











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

Received: 1 November 2022

Accepted: 8 February 2023

Published online: 23 March 2023

 Check for updates

Todd Hollon ^{1,2}✉, Cheng Jiang ^{1,2}, Asadur Chowdury ¹, Mustafa Nasir-Moin³, Akhil Kondepudi¹, Alexander Aabedi ⁴, Arjun Adapa¹, Wajd Al-Holou⁵, Jason Heth⁵, Oren Sagher⁵, Pedro Lowenstein⁴, Maria Castro⁴, Lisa Irina Wadiura ⁶, Georg Widhalm ⁶, Volker Neuschmelting ⁷, David Reinecke ⁷, Niklas von Spreckelsen ⁷, Mitchel S. Berger⁴, Shawn L. Hervey-Jumper⁴, John G. Golfinos³, Matija Snuderl⁸, Sandra Camelo-Piragua⁹, Christian Freudiger¹⁰, Honglak Lee¹¹ & Daniel A. Orringer ³

Molecular classification has transformed the management of brain tumors by enabling more accurate prognostication and personalized treatment. However, timely molecular diagnostic testing for patients with brain tumors is limited, complicating surgical and adjuvant treatment and obstructing clinical trial enrollment. In this study, we developed DeepGlioma, a rapid (<90 seconds), artificial-intelligence-based diagnostic screening system to streamline the molecular diagnosis of diffuse gliomas. DeepGlioma is trained using a multimodal dataset that includes stimulated Raman histology (SRH); a rapid, label-free, non-consumptive, optical imaging method; and large-scale, public genomic data. In a prospective, multicenter, international testing cohort of patients with diffuse glioma ($n = 153$) who underwent real-time SRH imaging, we demonstrate that DeepGlioma can predict the molecular alterations used by the World Health Organization to define the adult-type diffuse glioma taxonomy (IDH mutation, 1p19q co-deletion and ATRX mutation), achieving a mean molecular classification accuracy of $93.3 \pm 1.6\%$. Our results represent how artificial intelligence and optical histology can be used to provide a rapid and scalable adjunct to wet lab methods for the molecular screening of patients with diffuse glioma.

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 using a handful of molecular markers¹. However, molecular subgrouping of diffuse gliomas requires laboratory techniques such as immunohistochemistry (IHC), cytogenetic testing and, often, next-generation sequencing that are not uniformly available at the centers where patients with brain

tumors are treated. Moreover, the expert interpretation of molecular data is increasingly challenging in the setting of a declining pathology workforce². Consequently, molecular diagnostic and sequencing techniques for brain tumors, when available, are commonly associated with long turnaround times even in well-resourced settings (days to weeks)³. Barriers to molecular diagnosis can result in suboptimal care for patients with brain tumors, complicating prognostic prediction,

A full list of affiliations appears at the end of the paper. ✉ e-mail: tocho@med.umich.edu

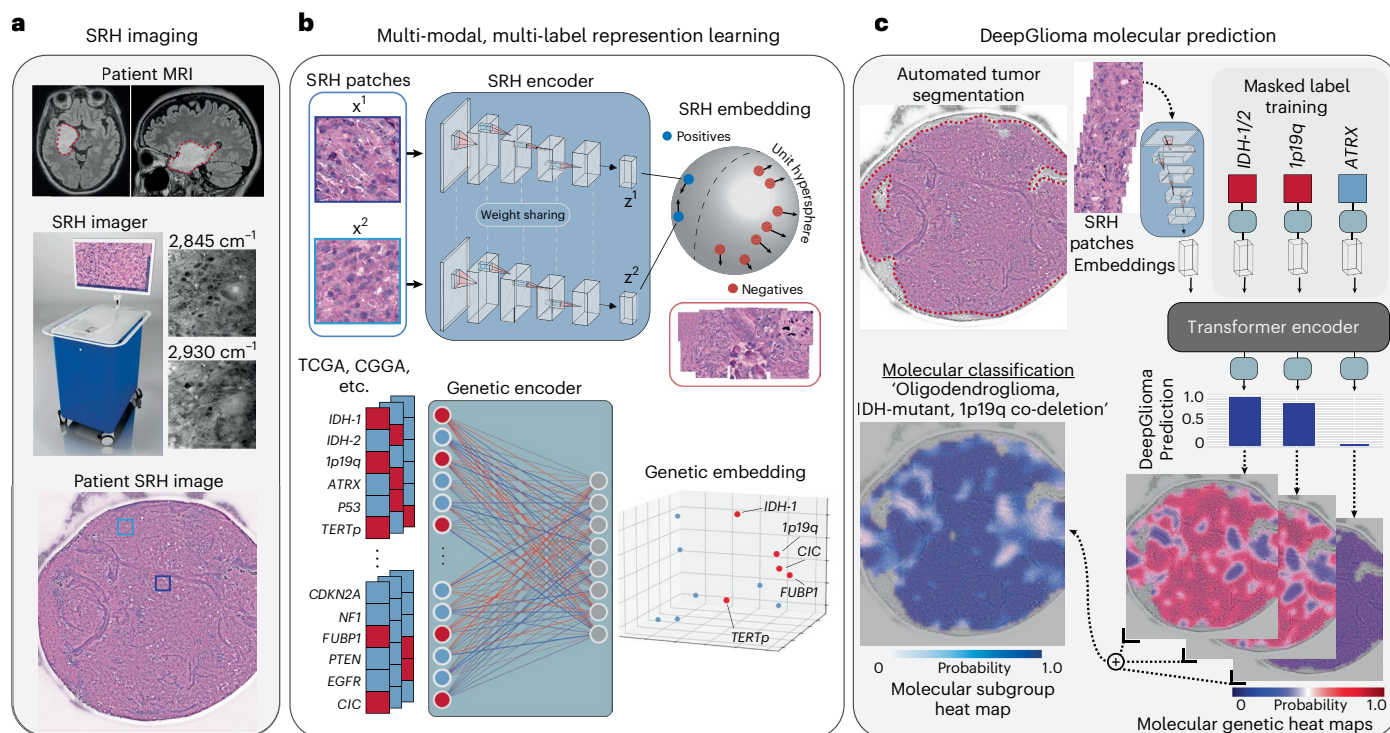


Fig. 1 | Bedside SRH and DeepGlioma workflow. **a**, A patient with a suspected diffuse glioma undergoes biopsy or surgical resection. The portable SRH imaging system is used to acquire histologic images in the operating room, performed by a single technician using simple touchscreen instructions. A freshly excised, unprocessed tissue specimen is loaded directly into a custom microscope slide (Extended Data Fig. 1). Raw SRH images are acquired at two Raman shifts, 2,845 cm^{-1} and 2,930 cm^{-1} , as strips. The time to acquire a $3 \times 3\text{-mm}^2$ SRH image is approximately 90 seconds. Raw optical images are rendered using a virtual H&E-like lookup table for clinician review⁵. **b**, SRH images are used to train a 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 (Supplementary Data Table 2) are used to train a genetic encoder to learn a genetic embedding (Extended Data

Fig. 5). **c**, DeepGlioma molecular prediction is achieved by using a pre-trained segmentation model⁶ to identify tumor regions, generate patches within those regions and perform a feedforward pass of tumor patches through the SRH encoder. The SRH and genetic encoders are integrated into a transformer model for multi-label prediction of diffuse glioma molecular diagnostic mutations. To improve DeepGlioma performance, we used masked label training to train the transformer encoder (Extended Data Fig. 5). DeepGlioma input is SRH images only during inference. Because our system uses patch-level predictions, spatial heat maps 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 Figs. 9 and 10).

surgical decision-making, extent of resection goals, selection of adjuvant chemoradiation regimens and clinical trial enrollment. Here we propose and prospectively validate an artificial-intelligence-based approach to simplify the molecular classification of diffuse gliomas through automated image analysis of rapid optical imaging of fresh, unprocessed surgical specimens.

Results

DeepGlioma is an artificial-intelligence-based diagnostic screening system that combines deep neural networks and stimulated Raman histology (SRH) to achieve rapid molecular screening of fresh glioma specimens (Fig. 1). Our approach predicts the most critical diagnostic genetic alterations in diffuse glioma using learned spectroscopic and histopathologic image features to inform patient care and guide downstream definitive molecular testing. Using SRH images only, DeepGlioma can achieve molecular classification in less than 2 minutes of tissue biopsy without the need for tissue processing or human interpretation (Extended Data Fig. 1). Although DeepGlioma can scale to an arbitrary number of diagnostic mutations, we focus on the major molecular diagnostic alterations used by the fifth edition of the World Health Organization Classification of Tumors of the Central Nervous System (WHO CNS5) to define the diffuse glioma subgroups: isocitrate dehydrogenase-1/2 (IDH) mutations, 1p19q chromosome co-deletion and ATRX loss^{1,4}.

The SRH workflow begins when a fresh, unprocessed surgical specimen is biopsied, and a small (3-mm) sample is placed into a custom 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 to generate two image channels: 2,845 cm^{-1} and 2,930 cm^{-1} (ref. 5). SRH patches are then sampled in a raster fashion from the whole-slide SRH image to generate non-overlapping, single-scale, high-resolution images for model training and inference. We used SRH images from 373 adult patients with diffuse glioma at the University of Michigan (UM) to train a deep convolutional neural network (CNN) as a visual encoder (Supplementary 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. Although previous studies have used linear classification layers trained end to end using cross-entropy⁶, we found that weakly supervised (that is, patient labels only) patch-based contrastive learning, or PatchCon, was ideally suited for whole-slide SRH classification (Fig. 1b and Extended Data Fig. 3)⁷. We developed a simple and general framework for multi-label contrastive learning of visual representations and trained an SRH encoder using this framework (Extended Data Fig. 4).

Next, we pre-trained a genetic embedding model using large-scale, public glioma genomic data (Fig. 1b and Supplementary Data Table 2). We aimed to learn a genetic embedding space that meaningfully

encodes the relationships between mutations to improve SRH classification. The co-occurrence of specific mutations in the same tumor defines the molecular subgroups of diffuse gliomas^{8,9}. The genetic embedding model learns to represent the co-occurrence dataset statistics using global vector embeddings¹⁰. The model learned a linear substructure that matches known molecular subgroups of diffuse gliomas (Extended Data Fig. 5). By pre-training 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.

Finally, the pre-trained SRH and genetic encoders are integrated into a transformer architecture for multi-label molecular classification (Fig. 1c)¹¹. During transformer training, the input tokens are the visual embedding of the SRH patch and the genetic embedding for the patient's tumor. Similarly to masked language modeling¹², we randomly mask a subset of the genes from the input, and the objective is to predict the masked genes. During inference, the transformer uses only the SRH patch embedding to predict the mutational status of each gene. We performed iterative hold-out cross-validation to show the advantage of PatchCon, genetic pre-training and masked label training through several ablation studies. We demonstrated that DeepGlioma was able to achieve a mean area under the receiver operator characteristic curve (mAUROC) of $92.6 \pm 5.4\%$ for molecular classification on held-out SRH data. Our multimodal training strategy that included the pre-trained genetic embedding model results in $\sim 5\%$ increase in overall classification performance (Extended Data Fig. 6b).

We tested DeepGlioma in a multicenter, prospective cohort of diffuse gliomas to evaluate how our model generalizes across different patient populations, patient care settings and SRH imaging systems. Model testing was designed as a non-inferiority diagnostic clinical trial. Four tertiary medical centers across the United States and Europe were included as testing recruitment centers. Patients were recruited as a consecutive cohort of adult (>18 years of age) patients with brain tumors who underwent biopsy or tumor resection for diffuse glioma. A total of 153 patients were included (Supplementary Data Table 3). DeepGlioma achieved a molecular classification accuracy for IDH mutation of 94.7% (95% confidence interval (CI): 90.0–97.7%), 1p19q co-deletion of 94.1% (95% CI: 89.1–97.3%) and ATRX mutation of 91.0% (95% CI: 85.1–94.9%), resulting in a mean accuracy of $93.3 \pm 1.6\%$. (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 co-deletion and ATRX, respectively.

Next, we performed a set of leave-institution-out cross-validation (LIOCV) experiments to (1) assess the stability of DeepGlioma performance across medical centers and (2) determine the effect of increasing training data on model performance (Fig. 2b). DeepGlioma demonstrated stability across each LIOCV iteration with molecular classification accuracy standard deviation range of ± 2.75 – 6.06% and an F1 score range of ± 1.71 – 4.70% . The prediction of ATRX mutations was consistently more challenging across our experiments. 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.

We compared the performance of DeepGlioma versus the gold standard molecular screening modality for glioma classification: IDH1-R132H IHC. Given that non-canonical (non-R132H) IDH mutations occur in 20–30% of IDH-mutant lower-grade gliomas¹³, IDH1-R132H IHC has known limitations in clinical practice. Due to the higher rates of lower-grade gliomas in young patients, genetic sequencing of IDH is recommended for glioma patients 55 years of age or younger¹⁴. Agnostic to IDH isoform, DeepGlioma generalizes to both canonical and non-canonical IDH mutations. IDH1-R132H IHC has a balanced accuracy of 91.4% (sensitivity 82.8%, specificity 100%). In our testing cohort,

DeepGlioma achieved a balanced accuracy of 94.2% (sensitivity 95.5%, specificity 93.0%). In patients 55 years of age or younger, IDH1-R132H has a balanced accuracy of 90.0%, and DeepGlioma achieved a balanced accuracy of 97.0% (Fig. 2c). Full patient demographic subgroup analyses can be found in Extended Data Figs. 7 and 8. All non-canonical mutations in our prospective cohort were correctly classified by DeepGlioma (Extended Data Fig. 7g).

Finally, DeepGlioma's prediction of the molecular genetics of diffuse gliomas enables direct classification of SRH images into three mutually exclusive diffuse glioma subgroups as defined by the WHO CNS5 classification scheme (IDH-wild-type, IDH-mutant and 1p19q co-deletion, IDH-mutant and 1p19q intact)¹. An algorithmic inference method was developed to classify each patient into a molecular subgroup (Algorithm 1). We established an artificial-intelligence-based performance benchmark motivated by our previous methods of SRH classification trained for multiclass classification⁶. DeepGlioma achieved a molecular subgroup classification accuracy of 91.5% (95% CI: 86.0–95.4%) (Fig. 2d) and demonstrated a $+4.6\%$ performance increase over our benchmark model (Extended Data Fig. 8 and Supplementary Data Table 4). The major performance gains of DeepGlioma are due to increased sensitivity for identifying IDH-mutant gliomas and modeling the co-occurrences of mutations within molecular subgroups. In patients 55 years of age or younger, our classification performance showed an overall increase ($+2.9\%$), obtaining a classification accuracy of 94.4% (95% CI: 87.3–98.2%) (Fig. 2d and Extended Data Fig. 8). DeepGlioma performance generalized well across multiple medical centers with distinct patient populations, clinical presentations, personnel and infrastructure. Molecular subgroup prediction heat maps for both canonical (Extended Data Fig. 9) and non-canonical (Extended Data Fig. 10) IDH mutations 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 at <https://deepglioma.mlins.org/>.

Discussion

We present DeepGlioma, a deep-learning-based screening system designed to streamline the detection of key molecular alterations in human gliomas. DeepGlioma accurately predicted IDH mutations, 1p19q co-deletion and ATRX mutations without the need for fluorescence in situ hybridization or genetic sequencing, enabling automated molecular subtyping of diffuse gliomas according to the WHO classification scheme¹.

Access to molecular diagnostic testing is uneven for patients who receive brain tumor care. DeepGlioma can streamline molecular testing by providing rapid molecular screening, enabling clinicians to focus on confirming the most likely diagnostic mutations only rather than using a diagnostic shotgun approach¹⁵. In addition, SRH is not consumptive and does not diminish diagnostic yield of tumor specimens, preserving scant clinical samples for definitive molecular testing.

Streamlining molecular classification could also have an immediate impact on the surgical care of patients with brain tumors. Surgical goals should be tailored based on molecular subgroups^{16,17}. Patients with molecular astrocytoma who undergo gross total resection achieve a 5-year increase in median survival compared to patients who receive subtotal resections (~ 12 -year versus >17 -year median survival). DeepGlioma creates an avenue for accurate and timely differentiation of diffuse glioma subgroups to define surgical goals with a better-calibrated risk–benefit analysis.

Even with optimal standard-of-care treatment, patients with diffuse glioma face limited treatment options. The development of novel therapies through clinical trials is essential. Unfortunately, fewer than 10% of patients with glioma are enrolled in clinical trials¹⁸. Clinical trials limit inclusion criteria to a specific subpopulation, often defined by molecular subgroups. DeepGlioma can initiate the process for trial

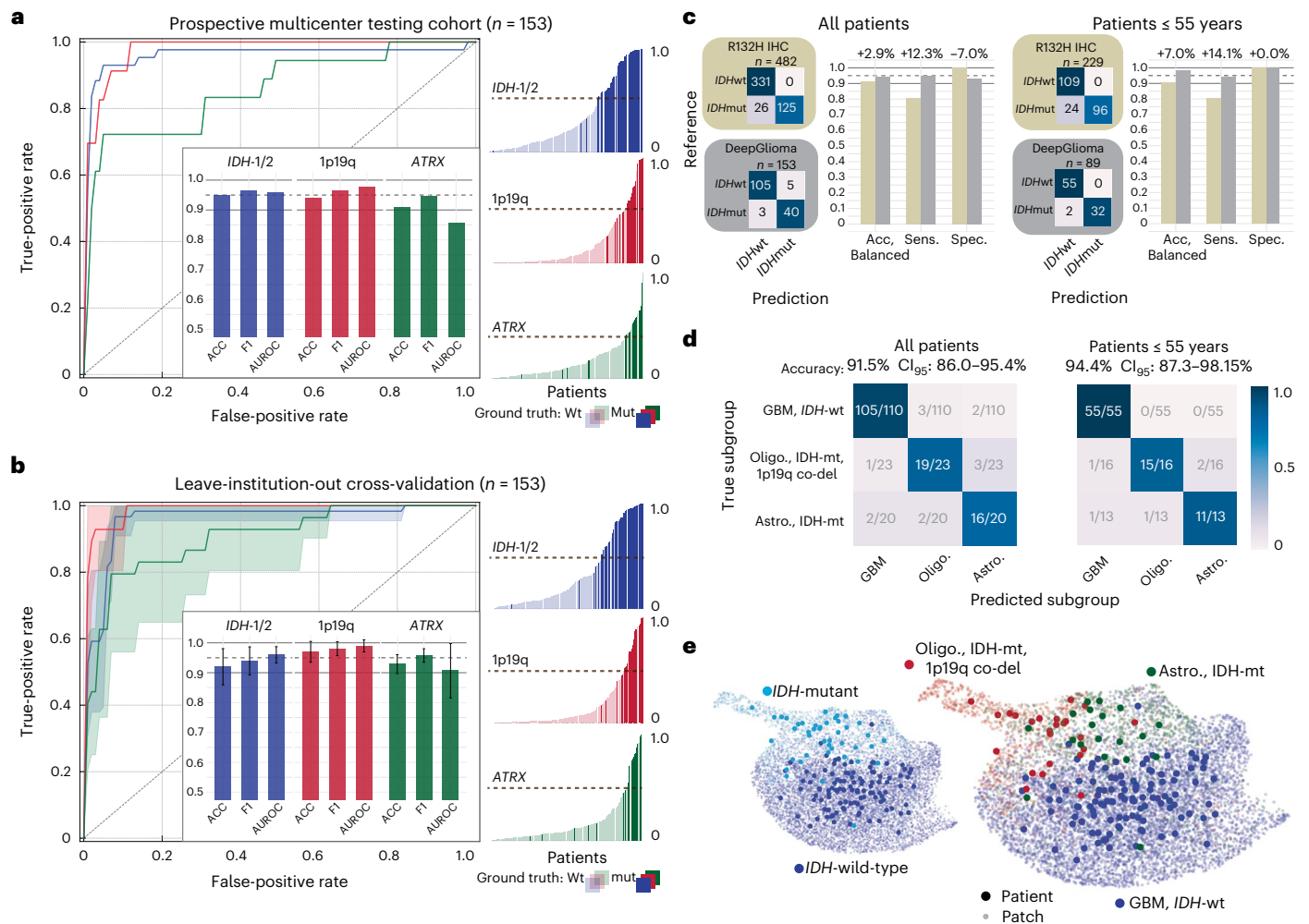


Fig. 2 | DeepGlioma molecular classification performance. **a**, Results from our prospective multicenter testing cohort of patients with diffuse glioma are shown. DeepGlioma was trained using UM data only ($n = 373$) and tested on our external medical centers ($n = 153$). All results are presented as patient-level predictions. Individual ROC curves for IDH (AUROC 95.9%), 1p9q (AUROC 97.7%) and ATRX (AUROC 85.7%) classification are shown. Bar plot inset shows the accuracy, F1 score and AUROC classification metrics for each mutation. 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 (mean \pm s.d.) 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. **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 of age or younger. 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 diffuse glioma taxonomy¹. Multiclass classification accuracy for all patients and patients 55 years of age or younger are shown. **e**, UMAP visualization of SRH representations from DeepGlioma. Small, semi-transparent points are patch representations, and large, solid points are patient representations (that is, average patch location). Representations are labeled according to their IDH subgroup and diffuse glioma molecular subgroup. ACC, accuracy; mut, mutant; UMAP, uniform manifold approximation and projection; wt, wild-type.

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, a major challenge in therapeutic delivery.

Limitations of our study include that the external testing cohort was restricted to the United States and Europe, potentially overfitting to this patient demographic. Although our subgroup analysis did not show a difference in performance across minority populations, DeepGlioma validation using a diverse, global demographic would improve model testing. Similarly to other deep neural networks, DeepGlioma is not directly interpretable. Uncovering the learned optical image features that predict molecular subgroups is an open question for future investigations.

In conclusion, our study demonstrates how artificial-intelligence-based screening methods have the potential to augment existing conventional diagnostic techniques to improve the access and speed of molecular diagnosis and improve the care of patients with brain tumors.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41591-023-02252-4>.

References

1. Louis, D. N. et al. The 2021 WHO Classification of Tumors of the Central Nervous System: a summary. *Neuro Oncol.* **23**, 1231–1251 (2021).
2. Metter, D. M., Colgan, T. J., Leung, S. T., Timmons, C. F. & Park, J. Y. Trends in the US and Canadian pathologist workforces from 2007 to 2017. *JAMA Netw. Open* **2**, e194337 (2019).
3. Damodaran, S., Berger, M. F. & Roychowdhury, S. Clinical tumor sequencing: opportunities and challenges for precision cancer medicine. *Am. Soc. Clin. Oncol. Educ. Book* **2015**, e175–e182 (2015).
4. Brat, D. J. et al. Molecular biomarker testing for the diagnosis of diffuse gliomas. *Arch. Pathol. Lab. Med* **146**, 547–574 (2022).
5. Orringer, D. A. et al. Rapid intraoperative histology of unprocessed surgical specimens via fibre-laser-based stimulated Raman scattering microscopy. *Nat. Biomed. Eng.* **1**, 0027 (2017).
6. Hollon, T. C. et al. Near real-time intraoperative brain tumor diagnosis using stimulated Raman histology and deep neural networks. *Nat. Med.* **26**, 52–58 (2020).
7. Chen, T., Kornblith, S., Norouzi, M. & Hinton, G. A simple framework for contrastive learning of visual representations. in *Proceedings of the 37th International Conference on Machine Learning* (eds Iii, H. D. & Singh, A.) 1597–1607 (PMLR, 2020).
8. Eckel-Passow, J. E. et al. Glioma groups based on 1p/19q, IDH, and TERT promoter mutations in tumors. *N. Engl. J. Med.* **372**, 2499–2508 (2015).
9. Cancer Genome Atlas Research Network et al. Comprehensive, integrative genomic analysis of diffuse lower-grade gliomas. *N. Engl. J. Med.* **372**, 2481–2498 (2015).
10. Pennington, J., Socher, R. & Manning, C. GloVe: global vectors for word representation. in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (Association for Computational Linguistics, 2014).
11. Vaswani, A. et al. Attention is all you need. in *Advances in Neural Information Processing Systems* (eds Guyon, I. et al.) 30 (Curran Associates, 2017).
12. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Volume 1 (Long and Short Papers) 4171–4186 (Association for Computational Linguistics, 2019).
13. DeWitt, J. C. 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**, 1640–1650 (2017).
14. Louis, D. N. 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**, 851–852 (2017).
15. Capper, D. et al. DNA methylation-based classification of central nervous system tumours. *Nature* **555**, 469–474 (2018).
16. Drexler, R. et al. DNA methylation subclasses predict the benefit from gross total tumor resection in IDH-wildtype glioblastoma patients. *Neuro Oncol.* **25**, 315–325 (2022).
17. Hervey-Jumper, S. L. et al. Interactive effects of molecular, therapeutic, and patient factors on outcome of diffuse low-grade glioma. *J. Clin. Oncol.* <https://doi.org/10.1200/JCO.21.02929> (2023).
18. Vanderbeek, A. M. et al. The clinical trials landscape for glioblastoma: is it adequate to develop new treatments? *Neuro Oncol.* **20**, 1034–1043 (2018).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

© The Author(s), under exclusive licence to Springer Nature America, Inc. 2023

¹Machine Learning in Neurosurgery Laboratory, Department of Neurosurgery, University of Michigan, Ann Arbor, MI, USA. ²Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI, USA. ³Department of Neurosurgery, New York University, New York, NY, USA. ⁴Department of Neurological Surgery, University of California, San Francisco, San Francisco, CA, USA. ⁵Department of Neurosurgery, University of Michigan, Ann Arbor, MI, USA. ⁶Department of Neurosurgery, Medical University Vienna, Vienna, Austria. ⁷Department of Neurosurgery, University Hospital Cologne, Cologne, Germany. ⁸Department of Pathology, New York University, New York, NY, USA. ⁹Department of Pathology, University of Michigan, Ann Arbor, MI, USA. ¹⁰Invenio Imaging, Inc., Santa Clara, CA, USA. ¹¹Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, MI, USA.

✉ e-mail: tocho@med.umich.edu

Methods

Study design

The main objectives of the study were (1) to develop a rapid molecular diagnostic screening tool for classifying adult-type diffuse gliomas into the taxonomy defined by the WHO CNS5 (ref.¹) using clinical SRH and deep-learning-based computer vision methods and (2) to 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, cytologic and 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 glioma specimens for specific mutations to inform intraoperative decision-making and potentially improve early clinical trial enrollment. DeepGlioma consists of two pre-trained separable modules—a visual encoder and a genetic encoder—that are integrated using a multi-headed attention mechanism for image classification¹¹. Inspired by previous work on deep visual-semantic embedding¹⁹ and text-to-image generation^{20–22}, 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 UM, 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 co-deletion and ATRX mutation.

SRH

The operating surgeon was instructed to provide a grossly lesional-appearing but viable tumor for SRH imaging. This strategy applies to all brain tumor biopsies to maximize the chance of sampling diagnostic tissue. Fiber-laser-based stimulated Raman scattering microscopy was used to acquire all images in our study^{23,24}. Detailed description of laser configuration was previously described⁵. In brief, surgical specimens were stimulated with a pump beam at 790 nm and a Stokes beam that has a tunable range from 1,015 nm to 1,050 nm. These laser settings allow for access to the Raman shift spectral range between 2,800 cm⁻¹ and 3,130 cm⁻¹. Images were acquired as 1,000-pixel strips with an imaging speed of 0.4 Mpixels per strip. We acquire two image channels sequentially at 2,845 cm⁻¹ (CH2 channel) and 2,930 cm⁻¹ (CH3 channel) Raman wavenumber shifts. A strong stimulated Raman signal at 2,845 cm⁻¹ corresponds to the CH2 symmetric stretching mode of lipid-rich structures, such as myelin. A Raman peak at 2,930 cm⁻¹ highlights protein-rich 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 image⁵. A virtual hematoxylin and eosin (H&E) color scheme can be applied to the two Raman channels to generate a three-channel, virtually stained RGB SRH image. These images provide a major advantage over conventional H&E histology because they allow for real-time pathologic review without degradation of diagnostic accuracy. Multiple studies have demonstrated near-perfect diagnostic concordance with 10× time savings^{5,25}. These images are used for clinician interpretation and are designed to replicate the image contrast seen in conventional H&E histology but are not used for model training. An overview of the SRH imaging workflow can be found in Extended Data Fig. 1.

Image data processing

All model training and inference was done using the raw, non-virtually colored SRH images. All images were acquired, processed and archived as 16-bit images to retain spectroscopic image features. Each strip has a 900-pixel width (that is, after edge clipping) and up to 6,000-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. After 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 raster window over the full image. SRH patches are then classified into one of three classes—tumor, normal brain or non-diagnostic tissue—using our previous trained whole-slide SRH segmentation model^{6,26}. Only tumor regions are used for DeepGlioma training and inference (Extended Data Fig. 1c).

Patient enrollment and training dataset generation

Clinical SRH imaging began at UM on 1 June 2015 after institutional review board approval (HUM00083059). Our imaging dataset was generated using two SRH imaging systems: an initial prototype SRH imager⁵ and the NIO Imaging System (Invenio Imaging)⁶. All patients with a suspected brain tumor were approached for intraoperative SRH imaging. Inclusion criteria for SRH imaging were as follows: patients who were undergoing surgery for (1) suspected central nervous system tumor and/or (2) epilepsy; (3) patient or durable power of attorney was able to provide consent; and (4) preoperative expectation that additional tumor tissue would be available beyond what is required for clinical pathologic diagnosis. Exclusion criteria were as follows: (1) insufficient diagnostic tissue as determined by surgeon or pathologist; (2) grossly inadequate tissue (for example, hemorrhagic, necrotic, fibrous, liquid, etc.); and (3) SRH imager malfunction. After intraoperative SRH imaging, inclusion criteria for the diffuse glioma training dataset were the following: (1) 18 years of age or older and (2) final pathologic diagnosis of an adult-type diffuse glioma as defined by WHO CNS5 (ref.¹) classification. Exclusion criterion 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 Supplementary Data Table 1 and Extended Data Fig. 2.

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 ResNet-50 architecture²⁷ for SRH feature extraction and did not find that larger models provided better performance. Although much of our previous work used conventional cross-entropy loss functions to train deep neural networks^{5,6,26}, we found that contrastive loss functions result in better visual representation learning^{7,28}. We trained our model using a supervised contrastive loss:

$$\mathcal{L}^{\text{sup}} = \sum_{i \in I} \mathcal{L}_i^{\text{sup}} = - \sum_{i \in I} \frac{1}{|P(i)|} \sum_{p \in P(i)} \log \frac{\exp\left(\frac{\text{sim}(g(\mathbf{z}_i), g(\mathbf{z}_p))}{\tau}\right)}{\sum_{n \in A(i)} \exp\left(\frac{\text{sim}(g(\mathbf{z}_i), g(\mathbf{z}_n))}{\tau}\right)} \quad (1)$$

where $\mathbf{z} = f(\mathbf{x}) \in \mathbb{R}^d$ is the d -dimensional feature vector of image \mathbf{x} after a feedforward pass through the visual encoder $f(\cdot)$. A linear projection layer $g(\cdot)$ maps the image feature vector \mathbf{z}_i to a 128-dimensional space where the contrastive objective is computed. \mathbf{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, $\text{sim}(\mathbf{u}, \mathbf{v}) = \frac{\mathbf{u} \cdot \mathbf{v}}{\|\mathbf{u}\| \|\mathbf{v}\|}$, to enforce that all feature vectors are on the unit hypersphere. We developed a novel framework for supervised contrastive learning to accommodate multi-label classification tasks. Each label is assigned a unique projection layer $g_\ell(\cdot)$ for computing a label-wise supervised contrastive objective. The final weighted multi-label supervised contrastive loss is:

$$\mathcal{L}_{\text{multi-label}}^{\text{sup}} = \sum_{\ell \in \mathcal{L}} \lambda_\ell \sum_{i \in I} \mathcal{L}_i^{\text{sup}}(i, g_\ell(\cdot), P_\ell(i)) \quad (2)$$

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 with an initial learning rate of 0.001, 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. After training, all projection layers were discarded, and the visual encoder $f(\cdot)$ was retained for multi-label classification training. We call the above visual representation learning strategy PatchCon for weakly supervised (for example, patient labels only), patch-based contrastive representation learning, and results can be found in Extended Data Fig. 4.

Diffuse glioma genetic embedding

A major component of our multimodal training method includes public genomic data from adult patients with diffuse glioma to pre-train a genetic embedding model. We aggregated genomic data from The Cancer Genome Atlas (TCGA), the Chinese Glioma Genome Atlas (CGGA)²⁹, the International Cancer Genome Consortium (ICGC)³⁰, the REMBRANDT brain cancer dataset³¹, the Memorial Sloan Kettering (MSK) Data Catalog³² and the Mayo Glioblastoma Xenograft National Resource. A total of 2,777 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 Supplementary Data Table 2. In brief, we selected common recurrent somatic mutations found in adult-type diffuse gliomas and encoded those mutations as either mutant or wild-type for each patient. Inspired by previous work on word embeddings¹⁰, 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.

$$\mathcal{L}_{\text{embed}} = \sum_{ij} f(\mathbf{X}_{ij}) (\mathbf{e}_i^\top \mathbf{e}_j - \log \mathbf{X}_{ij})^2$$

$\mathbf{X} \in \mathbb{R}^{2n \times 2n}$ is the pairwise gene co-occurrence matrix for our dataset, where \mathbf{X}_{ij} 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 \mathbf{e}_i and \mathbf{e}_j are updated to match the gene co-occurrence in our dataset. $f(\cdot)$ is a weighting function as previously described to avoid overweighing the most common co-occurrence pairs¹⁰. We found that GloVe embeddings perform better than Gene2Vec embedding models³³. The embedding model is trained for 10,000 epochs with a batch size of 60. The Adam optimizer was used with a learning rate of 5×10^{-5} . Pre-trained genetic embedding results can be found Extended Data Fig. 5. This method of using multimodal datasets can be extended to other clinical or imaging modalities, such as patient demographics or preoperative/intraoperative magnetic resonance imaging.

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 pre-trained visual features using an attention mechanism. Similarly to bi-directional masked language modeling in BERT-style pre-training¹², 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. After the training of our visual encoder $f(\cdot)$ using supervised contrastive learning, the weights are fixed, and a multi-layer perceptron (MLP) that contains a single linear layer is added and trained for multi-label classification.

$$\hat{\mathbf{y}}_\ell = \text{MLP}_\ell(f(\mathbf{x})) = \sigma((\mathbf{W}_\ell \cdot f(\mathbf{x})) + \mathbf{b}_\ell) \quad (3)$$

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

$$\mathcal{L}(\mathbf{y}, \hat{\mathbf{y}}) = \sum_{\ell=1}^{|\mathcal{L}|} \lambda_\ell [\mathbf{y}_\ell \log(\hat{\mathbf{y}}_\ell) + (1 - \mathbf{y}_\ell) \log(1 - \hat{\mathbf{y}}_\ell)] \quad (4)$$

Transformer-based strategy. A transformer encoder is used that includes our pre-trained genetic embedding layer \mathbf{W}_ℓ . The labels $[\ell_j, \dots, \ell_k]$ are embedded such that $\mathbf{e}_k = \mathbf{W}_\ell \cdot \ell_k$, where the k -th column of \mathbf{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)³⁴. The image feature vector \mathbf{z} and embedded genetic labels are concatenated and input into the transformer encoder:

$$\mathbf{H} = [\mathbf{h}_j, \dots, \mathbf{h}_k] = \text{MultiHeadAttention}([\mathbf{z}, \mathbf{e}_j, \dots, \mathbf{e}_k]) \quad (5)$$

where $\mathbf{H} = [\mathbf{h}_j, \dots, \mathbf{h}_k]$ is 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 done in conventional transformer architectures^{11,34,35}, we enforce that the output latent space of the transformer encoder is the same as the pre-trained genetic embedding space, such that:

$$\hat{\mathbf{y}} = \sigma(\text{diag}(\mathbf{H}\mathbf{W}_\ell^\top)) \quad (6)$$

where $\mathbf{H}\mathbf{W}_\ell^\top$ is in $\mathbb{R}^{\ell \times \ell}$ 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. 5.

Whole-slide segmentation, patient inference and molecular subgrouping

Patch-based image classification requires an 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 non-diagnostic tissue using an argmax operation. 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 re-normalized 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:

$$p^{\text{patient}}(\mathbf{y}|x) = \frac{1}{Z} \sum_{j=1}^{|\mathcal{X}|} \mathbb{1}(\arg \max p(\mathbf{y}_j | \mathbf{x}_j, \phi) = k_{\text{tumor}}) p(\mathbf{y}_j | \mathbf{x}_j, \theta) \quad (7)$$

where x is a set of patches from a patient; \mathbf{x}_j is the j -th patch; $p(\mathbf{y}_j | \mathbf{x}_j, \phi)$ is the patch output from the tumor segmentation model f_ϕ ; $p(\mathbf{y}_j | \mathbf{x}_j, \theta)$ is the DeepGlioma g_θ output; and $Z = \sum_{j=1}^{|\mathcal{X}|} \mathbb{1}(\arg \max p(\mathbf{y}_j | \mathbf{x}_j, \phi) = k_{\text{tumor}})$ is the number of patches classified as tumor. Mutually exclusive molecular subgroup prediction is achieved algorithmically from the above patient-level molecular genetic predictions $p^{\text{patient}}(\mathbf{y}|x)$, as shown in Algorithm 1.

Algorithm 1 DeepGlioma patient-level molecular subgroup prediction

Require: $p^{\text{patient}}(\mathbf{y}|x), \tau, \psi, \epsilon \triangleright \tau = 0.5, \psi = 1$ for DeepGlioma experiments
1 if $p(k_{\text{IDH}} | x) < \tau$ **then**
2 return "Glioblastoma, IDH-wild-type"
3 else if $p(k_{\text{IDH}} | x) \geq \tau$ & $\frac{p(k_{\text{1p19q}} | x)}{p(k_{\text{ATRX}} | x) + \epsilon} > \psi$ **then**
4 return "Oligodendroglioma, IDH-mutant and 1p19q co-deleted"
5 else if $p(k_{\text{IDH}} | x) \geq \tau$ & $\frac{p(k_{\text{1p19q}} | x)}{p(k_{\text{ATRX}} | x) + \epsilon} \leq \psi$ **then**
6 return "Astrocytoma, IDH-mutant"
7 end if

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 multi-label 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 per set) that contained a balanced number of IDH mutant ($n = 10$) and wild-type ($n = 10$) tumors. Results are shown in Extended Data Fig. 6. For (1), we trained a ResNet-50 model using conventional cross-entropy versus a weakly supervised patch-based contrastive learning, or PatchCon. Both models were initialized using ImageNet pre-trained weights and trained for ten epochs without additional hyperparameter tuning. For (2), the PatchCon pre-trained ResNet model from (1) was held fixed, and we trained a single linear classification layer versus a transformer model with three multi-headed attention layers. Each model was trained for ten epochs. For (3), only the transformer model was re-trained 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. 6.

Molecular heat map generation

Leveraging our previous work on semantic segmentation of SRH images^{6,26}, 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 heat map for both molecular genetic and molecular subgroup predictions. The tumor segmented regions are retained, and the normal/non-diagnostic regions are converted to grayscale to indicate that these regions were not candidates for molecular prediction. Each molecular genetic heat map is generated by averaging the output predictions from patches that overlap for any given pixel in the heat map. Molecular subgroup heat maps 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 heat map. The most straightforward molecular subgroup heat map is for glioblastoma, IDH-wild-type heat map, generated as:

$$p_{ij}^{\text{GBM}} = 1 - p_{\text{IDH}}(\mathbf{x}_{ij}) \quad (8)$$

such that ij corresponds to the whole-slide height and width indices, and $p_{\text{IDH}}(\mathbf{x}_{ij})$ 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 heat maps are generated as:

$$p_{ij}^{\text{Oligo.}} = \frac{[p_{\text{IDH}}(\mathbf{x}_{ij}) > \tau \wedge p_{\text{1p19q}}(\mathbf{x}_{ij}) > \phi]}{\text{Conditional molecular mask}} \cdot p_{\text{IDH}}(\mathbf{x}_{ij}) \quad (9)$$

with the binarized conditional molecular mask identifying heat map regions that are above hyperparameter thresholds τ and ϕ for IDH and 1p19q co-deletion, respectively. Molecular astrocytomas heat maps are generated as:

$$p_{ij}^{\text{Astro.}} = \frac{[p_{\text{IDH}}(\mathbf{x}_{ij}) > \tau \wedge [p_{\text{1p19q}}(\mathbf{x}_{ij}) < \phi \vee p_{\text{ATRX}}(\mathbf{x}_{ij}) > \pi]]}{\text{Conditional molecular mask}} \cdot p_{\text{IDH}}(\mathbf{x}_{ij}) \quad (10)$$

where τ, ϕ, π are all hyperparameter thresholds. All thresholds were set to 0.5 in our model without hyperparameter tuning to avoid overfitting. 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 heat maps can be found in Extended Data Figs. 9 and 10. Molecular heat maps allowed for the evaluation of classification performance in different molecular settings. For example, DeepGlioma was able to correctly predict IDH-wild-type status in patients with recurrent mutations found in molecular glioblastomas, such as CDKNA1 and TERT promotor mutations (Supplementary Fig. 1). Molecular heat maps were also used to identify sources of DeepGlioma’s classification errors. Potential sources, including low tumor infiltration and image quality, are presented in Supplementary Fig. 2. Interactive web-based interface for DeepGlioma predictions can be found at <https://deepglioma.mlins.org/>.

Prospective multicenter clinical testing and sample size calculation

We elected to perform prospective, international, multicenter clinical testing of DeepGlioma to adhere to the rigorous standards of responsible machine learning in healthcare³⁶. Our prospective clinical testing was designed using the same principles as a non-inferiority diagnostic clinical trial²⁶. 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 ($\frac{\text{sensitivity} + \text{specificity}}{2}$) for diffuse glioma IDH mutational status. The control arm was conventional first-line

laboratory molecular screening, and the experimental arm was DeepGlioma predictions. IDH-1 IHC for somatic mutations at residue R132H is the most common first-line molecular diagnostic screening test. DeWitt et al.¹³ performed the largest and most clinically representative analysis of IDH mutation detection via IHC and sequencing methods and determined that IDH1-R132H IHC has a balanced diagnostic accuracy of 91.4% for adult-type diffuse gliomas (see Fig. 2c for contingency tables). 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 a 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). Most patients in our prospective cohort did not undergo both IHC and sequencing; therefore, an accuracy value cannot be calculated for this group.

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^{6,26}. End-to-end representation learning and classification can yield patch-based classification results that approach pathologist-level performance for histologic brain tumor classification. However, our early experiments on molecular classification indicated that contrastive pre-training and label embedding were advantageous for multi-label classification. Therefore, as a secondary endpoint, the control arm was established by training a ResNet-50 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. Our experimental arm was DeepGlioma molecular subgroup predictions as shown in Algorithm 1. Secondary endpoint metric was overall multiclass classification accuracy (Fig. 2d).

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 post-processed image patches were saved as 16-bit TIFF images and handled using the tiff file 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. CNNs/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 pre-trained ResNet-50 model from torchvision (0.10.10). 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 at our GitHub page (<https://github.com/MLNeurosurg/deepglioma>).

Reporting Summary

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

Data availability

The genomic data for training the genetic embedding model are publicly available through the above-mentioned data repositories, and all genomic data are provided in Supplementary Data Table 2. Institutional review board (IRB) approval was obtained from all participating institutions for SRH imaging and data collection. Restrictions apply to the availability of 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. Contact the corresponding author (T.H.) for any requests for data sharing. All requests will be evaluated based on institutional and departmental policies to determine whether the data requested are subject to intellectual property or patient privacy obligations. Data can be shared only for non-commercial academic purposes and will require a formal material transfer agreement. Generally, all such requests for access to SRH data will be responded to within 1 week.

Code availability

All code was implemented in Python (version 3.8) using PyTorch (1.9.0) as the primary machine learning framework. The following packages were used for complete data analysis: pydicom (2.0.0), tiff file (2021.1.14), torchvision (0.10.10), scikit-learn (1.0.1), pandas (1.3.4), numpy (1.20.3), matplotlib (3.5.0), scikit-image (0.18.3) and opencvpython (4.6.0.66). For data visualization and scientific plotting, the following packages were used: R (3.5.2) packages ggplot2 (3.3.5), dplyr (2.1.1) and tidyverse (1.3.1). All code and scripts to reproduce the experiments of this paper are available on GitHub at <https://github.com/MLNeurosurg/deepglioma>.

References

- Frome, A. et al. DeViSE: a deep visual-semantic embedding model. in *Advances in Neural Information Processing Systems* (eds Burges, C. J. C., Bottou, L., Welling, M., Ghahramani, Z. & Weinberger, K. Q.) 26 (Curran Associates, 2013).
- Ramesh, A. et al. Zero-shot text-to-image generation. in *Proceedings of the 38th International Conference on Machine Learning* (eds Meila, M. & Zhang, T.) 8821–8831 (PMLR, 2021).
- Saharia, C. et al. Photorealistic text-to-image diffusion models with deep language understanding. Preprint at *arXiv* <https://doi.org/10.48550/arXiv.2205.11487> (2022).
- Radford, A. et al. Learning transferable visual models from natural language supervision. in *Proceedings of the 38th International Conference on Machine Learning* Vol. 139 (eds. Meila, M. & Zhang, T.) 8748–8763 (PMLR, 2021).
- Freudiger, C. W. et al. Label-free biomedical imaging with high sensitivity by stimulated Raman scattering microscopy. *Science* **322**, 1857–1861 (2008).
- Freudiger, C. W. et al. Stimulated Raman scattering microscopy with a robust fibre laser source. *Nat. Photonics* **8**, 153–159 (2014).
- Hollon, T. C. et al. Rapid intraoperative diagnosis of pediatric brain tumors using stimulated Raman histology. *Cancer Res.* **78**, 278–289 (2018).
- Hollon, T. C. et al. Rapid, label-free detection of diffuse glioma recurrence using intraoperative stimulated Raman histology and deep neural networks. *Neuro Oncol.* **23**, 144–155 (2021).
- He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). <https://doi.org/10.1109/CVPR.2016.90> (2016).
- Jiang, C. et al. Rapid automated analysis of skull base tumor specimens using intraoperative optical imaging and artificial intelligence. *Neurosurgery* **90**, 758–767 (2022).
- Zhao, Z. et al. Chinese Glioma Genome Atlas (CGGA): a comprehensive resource with functional genomic data from Chinese glioma patients. *Genomics Proteomics Bioinformatics* **19**, 1–12 (2021).
- Zhang, J. et al. The International Cancer Genome Consortium Data Portal. *Nat. Biotechnol.* **37**, 367–369 (2019).
- Gusev, Y. et al. The REMBRANDT study, a large collection of genomic data from brain cancer patients. *Sci. Data* **5**, 180158 (2018).

32. Jonsson, P. et al. Genomic correlates of disease progression and treatment response in prospectively characterized gliomas. *Clin. Cancer Res.* **25**, 5537–5547 (2019).
33. Du, J. et al. Gene2vec: distributed representation of genes based on co-expression. *BMC Genomics* **20**, 82 (2019).
34. Lanchantin, J., Wang, T., Ordonez, V. & Qi, Y. General multi-label image classification with transformers. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). <https://doi.org/10.1109/cvpr46437.2021.01621> (2021).
35. Dosovitskiy, A. et al. An image is worth 16×16 words: transformers for image recognition at scale. Preprint at *arXiv* <https://doi.org/10.48550/arXiv.2010.11929> (2020).
36. Wiens, J. et al. Do no harm: a roadmap for responsible machine learning for health care. *Nat. Med.* **25**, 1337–1340 (2019).

Acknowledgements

The results presented here are, in whole or in part, based upon data generated by the TCGA Research Network: <https://www.cancer.gov/tcga>. We would like to thank K. Eddy, L. Wang and A. Marshall for providing technical support and T. Cichonski for editorial assistance. This work was supported by the following: grants NIH R01CA226527 (D.A.O.), NIH/NIGMS T32GM141746 (C.J.) and NIH K12 NS080223 (T.C.H.). It was also supported by the Cook Family Brain Tumor Research Fund (T.C.H.), the Mark Trauner Brain Research Fund, the Zenkel Family Foundation (T.C.H.), Ian's Friends Foundation (T.C.H.) and the UM Precision Health Investigators Awards grant program (T.C.H.). This research was also supported, in part, through computational resources and services provided by Advanced Research Computing, a division of Information and Technology Services at the University of Michigan.

Author contributions

T.H., C.J., A.C., C.F. and D.O. contributed to the conceptualization, study design and analysis of results. T.H., C.J., A.C., A.K. and M.N.-M.

contributed to the experimentation, acquisition, analysis and interpretation of data. T.H., C.J., A.C., M.N.-M., A.K., A. Aabedi and A. Adapa contributed to generating the figures and tables for the manuscript. T.H., W.A.-H., J.H., O.S., L.I.W., G.W., V.W., D.R., N.V., M.B., S.H.-J., J.G. and D.O. contributed to obtained tissue for SRH imaging. M.S. and S.C.-P. provided pathologic evaluation of tissue. All authors were involved in the editing, analysis and review of all data and manuscript versions.

Competing interests

T.H., C.F. and D.O. are shareholders in Invenio Imaging, Inc. C.J., A.C., M.N.-M., A.K., A. Aabedi, A. Adapa, W.A.-H., J.H., O.S., P.L., M.C., L.I.W., G.W., V.N., D.R., N.V.S., M.B., S.H.-J., J.G., M.S., S.C.-P. and H.L. do not have any competing interests.

Additional information

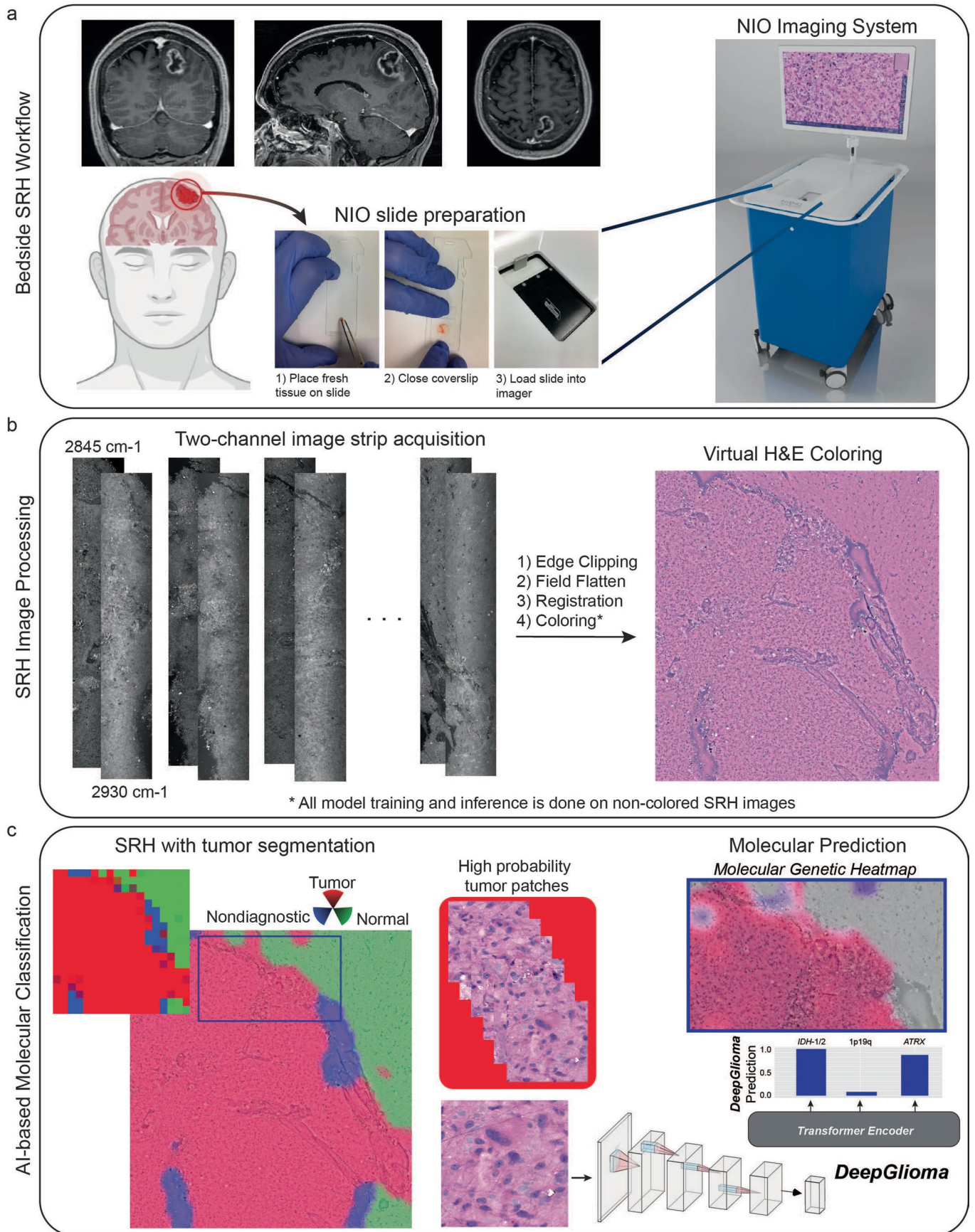
Extended data is available for this paper at <https://doi.org/10.1038/s41591-023-02252-4>.

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41591-023-02252-4>.

Correspondence and requests for materials should be addressed to Todd Hollon.

Peer review information *Nature Medicine* thanks Anant Madabhushi, Stephen Yip and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Primary handling editor: Ulrike Harjes, in collaboration with the *Nature Medicine* team.

Reprints and permissions information is available at www.nature.com/reprints.

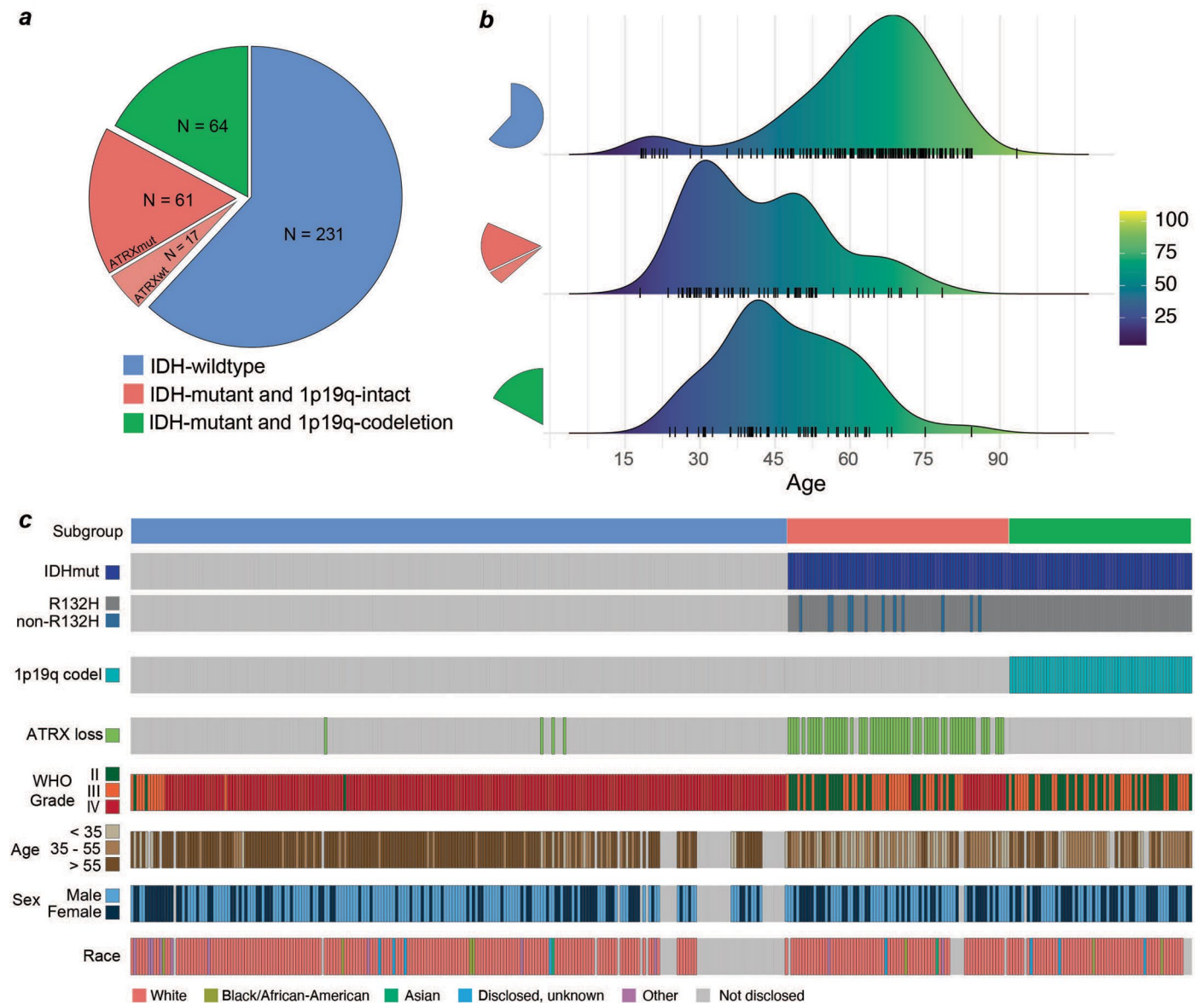


Extended Data Fig. 1 | See next page for caption.

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

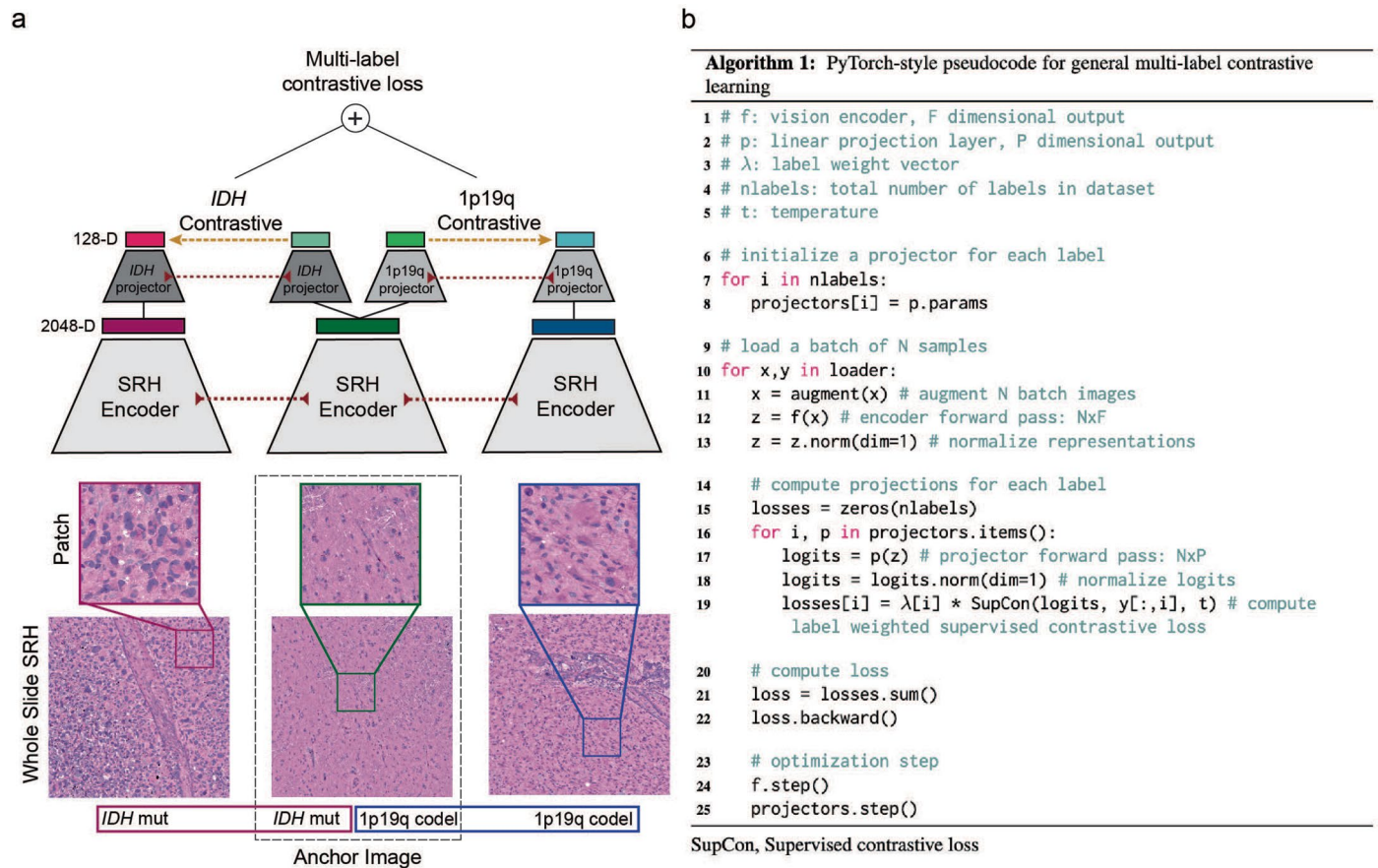
DeepGlioma. a, DeepGlioma for molecular prediction is intended for patients with clinical and radiographic evidence of a diffuse glioma who are undergoing surgery for tissue diagnosis and/or tumor resection. The surgical specimen is sampled from the patient's tumor and directly loaded into a premade, disposable microscope slide with an 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 image sizes range from $1\text{-}5\text{ mm}^2$

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, which are then used for both DeepGlioma training and inference. Additionally, whole slide images can be colored using a custom virtual H&E color scheme for review by the surgeon or pathologist [5]. **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 [25]. The tumor patches are then used by DeepGlioma at both training and inference to predict the molecular status of the patient's tumor.



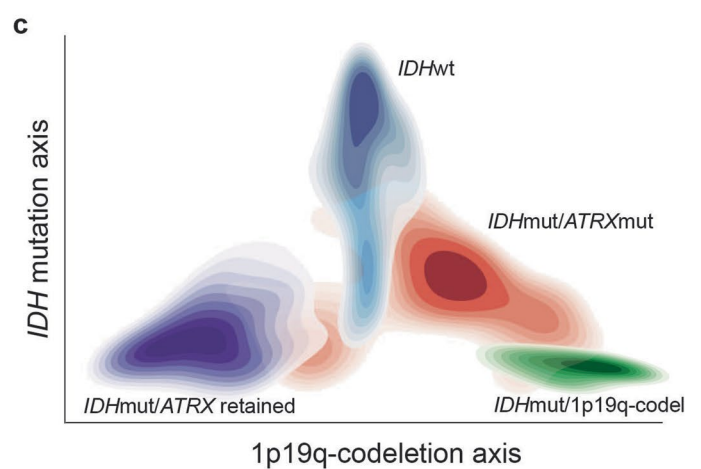
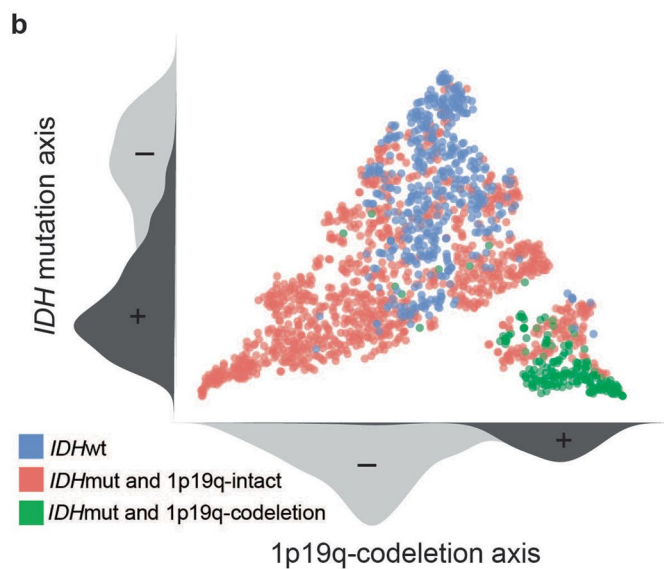
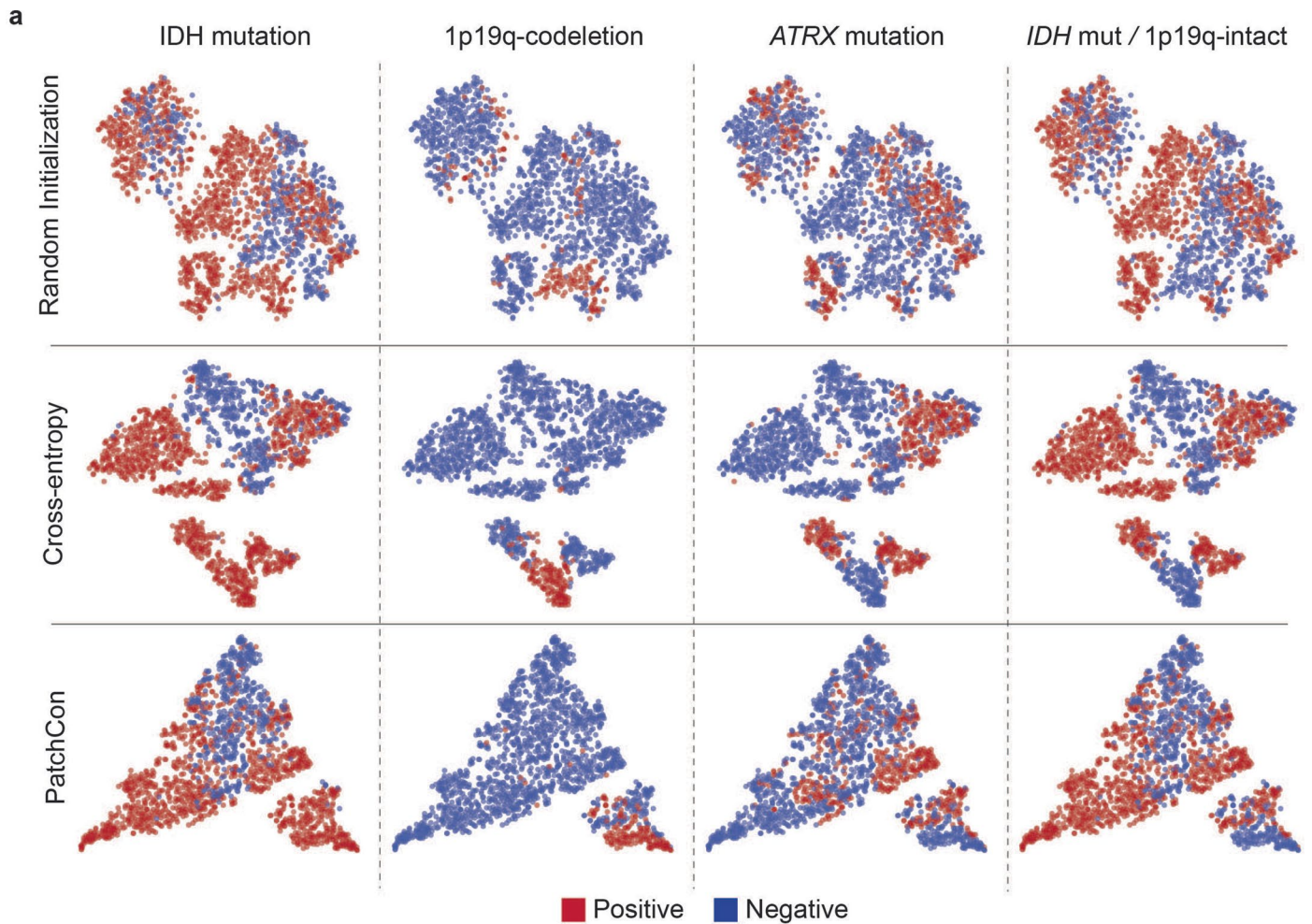
Extended Data Fig. 2 | Training dataset. The UM adult-type diffuse gliomas dataset used for model training. The UM training set consisted of a total of 373 patients who underwent a biopsy or brain tumor resection. Dataset generation occurred over a 6-year period, from November 2015 through November 2021. **a**, The distribution of patients by molecular subgroup. 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). Our dataset distribution of molecular subgroups is consistent with reported distributions in large-scale population studies. ATRX mutations were found in the majority of IDH-mutant/1p19q-intact patients (78%), also concordant with previous studies [9]. **b**, The age distribution for

each of the molecular subgroups are shown. The average age of IDH-wildtype patients was 62.6 ± 15.4 years and IDH-mutant patients was 44.6 ± 13.8 years. The average patient age of IDH-mutant/1p19q-codeleted group was 47.0 ± 12.9 years, and that of IDH-mutant/1p19q-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 [9] and is a major motivation for the present study.



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 [7]. 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. Unfortunately, no straightforward extension of existing contrastive loss functions, such as InfoNCE and NT-Xent [7], 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.



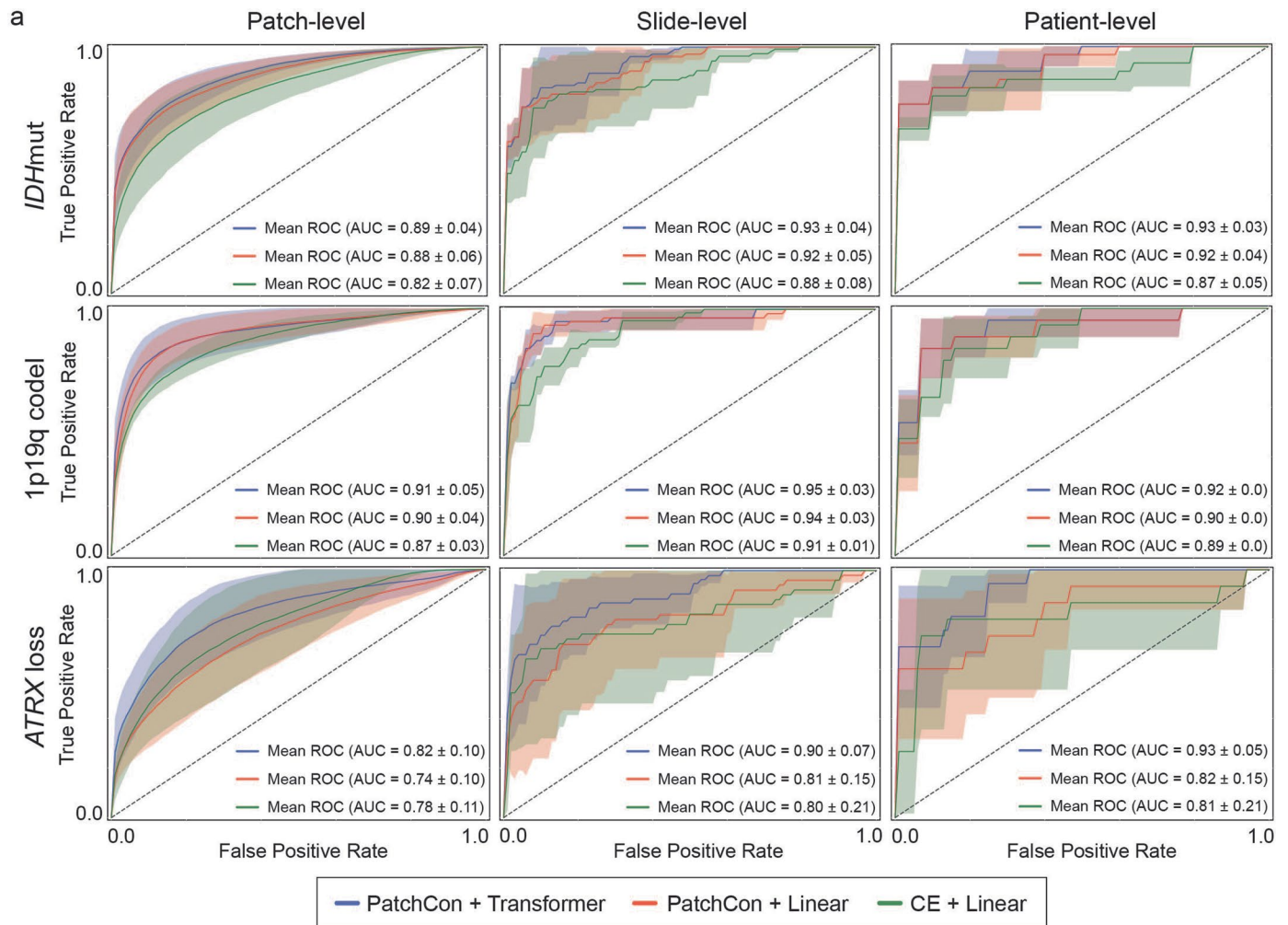
Extended Data Fig. 4 | See next page for caption.

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. **b**, Qualitative analysis of the SRH patch embeddings indicates that data are 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 the prospective testing cohort shown in Fig. 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 [10]. **a**, A subset of the most common mutations in diffuse gliomas is shown in the co-occurrence matrix. **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). Radial traversal of the embedding space around the wildtype genes defines clinically meaningful linear substructures [10] 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. **g**, 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 [34], 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.



b

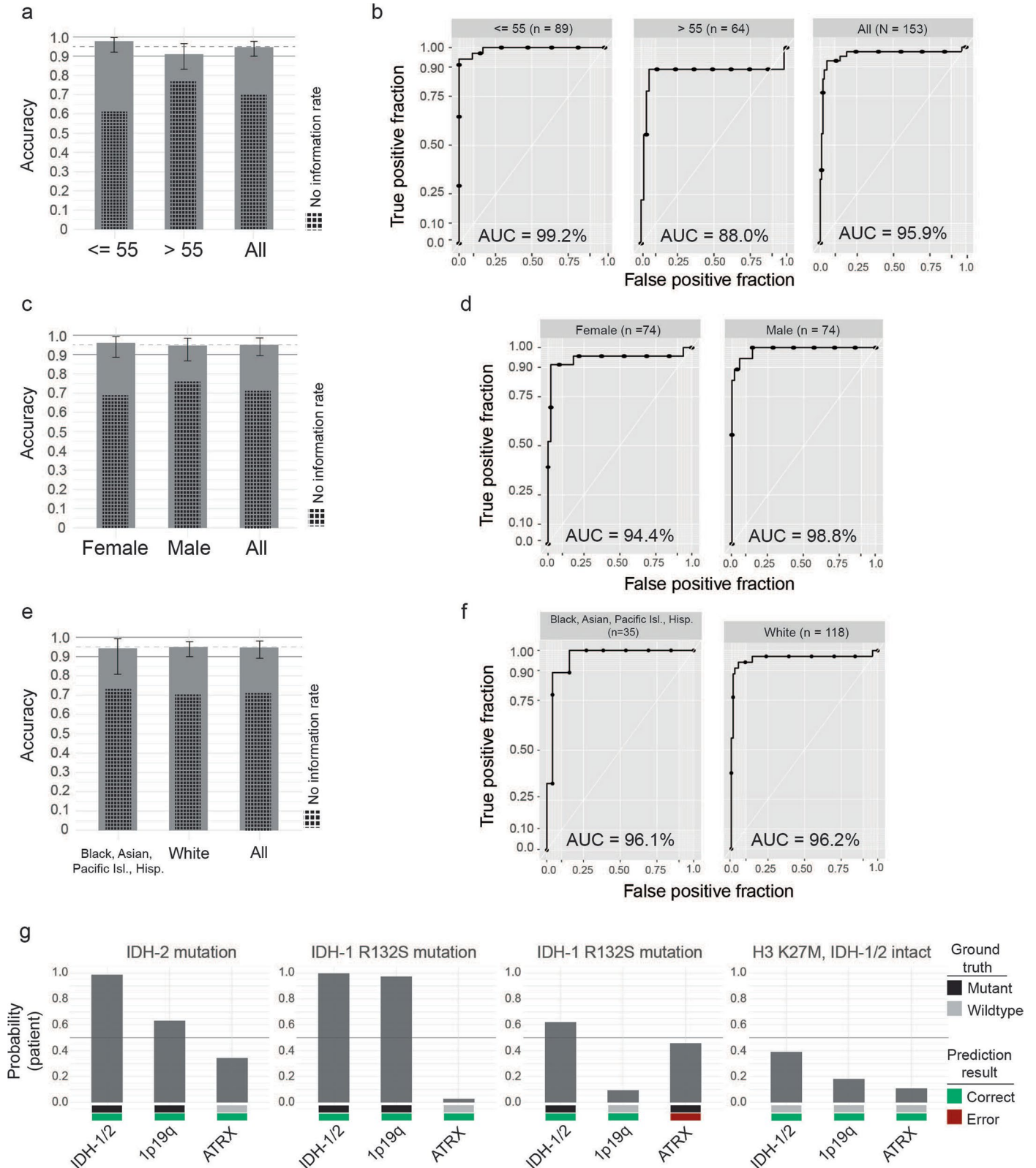
Level	% masked	Randomly initialized embedding					
		mAcc	mAP	mAUC	SubAcc	ebF1	micF1
Patch	100%	79.9 (1.4)	68.7 (6.7)	86.9 (1.8)	59.2 (2.0)	53.1 (1.4)	68.0 (1.7)
	66%	80.7 (0.2)	69.2 (2.1)	87.6 (0.1)	60.4 (0.3)	51.9 (1.3)	68.0 (1.0)
	33%	75.6 (1.4)	39.8 (0.8)	64.7 (1.0)	38.6 (12)	34.8 (4.7)	46.3 (7.5)
Slide	100%	87.4 (2.8)	86.8 (3.4)	94.9 (0.2)	72.7 (4.9)	66.5 (5.6)	77.1 (4.5)
	66%	88.0 (0.2)	90.0 (3.1)	95.2 (0.3)	75.3 (1.1)	66.4 (2.2)	78.0 (0.6)
	33%	76.9 (2.3)	40.4 (2.8)	61.2 (2.8)	42.9 (13)	31.4 (0.6)	43.0 (4.2)
Patient	100%	86.1 (2.8)	92.1 (1.0)	93.2 (1.0)	73.3 (4.7)	67.7 (6.2)	76.1 (5.4)
	66%	87.8 (0.8)	93.8 (0.2)	93.9 (0.9)	76.7 (2.4)	67.2 (2.4)	79.3 (1.0)
	33%	75.0 (1.4)	46.1 (2.5)	60.7 (3.1)	38.3 (14)	34.7 (4.5)	45.5 (7.9)
Level	% masked	Pretrained genetic embedding					
		mAcc	mAP	mAUC	SubAcc	ebF1	micF1
Patch	100%	85.7 (0.2)	68.8 (0.6)	86.2 (0.8)	69.2 (0.1)	58.4 (0.3)	73.2 (0.1)
	66%	84.9 (0.2)	68.8 (0.4)	85.6 (0.8)	67.9 (0.5)	57.7 (2.6)	72.5 (1.3)
	33%	85.9 (0.1)	66.3 (0.9)	85.6 (0.9)	66.8 (1.0)	59.3 (0.8)	72.9 (0.4)
Slide	100%	91.2 (0.4)	89.4 (1.1)	95.2 (0.4)	79.2 (1.1)	68.0 (1.0)	81.4 (1.1)
	66%	91.6 (0.2)	87.6 (2.1)	95.0 (0.6)	80.5 (1.1)	71.7 (2.7)	82.9 (0.3)
	33%	91.2 (0.2)	85.3 (2.7)	95.0 (0.3)	79.2 (0.0)	73.8 (1.4)	81.7 (0.3)
Patient	100%	88.9 (0.8)	93.3 (0.5)	92.4 (0.9)	76.7 (2.4)	61.6 (1.4)	78.7 (1.8)
	66%	91.1 (0.8)	93.1 (0.1)	92.0 (0.2)	80.0 (0.0)	73.7 (2.9)	83.6 (1.7)
	33%	88.3 (0.0)	82.4 (7.6)	92.0 (0.4)	75.0 (0.0)	66.7 (0.0)	77.9 (0.6)

Extended Data Fig. 6 | See next page for caption.

Extended Data Fig. 6 | 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) randomly initialized versus pretrained genetic embedding. **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 ResNet 50 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 (\pm 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

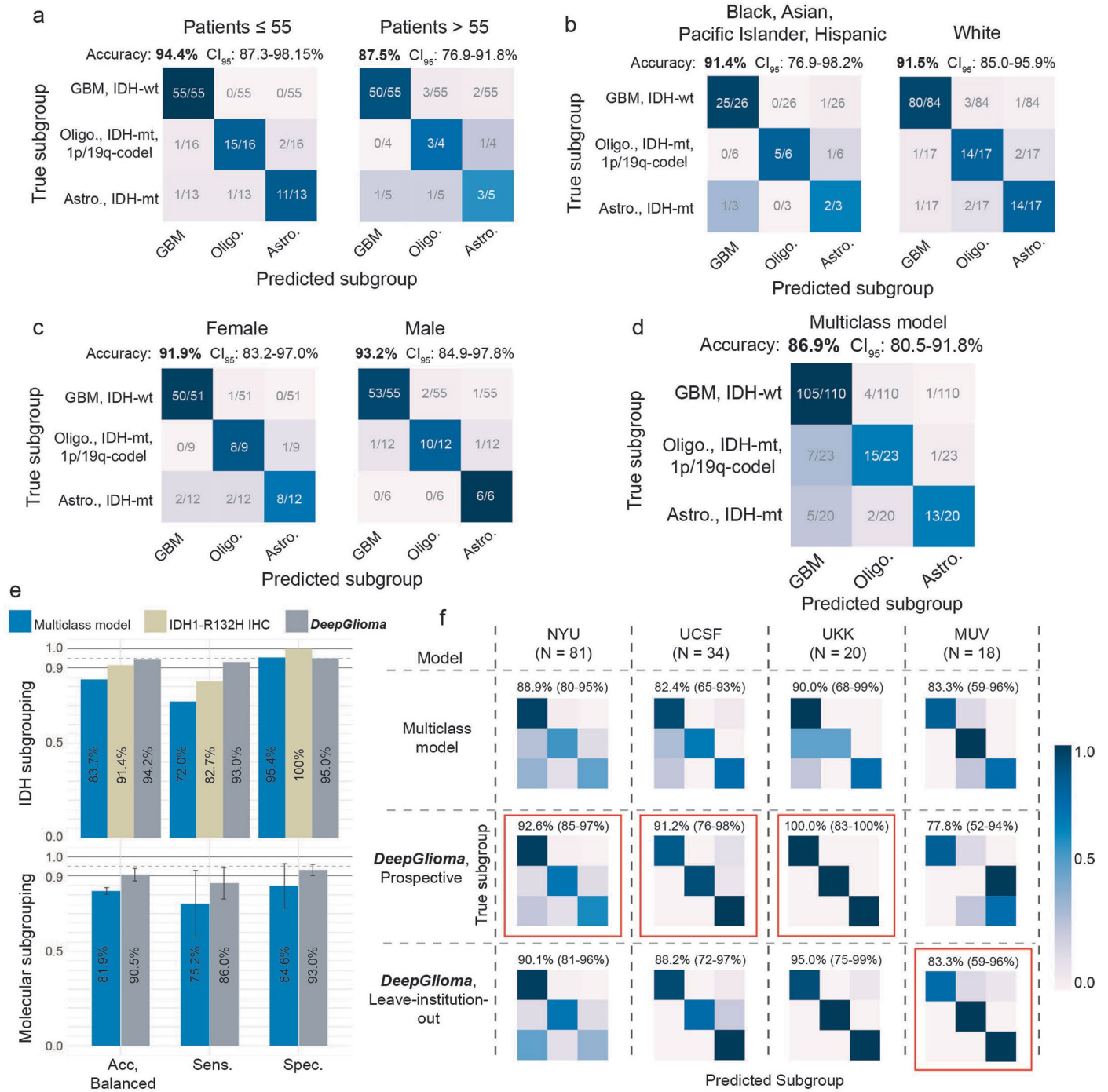
the linear classifier for multi-label classification. Note that the boost in performance of the transformer classifier over the linear model is due to the deep multi-headed attention mechanism learning conditional dependencies between labels in the context of specific SRH image features (i.e., not improved image feature learning because the SRH encoder weights are fixed). **b**, We then aimed to evaluate (3). A single ResNet50 model was trained using patchcon and the encoder weights were fixed for the following ablation study to isolate the contribution of random initialization versus pretraining of the genetic embedding layer. Three mask label training regimes were tested and are presented in the tables: all input labels masked (100%), two labels randomly masked (66%), and one label randomly masked (33%). The first row in the first table (100% masked) is non-multimodal training, where no genetic information is provided at any point during training or inference. We found that 66% input label masking, or randomly masking two of three diagnostic mutations, showed the best overall classification performance. 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. mAcc, mean label accuracy; mAP, mean average precision; mAUC, mean area under ROC curve; SubAcc, subset accuracy; ebF1, example based F1 score; micF1, micro-F1 score.



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

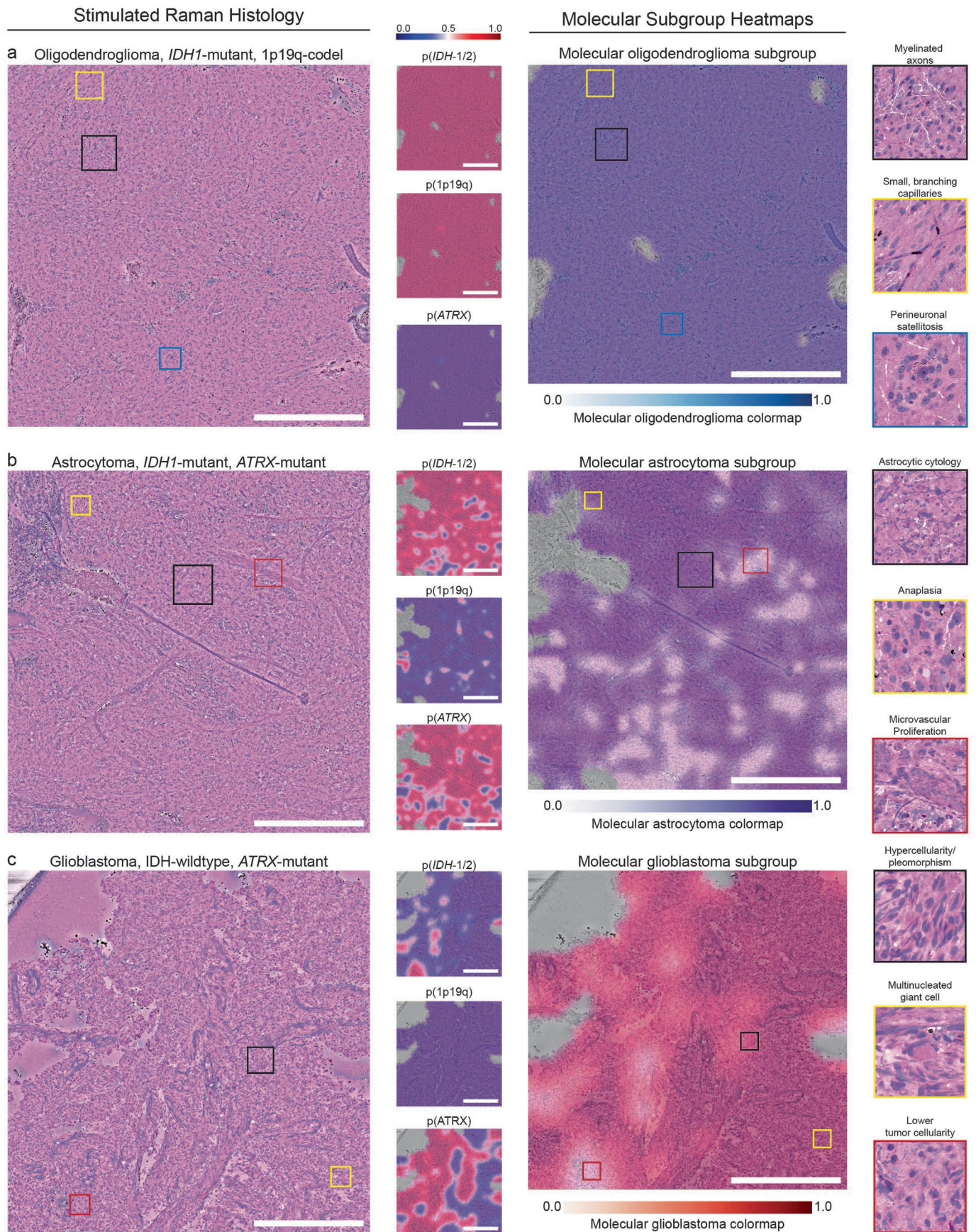
Extended Data Fig. 7 | Patient demographic subgroup analysis of DeepGlioma IDH classification performance. **a, b,** DeepGlioma performance for classifying IDH mutations stratified by patient age. Bar charts are showing patients classification accuracy (\pm standard deviation). Classification performance remains high in patients less than ($n = 89$) and greater than 55 years ($n = 64$). IDH mutations are less common in patients greater than 55 years, causing class imbalance and resulting in a greater proportional drop in classification performance with false negative predictions. **(c, d,)** Classification performance stratified by sex (male = 74, F = 74) and **(e, f)** racial groups (non-white = 35, white = 118) as defined by the National Institute of Health (NIH). Bar charts are

showing patients classification accuracy (\pm standard deviation). Classification performance remains high across all subgroup analyses. No information rate in the accuracy achieved by classifying all examples into the majority class. **g,** 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 pediatric-type diffuse high-grade gliomas, including diffuse midline glioma, H3 K27-altered, in a zero-shot fashion as these tumors 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.



Extended Data Fig. 8 | DeepGlioma molecular subgroup analysis. Multiclass classification performance for molecular subgroup prediction by DeepGlioma stratified by patient demographic information and prospective testing site is shown. Results stratified by (a) age, (b) race, and (c) sex are shown. Multiclass classification performance remained high in each patient demographic compared to the entire cohort. DeepGlioma was trained to generalize to all adult patients and to be agnostic to patient demographic information. d, Confusion matrix of our benchmark multiclass model trained using categorical cross-entropy. DeepGlioma outperformed the multiclass model by +4.6% in overall patient-level diagnostic accuracy with a substantial improvement in differentiating molecular astrocytomas and oligodendrogliomas. e, Direct comparison of subgrouping performance for our benchmark multiclass model,

IDH1-R132H IHC, and DeepGlioma. Performance metrics values are displayed. Molecular subgroupings mean and standard deviations are plotted for both IDH subgrouping and molecular subgrouping. These results provide evidence that multimodal training and multi-label prediction provide a performance boost over multi-class modeling. f, DeepGlioma molecular subgroup classification performance for each of the prospective testing medical centers is shown. Accuracy values with 95% confidence intervals (in parentheses) 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; however, some improvement was observed during the LIOCV experiments. Red indicates the best performance.

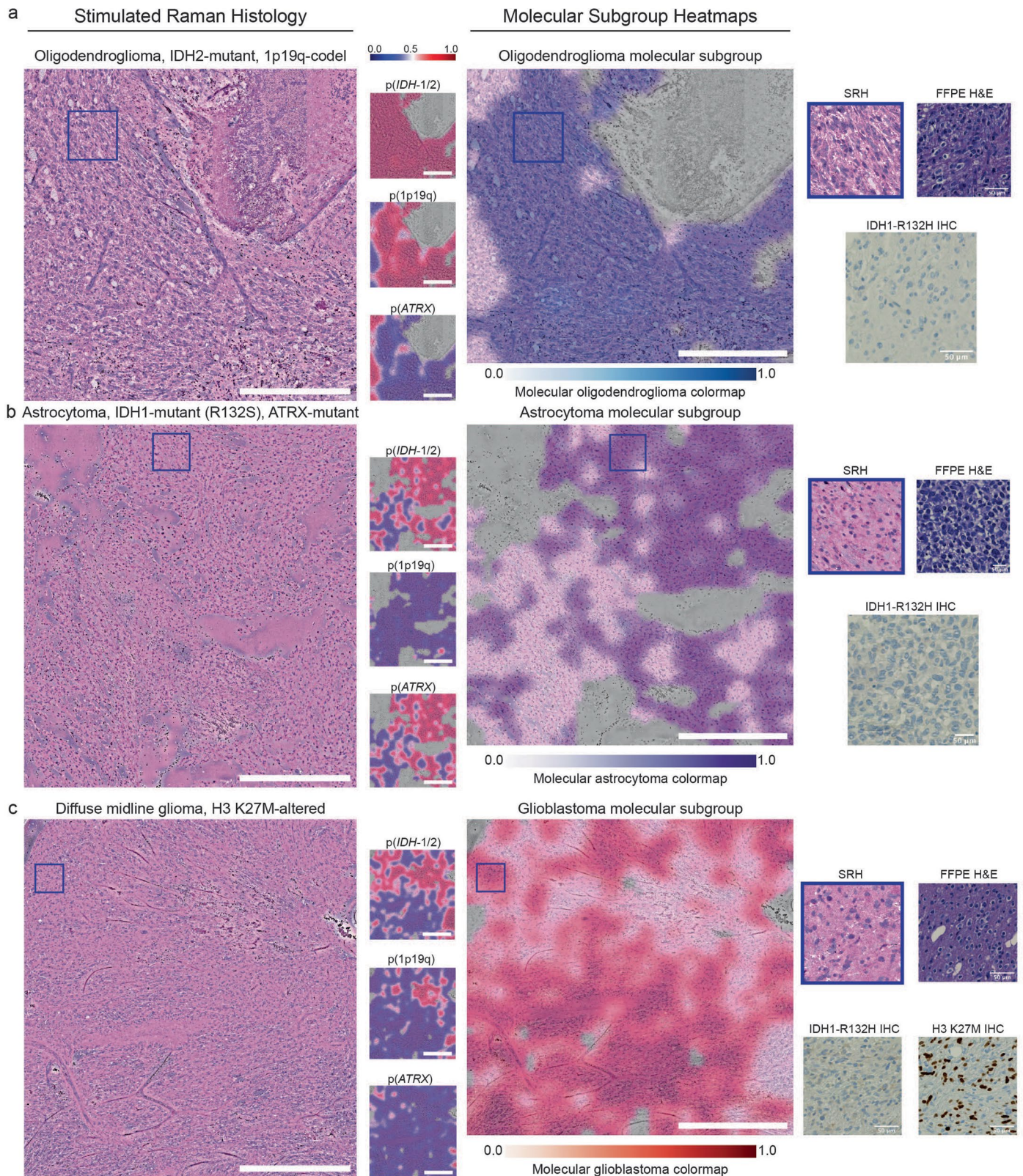


Extended Data Fig. 9 | See next page for caption.

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 probability 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 known to occur in IDH-wildtype tumors [9]. Despite the high co-occurrence 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 affect 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. Scale bar, 1 mm.



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 predicted 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 IDH1-R132S and ATRX mutations. DeepGlioma correctly identifies both mutations. **c.** A patient with a suspected adult-type diffuse glioma met inclusion criteria for the prospective clinical testing set. The 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 fashions. Scalebar, 1 mm.

Reporting Summary

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

Statistics

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

n/a | Confirmed

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

Our web collection on [statistics for biologists](#) contains articles on many of the points above.

Software and code

Policy information about [availability of computer code](#)

Data collection The software and code used in this study for development of DeepGlioma model is publicly available at <https://github.com/MLNeurosurg/deepglioma>. Proprietary software NIO (1.7.1) used in the NIO Imaging System (Invenio Imaging, Inc) was used for optical imaging.

Data analysis Data analysis was done using Python (version 3.8). The following packages were used for data analysis: pydicom (2.0.0), tifffile (2021.1.14), PyTorch (1.9.0), torchvision (0.10.10), scikit-learn (1.0.1), pandas (1.3.4), numpy (1.20.3), matplotlib (3.5.0), scikit-image (0.18.3), opencv-python (4.6.0.66). For data visualization and scientific plotting, we used R (3.5.2) packages ggplot2 (3.3.5), dplyr (2.1.1), and the tidyverse (1.3.1).

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

Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

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

Data availability statement is included in our manuscript. 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 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.) 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. Generally, all such requests for access to SRH data will be responded to within a week.

Human research participants

Policy information about [studies involving human research participants and Sex and Gender in Research](#).

Reporting on sex and gender	All reported findings apply to patients of any sex or gender.
Population characteristics	See 'Life sciences study design'. We have included all diffuse glioma patients >18. It includes all race and gender as well.
Recruitment	The patients are all brain tumor surgical candidates who had consented prior to surgery. Inclusion criteria for SRH imaging were patients who were undergoing surgery for (1) suspected central nervous system tumor and/or (2) epilepsy, (3) subject or durable power of attorney was able to provide consent, and (4) preoperative expectation that additional tumor tissue would be available beyond what is required for clinical pathologic diagnosis.
Ethics oversight	University of Michigan Institutional Review Board (HUM00083059)

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

Field-specific reporting

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

- Life sciences Behavioural & social sciences Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/documents/nr-reporting-summary-flat.pdf](https://www.nature.com/documents/nr-reporting-summary-flat.pdf)

Life sciences study design

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

Sample size	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. Our prospective clinical testing was designed using the same principles as a non-inferiority diagnostic clinical trial. We set the expected accuracy for both the control and experimental arms to be 91.4%, the equivalence limit was set to 10%, power to 90%, and alpha to 0.05%, resulting in a 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).
Data exclusions	Exclusion criteria for SRH imaging was (1) insufficient diagnostic tissue as determined by surgeon or pathologist, (2) grossly inadequate tissue (e.g. hemorrhagic, necrotic, fibrous, liquid, etc.), and (3) SRH imager malfunction. Exclusion criterion for DeepGlioma prediction was less than 10% area segmented as tumor by our trained SRH segmentation model.
Replication	We performed a leave-institution-out cross-validation, such that our methods were replicated across each institution individually. We were able to replicate DeepGlioma performance results across each medical institution.
Randomization	Randomization was not relevant for this study because patients were prospectively enrolled into the clinical testing cohort. There was no experimental vs control arms for randomization. Surgical specimens imaged in the operating room were selected by the clinician and the remainder of the clinical brain tumor specimen was sent for final pathologic analysis, including molecular classification.
Blinding	Blinding was not relevant for this study. Patients in our prospective testing cohort were included prior to brain tumor surgery. Several of the study authors (T.C.H, S.H-J., D.A.O, M.B., V.N., G.W., W.A-H., J.H., O.S., L.W.) were the treating surgeons for included patients and knowledge of their molecular diagnosis was essential for clinical treatment.

Reporting for specific materials, systems and methods

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

Materials & experimental systems

n/a	Included in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> Antibodies
<input checked="" type="checkbox"/>	<input type="checkbox"/> Eukaryotic cell lines
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology and archaeology
<input checked="" type="checkbox"/>	<input type="checkbox"/> Animals and other organisms
<input checked="" type="checkbox"/>	<input type="checkbox"/> Clinical data
<input checked="" type="checkbox"/>	<input type="checkbox"/> Dual use research of concern

Methods

n/a	Included in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input checked="" type="checkbox"/>	<input type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging