained on natural image datasets, but do not show sufficient superiority.

Visualization

The following two visualization procedures were employed at the slide and ROI levels, respectively.

Slide-level visualization. Attention scores of specific classes for an ROI were visualized so that we could gain intuitive insights into the contribution of a tissue region within the ROI to the predicted label of a slide. To achieve this objective, attention scores of ROIs were computed from the gated attention network during the inference using the model, and the scores were converted into RGB colours using a diverging colour map and applied to the corresponding positions in the slide to provide an intuitive interpretation of the mechanism of model prediction. In brief, regions with high attention scores were coloured in red, and regions with low attention scores were coloured in blue. Consequently, an WSI was partitioned into 2,048 × 2,048 ROIs with overlapping to obtain a fine-grained visualization heat map, which was overlaid on the original WSI with a transparency of 0.4 to facilitate the observation of morphological histopathological structures in different stained regions. Given the considerable size of the ROI regions, the difference in the generated heat maps with an overlap of 0.5 or higher, after applying boundary smoothing, was minimal (Supplementary Fig. 21). Subsequently, an overlapping threshold of 0.5 was applied to all subsequent auxiliary diagnoses to improve efficiency.

ROI-level visualization. Slide-level visualization provides a general insight into the contribution of tissue regions to the diagnosis of a slide. However, due to the large size of the ROI regions, the precision of a heat map is limited, as it can hardly capture the contribution of specific regions within an ROI to the diagnosis based on attention scores alone. Using the inherent properties of the SA mechanism, a method called Grad-Rollout, which is based on layer-wise relevance propagation in transformer[27,39], was employed to interpret the ROI-level visualization. The first row of an SA matrix represents precisely the attention allocated by the class token to each region within the ROI. To obtain the attention score of a region within an ROI to a specific category, a gradient matrix was obtained by computing the cross-entropy loss between the outputs of the model and the labels of the target class. Element-wise multiplication between the SA matrix of each layer and the corresponding gradient matrix was then performed to generate class-specific SA scores. Attention masks were generated at the three magnifications using reshaping and scaling operations. These masks were then combined by weighted summation using weight coefficients corresponding to each scale used during model training. Using the ROI-level visualization, some ROIs with high attention scores were visualized to gain insights into which structures or morphologies within the ROI contributed critically to its high attention scores, as these had a greater impact on the predictions of the target class. Additionally, other visualization methods designed for transformers based on Grad-CAM[40] were also tried. However, these methods were not as stable or as intuitive as the method based on Grad-Rollout. More details are provided in Supplementary Fig. 22. Our implementation of the ROI-level visualization is outlined in Supplementary Note 6.

Baseline methods

ROAM was compared with five computational pathology methods based on MIL. CLAM[18] and TransMIL[19] are two weakly supervised learning methods that perform well in the classification of histopathological images and have been used as baseline methods in previous studies[19,21]. CLAM enhances attention-based MIL[31] by constructing an attention network along with K independent classifiers to predict probabilities that a WSI is classified into the K categories. Besides slide-level supervision, this method further introduces a mechanism called instance-level clustering that assigns pseudo-labels to patches. TransMIL adopts a framework based on transformer[29] to obtain feature embeddings, and then it aggregates instances by considering their correlations to address the problem of overlooking the relevance between instances that has occurred in previous studies. Both methods adopt small image patches (224 × 224 or 256 × 256) as instances and, thus, lack the power to extract large-scale features that extend beyond the field of view of a single patch. H^2MIL (ref. 24), GTP[25] and TEA-graph[26] are three representative approaches based on graph neural networks. These methods were designed to enhance the receptive field of small image patches and offer another solution to the issue of insufficient information in small image blocks. H^2MIL (ref. 24) utilizes patches at different magnification levels to construct a pyramid graph spanning different scales and spatial positions, thus enabling the effective learning of hierarchical representations of pathological images. Nevertheless, the low resolution of the patches used in graph construction hinders the learning of small features at the cellular level, and the thumbnail features offer limited assistance in extracting features at large scales in pathological images. GTP[25] combines graph neural networks with transformers to more effectively leverage the spatial connectivity features of image patches. TEA-graph[26] abstracts adjacent image patches with similar

features into a super-patch and constructs a graph based on super-patches, thereby effectively increasing the receptive field of small patches and enabling the learning of large-scale texture features. However, both methods are less effective in learning large-scale homogeneous pathological features, although they are able to learn certain large-scale morphological texture features by enlarging their receptive fields.

These baseline methods were trained with the default parameters provided in their original papers. For CLAM and TransMIL, we used patches of size 256 × 256 from the ROIs utilized for ROAM instead of directly patching on the original WSIs. Although this modification may cause the loss of some patches at the edges of tissue regions, the results for CLAM and TransMIL, as shown in Supplementary Tables 4 and 5, indicate that the impact on the performance of these models was negligible. For the other three graph-based methods, we adhered strictly to the procedures provided in the respective papers for graph construction.

Hardware and software

Tissue segmentation, ROI extraction, feature extraction and visualization of results were conducted on a workstation with an Intel Core i9-12900K CPU, NVIDIA RTX A6000 GPU and 64 GB memory. Model evaluation experiments were conducted on a server with NVIDIA RTX 3090 GPUs. The WSI processing pipeline was implemented in Python (v.3.8.13). Several image-processing packages were used to support the preprocessing pipeline, including openslide (v.3.4.1), opencv (v.4.5.5) and pillow (v.6.2.1). Vahadane's method[41] was used for stain normalization. The deep-learning library PyTorch (v.1.12.1) was used for data loading and model training. Scikit-learn (v.1.0.2) was used to calculate the AUC. 95% confidence intervals of the test AUC in clinical practice were calculated using bootstrapping with nonparametric, unstratified resampling of 1,000 iterations[42]. Matplotlib (v.3.5.2) and seaborn (v.0.11.2) were used to generate plots. The online diagnostic platform was developed by the Fuzhou Institute for Data Technology. Pathologists can use this platform to view and annotate pathological slides, make diagnoses, refine diagnoses based on AI predictions and analyse the diagnostic basis given by our model, thereby drawing new insights from these processes.

Ethics statement

The study was approved by the Medical Ethics Committee of Xiangya Hospital, Central South University, under protocol number 202310205. Informed consent was waived for this retrospective study and participants were not compensated.

Reporting summary

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

Data availability

The TCGA public glioma WSI dataset is available from the National Institutes of Health's genomic data commons (https://portal.gdc.cancer.gov). Detailed information about the dataset, including sample IDs and labels, can be obtained from Zenodo via https://doi.org/10.5281/zenodo.11469546 (ref. 43). The Xiangya glioma WSI dataset is not publicly available in accordance with institutional requirements governing the protection of the privacy of human subjects. The in-house Xiangya dataset is not publicly available due to privacy concerns regarding patient information.

Code availability

The ROAM project, including detailed documents and instructions, is available on GitHub (https://github.com/whiteyunjie/ROAM)[44]. The source code is also available on Zenodo via https://doi.org/10.5281/zenodo.11469423 (ref. 45).

References

1. Ostrom, Q. T. et al. The epidemiology of glioma in adults: a 'state of the science' review. *Neuro-oncology* **16**, 896–913 (2014).
2. Weller, M. et al. Glioma. *Nat. Rev. Dis. Primers* **1**, 15017 (2015).
3. Chen, R., Smith-Cohn, M., Cohen, A. L. & Colman, H. Glioma subclassifications and their clinical significance. *Neurotherapeutics* **14**, 284–297 (2017).
4. Louis, D. N. et al. The 2021 WHO classification of tumors of the central nervous system: a summary. *Neuro-oncology* **23**, 1231–1251 (2021).
5. Horbinski, C. et al. NCCN guidelines® insights: central nervous system cancers, version 2.2022: featured updates to the NCCN guidelines. *J. Natl Compr. Cancer Netw.* **21**, 12–20 (2023).
6. Van Den Bent, M. J. Interobserver variation of the histopathological diagnosis in clinical trials on glioma: a clinician's perspective. *Acta Neuropathol.* **120**, 297–304 (2010).
7. Ker, J., Bai, Y. Q., Lee, H. Y., Rao, J. & Wang, L. P. Automated brain histology classification using machine learning. *J. Clin. Neurosci.* **66**, 239–245 (2019).
8. Bera, K., Schalper, K. A., Rimm, D. L., Velcheti, V. & Madabhushi, A. Artificial intelligence in digital pathology – new tools for diagnosis and precision oncology. *Nat. Rev. Clin. Oncol.* **16**, 703–715 (2019).
9. Niazi, M. K. K., Parwani, A. V. & Gurcan, M. N. Digital pathology and artificial intelligence. *Lancet Oncol.* **20**, e253–e261 (2019).
10. Echle, A. et al. Deep learning in cancer pathology: a new generation of clinical biomarkers. *Br. J. Cancer* **124**, 686–696 (2021).
11. Mobadersany, P. et al. Predicting cancer outcomes from histology and genomics using convolutional networks. *Proc. Natl Acad. Sci. USA* **115**, E2970–E2979 (2018).
12. Ertosun, M. G. & Rubin, D. L. Automated grading of gliomas using deep learning in digital pathology images: a modular approach with ensemble of convolutional neural networks. *AMIA Annu. Symp. Proc. Arch.* **2015**, 1899–1908 (2015).
13. Jin, L. et al. Artificial intelligence neuropathologist for glioma classification using deep learning on hematoxylin and eosin stained slide images and molecular markers. *Neuro-oncology* **23**, 44–52 (2021).
14. Im, S. et al. Classification of diffuse glioma subtype from clinical-grade pathological images using deep transfer learning. *Sensors* **21**, 3500 (2021).
15. Ocampo, P. et al. Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning. *J. Thorac. Oncol.* **13**, S562 (2018).
16. Hsu, W. W. et al. A weakly supervised deep learninÿg-based method for glioma subtype classification using WSI and mpMRIs. *Sci. Rep.* **12**, 6111 (2022).
17. Campanella, G. et al. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nat. Med.* **25**, 1301–1309 (2019).
18. Lu, M. Y. et al. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nat. Biomed. Eng.* **5**, 555–570 (2021).
19. Shao, Z. C. et al. TransMIL: transformer based correlated multiple instance learning for whole slide image classification. In *Proc. 35th Conference on Neural Information Processing Systems* (eds Ranzato, M. et al.) 2136–2147 (Curran Associates, 2021).
20. Li, B., Li, Y. & Eliceiri, K. W. Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 14318–14323 (IEEE, 2021).
21. Zhang, H. R. et al. DTFD-MIL: double-tier feature distillation multiple instance learning for histopathology whole slide image classification. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 18802–18812 (IEEE, 2022).

22. Zhang, J. et al. Attention multiple instance learning with transformer aggregation for breast cancer whole slide image classification. In *Proc. 2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 1804–1809 (IEEE, 2022).
23. Chen, R. J. et al. Scaling vision transformers to gigapixel images via hierarchical self-supervised learning. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition* 16144–16155 (IEEE, 2022).
24. Hou, W. et al. H^2-MIL: exploring hierarchical representation with heterogeneous multiple instance learning for whole slide image analysis. In *Proc. AAAI Conference on Artificial Intelligence* 933–941 (AAAI Press, 2022).
25. Zheng, Y. et al. A graph-transformer for whole slide image classification. *IEEE Trans. Med. Imaging* **41**, 3003–3015 (2022).
26. Lee, Y. et al. Derivation of prognostic contextual histopathological features from whole-slide images of tumours via graph deep learning. *Nat. Biomed. Eng.* https://doi.org/10.1038/s41551-022-00923-0 (2022).
27. Chefer, H., Gur, S. & Wolf, L. Transformer interpretability beyond attention visualization. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 782–791 (IEEE, 2021).
28. Dosovitskiy, A. et al. An image is worth 16×16 words: transformers for image recognition at scale. Preprint at https://arxiv.org/abs/2010.11929 (2020).
29. Vaswani, A. et al. Attention is all you need. In *Proc. 30th Conference on Neural Information Processing Systems* 5998–6008 (eds Guyon, I. et al.) (Curran Associates, 2017).
30. Liu, Z. et al. Swin transformer: hierarchical vision transformer using shifted windows. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 10012–10022 (IEEE, 2021).
31. Ilse, M., Tomczak, J. & Welling, M. Attention-based deep multiple instance learning. In *Proc. 35th International Conference on Machine Learning* (eds Dy, J. & Krause, A.) 2127–2136 (PMLR, 2018).
32. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition* 770–778 (IEEE, 2016).
33. Russakovsky, O. et al. Imagenet large scale visual recognition challenge. *Int. J. Comput. Vis.* **115**, 211–252 (2015).
34. Louis, D. N. et al. The 2016 World Health Organization classification of tumors of the central nervous system: a summary. *Acta Neuropathol.* **131**, 803–820 (2016).
35. Wang, X. et al. RetCCL: clustering-guided contrastive learning for whole-slide image retrieval. *Med. Image Anal.* **83**, 102645 (2023).
36. Yang, P. et al. CS-CO: a hybrid self-supervised visual representation learning method for H&E-stained histopathological images. *Med. Image Anal.* **81**, 102539 (2022).
37. Wang, X. et al. Transformer-based unsupervised contrastive learning for histopathological image classification. *Med. Image Anal.* **81**, 102559 (2022).
38. Ciga, O., Xu, T. & Martel, A. L. Self supervised contrastive learning for digital histopathology. *Mach. Learn. Appl.* **7**, 100198 (2022).
39. Bach, S. et al. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLoS ONE* **10**, e0130140 (2015).
40. Selvaraju, R. R. et al. Grad-cam: visual explanations from deep networks via gradient-based localization. In *Proc. IEEE Conference on Computer Vision* 618–626 (IEEE, 2017).
41. Vahadane, A. et al. Structure-preserving color normalization and sparse stain separation for histological images. *IEEE Trans. Med. Imaging* **35**, 1962–1971 (2016).
42. Carpenter, J. & Bithell, J. Bootstrap confidence intervals: when, which, what? A practical guide for medical statisticians. *Stat. Med.* **19**, 1141–1164 (2000).
43. Yin, X. TCGA glioma WSI dataset used for evaluating ROAM. *Zenodo* https://doi.org/10.5281/zenodo.11469546 (2024).
44. Yin, X. ROAM. *GitHub* https://github.com/whiteyunjie/ROAM (2024).
45. Yin, X. A transformer-based weakly supervised computational pathology method for clinical-grade diagnosis and molecular state revelation of gliomas (v1.0.0). *Zenodo* https://doi.org/10.5281/zenodo.11469423 (2024).

Acknowledgements

This work was supported by the National Key Research and Development Program of China (Grant Nos. 2023YFF1204802 and 2021YFF1200902 to R.J. and 2022YFF1202403 to H.L.) and the National Natural Science Foundation of China (Grant Nos. 62273194 to R.J. and 62250005 to X.Z.).

Author contributions

R.J., X.Z., Z.H. and H.L. conceived the study and supervised the research. R.J., X.Y. and P.Y. designed, implemented and validated the ROAM project. Y.W., J.H., J.Y., Z.H. and L.C. collected, curated and annotated the data and helped with analysing the results. F.C. provided technical support for the online platform. X.F., L.S., L.L., W.L., Z.H., L.C., Y.W. and H.Z. participated in the clinical-grade assessment of ROAM and the auxiliary diagnosis experiments. Z.H. and L.C. participated in the summarization and discovery of molecular and morphological biomarkers. R.J., X.Y., P.Y. and H.L. wrote the paper. L.C., R.J., X.Z., Z.H. and H.L. provided valuable comments on the paper.

Competing interests

The authors declare no competing interests.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s42256-024-00868-w.

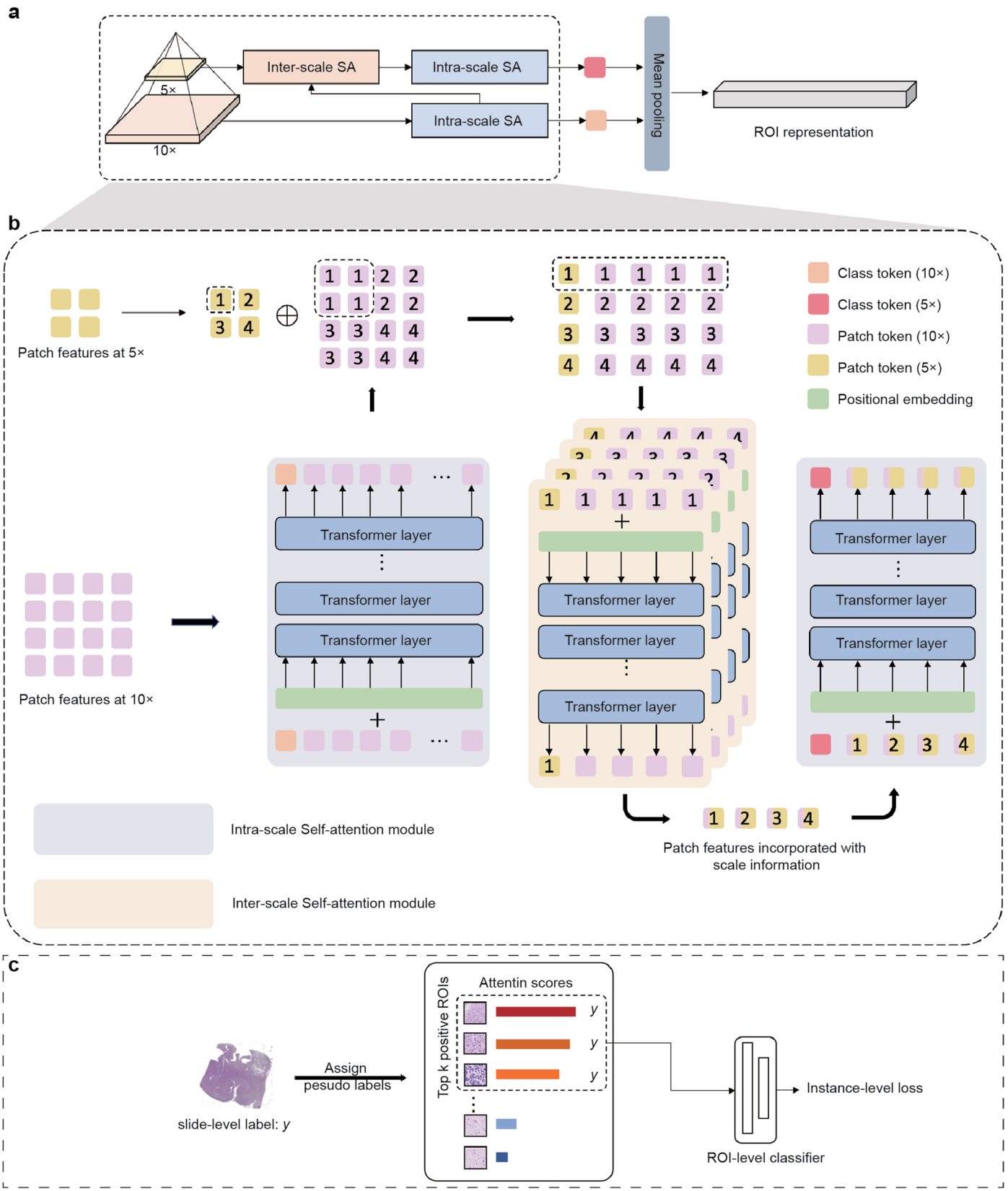
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s42256-024-00868-w.

Correspondence and requests for materials should be addressed to Xuegong Zhang, Zhongliang Hu or Hairong Lv.

Peer review information *Nature Machine Intelligence* thanks the anonymous reviewers for their contribution to the peer review of this work.

[1]Ministry of Education Key Laboratory of Bioinformatics, Bioinformatics Division at the Beijing National Research Center for Information Science and Technology, Center for Synthetic and Systems Biology, Department of Automation, Tsinghua University, Beijing, China. [2]JIUTIAN AI Center, China Mobile Research, Beijing, China. [3]Xiangya Hospital Central South University, Changsha, China. [4]Xiangya School of Medicine Central South University, Changsha, China. [5]Department of Pathology, Changsha Medical University, Changsha, China. [6]Department of Pathology, Yueyang Central Hospital, Yueyang, China. [7]Fuzhou Institute for Data Technology, Fuzhou, China. [8]These authors contributed equally: Rui Jiang, Xiaoxu Yin, Pengshuai Yang, Lingchao Cheng.
✉e-mail: zhangxg@mail.tsinghua.edu.cn; huzhongliang@csu.edu.cn; lvhairong@tsinghua.edu.cn

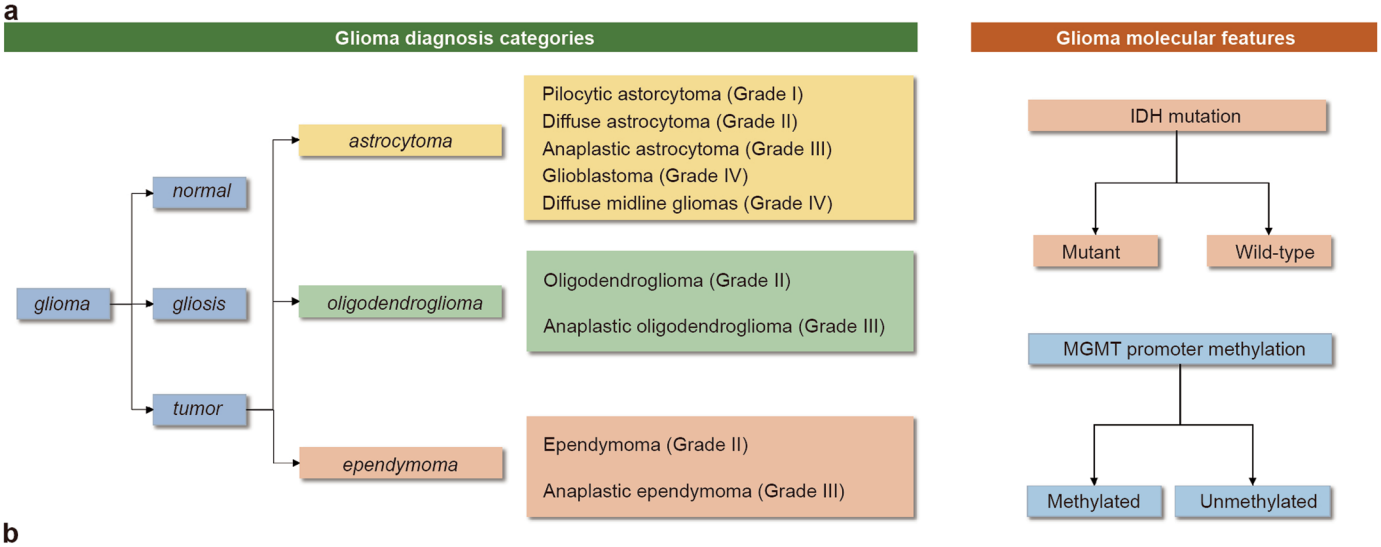
https://doi.org/10.1038/s42256-024-00868-w



Extended Data Fig. 1 | Workflow of the Pyramid Transformer and ROI-level supervision. a-b, the implementation process of the multi-scale self-attention module that incorporates visual features of ROI at 10× with features of ROI at 5×. Patch tokens with the same numerical ID represent that their corresponding tissue regions locate at the identity location within ROI, differing only in their magnification scales. **c**, the process of instance-level supervision. The top *k* ROIs with the highest attention scores are assigned the same label as their corresponding slide. Classification is performed using an ROI-level classifier, and the instance-level loss is combined with the slice-level loss to train the entire model.

https://doi.org/10.1038/s42256-024-00868-w

a



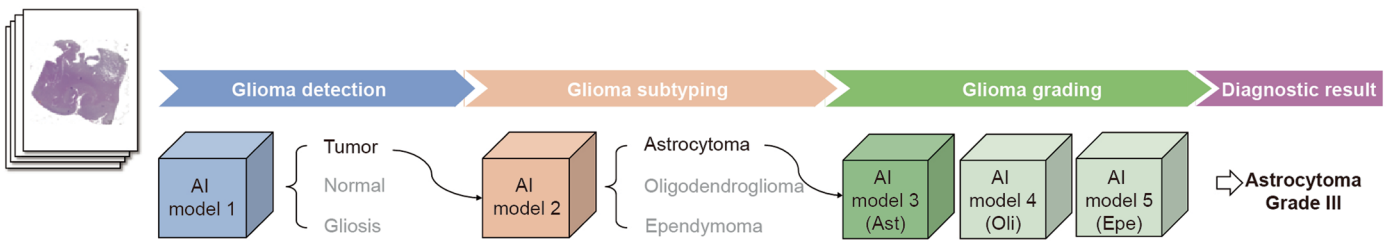
b

Task	Slides	Categories	Slides per class	Scan magnification
Glioma detection	1,109	Normal/Gliosis/Tumor	95/47/967	
Glioma subtyping	967	Astrocytoma/Oligodendroglioma/Ependymoma	530/224/213	
Astrocytoma grading	530	Grade I/II/III/IV/IV(Diffuse Midline Glioma)	101/106/100/121/102	
Oligodendroglioma grading	224	Grade II/III	114/110	40×
Ependymoma grading	213	Grade II/III	110/103	
IDH mutation prediction	634	Mutant/Wild-type	275/359	
MGMT promoter methylation prediction	641	Methylated/Unmethylated	284/357	
Total	1,109			

c

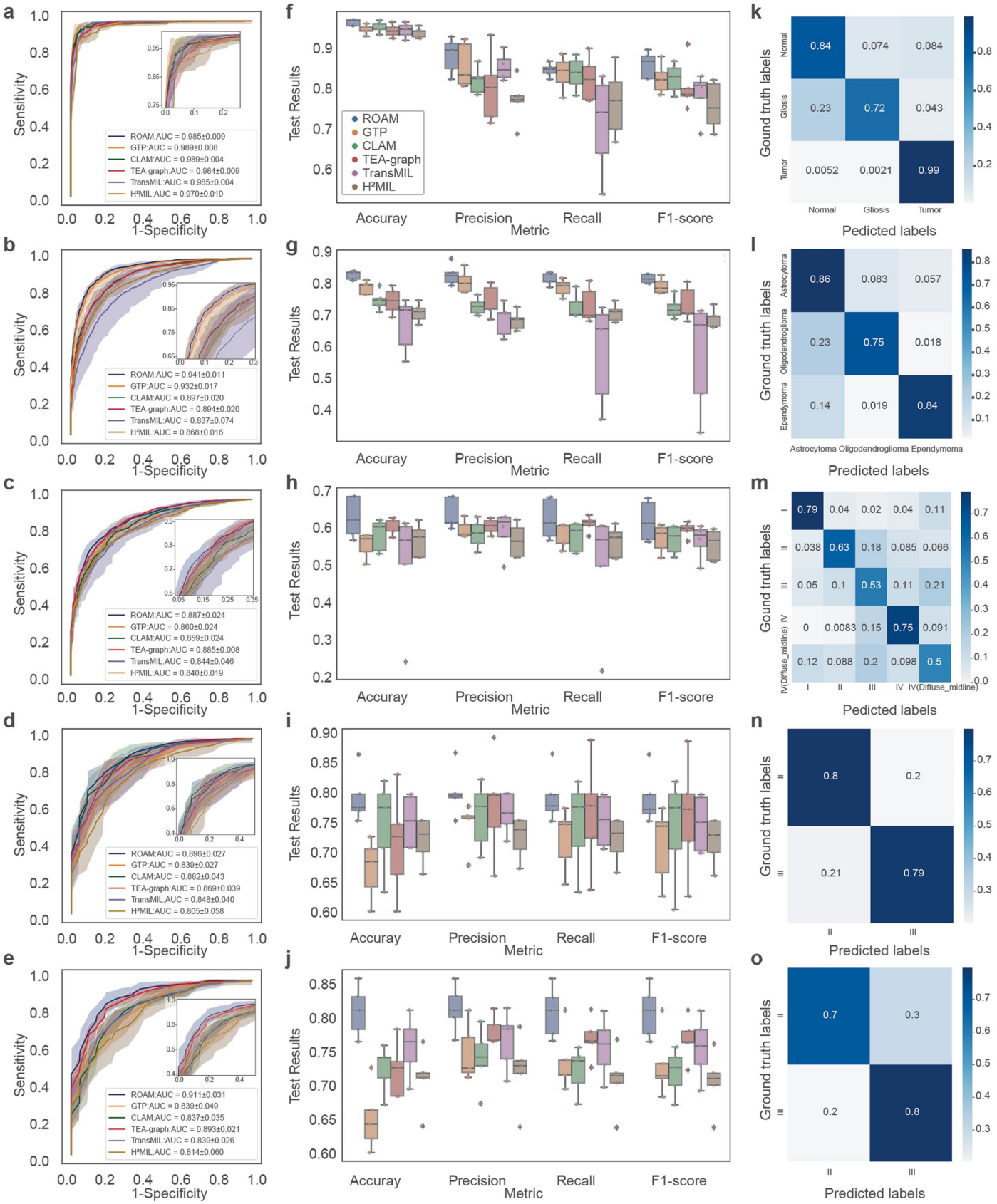
Task	Slides	Categories	Slides per class	Scan magnification
External glioma subtyping	618	Astrocytom/Oligodendroglioma	462/156	
External Oligodendroglioma grading	156	Grade II/III	88/68	40×/20×
External IDH mutation prediction	618	Mutant/Wild-type	251/367	
External MGMT promoter methylation prediction	618	Methylated/Unmethylated	169/449	
Total	618			

d



Extended Data Fig. 2 | Overview of the glioma data and the cascade diagnostic system. a, Detailed categories and crucial molecular features of glioma. Glioma can be broadly classified into normal, gliosis and tumor. The subtypes of glioma include astrocytoma, oligodendroglioma, and ependymoma, each having various grades. IDH mutation, and MGMT promoter methylation are two of the most critical molecular features associated with glioma diagnosis. **b**, Information of the in-house Xiangya glioma dataset. The dataset consists of a total of 1109 slides and supports 7 distinct classification tasks based on the provided labels. **c**, Details of the external TCGA glioma dataset. The dataset consists of 618 slides and supports 4 external classification tasks. **d**, The cascade diagnostic system of glioma based on ROAM. For instance, in the case of astrocytoma, the final diagnosis involves a sequential process utilizing 3 diagnostic models: model 1, model 2 and model 3.

https://doi.org/10.1038/s42256-024-00868-w

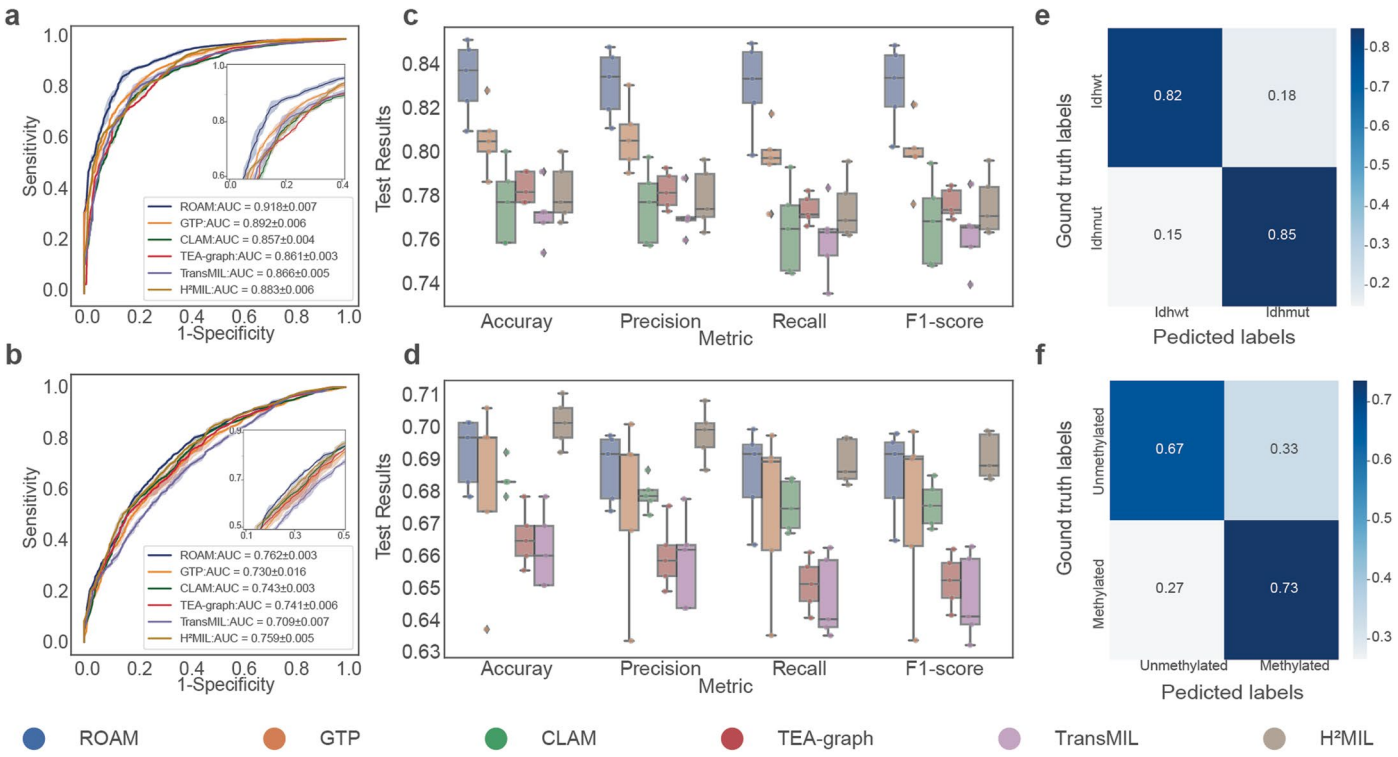


Extended Data Fig. 3 | See next page for caption.

https://doi.org/10.1038/s42256-024-00868-w

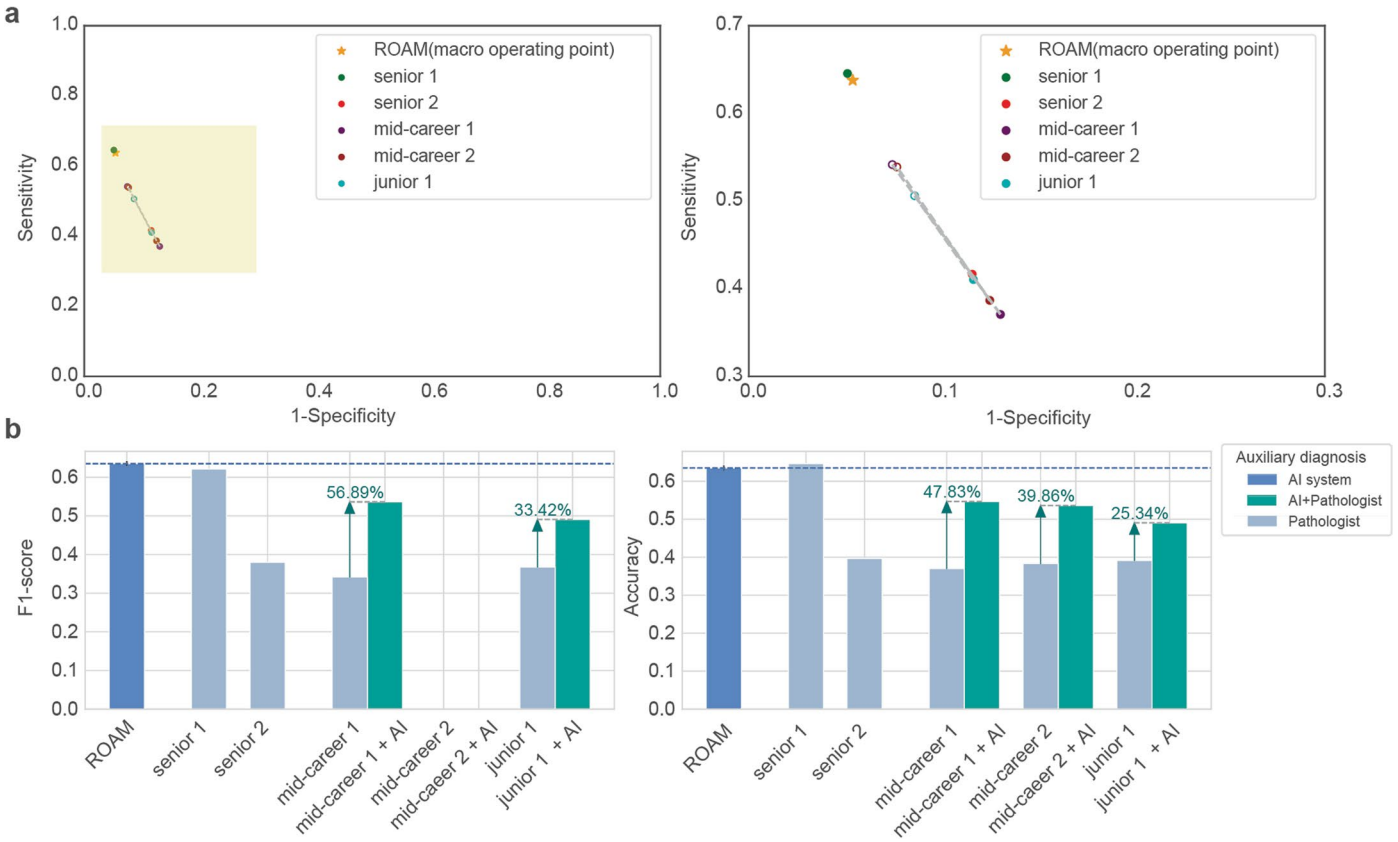
Extended Data Fig. 3 | Cross-validation of ROAM and baseline methods in glioma diagnosis without the use of the ensemble strategy. Tasks of glioma diagnosis include glioma detection (**a**,**f**,**k**, n = 222), glioma subtyping (**b**,**g**,**l**, n = 193), astrocytoma grading (**c**,**h**,**m**, n = 106), oligodendroglioma grading (**d**,**i**,**n**, n = 45), and ependymoma grading (**e**,**j**,**o**, n = 43). **a**-**o**, Results are derived from 5-fold cross-validation experiments on the Xiangya dataset. Models are trained without the use of the ensemble strategy. **a**-**e**, Receiver operating characteristic (ROC) curves and the corresponding area under the curves (AUC±s.d). The confidence bond shows ± 1 s.d for a curve. Inserts: zoomed-in view of the curves. **f**-**j**, Averaged Accuracy, macro recall, macro precision, and macro F1-score. The metrics are plotted using box plot and each box ranges from the upper to lower quartiles with the median as the horizontal line, whiskers extend to 1.5 times the interquartile range, and diamond points represent outliers. All the legends are consistent with the legend in **f**. **k**-**o**, Mean normalized confusion matrices of ROAM.

https://doi.org/10.1038/s42256-024-00868-w



Extended Data Fig. 4 | Performance of ROAM and baseline methods in the prediction of molecular features. Tasks include the prediction of IDH mutation (**a**, **c**, **e**, n = 216) and that of MGMT promoter methylation (**b**, **d**, **f**, n = 218). **a**-**f**, Results are derived from five ensemble models that are trained using the in-house training dataset and tested on the in-house test dataset, both split from the Xiangya dataset at the ratio of 2:1. **a**-**b**, Receiver operating characteristic (ROC) curves and the corresponding area under the curves (AUC±s.d). The confidence bond shows ±1 s.d for a curve. Inserts: zoomed-in view of the curves. **c**-**d**, Accuracy, macro recall, macro precision, and macro F1-score. The metrics are plotted using box plot and each box ranges from the upper to lower quartiles with the median as the horizontal line, whiskers extend to 1.5 times the interquartile range, and diamond points represent outliers. **e**-**f**, Mean normalized confusion matrices of ROAM.

https://doi.org/10.1038/s42256-024-00868-w



Extended Data Fig. 5 | Clinical-grade performance in the overall cascade diagnostic task. a, Performance of the cascade diagnostic system and the five pathologists for the overall glioma diagnosis task. Macro-averaged one-versus-rest operating points are plotted. There is no ROC curve because the prediction for the task is integrated based on the outcomes of multiple sub-tasks. **b**, Macro-averaged F1-score and accuracy of ROAM and pathologists (before and after AI assistance) for the overall glioma diagnosis task. The F1-score of mid-career 2 not being displayed implies that this pathologist made completely incorrect predictions for at least one category during the diagnosis.

nature portfolio

Corresponding author(s): Hairong Lv, Zhongliang Hu, Xuegong Zhang

Last updated by author(s): Jun 8, 2024

Reporting Summary

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

Statistics

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

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

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

Software and code

Policy information about availability of computer code

Data collection	In-house glass slides were digitized using the KFBIO KF-PRO-005 scanner at a resolution of 0.25 microns per pixel, and were accessed through openslide (3.4.1). Code for data preprocessing was implemented in Python (3.8.13), and is publicly available at https://github.com/whiteyunjie/ROAM.
Data analysis	The implementation of the project for model training and validating is publicly available at https://github.com/whiteyunjie/ROAM. We used Python (3.9.0) and Pytorch (1.12.1) for deep learning. These additional Python libraries were also used for model training, statistical analysis: pillow (6.2.1), numpy (1.23.1), pandas (1.4.3), scikit-learn (1.0.2), matplotlib (3.5.2) and seaborn (0.11.2).

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

Data

Policy information about availability of data

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

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

The TCGA public glioma WSI dataset data is available at the NIH genomic data commons (https://portal.gdc.cancer.gov) and detailed information about the dataset, including sample IDs and labels, can be obtained from Zenodo at https://doi.org/10.5281/zenodo.11469546. The Xiangya glioma WSI dataset is not publicly available in accordance to institutional requirements governing human subject privacy protections.

Human research participants

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

Reporting on sex and gender	Patient sex was not considered in the study design. No sex-based a priori analysis was done. The term sex and gender have not been used in interchangeably in the study. For in-house Xiangya glioma WSI dataset, we provide the aggregate distribution of self-reported sex as follows: 500 Female, 609 Male. For external TCGA public glioma WSI dataset, the distribution is as follows: 242 Female, 324 Male, 52 Unspecified.
Population characteristics	In-house xiangya glioma data: All patient cases were collected from Xiangya Hospital, Central South University, from 2019 to 2022, and the test data was randomly sampled from this patient population. The dataset includes 204 pediatric glioma cases (≤14 years) and 905 adolescent and adult glioma cases (>14 years). TCGA external validation data: The data were collected from a diverse population representing multiple hospitals, including 1 pediatric glioma case (≤14 years), 565 adolescent and adult glioma cases (>14 years), and 52 cases with unknown age.
Recruitment	No patient recruitment was necessary for the use of histology whole-slide images retrospectively.
Ethics oversight	The Medical Ethics Committee of Xiangya Hospital Central South University approved the study (Approval 202310205).

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

Field-specific reporting

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

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

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

Life sciences study design

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

Sample size	No statistical methods were used to determine the sample size. We used all available data for model training and evaluation, with the sample size determined solely by the number of high-quality pathological slides. We obtained these high-quality slides by excluding samples with missing content or incomplete diagnostic results. Details of both datasets in our study were implemented in the manuscript (Extended Data Fig 2).
Data exclusions	Data with equviocal diagnosis and low image quality are excluded. The details are implemented in Supplementary Figure 19.
Replication	We provided the complete code for replication. There may be slight variations due to hardware differences and non-determinism in GPU-accelerated code. We believe that these deviations are minimal and can be disregarded in terms of the experimental conclusions. Our code provided trained models and test example results was successfully replicated.
Randomization	Data was randomly divided into training, validation and test sets for in-house dataset. No other covariates were controlled for.
Blinding	For the human-AI diagnosis comparison and auxiliary diagnosis experiments, pathologists who participated in the diagnosis were blinded to group allocation during data analysis. For other experiments involving only the deep learning models, blinding was unnecessary as the analysis was based solely on the objective assessment of digitized pathology slide images using the deep learning models.

Reporting for specific materials, systems and methods

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

Materials & experimental systems

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

Methods

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