Cross-Modal Deep Learning for Integrated Molecular Diagnosis of Cancer: A Prospective Multicenter Study

Xiaofei Wang[1*]; Yinyan Wang[2*]; Wanming Hu[3*]; Mayen Briggs[4]; Zeya Yan[2]; Jie Hu[2]; Yupei Zhang[1]; Hao Duan[5]; Stephen Price[1]; Chao Li[1678^]

1. Department of Clinical Neurosciences, University of Cambridge, UK

2. Department of Neurosurgery, Beijing Tiantan Hospital, Capital Medical University, Beijing, China

3. Department of Pathology, State Key Laboratory of Oncology in South China, Guangdong Provincial Clinical Research Center for Cancer, Sun Yat-Sen University Cancer Center, Guangzhou, China

4. Cambridge University Hospitals NHS Foundation Trust

5. Department of Neurosurgery, State Key Laboratory of Oncology in South China, Guangdong Provincial Clinical Research Center for Cancer, Sun Yat-sen University Cancer Center, Guangzhou, China

6. Department of Applied Mathematics and Theoretical Physics, University of Cambridge, UK

7. School of Science and Engineering, University of Dundee, UK

8. School of Medicine, University of Dundee, UK

*equal contributions ^corresponding author (cl647@cam.ac.uk)

Abstract:

Accurate integration of histological and molecular features is central to modern cancer diagnostics, but it is often hampered by extended turnaround times and resource-intensive parallel workflows, which can delay diagnosis and introduce variability between pathologists. We present CAMPaS (Cross-modal AI for Integrated Molecular Pathology Diagnosis and Stratification), a trustworthy AI framework that jointly predicts glioma histology, molecular markers, and WHO 2021 integrative diagnoses from hematoxylin and eosin-stained slides. CAMPaS introduces a novel architecture purposefully designed to address challenges in real-world translation. Trained on 3,367 patients (6,043 slides) across eight cohorts (six retrospective, two prospective), CAMPaS achieved high diagnostic accuracy (AUC 0.895-0.916 in training; 0.946-0.955 in prospective cohorts) and generalized robustly across diverse settings. Its interpretable cross-modal predictions aligned with expert annotations and genomic profiles, revealing biologically coherent features. Crucially, CAMPaS stratifies prognosis and treatment response, offering a scalable and biologically grounded solution to accelerate precision oncology in routine care.

Introduction

Accurate and timely cancer diagnosis is crucial for guiding treatment and improving patient outcomes. While pathology remains the diagnostic gold standard, traditional histopathology requires substantial expertise and is subject to interobserver variability[1]. Over the past decade, molecular profiling has revolutionized cancer diagnosis[2]. The 2021 WHO classification of central nervous system tumors reflects this paradigm shift, embedding key molecular markers into diagnostic criteria[3]. For instance, glioblastoma is now diagnosed solely based on *IDH* mutation status, rather than histological features[3–5]. Nonetheless, histology remains important. A category of grade 4 astrocytoma is newly defined by *IDH* mutations, *1p/19q* co-deletion (CODEL) and *CDKN2A/B* homozygous deletion (HOMDEL), alongside necrosis and microvascular proliferation.

While this integrative approach enhances diagnostic precision, it also complicates routine clinical workflows[6]. First, molecular testing is costly and may be infeasible due to limited tissue availability[7]. Second, histological assessment and molecular testing follow separate, time-consuming pipelines, hindering timely integration for final diagnosis. Third, incorporating molecular markers presents challenges for practitioners trained primarily in morphology, increasing the risk of diagnostic discrepancies among different experience levels[8].

Artificial Intelligence (AI) has enabled the prediction of molecular markers directly from hematoxylin and eosin (H&E) stained slides, bridging histological and molecular diagnostics. However, current AI approaches[10-13] typically treat molecular and histological prediction as separate tasks, overlooking their biological interaction. Additionally, most models are trained on small, homogeneous cohorts, limiting their accuracy and robustness. This limitation is particularly evident in heterogeneous cancers like glioma, where rarer but clinically critical subtypes, such as the newly classified Grade 4 astrocytoma, are often underrepresented.

Emerging studies show the feasibility of AI modelling the complex relationships between histological and molecular features, offering the potential for more accurate integrated diagnoses[13,14]. More importantly, real-world challenges, including heterogeneous

datasets and missing molecular data, often compromise model robustness. Although AI models can generate spatial maps and surrogate molecular signatures to support pathologists in interpreting tissue holistically[9,14], few have rigorously validated their pathological or biological coherence, particularly in the context of multimodal integration.

Here, we introduce CAMPaS (Cross-modal AI for integrated Molecular Pathology Diagnosis and Stratification), a trustworthy AI framework tailored for real-world cancer diagnosis. To ensure *accuracy*, *robustness,* and *interpretability*, our **primary** objective is to develop our previously benchmarking-validated model that jointly models molecular markers and histology[13,14], specifically addressing real-world challenges. Our **secondary** objective is to rigorously validate the model using a large-scale dataset of 3,367 glioma patients (6,043 slides) from eight cohorts, including two prospective studies, demonstrating strong accuracy and robust generalization. CAMPaS shows interpretability at histological, molecular and cross-modal levels, aligning with established knowledge and contributing to a deeper understanding of glioma. Finally, CAMPaS demonstrates clinical relevance by predicting patient prognosis and treatment benefit beyond conventional metrics.

Together, our results underscore the translational potential of a trustworthy AI framework, purposefully designed and rigorously validated, to aid cancer diagnosis under molecular pathology criteria and support precision oncology.

Results

2.1 Study design

We extended our previously validated models[13,14] that jointly predict molecular markers and histology, with targeted innovations addressing real-world challenges, focusing on enhancing clinical translation (Fig. 1a-c). In designing and evaluating the proposed CAMPaS (Cross-modal AI for Integrated Molecular Pathology Diagnosis and Stratification), we followed the principles of trustworthy AI, emphasizing *accuracy*, *robustness*, *interpretability*, and *prognostic efficacy*. Specifically, CAMPaS was trained and validated on six retrospective and two prospective cohorts, with interpretability rigorously assessed at histological, molecular, and cross-modal levels using expert annotations and spatial transcriptomics. Further, to establish clinical utility, we performed survival analysis in the retrospective and prospective cohorts (Fig. 1d).

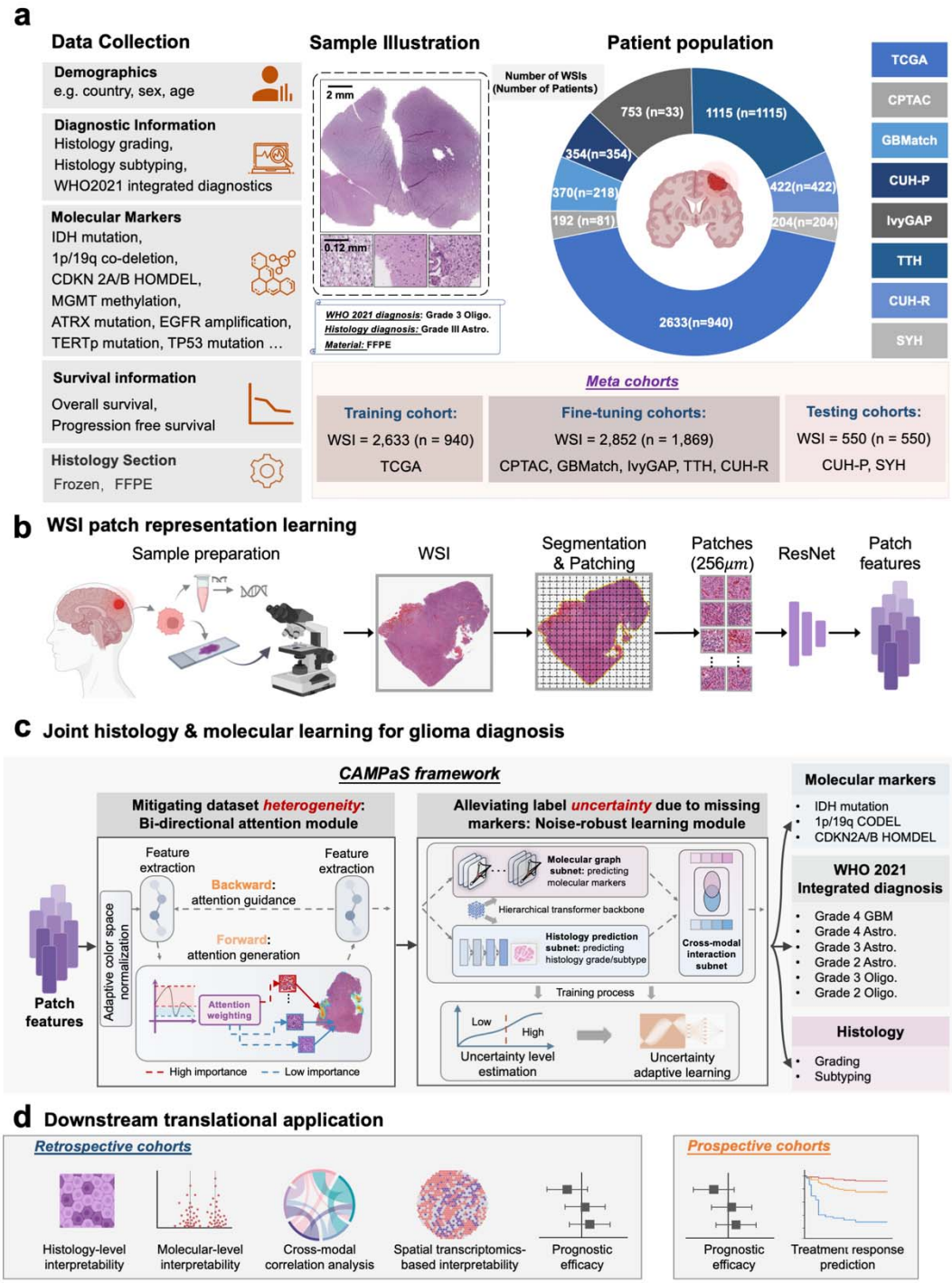


Fig. 1 | Data, model design and validation. **a,** We include multi-modal data, including demographics, diagnosis, molecular markers from either immune-staining or sequencing, survival information and whole slide images (WSIs). These data sources, whenever available, are aggregated from eight independent cohorts: TCGA, CPTAC, GBMatch, IvyGAP, Cambridge University Hospital (retrospective, CUH-R; prospective, CUH-P), Beijing Tian Tan Hospital (TTH), and Sun Yat-sen Hospital (SYH) (Tables S1). We use the large-scale TCGA cohort for model training. We then fine-tune and validate our model on the retrospective datasets, including CPTAC, GBMatch, IvyGAP, CUH-R and TTH

cohorts. We recruit CUH-P and SYH cohorts for prospective validation. **b**, Before being fed into the AI model, tumor region is segmented from the WSIs, subsequently patched at 256 μm, with the background and low staining quality (e.g., inked) regions filtered. A pretrained ResNet[15] is utilized to extract patch-level features, with the last fully connected layers of the network excluded. **c**, CAMPaS jointly predicts histology grading and molecular markers, finally integrated for glioma classification in a unified framework, with two specially designed modules, i.e., bi-directional attention module and noise-robust learning module, for mitigating challenges of dataset heterogeneity and label uncertainty in real-world datasets. Specifically, CAMPaS is generally a multi-task multi-instance learning framework built upon a hierarchical vision transformer[16] backbone. For each input WSI, the extracted WSI features from (**b**) are first fed into the bi-directional attention module to extract robust features. The bi-directional attention module dynamically weights the importance of WSI patches and therefore generates patch attention to identify and select critical regions of WSIs from heterogeneous cohorts. The generated attention is then used to guide the feature extraction in a forward-backward paradigm. Additionally, this module incorporates an adaptive color-space normalization strategy to mitigate variability introduced by WSI processing procedures. Further, the extracted features are refined under the noise-robust learning module, which enables them to alleviate the label uncertainty caused by missing molecular markers. During the noise-robust learning, we incorporate our previously proposed molecular graph, histology prediction and cross-modal interaction subnets[13,14] to model molecular markers, predict histology grading/subtyping and achieve cross-modal interaction, respectively. During training, to account for the label uncertainty in calculation of the loss function, the uncertainty level of each WSI is first estimated based on the missing molecular information, followed by an uncertainty adaptive learning process to account for the resulting label noise. Details of CAMPaS can be found in Methods and Extended Fig. 3. Astro.: astrocytoma; Oligo.: oligodendroglioma; GBM: glioblastoma. **d,** Downstream translational applications of AI-assisted diagnosis include histological-level, molecular-level, cross-modal correlation and spatial transcriptomics-based interpretability analysis, prognostic efficacy validation and treatment response prediction on both retrospective and prospective cohorts.

2.2 Cohorts

We assembled a multi-institutional dataset of 3,367 patients with 6,043 WSIs. We collected patient demographics, diagnoses, molecular markers from either immune staining or sequencing, and survival across both retrospective and prospective cohorts representing diverse ethnical backgrounds (Table S1).

- Retrospective cohorts: Six independent cohorts were used for model training and validation: four public cohorts of TCGA[17] (1,122), CPTAC[18] (99), IvyGAP[19] (41), and GBMatch[20] (225), and two in-house cohorts from Cambridge University Hospital (CUH-R, 590, Jan 2019-Oct 2021), and Beijing TianTan Hospital (TTH, 1,121, Feb 2015-Jan 2018). After excluding patients with missing or poor-quality minimum required data, we retained 2,809 patients (5,485 WSIs).
- Prospective cohorts: We recruited 369 patients from CUH (CUH-P, Nov 2021-May 2024), and 210 from Sun Yat-sen Hospital (SYH, Aug 2021-July 2024). After exclusion, we analyzed 558 patients (558 WSIs) in total, all diagnosed under the WHO 2021 criteria.

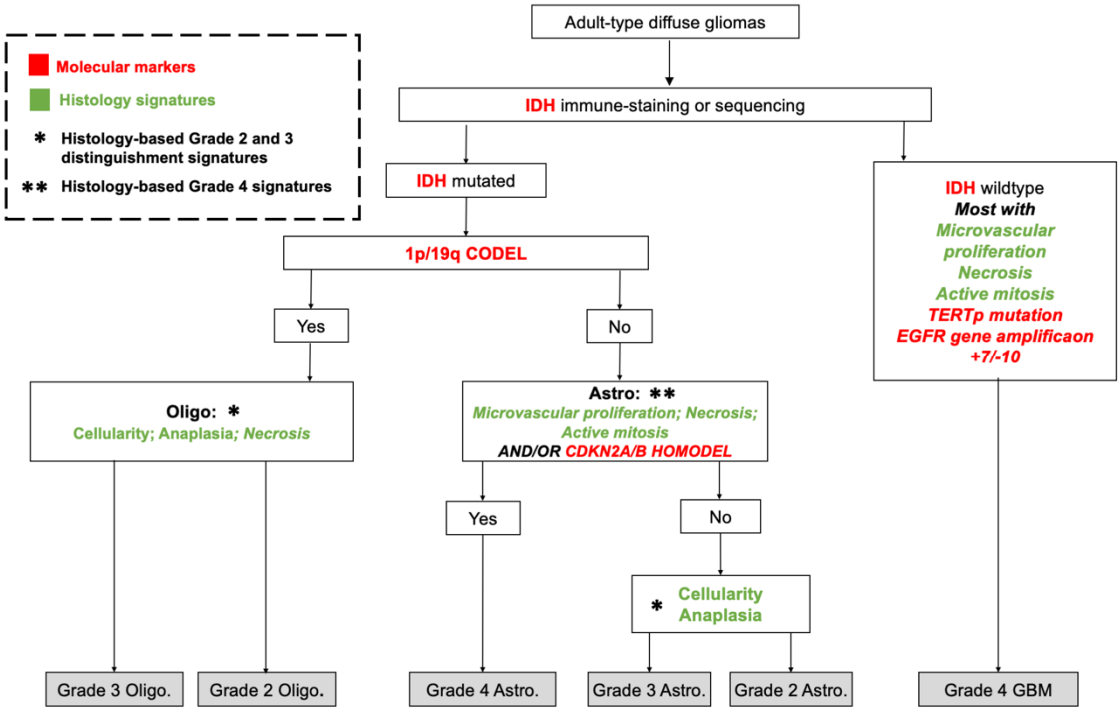
WHO 2021 glioma diagnosis workflows and inclusion/exclusion criteria and were detailed in Extended Fig.1 and Fig. S1.

2.3 Model design and performance

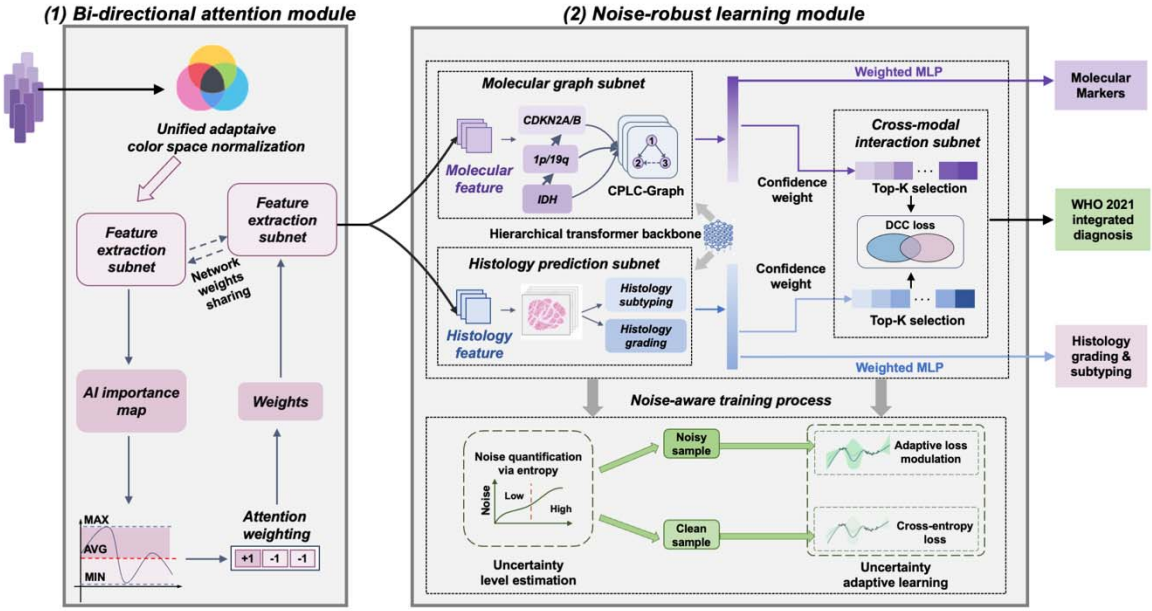
2.3.1 Model design for mitigating translational challenges

Aligned with the WHO 2021 glioma diagnosis criteria, CAMPaS (Fig. 1b-c, Extended Fig. 2; Methods) jointly predicted histology grades (Grade 2, 3 and 4), subtypes (Oligodendroglioma, Astrocytoma and GBM), and molecular markers (*IDH* mutation, *1p/19q* CODEL and *CDKN2A/B* HOMDEL). These outputs were integrated for the final glioma diagnosis. CAMPaS was developed upon our previous benchmarking-validated designs of a molecular graph subnet, a histology prediction subnet and cross-modal interaction subnet with a dynamic confidence constraint, enabling to capture relationships among molecular markers and guide cross-modal interactions, respectively. Beyond re-evaluating their predictive performance, this study further validated the robustness of these designs.

To enhance real-world translation, CAMPaS addressed two major challenges. First, to mitigate WSI heterogeneity (e.g., variation in staining, cell density and dispersion; Figs. S2-3), we designed a bi-directional attention module to effectively identify and select critical WSI regions. In this module, an adaptive color-space normalization strategy was embedded to reduce the effects of WSI procedures. Second, to handle incomplete molecular markers (statistical results of missing markers of all cohorts in Table S2), especially with limited tissue or resources, we designed a noise-robust learning module. This module estimated the uncertainty of each WSI using Shannon entropy and incorporated an uncertainty-adaptive learning strategy to reduce the model's sensitivity to missing labels during training.



Extended Fig. 1 | Integrated molecular-histology glioma diagnostic pipeline under WHO 2021 criteria. Astro., Oligo. and GBM denote astrocytoma, oligodendroglioma and glioblastoma, respectively.



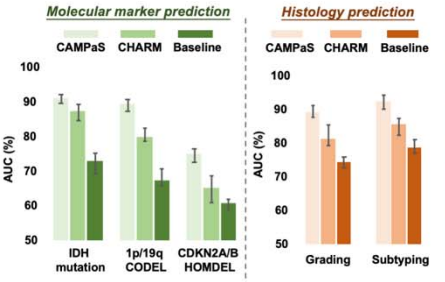
Extended Fig. 2 | Illustration of the CAMPaS framework. Taking the WSI patch features as input, CAMPaS can output the predictions of molecular markers (i.e., *IDH* mutation, *1p/19q* CODEL and *CDKN2A/B* HOMDEL), histology grading (Grade 2, 3 and 4), and subtyping (Oligodendroglioma, Astrocytoma and GBM), as well as the final WHO 2021 glioma diagnosis. Based on the backbone of hierarchical transformer, CAMPaS comprises two key modules for effective translation: a bi-directional attention module and a noise-robust learning module, detailed as follows. **(1)** The bi-directional attention module is designed to extract robust WSI features for both molecular markers prediction and histology grading and subtyping. To address dataset heterogeneity, critical WSI regions are dynamically selected by thresholding the AI importance map. These regions are then used to reweight WSI patches, effectively filtering out redundant or noisy regions. **(2)** Based on the robust features extracted in the bi-directional attention module, we further propose a novel noise-robust learning module to mitigate the label uncertainty caused by partial molecular labels missing in real-world datasets. The key idea of this learning strategy is to first estimate the uncertainty level of each WSI based on the missing molecular markers and then perform uncertainty adaptive learning process to account for the degree of label noise. Moreover, in this process, the molecular graph, histology prediction and cross-modal interaction subnets designed from our previous work[13,14] are incorporated for predicting molecular markers, histology grading/subtyping, and modelling histology-molecular interaction. In the molecular graph subnet, a Co-occurrence Probability and Label-Correlation Graph (CPLC-Graph)[13] is designed to capture the intrinsic dependencies among molecular markers. Meanwhile, in the cross-modal interaction subnet, patch-level importance scores from both molecular and histology predictions (termed "confident weights") are subjected to a Top-K (where K is a hyperparameter) selection, retaining the most reliable patches for further analysis. A Dynamic Confidence Constraint (DCC) loss is applied to guide multimodal feature interaction and reduce the impact of unreliable regions. Finally, we propose a noise-aware training process. During model training, the first step involves estimating the uncertainty level of each WSI based on the availability of corresponding molecular marker labels. This is achieved through entropy-based quantification, where high entropy reflects greater ambiguity or missingness in molecular annotations, indicating a higher likelihood of label noise. WSIs are thus categorized into noisy or clean samples accordingly. For noisy samples, an adaptive loss modulation strategy is employed to down-weight the influence of unreliable labels, mitigating the risk of overfitting to noise. In contrast, clean samples are trained using standard cross-entropy loss, enabling accurate supervision. This process constitutes an uncertainty adaptive learning framework, which dynamically calibrates the learning signal based on estimated label quality. Detailed description of CAMPaS can be found in Methods.

a

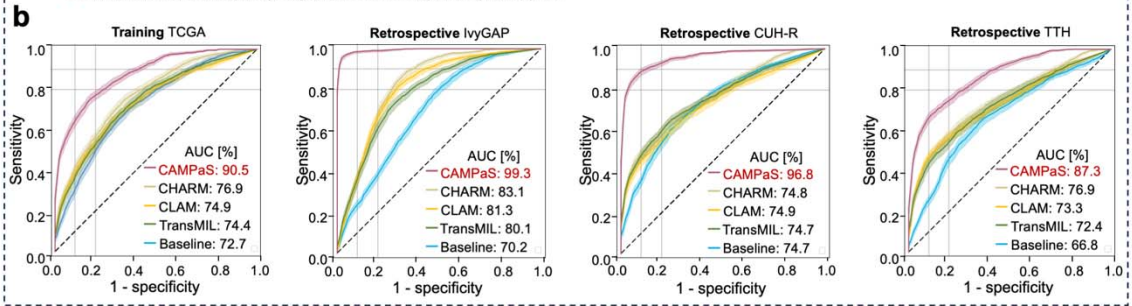
	Methods	*Training* (3-fold cross-validation)	*Retrospectively fine-tuning*		
		AUC [95% CI] on TCGA	AUC [95% CI] on IvyGAP	AUC [95% CI] on CUH-R	AUC [95% CI] on TTH
Performance comparison	CAMPaS (Ours)	**0.905 [0.895-0.916]**	**0.993 [0.968-0.998]**	**0.968 [0.944-0.992]**	**0.873 [0.843-0.881]**
	CHARM (ViT-based) •	0.769 [0.750-0.788]	0.831 [0.810-0.852]	0.747 [0.728-0.766]	0.774 [0.725-0.762]
	CLAM	0.749 [0.727-0.752]	0.813 [0.796-0.832]	0.749 [0.726-0.770]	0.733 [0.730-0.757]
	TransMIL	0.744 [0.717-0.749]	0.801 [0.766-0.822]	0.747 [0.716-0.759]	0.724 [0.706-0.739]
	Baseline (CNN-based) •	0.727 [0.709-0.745]	0.702 [0.684-0.720]	0.748 [0.729-0.766]	0.668 [0.651-0.684]
Component validation (w/o: without)	*w/o* molecular graph	0.856 [0.835-0.862]	0.902 [0.886-0.908]	0.900 [0.895-0.919]	0.823 [0.809-0.839]
	w/o bi-directional attention	0.834 [0.823-0.846]	0.887 [0.863-0.896]	0.883 [0.876-0.898]	0.812 [0.806-0.826]
	w/o cross-modal interaction	0.825 [0.813-0.840]	0.896 [0.838-0.916]	0.862 [0.856-0.873]	0.785 [0.769-0.798]
	w/o noise-robust learning	0.838 [0.826-0.849]	0.908 [0.863-0.940]	0.874 [0.822-0.903]	0.826 [0.816-0.839]

• These two methods are both specially designed for WHO 2021 integrated diagnosis of glioma

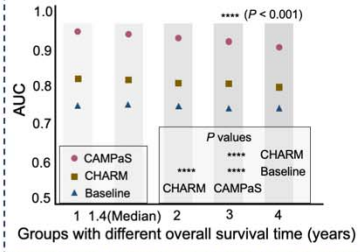
c



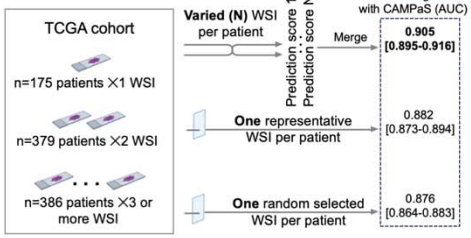
b



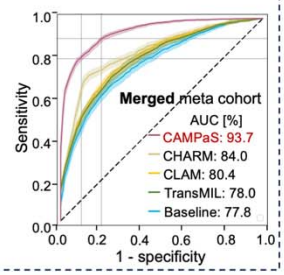
d CAMPaS diagnosis performance on various patient survival groups



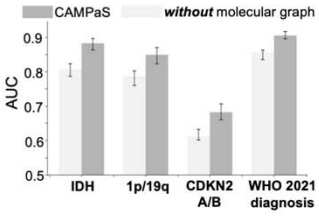
e Increasing WSIs per patient in training



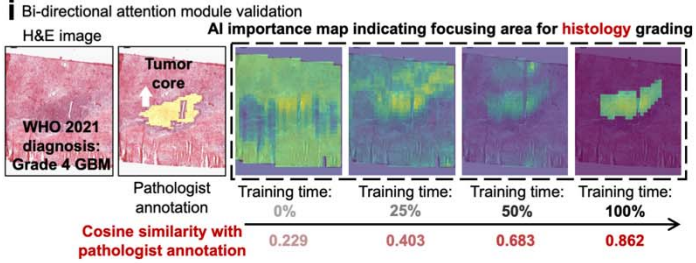
f Increasing patient number in training



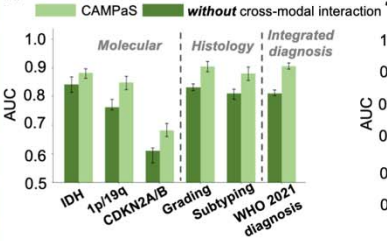
g Molecular graph subnet validation



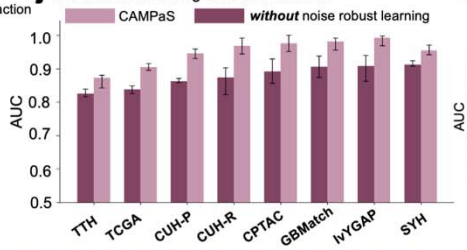
i Bi-directional attention module validation



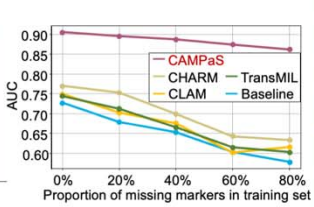
h Cross-modal interaction subnet validation



j Noise-robust learning module validation



k Glioma classification performance with varied amount of missing markers in TCGA cohort



Model accuracy analysis

Robustness analysis

Component validation

Fig. 2 | Validation on multicenter retrospective cohorts. **(a-c), Model accuracy analysis. a,** *Top panel*: Comparison of CAMPaS against CHARM, CLAM, TransMIL, and a baseline CNN-based method. *Bottom panel:* Ablation studies evaluating the contribution of the four key components proposed in CAMPaS. **b,** ROC curves of CAMPaS, CHARM, CLAM, TransMIL and the CNN baseline. **c,** Bar charts illustrating AUC values (with error bars indicating 95% CIs) for the prediction of molecular markers, histology grading and subtyping, comparing CAMPaS with other methods. **(d-f), Robustness analysis. d,** Performances of CAMPaS and comparison methods in glioma classification in the TCGA cohort, measured by AUC, across patient subgroups stratified by overall survival (OS). Each subgroup includes patients with OS less than the specified threshold. CAMPaS consistently outperforms other methods on different survival subgroups, demonstrating robust performance. **e,** Comparing our default input setting, i.e., using multiple WSIs per patient, with alternative WSI input strategies: (i) one randomly selected WSI, and (ii) the most representative (i.e., with the smallest prediction uncertainty) WSI on the TCGA cohort (n = 940). **f,** ROC curves on the merged cohort comparing CAMPaS and compared methods, highlighting the superiority of CAMPaS trained on all cohorts over its TCGA-training-only variant and other methods, further demonstrating its robustness. **(g-k), Model component validation. g,** Evaluation of the proposed molecular graph subnet on both molecular marker prediction and final diagnosis, demonstrating its effectiveness in modeling correlations of molecular markers to enhance integrated diagnosis. **h,** Evaluation of the proposed cross-modal interaction subnet. Removing this module leads to substantial performance drops across molecular marker prediction, histology grading/subtyping, and integrated diagnosis, highlighting its critical role in integrating multimodal information for improved diagnostic accuracy. **i,** Assessment of the bi-directional attention module. As training progresses, AI-generated importance maps for histology grading become progressively denoised and more focused on the tumor core, indicating improved patch selection and feature robustness. **j,** Validation of the noise-robust learning module. Glioma diagnosis performance declines across all cohorts when the module is removed, confirming its effectiveness in addressing label uncertainty due to the missing molecular data in real-world scenarios. **k,** Glioma classification performance under increasing proportions of missing molecular markers in the TCGA training set. CAMPaS consistently outperforms other methods (CHARM, TransMIL, CLAM, and Baseline) across all levels of missingness, demonstrating the superiority of the proposed noise-robust learning module in real-world scenarios of partial molecular markers missing.

2.3.2 Retrospective validation on multicenter cohorts

Model accuracy analysis

We first trained CAMPaS on TCGA, achieving a mean AUC of 0.905 (95% CI: 0.895-0.916, *P*=0.041) for glioma classification, based on three-fold cross-validation with bootstrapping (Fig. 2a; Table S3). AUCs for the six WHO 2021 glioma subtypes ranged from 0.848 to 0.918 (all *P*<0.050; Table S4).

To enhance generalizability on heterogeneous datasets, we fine-tuned CAMPaS on five additional retrospective cohorts, observing high and consistent performance across diverse datasets (Fig. 2a and Table S3): CPTAC (AUC: 0.976, 95%CI: 0.952-1.000, *P*=0.022), GBMatch (AUC: 0.981, 95%CI: 0.956-0.992, *P*=0.013), IvyGAP (AUC: 0.993,

95%CI: 0.968-0.998, *P*=0.018), TTH (AUC: 0.873, 95%CI: 0.843-0.881, *P*=0.040), and CUH-R(AUC: 0.968, 95%CI: 0.944-0.992, *P*=0.016).

CAMPaS was benchmarked against four comparison methods: CHARM[11] (a hierarchical vision transformer[16] [ViT]-based model as the SOTA comparator); CLAM[21] and TransMIL[22] (widely used multi-instance learning [MIL] methods), and a CNN[15]-based baseline (work[9] proposed for integrated glioma classification). On TCGA, CAMPaS outperformed all methods, with a ≥13.6% (*P*<0.001) AUC improvement and a 6.1% (*P*<0.001) improvement on all five fine-tuned cohorts (Fig. 2a;Table S3). ROC curves (Fig. 2b; Fig. S4; all *P*<0.001; Delong test) and forest plots (Fig. S5) further confirmed its superiority.

CAMPaS also performed strongly on auxiliary tasks of predicting histology grades/subtypes and molecular markers (Fig. 2c; Table S5). For example, on TCGA it achieved AUCs of 0.911 (*IDH* mutations, 95% CI: 0.899-0.925, *P*=0.015), 0.893 (*1p/19q* CODEL, 95% CI: 0.879-0.913, *P*=0.017) and 0.749 (*CDKN2A/B* HOMDEL, 95% CI: 0.733-0.771, *P*=0.022), consistently outperforming other methods (*P*<0.001).

Robustness analysis

To evaluate robustness, we performed subgroup analyses based on sex, histology subtype and grade, section type, survival time, and diagnostic grade shift[1] in both training (TCGA) and retrospective validation cohorts (TTH, CPTAC, GBMatch, IvyGAP) (Fig. 2d; Fig. S6; Tables S6-11). For example, the model remained comparable between grade-updated (AUC 0.896) and grade-keeping (AUC 0.913) patients in TCGA, indicating its effectiveness and robustness in reflecting the latest diagnostic criteria. Additionally, CAMPaS showed high performance on both frozen (AUC 0.900) and FFPE (AUC 0.940) sections in TCGA, demonstrating robustness in both scenarios of intraoperative and postoperative diagnosis.

[1] Patients with their glioma grading updated or kept, when transferring from previous diagnosis criteria to the current WHO 2021 scheme.

We also tested model robustness on TCGA patients with multiple WSIs (n=765, 81.4%). Using a single random slide per patient yielded an AUC of 0.876, slightly improving to 0.882 with the most representative slide (Methods). Notably, aggregating predictions across all WSIs (i.e., default setting in CAMPaS) improved AUC to 0.905 (Fig. 2e), demonstrating that increasing data diversity enhances model robustness.

Finally, we trained CAMPaS on the merged dataset spanning all six cohorts, to leverage the advantages of scale and diversity for enhancing model performance and generalization. CAMPaS outperformed all compared methods on all metrics, with an AUC of 0.937 (95% CI: 0.929-0.953, *P*=0.024), surpassing CAMPaS trained solely on TCGA (AUC 0.905), CHARM (AUC 0.912), CLAM (AUC 0.804), TransMIL (AUC 0.780) and CNN baseline (AUC 0.840) (Fig. 2f; Table S12). This indicated that expanding the training size enhanced model robustness and generalization.

Component validation

We validated the effectiveness of key components in CAMPaS through ablation studies, removing them in turn (Fig. 2a,g-k). Removing the molecular graph subnet and cross-modal interaction subnet resulted in significant performance drops (Fig. 2g-h), with AUC decreasing from 0.968 to 0.900 (*P*=0.012) and 0.862 (*P*=0.024) in CUH-R, respectively, highlighting their roles in modeling molecular correlations and histology-to-molecular interactions.

Ablating the bi-directional attention module led to a drop of AUC on TCGA from 0.905 (95% CI: 0.895-0.916, *P*=0.045) to 0.834 (95% CI: 0.823-0.846, *P*=0.047) (Fig. 2a). During training, AI demonstrated gradual attention to the tumor core, aligning well with expert annotations (Fig. 2i), confirming the role of bi-directional attention module in identifying key regions with robust features and accurate prediction.

Further, removing the noise-robust learning module consistently reduced performance across all cohorts, such as CPTAC (AUC from 0.873 to 0.826) and CUH-R (AUC from 0.993 to 0.908) (Fig. 2j). Additionally, when trained with increasing levels of missing molecular markers (from 0% to 80% in TCGA), CAMPaS exhibited only a slight performance drop (AUC from 0.905 to 0.862), consistently outperforming all compared

methods ($P < 0.001$; Fig. 2k).This suggests the strategy effectively mitigates label noise by down-weighting noisy or ambiguous samples, thereby preventing overfitting introduced by label uncertainty and improving model stability under real-world conditions.

2.4 Model interpretability analysis

2.4.1 AI-derived features are associated with clinicopathological signatures

To assess the interpretability of CAMPaS, we analyzed correlations between AI-derived features and clinicopathological signatures. Following prior work[23,24], we stratified TCGA patients into low (n=305), intermediate (n=102), and high (n=533) malignancy groups, using agglomerative hierarchical clustering[25] on AI-predicted malignancy scores (Methods; Extended Fig. 3).

Multiple linear regression (Fig. 3a-b) showed that higher AI-predicted malignancy was positively associated with pathology signatures indicating poor prognosis, including necrosis and angiogenesis, as well as IDH wildtype. These associations were also observed in the prospective SYH and CUH-P cohorts (Fig. S7), reinforcing the clinical alignment and interpretability of CAMPaS predictions.

We assessed associations with established diagnostic markers not yet included in the WHO 2021 classification (e.g., *ATRX, TERTp, PTEN, TP53*), as well as other reported prognostic markers (e.g., *MGMT, MTAP, SEC61G*) (Fig. 3a-b; Table S13). Notably, *MGMT* methylation (P=0.023) were linked to lower AI-predicted malignancy scores, consistent with their roles in IDH-mutated astrocytoma. *TERTp* mutation ($P=2.22\times10^{-9}$) was associated with higher scores, likely reflecting its co-occurrence with IDH wildtype in glioblastoma. Of note, most markers were biologically related to *IDH* mutation status, used in training CAMPaS, suggesting the model captures relevant molecular features beyond those directly supervised markers.

We expanded the analysis using the top 5% most prevalent molecular alterations from cBioPortal[26] (Fig. S8). Several markers showed significant associations with AI-predicted malignancy. For instance, alterations in the well-established RTK/RAS/PI3K pathway[27] (e.g., *PIK3CA*, *NF1*, *RB1*) were enriched in the high malignancy subgroup,

consistent with their roles in glioma aggressiveness and thereby verifying our interpretability. Beyond established pathways, we also observed enrichment of *CIC* and *FUBP1* alterations, implicated in the pathways of SWI/SNF (BAF) complex and histone acetylation/deacetylation, in the intermediate malignancy subgroup, indicating the potential in indicating glioma invasion and progression. This finding is largely attributed to our accurate prediction of the hallmark markers, e.g., IDH mutation, and robust features learned with cancer priors. Collectively, these findings indicated that CAMPaS may reveal candidate diagnostic or prognostic markers.

2.4.2 AI-derived features correlated with histological features

To validate histological relevance, we extracted AI importance maps[2] (Methods) for histology grading and calculated Pearson's correlation coefficients (PCCs) between AI importance maps with 50 pathologist-annotated histological features in the GBMatch cohort (Fig. 3c; Fig.S9). Twenty-seven features showed significant correlations ($P<0.05$), including established markers such as tumor area ($P=0.0013$) and necrosis area ($P=0.0025$) (Fig. 3c). Compared to other models, CAMPaS consistently demonstrated stronger correlations with histological metrics, using its top important (from top 50% to 5%) features[3] (Fig. S9), supporting its alignment with expert annotations.

We further validated these findings in TCGA by comparing AI-defined malignancy subtypes. The high-malignancy group consistently demonstrated higher necrosis area and more tumor nuclei than the intermediate and low subtypes (Fig. 3d-e, Fig. S10), demonstrating the ability of CAMPaS to capture biologically and clinically relevant histological patterns.

[2] AI importance maps highlighted regions in WSIs relevant to predicting molecular markers and histology grading/subtyping, while AI-predicted malignancy scores reflected the likelihood of high-grade tumors based on WHO 2021 criteria.

[3] The top 5% important features were further used for prognostic validation in Section 2.5.2

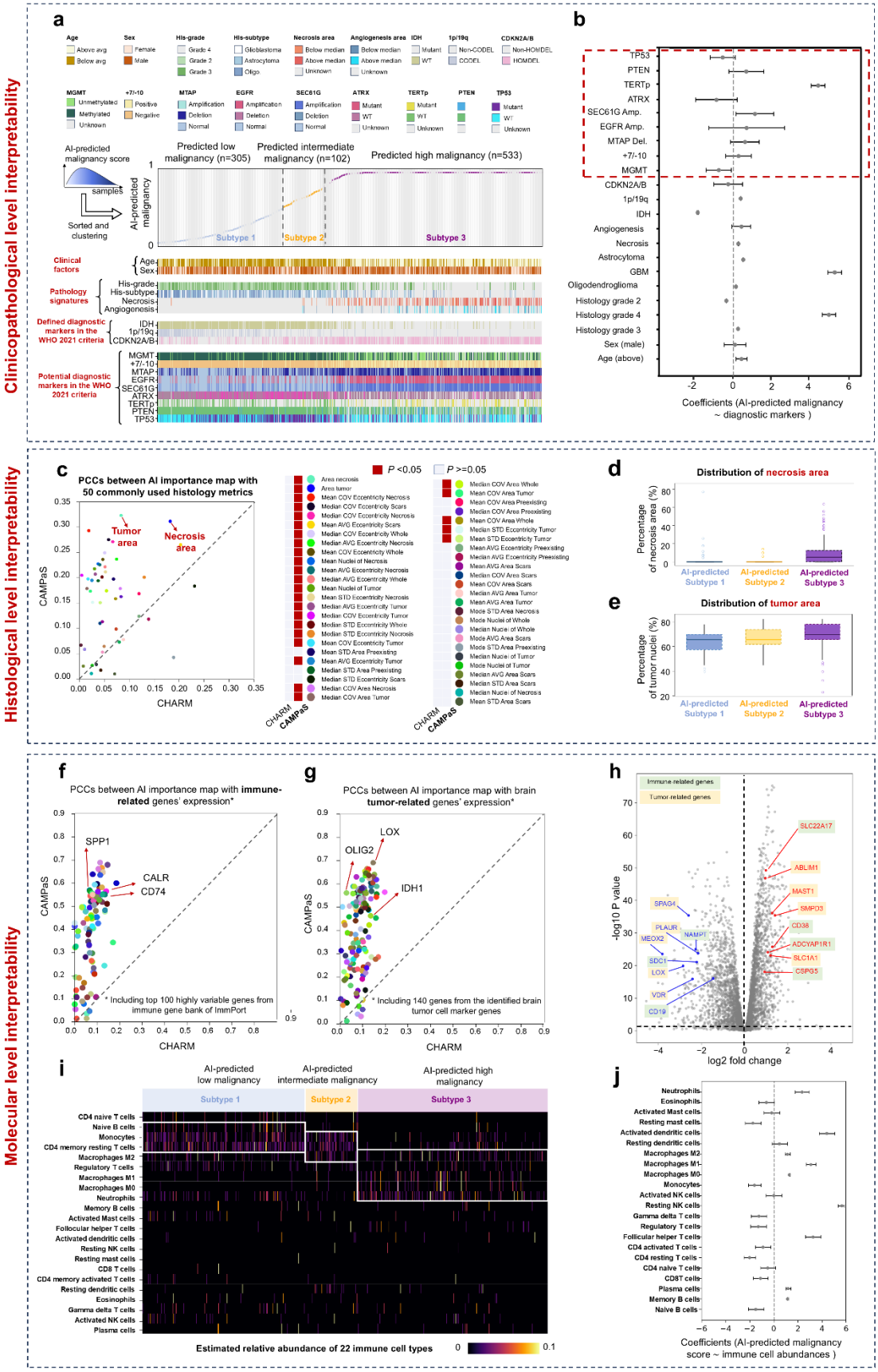

Clinicopathological level interpretability
Histological level interpretability
Molecular level interpretability
a
b
Age
Sex
His-grade
His-subtype
Necrosis area
Angiogenesis area
IDH
1p/19q
CDKN2A/B
MGMT
+7/-10
MTAP
EGFR
SEC61G
ATRX
TERTp
PTEN
TP53
AI-predicted malignancy score
samples
Sorted and clustering
Predicted low malignancy (n=305)
Predicted intermediate malignancy (n=102)
Predicted high malignancy (n=533)
AI-predicted malignancy
Subtype 1
Subtype 2
Subtype 3
Clinical factors
Pathology signatures
Defined diagnostic markers in the WHO 2021 criteria
Potential diagnostic markers in the WHO 2021 criteria
Coefficients (AI-predicted malignancy ~ diagnostic markers)
c
PCCs between AI importance map with 50 commonly used histology metrics
Tumor area
Necrosis area
CAMPaS
CHARM
P <0.05
P >=0.05
d
Distribution of necrosis area
Percentage of necrosis area (%)
AI-predicted Subtype 1
AI-predicted Subtype 2
AI-predicted Subtype 3
e
Distribution of tumor area
Percentage of tumor nuclei (%)
f
PCCs between AI importance map with immune-related genes' expression*
SPP1
CALR
CD74
* Including top 100 highly variable genes from immune gene bank of ImmPort
g
PCCs between AI importance map with brain tumor-related genes' expression*
LOX
OLIG2
IDH1
* Including 140 genes from the identified brain tumor cell marker genes
h
Immune-related genes
Tumor-related genes
-log10 P value
log2 fold change
SLC22A17
ABLIM1
MAST1
SMPD3
CD38
ADCYAP1R1
SLC1A1
CSPG5
SPAG4
PLAUR
NAMPT
MEOX2
SDC1
LOX
VDR
CD19
i
AI-predicted low malignancy
AI-predicted intermediate malignancy
AI-predicted high malignancy
Estimated relative abundance of 22 immune cell types
j
Coefficients (AI-predicted malignancy score ~ immune cell abundances)

Fig. 3 | Model interpretability analysis. (a-c) Clinicopathological level interpretability. a, Integrated visualization of AI-predicted malignancy scores and clinical factors, pathology features, and molecular markers in the TCGA cohort (n = 940). The top panel shows the AI-predicted malignancy score curve across patients, stratified into three subtypes using agglomerative clustering. The bottom panel includes heatmaps of key clinical, histological, and molecular features, including defined diagnostic markers and potential markers according to the WHO 2021 criteria. *+7/-10* denotes the presence of concomitant gain of chromosome 7 and loss of chromosome 10. **b,** Associations of AI-predicted malignancy scores with clinical factors, pathology features, and molecular markers, using multiple linear regression with the malignancy scores as the dependent variable. Data are presented as the coefficients of the linear regression and 95% CIs (n = 940). Amp.: amplification; Del.: deletion. **(c-e) Histological level interpretability. c,** Comparison of CAMPaS with CHARM (ViT-based model) in histological interpretability. Scatter plots display Pearson correlation coefficients (PCCs) comparing the correlations between 50 expert-annotated histology metrics and AI importance maps by CAMPaS with CHARM (see Methods) in GBMatch. Each colored dot represents a single histological feature, with notable examples such as Tumor area and Necrosis area. Histology metrics significantly correlated with AI importance maps ($P < 0.05$) are indicated by red squares in the adjacent legend. **d-e,** Distribution of histological features (**d,** percentage of necrosis area and **e,** percentage of tumor nuclei) across CAMPaS-predicted malignancy subtypes in the TCGA cohort. The box plots are defined by the center tick as the median value, the lower and upper parts of the box Q1 and Q3 quartiles, respectively, and the bounds of whiskers are (Q1 − 1.5 × IQR, Q3 + 1.5 × IQR). Any outlier points beyond the whiskers are displayed with point marks. **(f-j) Molecular level interpretability. f-g,** Correlation analysis between the AI importance maps of molecular markers and expression levels of 140 tumor-related (well-established glioma marker genes from literature[3,28]) (**f**) and 100 immune-related (highly variable genes from immune gene bank of ImmPort[29]) (**g**) genes in TCGA. These plots indicate that CAMPaS more strongly associates with gene expression signatures (both immune and tumor-related) compared to CHARM, supporting its enhanced molecular-level interpretability. **h,** Differential expression analysis between CAMPaS predicted high malignancy versus low malignancy in TCGA. It shows that tumors predicted as high or low malignancy by CAMPaS have distinct gene expression patterns, especially in immune- and tumor-related genes. P values of the likelihood ratio test were adjusted using the Benjamini-Hochberg FDR and statistical significance accepted < 0.050 (i.e., above the horizontal dash line). **i-j,** Immune landscape across AI-predicted malignancy in TCGA gliomas. **i,** Hierarchical clustering heatmap of abundance of 22 common human immune cells for TCGA patients (n = 940), ordered by three clusters representing the three subtypes in terms of AI-predicted malignancy. Notably, naive B cells, monocytes, and CD4 memory resting T cells are more abundant in tumors with low AI-predicted malignancy score, whereas monocytes, CD4 memory resting T cells, and M2 macrophages are abundant in tumors with intermediate AI-predicted malignancy score. In contrast, M0, M1, M2 macrophages, regulatory T cells and neutrophils are relatively enriched in tumors with high AI-predicted malignancy score. These patterns are consistent with prior glioma knowledge[30], suggesting that as malignancy increases, the immune environment becomes more suppressive and less cytotoxic, potentially facilitating immune evasion and tumor progression. **j,** Multiple single linear regressions reveal associations between AI-predicted malignancy score and immune cell abundances. Normalized regression coefficients ($\log_2$ fold change with 95% CI) indicate that neutrophils, M1 and M0 macrophages, activated dendritic cells, and resting NK cells are positively associated with AI-predicted malignancy scores, while naive B cells, CD 4 resting T cells, and CD8 T cells show negative associations.

2.4.3 AI-derived features reflected tumor- and immune-related transcriptions

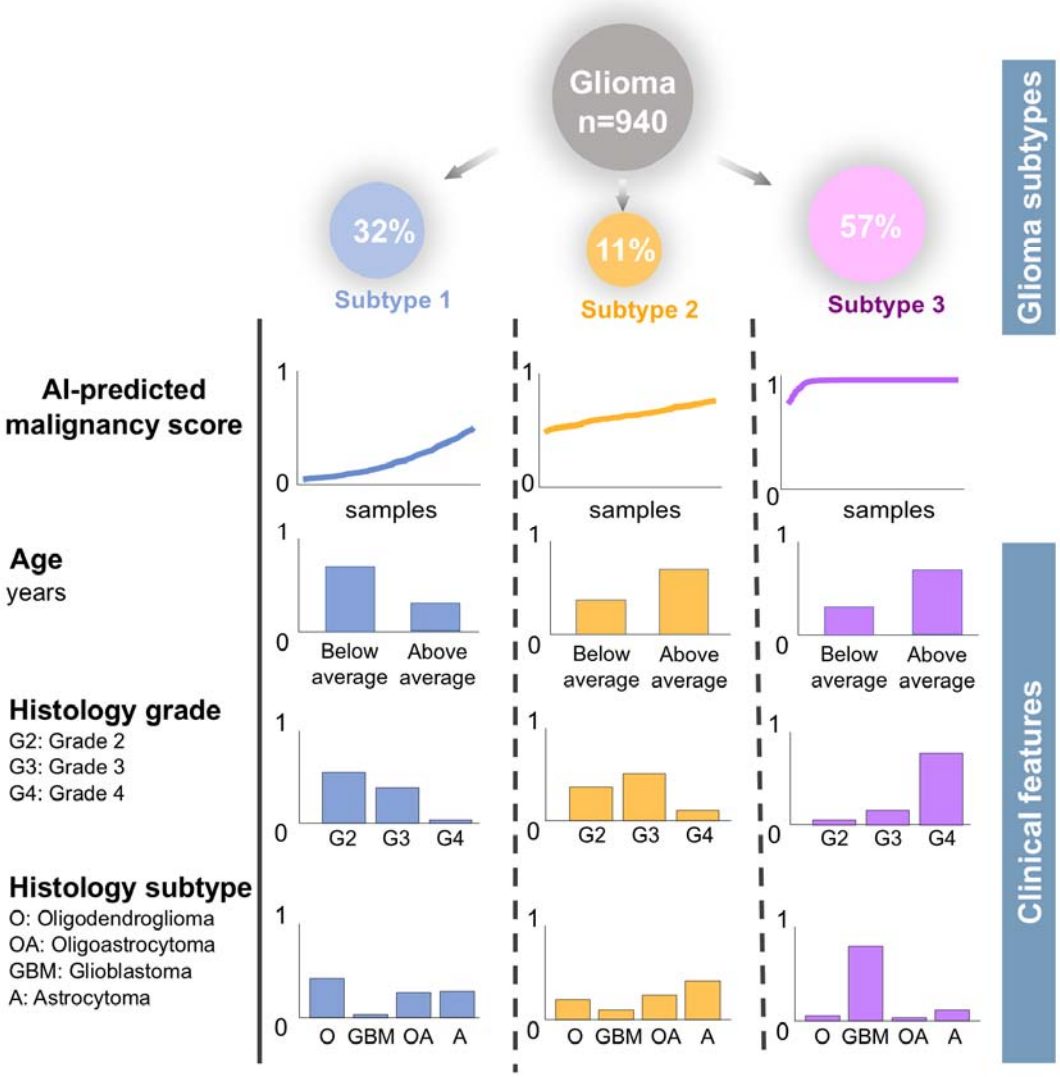
To assess model interpretability at the molecular level, we next analyzed transcriptional correlates of AI-predicted malignancy in TCGA (Fig. 3f-g, Figs. S11-12), focusing on common tumor-related and immune-related signature genes (Methods). CAMPaS consistently showed stronger correlations over comparison models across both gene sets. Alongside the known genes such as *IDH* (PCC=0.45, *P*=0.006), *OLIG2* (PCC=0.56, *P*=0.003), other significant tumor-related genes included LOX (PCC=0.69, *P*=0.002), involved in extracellular matrix (ECM) remodeling and tumor invasion. For immune-

related genes, significant correlations were observed for *SPP1* (PCC=0.56, *P*=0.002), *CALR* (PCC=0.56, *P*=0.001), *CD74* (PCC=0.55, *P*=0.005), all known as immune-modulatory genes linked to tumor immune evasion. These results indicated CAMPaS's ability to uncover potential markers of glioma invasiveness from morphological features.

Differential expression analysis showed that high AI-predicted malignancy was associated with upregulated genes (Fig. 3h), such as *ABLIM1*, *MAST1*, and *SLC1A1*, linked to tumor invasion and progression. Conversely, the AI-predicted low-malignancy group showed enriched genes associated with hormone signaling, such as *VDR*, involved in Vitamin D signaling and anti-proliferative effects. Gene ontology enrichment analysis showed that the immune-related genes were involved in antigen presentation and immune response, while tumor-related genes mapped to tissue remodeling, key features of tumor growth and invasion (Fig. S13).

To assess immune infiltration, we employed CIBERSORT[31] to estimate the relative abundances of 22 immune cells and their associations with AI-predicted malignancy scores (Fig. 3i-j, Table S14; Methods). Multiple linear regression revealed that higher malignancy scores were positively correlated with neutrophils (*P*=0.009), activated dendritic cells (*P*=0.006), macrophages M0 (*P*=0.01), M1 (*P*=0.012), resting natural killer cells (*P*=0.007), follicular helper T cells (*P*=0.012), plasma cells (*P*□=□0.007) and memory B cells (*P*=0.003), indicating that higher AI-predicted malignancy is associated with elevated infiltration by immunosuppressive and inflammatory cells, which could guide patient selection for immune-targeted interventions.

These results collectively suggested that CAMPaS captures transcriptional and immunological patterns aligned with tumor aggressiveness, demonstrating interpretability at the molecular level.



Extended Fig. 3 | Overview of three subtypes stratified via AI-predicted malignancy scores. Integrative analysis of 940 gliomas in TCGA results in three different subtypes with distinct clinicopathological characteristics. The three subtypes are identified by extracting the top-ranked features per subtype using AI-predicted malignancy scores with agglomerative hierarchical clustering (Methods). Subtype 1 (32%) shows low AI-predicted malignancy scores and is enriched for lower histology grade tumors, predominantly astrocytoma, oligodendroglioma and oligoastrocytoma, with a younger patient profile. Subtype 2 (11%) represents an intermediate group with moderate AI-predicted malignancy scores, mixed histology grades. Subtype 3 (57%) exhibits the highest AI-predicted malignancy scores and is dominated by high-grade glioblastoma, often in older patients, reflecting more aggressive disease. These trends suggest that AI-predicted malignancy scores effectively stratify glioma patients into clinically and biologically meaningful subgroups.

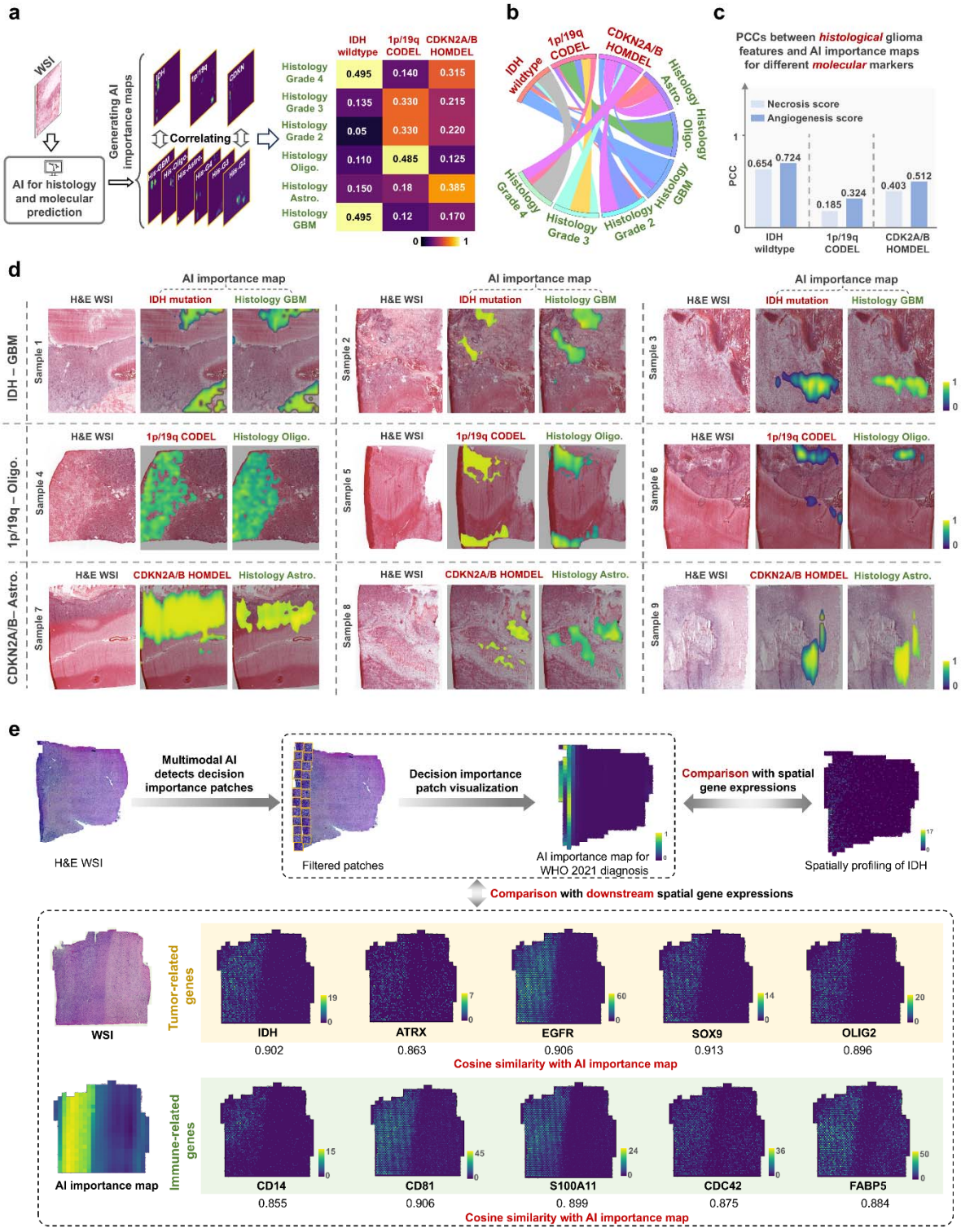


Fig. 4 | Cross-modal interpretability. a, Heatmap showing the pairwise spatial correlations between AI importance maps of molecular markers prediction (i.e., *IDH* mutation, *1p/19q* CODEL, *CDKN2A/B* HOMDEL) and histological predictions (i.e., histology grades and subtypes). Oligo.: oligodendroglioma; Astro.: astrocytoma. **b,** Chord diagram

visualizing the relationships in (**b**), highlighting the connections between AI-predicted molecular markers and histology features. The consistence with the prior knowledge, e.g., subtype GBM and *IDH* wildtype, indicates the cross-modal interpretability of our model. **c,** Bar plots showing Pearson correlation coefficients (PCCs) between histological features (angiogenesis score, necrosis score) and AI importance maps for predicting different molecular markers. For the AI importance map of *IDH* wildtype, correlations with necrosis and angiogenesis are both high, consistent with the aggressive and heterogeneous nature of high-grade glioma. In contrast, for the AI importance map of *1p/19q* CODEL, correlations with both metrics are low, reflecting the less aggressive, infiltrative tumor biology, further supporting the biological validity of the cross-modal spatial associations. **d,** Examples with strong spatial correlations observed between AI importance maps for molecular and histology prediction: spatial features associated with *IDH* wildtype status tend to align with those of histology GBM, spatial features of *1p/19q* CODEL align with histology oligodendroglioma, and spatial features of CDKN2A/B HOMDEL aligns with histology astrocytoma (quantified in **a**). These paired maps exhibit consistent spatial patterns, reflecting known molecular biology and supporting the model's cross-modal interpretability. **e,** Workflow for validating AI importance maps using spatial transcriptomics data[32] (See Methods for data processing details): the model identifies decision-relevant patches, whose spatial correlations with spatial transcriptomic profiles are measured by cosine similarity. The lower panel shows spatial expression maps of selected genes. Tumor-related genes (*IDH*, *ATRX*, *EGFR*, *SOX9*, *OLIG2*) exhibit high spatial similarity (cosine similarity > 0.86) with AI importance maps, supporting that the model focuses on biologically meaningful tumor regions. These genes are key drivers of tumorigenesis and are commonly altered in molecular subtypes defined by WHO 2021 criteria. Similarly, immune-related genes (*CD14*, *CD81*, *S100A11*, *CDC42*, *FABP5*), associated with macrophage activation, immune signaling, and tumor–immune interactions, also show strong spatial concordance with AI-derived regions. This highlights CAMPaS's ability to capture not only tumor-intrinsic features but also immune microenvironmental signals, reinforcing the biological interpretability of AI decisions at both tumor and immune levels.

2.4.4 CAMPaS showed cross-modal interpretability

To validate the cross-modal interpretability of CAMPaS, we generated AI importance maps from histology grade, subtype, and molecular marker predictions. Spatial correlation analyses demonstrated strong spatial alignment of these maps (Fig. 4a-b,d). For instance, AI importance maps for *IDH* wildtype spatially overlapped with those for Grade III and IV (PCC=0.495), while maps for *1p/19q* co-deletion closely aligned with those for oligodendroglioma (PCC=0.485). Further, compared to *1p/19q* CODEL and *CDKN2A/B* HOMDEL, the AI importance map for *IDH* wildtype showed strong correlations with necrosis and angiogenesis, aligning with the biological priors[33,34] of high-grade gliomas (Fig. 4c). These spatial co-occurrence patterns suggest that the cross-modal interaction subnet could effectively model the relationships between histological and molecular features, enhancing cross-modal interpretability.

We further employ sequencing-based spatial transcriptomics and observe that decision-relevant spatial regions, highlighted by CAMPaS, showed strong spatial concordance with the expressions of tumor-related genes (e.g., *ATRX*, *EGFR*, *SOX9*, *OLIG2*) and immune-related genes (e.g., *CD14*, *CD81*, *S100A11*, *CDC42*, *FABP5*) (Fig. 4e; Fig. S14). Particularly, AI-highlighted regions exhibited strong spatial alignment with the

expressions of many downstream genes of *IDH*, with cosine similarity scores ranging from 0.855 (*CD14*) to 0.913 (*SOX9*).

These findings highlight the ability of CAMPaS to produce spatially coherent and biologically relevant predictions across histological and molecular levels, reinforcing its interpretability and translational potential.

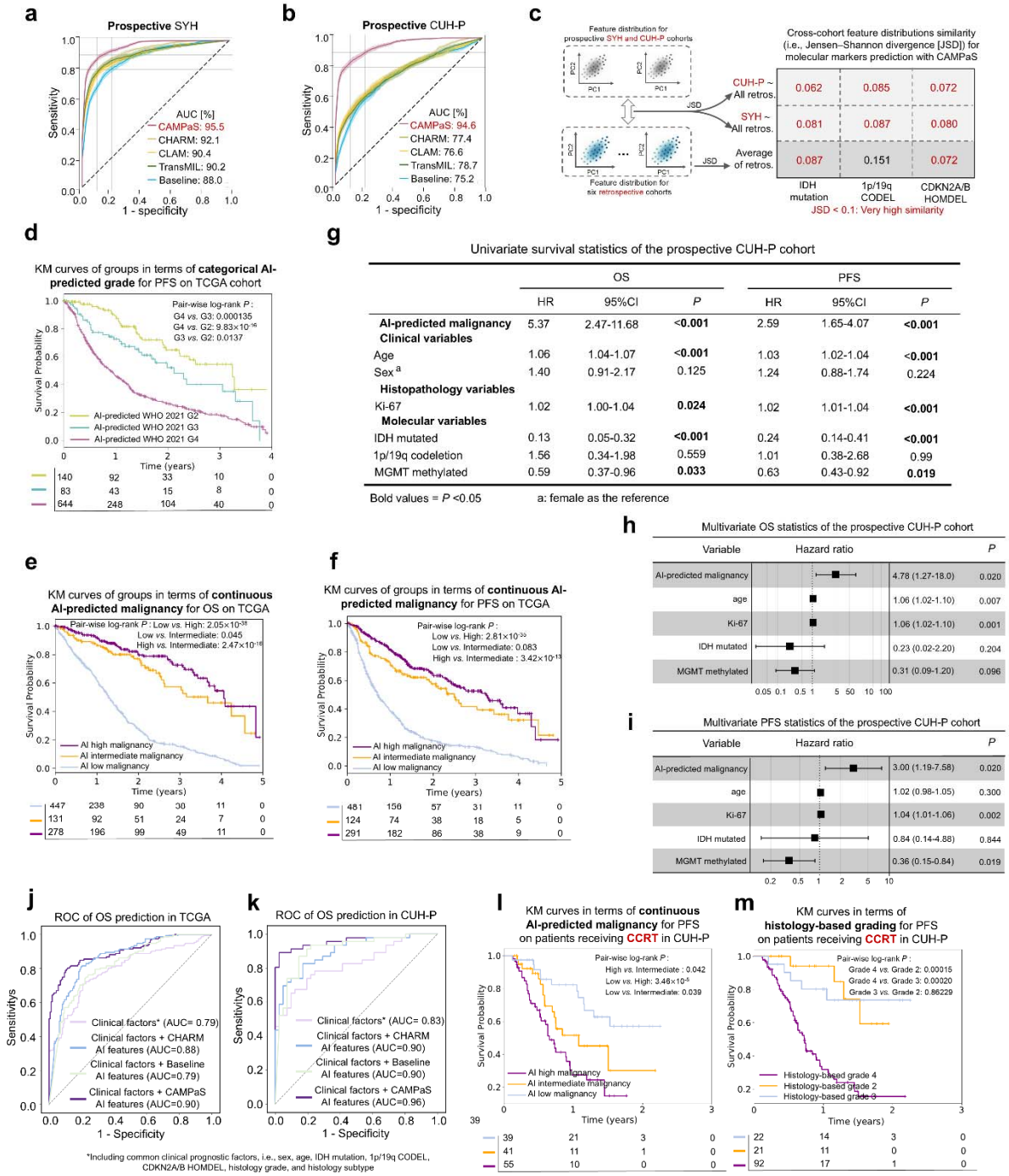


Fig. 5 | Prospective validation of the CAMPaS's performance and its prognostic value. (a-b), ROC curve of CAMPaS, CHARM, CLAM, TransMIL, and the CNN baseline demonstrate that CAMPaS consistently achieves superior glioma classification performance in the prospective SYH (**a**) and CUH-P (**b**). **c**, Jensen-Shannon divergence (JSD[35,36]) measuring how CAMPaS's latent feature space (derived via principal component analysis) for molecular markers prediction shifts from one cohort to another, including prospective SYH/CUH-P to the overall retrospective cohorts (TCGA, CAPTAC, IvyGAP, GBMatch, TTH, and CUH-R), and the averaged pair-wise comparison among retrospective cohorts. Across all three markers of *IDH* mutation, *1p/19q* CODEL and *CDKN2A/B* HOMDEL, the

prospective-retrospective JSD values are comparable to the retrospective mean, indicating that CAMPaS extracts feature representations in prospective cohorts that closely match those in retrospective data, underscoring its strong translational generalization capability. PC: principal component. **d,** Survival analysis based on AI-predicted WHO 2021 histology grades. Kaplan-Meier (KM) curves show groups stratified by AI-predicted histology grades (G2: grade 2, G3: grade 3, G4: grade 4) in the TCGA cohort show significant differences in progression-free survival (PFS), supporting the prognostic value of AI-predicted grading. **(e-f),** Prognostic value of AI-predicted malignancy scores in the TCGA cohort. KM survival curves stratified by AI-predicted malignancy scores (low, intermediate, high) show significant differences in both (**e**) OS and (**f**) PFS. Pair-wise p-values of the log-rank test are shown. (**g-i**) Univariate and multivariate models including AI-predicted malignancy score and available established clinical, histopathological and molecular prognostic factors in the prospective CUH-P cohort. **g**, Univariate Cox regression analysis of OS (left) and PFS (right). AI-predicted malignancy scores show strong and significant associations with both OS (HR = 2.62, P = 0.003) and PFS (HR = 3.84, P < 0.001). (**h-i**) Multivariate Cox regression analyses further confirm that AI-predicted malignancy remains an independent prognostic factor for both OS (**h**, HR = 2.62, P = 0.003) and PFS (**i**, HR = 2.34, P = 0.038), after adjusting for known prognostic variables such as age, *Ki-67*, *IDH* mutation, and *MGMT* methylation, which are all significant in univariate Cox regression analysis. **(j-k),** The model that integrates clinical factors with AI importance maps from CAMPaS achieves a higher AUC for OS prediction compared to models using clinical factors with AI importance maps from CHARM or the CNN-based baseline, as well as clinical factors alone, on the retrospective TCGA cohort (**j**) and the prospective CUH-P cohort (**k**). **(l-m),** KM analysis of AI-predicted malignancy scores (**l**) versus conventional prognostic factor of histology-based grades (**m**) on PFS in patients treated with concurrent chemoradiotherapy (CCRT) in CUH-P cohort. Pairwise log-rank P-values show that AI-predicted malignancy scores significantly distinguish all three risk groups, while histology-based grades show no significant difference between Grade 2 and Grade 3, indicating superior value of CAMPaS's feature as a strong prognostic factor over conventional established prognostic factor of histology-based grading.

2.5 Prospective validation of model performance and clinical efficacy

2.5.1 CAMPaS showed high performance on prospective cohorts

We assessed CAMPaS's generalizability through zero-shot validation, training only on TCGA and testing on two independent prospective cohorts (CUH-P and SYH) without fine-tuning (Fig. 5a-b). Despite differences in tissue processing, scanner vendors, and patient demographics, CAMPaS achieved high diagnostic performance (AUC: 0.946 for CUH-P, 0.955 for SYH), demonstrating robustness across real-world clinical settings. Additionally, following prior works[35,36], we quantified shifts in latent feature distributions for molecular-marker prediction using Jensen–Shannon divergence (JSD). Prospective-to-retrospective JSD values were comparable or lower than retrospective-to-retrospective values (Fig. 5c), highlighting CAMPaS's stability and prospective generalizability.

2.5.2 CAMPaS showed strong prognostic value

We validated the prognostic value of CAMPaS using both AI-predicted grades and continuous malignancy scores across the TCGA and prospective CUH-P (Fig. 5d-m; Fig. S15-18). For example, in TCGA, Kaplan-Meier (KM) curves demonstrated significantly

poorer outcomes for patients with predicted Grade 4 (median PFS: 330 days), compared to Grade 3 (median PFS: 768 days) and Grade 2 (median PFS: 1,182 days) (pairwise log-rank tests, all $P<0.050$; Fig. 5d), suggesting the prognostic value of AI-predicted glioma grading.

We further stratified patients into low-, intermediate-, and high-risk groups based on AI-predicted malignancy scores. KM analyses showed that the three groups had significantly different OS and PFS (pairwise log-rank tests, all $P<0.050$; Fig. 5e-f), confirming the prognostic value of the malignancy scores.

In CUH-P, univariate Cox regression identified AI-predicted malignancy scores as a strong predictor of survival (OS: hazard ratio [HR] 5.37, 95% CI: 2.47-11.68, $P<0.001$; PFS: HR 2.59, 95% CI: 1.65-4.07, $P<0.001$; Fig. 5g). In multivariate models adjusted for clinical variables, the malignancy score retained independent significance (OS: HR=4.78, 95% CI=1.27-18.0, $P=0.020$; PFS: HR=3.00, 95% CI=1.19-7.58, $P=0.020$; Fig. 5h-i).

In TCGA, integrating AI features with clinical factors (AUC 0.90) significantly improved OS prediction compared to models using clinical factors with AI features from CHARM (AUC 0.88) and the CNN-based baseline (AUC 0.79), as well as clinical factors alone clinical factors alone (AUC 0.79; Fig. 5j), Similar improvements were observed in CUH-P, increasing AUC from 0.83 to 0.96 (Fig. 5k).

Finally, we evaluated prognostic value in a subset (n=135) of CUH-P patients treated with standard concurrent chemoradiotherapy (CCRT) with temozolomide. KM curves showed significantly worse outcomes in the AI-predicted high-malignancy group compared to intermediate- and low-malignancy groups (pairwise log-rank tests, all $P<0.050$; Fig. 5l). In contrast, conventional prognostic factors (i.e., histology-based grade) showed similar trends but lacked statistical significance (Fig. 5m). These results indicated that CAMPaS offers robust, independent prognostic power across both retrospective and prospective cohorts. Moreover, the model effectively stratifies treatment response to standardized CCRT, highlighting its potential for guiding personalized clinical decision-making.

Discussion:

The landscape of cancer diagnostics has transitioned from relying solely on histological features to integrating molecular markers, exemplified by the WHO 2021 classification for diffuse gliomas. While this integration enhances diagnostic accuracy and biological relevance, it introduces practical challenges, such as delayed diagnoses from separate histology and molecular workflows and increased complexity for pathologists trained by histology inspection, increasing the risk of misdiagnosis.

To address these challenges, we developed CAMPaS, a trustworthy AI system trained and validated on 3,367 patients (6,043 WSIs) that jointly predicts histology grading/subtyping, molecular markers and integrated diagnosis, using H&E-stained slides. CAMPaS obtained AUC of 0.946 and 0.955 in two prospective cohorts for integrated glioma diagnosis, comparable to state-of-the-art AI diagnostic tools in other cancers, such as breast cancer[37] (AUC 0.943) and endometrial cancer[38] (AUC 0.96).

CAMPaS demonstrated robust performance across different numbers of WSIs per patient and varying training cohort sizes. It also performed consistently across subgroups, including FFPE and frozen sections, supporting its use in intraoperative settings where rapid diagnosis with fresh frozen samples is needed. Moreover, CAMPaS remained reliable even as the proportion of missing molecular markers increased during training, highlighting its effectiveness in real-world scenarios with incomplete molecular data.

CAMPaS's strong performance stems from our benchmarking-validated architecture with further innovations tailored to handle real-world challenges such as dataset heterogeneity and incomplete molecular information, trained on six large-scale multicenter cohorts. In addition, by considering co-occurrence patterns and cross-modal relationships, CAMPaS captured intrinsic relationship among molecular markers and histology-molecular marker correlations, respectively. Ablation studies further confirmed each component's effectiveness.

Few previous studies have developed models for glioma diagnosis, aligning with the WHO 2021 criteria. Despite success, most existing attempts were limited by small sample sizes and narrow scopes. Studies like DeepGlioma[10], involving only ~200 samples, were insufficiently validated. CHARM[11] lacked integration of cross-modal interactions and excluded key subtypes like Grade 4 astrocytoma with *CDKN2A/B* HOMDEL. Similarly, the ResNet[15]-based method[9] could misclassify Grade 4 GBM by applying redundant criteria other than IDH wildtype, and only predicted final glioma classes without predicting molecular markers and histological grading/subtyping. Other studies[12,39] were validated using single cohorts or lacked molecular-histological interpretability, limiting clinical utility.

Our interpretability analyses provide insights into understanding glioma. Key molecular markers and histological features were consistently linked to high AI-predicted malignancy, reinforcing the model's pathological grounding. Spatial transcriptomics further revealed strong spatial alignment between AI importance maps for predicting molecular markers and histology grades/subtypes. These findings highlight the model’s ability to capture clinically meaningful multimodal relationships, with notable examples such as the associations of *IDH* wildtype with high-grade histology and *1p/19q* CODEL with oligodendroglioma.

Interestingly, genomic alterations in several tumor-related genes, such as *ATRX* and *OLIG2*, which are downstream of or frequently co-occur with *IDH* mutations, showed strong associations with our AI-predicted malignancy, despite not being included in current diagnostic criteria or the training process. This may reflect the strength of our molecular graph modeling, which captures intrinsic relationships among molecular markers and holds promises for predicting more candidate markers based on histology.

Beyond diagnostics, CAMPaS provided prognostic value. AI-derived malignancy scores successfully stratified patients across multiple cohorts. In the subgroup of CUH-P cohort receiving concurrent chemoradiotherapy, AI-derived malignancy scores provided superior prognostics compared to conventional clinical metrics, highlighting the model’s potential for personalized therapeutic guidance. Further, our results imply that the AI

model could be used for treatment stratification. For instance, high CAMPaS malignancy was associated with *CD38* expression, a chimeric antigen receptor T cell target[40]. These results warrant further investigation in larger prospective studies or randomized clinical trials. CAMPaS-derived malignancy scores, trained solely for integrated diagnosis, independently predicted survival and were associated with transcriptional and immune cell profiles, underscoring the model's transferability to downstream prognostic and biological inference tasks without retraining.

Given its cross-modal architecture and strong out-of-distribution performance, CAMPaS could serve as a pretrained backbone for transfer learning in other molecularly defined cancers, including endometrial cancer[41] and renal neoplasia[42]. The modular design of CAMPaS, integrating vision transformers, molecular graphs, and uncertainty-aware multi-instance learning, supports plug-and-play adaptation to other diagnostic workflows beyond glioma. Future work will extend CAMPaS to other tumor types, potentially establishing a foundation model paradigm for computational molecular pathology.

Limitations include our focus on defined WHO diagnostic markers (*IDH*,*1p/19q* and *CDKN2A/B*), rather than broader panels that might boost diagnosis performance. We could further tune our model by including more molecular markers. In addition, our cross-modal interaction approach could benefit from different fusion methods or unsupervised pretraining to enhance generalizability. Furthermore, some morphological correlates observed in the H&E regions (e.g., structural changes) were not quantified due to the lack of labelled datasets. Finally, continuous prospective validation, in light of the evolving cancer therapeutics, remains essential for treatment stratification.

In conclusion, CAMPaS is a robust AI model for integrated glioma diagnosis, trained on the largest glioma cohorts to date. It addresses key real-world challenges through advanced cross-modal AI methodology. It demonstrates clinical utility through improved diagnostic accuracy, interpretability, and prognostic stratification.

REFERENCE

1. Komori, T. Grading of adult diffuse gliomas according to the 2021 WHO Classification of Tumors of the Central Nervous System. *Lab. Invest.* **102**, 126–133 (2022).
2. Chakravarty, D. & Solit, D. B. Clinical cancer genomic profiling. *Nat. Rev. Genet.* **22**, 483–501 (2021).
3. Louis, D. N. *et al.* The 2021 WHO Classification of Tumors of the Central Nervous System: a summary. *Neuro-Oncol.* **23**, 1231–1251 (2021).
4. Louis, D. N. *et al.* The 2007 WHO Classification of Tumours of the Central Nervous System. *Acta Neuropathol. (Berl.)* **114**, 97–109 (2007).
5. Louis, D. N. *et al.* The 2016 World Health Organization Classification of Tumors of the Central Nervous System: a summary. *Acta Neuropathol. (Berl.)* **131**, 803–820 (2016).
6. Horbinski, C., Berger, T., Packer, R. J. & Wen, P. Y. Clinical implications of the 2021 edition of the WHO classification of central nervous system tumours. *Nat. Rev. Neurol.* **18**, 515–529 (2022).
7. De Las Casas, L. E. & Hicks, D. G. Pathologists at the Leading Edge of Optimizing the Tumor Tissue Journey for Diagnostic Accuracy and Molecular Testing. *Am. J. Clin. Pathol.* **155**, 781–792 (2021).
8. Sejda, A. *et al.* WHO CNS5 2021 classification of gliomas: a practical review and road signs for diagnosing pathologists and proper patho-clinical and neuro-oncological cooperation. *Folia Neuropathol.* **60**, 137–152 (2022).
9. Wang, W. *et al.* Neuropathologist-level integrated classification of adult-type diffuse gliomas using deep learning from whole-slide pathological images. *Nat. Commun.* **14**, 6359 (2023).
10. Hollon, T. *et al.* Artificial-intelligence-based molecular classification of diffuse gliomas using rapid, label-free optical imaging. *Nat. Med.* **29**, 828–832 (2023).
11. Nasrallah, M. P. *et al.* Machine learning for cryosection pathology predicts the 2021 WHO classification of glioma. *Med* **4**, 526-540.e4 (2023).
12. Zuo, M. *et al.* Weakly supervised deep learning-based classification for histopathology of gliomas: a single center experience. *Sci. Rep.* **15**, 265 (2025).
13. Wang, X., Price, S. & Li, C. Multi-task Learning of Histology and Molecular Markers for Classifying Diffuse Glioma. in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023* (eds. Greenspan, H. et al.) 551–561 (Springer Nature Switzerland, Cham, 2023). doi:10.1007/978-3-031-43990-2_52.
14. Wang, X. *et al.* Joint modeling histology and molecular markers for cancer classification. *Med. Image Anal.* **102**, 103505 (2025).
15. He, K., Zhang, X., Ren, S. & Sun, J. Deep Residual Learning for Image Recognition. in 770–778 (2016).
16. Dosovitskiy, A. *et al.* An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. Preprint at https://doi.org/10.48550/arXiv.2010.11929 (2021).
17. Ceccarelli, M. *et al.* Molecular Profiling Reveals Biologically Discrete Subsets and Pathways of Progression in Diffuse Glioma. *Cell* **164**, 550–563 (2016).
18. Wang, L.-B. *et al.* Proteogenomic and metabolomic characterization of human glioblastoma. *Cancer Cell* **39**, 509-528.e20 (2021).
19. An anatomic transcriptional atlas of human glioblastoma | Science. https://www.science.org/doi/10.1126/science.aaf2666.

20. Klughammer, J. *et al.* The DNA methylation landscape of glioblastoma disease progression shows extensive heterogeneity in time and space. *Nat. Med.* **24**, 1611–1624 (2018).
21. Lu, M. Y. *et al.* Data-efficient and weakly supervised computational pathology on whole-slide images. *Nat. Biomed. Eng.* **5**, 555–570 (2021).
22. Shao, Z. *et al.* TransMIL: Transformer based Correlated Multiple Instance Learning for Whole Slide Image Classification. in *Advances in Neural Information Processing Systems* vol. 34 2136–2147 (Curran Associates, Inc., 2021).
23. Su, C. *et al.* Identification of Parkinson's disease PACE subtypes and repurposing treatments through integrative analyses of multimodal data. *Npj Digit. Med.* **7**, 1–22 (2024).
24. M. Amaral, D. *et al.* Temporal stratification of amyotrophic lateral sclerosis patients using disease progression patterns. *Nat. Commun.* **15**, 5717 (2024).
25. Müllner, D. Modern hierarchical, agglomerative clustering algorithms. Preprint at https://doi.org/10.48550/arXiv.1109.2378 (2011).
26. Gao, J. *et al.* Integrative Analysis of Complex Cancer Genomics and Clinical Profiles Using the cBioPortal. *Sci. Signal.* **6**, pl1–pl1 (2013).
27. Pearson, J. R. D. & Regad, T. Targeting cellular pathways in glioblastoma multiforme. *Signal Transduct. Target. Ther.* **2**, 17040 (2017).
28. Wang, Q. *et al.* Tumor Evolution of Glioma-Intrinsic Gene Expression Subtypes Associates with Immunological Changes in the Microenvironment. *Cancer Cell* **32**, 42-56.e6 (2017).
29. Bhattacharya, S. *et al.* ImmPort, toward repurposing of open access immunological assay data for translational and clinical research. *Sci. Data* **5**, 180015 (2018).
30. Thorsson, V. *et al.* The Immune Landscape of Cancer. *Immunity* **48**, 812-830.e14 (2018).
31. Newman, A. M. *et al.* Robust enumeration of cell subsets from tissue expression profiles. *Nat. Methods* **12**, 453–457 (2015).
32. Ravi, V. M. *et al.* Spatially resolved multi-omics deciphers bidirectional tumor-host interdependence in glioblastoma. *Cancer Cell* **40**, 639-655.e13 (2022).
33. Orasanu, C. I. *et al.* Implications of Cellular Immaturity in Necrosis and Microvascularization in Glioblastomas IDH-Wild-Type. *Clin. Pract.* **12**, 1054–1068 (2022).
34. Kamoun, A. *et al.* Integrated multi-omics analysis of oligodendroglial tumours identifies three subgroups of 1p/19q co-deleted gliomas. *Nat. Commun.* **7**, 11263 (2016).
35. Hsiang, J.-C., Shen, N., Soto, F. & Kerschensteiner, D. Distributed feature representations of natural stimuli across parallel retinal pathways. *Nat. Commun.* **15**, 1920 (2024).
36. Wang, G. *et al.* scMODAL: a general deep learning framework for comprehensive single-cell multi-omics data alignment with feature links. *Nat. Commun.* **16**, 4994 (2025).
37. Liu, T. *et al.* A Hybrid Deep Learning Model for Predicting Molecular Subtypes of Human Breast Cancer Using Multimodal Data. *IRBM* **43**, 62–74 (2022).

38. Wang, C.-W. *et al.* Deep learning to assess microsatellite instability directly from histopathological whole slide images in endometrial cancer. *Npj Digit. Med.* **7**, 1–13 (2024).
39. Le, M.-K. *et al.* Glioma Image-Level and Slide-Level Gene Predictor (GLISP) for Molecular Diagnosis and Predicting Genetic Events of Adult Diffuse Glioma. *Bioengineering* **12**, 12 (2025).
40. CD38-targeted therapy in glioblastoma: A step forward. | Journal of Clinical Oncology. https://ascopubs.org/doi/abs/10.1200/JCO.2018.36.15_suppl.e14030.
41. Imboden, S. *et al.* Implementation of the 2021 molecular ESGO/ESTRO/ESP risk groups in endometrial cancer. *Gynecol. Oncol.* **162**, 394–400 (2021).
42. Trpkov, K. *et al.* New developments in existing WHO entities and evolving molecular concepts: The Genitourinary Pathology Society (GUPS) update on renal neoplasia. *Mod. Pathol.* **34**, 1392–1424 (2021).

Methods

1. Ethics and data privacy

The TCGA (phs000178), CPTAC (phs001287), GBMatch (EK078-2004, EK550/2005, EK1412/2014, EK 27-147/2015) and IvyGAP (PRJNA420741) study protocols were approved by the Medical Ethical Committee at their participating centers. Protocols for the four in-house cohorts (i.e., TTH, SYH, CUH-R and CUH-P) were approved by the Ethics Committee of Tiantan Hospital (KY2023-160-02), Sun Yat-sen Hospital (B2024-346-01) and Cambridge University Hospital (PRaM-GBM: CRUK/A19732; SIND: NCT04007185). Notably, both low-grade and high-grade glioma patients were included in the PRaM-GBM and SIND programs for our study. For those enrolled in the prospective study, written consent was obtained, while for the retrospective study, a waiver of consent was signed when applicable. All research adhered to ethical standards, including the Declaration of Helsinki. Patients were informed about the use of AI software during screening, and their data were handled according to applicable privacy regulations, including the General Data Protection Regulation. Data from in-house cohorts were processed separately at each participating hospitals.

2. Cohorts

We included patients with primary brain tumors (excluding recurrent cases) using frozen and formalin-fixed paraffin-embedded (FFPE) tumor samples, along with clinicopathological data and molecular test results from four public cohorts and four in-house clinical cohorts (Table S1).

The TTH cohort recruited patients in China between February 2015 and January 2018. The CUH-R cohort included patients diagnosed in the UK from January 2019 to October 2021. The prospective CUH-P cohort recruited UK patients from November 2021 to February 2024, following the implementation of the WHO 2021 glioma diagnostic criteria. For both CUH-R and CUH-P, patients underwent primary tumor resection (complete or partial) and were assigned to treatment with concurrent chemoradiotherapy (CCRT), short-course radiotherapy (SCRT), chemotherapy, or palliative care. The prospective

SYH cohort recruited patients in China between January 2022 and July 2024, with all participants receiving molecular pathology testing and a WHO 2021-based diagnosis.

Among the public cohorts, the TCGA included 1,122 patients with diffuse gliomas diagnosed from 1989 to 2013, mostly from the United States, with smaller groups from Sweden, Italy, Canada, Brazil, Germany, and Australia. Most patients had tumor resection followed by radiotherapy and chemotherapy. The CPTAC cohort recruited 99 glioma patients between 2016 and 2021, primarily from Russia, China, Poland, and the United States, with some from the Philippines and Bulgaria. These patients underwent tumor resection and received various treatments, including radiotherapy, chemotherapy, molecular therapy, and immunotherapy. The IvyGAP cohort included 41 patients recruited between 2006 and 2013, while the GBMatch cohort recruited patients between 2001 and 2016; all underwent tumor resection surgery.

3. Datasets

In the present study, some cases were excluded from the supervised training, fine-tuning and testing of CAMPaS for both the public and in-house (both retrospective and prospective) cohorts based on the following criteria (Fig. S1). Generally, for all cohorts, the exclusion criteria include: (1) without available WSIs, (2) without any molecular markers of *IDH* mutation, *1p/19q* CODEL and *CDKN2A/B* HOMDEL, (3) WSIs with low image quality reviewed by our expert pathologist, i.e., no tumor, poor tissue quality or out-of-focus scanning. Besides the general exclusion criteria, for our prospective in-house cohorts (i.e., CUH-P and SYH), patients were further excluded with the following criteria: (1) suspicious histology of midline glioma, (2) without the integrated diagnosis according to the WHO 2021 criteria and (3) patients referred to other hospitals for treatment so that no prognostic history could be retrieved.

For all cohorts, at least one representative H&E-stained slide of the glioma resection or biopsy specimen was included for each patient depending on the availability of the tumor material. For the in-house cohorts, we collected one diagnostic H&E-stained tumor slides per patient case with glioma, each from an FFPE tumor tissue block. H&E slides were scanned at x40 magnification using Leica Aperio scanner AT-2 (resolution

0.2468 um per pixel), EasyScan scanner SZ-20 (resolution 0.2650 um per pixel) and Nanozoomer scanner s60 (resolution 0.2521 um per pixel), for CUH, SYH and TTH, respectively.

Following the aforementioned criteria, 3,367 patients (6,043 WSIs) were in our study: 940 (2,633 WSIs) from TCGA, 81 (192 WSIs) from CPTAC, 218 from (370 WSIs) GBMatch, 33 from (753 WSIs) IvyGAP, 1,115 (1,115 WSIs) from TTH, 422 (422 WSIs) from CUH-R, 354 (354 WSIs) from CUH-P and 204 (204 WSIs) from SYH. In addition, for the experimental setting of the enlarged training set, all cases of the eight cohorts were merged, named as the merged cohort.

4. Model performance evaluation

The threefold cross-validation was performed on the 2,633 WSIs reserved for supervised training, and the best-performing architecture and hyperparameters were selected based on the highest mean AUC across the folds. A receiver operating characteristic (ROC) curve analysis and an AUC analysis were done with 95% Wald CIs using the continuous AI-predicted malignancy score with bootstrapping (300 sample size and 50 sample times)

For further evaluation, six additional metrics with 95% confidence intervals were reported: accuracy, sensitivity, specificity, precision, recall, and F1-score. Notably, for all comparison methods, including CHARM, CLAM, TransMIL, and the CNN-based baseline, separate models were trained independently for histology grading/subtyping, molecular marker prediction, and integrated glioma diagnosis, as these methods could not jointly handle multiple tasks.

To evaluate the value of multiple WSIs per patient in the TCGA cohort, we further tested model robustness by training and validating with only a single WSI per patient. This WSI was either randomly selected or chosen as the most representative, defined as the slide with the lowest prediction uncertainty (i.e., the lowest Shannon entropy among a patient’s WSIs).

5. Cross-modal AI diagnostic model

8.1 Framework:

As shown in Extended Fig. 3, given the extracted representation features $\{\mathbf{F}_i\}_{i=1}^{N} \in \mathbb{R}^{N\times M}$ (with N = 2,500 patches and M = 1,024 for feature dimension) from the cropped patches $\{\mathbf{X}_i\}_{i=1}^{N} \in \mathbb{R}^{N\times H\times W\times 3}$ (with height H = 224 pixel, width W = 224 pixel and 3 for RGB channels), CAMPaS predicted: (1) Molecular markers, including *IDH* mutation ($\hat{l}_{idh} \in \mathbb{R}^2$), *1p/19q* CODEL ($\hat{l}_{1p/19q} \in \mathbb{R}^2$) and *CDKN2A/B* HOMDEL ($\hat{l}_{cdkn} \in \mathbb{R}^2$); (2) Histology grading ($\hat{l}_{grading} \in \mathbb{R}^3$) and subtyping ($\hat{l}_{subtyping} \in \mathbb{R}^3$); (3) Final integrated diagnosis of glioma ($\hat{l}_{glio} \in \mathbb{R}^6$). CAMPaS adopted a hierarchical multi-task learning framework based on a top-down vision transformer backbone. For translational efficacy, in this paper, we specially designed a bi-directional attention module and noise-robust learning module to solve the real-world challenges of dataset heterogeneity and partial missing of molecular markers, respectively. Details of these two modules are as follows.

8.2 Bi-directional attention module:

The bi-directional attention module was engineered to extract robust features from WSIs for predicting molecular markers as well as for histology-based grading and subtyping (Extended Fig. 3; Fig. S20). To mitigate dataset heterogeneity, this module dynamically identified critical WSI regions by applying thresholds to AI-generated importance maps. These selected regions then guided the reweighting of WSI patches, thereby suppressing redundant or noisy areas.

Specifically, the input patch-level features $\{\mathbf{F}_i\}_{i=1}^{N}$ were first processed through three cascaded transformer blocks to effectively extract WSI features. After an initial warm-up training stage (lasting for several epochs, i.e., 10 in this paper), the bi-directional attention mechanism was repeatedly applied throughout the remaining training phases. Within each iteration of the bi-directional attention guidance, AI importance maps highlighting the most discriminative regions for each target class were generated (See section 11 in Methods). These maps subsequently served as attention masks, directly multiplied with the features $\{\mathbf{F}_i\}_{i=1}^{N}$ in the subsequent epochs. By doing so, the model

progressively focused more precisely on class-specific informative regions for the downstream tasks, while simultaneously suppressing noisy regions arising from intra-tumor heterogeneity. For instance, in the histology grading task, dominant features indicative of high-grade tumor regions were prioritized, enabling the model to explicitly filter out distracting low-grade features within the WSIs. Details of generating the binarized attention guidance are elaborated as follows.

Taking the histology grading task as an example, we employed a channel-adaptive threshold truncation strategy on the AI importance maps of grade 2, 3 and 4, i.e., S_{G2}, S_{G3}, S_{G4}. For each channel c of the three grades, we computed the mean saliency conditioned on the network's predicted class y_{pred}:

$$\mu_{\text{pred}}^{(c)} = \mathbb{E}[S^{(c)} \mid y_{\text{pred}}]. \qquad (1)$$

Pixels in the raw importance map $S_{\text{raw}}^{(c)}$ of each grade exceeding this threshold were retained, whereas those below were set to the channel's minimum saliency $S_{\text{min}}^{(c)}$:

$$S_{\text{filtered}}^{(c)} = \begin{cases} S_{\text{raw}}^{(c)}, & S_{\text{raw}}^{(c)} \geq \mu_{\text{pred}}^{(c)}, \\ S_{\text{min}}^{(c)}, & S_{\text{raw}}^{(c)} < \mu_{\text{pred}}^{(c)}. \end{cases} \qquad (2)$$

A subsequent linear normalization projected the truncated map onto the unit interval:

$$S_{\text{norm}}^{(c)} = \frac{S_{\text{filtered}}^{(c)} - S_{\text{min}}^{(c)}}{S_{\text{max}}^{(c)} - S_{\text{min}}^{(c)}}, \qquad (3)$$

where $S_{\text{max}}^{(c)}$ denoted the channel's maximum saliency. This step simultaneously removed weak responses and unified saliency scales.

We then defined a fusion weight function based on the predicted grading category c, leveraging indicator functions $\mathbb{I}(\cdot)$ to assign full attention to the predicted class and partial attention to non-predicted classes:

$$W_{\text{saliency}}(x) = 1.0\ \mathbb{I}(S_{\text{norm}}^{(c)}(x) > 0) + 0.5 \sum_{k \neq c} \mathbb{I}(S_{\text{norm}}^{(k)}(x) > 0). \qquad (4)$$

Finally, the fused weight map W_{saliency} was applied to the initial feature map F_{init} via the Hadamard product to produce the binarized attention guidance F_{attn} of histology grading task:

$$F_{\text{attn}} = F_{\text{init}} \otimes W_{\text{saliency}}. \qquad (5)$$

Similarly, the binarized attention guidance for the histology subtyping task can also be derived.

Notably, the attention module incorporated a unified adaptive color-space normalization strategy[43] to mitigate domain shifts across datasets. This normalization dynamically adjusted the color distribution of input histology images to align with a reference domain, thereby ensuring consistent visual features across varying staining protocols and scanner conditions.

8.3 Noise-robust learning module:

Building on the robust features extracted by the bi-directional attention module, we introduced a novel noise-robust learning module to address label uncertainty caused by incomplete molecular annotations in real-world datasets. Within this framework, we incorporated the molecular graph subnet, histology prediction subnet, and cross-modal interaction subnet from our previous work[13,14] to support molecular marker prediction, histology grading/subtyping, and modeling histology-molecular relationships (Figs. S21-22).

Specifically, in the noise-aware training process, we first estimated the uncertainty level of each WSI based on the availability of its molecular marker labels. We achieved this through entropy-based quantification (including three steps of dual-network prediction, consistency-based categorization and entropy-guided noise assessment), where higher entropy indicated greater ambiguity or missing annotations, suggesting a higher likelihood of label noise. Based on these estimates, we categorized WSIs as either noisy or clean. Then, we adopted uncertainty-adapt learning. For noisy samples, we applied an adaptive loss modulation strategy to down-weight the impact of unreliable labels, thereby reducing the risk of overfitting to noise. In contrast, we trained clean

samples using standard cross-entropy loss to provide accurate supervision. This overall procedure formed an uncertainty-adaptive learning framework, dynamically calibrating the learning signal according to the estimated label quality.

6. Model training and validation strategy

We trained and assessed our model in three stages as follows.

Stage I: Training. To empower the AI model for effective cross-modal modeling under the WHO 2021 criteria, we first trained our model on TCGA, a large-scale multicenter cohort with both molecular markers and histology grading..

Stage II: Fine-tuning. To enhance the generalizability of our model on heterogeneous datasets, we fine-tuned and tested our model on five retrospective cohorts, which demonstrate heterogeneity in color and cellular distributions. In model finetuning, we included all samples meeting the general inclusion and exclusion criteria.

Stage III: Testing. To further evaluate the model in real-world clinical scenarios, we recruited two prospective cohorts diagnosed under WHO 2021 criteria.

7. Association with multimodal diagnostic factors analysis

For analyzing the association of the AI-predicted malignancy scores with diagnostic factors (i.e., clinicopathological factors and molecular markers), patients were first clustered into three subgroups according to their AI predicted malignancy with agglomerative clustering[25]. This unsupervised approach iteratively merged with the closest clusters based on score similarity, minimizing within-cluster variance. We determined the optimal number of clusters (k = 3) based on the dendrogram structure and interpretability in clinical context, resulting in low, intermediate, and high malignancy groups.

Then, integrated visualization of AI-predicted malignancy scores and related clinical, molecular, and histopathological features (for TCGA, SYH and CUH-P cohorts) was performed (Fig. 3a and Fig. S7). We further performed multiple single linear regression analyses using the CAMPaS continuous malignancy scores as the dependent variable and the clinicopathological data as the regressor. Statistical tests were two sided with

statistical significance accepted with *P* values < 0.050. Regression coefficients and exact *P* values have been reported in Table S13. Of note, in Fig. S7, the potential diagnostic molecular markers (both CNV- and SNP-level) were selected as the markers with at least 5% alteration proportion, from the cBioPortal.

8. Generating AI importance maps under MIL settings

To obtain the AI importance map, following a well-recognized MIL-based network visualization method[44], we utilized the mechanisms of class activation map (CAM) method[45] and extended it to the MIL settings for visualizing the decisions of the AI model for both tasks of histology grading, subtyping and molecular markers prediction. For WSIs with fewer than 2,500 original patches, we averaged the importance scores of the 2,500 patches assigned to each original patch. For WSIs with more than 2,500 original patches, we applied bilinear interpolation to align the 2,500 importance scores with the original patch grid, ensuring accurate and consistent visualization.

9. AI-predicted malignancy score

Based on the AI prediction of both molecular markers and histology grades/subtypes, the AI-predicted malignancy score indicated the probability of being high grade in WHO 2021 criteria. Specifically, it can be derived following the diagnostic scheme of WHO 2021, i.e., the predicted probability of Grade 4 GBM and Grade 4 Astrocytoma. Formally, the malignancy score $\hat{l}_{malig} \in \mathbb{R}^1$) was defined as follows.

$$\hat{l}_{malig} = \hat{l}_{idh}^{wt} + \hat{l}_{idh}^{mut} \times \hat{l}_{1p/19q}^{non-codel} \times (\hat{l}_{grading}^{IV} \times \hat{l}_{cdkn}^{non-homdel} + (1 - \hat{l}_{grading}^{IV}) \times \hat{l}_{cdkn}^{homdel} + \hat{l}_{grading}^{IV} \times \hat{l}_{cdkn}^{homdel}) \quad (6)$$

In the above equation, $\hat{l}_{idh}^{wt}$, $\hat{l}_{idh}^{mut}$, $\hat{l}_{1p/19q}^{non-codel}$, $\hat{l}_{cdkn}^{homdel}$, $\hat{l}_{cdkn}^{non-homdel}$ and $\hat{l}_{grading}^{IV}$ indicated the AI predicted probability of *IDH* wildtype, *IDH* mutated, *1p/19q* CODEL, *CDKN2A/B* HOMDEL, *CDKN2A/B* non-HOMDEL and histology high-grade, respectively.

10. Morphological interpretability analysis

To investigate morphological correlations underlying the AI decision, AI importance maps generated by the AI model for histology grading were analyzed. Specifically, the

PCC was calculated between the spatially averaged AI importance maps and 50 commonly utilized histopathological metrics, as annotated by expert pathologists in the GBMatch cohort (Fig. 3d, Fig. S9). To rigorously evaluate the interpretability and precision of our model, we compared its performance with CHARM (ViT-based) and a CNN-based model. Notably, in subsequent association analyses utilizing a subset of the most important AI features, the top K% (K = 5, 25, or 50) elements of the layer prior to the MLP layers were selected according to their values.

To further validate these morphological insights, we quantitatively assessed neuropathologist-annotated histological features, particularly the percentage of tumor nuclei and necrosis area, against AI-defined malignancy subtypes in the TCGA cohort (Fig. 3e-f). Subgroup analyses stratified by clinical parameters, including age and sex, consistently reinforced these observations, further underscoring the clinical relevance and robustness of the identified morphological associations (Fig. S10).

11. Bulk-level and spatially resolved transcriptomic correlation analysis

Bulk gene expression-level association. To explore bulk-level transcriptomic correlates of AI's decision on molecular markers, we performed comprehensive analyses integrating AI importance maps with gene expression data (normalized with unstranded Transcripts Per Million (TPM)) from the TCGA cohort. Initially, bulk-level transcriptomic associations were evaluated by calculating PCC between the spatially averaged AI importance maps and expression levels of selected brain tumor-related and immune-related genes. Gene sets utilized for these analyses included 140 brain tumor-associated genes (including potential diagnostic marker genes in WHO 2021[3], and transcriptomics-level GBM subtyping marker genes[28]) and top 100 highly variable immune-related genes derived from established databases, such as ImmPort[29] (Fig. 3f-g; Figs. S11-12). Gene-wise correlation significance was assessed using the PCC metric, and statistical significance was established using a threshold of $P < 0.05$.

Immune cells-level association. CIBERSORT[31] was used for estimating the relative abundance of 22 common human immune cell types, using the bulk transcriptomics data in TCGA cohort. Then, the association between the AI-predicted malignancy

scores and each immune cell subset was performed using the log2 (transformed proportion of the immune cell subset) as a fraction of the whole tumor, using the leukocyte fraction values. Linear regressions were performed with the CAMPaS continuous risk scores as the independent variable. Two-sided P values <0.050 are accepted as significant. Regression coefficients and exact P values have been reported in Table S14.

Differentially expressed genes were assessed between CAMPaS high-risk and CAMPaS low-risk cases by Scipy Python package (v.1.11.1) and statsmodels Python package (v.0.14.1). Genes with a likelihood ratio test *P* value adjusted using a Benjamini-Hochberg false discovery rate (FDR) were accepted if <0.050.

Spatial transcriptomics-level association. For spatially resolved transcriptomic correlation, we utilized brain tumor spatial transcriptomic (ST) profiling data, along with spatially aligned H&E-stained WSIs, from a published work[32], generated using the 10X Visium[46] platform. Raw data preprocessing included normalization, scaling, and dimensionality reduction to produce spatial gene expression profiles. The paired WSIs were preprocessed for AI prediction using the same procedures as those for the other cohorts. Decision-relevant spatial regions (for molecular marker prediction) identified by AI from histology were then compared to these spatial gene expression patterns (Fig. 4d). Specifically, we computed cosine similarity scores between the AI importance maps and the spatial expression patterns of key tumor-related and immune-related genes, thereby quantitatively assessing the spatial coherence between AI predictions and gene expression profiles.

12. Prognostic efficacy analysis

The prognostic efficacy of AI-predicted malignancy features was systematically assessed through survival analyses in both retrospective TCGA cohort and prospective clinical CUH-P cohorts. The prognostic significance was evaluated for OS and PFS using KM survival analysis[47] and two-sided log-rank tests with statistical significance accepted for $P < 0.050$. In continuous AI malignancy survival analysis, cutoffs for the CAMPaS malignancy groups were defined by taking the tertiles of the distribution of

CAMPaS predicted malignancy scores in the TCGA and CUH-P cohort. Survival outcomes across these groups were compared, and pairwise log-rank tests were conducted to establish significant differences among these groups.

Additionally, univariate and multivariate Cox proportional hazards regression analyses were performed to quantify the prognostic impact of continuous AI-predicted malignancy scores and other clinical (sex and age), histopathological (*Ki-67* index), and molecular variables (*IDH* mutation, *1p/19q* CODEL and *MGMT* methylation) established for glioma diagnosis. Hazard ratios (HRs), along with 95% CIs and corresponding *P* values, were calculated to determine the strength and independence of AI-predicted malignancy scores as prognostic indicators.

To evaluate the incremental prognostic value of AI-derived features, a logistic regression model taking the learned CAMPaS features and clinical diagnostic factors (including sex, age, histology grading and subtyping, *IDH* mutation *1p/19q* CODEL and *CDKN2A/B* HOMDEL status) was trained to predict median OS in both TCGA and prospective CUH-P cohorts. The model was trained with a learning rate of 0.003 and a batch size of 4 on the same device as in training the CAMPaS. The compared models were logistic regression models taking only the clinical diagnostic factors, the clinical diagnostic factors and the learned CHARM features, or the clinical diagnostic factors and the learned baseline CNN-based features as input. Predictive performance was quantified using ROC analysis, with statistical confidence assessed using 95% CIs.

To evaluate the prognostic relevance of the features learned by CAMPaS, we computed malignancy scores for patients in the CUH-P treatment arm who received standard CCRT (n=135). PFS probabilities were analyzed using the KM-based method. To assess treatment outcomes, patients were stratified into malignancy groups based on CAMPaS-predicted scores. Prognostic discrimination was compared against conventional histology-based grading, serving as a baseline reference.