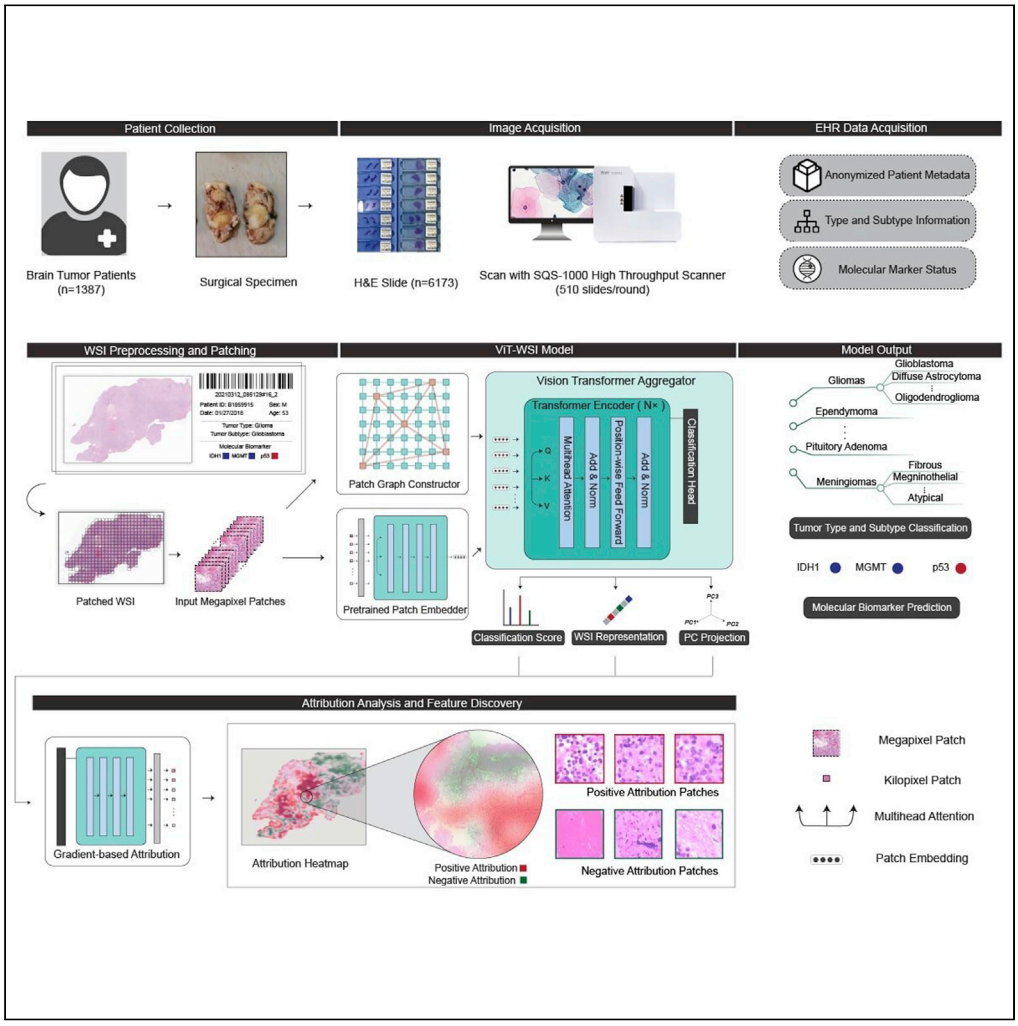
iScience

CellPress
OPEN ACCESS

Article

Vision transformer-based weakly supervised histopathological image analysis of primary brain tumors



Zhongxiao Li, Yuwei Cong, Xin Chen, ..., Yupeng Chen, Shiguang Zhao, Xin Gao

qijiping2003@163.com (J.Q.)
larry.carin@kaust.edu.sa (L.C.)
chenyp@sustech.edu.cn (Y.C.)
guangsz@hotmail.com (S.Z.)
xin.gao@kaust.edu.sa (X.G.)

Highlights

ViT-WSI, is suitable for weakly supervised learning on histopathological images

ViT-WSI performs tumor typing, subtyping and molecular marker prediction

ViT-WSI automatically discovers brain tumor histological features

Li et al., iScience 26, 105872
January 20, 2023 © 2022 The Author(s).
https://doi.org/10.1016/j.isci.2022.105872

iScience

CellPress
OPEN ACCESS

Article

Vision transformer-based weakly supervised histopathological image analysis of primary brain tumors

Zhongxiao Li,[1,2,10] Yuwei Cong,[3,10] Xin Chen,[4,10] Jiping Qi,[3,10,*] Jingxian Sun,[4] Tao Yan,[4] He Yang,[4] Junsi Liu,[4] Enzhou Lu,[4] Lixiang Wang,[4] Jiafeng Li,[4] Hong Hu,[4] Cheng Zhang,[5] Quan Yang,[4] Jiawei Yao,[4] Penglei Yao,[4] Qiuyi Jiang,[4] Wenwu Liu,[4] Jiangning Song,[6,7] Lawrence Carin,[1,*] Yupeng Chen,[8,*] Shiguang Zhao,[4,9,*] and Xin Gao[1,2,11,*]

SUMMARY

Diagnosis of primary brain tumors relies heavily on histopathology. Although various computational pathology methods have been developed for automated diagnosis of primary brain tumors, they usually require neuropathologists' annotation of region of interests or selection of image patches on whole-slide images (WSI). We developed an end-to-end Vision Transformer (ViT) – based deep learning architecture for brain tumor WSI analysis, yielding a highly interpretable deep-learning model, ViT-WSI. Based on the principle of weakly supervised machine learning, ViT-WSI accomplishes the task of major primary brain tumor type and subtype classification. Using a systematic gradient-based attribution analysis procedure, ViT-WSI can discover diagnostic histopathological features for primary brain tumors. Furthermore, we demonstrated that ViT-WSI has high predictive power of inferring the status of three diagnostic glioma molecular markers, *IDH1* mutation, *p53* mutation, and *MGMT* methylation, directly from H&E-stained histopathological images, with patient level AUC scores of 0.960, 0.874, and 0.845, respectively.

INTRODUCTION

Definitive diagnosis of primary brain tumor almost always requires histopathology. Histopathological diagnosis of tumors usually requires pathologists with years of experience to manually examine histological details at various levels of magnification and is inherently laborious. This is further complicated by the diversity of brain tumor histological subtypes and the subtlety to differentiate among them. In this way, pathologists are met with complicated criteria for diagnosis, and subjectivity becomes unavoidable. Therefore, significant interobserver variability has been observed retrospectively in brain tumor diagnosis, especially among gliomas and meningiomas.[1,2] In recent years, because of the improved understanding of tumorigenesis and advances in molecular biology experimental techniques, molecular biology assays are having a more important role in the brain tumor diagnosis workflow. As specified in the 2016 WHO classification of tumors of the central nervous system (CNS)[3], some molecular biomarkers, such as somatic mutations in isocitrate dehydrogenase (IDH) and 1p/19q co-deletion, have become essential in the diagnosis of certain subtypes of glioma. Compared to histopathological analysis, molecular biology assays, although much more objective and reliable, are often costly and may have less availability to economically underdeveloped regions.[4]

In recent years, significant advances in digital pathology hardware have pushed the field of histopathology into the 'big data' era. More and more healthcare institutes worldwide have adopted the usage of high-throughput microscopic slide scanners in their daily workflow, in which histopathology slides are digitalized into whole-slide images (WSIs) for reliable long-term data storage.[5] Correspondingly, on the software side, new computational methods have begun to emerge for automatic and objective histopathology image analysis.

With the advance of deep learning-based computer vision algorithms, much research has been conducted to automate pathologists' diagnosis from different aspects, from the early works on tumor segmentation[6] to histopathological subtyping,[7] grading,[8] and prognosis.[9] More recently, deep learning-based methods

[1]Computer Science Program, Computer, Electrical and Mathematical Sciences and Engineering (CEMSE) Division, King Abdullah University of Science and Technology (KAUST), Thuwal 23955-6900, Kingdom of Saudi Arabia

[2]KAUST Computational Bioscience Research Center (CBRC), King Abdullah University of Science and Technology (KAUST), Thuwal 23955-6900, Kingdom of Saudi Arabia

[3]Department of Pathology, The First Affiliated Hospital of Harbin Medical University, 23 Youzheng Street, Nangang District, Harbin 150001, People's Republic of China

[4]Department of Neurosurgery, The First Affiliated Hospital of Harbin Medical University, Harbin, Heilongjiang Province 150001, China

[5]Suffolk University, Boston, MA, USA

[6]Biomedicine Discovery Institute and Department of Biochemistry and Molecular Biology, Monash University, Melbourne, VIC 3800, Australia

[7]Monash Data Futures Institute, Monash University, Melbourne, VIC 3800, Australia

[8]School of Microelectronics, Southern University of Science and Technology, Shenzhen 518055, PR China

[9]Department of Neurosurgery, Shenzhen University General Hospital, Shenzhen, Guangdong Province 518100, China

[10]These authors contributed equally

Continued

This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

have been shown to 'see the unseen' from H&E-stained histopathology slides, in which they are shown to have predictive power over the presence of underlying biomarkers, such as somatic mutations,[10] microsatellite instability,[11] and tumor mutational burden.[12] In this way, the computational methods can serve as a low-cost alternative or secondary verification of the costly molecular biology assays.

From the algorithmic point of view, early works on computational pathology are mainly based on supervised learning algorithms.[10,13] They usually require pathologists' annotation of region of interests (ROIs) or selection of image patches on WSIs. The selected ROIs or image patches are then used to train a supervised machine learning algorithm for inference tasks on histopathology images. Owing to the requirement of supervision from pathologists, such methods have limitations in real-world applications because accurate annotations from pathologists are difficult to obtain. In light of the weaknesses of such methods, a lot of the so-called 'weakly supervised learning algorithms' have been developed in recent years. Instead of requiring pathologists' detailed annotation at patch-level or ROI-level, such algorithms only require one annotation per slide, which can be easily obtained directly from the patients' electronic health records (EHRs). Such methods have already had successful applications in tumor detection,[14] histopathological subtyping,[15] and tumor origin prediction.[16]

The development of computational pathology methods on primary brain tumors has been lagging considerably behind other tumor types. This is in part because of the difficulty in obtaining large annotated pathology datasets for brain tumors. The incidence of brain tumors and other CNS tumors is generally lower than tumors of other origins.[17] There is also a lack of publicly available high-quality brain tumor histopathology datasets, with the exception of several glioma subtypes in The Cancer Genome Atlas (TCGA).[18] Therefore, existing works have mainly focused on the classification tasks in a few primary brain tumor subtypes, mostly only within glioma subtypes.[19,20] Despite the demonstrated success in other brain tumor imaging modalities,[21,22] there is a lack of systematic investigation of applying weakly supervised learning algorithms on primary brain tumor H&E histopathology.

To fill the gap in brain tumor computational pathology, we here propose ViT-WSI, a highly interpretable and weakly supervised model, leveraging the state-of-the-art Vision Transformer architecture in an end-to-end manner, for brain tumor WSI analysis. ViT-WSI achieves the task of weakly supervised learning through the self-attention mechanism[23] and a constructed patch-level graph of the WSI, which benefits ViT-WSI by modeling the relationship between WSI patches in a context-aware fashion. Importantly, we demonstrate how a weakly supervised ViT-WSI, learned from H&E WSI and labels purely extracted from electronic health records (EHRs) and without any additional pathologist supervision, accomplishes the task of major primary brain tumor type and subtype classification. We develop a systematic procedure to interpret the ViT-WSI model and show how ViT-WSI automatically discovers diagnostic histopathological features directly from WSI. Finally, the ViT-WSI model fine-tuned on additional evidence from Immunohistochemistry (IHC) and molecular biology assays, can 'see the unseen' from H&E histopathology and precisely predict the status of three diagnostic glioma molecular markers.

RESULTS

Vision transformer (ViT)–based weakly supervised whole-slide image analysis model (ViT-WSI)

To leverage the strong performance of visual recognition of Vision Transformers,[24] we designed the ViT-WSI model for weakly-supervised histopathology image analysis (Figure 1). ViT-WSI can be trained in a weakly-supervised fashion to perform various brain tumor classification tasks. ViT-WSI first segments a gigapixel WSI into 1024×1024 megapixel patches (Figure 1). The first part of ViT-WSI, the Pretrained Patch Embedder (Figure 1), extracts image features of each megapixel patch and embeds them into a 1024-dimensional vector. The Pretrained Patch Embedder is a ViT-L-16 model pretrained on the ImageNet-21k dataset.[25] Each embedded megapixel patch is then sent to the second part of ViT-WSI, the Vision Transformer Aggregator (Figure 1), which performs whole-slide aggregation for the weakly supervised learning task. Vision Transformer Aggregator is also designed as a transformer architecture, aiming to better utilize the extracted features from the Pretrained Patch Embedder. During this aggregation phase, the self-attention mechanism of the Vision Transformer Aggregator models the interaction and relationship among patches. Feature similarity and closeness of physical locations between the patches are also taken into consideration by constructing a nearest neighbor graph of the patches and are input to the Vision Transformer Aggregator via embedding and attention computation (Figure 1). In this way, the

[11]Lead contact

*Correspondence:
qijiping2003@163.com (J.Q.),
larry.carin@kaust.edu.sa (L.C.),
chenyp@sustech.edu.cn (Y.C.),
guangsz@hotmail.com (S.Z.),
xin.gao@kaust.edu.sa (X.G.)

https://doi.org/10.1016/j.isci.2022.105872

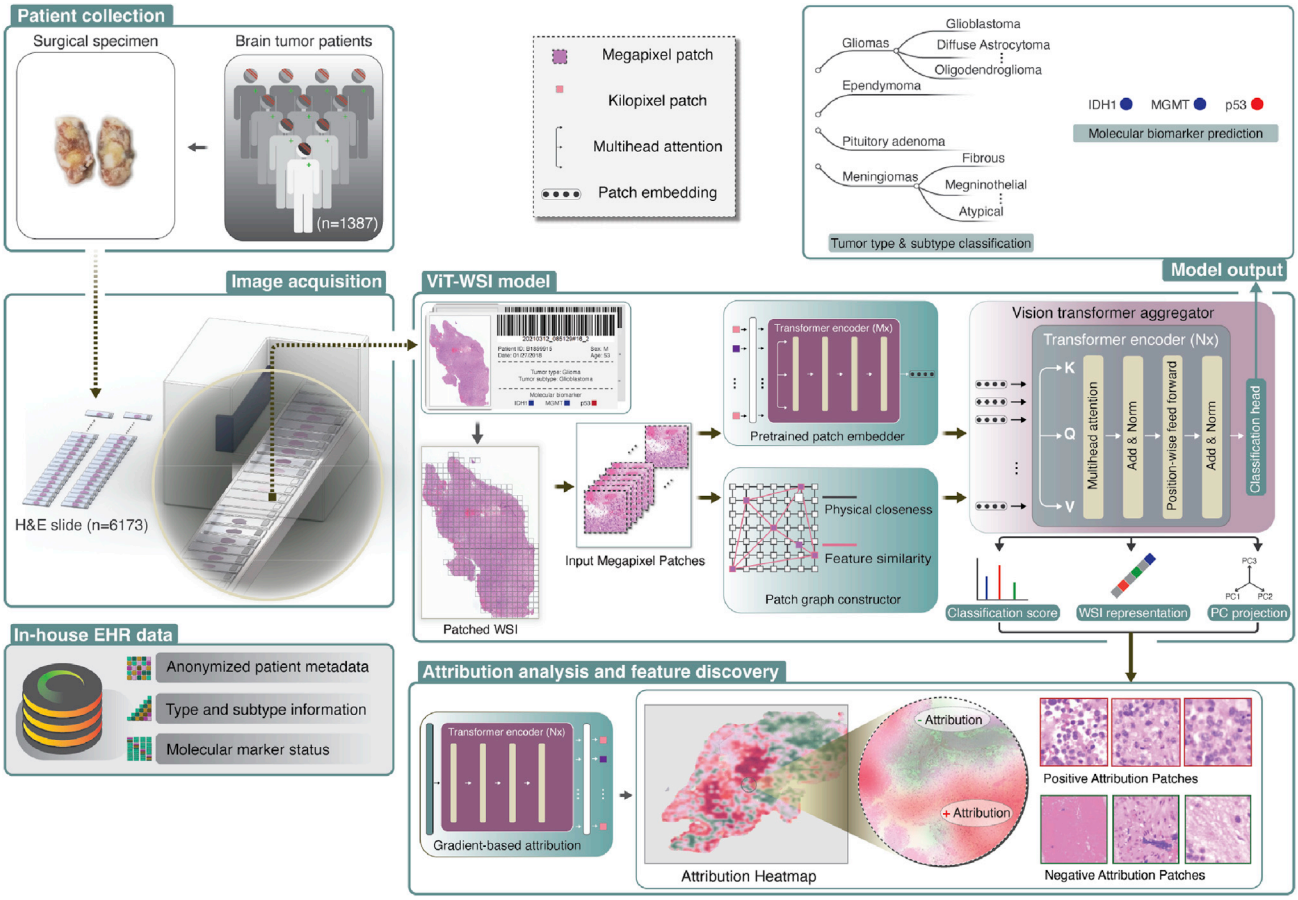


Figure 1. An overview of the primary brain tumor dataset, analysis pipeline, and architecture of ViT-WSI

Left: A total number of 6,173 H&E slides of surgical specimens from 1,387 patients were retrieved from inventory for scanning. The H&E slides were scanned with an SQS-1000 high-throughput scanner which can handle 510 slides per round. Anonymized patient metadata (slide identifiers, primary diagnosis, age, sex, etc.), type and subtype information, and molecular marker status were exported and extracted from the patients' EHR data. The digitalized H&E slides and extracted information were used for model training and evaluation. Right: After a WSI's tissue region is segmented into patches, the analysis pipeline of the ViT-WSI model (the 'ViT-WSI model' box) can be divided into two stages: The first stage utilizes a Pretrained Patch Embedder that is responsible for transforming the WSI patches ('Input Megapixel Patches') into token embeddings that will be used in the second stage. Internally, it divides the patches into kilopixel patches. A Patch Graph Constructor constructs a nearest-neighbor graph of the patches based on their physical closeness (white edges) and feature similarity (pink edges). In the second stage, a Vision Transformer Aggregator aggregates the information across the WSI patches and summarizes them into a single, slide-level prediction. Both the Pretrained Patch Embedder and Vision Transformer Aggregator consist of multiple stacked Transformer Encoders (see STAR Methods). The multi-head self-attention in the Vision Transformer Aggregator, together with the Graph Constructor, aggregate patch-level information from WSIs in a context-aware fashion. Classification output is produced by the classification head attached to the last layer of the network and is used for the slide-level prediction tasks (summarized in the 'Model output' box). Finally, various intermediate quantities (e.g., the classification score, dimensions of WSI representation, and PC projections) are used for downstream interpretation of the model (the 'Attribution analysis and feature discovery' box). A gradient-based attribution algorithm is run through the model in a backward pass and assigns attribution scores to the input megapixel patches. Histological image feature discovery can be performed by inspecting the patches with high positive/negative attribution scores.

self-attention mechanism in the Vision Transformer Aggregator is able to aggregate the patch-level information in a context-aware manner. Finally, the classification scores corresponding to each class can be outputted from the classification head, and the output from the penultimate layer under the classification head is used as the whole-slide representation vector (WSRV) of the WSI. Detailed model configurations, design principles, and training protocol are discussed in STAR Methods.

Acquisition of the primary brain tumor histopathology WSI dataset

To overcome the lack of publicly available primary brain tumor datasets, we retrieved 6,173 primary brain tumor H&E slides from The First Hospital of Harbin Medical University (Figure 1). The slides were formalin-fixed paraffin-embedded (FFPE) and prepared between April 2017 and April 2020. They were scanned with

a high-throughput slide scanner, SQS-1000, developed by Shenzhen ShengQiang Technology Co., Ltd (https://www.sqray.com/), and then the digitalized H&E slides were converted to public file format (the 'svs' format) for downstream processing (Figure 1, STAR Methods). Our in-house dataset covers a total of eight brain tumor types, including glioma, meningioma, pituitary adenoma, ependymoma, craniopharyngioma, CNS lymphoma, chordoma, and germ cell tumor (Figure S1). Table 1 summarizes multiple statistics of this in-house cohort. We retrieved the corresponding EHRs of the patients which contain the diagnosis of the patient by the department of pathology of The Affiliated First Hospital of Harbin Medical University. All cases were carefully reviewed by a panel of seasoned neuropathologists at the hospital and the diagnosis was the one on which subsequent treatment was based. We kept only the slides that could be matched to their patient metadata and whose brain tumor type could be determined from the EHRs (Table 1, Figure S1). In total, we assembled a cohort of 5,216 slides from 1,211 patients spanning the eight brain tumor types (Table 1, Figure S1). The consent forms of the patients were waived before this research was carried out under the retrospective research protocol of the institution. Information on eleven glioma and meningioma subtypes and glioma molecular biomarkers were further extracted from the EHR for a subset of 597 patients (Table 1, Figure S1). The eleven subtypes span five glioma subtypes (Diffusive Astrocytoma (DA), Anaplastic Astrocytoma (AA), Oligodendroglioma (O), Anaplastic Oligodendroglioma (AO), Glioblastoma (GBM)) and six meningioma subtypes (Fibrous, Meningothelial, Transitional, Angiomatous, Atypical, Anaplastic). The hierarchy of the brain tumor types and subtypes is shown in Figure S2. Consistent with prior knowledge, the glioma subset contains more male patients (55.21%) than female patients (44.78%), whereas the meningioma subset contains more female patients (74.30%) than male patients (25.69%) (Table 1). For the training and evaluation of ViT-WSI, we subsequently refer to the classification of 8-class brain tumor types as the '8-class top-level type classification task', and the classification of 11-class meningioma and glioma subtypes as the '11-class subtyping task' (statistics shown in Table 1).

Previous work in lung adenocarcinoma suggests a general possibility of predicting molecular biomarkers directly from the H&E histopathology.[10] In glioma, three molecular biomarkers are of particular interest to pathologists: Isocitrate dehydrogenase (IDH), Tumor suppressor p53 (*TP53*), and O^6-methylguanine DNA methyltransferase (*MGMT*). Of the mutations in IDH, the most common one is the R132H mutation in the *IDH1* allele and is a frequent somatic mutation event in multiple astrocytic and oligodendroglial subtypes that imply better prognosis.[3] *TP53* is the tumor suppressor gene that frequently mutates in high-grade gliomas that are associated with higher malignancy and poorer prognosis.[26] Methylation at the promoter region of *MGMT*, a gene crucial in DNA repair pathways, silences its expression and makes tumor cells more sensitive to alkylating agents used in chemotherapy.[27] Recent work showed that patch-level classifiers trained on glioma histopathology images could predict IDH mutation with decent accuracy.[28,29] Glioma MRI images are also shown to have certain predictive power of *MGMT* methylation.[30] These suggest the possibility of phenotype-to-genotype inference in glioma subtypes. To investigate the possibility of ViT-WSI to make such an inference from histopathology slides, we further extracted the status of molecular biomarkers from the glioma patients' EHR data (statistics shown in Table 2). A total of 304 cases presented with their *IDH1*mutation status determined either by IHC (108 positive cases, 196 negative cases, with positivity indicating mutation in the *IDH1* gene) or Sanger sequencing (13 cases with mutation and 13 cases without mutation). 219 cases presented with their *TP53*mutation status, either by IHC (104 positive cases, 115 negative cases, with positivity indicating mutation in the *TP53* gene) or Sanger sequencing (showing 3 cases without *TP53*mutation). 71 cases presented with *MGMT* promoter methylation status, either by IHC (46 positive cases, 25 negative cases, with negativity suggesting *MGMT* promoter methylation) or Methylation Specific PCR[31] (MSP) (11 cases with methylation and 7 cases without methylation). These cases and their associated slides are then used to train and evaluate ViT-WSI for the molecular biomarker prediction task. As there are much more patients with IHC testing than Sanger sequencing or MSP, we use the status of the molecular biomarker determined from IHC for training and performance evaluation, and additionally validate them using the status of the biomarker reported by other methods.

We further retrieved 1,703 glioma FFPE slides of 879 patients from the subdirectories TCGA-GBM (for glioblastoma cases) and TCGA-LGG (for lower grade glioma cases) of The Cancer Genome Atlas (TCGA) as an independent data source (statistics shown in Table 1). Excluding the ambiguous subtype 'mixed glioma', the slides from the other subtypes (AA, AO, DA, GBM, and O) are assembled into a 5-class classification task dataset, hereafter referred to as the 'TCGA glioma subtyping task'. To get the patients' somatic mutation status in *IDH1* and *TP53*, we used somatic mutations that are independently called by four methods, MuSE,[32] MuTect2,[33] SomaticSniper,[34] and Varscan2[35] on their associated whole exome sequencing data.

Table 1. Statistics of histopathology datasets used in this study

	Patient level summary	Slide level summary
In-house primary brain tumor dataset statistics:		
Slide Scan Statistics:		
Retrieved from Inventory	1387	6173
Scanned WSI w/EHR Found	1247	5502
Scanned WSI w/Good Visual Quality	1221	5297
Primary Brain TumorType Statistics (8-class top-level type classification task):		
Scanned WSI w/Primary Brain TumorType Information	1211	5216
Glioma	326 (26.92%)	1683 (32.27%)
Meningioma	471 (38.89%)	2691 (51.59%)
Pituitary Adenoma	263 (21.71%)	424 (8.13%)
Ependymoma	38 (3.14%)	155 (2.97%)
Craniopharyngioma	69 (5.70%)	144 (2.76%)
CNS Lymphoma	14 (1.15%)	52 (1.00%)
Chordoma	18 (1.49%)	39 (0.75%)
Germ Cell Tumor	12 (0.991%)	28 (0.54%)
Used in the '8-class top-level type classification task'	1211 (100%)	5216 (100%)
Patient Statistics		
Sex:		
Female	698 (57.64%)	
Male	504 (41.62%)	
Not Available	9 (0.74%)	
Age:	51.62 $\pm$ 13.13 years	
Glioma Subtype Statistics (11-class subtyping task):		
All Scanned Gliomas	326	1683
Anaplastic Astrocytoma (AA)	10 (3.06%)	43 (2.55%)
Anaplastic Oligodendroglioma (AO)	28 (8.58%)	138 (8.19%)
Diffuse astrocytoma (DA)	29 (8.89%)	187 (11.11%)
Glioblastoma (GBM)	137 (42.02%)	650 (38.62%)
Oligodendroglioma (O)	27 (8.28%)	89 (5.28%)
Used in the '11-class subtyping task'	231 (70.85%)	1107 (65.77%)
Others/Not Available	95 (29.14%)	576 (34.22%)
Patient Statistics:		
Sex:		
Female	146 (44.78%)	
Male	180 (55.21%)	
Age:		
<20 years	5 (1.53%)	
20–30 years	14 (4.29%)	
30–40 years	54 (16.56%)	
40–50 years	90 (27.6%)	
50–60 years	88 (26.99%)	
60–70 years	63 (19.32%)	
>70 years	12 (3.68%)	

(*Continued on next page*)

Table 1. *Continued*

	Patient level summary	Slide level summary
Meningioma Subtype Statistics (11-class subtyping task):		
All Scanned Meningiomas	471	2691
Meningothelial Meningioma	74 (15.71%)	295 (10.96%)
Fibrous Meningioma	94 (19.95%)	410 (15.23%)
Transitional Meningioma	87 (18.47%)	330 (12.26%)
Angiomatous Meningioma	31 (6.58%)	94 (3.49%)
Atypical Meningioma	56 (11.88%)	462 (17.16%)
Anaplastic Meningioma	24 (5.09%)	56 (2.08%)
Used in the '11-class subtyping task'	366 (77.7%)	1647 (61.2%)
Others/Not Available	105 (22.29%)	1044 (38.79%)
Patient Statistics:		
Sex:		
Female	350 (74.30%)	
Male	121 (25.69%)	
Age		
<20 years	1 (0.21%)	
20–30 years	11 (2.33%)	
30–40 years	32 (6.79%)	
40–50 years	98 (20.80%)	
50–60 years	177 (37.57%)	
60–70 years	131 (27.81%)	
>70 years	21 (4.45%)	
TCGA Glioma Dataset Statistics:		
All TCGA Gliomas (with FPPE slides)	879	1703
TCGA-GBM		
Glioblastoma (GBM)	389 (42.45%)	860 (50.49%)
TCGA-LGG		
Anaplastic Astrocytoma (AA)	122 (13.88%)	164 (9.63%)
Anaplastic Oligodendroglioma (AO)	72 (8.19%)	155 (9.10%)
Diffuse astrocytoma (DA)	59 (6.71%)	104 (6.11%)
Oligodendroglioma (O)	107 (12.17%)	204 (11.97%)
Mixed Glioma	130 (14.78%)	216 (12.68%)
Used in the 'TCGA glioma subtyping task'	749 (85.21%)	1487 (87.31%)
Patient Statistics:		
Sex:		
Female	367 (41.75%)	
Male	512 (58.24%)	
Age	50.54 ± 15.54 years	
<20 years	10 (1.14%)	
20–30 years	78 (8.87%)	
30–40 years	180 (20.48%)	
40–50 years	157 (17.86%)	
50–60 years	197 (22.41%)	
60–70 years	160 (18.20%)	
>70 years	96 (10.92%)	
Not Available	1 (0.11%)	

Table 2. Statistics of available molecular biomarkers in glioma datasets

Glioma Cases Among In-house Primary Brain Tumor Dataset Statistics:

	IDH1		*TP53*		*MGMT*	
	IHC	Sanger	IHC	Sanger	IHC	MSP
Anaplastic Astrocytoma (AA)	+ (4) - (6)		+ (7)		+ (2)	
Anaplastic Oligodendroglioma (AO)	+ (21) - (7)		+ (3) - (13)		+ (1) - (3)	
Diffuse astrocytoma (DA)	+ (22) - (7)	mut (7)	+ (12) - (9)		- (1)	w/met (1)
Glioblastoma (GBM)	+ (12) - (117)	mut (2) wt (9)	+ (57) - (33)	wt (2)	+ (36) - (17)	w/met (6) w/o met (5)
Oligodendroglioma (O)	+ (9) - (7)	mut (1)	- (13)	wt (1)	+ (1)	
Others	+ (38) - (54)	mut (3) wt (4)	+ (25) - (45)		+ (6) - (4)	w/met (5) w/o met (1)
Total	+ (106) - (198)	mut (13) wt (13)	+ (104) - (115)	wt (3)	+ (46) - (25)	w/met (11) w/o met (7)

TCGA Glioma datasets statistics:

	IDH1	TP53	MGMT
Anaplastic Astrocytoma (AA)	mut (70) wt (48)	mut (56) wt (50)	w/met (21) w/o met (11)
Anaplastic Oligodendroglioma (AO)	mut (54) wt (15)	mut (14) wt (55)	w/met (17) w/o met (2)
Diffuse astrocytoma (DA)	mut (46) wt (9)	mut (32) wt (13)	w/met (8) w/o met (7)
Glioblastoma (GBM)	mut (14) wt (227)	mut (56) wt (166)	w/met (23) w/o met (25)
Oligodendroglioma (O)	mut (81) wt (16)	mut (21) wt (78)	w/met (35) w/o met (3)
Others	mut (99) wt (22)	mut (60) wt (57)	w/met (28) w/o met (5)
Total	mut (364) wt (337)	mut (239) wt (419)	w/met (133) w/o met (53)

Only when all four methods detected (or do not detect) mutation(s) in a gene, did we regard the case as unambiguous and keep it. All other ambiguous cases were discarded. For the *MGMT* status, we adopted the labels as having been used in[30] based on the Infinium methylation assay.[36] The statistics of the three molecular biomarkers of the TCGA cases are summarized in Table 2.

For all three tasks, we randomly make ten independent splits with a train/test ratio of 7:3 for 10-fold cross-validation. The performance statistics are reported as the average of the ten trained models on the ten splits. We also make one more independent split, on which we will perform downstream analysis and case study of the model. The patient level and slide level statistics of the splits are provided in Table S1.

Attribution analysis of ViT-WSI

After a ViT-WSI model is trained, we systematically interpreted the trained model with the gradient-based attribution analysis algorithms (Figure 1). We aim to investigate whether ViT-WSI is able to discover histological meaningful features using such procedures. Attribution analysis aims to quantitatively evaluate the

contribution of a particular input to an output of a deep learning model, and it serves as a cornerstone for interpretable machine learning.[37] A large family of attribution analysis algorithms achieves this goal by comparing what the network outputs from a given input to what it outputs from a 'baseline' input (which is usually chosen to be an all-zero input or a random input) and assigning attribution scores to the elements of the given input according to their contribution to the difference.[38] These attribution algorithms have been widely used to interpret Transformer-based models.[39] We applied one of the above-mentioned attribution analysis methods, Integrated Gradients (IG),[40] to the ViT-WSI aggregator. Suppose that the input to the ViT-WSI aggregator (represented as function F) is $\mathbf{X}^{(\mathrm{inp})} = [X_1^{(inp)}, X_2^{(inp)}, \ldots, X_n^{(inp)}]$ (where n is the number of megapixel patches) and the output is a scalar $F(\mathbf{X}^{(\mathrm{inp})})$. IG computes the attribution of the kth input element as follows:

$$\mathrm{IG}_k\left(\mathbf{X}^{(\mathrm{inp})}\right) := \left(\mathbf{X}_k^{(\mathrm{inp})} - \mathbf{X}_k^{(\mathrm{baseline})}\right) \int_{\alpha=0}^{1} F_k'\left(\mathbf{X}^{(\mathrm{baseline})} + \alpha\left(\mathbf{X}^{(\mathrm{inp})} - \mathbf{X}^{(\mathrm{baseline})}\right)\right) d\alpha$$

where F_k' is the partial derivative of the network F w.r.t. the kth input element, $\mathbf{X}^{(\mathrm{baseline})}$ is the above-mentioned baseline input and α is the interpolating factor which varies from 0 to 1. This guarantees that the sum of the attribution of each input element equals the total output difference, which can be expressed as:

$$F\left(\mathbf{X}^{(\mathrm{inp})}\right) - F\left(\mathbf{X}^{(\mathrm{baseline})}\right) = \sum_k \mathrm{IG}_k\left(\mathbf{X}^{(\mathrm{inp})}\right)$$

We computed the contribution of the input patches to various network output quantities, including the classification score of a particular class (the 'class attribution'), specific dimensions of the WSRV (the 'WSRV attribution'), as well as on the projections of WSRV on its specific principal components PC (the 'PC attribution'). In this way, we hope to discover subtype-specific histological features, as well as histological patterns that are managed by specific dimensions of the WSRV, or certain linear combinations of them. The result of the attribution analysis is plotted in a heatmap that highlights regions with positive/negative contributions. The histological features in those regions are then reviewed by two pathologists with ten years of experience (YC and XC). Details of the formulation of attribution analysis are discussed in STAR Methods.

Performance evaluation of the ViT-WSI model

Table 3 summarizes the performance of ViT-WSI on the 8-class top-level type classification task, the 11-class subtyping task, as well as the TCGA glioma subtyping task. We compared the performance of ViT-WSI [ViT (ViT-L-16) + ViT-WSI aggregator] with various other weakly supervised methods including Max Pooling[14] and CLAM-MB.[15] We used various pretrained patch embedders for performance evaluation, including ResNet50, Inception v3 used by Coudray et al.,[10] and the GNN network used by Jaume et al.[41] As previous methods were developed without the presence of pretrained Vision Transformers, to make the comparison fair, we also evaluated the performance of other methods using both ViT-L-16 as the pretrained patch embedder. We also evaluated the patch embedder performance of the ViT-S-16 and ViT-L-16 models when they are self-supervised by DINO[42] instead of pretrained on the ImageNet as used in Chen and Krishnan.[43] We further compared the performance of the above weakly supervised methods with non-weakly supervised methods by training a patch-level version of the patch embedders directly for classification using the slide-level labels as ground truth.

In terms of the area under curve of the one versus rest receiver operating characteristics macro-averaged across each class (Macro AUC) and the Matthews' correlation coefficient (MCC), ViT-WSI with patch graph information (ViT-WSI + Graph) achieves the highest performance (Macro AUC 0.9408, 0.8867 and 0.9313, MCC 0.8757, 0.5628 and 0.8344, under 10-fold cross-validation) on all three tasks compared to other aggregators that use ViT-L-16 as the pretrained patch embedder. Compared to the weakly-supervised learning methods (MIL,[44] CLAM_MB,[15] ViT-WSI, ViT-WSI + Graph), using the slide-level label to train patch-level classifiers achieves the worst performance on the three tasks. There is a larger performance gap between this non-weakly-supervised method and other weakly supervised methods in the more difficult 11-class subtyping task than the 8-class top-level type classification task and the TCGA glioma subtyping task, showing the greater advantage of developing weakly-supervised methods on more fine-grained tumor classification tasks. Adding patch-level graph information to the ViT-WSI model (ViT-WSI + Graph) improves the performance compared to the counterpart that is without (ViT-WSI). Using ViT-WSI-based aggregation performs better than the other methods regardless of the pretrained patch embedder that is used. Consistent with the previous observation,[43] only on the relatively easier tasks, i.e., the 8-class top-level type classification task and TCGA-glioma subtyping task, there are further improvements by using the vision transformers ViT-S-16 or ViT-L-16 self-supervised by DINO as the patch embedders, although

Table 3. Performance of ViT-WSI and comparing methods on three brain tumor type and subtype classification tasks

Dataset	Method		Reference	Macro AUC (mean ± std)	Matthew's correlation coefficient (MCC) (mean ± std)
	Pretrained Patch Embedder	Aggregator			
8-class top-level type classification	ResNet50 (non-weakly-supervised)	N/A		0.867 ± 0.036	0.758 ± 0.032
	ResNet50	CLAM-MB	Lu et al.[15]	0.910 ± 0.020	0.835 ± 0.016
	ResNet50	ViT-WSI aggregator w/graph		0.934* ±0.018	0.864* ± 0.009
	ViT (ViT-L-16, non-weakly-supervised)	N/A		0.875 ± 0.054	0.804 ± 0.068
	ViT (ViT-L-16)	Max Pooling		0.899 ± 0.018	0.826 ± 0.026
	ViT (ViT-L-16)	CLAM-MB		0.930 ± 0.020	0.860 ± 0.026
	ViT (ViT-L-16)	ViT-WSI aggregator w/graph		0.941* ±0.022	0.876* ± 0.024
	Inception v3 (non-weakly-supervised)	N/A	Coudray et al.[10]	0.860* ±0.023	0.745* ± 0.024
	GNN (non-weakly-supervised)	N/A	Jaume et al.[41]	0.832 ± 0.027	0.723 ± 0.025
	ViT (ViT-S-16, DINO)	ViT-WSI aggregator w/graph	Chen and Krishnan[43]	0.940 ± 0.004	0.870 ± 0.012
	ViT (ViT-L-16, DINO)	ViT-WSI aggregator w/graph		0.942* ±0.027	0.872 ± 0.031
	Human Performance: Macro FPR = 0.0901 ± 0.078, Macro TPR = 0.9557 ± 0.042				
11-class subtyping	ResNet50 (non-weakly-supervised)	N/A		0.745 ± 0.022	0.414 ± 0.029
	ResNet50	CLAM-MB	Lu et al.[15]	0.845 ± 0.021	0.536 ± 0.027
	ResNet50	ViT-WSI aggregator w/graph		0.873* ±0.017	0.556* ± 0.031
	ViT (ViT-L-16, non-weakly-supervised)	N/A		0.753 ± 0.027	0.425 ± 0.045
	ViT (ViT-L-16)	Max Pooling		0.837 ± 0.019	0.480 ± 0.031
	ViT (ViT-L-16)	CLAM-MB		0.860 ± 0.022	0.551 ± 0.041
	ViT (ViT-L-16)	ViT-WSI aggregator w/graph		0.887* ±0.024	0.563* ± 0.030
	Inception v3 (non-weakly-supervised)	N/A	Coudray et al.[10]	0.695* ±0.020	0.373 ± 0.029
	GNN (non-weakly-supervised)	N/A	Jaume et al.[41]	0.672 ± 0.019	0.399 ± 0.028
	ViT (ViT-S-16, DINO)	ViT-WSI aggregator w/graph	Chen and Krishnan[43]	0.866 ± 0.017	0.558 ± 0.028
	ViT (ViT-L-16, DINO)	ViT-WSI aggregator w/graph		0.880* ±0.018	0.560 ± 0.026

(*Continued on next page*)

Table 3. *Continued*

Dataset	Method		Reference	Macro AUC (mean ± std)	Matthew's correlation coefficient (MCC) (mean ± std)
	Human Performance: Macro FPR = 0.0509 ± 0.042, Macro TPR = 0.9309 ± 0.074				
TCGA-glioma subtyping	ViT (ViT-L-16, non-weakly-supervised)	N/A		0.845 ± 0.021	0.755 ± 0.033
	ViT (ViT-L-16)	Max Pooling		0.875 ± 0.021	0.824 ± 0.031
	ViT (ViT-L-16)	CLAM-MB		0.916 ± 0.018	0.826 ± 0.030
	ViT (ViT-L-16)	ViT-WSI aggregator w/graph		0.931* ±0.019	0.834* ± 0.030
	Inception v3 (non-weakly-supervised)	N/A	Coudray et al.[10]	0.804* ±0.020	0.723* ± 0.027
	GNN (non-weakly-supervised)	N/A	Jaume et al.[41]	0.762 ± 0.018	0.693 ± 0.030
	ViT (ViT-S-16, DINO)	ViT-WSI aggregator w/graph	Chen and Krishnan[43]	0.932 ± 0.017	**0.843* ± 0.031**
	ViT (ViT-L-16, DINO)	ViT-WSI aggregator w/graph		0.930 ± 0.020	0.830 ± 0.029
	Human Performance: Macro FPR = 0.066 ± 0.065, Macro TPR = 0.936 ± 0.045				

The performance of the full-fledged ViT-WSI model is 'ViT (ViT-L-16)' as the pretrained patch embedder and 'ViT-WSI aggregator w/graph' as the aggregator. We used various pretrained patch embedders for performance evaluation, including ResNet50, ViT (ViT-L-16), Inception v3 used by Coudray et al.,[10] the GNN network used by Jaume et al.,[41] the ViT-S-16 and ViT-L-16 self-supervised by DINO[42] used in Chen and Krishnan.[43] We also compared the performance of various aggregators for weakly supervised learning on histopathology images, including Max Pooling, CLAM-MB,[15] the ViT-WSI aggregator with or without patch graph information. We also compared the weakly supervised algorithms' performance against the non-weakly-supervised patch-level classifier. Bold face indicates the highest performance. Underline indicates the highest performance within each group. * indicates statistically significant ($p<0.05$) performance improvement over the second highest-performing method in the group (Wilcoxon signed-rank test across the 10-folds). Full results are available in Tables S2–S4.

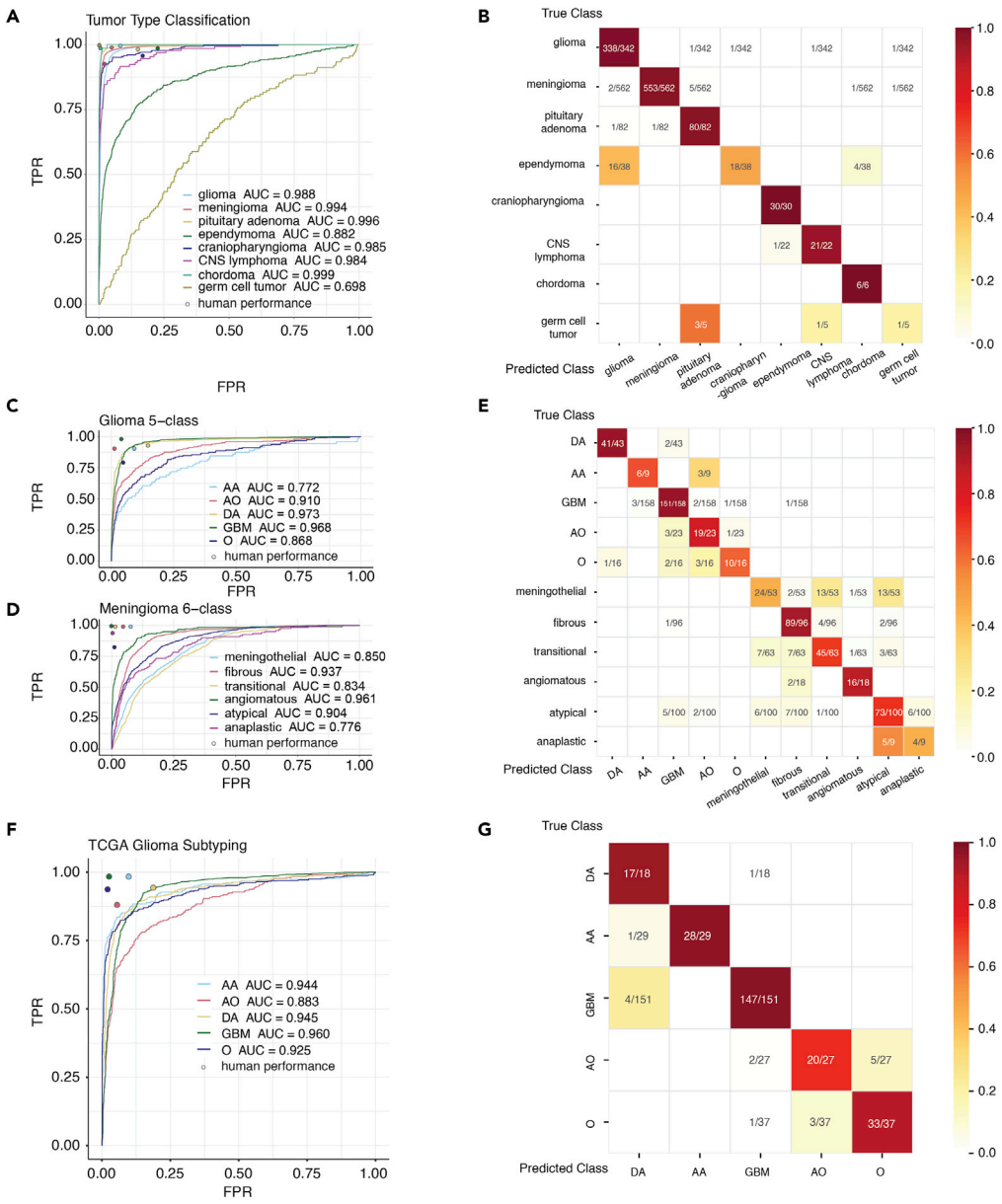


Figure 2. Per-class performance of ViT-WSI

Per-class performance of ViT-WSI on the 8-class top-level type classification task. (A and B), the 11-class subtyping task (C–E), and the TCGA glioma subtyping task (F–G): (1) the per-class one vs. rest receiver operating characteristics and human performance on sampled slides (A, C, D, and F), and (2) confusion matrix on the left-out test set (B, E, and G). For the confusion matrix, the diagonal represents correct predictions, each row represents a ground truth class, and each column represents a predicted class.

not statistically significant ($p = 0.15$ for the 8-class top-level type classification task and $p = 0.32$ for the TCGA-glioma subtyping task, by Wilcoxon signed-rank test). Because of this and to be consistent with previous weakly supervised learning methods, the rest of the analyses are still performed using the ViT-WSI model with ViT-L-16 pretrained on ImageNet as the patch embedder.

We next analyzed the per-class performance of the ViT-WSI + Graph model on the three tasks (Figures 2A–2G). As an additional reference, we sampled 100 slides from each of the three tasks and asked one pathologist (C.W.) to perform a brief and independent pass through them. We evaluated the performance (Macro-average FPR and TPR) of the pathologist by comparing the pathologist's classification to the original diagnosis in the EHRs which

was based on the consensus of a group of pathologists and used this as the 'human performance' of the task (Table 3, Figures 2C,2D, and 2F). Because of class imbalance of the dataset, the model generally has a lower performance on classes that are under-represented (e.g., ependymoma, germ cell tumor). For the 8-class top-level type classification task, there is a very low misclassification rate among the three major primary brain tumor types (glioma, meningioma, and pituitary adenoma) and they are on par with human-level performance (Figures 2A and 2B). However, there is a larger misclassification rate between glioma and ependymoma (Figure 2B). This is because ependymomas, arising from ependymal cells, are also a type of glial cells, and ependymomas can often be considered gliomas in a broad sense.[45] A higher misclassification rate indicates their similarity in histopathology. For the 11-class subtyping task, misclassification is rare between the glioma subtypes and the meningioma subtypes, with one exception (several atypical meningioma slides misclassified as several glioma subtypes) (Figure 2E). This is probably because both tumor subtypes have a higher level of malignancy. In both the 11-class subtyping task (Figure 2E) and the TCGA glioma subtyping task (Figure 2G), misclassification is in general within glioma or within meningioma subtypes, especially between oligodendroglioma (O) and anaplastic oligodendroglioma (AO), as well as atypical and anaplastic meningiomas, which shows the difficulty of classifying tumors of similar origin and cell type but with only slightly different levels of malignancy. There is in general more gap between the model performance and human performance on the more difficult 11-class subtyping task and TCGA glioma subtyping task than on the 8-class top-level type classification task.

The whole representation vectors (WSRV) extracted from the penultimate layer of the ViT-WSI model are visualized in 2D using t-distributed stochastic neighbor embedding (t-SNE)[46] (Figures S3A and S3B, Supplementary Notes Section S1). For the 8-class top-level type classification task (Figure S3A), one can observe a Clear clustering of the glioma, meningioma, pituitary adenoma, and craniopharyngioma slides in the t-SNE space. Ependymoma slides are observed to be closer to the glioma slides which agrees well with their higher misclassification rate and their histopathological similarity. For the 11-class subtyping task, there is a clear separation of all glioma subtypes and meningioma subtypes (Figure S3B). However, a large overlap is observed for meningioma subtype pairs with higher histological similarities like fibrous and transitional, as well as atypical and anaplastic.

ViT-WSI distinguishes itself from other methods by incorporating the self-attention mechanism for whole-slide aggregation. The self-attention mechanism models the relationships between different patches in a WSI. It discovers semantically similar regions (Figure S4, Supplementary Notes Section S2), and has a greater diversity and coverage than other attention-based methods (Figure S5), which results in its greater prediction confidence (Figure S6).

ViT-WSI automatically discovers diagnostic histopathological features

We systematically interpreted the ViT-WSI model trained on the '11-class subtyping task' using gradient-based attribution algorithms. As an example, we first performed attribution analysis of two meningioma slides reported by pathologists as having histological features of multiple subtypes. The first is a mixed type of fibrous and transitional meningioma (Figure 3A). Attribution to the 'fibrous' class (based on the 'class attribution' method, STAR Methods) highlights the regions in the upper part of the slide with cells having typical narrow, rod-shaped histopathological features.[47] Attribution to the 'transitional' class (based on the 'class attribution' method) highlights the typical whorl structures of transitional meningioma.[47] In some parts of the slide where the 'transitional' whorl features are scattered, the attribution heatmap is even more capable of highlighting them individually (Figure 3A, zoom-in view). The second is a meningothelial slide showing features of angiomatous meningioma (Figure 3B). Individual attribution maps for the 'meningothelial' and 'angiomatous' classes (based on the 'class attribution' method) again 'pick out' histological structures of their corresponding subtypes.

Motivated by the effectiveness of the above analysis, we further extended the attribution procedure to various other network outputs. Instead of performing attribution analysis on the network's final layer (the classification output probabilities), we moved our focus to the penultimate layer, which outputs the WSRV that has been visualized in Figure S3. From here, we wished to find what histopathological features are being attended to by the different dimensions of the WSRV. To achieve this goal, we could exhaustively perform the analysis on each dimension of the representation vector one by one and observe their correlation with some histopathological-meaningful patterns. However, this approach is inefficient and uninformed because the dimension of the representation vector is large (1,024 dimensions), and some dimensions may have overlapping or repetitive semantics.

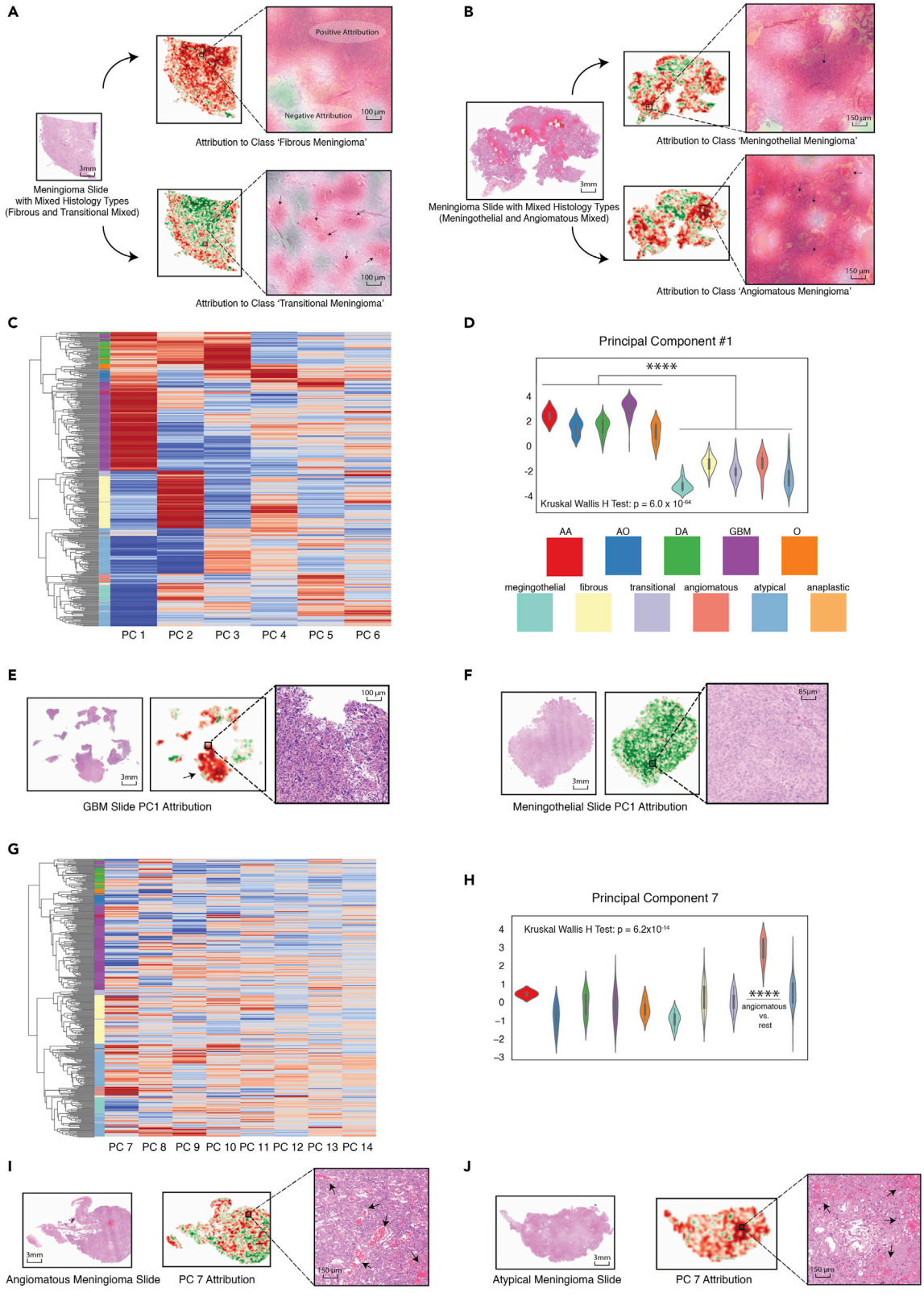

A
Positive Attribution
Negative Attribution
100 µm
3mm
Meningioma Slide
with Mixed Histology Types
(Fibrous and Transitional Mixed)
Attribution to Class 'Fibrous Meningioma'
Attribution to Class 'Transitional Meningioma'
B
3mm
Meningioma Slide with Mixed Histology Types
(Meningothelial and Angiomatous Mixed)
150 µm
Attribution to Class 'Meningothelial Meningioma'
Attribution to Class 'Angiomatous Meningioma'
C
PC 1 PC 2 PC 3 PC 4 PC 5 PC 6
D
Principal Component #1
****
Kruskal Wallis H Test: p = 6.0 x 10-64
AA AO DA GBM O
megingothelial fibrous transitional angiomatous atypical anaplastic
E
100 µm
3mm
GBM Slide PC1 Attribution
F
85µm
3mm
Meningothelial Slide PC1 Attribution
G
PC 7 PC 8 PC 9 PC 10 PC 11 PC 12 PC 13 PC 14
H
Principal Component 7
Kruskal Wallis H Test: p = 6.2x10-14
****
angiomatous
vs.
rest
I
3mm
Angiomatous Meningioma Slide
PC 7 Attribution
150 µm
J
3mm
Atypical Meningioma Slide
PC 7 Attribution
150 µm

Figure 3. Attribution analysis automatically discovers histological features

(A) Class attribution to a slide with mixed histology types of fibrous and transitional meningioma. Attribution to the 'fibrous' class highlights typical rod-shaped fibrous meningioma tumor cells. Attribution to the 'transitional' class highlights the typical whorl structures of transitional meningioma. In some parts of the slide where the transitional whorl features are scattered, the attribution heatmap is even more capable of highlighting them individually (black arrows pointed).

(B) Class attribution to a meningothelial slide showing features of angiomatous meningioma. Attribution to the 'meningothelial' class highlights the syncytial cell structures (black arrow pointed). Attribution to the 'angiomatous' class highlights the abundant blood vessels present in the tissue slice (black arrows pointed).

(C) Clustered heatmap for the first six PCs for test set slides that are confidently and correctly predicted by ViT-WSI in the '11-class subtyping task'. The cluster dendrogram and identity color bar show the subtypes with similar histology clustered together. For example, all glioma subtypes are clustered at the top, whereas all meningioma subtypes are clustered at the bottom. Most oligodendroglioma (O) and anaplastic oligodendroglioma (AO) are under a same subcluster, and so are fibrous and transitional meningiomas. The heatmap shows the activation of the first six components that are mostly subtype-specific.

(D) Violin plot showing the activation of PC1 being very different among subtypes (Kruskal-Wallis H Test: $p = 6.0 \times 10\text{-}64$) mainly among glioma subtypes (Mann-Whitney U Test: $p = 1.5 \times 10\text{-}62$) and especially in glioblastoma. **** for $p<0.0001$.

(E) PC1 attribution to a glioblastoma slide's input patches shows strong positive attribution (red) to most slide locations, with one tissue piece having the highest density of malignant cells showing the most (black arrows pointed).

(F) PC1 projection to a meningothelial slide's input patches shows general negative attribution (green) to most slide locations.

(G) The heatmap shows the activation of the PC7-PC14 is gradually distributed more diversely among subtypes.

(H) Violin plot showing the activation of PC7 is slightly higher in fibrous, transitional, and atypical meningiomas but has a much higher activation in angiomatous subtypes (Mann-Whitney U Test: $p = 1.8 \times 10\text{-}7$ (angiomatous meningioma versus rest)). Its distribution is more diverse than previous PCs (Kruskal Wallis H Test: $p = 6.2 \times 10\text{–}14$, which is less significant than the same test for PC1, PC2, and PC3). **** for $p<0.0001$ by Mann-Whitney U Test.

(I) Attribution analysis of PC7 attributes its activation to blood vessel structures (arrow pointed) in angiomatous meningioma.

(J) Attribution analysis of PC7 attributes its activation also to angiomatous characteristics (arrow pointed) within atypical meningioma slides (as reported in the EHR by pathologists).

We thus used a principal component (PC)-guided attribution procedure to interpret the trained ViT-WSI model (the 'PC attribution' method, STAR Methods). Without performing attribution analysis on each individual dimension, we instead performed on the WSRV's projections to their PCs. The principal components are computed using all the representation vectors of the slides in the test set of the independent split of the '11-class subtyping task' and ordered in their amount of explained variance (Figure S7).

The PC-projected whole slide representation vectors are plotted for each slide in the '11-class subtyping task' test set, separately for the first six PCs (PC1-6, Figure 3C) and the next eight PCs (PC7-14, Figure 3G). The slides are hierarchically clustered using the 'average linkage' method on their first 20 PCs. One can observe a high-level agreement between the clustering result and their true histopathological subtypes, with all the subtypes almost exclusively clustered together and the glioma slides consistently clustered on the top, and the meningioma slides at the bottom. We further dropped the anaplastic meningioma subtype in the subsequent analyses because of its limited number of slides in the test set ($n = 9$) and its largely overlapping clustering semantics as the atypical meningioma subtype.

For the first six components, there is a general trend that projections are highly activated for a specific subtype or a few subtypes (Figure 3C, summarized in the first half of Table 4). As an example, PC1 is highly activated in the glioma subtypes and specifically in the glioblastoma subtype (Figure 3D). On the contrary, it is mostly negative among the meningioma subtypes and specifically in the meningothelial subtype (Figure 3D). Attribution of PC1's activation to a GBM slide's input patches shows strong positive attribution to most slide locations (Figure 3E), with one tissue piece possessing the highest density of malignant cells showing the most (Figure 3E, black arrow pointed). Attribution of PC1's activation to a meningothelial slide's input patches shows general negative attribution to most slide locations (Figure 3F).

Next, PC2, with high activation in fibrous meningioma (Figures 3C and S8D), also has high attribution to most regions on fibrous meningioma slides (Figures S8B and S8C). PC2 has the highest-loaded dimension #261 (Figures S8A and S8E) attributed to tumor cells with rod-shaped nuclei (Figures S8G and S8I). PC3 is also highly activated in glioma populations, but unlike PC1, PC3 is mainly activated in low-grade gliomas like diffusive astrocytoma and oligodendroglioma (Figures 3C and S9B). Projections on PC3 can be mostly attributed to typical astrocytoma lesions (Figure S9C) and oligodendroglioma's 'fried-egg-shaped' glial cells (Figure S9D). Some of its highest-loaded dimensions, e.g., #838, are responsible for astrocytoma components (Figures S9E and S9G), whereas others, e.g., #335, are responsible for oligodendroglioma components (Figures S9F and S9H). Attribution of these dimensions on the slide with the 'wrong' subtype is more negative (attribution of #838 on the oligodendroglioma slide, Figure S9I) or less consistent with the full PC3 attribution (attribution of #335 on the astrocytoma slide, comparing Figures S9H and S9C).

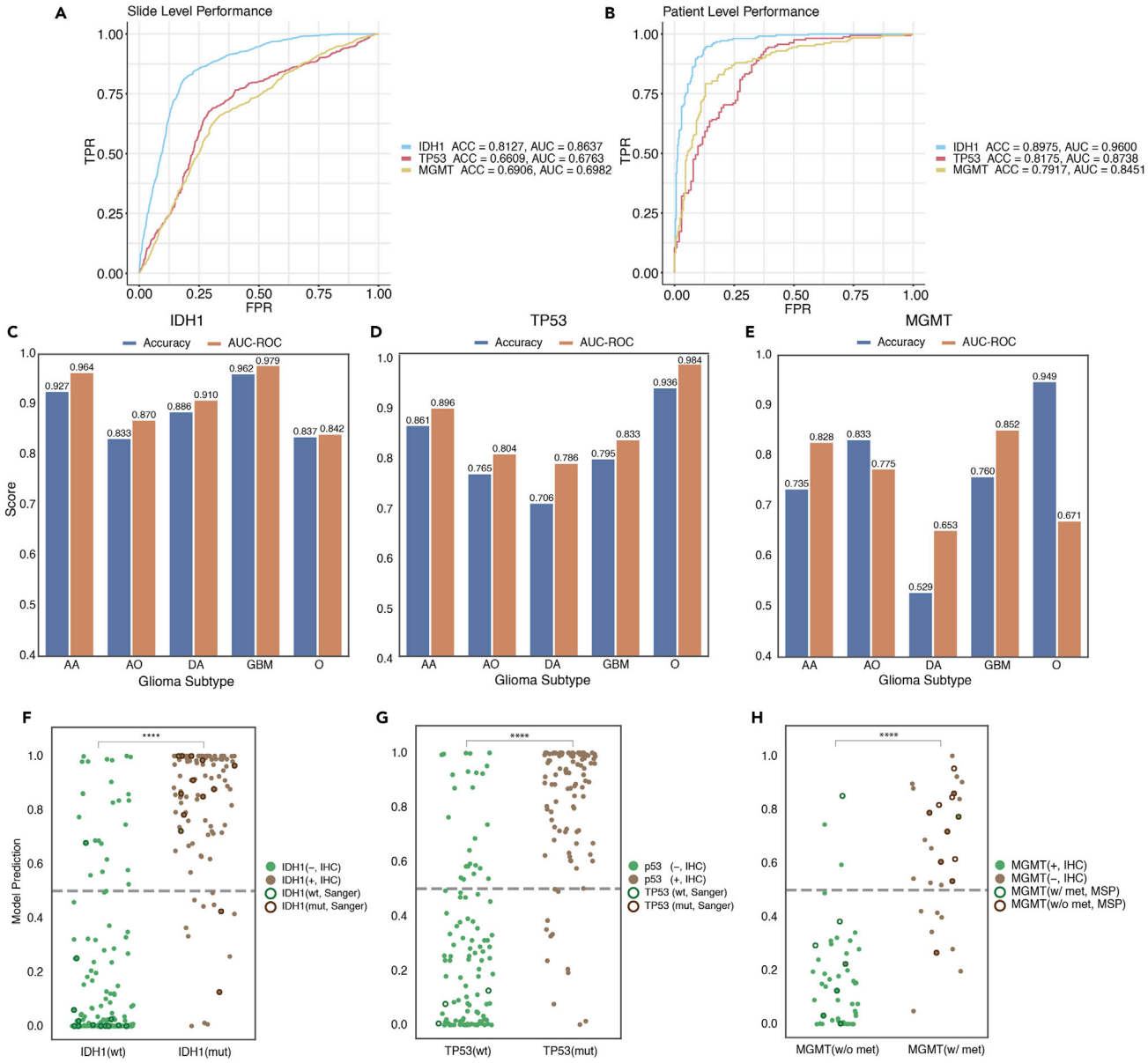


Figure 4. Diagnostic molecular biomarkers prediction of ViT-WSI in glioma

(A) ROC curves at the slide-level for the prediction of IDH1 somatic mutation, TP53 somatic mutation, and MGMT promoter methylation.

(B) ROC curves at the patient level.

(C–E) Performance evaluation of ViT-WSI in terms of the accuracy and AUC-ROC per each glioma subtype for IDH1 (C), TP53 (D), and MGMT (E).

(F–H) Comparison of molecular biomarker status predicted by ViT-WSI, status reported by IHC, and status reported by molecular biology assays of in-house patients. Each dot represents a patient whose fill color (green for negative and brown for positive) is determined by the biomarker status reported by IHC, and edge color (green for wild type or w/o methylation and brown for mutant or w/methylation) is determined by the biomarker status reported by molecular biology assays (Sanger sequencing for IDH1 and TP53 and MSP for MGMT). An edgeless point means molecular biology assay is unavailable, while a point without fill color means the IHC result is unavailable. The patients are grouped (along the xaxis) by the final diagnosis for a biomarker, which is determined from the IHC report or the molecular biology assay report, whichever is available. Whenever there is a disagreement between the two methods, molecular biology assay takes precedence. The yaxis indicates ViT-WSI's predicted probability of each patient for the positive class (somatic mutation in IDH1 and TP53 or methylation in MGMT), for which 0.5 is used as a prediction cutoff threshold (indicated by the gray dashed line). ViT-WSI's prediction agrees with 23 of the 26 cases with Sanger sequencing of IDH1 (F), all 3 cases with Sanger sequencing of TP53 (G), and 16 of the 18 cases with MSP of MGMT (H). ViT-WSI's prediction agrees with the 4 cases (2 cases in IDH1 and 2 cases in MGMT) whose molecular biological diagnosis disagrees with IHC assays (F, H). **** for $p<0.0001$ by Mann-Whitney U Test.

Table 4. List of histological features discovered by the gradient-based attribution analysis of ViT-WSI

	Histological features	Active subtypes	Relevant figures
Histological features in specific subtypes			
PC 1	Glioma specific features	Multiple glioma subtypes, especially in glioblastoma	Figures 3B–3D
PC 2	Fibrous meningioma features	Fibrous meningioma	Figure S8
PC 3	Low-grade glioma features	Diffuse astrocytoma, oligodendroglioma	Figure S9
Histological features shared across subtypes			
PC 7	Blood vessels	Multiple meningioma subtypes, especially in angiomatous meningioma subtype	Figures 3E–3H
PC 10	Hemorrhage	Various glioma and meningioma subtypes	Figure S10
PC 11	Calcification regions	Various meningioma subtypes	Figure S11
PC 13	Necrosis regions	Glioblastoma and atypical meningioma	Figure S12

We list the primary histological features and the particular subtype(s) that result in high activation of a specific PC of the WSRV.

Projections on components on the next eight PCs (PC7-14, Figure 3G, summarized in the second half of Table 4) are generally lower in magnitude but are distributed more diversely among subtypes and tend to be activated by histopathological features that are shared across the subtypes. PC7, for example, has a slightly higher activation in fibrous, transitional, and atypical meningiomas but has a much higher activation in the angiomatous meningioma subtype (Figure 3H). Attribution analysis on this component shows that it attends not only to blood vessel structures in angiomatous meningioma (Figure 3I) but also to atypical meningioma slides with angiomatous characteristics (Figure 3J). PC10, also being less specific to a particular subtype (Figure S10A), attends to blood cells in hemorrhage that are less structured than those in the blood vessels among several different glioma and meningioma subtypes (Figures S10B–S10E). PC11, still non-specific to a subtype (Figure S11A), has its attribution map that almost exclusively covers the calcification regions, representing a common histological feature in various meningioma subtypes (Figures S11B–S11D).[47,48] Similarly, PC13 has been found to attend to various necrosis regions in glioblastoma, diffusive astrocytoma, and atypical meningioma (Figures S12B–S12D). The activation of PC13, being slightly higher in glioblastoma and atypical meningioma, suggests that tumors in these two subtypes more frequently display necrosis (Figure S12A), indicating their higher malignancy.

ViT-WSI predicts diagnostic molecular markers directly from H&E histopathology slides

We next investigated whether ViT-WSI, trained in a weakly-supervised fashion, can predict common diagnostic molecular markers from glioma H&E slides. We further assessed if ViT-WSI can be either used clinically as a secondary check or as a low-cost surrogate for real molecular biology experiments.

We systematically explored the potential of ViT-WSI for inferring the aforementioned molecular biomarkers in a weakly supervised setting. A ViT-WSI model was fine-tuned from the trained 11-class subtyping model for each of the three binary classification tasks (somatic mutation versus wild type for *IDH1* and *TP53*, with methylation versus without methylation for *MGMT*) under a 10-fold cross-validation scheme using the combined dataset from the in-house and TCGA glioma slides that are identified with the three biomarkers. Overall, at the slide level, on the combined cohort of the in-house data and TCGA, ViT-WSI achieves accuracy scores of 0.8127 (ROC-AUC = 0.8637), 0.6609 (ROC-AUC = 0.6763), and 0.6906 (ROC-AUC = 0.6981), respectively for the three tasks (Figure 4A). However, once the result is aggregated by averaging the prediction across multiple slides of a patient, ViT-WSI achieves patient-level accuracy scores of 0.8975 (ROC-AUC = 0.9600), 0.8175 (ROC-AUC = 0.8738), and 0.7916 (ROC-AUC = 0.8451), respectively for the three tasks (Figure 4B). As the most frequent subtype, the performance on glioblastoma among the three tasks is generally closer to the overall performance than the other subtypes (Figures 4C–4E). In contrast, diffuse astrocytoma, being the most under-represented subtype for *TP53* and *MGMT* prediction, has a much lower performance. Some subtypes, such as oligodendroglioma in *MGMT* prediction, are observed with a much higher accuracy score than their ROC-AUC score because of the severe class imbalance within the particular subtype.

In the in-house dataset, there are 29 cases whose somatic mutation is validated by Sanger sequencing in addition to IHC staining (26 cases for *IDH1*, Figure 4F, and 3 cases for *TP53*, Figure 4G). There are also 18

cases whose *MGMT* methylation status is also validated by methylation-specific PCR (MSP)[31](Figure 4H). Despite the higher costs, these molecular biological diagnostic methods have shown higher sensitivity and specificity than the IHC-based diagnosis.[49,50] ViT-WSI's predictions agree with 23 of the 26 cases with Sanger sequencing of *IDH1* (Figure 4F), all 3 cases with Sanger sequencing of *TP53* (Figure 4G), and 16 of the 18 cases with MSP of *MGMT* (Figure 4H). Notably, ViT-WSI's predictions agree with the 4 cases (2 cases in *IDH1* and 2 cases in *MGMT*) whose molecular biological diagnosis disagrees with IHC assays (Figures 4F and 4H).

For *IDH1* and *TP53*, attribution analysis of ViT-WSI's output probability for the 'somatic mutation' class prioritizes out regions having typical malignancy in two anaplastic astrocytoma examples (Figures S13A–S13H). For *MGMT*, attribution analysis of ViT-WSI's output probability for the 'without methylation' class (which implies *MGMT* expression) shows different attribution levels on different tissue patches of the same slide (Figures S13I–S13L). Alongside the attribution heatmap, we illustrated the IHC-stained slide of an adjacent tissue section that is sliced out from the same specimen as the H&E slide used by ViT-WSI (Figures S13C, S13G, S13K, S13D, S13H, and S13L). Because of the adjacency of the H&E and the IHC tissue sections, the two slides show similarly shaped tissue boundaries and tissue components. A strong agreement can be observed between the attribution heatmap and the IHC-stained positive regions, indicating that the regions that strongly contributed to ViT-WSI's decision-making indeed presented the underlying molecular mechanism.

DISCUSSION

We have developed a Vision Transformer-based deep learning architecture, termed ViT-WSI, for weakly supervised histopathology WSI analysis in brain tumors. We showed how ViT-WSI achieved superior performance in brain tumor type and subtype classification compared to the currently available methods. We also showed how a fully differentiable ViT-WSI aggregator is amenable to attribution-based interpretation and how it automatically discovers diagnostic histopathological features. We finally showed how ViT-WSI, in a completely weakly-supervised fashion, predicts diagnostic molecular markers with decent accuracy directly from H&E images.

As a two-stage framework, the pretrained patch embedder of ViT-WSI leverages the state-of-the-art Vision Transformer for feature extraction. The ViT-WSI aggregator, also being a Vision Transformer itself, allows better utilization of the extracted features and easier optimization because it simply contains additional Transformer layers on top of the patch embedder, and optimization of deep transformer models has proven to be successful.[51,52] In contrast to many previous WSI weakly supervised learning algorithms, within ViT-WSI we aggregate patches in a context-aware way using self-attention and a nearest-neighbor graph that takes into account patch feature similarity and closeness of their physical locations. In this way, the model is capable of discovering semantically similar regions in the WSIs at its information processing stage (Figure S4). The model's attention region is generally more diverse and greater in size, and its prediction is more confident than aggregation methods that deal with patches independently (Figures S5 and S6).

Misclassification of ViT-WSI between oligodendroglioma (O) and anaplastic oligodendroglioma (AO), as well as atypical and anaplastic meningiomas, suggested a general difficulty in classifying tumors of similar origin and cell type but with only different levels of malignancy. The latest 2021 WHO Classification of Tumors of the CNS dropped the usage of 'anaplastic' in gliomas completely and considered them only as a grading difference from their non-anaplastic counterparts, acknowledging the subtlety of histological difference between them.[53] In this study, we still adhered to the 2016 WHO classification system, as the 2021 WHO classification system is still too new to be adopted by most healthcare institutions worldwide.

Having a bulkier input, interpreting weakly supervised deep learning models on gigapixel WSIs is even more challenging than interpreting common computer vision models on megapixel images. We showed how ViT-WSI is amenable to attribution analysis that interprets ViT-WSI with attribution scores assigned to input patches. We performed attribution analysis on multiple types of network outputs. Attribution analysis on class probability output highlights regions on WSI belonging to a particular subtype (Figures 3A and 3B) or regions presenting a particular molecular mechanism (Figures S13C, S13G, and S13K). Attribution on principal components highlights different histopathological features enriched within a particular subtype (Figures 3, S8, and S9) or shared between different subtypes (Figures S10–S12).

Earlier work has shown the penultimate layer output of a vision network (VGG19 [54]), trained on brain tumor histopathology image at the patch level, highly preserves image semantics.[55] Hierarchical clustering on the penultimate layer output, which is a high-dimensional representation vector for one histopathology image patch, automatically re-discovered a large proportion of brain tumor ontological relationships. The penultimate layer of ViT-WSI also outputs a high-dimensional representation vector, i.e., the WSRV. However, it represents a gigapixel WSI instead of just one image patch. Hierarchical clustering of the ViT-WSI's WSRV also re-discovered similar ontological relations among the meningioma and glioma subtypes that are consistent with prior knowledge (Figure 3C). When interpreting the high-dimensional representation vector, Faust et al. selected the dimensions of the feature vector with different activation distributions among subtypes and exhaustively visually examined them to find out possible correlations with known histopathological features.[55] In ViT-WSI, we instead employed a more informative PC-guided approach for feature interpretation. In this way, each PC summarizes a particular histopathological feature, whose loadings display the contribution of each dimension to the PC, and as such, important dimensions can be readily identified according to the magnitude of their loadings.

Without spatial profiling,[56] training H&E WSI-based predictive models of molecular biomarkers is inherently a weakly supervised learning problem. Previous works with machine learning models trained at the patch level have shown predictive power of H&E images over common somatic mutations such as lung tumor,[10] liver tumor,[57] and microsatellite instability in gastrointestinal tumor.[58] In this work, our proposed ViT-WSI method, weakly supervised at the slide level, also showed promising potential to 'see the unseen' from the histopathological images and predicted the three highly informative molecular biomarkers in glioma, and especially somatic mutation in *IDH1*, with decent accuracy.

Limitations of the study

A limitation of ViT-WSI concerns its large memory consumption, typical for Transformer-based models. Although this problem is generally manageable during model training and evaluation, it may arise during attribution analysis that quickly depletes GPU memory. This is especially the case when being used in combination with gradient-based attribution algorithms such as Integrated Gradients,[40] as it requires multiple iterations through the network. We, therefore, had to perform all the attribution analysis by running the attribution algorithm purely with CPU on large-memory (~512 GB) computational nodes. This also limits the resolution of attribution analysis. Currently, only the ViT-WSI aggregator participated in the attribution analysis, and only one attribution score is assigned to each WSI patch. Applying attribution analysis to both the pretrained patch embedder and aggregator as a whole and attributing to each pixel of the WSI is currently not feasible with ordinary computational infrastructure. Future work that involves experimentation with more memory-efficient Transformer models and attribution methods should be encouraged and appreciated.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
- METHOD DETAILS
  - Data preparation and preprocessing
  - Vision transformer (ViT)–based weakly-supervised whole-slide image analysis model (ViT-WSI)
  - Training and evaluation of ViT-WSI
  - Whole-slide representation vector (WSRV) produced by ViT-WSI
  - Self-attention visualization of ViT-WSI
  - Attribution analysis of ViT-WSI
- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2022.105872.

ACKNOWLEDGMENTS

Figure 1 was created by Heno Hwang, scientific illustrator at King Abdullah University of Science and Technology (KAUST). The results shown here are in whole or part based on data generated by the TCGA Research Network: https://www.cancer.gov/tcga. This work was supported by Office of Research Administration (ORA) at KAUST under award numbers FCC/1/1976-44-01, FCC/1/1976-45-01, URF/1/4098-01-01, URF/1/4352-01-01, REI/1/5202-01-01, REI/1/4940-01-01, RGC/3/4816-01-01, and REI/1/0018-01-01.

AUTHOR CONTRIBUTIONS

Z.L., S.Z., and X.G. conceived the project. Z.L. and X.G. developed the ViT-WSI model and did the computational experiments. Y.C. and X.C. collected and scanned the histopathology slides, retrieved and sorted the EHR of the patients, and reviewed the discovered histopathological features in selected cases. J.S., T.Y., H.Y., J.L., E.L., L.W., J.L., H.H., C.Z., Q.Y., J.Y., P.Y., Q.J., and W.L. assisted in the collection and scanning of the histopathology slides. J.Q., J.S., Y.C., and S.Z. provided additional pathological insights into the experimental results. All authors read and approved the final manuscript.

DECLARATION OF INTERESTS

The authors have declared no competing interests.

INCLUSION AND DIVERSITY

We support inclusive, diverse, and equitable conduct of research.

Received: September 2, 2022
Revised: December 3, 2022
Accepted: December 21, 2022
Published: January 20, 2023

REFERENCES

1. Van den Bent, M.J. (2010). Interobserver variation of the histopathological diagnosis in clinical trials on glioma: a clinician's perspective. Acta Neuropathol. *120*, 297–304.

2. Willis, J., Smith, C., Ironside, J.W., Erridge, S., Whittle, I.R., and Everington, D. (2005). The accuracy of meningioma grading: a 10-year retrospective audit. Neuropathol. Appl. Neurobiol. *31*, 141–149.

3. Louis, D.N., Perry, A., Reifenberger, G., von Deimling, A., Figarella-Branger, D., Cavenee, W.K., Ohgaki, H., Wiestler, O.D., Kleihues, P., and Ellison, D.W. (2016). The 2016 world health organization classification of tumors of the central nervous system: a summary. Acta Neuropathol. *131*, 803–820. https://doi.org/10.1007/s00401-016-1545-1.

4. Rapôso, C., Vitorino-Araujo, J.L., and Barreto, N. (2021). In Molecular Markers of Gliomas to Predict Treatment and Prognosis: Current State and Future Directions, W.D. Gliomas, ed. (Exon Publications). chapter10. https://doi.org/10.36255/exonpublications.gliomas.2021.

5. Fraggetta, F., Garozzo, S., Zannoni, G.F., Pantanowitz, L., and Rossi, E.D. (2017). Routine digital pathology workflow: the Catania experience. J. Pathol. Inf. *8*, 51.

6. Litjens, G., Bandi, P., Ehteshami Bejnordi, B., Geessink, O., Balkenhol, M., Bult, P., Halilovic, A., Hermsen, M., van de Loo, R., Vogels, R., et al. (2018). 1399 H& E-stained sentinel lymph node sections of breast cancer patients: the CAMELYON dataset. GigaScience *7*, giy065. https://doi.org/10.1093/gigascience/giy065.

7. Couture, H.D., Williams, L.A., Geradts, J., Nyante, S.J., Butler, E.N., Marron, J.S., Perou, C.M., Troester, M.A., and Niethammer, M. (2018). Image analysis with deep learning to predict breast cancer grade, ER status, histologic subtype, and intrinsic subtype. npj Breast Cancer *4*, 30. https://doi.org/10.1038/s41523-018-0079-1.

8. Strom, P., Kartasalo, K., and Olsson, H. (2020). Artificial intelligence for diagnosis and grading of prostate cancer in biopsies: a population-based, diagnostic study. Lancet Oncol. *21*, E70.

9. Courtiol, P., Maussion, C., Moarii, M., Pronier, E., Pilcer, S., Sefta, M., Manceron, P., Toldo, S., Zaslavskiy, M., Le Stang, N., et al. (2019). Deep learning-based classification of mesothelioma improves prediction of patient outcome. Nat. Med. *25*, 1519–1525.

10. Coudray, N., Ocampo, P.S., Sakellaropoulos, T., Narula, N., Snuderl, M., Fenyö, D., Moreira, A.L., Razavian, N., and Tsirigos, A. (2018). Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning. Nat. Med. *24*, 1559–1567. https://doi.org/10.1038/s41591-018-0177-5.

11. Huang, K.K., Huang, J., Wu, J.K.L., Lee, M., Tay, S.T., Kumar, V., Ramnarayanan, K., Padmanabhan, N., Xu, C., Tan, A.L.K., et al. (2021). Long-read transcriptome sequencing reveals abundant promoter diversity in distinct molecular subtypes of gastric cancer. Genome Biol. *22*, 44.

12. Jain, M.S., and Massoud, T.F. (2020). Predicting tumour mutational burden from histopathological images using multiscale deep learning. Nat. Mach. Intell. *2*, 356–362. https://doi.org/10.1038/s42256-020-0190-5.

13. Ehteshami Bejnordi, B., Veta, M., Johannes van Diest, P., van Ginneken, B., Karssemeijer, N., Litjens, G., van der Laak, J.A.W.M., the CAMELYON16 Consortium, Hermsen, M., Manson, Q.F., et al. (2017). Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. JAMA *318*, 2199–2210. https://doi.org/10.1001/jama.2017.14585.

14. Campanella, G., Hanna, M.G., Geneslaw, L., Miraflor, A., Werneck Krauss Silva, V., Busam, K.J., Brogi, E., Reuter, V.E., Klimstra, D.S., and Fuchs, T.J. (2019). Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. Nat. Med. *25*, 1301–1309. https://doi.org/10.1038/s41591-019-0508-1.

15. Lu, M.Y., Williamson, D.F.K., Chen, T.Y., Chen, R.J., Barbieri, M., and Mahmood, F. (2021). Data-efficient and weakly supervised computational pathology on whole-slide images. Nat. Biomed. Eng. *5*, 555–570.

https://doi.org/10.1038/s41551-020-00682-w.

16. Lu, M.Y., Chen, T.Y., Williamson, D.F.K., Zhao, M., Shady, M., Lipkova, J., and Mahmood, F. (2021). AI-based pathology predicts origins for cancers of unknown primary. Nature *594*, 106–110. https://doi.org/10.1038/s41586-021-03512-4.

17. Ostrom, Q.T., Patil, N., Cioffi, G., Waite, K., Kruchko, C., and Barnholtz-Sloan, J.S. (2020). CBTRUS statistical report: primary brain and other central nervous system tumors diagnosed in the United States in 2013–2017. Neuro Oncol. *22*, iv1–iv96. https://doi.org/10.1093/neuonc/noaa200.

18. Cancer Genome Atlas Research Network, Brat, D.J., Verhaak, R.G.W., Aldape, K.D., Yung, W.K.A., Salama, S.R., Cooper, L.A.D., Rheinbay, E., Miller, C.R., Vitucci, M., et al. (2015). Comprehensive, integrative genomic analysis of diffuse lower-grade gliomas. N. Engl. J. Med. *372*, 2481–2498.

19. Ertosun, M.G., and Rubin, D.L. (2015). Automated Grading of Gliomas Using Deep Learning in Digital Pathology Images: A Modular Approach with Ensemble of Convolutional Neural Networks (American Medical Informatics Association), p. 1899.

20. Jin, L., Shi, F., Chun, Q., Chen, H., Ma, Y., Wu, S., Hameed, N.U.F., Mei, C., Lu, J., Zhang, J., et al. (2021). Artificial intelligence neuropathologist for glioma classification using deep learning on hematoxylin and eosin stained slide images and molecular markers. Neuro Oncol. *23*, 44–52. https://doi.org/10.1093/neuonc/noaa163.

21. Liu, M., Zhang, J., Lian, C., and Shen, D. (2020). Weakly supervised deep learning for brain disease prognosis using MRI and incomplete clinical scores. IEEE Trans. Cybern. *50*, 3381–3392.

22. Ji, Z., Shen, Y., Ma, C., and Gao, M. (2019). Scribble-based Hierarchical Weakly Supervised Learning for Brain Tumor Segmentation (Springer), pp. 175–183.

23. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., and Polosukhin, I. (2017). Attention Is All You Need, pp. 5998–6008.

24. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. (2020). An image is worth 16x16 words: transformers for image recognition at scale. Preprint at arXiv. https://doi.org/10.48550/arXiv.2010.11929.

25. Ridnik, T., Baruch, E.B., Noy, A., and Zelnik-Manor, L. (2021). ImageNet-21K pretraining for the masses. Preprint at arXiv. https://doi.org/10.48550/arXiv.2104.10972.

26. Rasheed, B.A., McLendon, R.E., Herndon, J.E., Friedman, H.S., Friedman, A.H., Bigner, D.D., and Bigner, S.H. (1994). Alterations of the TP53 gene in human gliomas. Cancer Res. *54*, 1324–1330.

27. Wick, W., Weller, M., van den Bent, M., Sanson, M., Weiler, M., von Deimling, A., Plass, C., Hegi, M., Platten, M., and Reifenberger, G. (2014). MGMT testing—the challenges for biomarker-based glioma treatment. Nat. Rev. Neurol. *10*, 372–385. https://doi.org/10.1038/nrneurol.2014.100.

28. Jiang, S., Zanazzi, G.J., and Hassanpour, S. (2021). Predicting prognosis and IDH mutation status for patients with lower-grade gliomas using whole slide images. Sci. Rep. *11*, 16849. https://doi.org/10.1038/s41598-021-95948-x.

29. Liu, S., Shah, Z., Sav, A., Russo, C., Berkovsky, S., Qian, Y., Coiera, E., and Di Ieva, A. (2020). Isocitrate dehydrogenase (IDH) status prediction in histopathology images of gliomas using deep learning. Sci. Rep. *10*, 7733. https://doi.org/10.1038/s41598-020-64588-y.

30. Yogananda, C.G.B., Shah, B.R., Nalawade, S.S., Murugesan, G.K., Yu, F.F., Pinho, M.C., Wagner, B.C., Mickey, B., Patel, T.R., Fei, B., et al. (2021). MRI-based deep-learning method for determining glioma MGMT promoter methylation status. AJNR. Am. J. Neuroradiol. *42*, 845–852. https://doi.org/10.3174/ajnr.A7029.

31. Herman, J.G., Graff, J.R., Myöhänen, S., Nelkin, B.D., and Baylin, S.B. (1996). Methylation-specific PCR: a novel PCR assay for methylation status of CpG islands. Proc. Natl. Acad. Sci. USA *93*, 9821–9826.

32. Fan, Y., Xi, L., Hughes, D.S.T., Zhang, J., Zhang, J., Futreal, P.A., Wheeler, D.A., and Wang, W. (2016). MuSE: accounting for tumor heterogeneity using a sample-specific error model improves sensitivity and specificity in mutation calling from sequencing data. Genome Biol. *17*, 178. https://doi.org/10.1186/s13059-016-1029-6.

33. Benjamin, D., Sato, T., Cibulskis, K., Getz, G., Stewart, C., and Lichtenstein, L. (2019). Calling somatic SNVs and indels with mutect2. Preprint at bioRxiv. https://doi.org/10.1101/861054.

34. Larson, D.E., Harris, C.C., Chen, K., Koboldt, D.C., Abbott, T.E., Dooling, D.J., Ley, T.J., Mardis, E.R., Wilson, R.K., and Ding, L. (2012). SomaticSniper: identification of somatic point mutations in whole genome sequencing data. Bioinformatics *28*, 311–317.

35. Koboldt, D.C., Zhang, Q., Larson, D.E., Shen, D., McLellan, M.D., Lin, L., Miller, C.A., Mardis, E.R., Ding, L., and Wilson, R.K. (2012). VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Res. *22*, 568–576.

36. Bibikova, M., Le, J., Barnes, B., Saedinia-Melnyk, S., Zhou, L., Shen, R., and Gunderson, K.L. (2009). Genome-wide DNA methylation profiling using Infinium® assay. Epigenomics *1*, 177–200.

37. Du, M., Liu, N., and Hu, X. (2019). Techniques for interpretable machine learning. Commun. ACM *63*, 68–77.

38. Ancona, M., Ceolini, E., Öztireli, C., and Gross, M. (2017). Towards better understanding of gradient-based attribution methods for deep neural networks. Preprint at arXiv. https://doi.org/10.48550/arXiv.1711.06104.

39. Chefer, H., Gur, S., and Wolf, L. (2021). Transformer Interpretability beyond Attention Visualization, pp. 782–791.

40. Sundararajan, M., Taly, A., and Yan, Q. (2017). Axiomatic Attribution for Deep Networks (PMLR), pp. 3319–3328.

41. Jaume, G., Pati, P., Bozorgtabar, B., Foncubierta, A., Anniciello, A.M., Feroce, F., Rau, T., Thiran, J.-P., Gabrani, M., and Goksel, O. (2021). Quantifying Explainers of Graph Neural Networks in Computational Pathology, pp. 8106–8116.

42. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., and Joulin, A. (2021). Emerging Properties in Self-Supervised Vision Transformers, pp. 9650–9660.

43. Chen, R.J., and Krishnan, R.G. (2022). Self-supervised vision transformers learn visual concepts in histopathology. Preprint at arXiv. https://doi.org/10.48550/arXiv.2203.00585.

44. Dietterich, T.G., Lathrop, R.H., and Lozano-Pérez, T. (1997). Solving the multiple instance problem with axis-parallel rectangles. Artif. Intell. *89*, 31–71.

45. Reni, M., Gatta, G., Mazza, E., and Vecht, C. (2007). Crit. Rev. Oncol. Hematol. *63*, 81–89. https://doi.org/10.1016/j.critrevonc.2007.03.004.

46. Maaten, L.v.d., and Hinton, G. (2008). Visualizing data using t-SNE. J. Mach. Learn. Res. *9*, 2579–2605.

47. International Agency for Research on Cancer (IARC) (2016). Section 10 meningiomas. In WHO Classification of Tumours of the Central Nervous System (IARC Press).

48. Lyndon, D., Lansley, J.A., Evanson, J., and Krishnan, A.S. (2019). Dural masses: meningiomas and their mimics. Insights Imaging *10*, 11. https://doi.org/10.1186/s13244-019-0697-7.

49. Preusser, M., Charles Janzer, R., Felsberg, J., Reifenberger, G., Hamou, M.F., Diserens, A.C., Stupp, R., Gorlia, T., Marosi, C., Heinzl, H., et al. (2008). Anti-O6-methylguanine-methyltransferase (MGMT) immunohistochemistry in glioblastoma multiforme: observer variability and lack of association with patient survival impede its use as clinical biomarker. Brain Pathol. *18*, 520–532.

50. Choi, J., Lee, E.Y., Shin, K.-J., Minn, Y.-K., Kim, J., and Kim, S.H. (2013). IDH1 mutation analysis in low cellularity specimen: a limitation of diagnostic accuracy and a proposal for the diagnostic procedure. Pathol. Res. Pract. *209*, 284–290.

51. Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. Preprint at arXiv. https://doi.org/10.48550/arXiv.2005.14165.

52. Zhai, X., Kolesnikov, A., Houlsby, N., and Beyer, L. (2021). Scaling vision transformers. Preprint at arXiv. https://doi.org/10.48550/arXiv.2106.04560.

53. Louis, D.N., Perry, A., Wesseling, P., Brat, D.J., Cree, I.A., Figarella-Branger, D., Hawkins, C., Ng, H.K., Pfister, S.M., Reifenberger, G., et al. (2021). The 2021 WHO classification of tumors of the central nervous system: a summary. Neuro Oncol. *23*, 1231–1251. https://doi.org/10.1093/neuonc/noab106.

54. Simonyan, K., and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. Preprint at arXiv. https://doi.org/10.48550/arXiv.1409.1556.

55. Faust, K., Bala, S., van Ommeren, R., Portante, A., Al Qawahmed, R., Djuric, U., and Diamandis, P. (2019). Intelligent feature engineering and ontological mapping of brain tumour histomorphologies by deep learning. Nat. Mach. Intell. *1*, 316–321. https://doi.org/10.1038/s42256-019-0068-6.

56. Ståhl, P.L., Salmén, F., Vickovic, S., Lundmark, A., Navarro, J.F., Magnusson, J., Giacomello, S., Asp, M., Westholm, J.O., Huss, M., et al. (2016). Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. Science *353*, 78–82.

57. Chen, M., Zhang, B., Topatana, W., Cao, J., Zhu, H., Juengpanich, S., Mao, Q., Yu, H., and Cai, X. (2020). Classification and mutation prediction based on histopathology H&E images in liver cancer using deep learning. NPJ Precis. Oncol. *4*, 14. https://doi.org/10.1038/s41698-020-0120-3.

58. Kather, J.N., Pearson, A.T., Halama, N., Jäger, D., Krause, J., Loosen, S.H., Marx, A., Boor, P., Tacke, F., Neumann, U.P., et al. (2019). Deep learning can predict microsatellite instability directly from histology in gastrointestinal cancer. Nat. Med. *25*, 1054–1056. https://doi.org/10.1038/s41591-019-0462-y.

59. Aperio Technologies, I. (2008). Digital Slides and Third-Party Data Interchange. http://www.aperio.com/documents/api/Aperio_Digital_Slides_and_Third-party_data_interchange.pdf.

60. Shenzhen Shengqiang Technology Co (2021). Reading Software. https://www.sqray.com/yprj.

61. Goode, A., Gilbert, B., Harkes, J., Jukic, D., and Satyanarayanan, M. (2013). OpenSlide: a vendor-neutral software foundation for digital pathology. J. Pathol. Inf. *4*, 27.

62. Yu, K.-H., Zhang, C., Berry, G.J., Altman, R.B., Ré, C., Rubin, D.L., and Snyder, M. (2016). Predicting non-small cell lung cancer prognosis by fully automated microscopic pathology image features. Nat. Commun. *7*, 12474. https://doi.org/10.1038/ncomms12474.

63. Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding (Association for Computational Linguistics), pp. 4171–4186.

64. Sun, C., Shrivastava, A., Singh, S., and Gupta, A. (2017). Revisiting Unreasonable Effectiveness of Data in Deep Learning Era, pp. 843–852.

65. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al. (2015). Imagenet large scale visual recognition challenge. Int. J. Comput. Vis. *115*, 211–252.

66. Ba, J.L., Kiros, J.R., and Hinton, G.E. (2016). Layer normalization. Preprint at arXiv. https://doi.org/10.48550/arXiv.1607.06450.

67. Kraus, O.Z., Ba, J.L., and Frey, B.J. (2016). Classifying and segmenting microscopy images with deep multiple instance learning. Bioinformatics *32*, i52–i59. https://doi.org/10.1093/bioinformatics/btw252.

68. Schrammen, P.L., Ghaffari Laleh, N., Echle, A., Truhn, D., Schulz, V., Brinker, T.J., Brenner, H., Chang-Claude, J., Alwers, E., Brobeil, A., et al. (2022). Weakly supervised annotation-free cancer detection and prediction of genotype in routine histopathology. J. Pathol. *256*, 50–60. https://doi.org/10.1002/path.5800.

69. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. (2015). Going Deeper with Convolutions, pp. 1–9.

70. He, K., Zhang, X., Ren, S., and Sun, J. (2015). Deep residual learning for image recognition. Preprint at arXiv. https://doi.org/10.48550/arXiv.1512.03385.

71. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. (2016). Rethinking the Inception Architecture for Computer Vision, pp. 2818–2826.

72. Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K.Q. (2017). Densely Connected Convolutional Networks, pp. 4700–4708.

73. Ying, C., Cai, T., Luo, S., Zheng, S., Ke, G., He, D., Shen, Y., and Liu, T.-Y. (2021). Do transformers really perform bad for graph representation?. Preprint at arXiv. https://doi.org/10.48550/arXiv.2106.05234.

74. Kingma, D.P., and Ba, J. (2014). Adam: a method for stochastic optimization. Preprint at arXiv. https://doi.org/10.48550/arXiv.1412.6980.

75. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al. (2019). PyTorch: An Imperative Style (High-Performance Deep Learning Library).

76. Ishizawa, K., Hirose, T., Sugiyama, K., Kageji, T., Nobusawa, S., Homma, T., Komori, T., and Sasaki, A. (2012). Pathologic diversity of glioneuronal tumor with neuropil-like islands: a histological and immunohistochemical study with a special reference to isocitrate dehydrogenase 1 (IDH1) in 5 cases. Clin. Neuropathol. *31*, 67–76.

77. International Agency for Research on Cancer (IARC) (2016). Section 1 Diffuse astrocytic and oligodendroglial tumours. In WHO Classification of Tumours of the Central Nervous System (IARC Press).

78. Louis, D.N., Froschm, M.P., Mena, H., Rushing, E.J., and Judkins, A.R. (2009). Section 2 microscopic neuropathology. In Non-Neoplastic Diseases of the Central Nervous System (American Registry of Pathology).

79. International Agency for Research on Cancer (IARC) (2016). Section 17 Tumours of the sellar region. In WHO Classification of Tumours of the Central Nervous System (IARC Press).

80. Shrikumar, A., Greenside, P., and Kundaje, A. (2017). Learning Important Features through Propagating Activation Differences (PMLR), pp. 3145–3153.

81. Lundberg, S.M., and Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions, pp. 4768–4777.

82. Kokhlikyan, N., Miglani, V., Martin, M., Wang, E., Alsallakh, B., Reynolds, J., Melnikov, A., Kliushkina, N., Araya, C., and Yan, S. (2020). Captum: a unified and generic model interpretability library for pytorch. Preprint at arXiv. https://doi.org/10.48550/arXiv.2009.07896.

83. Virtanen, P., Gommers, R., Oliphant, T.E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., et al. (2020). SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat. Methods *17*, 261–272.

84. Wightmann, R. (2019). PyTorch image models. GitHub repository. https://doi.org/10.5281/zenodo.4414861.

85. Harris, C.R., Millman, K.J., Van Der Walt, S.J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N.J., et al. (2020). Array programming with NumPy. Nature *585*, 357–362.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
The Cancer Genome Atlas (TCGA)	National Cancer Institute	https://portal.gdc.cancer.gov/
In-house Primary Brain Tumor Dataset	This Paper	https://doi.org/10.5281/zenodo.7392203
Software and algorithms		
Numpy	Harris et al.[85]	https://numpy.org/
Scipy	Virtanen et al.[83]	https://scipy.org/
Pytorch	Paszke et al.[75]	https://pytorch.org/
PyTorch Image Models	Wightmann, R.[84]	https://github.com/rwightman/pytorch-image-models
ViT-WSI	This paper	https://github.com/lzx325/ViT-WSI-repo
Other		
NVIDIA A100 GPU	KAUST Ibex HPC Cluster	https://www.hpc.kaust.edu.sa/ibex
SQS-1000	Shenzhen Shengqiang Technology Co.	https://www.sqray.com/

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources should be directed to and will be fulfilled by the lead contact: Xin Gao (xin.gao@kaust.edu.sa).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- Histopathology Data from TCGA used in this study are available from the Genomic Data Commons Portal of the National Cancer Institute (https://gdc.cancer.gov/). In-house data from the First Affiliated Hospital of Harbin Medical University is deposited at Zenodo and is available from the lead contact upon reasonable request. The accession numbers are listed in the key resources table.
- The source code used in this study is available at https://github.com/lzx325/ViT-WSI-repo.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

To overcome the lack of publicly available primary brain tumor datasets, we retrieved 6,173 primary brain tumor H&E slides from The First Hospital of Harbin Medical University (Figure 1, Table 1). The consent forms of the patients were waived before this research was carried out under the retrospective research protocol of the institution. Our in-house dataset covers a total of 8 brain tumor types, including glioma, meningioma, pituitary adenoma, ependymoma, craniopharyngioma, CNS lymphoma, chordoma, and germ cell tumor. We retrieved the corresponding EHRs of the patients and kept only the slides that can be matched to their patient metadata and whose brain tumor type can be determined from the EHRs. In total, we assembled a cohort of 5,216 slides from 1,211 patients spanning the eight brain tumor types (Figure S1, Table 1). Summary statistics including sex and age of the studies patients are available in Table 1. Information of the subtypes and molecular biomarkers for some tumor types are further extracted from the EHR for a subset of the above-mentioned patients (Tables 1 and 2).

METHOD DETAILS

Data preparation and preprocessing

Preliminary format conversion, clean-up, and quality control are performed on the in-house brain tumor dataset. The dataset generated by the SQS-1000 scanner is in its proprietary sdpc format. They are first converted to the standard Aperio format (svs)[59] using the proprietary tool 'sdpc2svs'[60] provided by the scanner manufacturer. The converted svs format has compression type 'JPEG' with compression quality 80 and tile size of 672× 672. The slide identifier of each slide is extracted from the 'slide identifier label image' associated with each WSI and is subsequently used to query the EHRs of the patients to match against their diagnostic information. We then further discarded slides whose identifier is unclear/missing or not found in the EHRs (resulting in 1,247 patients and 5,502 slides, Figure S1) and whose scanning quality is poor after a brief manual inspection (resulting in 1,221 patients and 5,297 slides). Tumortype information was successfully extracted from their respective EHRs associated with each slide for a total of 1,211 patients and 5,216 slides, with their per-type statistics shown in Figure S1. These slides subsequently served as the '8-class top-level type classification' task (statistics in Table 1). Furthermore, meningioma and glioma subtype information was successfully extracted for 597 patients. The meningioma and glioma subtypes cover five glioma subtypes (anaplastic astrocytoma, anaplastic oligodendroglioma, diffuse astrocytoma, glioblastoma, and oligodendroglioma) and six meningioma subtypes (meningothelial, fibrous, transitional, angiomatous, atypical, and anaplastic). These slides subsequently served as the '11-class subtyping task' (statistics in Table 1). The hierarchy showing the relationship between the type and subtypes is shown in Figure S2. *IDH1* immunohistochemistry (IHC) tests status for 304 glioma patients (106 positives, 198 negatives, with positivity indicating mutation in the *IDH1* gene), *TP53* immunohistochemistry (IHC) tests status for 219 glioma patients (104 positives, 115 negatives, with positivity indicating mutation in *TP53* gene), and *MGMT* immunohistochemistry (IHC) tests status for 71 glioma patients (49 positives, 29 negatives, with negativity suggesting methylation), respectively (Figure 1). They are subsequently used for the molecular marker prediction tasks. For the TCGA slides, there is no need for format conversion because they are already in standard WSI formats (tiff or svs). All slides can be readily read programmatically with the OpenSlide whole slide imaging library.[61]

The WSIs were then segmented for tissue regions from the empty slide background. At a 32× down-sampled level of each WSI, the down-sampled image was converted from the RGBA space to the HSV space. The image was then converted to a binary mask using Otsu's thresholding method on the Gaussian blurred version of the saturation channel. Small artifacts such as holes and isolated points were further removed in order to carve out a representative connected component which covers the main tissue area. Subsequently, using a sliding window-based approach, the carved-out tissue area is completely covered with a number of image patches that are 1024×1024 in size, which serve as units for subsequent WSI analysis (Figure 1). This is because WSIs are typically gigapixel images that are too large to fit into the CPU/GPU memory when they are analyzed as a whole. Because the patches are created from the WSI directly and are approximately thousands by thousands of pixels in size, we will hereafter refer to them as 'WSI patches' or 'megapixel patches'.

Vision transformer (ViT)–based weakly-supervised whole-slide image analysis model (ViT-WSI)

For a given WSI (W_i), after the proper tissue segmentation and patching steps described in the previous section, they can now be represented as a list of patches, i.e., $W_i = [X_{ij}]_{j=1}^{n}$, where X_{ij} is a particular patch in the segmented WSI, and n is the total number of patches that the WSI has. In a fully-supervised setting, each of the patches is associated with a ground truth label Y_{ij}. In a weakly-supervised setting, however, one WSI is only associated with a single slide-level label, Y_i. A model's goal is to infer the slide-level label given the patches $[X_{ij}]_{j=1}^{n}$. There is a trivial solution to the weakly-supervised problem, which is to assign a patch-level label Y_{ij} for each patch to be the same as the slide-level label Y_i, as has been done in the previous works.[58,62] However, this solution could over-simplify the problem and assign incorrect labels to too many patches, which will confuse and lead to biased training of the model.

Transformers[23] are attention-based deep learning architectures that are originally designed for natural language processing (NLP) and have since achieved state-of-the-art performance on many NLP tasks.[51,63] More recently, Transformers have begun revolutionizing the computer vision field, where it achieves superior performance on several computer vision benchmark tasks. The Vision Transformer (ViT)[24] achieves state-of-the-art performance by supervised pretraining on a large, labeled dataset (e.g., JFT[64]) and then transfers the knowledge to the dataset of the task (e.g., the ImageNet 2012 dataset[65]). ViT processes images by splitting them into a number of patches (e.g., 16x16 patches) and treats the patches as if they were

tokens in an NLP task. As in the traditional Transformers, ViT produces representations of each patch (token), one layer after another, based on the attention mechanism. The prediction of the whole image is then aggregated at the last layer of the network.

The high performance of ViT on image recognition tasks demonstrates that Transformers' outstanding representation learning ability is generally applicable to computer vision tasks. Inspired by the success of ViT on natural image recognition and the similarity of its image processing procedure to the WSI processing procedure, i.e., by first dividing a large image into small patches, we are particularly interested in investigating whether a Transformer-based architecture can be useful for histopathological image analysis.

Specifically, we designed a Transformer-based model, termed ViT-WSI, to address the above weakly-supervised learning task. Suppose that we have an input WSI, $W_i = [X_{ij}]_{j=1}^{n}$. The ViT-WSI architecture is composed of the 'Encoder Layers,'which are based on the original Transformer encoder architecture used in BERT.[63] For a specific layer l ($1 \leq l \leq L$), it takes in n representations from the one layer below and outputs n representations as the current layer's output:

$$\left[X_{i1}^{(l)}, X_{i2}^{(l)}, \ldots, X_{in}^{(l)}\right] = \text{EncoderLayer}\left(\left[X_{i1}^{(l-1)}, X_{i2}^{(l-1)}, \ldots, X_{in}^{(l-1)}\right], \boldsymbol{\Theta}^{(l)}\right)$$

where each of the n input representations ($[X_{i1}^{(l-1)}, X_{i2}^{(l-1)}, \ldots, X_{in}^{(l-1)}]$) and n output representations ($[X_{i1}^{(l)}, X_{i2}^{(l)}, \ldots, X_{in}^{(l)}]$ corresponds to a specific patch in the input. $\boldsymbol{\Theta}^{(l)}$ contains the learnable layer parameters for the lth layer. Each Encoder Layer is made up of:

- A multi-head self-attention layer

$$\mathbf{X}' = \text{MultiHeadSelfAttention}(\mathbf{X}) = [\text{head}_1, \ldots, \text{head}_H]\mathbf{W}^O \quad \textit{where}$$

$$\text{head}_h = \text{Attention}\left(\mathbf{X}\mathbf{W}_h^Q, \mathbf{X}\mathbf{W}_h^K, \mathbf{X}\mathbf{W}_h^V\right)$$

$$\text{Attention}(\mathbf{K}, \mathbf{Q}, \mathbf{V}) = \frac{\text{Softmax}(\mathbf{Q}\mathbf{K}^\top)}{\sqrt{\text{d}}}\mathbf{V}$$

- A position-wise feed-forward layer

$$\mathbf{X}' = \text{MLP}\left(\mathbf{X}, \mathbf{W}^{\text{MLP}}\right)$$

- Layer Normalization[66] layers following the above two layers

$$\mathbf{X}' = \text{LayerNorm}(\mathbf{X})$$

In each of the above equations, $\mathbf{X}$ serves as a shorthand notation for the layer's input, which is formed by row-wise-stacking up the internal intermediate representations of each patch from its preceding layer, and $\mathbf{X}'$ as the notation for its output. $\boldsymbol{\Theta}^{(l)}$, the learnable layer parameters of the layer l, is a collection of the parameters above $\{\mathbf{W}^O, \mathbf{W}_h^Q, \mathbf{W}_h^K, \mathbf{W}_h^V, \mathbf{W}^{\text{MLP}}\}$, while d is the embedding dimension used throughout the two parts of the model.

The pipeline of ViT-WSI can be divided into two stages: The first stage consists of a *pretrained patch embedder* (Figure 1) that is responsible for transforming the WSI patches into token embeddings that will be used in the second stage. In the second stage, a *Vision Transformer aggregator* (Figure 1) aggregates the information across the WSI patches and summarizes them into a single, slide-level prediction. Both the pretrained patch embedder and the aggregator consist of the Encoder Layers as described above. For the

pretrained patch embedder, the parameters of its Encoder Layers are extracted from the ViT-L-16 model that is pretrained on the full Image21k dataset.[65] It is, therefore, able to produce high-level embedding for a specific WSI patch. As for the aggregator, it contains much fewer Encoder Layers compared to the patch embedder, with randomly initialized parameters.

A WSI is first segmented and split into megapixel patches as described in the previous section. Then, each patch is sent out to the patch embedder for its embedding computation. Inside the patch embedder, each patch is further split into even smaller patches, which are referred to as 'kilopixel patches' (Figure 1), as they are usually tens of pixels by tens of pixels in size. Inside the pretrained patch embedder, each kilopixel patch is a token. After being embedded with a linear embedder layer which is also pretrained,[24] they are sent to the Encoder Layers of the pretrained patch embedder for feature extraction. The output of the patch embedder is a patch embedding that holds the extracted features for one megapixel patch, which serves as *one token* in the aggregator. When the embeddings of all the megapixel patches of the WSI are computed and generated, they are input together to the aggregator for aggregation and prediction. The aggregator's Encoder Layers then process the patch embeddings layer by layer and finally summarize the representation of all megapixel patches of the WSI with a global average pooling (GAP) operation. The resulting vector serves as a *whole slide representation vector* (WSRV) and is attached to an MLP classification head for generating the final classification outcome.

The advantages of developing ViT-WSI to solve the weakly-supervised WSI classification tasks are three-fold. Firstly, the aggregator serves as a learnable aggregation function of the WSI megapixel patches. This is in contrast to the multiple instance learning (MIL)–based solution[44,67] of the weakly-supervised learning problem. In the latter, a fixed operation, typically a MAX operation, selects the prediction score from a single patch in a WSI image as the whole slide-level prediction. This can be inefficient for model training, as back-propagation is performed using only one patch and is also ineffective for making the prediction, as only one patch is actually used for the whole-slide prediction.[14] Most importantly, a fixed aggregation operation rules out the opportunity of the aggregator to learn a complex aggregation function over the WSI patches in a data-driven approach. On the contrary, our aggregator is not only a learned network but also makes full use of all patches within a WSI. The aggregator function is also fully differentiable, thereby making it more amenable to gradient-based interpretability analysis. Secondly, the ViT encoder-based aggregator fits nicely into the aggregation procedure of the WSI weakly-supervised learning problem, as (i) it naturally allows each input example to have a variable number (up to a certain memory limit) of tokens (i.e., WSI patches) and (ii) its inference procedure respects the pairwise relationship between patches by modeling them using self-attention. This context-aware approach is in contrast to the majority of previous works on WSI weakly-supervised learning, which aggregate patches by weighting them with attention scores that are calculated independently.[14,68] Based on the self-attention between patches, the ViT-WSI model's attention generally covers a greater area and has greater diversity than the aggregation methods that deal with patches independently (Figures S5 and S6); Thirdly, the ViT-based pretrained patch embedder is more powerful in performance than previous vision networks (e.g., VGG,[54] Inception,[69] and ResNet[70]), which are mostly based on convolutional neural networks (CNN). Moreover, designing the aggregator to be in a similar architecture as the pretrained patch embedder offers the advantage of introducing more compatibility between the parts, as demonstrated by a lot of network design cases with repeated blocks.[70–72]

To further improve the utilization of the topological structure of the WSI patches in the aggregation step, we built a WSI patch graph based on the patches' embedding similarity and inter-patch distances within the original WSI image. Then, the computation in each of the ViT encoder layers is modified to incorporate this graph information. Specifically, a nearest-neighbor graph $G_i = (V_i, E_i)$ is constructed with a WSI, W_i. $V_i = \{X_1, X_2, \ldots, X_n\}$ contains the WSI patches. And for each vertex, the k-nearest neighbors (kNN) in terms of location (coordinates) in the original WSI and the cosine similarity of the patch embeddings are connected with edges. Inspired by the recent Graphormer work,[73] the topological information in the graph G_i can be added to the aggregator as follows:

- The first layer's input of the aggregator is added with a node degree centrality embedding,[74] $z_{\deg(X_i)}$, i.e.

$$X_i^{(0)\prime} = X_i^{(0)} + z_{\deg(X_i)}$$

where $z_{\deg(X_i)}$ is a learned embedding, which is shared across the nodes that have the same degree centrality. Its dimension is the same as the embedding dimension of the patch embedder.

- The computation of self-attention is added with a bias term that depends on the shortest distance between the two nodes calculated by the Floyd-Warshall algorithm 74. Concretely, the attention between patch X_k and X_l can be calculated as:

$$\text{Attention}(X_k, X_l) = \frac{Q(X_k) \cdot K(X_l)}{\sqrt{d}} + b_{\varphi(X_k, X_l)}$$

where $Q(X_k)$ and $K(X_l)$ are the query vector and key vector used for the attention computation, respectively, while $b_{\varphi(X_k, X_l)}$ is the learned scalar bias term that is shared by the node pairs with the same shortest path distance.

Training and evaluation of ViT-WSI

For each of the prediction tasks, whether the 8-class top-level type classification task, the 11-class subtyping task, the TCGA glioma subtype classification task or the molecular biomarker prediction tasks, the datasets are randomly partitioned into ten folds for cross-validation and fine-tuning of hyperparameters. The performance is reported as the averaged performance over the ten folds. Another independent train-test split with a ratio of 7:3 is done on each task dataset. The model trained on this separate split is used for WSRV generation and downstream attribution analysis. Slides are carefully managed in the splitting process to ensure that all folds or splits have slides from distinct sets of patients. In all cases, the network is trained with the adaptive moment estimation (Adam) optimizer.[74] Each sample is weighted in the cross-entropy loss to overcome the class imbalance issue. We construct the network using the PyTorch[75] deep learning framework and utilize NVIDIA A100 Tensor Core GPU as the hardware platform.

Whole-slide representation vector (WSRV) produced by ViT-WSI

The whole representation vector (WSRV) is the network value extracted from the penultimate layer (before the classification head) of ViT-WSI. The WSRVs of the slides from the test set the type and subtype classification tasks are visualized in 2D using t-distributed stochastic neighbor embedding (t-SNE)[46] (Figures S3A and S3B). For the 8-class top-level type classification task (Figure S3A), one can observe a clear clustering of the glioma, meningioma, pituitary adenoma, and craniopharyngioma slides in the t-SNE space. Ependymoma slides are observed to be closer to the glioma slides, which agrees well with their higher misclassification rate and their histopathological similarity. For the 11-class subtyping task, there is a clear separation of all glioma subtypes and meningioma subtypes.

Closeness in the t-SNE space between oligodendroglioma (O) and anaplastic oligodendroglioma (AO), as well as atypical and anaplastic meningiomas, suggested a general difficulty in classifying tumors of similar origin and cell type but with only different levels of malignancy. The latest 2021 WHO Classification of Tumors of the Central Nervous System dropped the usage of 'anaplastic' in gliomas completely and considered them only as a grading difference from their non-anaplastic counterparts, acknowledging the subtlety of histological difference between them.[53] In this study, we still adhered to the 2016 WHO classification system, as the 2021 WHO classification system is still too new to be adopted by most healthcare institutions worldwide.

Self-attention visualization of ViT-WSI

One of the distinctive features of ViT-WSI lies in its Vision Transformer-based aggregator. It aggregates patch-level embeddings using the multi-head self-attention. The multi-head attention aggregates patches through several independent heads, creating an 'ensemble effect' that is beneficial to the model's performance and generalizability. The self-attention mechanism models the interdependence between patches, in contrast to most previous weakly-supervised WSI learning algorithms that process each WSI patch independently.[14,15]

The multi-head self-attention of ViT-WSI is illustrated with three examples, including one glioblastoma (GBM) example (Figure S4A), one transitional meningioma example (Figure S4B), and one craniopharyngioma example (Figure S4C). For each slide, we observed high agreement between the aggregated

attention map across different heads (Figure S4, the third column, 'Aggregated Attention Map') with at least one of the pathologist-annotated lesions (Figure S4, the first column). Different regions of the slide are attended to differently by different heads (Figure S4, the second column, 'Per-head Self-Attention Map'). When attention scores from multiple heads are aggregated, some parts cancel out, while others are reinforced.

Figure S3 also illustrates how the ViT-WSI self-attention mechanism allows the model to discover semantically similar regions in an unsupervised fashion. By visualizing self-attention between a specific query patch and all other patches in the WSI, semantically similar regions are specifically highlighted (Figure S4, the second column of the lower panel, 'Per-patch Self-Attention Map'). Concretely, 'neuropil structures' and 'blood cell scattered between malignant cells' due to hemorrhage are recognized in glioblastoma[76,77]; 'whorl structures' and 'arachnoidal hyperplasia' are recognized in transitional meningioma[47,78]; 'calcification' and 'fibrocyte/fibroblast region' are recognized in craniopharyngioma.[79] The patches which are highly attended in self-attention are shown to have similar semantics to the query patch in attention computation (Figure S4, the third column of the lower panel).

Attribution analysis of ViT-WSI

Attribution analysis has been adopted as an important strategy for neural network model explanation and interpretation. It contains a family of algorithms that achieve this goal by calculating real-valued importance scores that 'attribute' the result of a network output to a network input.[37] The importance scores, which serve as the attribution, form an array that has exactly the same shape as the network input, with each being the attribution for each input element. There are a variety of attribution algorithms, including Integrated Gradients,[40] DeepLIFT,[80] and SHAP.[81] All of the above three algorithms can calculate the attribution by comparing what the network outputs from a given input to what it outputs from a 'baseline' input (which is usually chosen to be an all-zero input or a random input) and assign attribution scores to the elements of the given input according to their contribution to the difference. A positive score indicates a positive contribution while a negative indicates a negative contribution.

In the attribution analysis of ViT-WSI, we use Integrated Gradients (hereafter abbreviated as 'IG') to perform attribution analysis. We will illustrate how ViT-WSI can be amenable to this gradient-based attribution analysis under various configurations. Suppose that the input to the ViT-WSI aggregator (represented as function F) is $\mathbf{X}^{(\mathrm{inp})} = [X_1^{(inp)}, X_2^{(inp)}, \ldots, X_n^{(inp)}]$ (where n is the number of megapixel patches) and the output is a scalar $F(\mathbf{X}^{(\mathrm{inp})})$. IG computes the attribution of the kth input element as follows:

$$\mathrm{IG}_k\left(\mathbf{X}^{(\mathrm{inp})}\right) = \left(\mathbf{X}_k^{(\mathrm{inp})} - \mathbf{X}_k^{(\mathrm{baseline})}\right) \int_{\alpha=0}^{1} F_k'\left(\mathbf{X}^{(\mathrm{baseline})} + \alpha\left(\mathbf{X}^{(\mathrm{inp})} - \mathbf{X}^{(\mathrm{baseline})}\right)\right) d\alpha$$

where F_k' is the partial derivative of the network F w.r.t. the kth input element, $\mathbf{X}^{(\mathrm{baseline})}$ is the above-mentioned baseline input and α is the interpolating factor which varies from 0 to 1. This guarantees that the sum of the attribution of each input element equals the total output difference, which can be expressed as:

$$F\left(\mathbf{X}^{(\mathrm{inp})}\right) - F\left(\mathbf{X}^{(\mathrm{baseline})}\right) = \sum_k \mathrm{IG}_k\left(\mathbf{X}^{(\mathrm{inp})}\right)$$

Overall, the following three types of outputs are used in the ViT-WSI attribution analysis.

- Class attribution. The classification score of a particular class is used as the network output value in the attribution analysis. This score quantifies the contribution of the input to the likelihood of a particular class. Examples are shown in Figures 3A and 3B.
- WSRV attribution. A particular dimension of WSRV is used as the network output value. This assesses the contribution of the input to a particular dimension in the WSI representation. Examples are illustrated in Figures S8G, S8I, S8H, and S8J.
- PC attribution. To systematically interpret the ViT-WSI model, we performed the Principal Component Analysis (PCA) on ViT-WSI computed WSI representations of the test set. The percentage of variance explained by the top PCs is plotted in Figure S7. To distinguish this from another PCA performed on the input to the ViT-WSI aggregator (described later), we termed this as 'output PCA

projection.' Suppose that the PCA-learned components matrix is $W_{comp}^{(out)} \in \mathbb{R}^{N_{comp} \times N_{dim}}$ and mean vector $b_{mean}^{(out)} \in \mathbb{R}^{N_{dim}}$. A WSI representation $r^{(out)} \in \mathbb{R}^{N_{dim}}$ can be projected onto the PCA principal components as follows:

$$r_{proj}^{(out)} = W_{comp}^{(out)} \left(r^{(out)} - b_{mean}^{(out)} \right)$$

where the ith dimension of $r_{proj}^{(out)}$, $r_{proj,i}^{(out)}$, is the column 'PC i' shown in Figures 3C and 3G. Its attribution to the input is computed in the same way as described above. Examples are shown in Figures 3E, 3F, 3I, and 3J.

Due to the large dimension (1024) of the feature vector produced by the ViT-WSI feature extractor, if they are used as the input as a whole, the IG attribution method will fail to converge under typical memory constraints. For all the three above-mentioned attribution analyses, we first precomputed another PCA projection on the input feature vector to get the 'projected version' of each input patch feature as follows:

$$r_{proj}^{(in)} = W_{comp}^{(in)} \left(r^{(in)} - b_{mean}^{(in)} \right)$$

During network forward propagation, a reconstructed feature vector can be computed from the projection:

$$r_{reconstruct}^{(in)} = W_{comp}^{(in)\ \top} r_{proj}^{(in)} + b_{mean}^{(in)}$$

Attribution is then performed from the three above-mentioned output values to $r_{proj}^{(in)}$. In this study, the attribution to a particular WSI megapixel patch is defined as the averaged attribution on the first 30 PCs. The whole attribution analysis is implemented in PyTorch with the utilization of the Captum interpretability library.[82] Finally, the attribution heatmap is plotted separately for the positively (red) and negatively (green) attributed patches with a quantile-normalized color map.

QUANTIFICATION AND STATISTICAL ANALYSIS

Student's t test, Mann-Whitney U Test, Wilcoxon signed-rank test, and Kruskal-Wallis H Test were performed using the statistical functions from the Scipy package (scipy.stats).[83] Statistical significance (p values) was reported in respective figures with * ($p < 0.05$), ** ($p < 0.01$), *** ($p < 0.001$), and **** ($p < 0.0001$).