HHS Public Access

Author manuscript

Nat Med. Author manuscript; available in PMC 2023 August 23.

Published in final edited form as:

Nat Med. 2023 April ; 29(4): 828–832. doi:10.1038/s41591-023-02252-4.

AI-based molecular classification of diffuse gliomas using rapid, label-free optical imaging

Todd Hollon[1,*], **Cheng Jiang**[1], **Asadur Chowdury**[1], **Mustafa Nasir-Moin**[2], **Akhil Kondepudi**[1], **Alexander Aabedi**[1], **Arjun Adapa**[1], **Wajd Al-Holou**[1], **Jason Heth**[1], **Oren Sagher**[1], **Pedro Lowenstein**[1], **Maria Castro**[1], **Lisa Irina Wadiura**[1], **Georg Widhalm**[1], **Misha Movahed-Ezazi**[1], **Volker Neuschmelting**[1], **David Reinecke**[1], **Niklas von Spreckelsen**[1], **Mitchell Berger**[1], **Shawn Hervey-Jumper**[1], **John Golfinos**[1], **Sandra Camelo-Piragua**[1], **Christian Freudiger**[1], **Honglak Lee**[1], **Daniel Orringer**[2,*]

[1]Machine Learning in Neurosurgery Laboratory, Department of Neurosurgery, University of Michigan, 1500 E. Medical Center Dr., Ann Arbor, 48105, MI, USA.

[2]Department of Neurosurgery, New York University, Street, City, 10587, State, Country.

Abstract

Molecular classification has transformed the management of brain tumors by enabling more accurate prognostication and personalized treatment. Access to timely molecular diagnostic testing for brain tumor patients is limited [1–3], complicating surgical and adjuvant treatment and obstructing clinical trial enrollment [4]. We developed a rapid (<90 seconds), AI-based diagnostic screening system that can provide molecular classification of diffuse gliomas and report its use in a prospective, multicenter, international testing cohort of diffuse glioma patients (N = 153). By combining stimulated Raman histology (SRH), a rapid, label-free, non-consumptive, optical imaging method [5–7], and deep learning-based image classification, we are able to predict the molecular features used by the World Health Organization (WHO) to define the adult-type diffuse glioma taxonomy [8]. We developed a transformer-based multimodal training strategy that uses a pretrained SRH image feature encoder and a large-scale, genetic embedding model to achieve optimal molecular classification performance. Using this system, called DeepGlioma, we were able to achieve an average molecular genetic classification accuracy of 93.2% and identify the correct diffuse glioma molecular subgroup with 91.5% accuracy. Our results represent how artificial intelligence and optical histology can be used to provide a rapid and scalable alternative to wet lab methods for the molecular diagnosis of brain tumor patients.

[*]Corresponding author(s). tocho@med.umich.edu; daniel.orringer@nyulangone.org.
Contributing authors: chengjia@med.umich.edu; achowdur@med.umich.edu; mustafa.nasir-moin@nyulangone.org; akhilk@umich.edu; alexander.aabedi@ucsf.edu; arjunra@med.umich.edu; wna@med.umich.edu; jheth@med.umich.edu; osagher@med.umich.edu; pedrol@med.umich.edu; mariacas@med.umich.edu; lisa.wadiura@meduniwien.ac.at; georg.widhalm@meduniwien.ac.at; misha.movahed-ezazi@nyulangone.org; volker.neuschmelting@uk-koeln.de; david.reinecke@uk-koeln.de; niklas.von-spreckelsen@uk-koeln.de; mitchel.berger@ucsf.edu; shawn.hervey-jumper@ucsf.edu; john.golfinos@nyulangone.org; sandraca@med.umich.edu; chris@invenio-imaging.com; honglak@eecs.umich.edu;

ToDo:
1. add authors contributions

Keywords

Molecular classification; artificial intelligence; optical imaging; diffuse gliomas; deep learning

1 Main

Molecular classification is increasingly central to the diagnosis and treatment of human cancers. Diffuse gliomas, the most common and deadly primary brain tumors, are now defined through a handful of molecular markers [9–11]. However, molecular subgrouping of diffuse gliomas requires immunohistochemistry, cytogenetic testing and, often, next-generation sequencing, that is not available in all centers caring for brain tumor patients. Moreover, the expert interpretation of molecular data is increasingly challenging in the setting of a declining pathology workforce [12]. Consequently, molecular diagnostic and sequencing techniques for brain tumors are not universally available and may be associated with long turnaround times even in well-resourced settings (days-weeks) [2, 13, 14]. Barriers to molecular diagnosis can result in suboptimal care for brain tumor patients complicating prognostic prediction, surgical decision-making, selection of adjuvant chemoradiation, and clinical trial enrollment. Computational approaches to predict molecular genetics based on clinical, radiographic and pathomic data have been proposed [15–17] but none have achieved the accuracy of conventional molecular diagnosis, been integrated into clinical use, or rigorously studied in prospective multi-institutional trials. Here, we propose and prospectively validate an approach to simplify molecular classification of diffuse gliomas through AI-based analysis of rapid optical imaging of fresh, unprocessed surgical specimens without the need for conventional pathology laboratory resources.

AI-based molecular diagnosis

DeepGlioma is an AI-based diagnostic system that combines deep neural networks and SRH to achieve rapid molecular classification of fresh glioma specimens (Fig. 1). Our approach predicts the most critical diagnostic genetic alterations in diffuse glioma and can achieve molecular classification in less than two minutes of tissue biopsy without the need for tissue processing or human interpretation (Extended Data Fig. 1). While DeepGlioma can scale to an arbitrary number of diagnostic mutations, we focus on the major molecular diagnostic features used by the WHO CNS5[8] to define the diffuse glioma subgroups: isocitrate dehydrogenase-1 and 2 (IDH) mutations, 1p19q chromosome co-deletion (1p19q-codel), and ATRX loss.

Patch contrastive learning for SRH representation learning

The SRH workflow begins when a fresh, unprocessed surgical specimen is biopsied from a brain tumor patient and a small (3×3mm) sample is placed into a premade microscope slide (Fig. 1a and Extended Data Fig. 1a). The slide is inserted into the SRH imager and images are acquired at two Raman shifts: 2,845cm^{-1} and 2,930cm^{-1} [6, 7]. SRH patches are then sampled from the whole slide SRH image to generate non-overlapping, single-scale, high-resolution, high-magnification images for model training and inference. Our molecular classification model was trained using a multimodal approach that included

two datasets: clinical SRH images and genomic sequencing data. First, we used SRH images from 373 adult diffuse glioma patients treated at the University of Michigan to train a deep convolutional neural network (CNN, ResNet50[18]) as a visual encoder for optical image feature learning (Extended Data Table 1 and Extended Data Fig. 2). Molecular classification is a multi-label classification task, such that the model must predict the mutational status of multiple genetic mutations. While previous studies have used linear classification layers trained end-to-end using cross-entropy [17, 19, 20], we found that weakly supervised (i.e. patient labels only) patch-based contrastive learning, or patchcon, was ideally suited for whole slide SRH classification (Fig. 1b and Extended Data Fig. 3) [21, 22]. We then developed a simple and general framework for multi-label contrastive learning of visual representations and trained our SRH encoder using this framework (Extended Data Fig. 4).

Pretrained genetic embedding using public genomic data

Next, we pretrained a genetic embedding model using large-scale, public glioma genomic data (Fig. 1b and Extended Data Table 2). Inspired by previous work on learning a joint semantic-visual embedding space for natural image classification[23, 24] and text-to-image generation[25, 26], we aimed to learn a genetic embedding space that meaningfully encodes the relationships between mutations in order to improve SRH classification. The co-occurrence of specific mutations in the same tumor defines the molecular subgroups of diffuse gliomas [9, 10, 27]. Our genetic embedding model learns to represent these co-occurrences using global vector embeddings [28]. The model learns embedding vectors for each mutation that match the global gene-gene co-occurrence statistics in the aggregated public genomic dataset. Our training strategy learned a linear substructure that matches known molecular subgroups of diffuse gliomas (Extended Data Fig. 5). By pretraining an embedding model using a large genomic dataset, DeepGlioma can be trained using the known genomic landscape of diffuse gliomas, allowing for efficient multi-label molecular classification using SRH image features.

Transformer-based molecular classification

Finally, the pretrained SRH and genetic encoders are integrated into a single transformer architecture for multi-label molecular classification (Fig. 1c). Transformers map an input sequence of data to a learned latent representation using a multi-headed self-attention mechanism [29]. During transformer training, the input tokens to the transformer are the visual embedding of the SRH patch and the genetic embedding for the patient's tumor. Similar to masked bidirectional training of transformers for language understanding, we randomly mask a subset of the genes from the input and the objective is to predict the masked genes [30, 31]. During inference, the transformer uses only the SRH patch embedding to predict the mutational status of each gene. To take full advantage of the genetic encoder pretraining and learned molecular subgroup substructure, we enforce that the transformer output is also the pre-trained embedding space (Extended Data Fig. 6). We performed iterative hold-out cross-validation to show the advantage of patchcon, genetic pretraining, and masked label transformer training through several ablation studies. Specifically, we demonstrated that DeepGlioma was able to achieve a mean area under the receiver operator characteristic curve (mAUROC) of 92.6 ± 5.4% for molecular

genetic classification on cross-validation experiments performed on held-out UM SRH data (Extended Data Fig. 7).

Molecular genetic prediction in a prospective testing cohort

We tested DeepGlioma in a multicenter, prospective cohort of primary, non-recurrent diffuse gliomas to evaluate how our model generalizes across different patient populations, clinicians, infrastructures, and SRH imaging systems. Model testing was designed as a non-inferiority diagnostic clinical trial to determine the minimal sample size (N = 135). Four tertiary medical centers across the United States (New York University, University of California San Francisco) and Europe (Medical University of Vienna, University Hospital Cologne) were included as recruitment centers for the external prospective testing cohort. Patients were recruited as a consecutive cohort of adult (>18) brain tumor patients who underwent biopsy or tumor resection for a suspected diffuse glioma. A total of 153 patients were included (Extended Data Table 3). DeepGlioma achieved a molecular diagnostic classification accuracy for IDH mutation of 94.7% (95% CI 90.0–97.7%), 1p19q-codeletion of 94.1% (95% CI 89.1–97.3%), and ATRX mutation of 91.0% (95% CI 85.1–94.9%). Classification performance on each molecular diagnostic mutation is shown in Fig. 2a. Despite training and testing dataset imbalance due to different incidences among each mutation, DeepGlioma achieved F1 scores of 96.3%, 96.6%, and 94.7% for IDH, 1p19q codeletion, and ATRX, respectively.

Leave-insitution-out cross-validation

To rigorously evaluate molecular classification performance, we performed a set of leave-institution-out cross-validation (LIOCV) experiments in order (1) to assess the stability of DeepGlioma performance across multiple medical centers and (2) to determine the effect of increasing training data on model performance. Results are presented in Fig. 2b. For each LIOCV iteration, one of the testing medical centers (NYU, UCSF, UKK, MUV) was left out as a validation set and the remaining medical centers were combined to form a multicenter training set. Each LIOCV model was trained for a fixed number of epochs without hyperparameter tuning. DeepGlioma demonstrated stability across each LIOCV iterations with molecular classification accuracy standard deviation range of ±2.75–6.06% and a F1 score range of ±1.71–4.70%. The prediction of ATRX mutations was consistently more challenging across our experiments. ATRX mAUROC variance was ±9.10%, which was larger than IDH (±2.73%) and 1p19q-codeletion (±2.02%). We hypothesize that this is related to ATRX mutations being variably present in IDH-mutant astrocytomas and can occur in IDH-wildtype glioblastomas [10]. However, our LIOCV results indicate that this challenge can be addressed with additional training data. DeepGlioma LIOCV classification performance of ATRX mutation improved by a minimum of +2% across all evaluation metrics compared to our prospective clinical testing results. ATRX mAUROC increased by +5.0% and mean accuracy increased by +2.2%. 1p19q codeletion classification performance also increased with additional training data, achieving a mean classification accuracy of 97.0 ±3.5%.

DeepGlioma performance compared to IDH-1 immunohistochemistry

Next, we compare the diagnostic performance of DeepGlioma versus the current gold-standard molecular screening modality for diffuse glioma classification: IDH1-R132H immunohistochemistry (IHC). IDH1-R132H IHC has known limitations due to the presence of non-canonical IDH-1 mutations, such as R132C and R132S, and IDH-2 mutations. Non-canonical mutations occur in around 70–80% of lower grade gliomas [11, 33]. Due to the higher rates of lower grade gliomas in young patients, both the US[34] and Europe[35] recommend genetic sequencing for patients 55 year or less to avoid false negative screening from IDH1-R132H IHC. DeepGlioma was trained to generalize to both canonical and non-canonical IDH mutations. We used the largest, unbiased dataset of diffuse glioma patients (n = 482) who underwent both IHC and DNA sequencing to determine IDH mutational status [33]. IDH1-R132H IHC has a balanced accuracy of 91.4% (sensitivity 82.8%, specificity 100%). In our prospective, multicenter, testing cohort, DeepGlioma achieved a balanced accuracy of 94.2% (sensitivity 95.5%, specificity 93.0%). In patients 55 years or less, IDH1-R132H has a balanced accuracy of 90.0% (sensitivity 80.0%, specificity 100%) and DeepGlioma achieved a balanced accuracy of 97.0% (sensitivity 94.1%, specificity 100%) (Fig. 2c). The increased classification performance over IDH1-R132H is due to DeepGlioma's ability to detect all IDH mutations, resulting in an increase in model sensitivity. All non-canonical mutations in our prospective cohort, which included IDH1-R132S and IDH-2 mutations, were correctly classified by DeepGlioma (Extended Data Fig. 8a).

Diffuse glioma molecular subgroup classification

Finally, DeepGlioma's accurate prediction of the molecular genetics of diffuse gliomas allows for direct classification of SRH images into the set of mutually exclusive adult-type diffuse glioma molecular subgroups as defined by the WHO CNS5 classification scheme [8]. Using the three molecular predictions from DeepGlioma, an algorithmic inference method was developed to classify each patient into a molecular subgroup (Algorithm 1). We established an AI-based performance benchmark motivated by our previous methods for SRH classification using a ResNet50 model trained using categorical cross-entropy for multiclass classification [7, 19]. In our full prospective testing cohort, DeepGlioma achieved a diffuse glioma molecular subgroup classification accuracy of 91.5% (95% CI 86.0–95.4%) (Fig. 2d) and demonstrated a +4.6% performance increase over our multiclass model (Extended Data Fig. 8b, c). The major performance gains of DeepGlioma are due to increased sensitivity for identifying IDH-mutant diffuse gliomas and explicitly modelling the co-occurences of mutations within molecular subgroups (e.g. co-occurence of IDH mutations and 1p19q co-deletions). In patient 55 years or less, our classification performance showed an overall increase (+2.9%), obtaining an classificaton accuracy of 94.4% (95% CI 87.3–98.2%) (Fig. 2d and Extended Data Fig. 8d). We then evaluated the performance of DeepGlioma across each of the external testing medical centers. Average medical center accuracy was 90.4% (range 77.8% – 100%) (Extended Data Fig. 8e). These results show that DeepGlioma performance generalized well to multiple medical centers despite distinct patient populations, clinical presentations, personnel, and infrastructure. Molecular subgroup prediction heatmaps for both canonical (see Extended Data Fig. 9) and non-canonical IDH mutations (see Extended Data Fig. 10) were generated to improve

model interpretability and map DeepGlioma predictions to SRH image features. High-resolution molecular genetic and molecular subgroup predictions can be accessed through our interactive DeepGlioma website (deepglioma.mlins.org).

Discussion

We present DeepGlioma, a deep learning-based diagnostic system that provides accurate molecular genetic predictions using rapid, label-free optical imaging of fresh diffuse gliomas surgical specimens. Utilizing a multimodal dataset, DeepGlioma was trained to predict molecular genetic mutations by analyzing SRH images within the contexst of the known genomic landscape of diffuse gliomas. Here, we show that using only SRH images as input, DeepGlioma can predict the genetic mutations that define the WHO classification of adult-type diffuse gliomas within minutes of tumor biopsy without the need for any tissue processing or conventional pathology laboratory infrastructure. In our study, DeepGlioma outperformed standard immunohistochemical staining for detecting canonical and noncanonical IDH mutations with an accuracy of 94.2%. DeepGlioma also predicted 1p19q-codeletion and ATRX mutations without the need for fluorescence in-situ hybridization or genetic sequencing enabling molecular subtyping of diffuse gliomas according to the WHO classification scheme with an accuracy of 91.5%.

Access to molecular diagnostic testing is uneven for patients who receive brain tumor care. In settings where timely molecular data is not readily available, send-out testing is required to establish a diagnosis and the optimal clinical management plan. DeepGlioma can streamline molecular testing by providing rapid molecular prediction, enabling clinicians to focus on confirming the most likely diagnostic mutations only, rather than using a shotgun approach [36]. In addition, SRH is not consumptive and does not diminish diagnostic yield of tumor specimens, preserving scant clinical samples for definitive molecular testing. In patient care environments where in-house or send-out testing is not available, DeepGlioma could serve as a fully autonomous option for tailoring brain tumor treatment.

Streamlining molecular classification could also have an immediate impact on the surgical care of brain tumor patients. A growing body of evidence supports that surgical goals should be tailored based on molecular subgroups [37–39]. Notably, extent of resection carries a greater impact on survival in molecular astrocytomas (IDH-mutant, 1p19q-intact) than oligodendrogliomas (IDH-mutant, 1p19q-codeleted). DeepGlioma creates an avenue for accurate and timely differentiation of diffuse glioma subgroups in a manner that can be used to define surgical goals with a better calibrated risk-benefit analysis.

Even with optimal standard-of-care treatment, diffuse glioma patients face limited treatment options. Consequently, the development of novel therapies through clinical trials is essential. Unfortunately, fewer than 10% of glioma patients are enrolled in clinical trials [40]. One of the major roadblocks lies in the identification of eligible candidates. Clinical trials limit inclusion criteria to a specific sub-population, often defined by genetic subgroups. In some cases, reliance on a central review of histologic and molecular data can be an insurmountable logistic barrier to trial enrollment. By providing an avenue for rapid molecular classification, DeepGlioma can initiate the process for trial enrollment at the earliest stages of patient care. Moreover, DeepGlioma can facilitate clinical trials that rely

on intraoperative local delivery of agents into the surgical cavity and circumvent the blood-brain barrier [41–44], a major challenge in therapeutic delivery.

In conclusion, our study demonstrates how AI-based diagnostic methods have the potential to augment our existing wet laboratory techniques to improve the access and speed of molecular diagnosis in the comprehensive care of brain tumor patients.

2 Methods

2.1 Study design

The main objectives of the study were to (1) develop a rapid molecular diagnostic screening tool for classifying adult-type diffuse gliomas into the taxonomy defined by the WHO CNS5 [8] using clinical SRH and deep learning-based computer vision methods and (2) test our molecular diagnostic screening tool in a large multicenter prospective clinical testing set. We aimed to demonstrate that key molecular diagnostic mutations produce learnable spectroscopic, cytology, histoarchitectural changes in SRH images that allow for accurate molecular classification. We aimed to make a clinical contribution by demonstrating that our trained diagnostic system, DeepGlioma, could robustly and reproducibly screen fresh diffuse gliomas specimens for specific mutations to inform intraoperative decision making and potentially improve early clinical trial enrollment. DeepGlioma consists of two pretrained separable modules, a visual encoder and a genetic encoder, that are integrated together using a multi-headed attention mechanism for image classification [29]. Inspired by previous work on deep visual-semantic embedding [23], our aim was to use multimodal data that included both imaging and genomic data to achieve optimal performance on a multi-label genetic classification task. The primary SRH dataset for model training and validation was from the University of Michigan and the prospective testing dataset was collected from four international institutions: (1) New York University (NYU), (2) University of California, San Francisco (UCSF), (3) Medical University of Vienna (MUV), and (4) University Hospital Cologne (UKK). We focused on predicting the most clinically important molecular aberrations in diffuse gliomas, but aimed to develop a model architecture that could scale to any number of recurrent mutations in human cancers. For the purposes of this study, we focused our classification task on three key molecular aberrations found in adult-type diffuse gliomas: IDH mutation, 1p19q-codeletion, and ATRX mutation.

2.2 Stimulated Raman histology

Fiber-laser-based stimulated Raman scattering microscopy was used to acquire all images in our study [5, 45]. Detailed description of laser configuration has been previously described [6]. In brief, surgical specimens are stimulated with a pump beam at 790nm and Stokes beam that has a tunable range from 1015nm-1050nm. These laser settings allow for access to the Raman shift spectral range between 2800cm^{-1} – 3130cm^{-1}. Images are acquired as 1000 pixel strips with an imaging speed of 0.4Mpixel(s) per strip. We acquire two image channels sequentially at 2845 cm^{-1} (CH2 channel) and 2930cm^{-1} (CH3 channel) Raman wavenumber shifts. A strong stimulated Raman signal at 2845cm^{-1} corresponds to the CH2 symmetric stretching mode of lipid-rich structures, such as myelin. A Raman peak at 2930cm^{-1} highlights protein- and nucleic acid-rich regions such as the cell nucleus.

The first and last 50 pixels on the long axis of each strip are removed to improve edge alignment and the strips are concatenated along the long dimension to generate a stimulated Raman histology image [6]. A virtual hematoxylin and eosin (H&E) colorscheme can be applied to the two Raman channels to generate a three-channel, virtually-stained RGB SRH image. These images are used for clinician interpretation and designed to replicate the image contrast seen in conventional HE histology, but are not used for model training. An overview of SRH imaging workflow can be found in Extended Data Fig. 1.

2.3 Image data processing

All model training and inference was done using the raw, non-virtually colored SRH images. All images are acquired, processed, and archived as 16-bit images to retain spectroscopic image features. Each strip has a 900 pixels width (i.e. after edge clipping) and up to 6000 pixel height. Field flattening correction is used to correct for variation in pixel intensities within image strips. To account for tissue shifts that occur during and between image channel acquisition, the sequentially acquired CH2 and CH3 strips are co-registered using a discrete Fourier transform-based technique for translation, rotation, and scale-invariant image registration [46]. Following registration, a pixel-wise subtraction between the CH3 and CH2 channels generates a third ‘red’ channel that highlights the cell nuclei and other protein-rich structures. The whole slide SRH images are finally split into 300×300 pixel patches without overlap using a sliding window over the full image. SRH patches are then classified into one of three classes, tumor, normal brain, or nondiagnostic tissue, using our previous trained whole slide SRH segmentation model [7, 19]. Only tumor regions are used for DeepGlioma training and inference (Extended Data Fig. 1c).

2.4 Patient enrollment and training dataset generation

Clinical SRH imaging began at the University of Michigan on 1 June 2015 following Institutional Review Board approval (HUM00083059). Our imaging dataset was generated using two SRH imaging systems. An initial prototype SRH imager[6] and the NIO Imaging System (Invenio Imaging, Inc., Santa Clara, CA) [7]. All patients with a suspected brain tumor are approached for intraoperative SRH imaging. Inclusion criteria for SRH imaging are patients who are undergoing surgery for (1) suspected central nervous system tumor and/or (2) epilepsy, (3) subject or durable power of attorney is able to provide consent, and (4) preoperative expectation that additional tumor tissue will be available beyond what is required for clinical pathologic diagnosis. Exclusion criteria were (1) insufficient diagnostic tissue as determined by surgeon or pathologist, (2) grossly inadequate tissue (e.g. hemorrhagic, necrotic, fibrous, liquid, etc.), or (3) SRH imager malfunction. Following intraoperative SRH imaging, inclusion criteria for the diffuse glioma training dataset were the following: (1) 18 years or older and (2) final pathologic diagnosis of an adult-type diffuse glioma as defined by WHO CNS5 classification [8]. Exclusion criteria was less than 10% area segmented as tumor by our trained SRH segmentation model. UM dataset generation was stopped on 11 November 2021 and a total of 373 patients were included for model training and validation. Patient demographics and molecular diagnostic information can be found in Extended Data Table 1 and Extended Data Fig. 2.

2.5 Multi-label contrastive visual representation learning

Visual representation learning entails learning a parameterized mapping from an input image to a feature vector that effectively represents the most important image features for a given computer vision task. We used a ResNet50 architecture[18] for SRH feature extraction and did not find that larger models provided better performance. While much of our previous work used conventional cross-entropy loss functions to train deep neural networks [6, 7, 19], we found that contrastive loss functions result in better visual representation learning [21, 22]. We trained our model using a supervised contrastive loss:

$$\mathscr{L}^{sup} = \sum_{i \in I} \mathscr{L}_i^{sup} = -\sum_{i \in I} \frac{1}{|P(i)|} \sum_{p \in P(i)} \log \frac{\exp(\mathrm{sim}(g(z_i), g(z_p))/\tau)}{\sum_{n \in A(i)} \exp(\mathrm{sim}(g(z_i), g(z_n))/\tau)} \quad (1)$$

where $z = f(x) \in \mathbb{R}^d$ is the d-dimensional feature vector of image x after a feedforward pass through the visual encoder $f(\cdot)$. A linear projection layer $g(\cdot)$ maps the image feature vector z_i to a 128-dimensional space where the contrastive objective is computed. z_p is a feature vector from the set of paired positive examples, $P(i)$, for feature vector z_i, and $A(i)$ is the set of all images in a minibatch. $\tau \in \mathbb{R}^+$ is a temperature hyperparameter. Paired positive examples are images sampled from the same label. The cosine similarity metric was used in the contrastive objective function, $\mathrm{sim}(\boldsymbol{u}, \boldsymbol{v}) = \frac{\boldsymbol{u} \cdot \boldsymbol{v}}{\|\boldsymbol{u}\|\|\boldsymbol{v}\|}$, to enforce that all feature vectors are on the unit hypersphere. We developed a novel framework for supervised contrastive learning to accommodate multilabel classification tasks. Each label is assigned a unique projection layer $g_l(\cdot)$ for computing a label-wise supervised contrastive objective. The final weighted multi-label supervised contrastive loss is:

$$\mathscr{L}^{sup}_{multi-label} = \sum_{l \in L} \lambda_l \sum_{i \in I} \mathscr{L}_i^{sup}(i, g_l(\cdot), P_l(i)) \quad (2)$$

where λ_l is the label weight coefficient. The PyTorch-style pseudocode for implementation can be found in Extended Data Fig. 3. All models were trained for 50 epochs using the Adam optimizer, a cosine annealing learning rate scheduler, and a temperature of 0.07. The batch size was 256. Data augmentation included random cropping, gaussian blur, flipping, and random erasing. Following training, all projection layers are discarded and the visual encoder $f(\cdot)$ is retained for multi-label classification training. We call the above visual representation learning strategy patchcon for weakly supervised (e.g. patient labels only), patch-based contrastive representation learning and results can be found in Extended Data Fig. 4.

2.6 Diffuse glioma genetic embedding

A major component of our multimodal training method includes public genomic data from adult diffuse glioma patients to pretrain a genetic embedding model. We aggregated genomic data from The Cancer Genome Atlas (TCGA), Chinese Glioma Genome Atlas (CGGA) [47], International Cancer Genome Consortium (ICGC) [48], Rembrandt brain cancer dataset [49], Memorial Sloan Kettering (MSK) Data Catalog [50], and the Mayo Glioblastoma Xenograft National Resource. A total of 2777 patients with diffuse gliomas

were aggregated and used for embedding model training. The data used to train our genetic embedding model can be found in Extended Data Table 2. Briefly, we selected common recurrent somatic mutations found in adult-type diffuse gliomas and encoded those mutations as either mutant or wildtype for each patient. Inspired by previous work on word embeddings [28], we used a global vector (GloVe) embedding loss function that minimizes the mean squared difference between the pairwise inner product of the learned gene embedding vectors and the co-occurrence of the genes mutational status.

$$\mathscr{L}_{embed} = \sum_{i,j} f(X_{i,j})\left(e_i^{\top} e_j - log X_{i,j}\right)^2 \tag{3}$$

$X \in \mathbb{R}^{2n \times 2n}$ is the pairwise gene co-occurrence matrix for our dataset, where $X_{i,j}$ is the number of times the mutational status of the i-th and j-th genes co-occurred in the same tumor. n is the number of genes. The vectors e_i and e_j are updated to match the gene co-occurrence in our dataset. $f(\cdot)$ is a weighting function as previously described to avoid overweighting the most common co-occurrence pairs [28]. We found that global vector embeddings perform better than Gene2Vec embedding models [51]. The embedding model is trained for 10K epochs with a batch size of 60. The Adam optimizer was used with a learning rate of 5E-5.

2.7 Multi-label molecular classification

Two multi-label molecular classification strategies were tested, a linear binary relevance strategy and a transformer-based strategy. Linear binary relevance involves splitting a multi-label classification task into multiple independent binary classifiers. The advantage of using a transformer-based strategy for multi-label classification is the ability to explicitly model complex label dependencies and the co-occurrence of specific genetic mutations in the context of pretrained visual features using an attention mechanism. Similar to bidirectional masked language modelling in BERT-style pretraining [30], we randomly mask a subset of the genetic mutations from the input and the objective is to predict the unknown or masked genes. Masked label training allows for more semantically informative supervision during model training that can improve multi-label classification performance.

Linear binary relevance strategy: Following the training of our visual encoder $f(\cdot)$ using supervised contrastive learning, the weights are fixed and a multilayer perceptron (MLP) that contains a single linear layer is added and trained for multi-label classification.

$$\hat{y}_l = MLP_l(f(x)) = \sigma((\mathbf{w}_l \cdot f(x)) + b_l) \tag{4}$$

where σ is a sigmoid activation function that outputs the probability for the l^{th} genetic mutation. This layer is trained using a weighted binary cross-entropy loss.

$$\mathscr{L}(y, \hat{y}) = \sum_{l=1}^{|L|} \lambda_l [y_l \log(\hat{y}_l) + (1 - y_l)\log(1 - \hat{y}_l)) \tag{5}$$

Transformer-based strategy: A transformer encoder is used that includes our pretrained genetic embedding layer W_l. The labels $[l_j, \ldots, l_k]$ are embedded such that $e_k = W_l \cdot l_k$ where the k^{th} column of W_l is the label embedding for the k^{th} label. A label mask is then sampled that randomly selects a subset of labels for transformer input and the remainder to be predicted as output. We used learnable state embeddings to encode whether a label was positive, negative, or unknown/masked (not included to simplify notation) [31]. The image feature vector z and embedded genetic labels are concatenated and input into the transformer encoder:

$$H = [h_j, \ldots, h_k] = \text{MultiHeadAttention}([z, e_j, \ldots, e_k]) \tag{6}$$

where $H = [h_j, \ldots, h_k]$ are the output representations of the genetic labels and the image token removed. Rather than using a position-wise linear feedforward network and/or a [class] token for label classification as is does in conventional transformer architectures [29, 31, 52], we enforce that the output latent space of the transformer encoder is the same as the pretrained genetic embedding space such that:

$$\hat{\boldsymbol{y}} = \sigma\left(\text{diagonal}\left(HW_l^\top\right)\right) \tag{7}$$

where $HW_l^\top$ is in $\mathbb{R}^{l \times l}$ matrix and the diagonal elements are the inner product between transformer output latents and the corresponding label embedding of the same label index. The transformer encoder model is trained using the same weighted binary cross-entropy loss function as above. The embedding layer weights are fixed during the transformer encoder training. The PyTorch-style pseudocode for implementation can be found in Extended Data Fig. 6.

2.8 Whole slide segmentation, patient inference, and molecular subgrouping

Patch-based image classification requires in inference function to aggregate patch-level predictions into a single whole slide-level or patient-level diagnosis. To accomplish this, whole slide SRH images are patched and each patch undergoes an initial feedforward pass through our previously trained segmentation model, f_ϕ, that classifies each patch into tumor, normal brain, or nondiagnostic tissue using an argmax operation [7]. If less than 10% of the image area is classified as tumor, the whole slide is excluded from inference for molecular classification. Our DeepGlioma model, g_θ, predicts on the tumor patches only. The patch-level model outputs are summed using soft probability density aggregation and each label is then renormalized to give a valid Bernoulli distribution for each label. For patients with multiple whole slide images, all patch-level predictions are aggregated and a single patient-level diagnosis is returned. The molecular genetic patient inference function is:

$$p^{patient}(y \mid \mathcal{X}) = \frac{1}{Z} \sum_{j=1}^{|\mathcal{X}|} \mathbb{1}(\text{argmax } p(y_j \mid x_j, \phi) = k_{\text{tumor}}) p(y_j \mid x_j, \theta) \tag{8}$$

where $\mathcal{X}$ is a set of patches from a patient, x_j is the j^{th} patch, $p(y_j \mid x_j, \phi)$ is the patch output from the tumor segmentation model f_ϕ, $p(y_j \mid x_j, \theta)$ is the DeepGlioma g_θ output, and $Z = \sum_{j=1}^{|\mathcal{X}|} \mathbb{1}(\text{argmax } p(y_j \mid x_j, \phi) = k_{\text{tumor}}$ is the number of patches classified as tumor.

Mutually-exclusive molecular subgroup prediction is achieved algorithmically from the above patient-level molecular genetic predictions $p^{patient}(y \mid \mathcal{X})$ as shown in Algorithm 1.

Algorithm 1

DeepGlioma patient-level molecular subgroup prediction

Require: $p(y \mid \mathcal{X}), \tau, \psi, \epsilon$ ▷ $\tau = 0.5, \psi = 1$ for DeepGlioma experiments

$\hat{y}_{mol} \Leftarrow \varnothing$

if $p(y = k_{IDH} \mid \mathcal{X}) < \tau$ **then**

  $\hat{y}_{mol} \Leftarrow$ 'Glioblastoma, IDH-wildtype'

else if $p(y = k_{IDH} \mid \mathcal{X}) \geq \tau$ & $\frac{p(y = k_{1p19q} \mid \mathcal{X})}{p(y = k_{ATRX} \mid \mathcal{X}) + \epsilon} > \psi$ **then**

  $\hat{y}_{mol} \Leftarrow$ 'Oligodendroglioma, IDH-mutant, and 1p19q-codeleted'

else if $p(y = k_{IDH} \mid \mathcal{X}) \geq \tau$ & $\frac{p(y = k_{1p19q} \mid \mathcal{X})}{p(y = k_{ATRX} \mid \mathcal{X}) + \epsilon} \leq \psi$ **then**

  $\hat{y}_{mol} \Leftarrow$ 'Astrocytoma, IDH-mutant'

end if

return $\hat{y}_{mol}$

2.9 Ablation studies

We conducted three main ablation experiments to test the importance of major training strategies and model architectural design choices: (1) cross-entropy versus contrastive learning for visual representation learning, (2) linear versus transformer-based multilabel classification, and (3) fully-supervised versus masked label training. Using the UM dataset only, we performed hold-out validation on three randomly sampled validation sets (n=20 patients/set) that contained a balanced number of IDH mutant (n=10) and wildtype (n=10) tumors. Results are shown in Extended Data Fig. 7. For (1), we trained a ResNet50 model using conventional cross-entropy versus a weakly supervised patch-based contrastive learning, or patchcon. Both models were initialized using ImageNet pretrained weights[53] and trained for 10 epochs without additional hyperparameter tuning. For (2), the patchcon pretrained ResNet model from (1) was held fixed and we trained a single linear classification layer versus a transformer model with 3 multi-headed attention layers. Each model was trained for 10 epochs. For (3), only the transformer model was retrained using variable percentages of labels masked. We tested 0%, 33%, and 66% of labels provided as input, which corresponded to 0, 1, and 2 labels provided for our dataset. Each model was trained using the same contrastive pre-trained ResNet SRH encoder to isolate the effect of label masked training on classifier performance. Results of ablation studies can be found in Extended Data Fig. 7.

2.10 Molecular heatmap generation

Leveraging our previous work on semantic segmentation of SRH images [7, 54], we densely sample patches at 100 pixel step size, which allows for local probability pooling from overlapping patch predictions. A major contribution of this work is the integration of our tumor segmentation model and DeepGlioma into a single interpretable heatmap for

both molecular genetic and molecular subgroup predictions. The tumor segmented regions are retained and the normal/nondiagnostic regions are converted to grayscale in order to indicate these regions were not candidates for molecular prediction. Each molecular genetic heatmap is generated by averaging the output predictions from patches that overlap for any given pixel in the heatmap. Molecular subgroup heatmaps are more challenging and require integrating the molecular genetic predictions that are necessary for subgroup classification. To address this challenge, we use a molecular subgroup-specific conditional mask combined with IDH predictions to generate an interpretable and spatially consistent molecular subgroup heatmap. The most straightforward molecular subgroup heatmap is for glioblastoma, IDH-wildtype heatmap, generated as:

$$p_{i,j}^{GBM} = 1 - p_{IDH}(x_{i,j}) \tag{9}$$

such that i, j corresponds to the whole slide height and width indices and $p_{IDH}(x_{i,j})$ is the IDH prediction at the corresponding spatial location. In contrast, molecular oligodendrogliomas and astrocytomas require a conditional molecular mask to segment regions that meet specific molecular subgroup criteria. Molecular oligodendroglioma heatmaps are generated as:

$$p_{i,j}^{Oligo.} = \underbrace{[p_{IDH}(x_{i,j}) > \tau \wedge p_{1p19q}(x_{i,j}) > \phi]}_{\text{Conditional molecular mask}} * p_{IDH}(x_{i,j}) \tag{10}$$

with the binarized conditional molecular mask identifying heatmap regions that are above hyperparameter thresholds τ and ϕ for IDH and 1p19q codeletion, respectively. Molecular astrocytomas heatmaps are generated as:

$$p_{i,j}^{Astro.} = \underbrace{[p_{IDH}(x_{i,j}) > \tau \wedge [p_{1p19q}(x_{i,j}) < \phi \vee p_{ATRX}(x_{i,j}) > \pi]]}_{\text{Conditional molecular mask}} * p_{IDH}(x_{i,j}) \tag{11}$$

where τ, ϕ, π are all hyperparameter thresholds. Conditional molecular masking encodes the spatial locations where the molecular subgroup conditions are instantiated and the IDH prediction provides the representative probability distribution for the molecular subgroup. Examples of molecular genetic and molecular subgroup heatmaps can be found in Extended Data Fig. 9 and 10. Interactive web-based interface for DeepGlioma predictions can be found at deepglioma.mlins.org.

2.11 Prospective multicenter clinical testing and sample size calculation

We elected to perform prospective, international, multicenter clinical testing of DeepGlioma in order to adhere to the rigorous standards of responsible machine learning in healthcare [55]. Our prospective clinical testing was designed using the same principles as a non-inferiority diagnostic clinical trial [19]. NYU, UCSF, MUV, and UKK were all included as medical centers for prospective patient enrollment.

Primary testing endpoint—Our primary diagnostic endpoint was balanced classification accuracy $\left(\frac{\text{sensitivity+specificity}}{2}\right)$ for diffuse glioma IDH mutational status. The control arm was conventional first-line laboratory molecular screening testing and the experimental arm was DeepGlioma predictions. IDH-1 immunohistochemistry (IHC) for somatic mutations at

residue R132H is the most common first-line molecular diagnostic screening test. Dewitt et al. performed the largest and most clinically representative analysis of IDH mutation detection via IHC and sequencing methods, and determining that IDH1-R132H IHC has balanced diagnostic accuracy of 91.4% for adult-type diffuse gliomas (see Fig. 2c for contingency tables) [33]. We used this value to set the expected accuracy for both the control and experimental arms, the equivalence limit was set to 10%, power to 90%, and alpha to 0.05%, resulting in sample size value of 135 patients. All sample size calculations were performed using the epiR package (version 2.0.46) in R (version 3.6.3).

Secondary testing endpoint—Our secondary endpoint was to achieve improved classification performance compared to our previous methods for training deep computer vision models on SRH images for multiclass classification [7, 19]. End-to-end representation learning and classification can yield patch-based classification results that approach pathologist-level performance for histologic brain tumor classifcation. However, our early experiments on molecular classification indicated that contrastive pretraining and label embedding was advantageous for multilabel classification. Therefore, as a secondary endpoint, the control arm was established by training a ResNet50 model to classify the three mutually exclusive molecular subgroups using a conventional categorical cross-entropy loss function. This is equivalent to our previous model training method with the exception of different labels [7, 19]. Our experimental arm was DeepGlioma molecular subgroup predictions. Secondary endpoint metric was overall multiclass classification accuracy (Fig. 2d).

2.12 Computational hardware and software

All SRH images were processed using an Intel Core i76700K Skylake QuadCore 4.0 central processing unit (CPU) using our custom Python-based (version 3.8) mlins-package. We used the pydicom package (version 2.0.0) to process the SRH images from the NIO Imaging System. All archived postprocessed image patches were saved as 16-bit TIFF images and handled using the tifffile package (version 2021.1.14). All models were trained using the University of Michigan Advanced Research Computing (ARC) Armis2 high-performance computing cluster. Armis2 is a high-performance distributed computing environment that aligns with HIPAA privacy standards. Convolutional neural networks/visual encoders were trained on four NVIDIA Titan V100 graphical processing units (GPUs). Our genetic embedding model and classifiers were trained on eight NVIDIA 2080Ti GPUs. All custom code for training and inference can be found in our open-source DeepGlioma repository. Our models were implemented in PyTorch (version 1.9.0). We used the ImageNet pretrained ResNet50 model from torchvision (0.10.0). Scikit-learn (version 1.0.1) was used to compute performance metrics on model predictions at both training and inference. Additional dependencies and specifications can be found in our DeepGlioma package.

2.13 Data availability

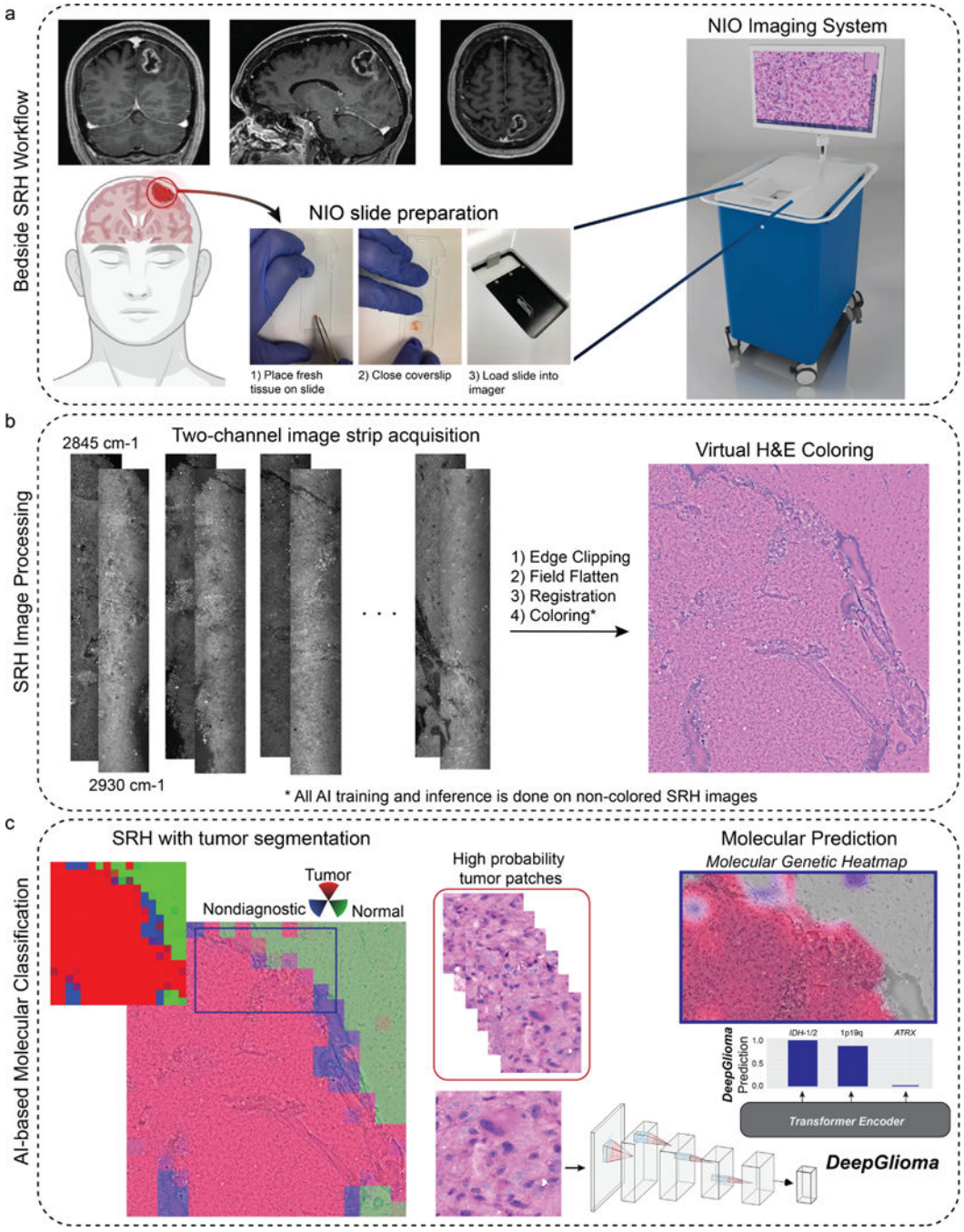
The genomic data for training the genetic embedding model are publicly available through the above mentioned data repositories and all genomic data used is provided in Extended Data Table 2. Institutional Review Board approval was obtained from all participating institutions for SRH imaging and data collection. Restrictions apply to the availability of the

raw patient imaging or genetic data, which were used with institutional permission through IRB approval for the current study, and are thus not publicly available. Please contact the corresponding authors (T.C.H. or D.A.O.) for any requests for data sharing. All requests will be evaluated based on institutional and departmental policies to determine whether the data requested is subject to intellectual property or patient privacy obligations. Data can only be shared for non-commercial academic purposes and will require a formal material transfer agreement.

2.14 Code availability

All code was implemented in Python using PyTorch as the primary deep learning package. All code and scripts to reproduce the experiments of this paper are available at GITHUB LINK HERE.

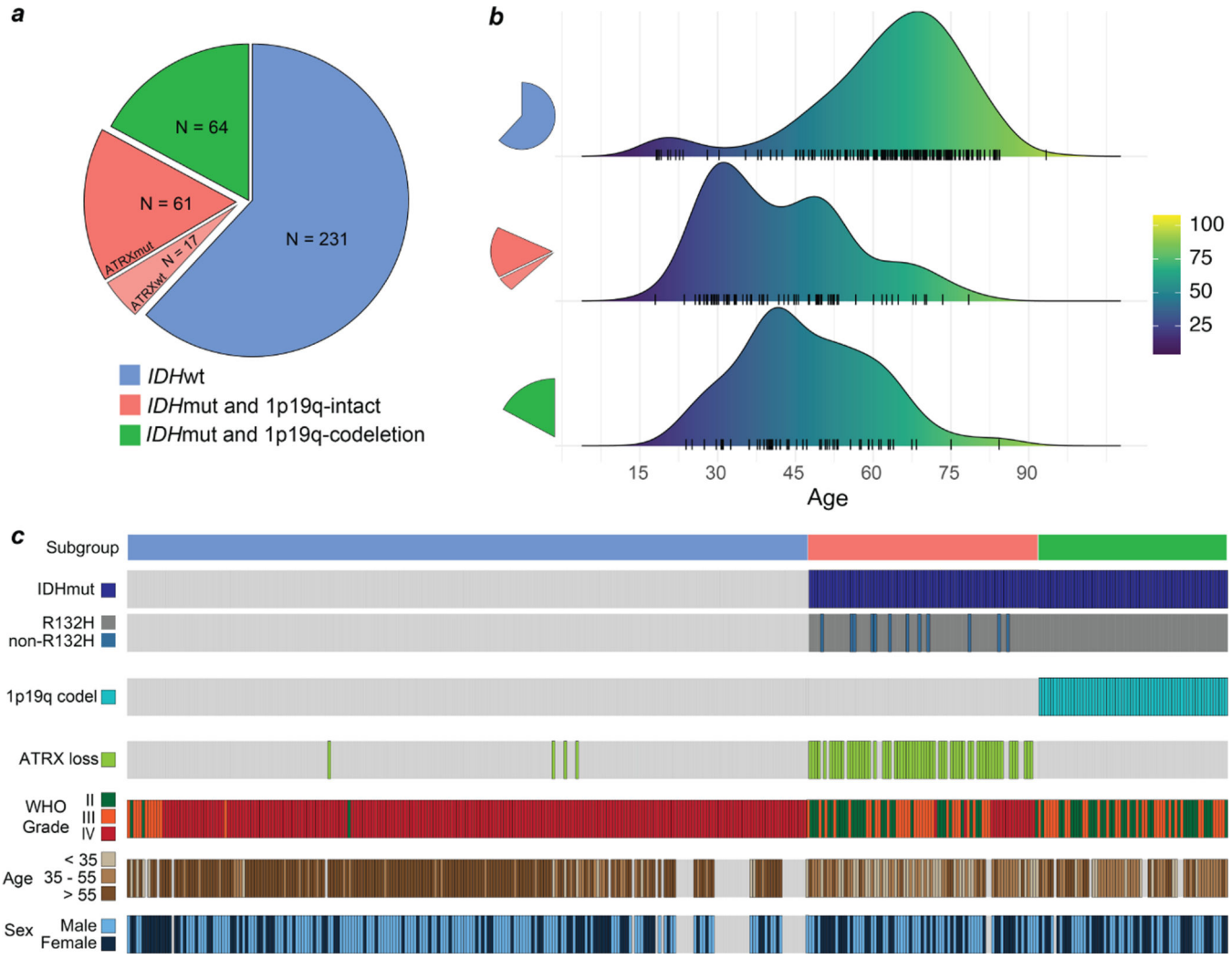
3 Extended Data



Extended Data Fig 1. Overall workflow of intraoperative SRH and DeepGlioma

a, DeepGlioma for molecular prediction is intended for adult (≥ 18 years) patients with clinical and radiographic evidence of a diffuse glioma who are undergoing surgery for tissue diagnosis and/or tumor resection. The surgical specimen is sampled from the patient's tumor and directly loaded into a premade, disposable microscope slide with attached coverslip. The specimen is loaded into the NIO Imaging System (Invenio Imaging, Inc., Santa Clara,

CA) for rapid optical imaging. **b**, SRH images are acquired sequentially as strips at two Raman shifts, 2845 cm^{-1} and 2930 cm^{-1}. The size and number of strips to be acquired is set by the operator who defines the desired image size. Standard images sizes range from 1–5mm^2 and image acquisition time ranges from 30 seconds to 3 minutes. The strips are edge clipped, field flattened, and registered to generate whole slide SRH images. These whole slide images are then used for both DeepGlioma training and inference. Additionally, whole slide images can be colored using a custom virtual H&E colorscheme for review by the surgeon or pathologist [6]. **c**, For AI-based molecular diagnosis, the whole slide image is split into non-overlapping 300×300 pixel patches and each patch undergoes a feedforward pass through a previously trained network to segment the patches into tumor regions, normal brain, and nondiagnostic regions [19]. The tumor patches are then used by DeepGlioma at both training and inference to predict the molecular status of the patient's tumor.

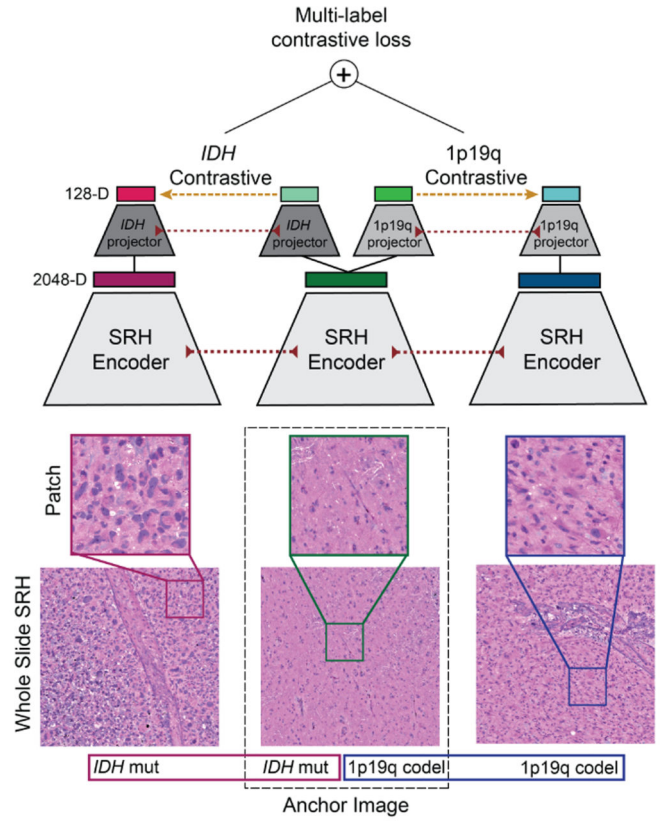


Extended Data Fig 2. Training dataset

The UM adult-type diffuse gliomas dataset used for model training. The UM training set consisted of a total of 373 patients who underwent a biopsy or brain tumor resection. Dataset generation occurred over a six-year period, starting in November 2015 and ended

in November 2021. **a**, The distribution of patients by molecular subgroup is shown. IDH-wildtype gliomas consisted of 61.9% (231/373) of the total dataset. IDH-mutant/1p19q-codeleted tumors consisted of 17.2% (64/373) and IDH-mutant/1p19q-intact tumors consisted of 21.% (78/373) of tumors in the dataset. Our dataset distribution of molecular subgroups is consistent with reported distributions in large-scale population studies [56]. ATRX mutations were found in the majority of IDH-mutant/1p19q-intact patients (78%), also concordant with previous studies [10]. **b**, The age distribution for each of the molecular subgroups is shown. The average age of IDH-wildtype patients was 62.6 ± 15.4 and IDH-mutant patients was 44.6 ± 13.8. The average patient age of IDH-mutant/1p19q-codel group was 47.0 ± 12.9 and IDH-mutant/1p19-intact was 42.5 ± 14.1 years. **c**, Individualized patient characteristics and mutational status are shown by molecular subgroups. We report the WHO grade based on pathologic interpretation at the time of diagnosis. Because many of the patients were treated prior to the routine use of molecular status alone to determine WHO grade, several patients have IDH-wildtype lower grade gliomas (grade II or III) or IDH-mutant glioblastomas (grade IV). The discordance between histologic features and molecular features has been well documented[10] and is a major motivation for the present study.

a



b

Algorithm 1: PyTorch-style pseudocode for general multi-label contrastive learning

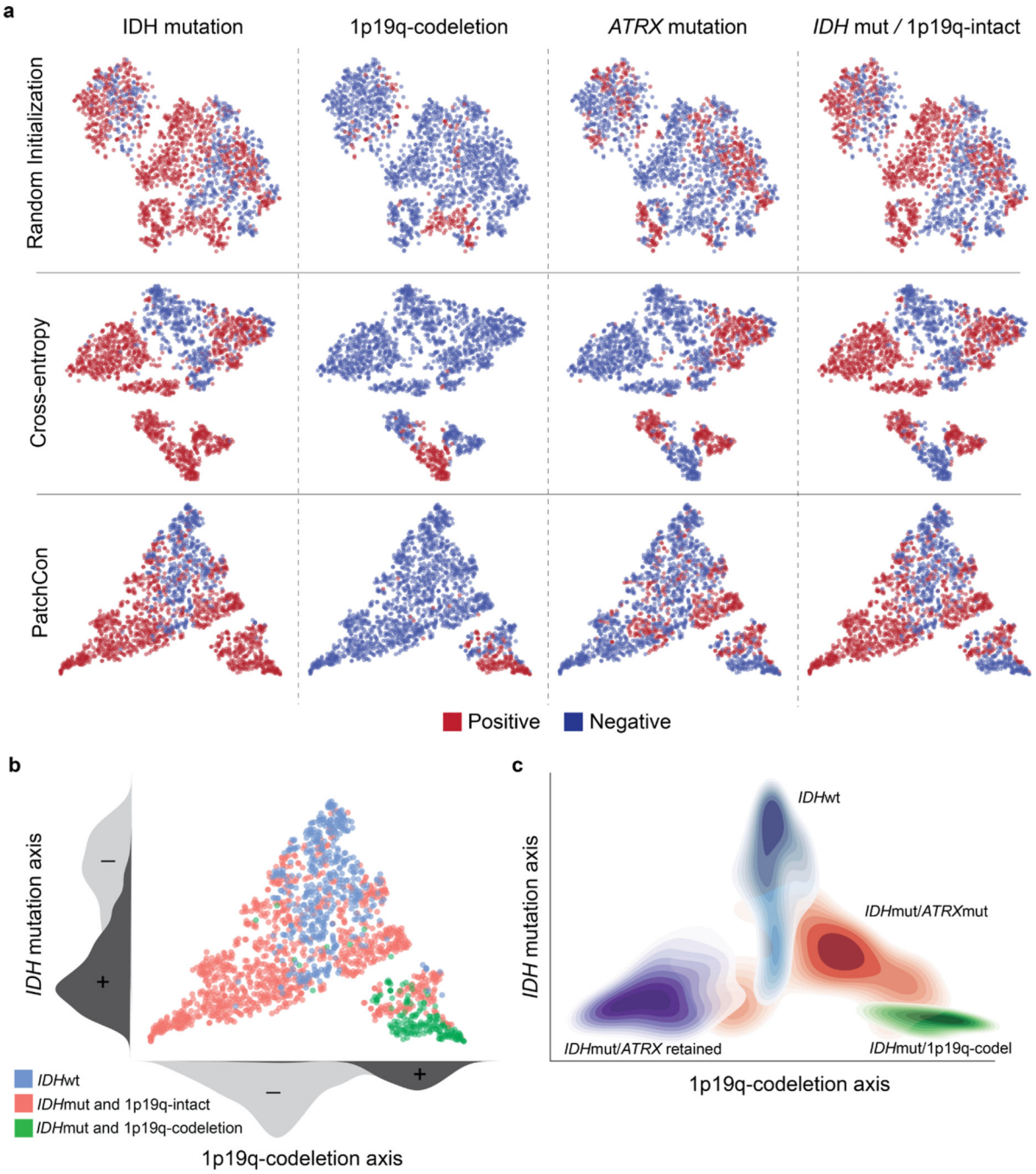
```
# f : vision encoder, F dimensional output
# p : linear projection layer, P dimensional output
# λ : label weight vector
# nlabels : total number of labels in dataset
# t : temperature

# initialize a projector for each label
for i in nlabels:
   projectors[i] = p.params

# load a batch of N samples
for x,y in loader:
   x = augment(x)  # augment N batch images
   z = f(x)  # encoder forward pass: NxF
   z = z.norm(dim = 1)  # normalize representations

   # compute projections for each label
   losses = zeros(nlabels)
   for i, p in projectors.items():
      logits = p(z) # projector forward pass: NxP
      logits = logits.norm(dim = 1) # normalize logits
      losses[i] = λ[i] * SupCon(logits, y[:,i], t) # compute
          label weighted contrastive loss

   # compute loss
   loss = losses.sum()
   loss.backward()

   # optimization step
   f.step()
   projectors.step()
```

SupCon, supervised contrastive loss

Extended Data Fig 3. Multi-label contrastive learning for visual representations

Contrastive learning for visual representation is an active area of research in computer vision [22, 57, 58]. While the majority of research has focused on self-supervised

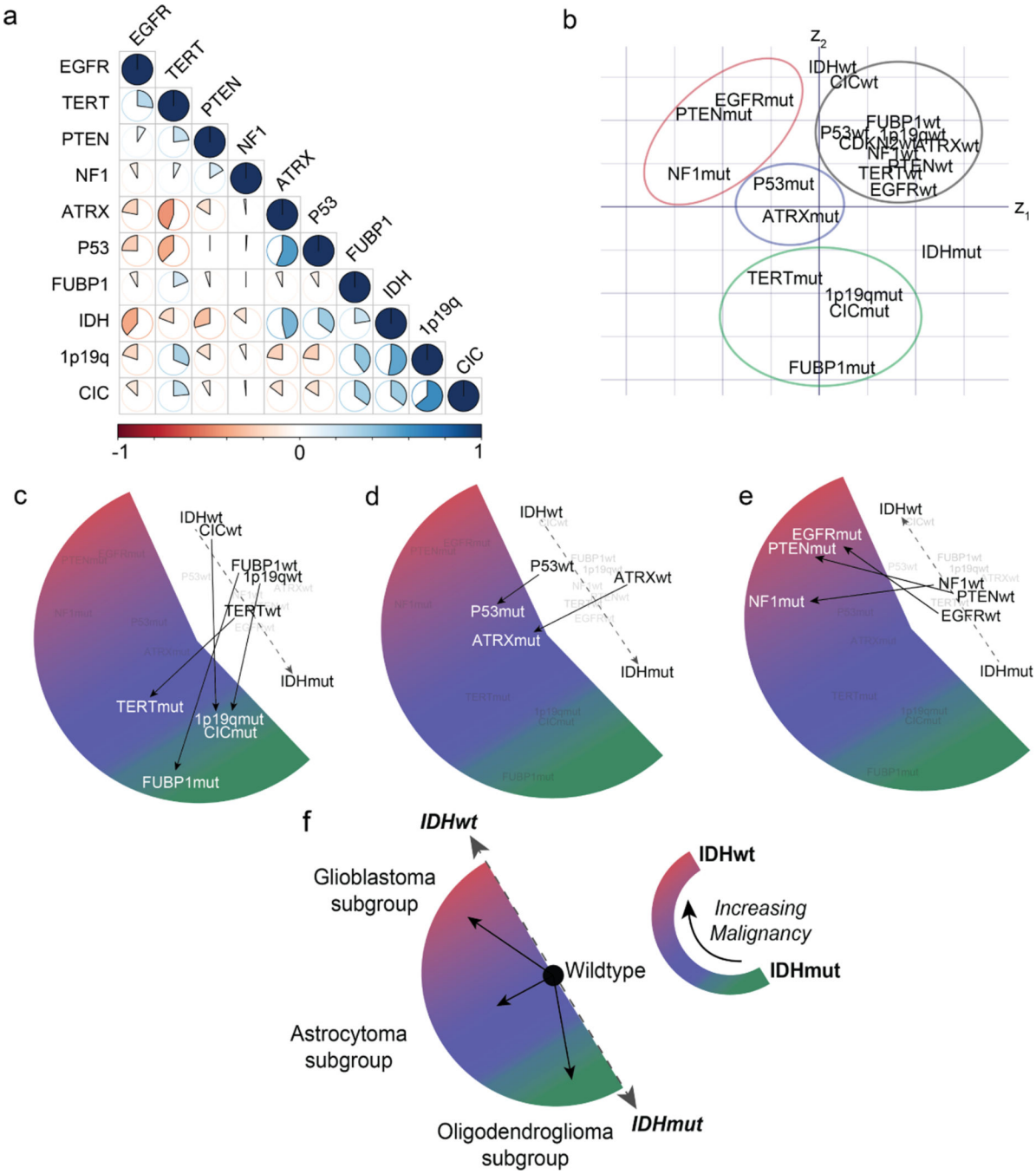
learning, supervised contrastive loss functions have been underexplored and provide several advantages over supervised cross-entropy losses [58, 59]. Unfortunately, no straightforward extension of existing contrastive loss functions, such as InfoNCE[60] and NT-Xent[61], can accommodate multi-label supervision. Here, we propose a simple and general extension of supervised contrastive learning for multi-label tasks and present the method in the context of patch-based image classification. **a**, Our multi-label contrastive learning framework starts with a randomly sampled anchor image with an associated set of labels. Within each minibatch a set of positive examples are defined for each label of the anchor image that shares the same label status. All images in the minibatch undergo a feedforward pass through the SRH encoder (red dotted lines indicate weight sharing). Each image representation vector (2048-D) is then passed through multiple label projectors (128-D) in order to compute a contrastive loss for each label (yellow dashed line). The scalar label-level contrastive loss is then summed and backpropagated through the projectors and image encoder. The multi-label contrastive loss is computed for all examples in each minibatch. **b**, PyTorch-style pseudocode for implementation of our proposed multi-label contrastive learning framework is shown. Note that this framework is general and can be applied to any multi-label classification task. We call our implementation *patchcon* because individual image patches are sampled from whole slide SRH images to compute the contrastive loss. Because we use a single projection layer for each label and the same image encoder is used for all images, the computational complexity is linear in the number of labels.



Extended Data Fig 4. SRH visual representation learning comparison

a, SRH patch representations of a held-out validation set are plotted. Patch representations from a ResNet50 encoder randomly initialized (top row), trained with cross-entropy (middle row), and PatchCon (bottom row) are shown. Each column shows binary labels for the listed molecular diagnostic mutation or subgroup. A randomly initialized encoder shows evidence of clustering because patches sampled from the same patient are correlated and can have similar image features. Training with a cross-entropy loss does enforce separability between some of the labels; however, there is no discernible lowdimensional

manifold that disentangles the label information. Our proposed multi-label contrastive loss produced embeddings that are more uniformly distributed in representation space than cross-entropy. Uniformity of the learned embedding distribution is known to be a desirable feature of contrastive representation learning [32]. **b**, Qualitative analysis of the SRH patch embeddings indicates that that data is distributed along two major axes that correspond to IDH mutational status and 1p19q-codeletion status. This distribution produces a simplex with the three major molecular subgroups at each of the vertices. These qualitative results are reproduced in our prospective testing cohort show in Figure 2e. **c**, The contour density plots for each of the major molecular subgroups are shown to summarize the overall embedding structure. IDH-wildtype images cluster at the apex and IDH-mutant tumors cluster at the base. Patients with 1p19q-intact are closer to the origin and 1p19q-codeleted tumors are further from the origin.



Extended Data Fig 5. Diffuse glioma genetic embedding using global vectors

Embedding models transform discrete variables, such as words or gene mutational status, into continuous representations that populate a vector space such that location, direction, and distance are semantically meaningful. Our genetic embedding model was trained using data sourced from multiple public repositories of sequenced diffuse gliomas (Extended Data Table 2). We used a global vector embedding objective for training [28]. **a**, A subset of the most common mutations in diffuse gliomas is shown in the co-occurrence matrix. Data was collected from multiple public repositories and aggregated to generate a single co-

occurrence matrix for global vector embedding training. **b**, The learned genetic embedding vector space with the 11 most commonly mutated genes shown. Both the mutant and wildtype mutational statuses (N=22) are included during training to encode the presence or absence of a mutation. Genes that co-occur in specific molecular subgroups cluster together within the vector space, such as mutations that occur in (**c**) IDH-mutant, 1p19q-codel oligodendrogliomas (green), (**d**) IDH-mutant, ATRX-mutant diffuse astrocytomas (blue), and (**e**) IDH-wildtype glioblastoma subtypes (red). Additionally, wildtype genes (black) form a single cluster with gene mutations organized in a radial pattern. Radial traversal of the embedding space defines clinically meaningful linear substructures [28] corresponding to molecular subgroups. **f**, Corresponding to the known clinical and prognostic significance of IDH mutations in diffuse gliomas, IDH mutational status determines the axis along which increasing malignancy is defined in our genetic embedding space.

Algorithm 1: PyTorch-style pseudocode for transformer-based multi-label classification

```
# f : vision encoder, F dimensional output
# e : genetic encoder, F dimensional output
# t : transformer, L dimensional output
# mask : label mask,[0,1]^L
# λ : label weight vector

# load a batch of N samples
for x, y, mask in loader:
    x = augment(x)  # augment N batch images
    h = f(x) # encoder forward pass: NxF
    h = h.norm(dim = 1)  # normalize representations

   l_k = mask * y    # mask label subset
   l_e = e(l_k)   # embed labels
   embed = cat(h, l_e)   # concatenate image and label
     embeddings

    outputs = t(embed)  # transformer forward pass: NxL
    x = bmm(outputs[:, 1:], e.weights.T.detach()) # batch
     matrix multiplication on label outputs and
     embedding layer weights: NxLxL
    logits = diagonal(x, dim1=-2, dim2=-1) # select the
     main diagonal logits: NxL

    # compute loss
    loss = λ * BCE(logits, y) * 1-mask
    loss.sum().backward()

    # optimization step
    f.step()# optional, f pretrained in our implementation
    e.step()# optional, e pretrained in our implementation
    t.step()
```

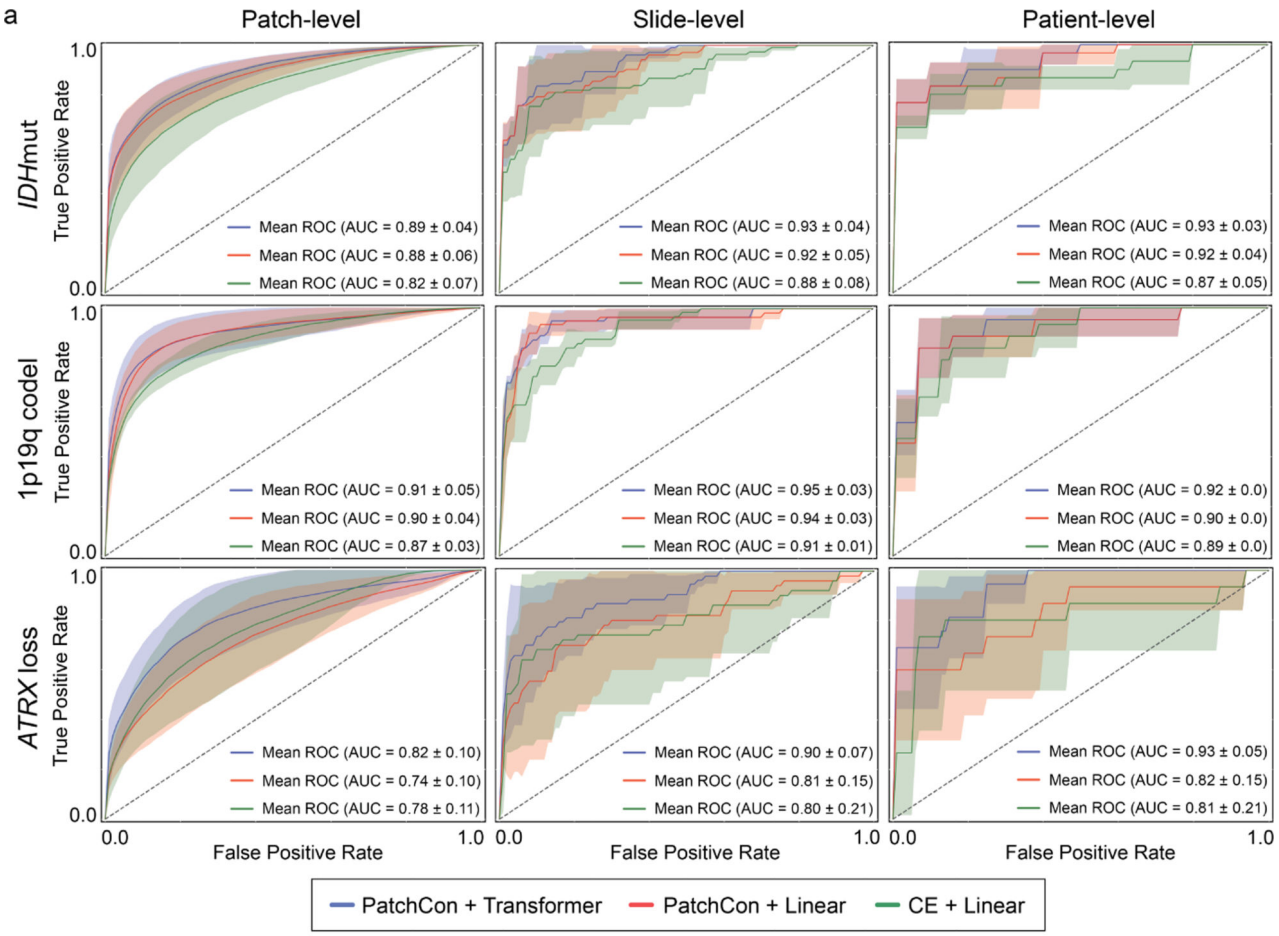
bmm, batch matrix multiplication; BCE, binary cross-entropy loss

Extended Data Fig 6. PyTorch-style pseudocode for transformer-based masked multi-label classification

Inputs to our masked multi-label classification algorithm are listed in lines 1–5. The vision encoder and genetic encoder are pretrained in our implementation but can be randomly initialized and trained end-to-end. The label mask is an L-dimensional binary mask with a variable percentage of the labels removed and subsequently predicted in each feedforward pass. An image x is augmented and undergoes a feedforward pass through the vision encoder f. The image representation is then ℓ^2 normalized. The labels are

embedded using our pretrained genetic embedding model and the label mask is applied. The label embeddings are then concatenated with the image embedding and passed into the transformer encoder as input tokens. Unlike previous transformer-based methods for multi-label classification [31], we enforce that the transformer encoder outputs into the same vector space as the pretrained genetic embedding model. We perform a batch matrix multiplication with the transformer outputs and the embedding layer weights. The main diagonal elements are the inner product between the transformer encoder output and the corresponding embedding weight values. We then compute the masked binary cross-entropy loss. In our implementation, this is used to train the transformer encoder model only.

a



b

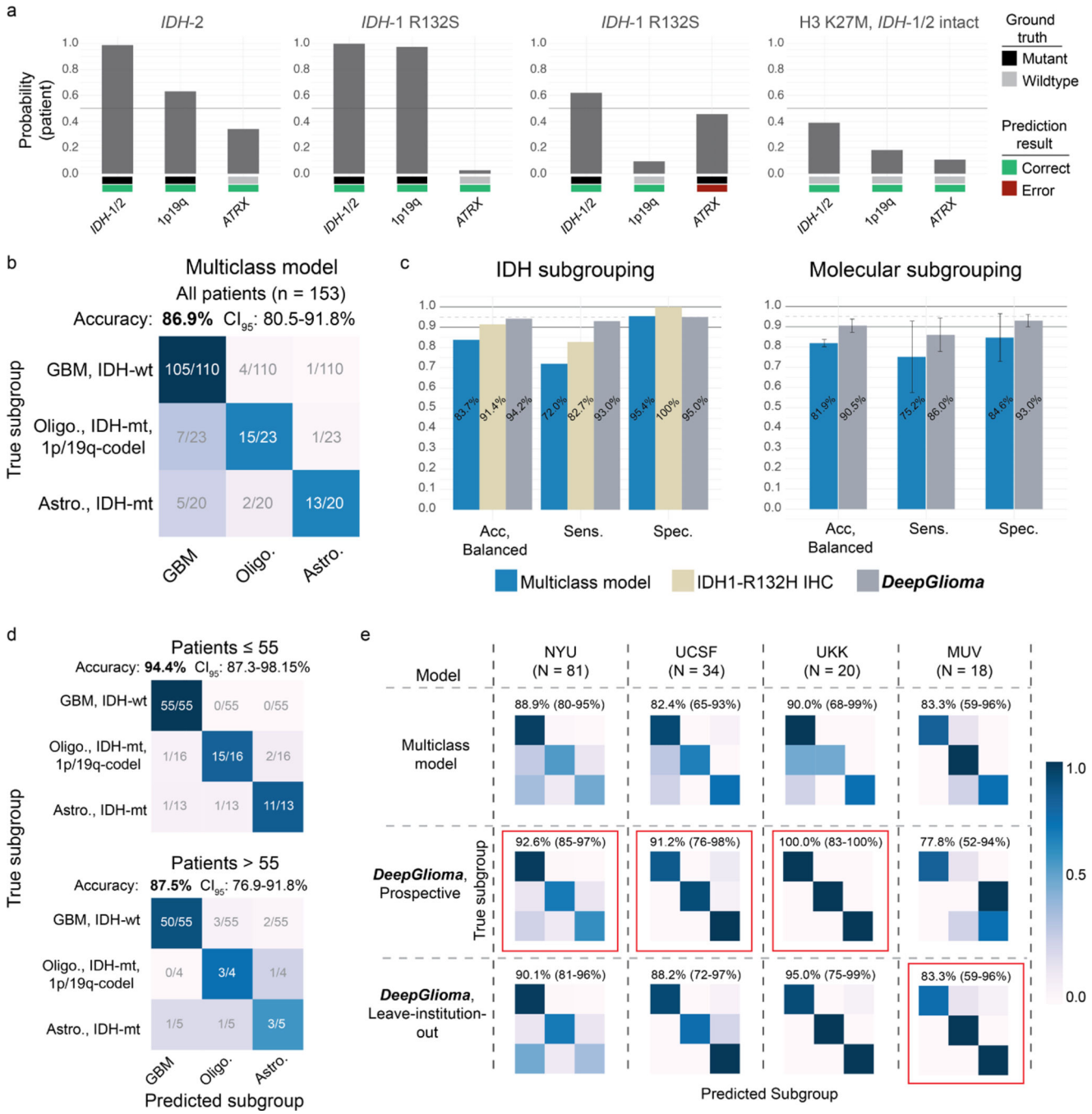
Level	% Input	Multi-label metrics						
		mAP	mAUC	Hamming	Subset Acc	ebF1	micF1	macF1
Patch	0%	68.7 ± 4.7	86.6 ± 9.3	$\mathbf{84.8 \pm 0.2}$	67.6 ± 0.7	55.3 ± 4.0	71.2 ± 1.8	64.5 ± 2.8
	33%	69.7 ± 1.3	$\mathbf{87.4 \pm 7.1}$	$\mathbf{84.8 \pm 0.9}$	$\mathbf{68.2 \pm 1.4}$	$\mathbf{57.4 \pm 2.9}$	$\mathbf{72.3 \pm 1.4}$	65.9 ± 1.9
	66%	$\mathbf{70.0 \pm 2.1}$	86.3 ± 1.5	84.7 ± 0.5	67.8 ± 0.8	54.6 ± 2.8	70.9 ± 1.7	$\mathbf{66.1 \pm 3.5}$
Slide	0%	87.4 ± 1.7	94.6 ± 0.3	91.2 ± 1.6	78.4 ± 4.1	68.9 ± 9.4	81.7 ± 4.6	77.9 ± 5.4
	33%	$\mathbf{91.0 \pm 0.5}$	$\mathbf{95.4 \pm 0.3}$	$\mathbf{91.5 \pm 0.9}$	$\mathbf{81.0 \pm 4.0}$	$\mathbf{71.0 \pm 8.3}$	$\mathbf{82.7 \pm 2.8}$	$\mathbf{80.2 \pm 4.1}$
	66%	90.4 ± 2.7	95.1 ± 0.7	89.9 ± 1.3	77.1 ± 4.0	62.8 ± 6.3	78.4 ± 3.8	77.3 ± 5.1
Patient	0%	92.9 ± 2.3	91.8 ± 1.9	89.4 ± 1.9	76.7 ± 2.9	65.7 ± 8.6	80.2 ± 4.8	73.9 ± 1.8
	33%	$\mathbf{93.4 \pm 3.9}$	$\mathbf{92.6 \pm 5.4}$	$\mathbf{90.0 \pm 0.0}$	$\mathbf{78.3 \pm 2.9}$	$\mathbf{70.7 \pm 6.3}$	$\mathbf{81.9 \pm 1.2}$	$\mathbf{74.4 \pm 1.7}$
	66%	93.2 ± 1.2	92.3 ± 3.5	87.2 ± 2.6	73.3 ± 5.8	61.6 ± 7.0	75.8 ± 4.8	60.4 ± 1.3

Figure 1: mAP, mean average precision (0.5 threshold); mAUC, mean area under ROC curve; Hamming, Hamming score; Subset Acc, subset accuracy; ebF1, example-based F1 score; micF1, micro-F1 score; macF1, macro-F1 score

Extended Data Fig 7. Ablation studies and cross-validation results

We conducted three main ablation studies to evaluate the following model architectural design choices and major training strategies: (1) cross-entropy versus contrastive loss for visual representation learning, (2) linear versus transformer-based multi-label classification, and (3) fully-supervised versus masked label training. **a**, The first two ablation studies are shown in the panel and the details of the cross-validation experiments are explained in the Methods section (see ‘Ablation Studies‘). Firstly, a ResNet50 model was trained using either cross-entropy or PatchCon. The PatchCon trained image encoder was then fixed.

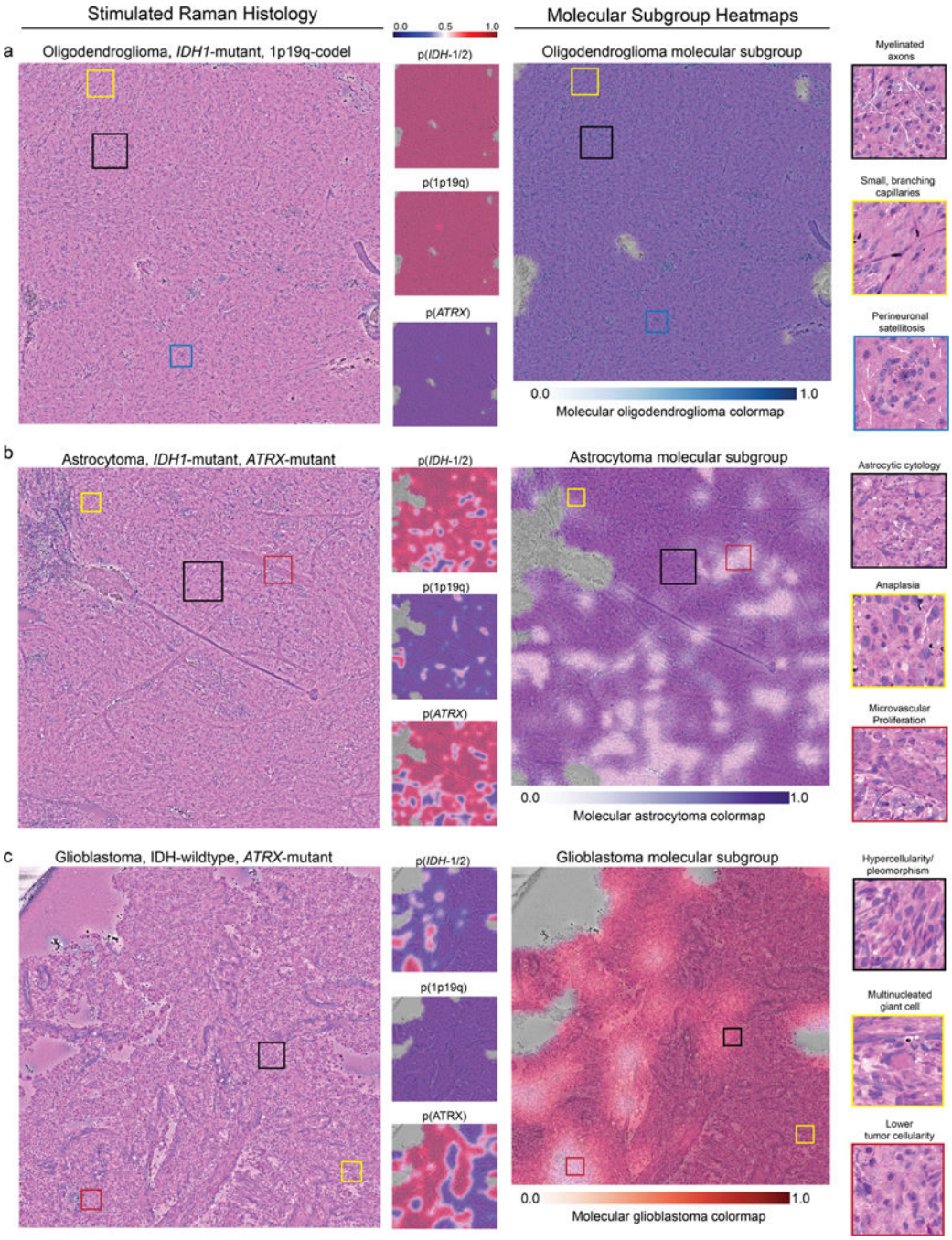
A linear classifier and transformer classifier were then trained using the *same* patchcon image encoder in order to evaluate the performance boost from using a transformer encoder. This ablation study design allows us to evaluate (1) and (2). The columns of the panel correspond to the three levels of prediction for SRH image classification: patch-, slide-, and patient-level. Each model was trained three times on randomly sampled validation sets and the average (± standard deviation) ROC curves are shown for each model. Each row corresponds to the three molecular diagnostic mutations we aimed to predict using our DeepGlioma model. The results show that PatchCon outperforms cross-entropy for visual representation learning and that the transformer classifier outperforms the linear classifier multi-label classification. Note that the boost in performance of the transformer classifier over the linear model is due to the deep multi-headed attention mechanism learning conditional dependencies between labels in the context of specific SRH image features (i.e. not improved image feature learning due to fixed encoder weights). **b**, We then aimed to evaluate (3). Similar to the above, a single ResNet50 model was trained using PatchCon and the encoder weights were fixed for the following ablation study to isolate the contribution of masked label training. Three training regimes were tested and are presented in the table: no masking (0%), 33% masking (one label randomly masked), and 66% (two labels randomly masked). To better investigate the importance of masked label training, we report multiple multi-label classification metrics. We found that 33% masking, or randomly masking one of three diagnostic mutations, showed the best results across all metrics at the slide-level and patient-level. We hypothesize that this results from allowing a single mutation to weakly define the genetic context while allowing supervision from the two masked labels to backpropagate through the transformer encoder.



Extended Data Fig 8. Patient subgroup analysis of DeepGlioma performance

a, Subset of patients from the prospective cohort with non-canonical IDH mutations and a diffuse midline glioma, H3 K27M mutation. DeepGlioma correctly classified all non-canonical IDH mutations, including IDH-2 mutation. Moreover, DeepGlioma generalized to a pediatric-type diffuse high-grade gliomas, including diffuse midline glioma, H3 K27-altered, in a zero-shot fashion as these tumor were not included in the UM training set. This patient was included in our prospective cohort because the patient was a 34 year old adult at presentation. **b**, Confusion matrix of our benchmark multiclass model trained

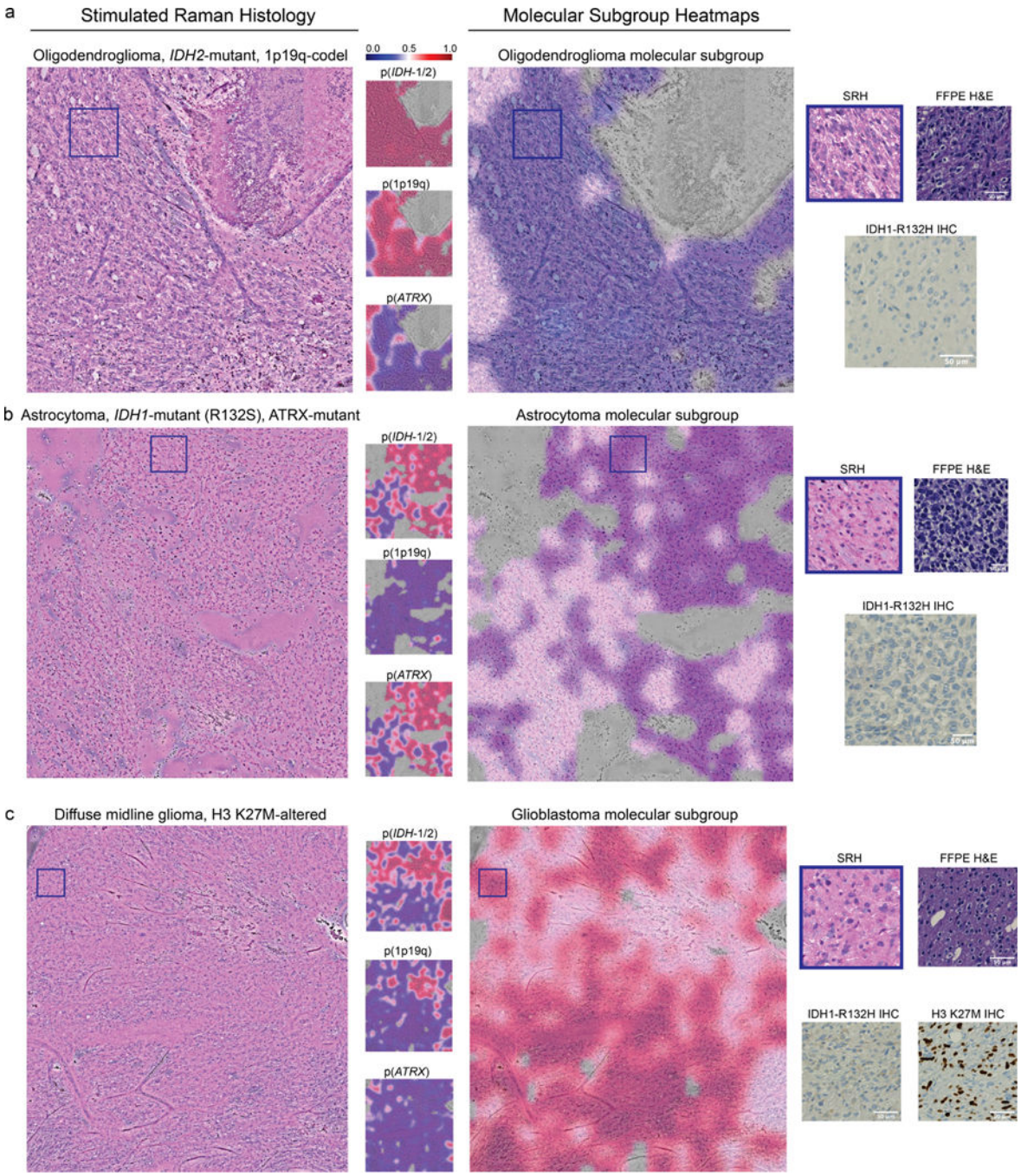
using categorical cross-entropy. DeepGlioma outperformed the multiclass model by +4.6% in overall diagnostic accuracy with a substantial improvement in differentiating molecular astrocytomas and oligodendrogliomas. **c**, Direct comparison of subgrouping performance for our benchmark multiclass model, IDH1-R132H IHC, and DeepGlioma. Performance metrics values are displayed. Molecular subgrouping mean and standard deviations are plotted. **d**, DeepGlioma molecular subgroup classification performance on patients 55 years or younger versus patient older than 55 years. The overall DeepGlioma performance remained high in the ≤ 55 cohort, maintaining a high multiclass accuracy compared the entire cohort. DeepGlioma was trained to generalize to all adult patients. b, DeepGlioma molecular subgroup classification performance for each of the prospective testing medical centers is shown. Accuracy (95% confidence intervals) are shown above the confusion matrices. Overall performance was stable across the three largest contributors of prospective patients. Performance on the MUV dataset was comparatively lower than other centers; however, some improvement was observed during the LIOCV experiments. Red indicates the best performance.



Extended Data Fig 9. Molecular genetic and molecular subgroup heatmaps

DeepGlioma predictions are presented as heatmaps from representative patients included in our prospective clinical testing dataset for each diffuse glioma molecular subgroup. **a**, SRH images from a patient with a molecular oligodendroglioma, IDH-mutant, 1p19q-codel. Uniform high probability prediction for both IDH and 1p19q-codel and corresponding low ATRX mutation prediction. SRH images show classic oligodendroglioma features, including small, branching ‘chicken-wire’ capillaries and perineuronal satellitosis. Oligodendroglioma molecular subgroup heatmap shows expected high prediction probablity throughout the

dense tumor regions. **b**, A molecular astrocytoma, IDH-mutant, 1p19q-intact and ATRX-mutant is shown. Astrocytoma molecular subgroup heatmap shows some regions of lower probability that may be related to the presence of image features found in glioblastoma, such as microvascular proliferation. However, regions of dense hypercellularity and anaplasia are correctly classified as IDH mutant. These findings indicate DeepGlioma's IDH mutational status predictions are not determined solely by conventional cytologic or histomorphologic features that correlate with lower grade versus high grade diffuse gliomas. **c**, A glioblastoma, IDH-wildtype tumor is shown. Glioblastoma molecular subgroup heatmap shows high confidence throughout the tumor specimen. Additionally, this tumor was also ATRX mutated, which is know to occur in IDH-wildtype tumors [10]. Despite the high co-occurence of IDH mutations with ATRX mutations, DeepGlioma was able to identify image features predictive of ATRX mutations in a molecular glioblastoma. Because ATRX mutations are not diagnostic of molecular glioblastomas, the ATRX prediction does not effect the molecular subgroup heatmap (see 'Molecular heatmap generation' section in Methods). Additional SRH images and DeepGlioma prediction heatmaps can be found at our interactive web-based viewer deepglioma.mlins.org.



Extended Data Fig 10. Evaluation of DeepGlioma on non-canonical diffuse gliomas

A major advantage of DeepGlioma over conventional immunohistochemical laboratory techniques is that it is not reliant on specific antigens for effective molecular screening. **a**, A molecular oligodendroglioma with an IDH2 mutation is shown. DeepGlioma correctly predicts the presence of both an IDH mutation and 1p19q-codeletion. IDH1-R132H IHC performed on the imaged specimen is negative. The patient was younger than 55 and, therefore, required genetic sequencing in order to complete full molecular diagnostic testing using our current laboratory methods. **b**, A molecular astrocytoma with an IDH1-R132S

and ATRX mutation. DeepGlioma correctly identifies both mutations. **c**, A patient with a suspected adult-type diffuse gliomas met inclusion criteria for the prospective clinical testing set. Patient was later diagnosed with a diffuse midline glioma, H3 K27-altered. DeepGlioma correctly predicted the patient to be IDH-wildtype without previous training on diffuse midline gliomas or other pediatric-type diffuse gliomas. We hypothesize that DeepGlioma can perform well on other glial neoplasms in a similar zero-shot fashion.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The results presented here are in whole or part based upon data generated by the TCGA Research Network: www.cancer.gov/tcga. We would like to thank Tom Cichonski for providing expert medical editing. We would like to thank Karen Eddy, Lin Wang, and Andrea Marshall for providing technical support.

References

[1]. Yadav H, Shah D, Sayed S, Horton S & Schroeder LF Availability of essential diagnostics in ten low-income and middle-income countries: results from national health facility surveys. The Lancet Global Health (2021).

[2]. Sullivan R et al. Global cancer surgery: delivering safe, affordable, and timely cancer surgery. Lancet Oncol. 16 (11), 1193–1224 (2015). [PubMed: 26427363]

[3]. Cheah P-L, Looi LM & Horton S Cost analysis of operating an anatomic pathology laboratory in a Middle-Income country. Am. J. Clin. Pathol. 149 (1), 1–7 (2018).

[4]. Horbinski C et al. The medical necessity of advanced molecular testing in the diagnosis and treatment of brain tumor patients. Neuro. Oncol. 21 (12), 1498–1508 (2019). [PubMed: 31276167]

[5]. Freudiger CW et al. Label-free biomedical imaging with high sensitivity by stimulated raman scattering microscopy. Science 322 (5909), 1857–1861 (2008). [PubMed: 19095943]

[6]. Orringer DA et al. Rapid intraoperative histology of unprocessed surgical specimens via fibre-laser-based stimulated raman scattering microscopy. Nat Biomed Eng 1 (2017).

[7]. Hollon TC et al. Near real-time intraoperative brain tumor diagnosis using stimulated raman histology and deep neural networks. Nat. Med. (2020).

[8]. Louis DN et al. The 2021 WHO classification of tumors of the central nervous system: a summary. Neuro. Oncol. (2021).

[9]. Eckel-Passow JE et al. Glioma groups based on 1p/19q, IDH, and TERT promoter mutations in tumors. N. Engl. J. Med. 372 (26), 2499–2508 (2015). [PubMed: 26061753]

[10]. Cancer Genome Atlas Research Network et al. Comprehensive, integrative genomic analysis of diffuse Lower-Grade gliomas. N. Engl. J. Med. 372 (26), 2481–2498 (2015). [PubMed: 26061751]

[11]. Yan H et al. IDH1 and IDH2 mutations in gliomas. N. Engl. J. Med. 360 (8), 765–773 (2009). [PubMed: 19228619]

[12]. Metter DM, Colgan TJ, Leung ST, Timmons CF & Park JY Trends in the US and canadian pathologist workforces from 2007 to 2017. JAMA Netw Open 2 (5), e194337 (2019).

[13]. Damodaran S, Berger MF & Roychowdhury S Clinical tumor sequencing: opportunities and challenges for precision cancer medicine. Am Soc Clin Oncol Educ Book e175–82 (2015). [PubMed: 25993170]

[14]. Fortin Ensign S, Hrachova M, Chang S & Mrugala MM Assessing the utility and attitudes toward molecular testing in neuro-oncology: a survey of the society for Neuro-Oncology members. Neurooncol Pract 8 (3), 310–316 (2021). [PubMed: 34055378]

[15]. Chen L et al. Predicting the likelihood of an isocitrate dehydrogenase 1 or 2 mutation in diagnoses of infiltrative glioma. Neuro. Oncol. 16 (11), 1478–1483 (2014). [PubMed: 24860178]

[16]. Bhandari AP, Liong R, Koppen J, Murthy SV & Lasocki A Non-invasive determination of IDH and 1p19q status of lower-grade gliomas using MRI radiomics: A systematic review. AJNR Am. J. Neuroradiol. 42 (1), 94–101 (2021). [PubMed: 33243896]

[17]. Kather JN et al. Pan-cancer image-based detection of clinically actionable genetic alterations. Nat Cancer 1 (8), 789–799 (2020). [PubMed: 33763651]

[18]. He K, Zhang X, Ren S & Sun J Deep residual learning for image recognition. Proc. IAPR Int. Conf. Pattern Recogn. (2016).

[19]. Hollon TC et al. Rapid, label-free detection of diffuse glioma recurrence using intraoperative stimulated raman histology and deep neural networks. Neuro. Oncol. (2020).

[20]. Coudray N et al. Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning. Nat. Med. 24 (10), 1559–1567 (2018). [PubMed: 30224757]

[21]. Jiang C et al. Rapid automated analysis of skull base tumor specimens using intraoperative optical imaging and artificial intelligence. Neurosurgery (2022).

[22]. Chen T, Kornblith S, Norouzi M & Hinton G A simple framework for contrastive learning of visual representations (2020). https://arxiv.org/abs/2002.05709 [cs.LG].

[23]. Frome A et al. Burges CJC, Bottou L, Welling M, Ghahramani Z & K. Q. (eds) DeViSE: A deep Visual-Semantic embedding model. (eds Burges CJC, Bottou L, Welling M, Ghahramani Z. & Weinberger KQ) Advances in Neural Information Processing Systems, Vol. 26 (Curran Associates, Inc., 2013).

[24]. Wang J et al. CNN-RNN: A unified framework for multi-label image classification (2016). https://arxiv.org/abs/1604.04573 [cs.CV].

[25]. Ramesh A et al. Zero-Shot Text-to-Image generation (2021). https://arxiv.org/abs/2102.12092 [cs.CV].

[26]. Saharia C et al. Photorealistic Text-to-Image diffusion models with deep language understanding (2022). https://arxiv.org/abs/2205.11487 [cs.CV].

[27]. Verhaak RGW et al. Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. Cancer Cell 17 (1), 98–110 (2010). [PubMed: 20129251]

[28]. Pennington J, Socher R & Manning C Glove: Global vectors for word representation (Association for Computational Linguistics, Stroudsburg, PA, USA, 2014).

[29]. Vaswani A et al. Guyon I et al. (eds) Attention is all you need. (eds Guyon I et al.) Advances in Neural Information Processing Systems, I Vol. 30 (Curran Associates, Inc., 2017).

[30]. Devlin J, Chang M-W, Lee K & Toutanova K BERT: Pre-training of ‘ deep bidirectional transformers for language understanding (2018). https://arxiv.org/abs/1810.04805 [cs.CL].

[31]. Lanchantin J, Wang T, Ordonez V & Qi Y General multi-label image ‘ classification with transformers (2020). https://arxiv.org/abs/2011.14027 [cs.CV].

[32]. Wang T & Isola P Understanding contrastive representation learning through alignment and uniformity on the hypersphere (2020). https://arxiv.org/abs/2005.10242 [cs.LG].

[33]. DeWitt JC et al. Cost-effectiveness of IDH testing in diffuse gliomas according to the 2016 WHO classification of tumors of the central nervous system recommendations. Neuro. Oncol. 19 (12), 1640–1650 (2017). [PubMed: 29016871]

[34]. Louis DN et al. cIMPACT-NOW (the consortium to inform molecular and practical approaches to CNS tumor taxonomy): a new initiative in advancing nervous system tumor classification. Brain Pathol. 27 (6), 851–852 (2017). [PubMed: 27997995]

[35]. Weller M et al. EANO guidelines on the diagnosis and treatment of diffuse gliomas of adulthood. Nat. Rev. Clin. Oncol. 18 (3), 170–186 (2021). [PubMed: 33293629]

[36]. Capper D et al. DNA methylation-based classification of central nervous system tumours. Nature 555 (7697), 469–474 (2018). [PubMed: 29539639]

[37]. Beiko J et al. IDH1 mutant malignant astrocytomas are more amenable to surgical resection and have a survival benefit associated with maximal surgical resection. Neuro. Oncol. 16 (1), 81–91 (2014). [PubMed: 24305719]

[38]. Cahill DP Extent of resection of glioblastoma: A critical evaluation in the molecular era. Neurosurg. Clin. N. Am. 32 (1), 23–29 (2021). [PubMed: 33223023]

[39]. Molinaro AM et al. Association of maximal extent of resection of Contrast-Enhanced and Non-Contrast-Enhanced tumor with survival within molecular subgroups of patients with newly diagnosed glioblastoma. JAMA Oncol (2020).

[40]. Vanderbeek AM et al. The clinical trials landscape for glioblastoma: is it adequate to develop new treatments? Neuro. Oncol. 20 (8), 1034–1043 (2018). [PubMed: 29518210]

[41]. Chiocca EA et al. Phase IB study of gene-mediated cytotoxic immunotherapy adjuvant to up-front surgery and intensive timing radiation for malignant glioma. J. Clin. Oncol. 29 (27), 3611–3619 (2011) [PubMed: 21844505]

[42]. Wheeler LA et al. Phase II multicenter study of gene-mediated cytotoxic immunotherapy as adjuvant to surgical resection for newly diagnosed malignant glioma. Neuro. Oncol. 18 (8), 1137–1145 (2016). [PubMed: 26843484]

[43]. Desjardins A et al. Recurrent glioblastoma treated with recombinant poliovirus. N. Engl. J. Med. 379 (2), 150–161 (2018). [PubMed: 29943666]

[44]. Brem H et al. Placebo-controlled trial of safety and efficacy of intraoperative controlled delivery by biodegradable polymers of chemotherapy for recurrent gliomas. Lancet 345 (8956), 1008–1012 (1995). [PubMed: 7723496]

[45]. Freudiger CW et al. Stimulated raman scattering microscopy with a robust fibre laser source. Nat. Photonics 8 (2), 153–159 (2014). [PubMed: 25313312]

[46]. Reddy BS & Chatterji BN An FFT-based technique for translation, rotation, and scale-invariant image registration. IEEE Trans. Image Process. 5 (8), 1266–1271 (1996). [PubMed: 18285214]

[47]. Zhou L, Chen G, Feng M & Knoll A Improving Low-Resolution image classification by Super-Resolution with enhancing High-Frequency content, 1972–1978 (2021).

[48]. Zhang J et al. The international cancer genome consortium data portal. Nat. Biotechnol. 37 (4), 367–369 (2019). [PubMed: 30877282]

[49]. Gusev Y et al. The REMBRANDT study, a large collection of genomic data from brain cancer patients. Sci Data 5, 180158 (2018).

[50]. Jonsson P et al. Genomic correlates of disease progression and treatment response in prospectively characterized gliomas. Clin. Cancer Res. 25 (18), 5537–5547 (2019). [PubMed: 31263031]

[51]. Du J et al. Gene2vec: distributed representation of genes based on co-expression. BMC Genomics 20 (Suppl 1), 82 (2019). [PubMed: 30712510]

[52]. Dosovitskiy A et al. An image is worth 16×16 words: Transformers for image recognition at scale (2020).

[53]. Deng J et al. ImageNet: A large-scale hierarchical image database, 248–255 (2009).

[54]. Hollon TC et al. Rapid, label-free detection of diffuse glioma recurrence using intraoperative stimulated raman histology and deep neural networks. Neuro. Oncol. (2020).

[55]. Wiens J et al. Do no harm: a roadmap for responsible machine learning for health care. Nat. Med. 25 (9), 1337–1340 (2019). [PubMed: 31427808]

[56]. Ostrom QT et al. CBTRUS statistical report: Primary brain and other central nervous system tumors diagnosed in the united states in 2013–2017. Neuro. Oncol. 22 (12 Suppl 2), iv1-iv96 (2020).

[57]. Li J, Zhou P, Xiong C & Hoi SCH Prototypical contrastive learning of unsupervised representations (2020). https://arxiv.org/abs/2005.04966 [cs.CV].

[58]. Wang F & Liu H Understanding the behaviour of contrastive loss (IEEE, 2021).

[59]. Khosla P et al. Supervised contrastive learning (2020). https://arxiv.org/abs/2004.11362 [cs.LG].

[60]. van den Oord A, Li Y & Vinyals O Representation learning with contrastive predictive coding (2018). https://arxiv.org/abs/1807.03748 [cs.LG].

[61]. Chen X, Xie S & He K An empirical study of training Self-Supervised vision transformers (2021). https://arxiv.org/abs/2104.02057 [cs.CV].

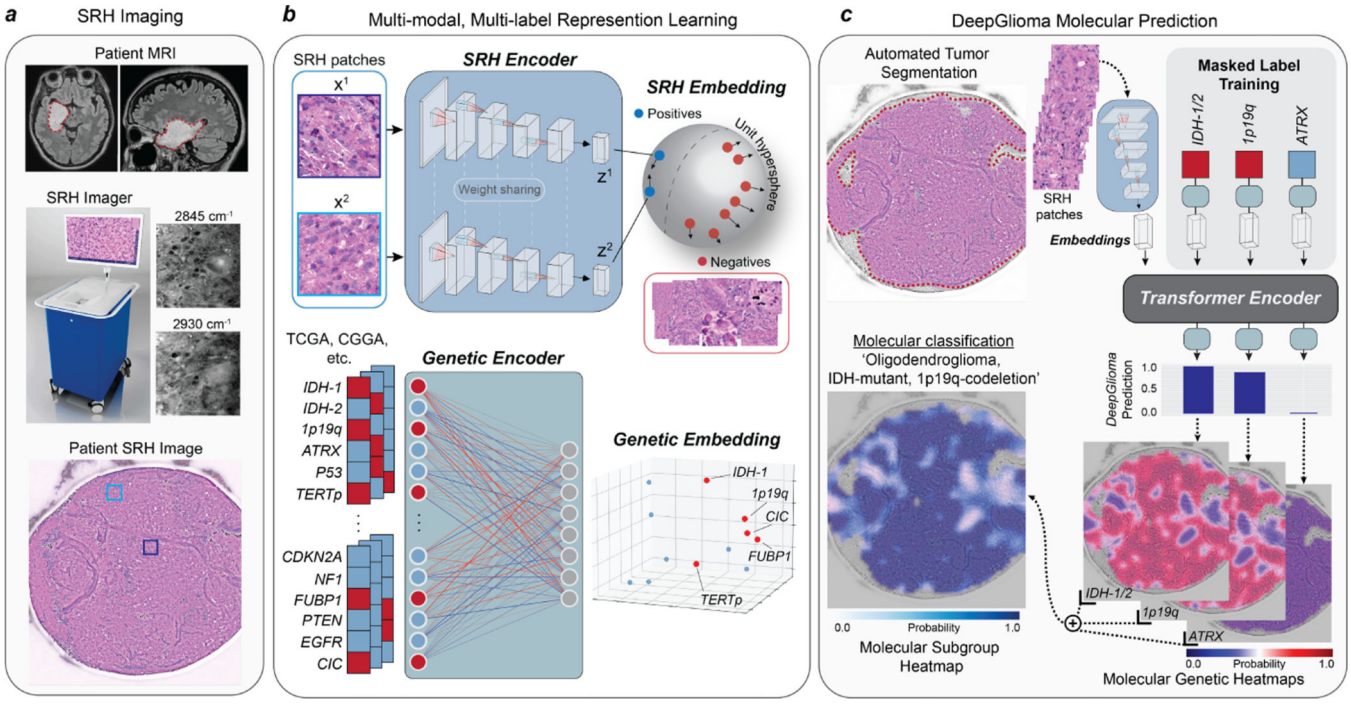


Fig. 1. Bedside SRH and DeepGlioma workflow.

a, A patient with a suspected diffuse glioma undergoes surgery for tumor biopsy or surgical resection. The SRH imaging system is portable and imaging takes place in the operating room, performed by a single technician using simple touch screen instructions. A freshly excised tissue specimen is loaded directly into a premade microscope slide and inserted into the SRH imager without the need for tissue processing (Extended Data Fig. 1). Raw SRH images are acquired at two Raman shifts, 2,845cm^{-1} and 2,930cm^{-1}, as strips. The time to acquire a 3×3mm^2 SRH image is approximately 90 seconds. Raw optical images can then be colored using a custom hematoxlyin and eosin (HE) virtual staining method for clinician review. **b**, DeepGlioma is trained using a multi-modal dataset. First, SRH images are used to train an CNN encoder using weakly supervised, multi-label contrastive learning for image feature embedding (Extended Data Fig. 3). Second, public diffuse glioma genomic data from TCGA, CGGA, and others (Extended Data Table 2) are used to train a genetic encoder to learn a genetic embedding that represents known co-occurrence relationships between genetic mutations (Extended Data Fig. 5). **c**, The SRH and genetic encoders are integrated into a single architecture using a transformer encoder for multi-label prediction of diffuse glioma molecular diagnostic mutations. We use masked label training to train the transformer encoder (Extended Data Fig. 6). Because our system uses patch-level predictions, spatial heatmaps can be generated for both molecular genetic and molecular subgroup predictions to improve model interpretability, identify regions of variable confidence, and associate SRH image features with DeepGlioma predictions (Extended Data Fig. 9 and 10).

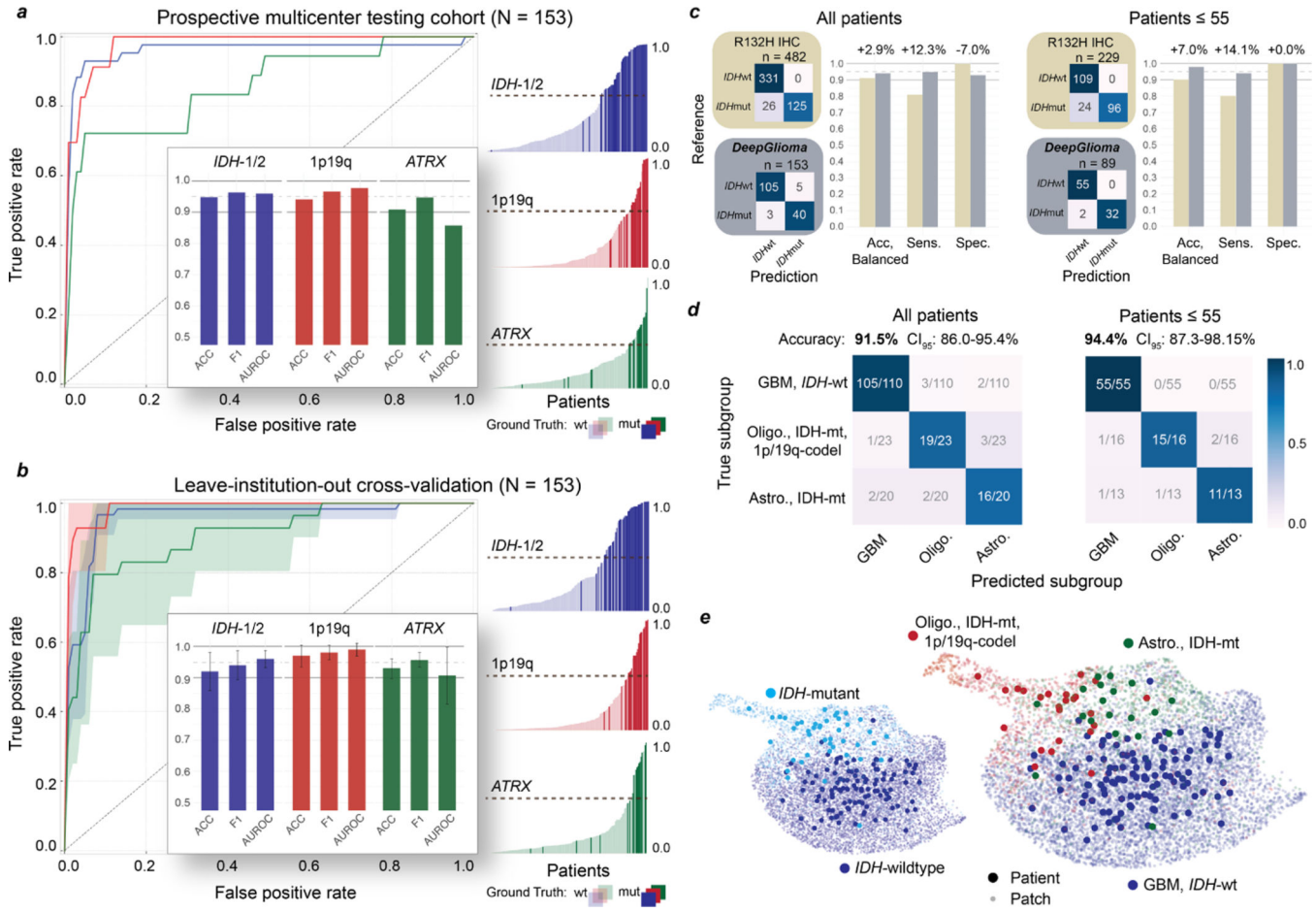


Fig. 2. DeepGlioma molecular classification performance

a, Results from our prospective multicenter testing cohort of diffuse glioma patients are shown. DeepGlioma was trained using UM data only and tested on our external medical centers. All results are presented as patient-level predictions. Individual ROC curves for IDH-1/2 (AUROC 95.9%), 1p9q-codeletion (AUROC 97.7%), and ATRX (AUROC 85.7%) classification are shown. Our AUROC values were highest for IDH-1/2 and 1p19q-codeletion prediction. Bar plot inset shows the accuracy, F1 score, and AUROC classification metrics for each of the mutations. Similar to our cross-validation experiments, ATRX mutation prediction was the most challenging as demonstrated by comparatively lower metric scores. Individual patient-level molecular genetic prediction probabilities are ordered and displayed. **b**, Results from the LIOCV experiments. Mean (solid line) and standard deviation (fill color) ROC curves are shown. Metrics are averaged over external testing centers to determine the stability of DeepGlioma classification results given different patient populations, clinical workflows, and SRH imagers. Including additional training data resulted in an increase in DeepGlioma performance, especially for 1p19q and ATRX classification. **c**, *Primary testing endpoint*: comparison of IDH1-R132H IHC versus DeepGlioma for IDH mutational status detection. DeepGlioma achieved a 94.2% balanced accuracy for the prospective cohort and a 97.0% balanced accuracy for patients 55 years or less. The major performance boost was due to the +10% increase in prediction

sensitivity over IDH1-R132H IHC due to DeepGlioma's detection of both canonical and non-canonical IDH mutations. **d**, *Secondary testing endpoint*: DeepGlioma results for molecular subgrouping according to WHO CNS5 adult-type diffuse glioma taxonomy. Multiclass classification accuracy for all patients and patients 55 years or less are shown. **e**, UMAP visualization of SRH representations from DeepGlioma. Small, semi-transparent points are SRH patch representations and large, solid points are patient representations (i.e. average patch location) from the prospective clinical cohort. Representations are labeled according to their IDH subgroup and diffuse glioma molecular subgroup. Our patch contrastive learning encourages the SRH encoder to learn representations that are uniformly distributed on the unit hypersphere [32].