Biologically-informed deep neural networks provide quantitative assessment of intratumoral heterogeneity in post-treatment glioblastoma

Hairong Wang[1*], Michael G Argenziano[2*], Hyunsoo Yoon[3], Deborah Boyett[2], Akshay Save[2], Petros Petridis[2,4], William Savage[2], Pamela Jackson[5], Andrea Hawkins-Daarud[5], Nhan Tran[6], Leland Hu[7], Osama Al Dalahmah[8], Jeffrey N. Bruce[2], Jack Grinband[9,10], Kristin R Swanson[5*], Peter Canoll[8*], Jing Li[1*]

[1]H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA, USA
[2]Department of Neurological Surgery, Columbia University Irving Medical Center, New York, NY, USA
[3]Department of Industrial Engineering, Yonsei University, Seoul, Republic of Korea
[4]Department of Psychiatry, New York University, New York, NY, USA
[5]Mathematical NeuroOncology Lab, Precision Neurotherapeutics Innovation Program, Mayo Clinic, Phoenix, AZ
[6]Department of Cancer Biology, Mayo Clinic, Phoenix, AZ
[7]Department of Radiology, Mayo Clinic, Phoenix, AZ
[8]Department of Pathology and Cell Biology, Columbia University Irving Medical Center, New York, NY, USA
[9]Department of Psychiatry, Columbia University Irving Medical Center, New York, NY, USA
[10]Department of Radiology, Columbia University Irving Medical Center, New York, NY, USA

Keywords: deep learning, recurrent glioblastoma, knowledge integration, intratumoral heterogeneity

Abstract

Intratumoral heterogeneity presents a major challenge to diagnosis and treatment of glioblastoma (GBM). Such heterogeneity is further exacerbated upon the recurrence of GBM, where treatment-induced reactive changes produce additional intratumoral heterogeneity that is ambiguous to differentiate on clinical imaging. There is an urgent need to develop non-invasive approaches to map the heterogeneous landscape of histopathological alterations throughout the entire lesion for each patient. We propose to predictively fuse MRI with the underlying intratumoral heterogeneity in recurrent GBM using machine learning (ML) by leveraging unique image-localized biopsies with their associated locoregional MRI features. To this end, we develop BioNet, a biologically informed multi-task framework combining Bayesian neural networks and semi-supervised adversarial autoencoders, to predict regional distributions of three tissue-specific gene modules: proliferating tumor, reactive/inflammatory cells, and infiltrated brain tissue. BioNet provides insight into how to integrate implicit and hierarchical domain knowledge, which is difficult to incorporate into ML models through existing methods. The proposed architecture further addresses challenges in exploiting latent feature structures from limited labeled image-localized biopsy samples, which lead to improvements in prediction accuracy. BioNet performs significantly better than existing methods on cross-validation and blind test datasets, shows generalizability that surpasses other models, and is adaptable to different types of data or tasks. Prediction maps of gene modules from BioNet provide accurate predictions of intratumoral heterogeneity, which can improve surgical planning and localization of diagnostic biopsies, as well as inform neuro-oncological treatment assessment for each patient. These results also highlight the emerging role of ML in precision medicine.

Significance Statement

Quantitative assessments of intratumoral heterogeneity are limited by sparse biopsy sampling but is crucial for diagnosis, clinical management and treatment of (recurrent) glioblastoma. We propose leveraging a unique cohort of image-localized biopsies and their associated locoregional imaging features to develop a deep learning model, BioNet, that takes as input patient MRIs to predict output maps of the regional distributions of tissue-states. BioNet is able to (1) amplify the signal to noise ratio of the intratumoral genetic and cellular heterogeneity and (2) augment the learning capability of deep learning (DL) models through integrating implicit, hierarchical, but hard to be mathematically formulated domain knowledge. Our method performs significantly better than existing methods and is able to be adapted to related diseases.

Introduction

Glioblastoma (GBM) exhibits pronounced intratumoral heterogeneity, which can confound diagnosis and clinical management, and is a leading driver of tumor recurrence recurrence (1, 2). Treatment-induced reactive changes futher exacerbate intratumoral heterogeneity (3, 4). Because histopathological and molecular analyses are limited by sparse surgical sampling, there is significant need to develop non-invasive approaches to map the heterogeneous landscape of histopathological alterations throughout the entire lesion in order to improve surgical targeting of confirmatory biopsies and non-invasive assessment of neuro-oncological treatment response. Radiogenomics is a growing research field, which seeks to develop machine learning (ML) models to predict cellular, molecular and genetic characteristics of tumors based on MRI and other imaging types (5–7). Radio(gen)omics methods have been shown to accurately predict not only diversity in tumor cell density associated with diffuse invasion into the brain parenchyma peripheral to the frank lesion seen on MRI (8–10), but also abnormalities in hallmark genes such as EGFR, PDGFRA, and PTEN (11–14), IDH mutation status (15–19), and MGMT methylation status based on radiographic features (14–16). These studies represent examples in which the prediction provides a single categorical label per tumor using imaging features that span the entire lesion.

However, precise representations of intratumoral heterogeneity require *voxel-wise* labels (e.g. image-localized biopsies) that reflect local or regional features of the tumor. A major challenge for such prediction is the lack of large image-localized biopsy datasets (20, 21) to train deep learning (DL) models that are well-known to be heavily-parameterized and data-hungry. Creation of large training datasets is limited by various factors such as the invasiveness and high expense of sample acquisition, need of highly-specialized experts to create accurate labels, rarity of some diseases in the population, and difficulty in patient recruitment (20). Moreover, the lack of large data sets has severely limited the number of studies focusing on *regional* tumor heterogeneity. A few studies have developed MRI-biology fusion models to predict cell density (22–27) or regional copy number variation of EGFR, PDGFRA, and PTEN in untreated, primary GBM (28, 29). But in recurrent GBM (recGBM), treatment-induced reactive changes lead to additional intratumoral heterogeneity and the related additional complexity in tissue composition makes prediction of tissue states more difficult (30), requiring even larger training data sets.

We have combined two novel computational approaches to address the problem of limited training data by (1) amplifying the signal to noise ratio of the genetic and cellular heterogeneity signal in the tissue by combining individual gene expression and cellular composition patterns into three *tissue states* and (2) augmenting the learning capability of DL models through integrating implicit, hierarchical, but hard-to-be-mathematically-formulated domain knowledge of three gene modules. The radiogenomic signal associated with individual genes is noisy, but it should be possible to improve prediction accuracy of any ML approach by taking advantage of clusters of correlated genes that represent different tissue states. We have identified three tissue-specific gene modules defined by cellular composition (proliferative/recurrent tumor, treatment-induced reactive/inflammatory cells, and infiltrated brain tissue) thereby reducing data complexity and improving prediction accuracy.

Second, integration of biological/biomedical domain knowledge, such as, biological principles, empirical models, simulations, and knowledge graphs, can provide a rich source of information (pseudo data) to help alleviate the data shortage in training DL models. Biologically- or biomedically-informed deep learning (BIDL) has been proposed to use domain knowledge to guide the design of DL architecture (31, 32), as attribution priors of features (33), to regularize the model predictions, coefficients, or latent feature representations (24, 34), and implicitly leveraging knowledge from other domains through transfer learning (35–38). However, these existing works lack the capability of integrating hierarchical domain knowledge which is hard to be described in mathematical formulation. To fully exploit information from this kind of domain knowledge, our work introduces a novel BIDL approach called BioNet, a multi-task DL framework composed of Bayesian neural networks and semi-supervised adversarial autoencoders, which enhances the learning capability of DL models that use small datasets for training. Specifically, our proposed framework acts like a protocol that describes conditions between different levels of the hierarchical relationship. For each level of the relationship, BioNet integrates the distributions of three gene modules via two mechanisms: using domain knowledge to create large virtual biopsy datasets to pre-train the

Bayesian neural networks and using domain knowledge to guide the design of novel loss functions to regularize the predictions from semi-supervised adversarial autoencoders.

In sum, by leveraging a unique cohort of image-localized biopsies in recGBM, BioNet (1) provides a method of integrating implicit and hierarchical domain knowledge into network design, effectively overcoming the barrier of limited biopsy data that impedes broader DL applications in biomedical fields, which results in (2) the first non-invasive MRI and DL approach able to quantitatively assess intratumoral heterogeneity and the histopathological and molecular alterations associated with proliferating tumor and reactive/inflammatory changes in the setting of post-treatment recGBM.

Results

Identification of three tissue-specific gene modules: recGBM_neuronal, recGBM_proliferative, recGBM_inflammatory. A total of 84 tissue biopsies were harvested from 31 patients (mean age = 56, 63% male). The majority of biopsies from Cohort A (48/84; 57%) had both RNA sequencing and immunohistochemical (IHC) staining for SOX2, CD68, Ki67, and NeuN. Hemotoxylin counterstain labeled all nuclei, providing a measure of total cell density. Correlation analysis between the normalized gene expression values and IHC stained cell counts identified 7779 statistically significant genes (p-value < 0.05). Hierarchical clustering determined three distinct clusters with mutually exclusive genes **(Fig. 1A)**. Gene module 1 (n=3688) were genes significantly positively correlated with the labeling indices of SOX2, Ki67, and total cell density; gene module 2 (n=1673) were genes strongly positively correlated with NeuN labeling index; gene module 3 (n=2418) were genes strongly positively correlated with CD68 labeling index **(Fig. 1A)**. Gene ontology analysis demonstrated that gene module 1 was associated with proliferation and cell cycle ontologies (recGBM_proliferative); module 2 was associated with neuronal signaling (recGBM_neuronal); and module 3 was associated with cytokine production and immune response (recGBM_inflammatory) (**Fig. 1B**). The complete list of genes for each gene cluster can be found in **Supplementary Table 1**. The three gene modules developed from the RNA-IHC clustering were applied to all 84 MRI-localized biopsies via GSVA, using the normalized PLATE-seq expression profile for each sample. Each MRI-localized biopsy had GSVA enrichment scores for each of the three gene modules (**Fig. 1C**), and were normalized as described in Methods**.**

Formulation of biologically-informed design principles for BioNet. The three gene modules have intrinsic biological relationships that reflect the tissue composition (**Fig. 2A**). BioNet leverages these relationships to improve the predictive strength of the ML model. We know that genes in the recGBM_neuronal module are enriched in normal brain and reduced in lesional brain tissue. Thus, samples that have high enrichment for the recGBM_neuronal gene signature likely have low enrichment for the gene signatures for recurrent tumor (recGBM_proliferative) and inflammatory response (recGBM_inflammatory). Furthermore, for samples with low enrichment of recGBM_neuronal, the lesional component of the tissue will be composed of some mixtures of proliferative tumor and inflammatory response, i.e., if a sample has more proliferative tumor, it is likely to have less inflammatory response, and vice versa, which means that enrichment scores of recGBM_proliferative and recGBM_inflammatory are negatively correlated. Our analysis of the MRI-localized biopsies provides empirical evidence for these relationships (**Fig. 2**). The biopsy samples are divided into two groups with above-zero ('high') and below-zero ('low') enrichment scores of recGBM_neuronal. Samples with high recGBM_neuronal scores have significantly lower average score of recGBM_proliferative and recGBM_inflammatory (**Fig. 2B)**, compared to samples with low recGBM_neuronal scores. Furthermore, for samples with low recGBM_neuronal scores there is a significant negative correlation between recGBM_proliferative and recGBM_inflammatory (**Fig. 2C**). These relationships, observed in the labeled biopsies, were assumed to exist in unsampled regions of the tumors.

The knowledge about the biological relationships between the three gene modules inspired the hierarchical design of BioNet (**Fig. 2D**). We first constructed a neural network (NN), BioNet_neu, to predict recGBM_neuronal. This model allowed us to stratify the large number of unlabeled samples into two groups with high and low predicted recGBM_neuronal scores. Within each stratum, there is different knowledge about the other two gene modules, recGBM_proliferative and recGBM_inflammatory (**Fig. 2A**). We constructed two additional NNs, BioNet_pro and BioNet_inf, that encode the knowledge in each stratum as a loss function to regularize the predictions on unlabeled samples while minimizing the prediction loss on

biopsy samples. In this way, we effectively integrated knowledge and unlabeled samples to augment biopsy/labeled samples to train robust models.

Construction of a neural network for recGBM_neuronal, BioNet_neu, with transfer learning and uncertainty quantification. We adopted several strategies to overcome the limitation of dimensionality and risk of overfitting in constructing BioNet_neu: (1) Pre-train a NN using a large number of virtual biopsy samples that are selected based on biological knowledge, and fine-tune the NN using real biopsy samples. This scheme is known as transfer learning (39). (2) Use Monte Carlo dropout (40) to create a Bayesian neural network, where dropout works as a Bayesian approximation of the posterior distribution of the parameters of the model. Also, the Bayesian neural network allows us to quantify the uncertainty of prediction for all the outputs by computing uncertainty measures in information theory, such as Predictive Entropy (PE) (41, 42). (3) Use a data augmentation strategy of including samples within a pre-defined neighborhood of each biopsy sample to enforce smoothness of gene module distributions.

BioNet_neu achieved an Area Under the Curve (AUC) of 0.77 based on patient Cohort A using 5-fold cross validation (**Fig. 3A**). Without Monte Carlo dropout, transfer learning and data augmentation, the AUC was reduced to 0.70, 0.64 and 0.56 (**Fig. 3A**), respectively. Details about BioNet_neu with data augmentation are provided in Section 1 of SI Appendix.
We further assessed the uncertainty quantification capability of BioNet_neu. This capability is deemed adequate if the model satisfies a "more certain more accurate (MCMA)" condition, i.e., the predictions that the model is more certain about are indeed more accurate. To check this, we computed the uncertainty measure PE for each biopsy sample, and further computed the accuracy on subsets/strata of samples above increasingly stringent certainty thresholds (**Fig. 3B**). **Fig. 3B** shows that the MCMA condition is indeed satisfied by BioNet_neu. For example, the accuracy increased to 0.84 when computed on the top 59% certain samples (corresponding to PE < 0.6).

Construction of neural networks for recGBM_proliferative and recGBM_inflammatory, BioNet_pro and BioNet_inf, with knowledge regularization and semi-supervised learning. Constructing BioNet_pro and BioNet_inf faces the same challenges as BioNet_neu. However, unlike BioNet_neu, we do not have biological knowledge to inform the gathering of a large virtual biopsy dataset to pre-train these models. We adopted different strategies to tackle the challenges: (1) knowledge regularization based on the biological relationships between recGBM_proliferative and recGBM_inflammatory, assuming that recGBM_neuronal is known (**Fig. 2A**), (2) semi-supervised learning (43) that integrates labeled and unlabeled data, (3) multi-task learning (44) that improves model generalization by sharing common information contained in these two closely related tasks, and (4) adversarial autoencoder (AAE) (45) that projects high-dimensional and noisy input feature spaces to robust latent spaces for two gene modules respectively. **Fig. 4** provides an overview of the architecture of BioNet_pro and BioNet_inf. The two models are trained together to minimize an overall loss function that combines the knowledge regularization (KR) loss, reconstruction loss, and prediction loss. The detailed model design is provided in Methods.

Finally, we want to point out an important strategy to ensure the training accuracy of the models. To regularize the predictions by the biological relationships (**Fig. 2A**) in the KR loss, two groups of unlabeled samples were stratified according to high and low predicted recGBM_neuronal scores. The quality of the KR loss largely relies on the accuracy of predicted recGBM_neuronal binary scales (i.e. high or low) for a large number of unlabeled samples. In order to guarantee the high accuracy of predicted recGBM_neuronal binary scales, we proposed to only include a subset of unlabeled samples in the KR loss whose predictions by BioNet_neu are above a certainty threshold. This subset is expected to have high accuracy by BioNet_neu because BioNet_neu satisfies the MCMA condition as demonstrated in a previous section (**Fig. 3D**). Making the certainty threshold more stringent increases the quality of the KR loss, as well as reduces the number of unlabeled samples. To strike a balance between these two aspects, the certainty threshold is treated as a tuning parameter for BioNet_pro and BioNet_inf.

Same as recGBM_neuronal, recGBM_proliferative and recGBM_inflammatory were dichotomized into a binary scale, 'high' vs. 'low'. BioNet_pro and BioNet_inf achieved 0.78 and 0.80 classification accuracy (ACC), 0.77 and 0.72 AUC based on leave-one-patient-out (LOPO) cross-validation (CV) using patient cohort A (**Fig. 5AB**). Each component of the models helped to achieve good performance. Without

knowledge regularization, the models were reduced to semi-supervised AAEs (semi-AAE), which have only 0.71 and 0.71 AAC, 0.62 and 0.59 AUC. Further excluding unlabeled samples, we tried three popular supervised learning models which are trained using labeled/biopsy samples alone, NN, Support Vector Regression (SVR), and Random Forest (RF). These models have only 0.54-0.61 and 0.54-0.68 range of ACC, 0.50-0.64 and 0.55-0.58 range of AUC. Additionally, we also computed the Mean Absolute Error (MAE) between the predicted and true scores of recGBM_proliferative and recGBM_inflammatory using their numerical scale. The MAE of BioNet_pro and BioNet_inf is smaller than the other methods except being slightly higher than semi-AAE for recGBM_proliferative.

BioNet as well as other methods under comparison were used to generate predictions of recGBM_proliferative and recGBM_inflammatory, respectively, for every unlabeled sample within the ROI, composing two prediction maps for each patient (**Fig. 6** gives an example of the prediction maps for one patient generated by BioNet). To assess the validity of the maps, we computed two knowledge concordance (KC) metrics, KC_{neu^+} and KC_{neu^-} **(Fig. 5C)**, corresponding to the knowledge that recGBM_proliferative and recGBM_inflammatory are likely to be low for unlabeled samples with high certainly predicted recGBM_neuronal scores, and that they are likely to be negatively correlated for unlabeled samples with low certainly predicted recGBM_neuronal scores. BioNet_pro and BioNet_inf have 63% KC_{neu^+} and 69% KC_{neu^-}. These values are moderately high, meaning that the knowledge regularization in these models worked. On the other hand, since knowledge has uncertainty, we do not want the regularization to be too strong, and force very high concordance. Rather, we want the models to balance knowledge regularization with the prediction loss and reconstruction loss. In this sense, moderately high concordance is indeed the most reasonable. BioNet_pro and BioNet_inf have moderately high KC_{neu^+} (63%), but other methods have relatively low KC_{neu^+}. This is reasonable because BioNet_pro and BioNet_inf are the only methods with knowledge regularization. Furthermore, all methods have relatively high KC_{neu^-}, indicating that this knowledge may be easier to learn by a model even without explicit regularization. KC_{neu^-} of BioNet_pro and BioNet_inf (69%) is not the highest, which illustrates that BioNet predictions are not only based on knowledge but also the labels of gene modules and the image features of both labeled and unlabeled samples. Additional interpretation of BioNet_pro and BioNet_inf is provided in Section 2 of SI Appendix.

Testing of BioNet in an Independent Cohort B. To test the generalizability of BioNet, we used an independent dataset from a recent published clinical trial (cohort B) ref from peter, which contained a total of 31 MRI-localized biopsy samples harvested from five patients with post-treatment recurrent GBM (median age = 56, 60% male; 1-10 samples per patient).

The commonly used train-test procedure is to apply the trained model (without any adjustment) directly on the test samples. We tried this procedure and obtained a test AUC of 0.70, 0.60, 0.64 for BioNet_neu, BioNet_pro, and BioNet_inf, respectively, on Cohort B. One reason of the reduced AUC performance may be that the MRI images of Cohort B were acquired at a higher resolution than cohort A. Even though we tried using downsampling to mitigate this discrepancy, it did not seem to help resolve the problem completely. To improve the test performance, we employed an approach inspired by domain adaptation (46) to retrain BioNet by including unlabeled samples from Cohort B (see details in Methods). This approach did not include any biopsy sample from Cohort B, so that the retrained BioNet was still "blind" to the ground-truth scores of the biopsy samples in Cohort B. This approach only used data from the pre-surgical MRI, and in principle could have been performed prior to surgery. Using this approach, the AUC for BioNet_neu, BioNet_pro, and BioNet_inf improved to 0.71, 0.78, and 0.72, respectively. These metrics are higher than the accuracy of semi-AE, NN, SVR, and RF on cohort B (**Fig. 7AB**). Furthermore, BioNet_pro and BioNet_inf as well as the other methods under comparison were used to generate prediction maps of recGBM_proliferative and recGBM_inflammatory within the ROI of each test patient. Based on the maps, the two knowledge concordance metrics previously defined, KC_{neu^+} and KC_{neu^-}, were computed (**Fig. 7C**). BioNet_pro and BioNet_inf have 68% KC_{neu^+} and 73% KC_{neu^-}, which are at similar levels as those based on Cohort A using CV. The other methods have significantly lower KC_{neu^+} (range: 9-21%) and similar KC_{neu^-} (rage: 49-79%). These trends are also similar to those found in Cohort A.

Discussion

Treatment-induced reactive changes (treatment-effect) add additional complexity to the tissue composition of recGBM making tissue classification based on MRI features more difficult. We have demonstrated a non-invasive approach to characterize intratumoral heterogeneity and identify gene modules at *voxel-level* resolution, which could improve surgical targeting of confirmatory biopsies and non-invasive assessment of neurooncological treatment response.

We systematically evaluated the performance of our approach through leave-one-patient-out CV on one dataset and blind testing on another independent dataset of recGBM patients from a clinical trial. Although our training set containes only very sparse biopsy samples, BioNet achieves 0.78 recGBM_proliferative prediction accuracy and 0.80 recGBM_inflammatory prediction accuracy from CV and 0.74 recGBM_proliferative prediction accuracy and 0.71 recGBM_inflammatory prediction accuracy from blind testing. In contrast to standard DL and ML methods that are not able to reach 0.65 accuracy from CV for both gene modules, BioNet provides a quantitative assessment of intratumoral heterogeneity in the context of post-treatment recurrent GBM, which is difficult for human experts to assess (47, 48). Furthermore, BioNet_pro and BioNet_inf have strong generalizability on the entire brain because of inclusion of abundant unlabeled samples in model training. Consequently, the well-trained model has the capability to generate prediction maps of recGBM_proliferative and recGBM_inflammatory throughout the whole ROI of each patient. These maps indicate the most likely location of recurrent tumor and specify the targets for confirmatory biopsies, which helps reveal the landscape of intratumoral heterogeneity and spatial distribution for proliferating tumor and reactive/inflammatory changes in the post-treatment setting, potentially assisting physicians for individualized assessment.

Due to the limited availability of labeled/biopsy samples, it was crucial to leverage the biological relationships of our three gene modules to complement training data. In recent years, ML approaches have demonstrated success in integrating biological/biomedical domain knowledge into the model design. Depending on the form of the domain knowledge, the approaches for integration vary. Some researchers proposed to use the knowledge of biological pathways to guide the architecture design of DL (31, 32). Moreover, in some biomedical fields, domain knowledge exists in the form of algebraic equations representing biological principles which are imposed on ML models (24, 34). Additionally, some researchers proposed to integrate the knowledge about how features should behave in generating prediction as attribution priors into DL training (33). Finally, noting that some knowledge learned from relative domains is also informative for the target tasks, transfer learning is an effective approach to leverage such knowledge in DL training (35–38).

However, the biological relationship of three gene modules is implicit, hierarchical, and hard to introduce into the model design through existing methods, which motivated the development of BioNet. The proposed framework linked BioNet_neu (a Bayesian NN), BioNet_pro and BioNet_inf (a multi-task semi-supervised AAE) according to the hierarchical relationships between the three gene modules. For each level of the relationship: (1) BioNet_neu leveraged the knowledge about recGBM_neuronal distributions; and (2) BioNet_pro and BioNet_inf leveraged the knowledge about relationships between recGBM_proliferative and recGBM_inflammatory, when the labels of recGBM_neuronal are known.

Although we fully exploited information from domain knowledge, it was still challenging to train an ensemble of networks based on very sparse labeled/biopsy samples. To strengthen the generalizability of BioNet, we introduced a large amount of unlabeled samples in model training by two coupled AAEs which were able to find representations of input features that were not only robust but also discriminative for gene module predictions.

Blind testing provides an unbiased evaluation of the model generalizability. However, a common problem when integrating multiple datasets is that there exists discrepancy caused by between-site differences in clinical measures between datasets. This problem also occurred in the BioNet blind testing. When we directly applied the BioNet fit on CV dataset to another independent blind test dataset, the prediction accuracy of recGBM_proliferative and recGBM_inflammatory on the blind test dataset was much lower than that on CV dataset. To mitigate the discrepancy between two datasets, we adjusted the model through unlabeled samples from the blind test, which provided an opportunity to learn new image representations. Since no biopsy samples from the blind test were included when retraining the model, the test dataset was

still blind to BioNet. Using this approach, the prediction accuracy was closer to that of CV, demonstrating generalizability of BioNet to independent datasets. Notably, this approach of adjusting the model with data extracted from patients' pre-surgical MRI could be performed prior to surgery as a way to improve the models' predictive strength and better guide surgical sampling. In sum, BioNet is a general, multi-task, semi-supervised framework that can be readily extended to other fields where domain knowledge can be hierarchically described and has limited labeled data.

We now discuss some potential improvements of our model. The biopsy sample size in our developmental cohort is limited. The model performance and robustness can potentially be further improved when having more biopsy samples. It is also of crucial importance to quantify model uncertainty for predictions, especially in healthcare. An interesting future direction is to develop probabilistic DL models which can offer confidence bounds on the predictions of all gene modules.

Methods

Patient inclusion and biopsy acquisition. Two independent patient samples were used to develop (cohort A) and test (cohort B) the model. Cohort A was acquired as part of a retrospective, observational study designed to study patients who had undergone repeat surgical resection for recurrence of high-grade gliomas after treatment with chemotherapy and radiation therapy (49). Cohort B was acquired as part of a prospective, clinical trial for convection enhanced delivery of topotecan chemotherapy (50). Both cohorts received standard of care neuroimaging performed at Columbia University Irving Medical Center within one month prior to surgery (Cohort A) or one day prior to surgery (Cohort B). Samples were obtained according to IRB-approved protocol from the operating room. Patient information was de-identified and maintained by a tissue broker who has designated clinical information. Details about biopsy acquisition are provided in Section 3 of SI Appendix.

Immunohistochemical staining (IHC) and quantification. A subset of the localized biopsies in Cohort A (n=48) and all biopsies in Cohort B underwent histological staining for H&E and immunohistochemical staining (IHC) for SOX2, Ki67, CD68, and NeuN. All slides were digitized. Total cell density was measured by counting the hematoxylin-stained nuclei. Subsequently, for each sample, each stain was quantified and its labeling index determined on a scale from 0 to 1 by normalizing to H&E. Accuracy of the labeling indices was checked by a neuropathologist (Peter Canoll).

RNAseq-IHC correlation analysis, clustering, and GSVA. For samples in Cohort A that had both RNAseq and IHC quantification, Pearson correlation was calculated between the normalized expression values for each gene and the IHC labeling index for each marker, building a correlation matrix of *m* genes by 5 IHC markers (4 IHC markers and total normalized cellularity from H&E images). A p-value cutoff of 0.05 was set for determining the significance of the IHC-gene expression correlation for each stain, and all genes with significant correlations with ≥1 marker were selected for downstream analysis. Hierarchical clustering using Euclidean distance was performed on the resulting correlation matrix and major clusters with mutually exclusive genes highly correlated to specific IHC markers were identified via hierarchical clustering. These mutually exclusive gene clusters were used as gene sets for downstream Gene Set Variation Analysis (GSVA) on a sample-by-sample basis, which produced three gene set enrichment scores for each MRI-localized biopsy. We call the three gene sets "gene modules" and use the enrichment scores as estimates of gene module expression. Analysis of the genes in each module suggested that the modules consisted of proliferative (recGBM_proliferative), inflammatory (recGBM_inflammatory), and neuronal (recGBM_neuronal) genes. The raw enrichment scores are between -1 and 1. In training the NNs in BioNet, the scores were transformed to [0,1] using a simple mapping of $f(x) = \frac{x+1}{2}$, so that the NNs could be trained as soft classifiers. The genes identified in Cohort A were then used to perform GSVA in Cohort B to generate enrichment scores for each of the three gene modules. RNA extraction and pooled library amplification for transcriptome expression RNA-sequencing (PLATE-Seq) are provided in Section 4 of SI Appendix.

Extraction of localized features from multiparametric MRI. The Region of Interest (ROI) for each patient was defined by combining the enhancing tumor and infiltrating tumor regions segmented based on T1Gd and T2/FLAIR images plus a 7mm margin. A 2D sliding window approach was used to create local regions

with a size of 5×5 voxels centered on each voxel within the ROI. The window size approximates the biopsy tissue size. The 2D sliding windows were moved in a sliding window manner from top left to bottom right with stride 1 on each MRI axial slice. From each sliding window, 280 statistical and texture features were computed from five multiparametric MRI images, including T1-weighted + Gd, T2-weighted, FLAIR, apparent diffusion coefficient (ADC), and susceptibility weighted imaging (SWI). As a result, for each patient we have a dataset $(x_i, y_i)_{i=1}^{l} \cup (x_i)_{i=l+1}^{l+u}$, where x_i contains the features extracted from sliding window i and $y_i = (y_{i,neu}, y_{i,pro}, y_{i,inf})$ contains the GSVA enrichment scores for the three gene modules (**Fig. 8**). The first part of the dataset, $(x_i, y_i)_{i=1}^{l}$, corresponds to biopsy samples (a.k.a. labeled-samples) whereas the second part, $(x_i)_{i=l+1}^{l+u}$, corresponds to unlabeled samples for which we want to predict the enrichment scores. The collective dataset of 31 patients includes 69 biopsy/labeled samples with spatially-matched multiparametric MRI features and enrichment scores (range: 1-3 per patient) and 15000 unlabeled samples (range: 200-800 per patient). Details about multiparametric MRI pre-processing, segmentation, and feature extraction are provided in Section 5 of SI Appendix.

Composing a virtual biopsy dataset to pre-train BioNet_neu. Our goal is to identify unlabeled samples that are likely to have high or low recGBM_neuronal based on biological knowledge, denoted as class $\tilde{1}$ or class $\tilde{0}$. The overhead '∼' indicates that knowledge has uncertainty and thus the samples in each class may have labeling errors. This is acceptable because this dataset is only used to pre-train the BioNet-Neu. Specifically, we have the knowledge that recGBM_neuronal should be high outside the ROI (i.e., in the normal brain areas), because the genes included recGBM_neuronal are predominantly involved in neuronal signaling. In theory we can sample anywhere outside the ROI. However, it is hard to avoid unwanted brain structures. Thus, we chose to include all samples on the ROI boundary in class $\tilde{1}$.

Furthermore, recGBM_neuronal is likely to be low within enhancing tumor, because it is known to be associated with tumor proliferation or immune response (51), and is less likely to be associated with neuronal signaling. Thus, we included all samples from enhancing tumor in class $\tilde{0}$. As a result, the virtual biopsy dataset included 7500 random selected samples in class $\tilde{1}$ and $\tilde{0}$ respectively from 31 patients.

BioNet_neu design, training, and evaluation. BioNet_neu included two hidden layers with 2048 nodes at each layer. The input to each layer, x_l, is linked to the output, x_{l+1}, through $x_{l+1} = f\big((\varepsilon \odot W_l) \quad x_l + b_l\big)$. W_l and b_l denote the weights and bias. f is the ReLU activation function. To add uncertainty quantification capability of the model, a dropout mask, ε, is used, which samples from a Bernoulli distribution according to a pre-specified drop rate p_ε (42). Repeatedly sampling from the Bernoulli distribution creates an ensemble of networks.

BioNet_pro and BioNet_inf design, training, and evaluation. BioNet_pro and BioNet_inf are two parallel NNs that are coupled together in the loss function. Each parallel NN includes an AAE reconstruction network (45, 52) that consists of an encoder En, a decoder De, and a discriminator Di, as well as a classification network Cl. There are three hidden layers in En, two in De, two in Di, and three in Cl, with 256 nodes for each layer. The dimension for the latent representation of the AAE, z, is 512 and 64 in BioNet_pro and BioNet_inf, respectively.

Specifically, we designed two loss functions to regularize the predictions on two groups of unlabeled samples stratified by BioNet_neu, respectively. Let $(x_i)_{i \in neu^+}$ and $(x_i)_{i \in neu^-}$ denote unlabeled samples with high and low predicted recGBM_neuronal scores. For the 'high' group, we have the knowledge that recGBM_proliferative and recGBM_inflammatory are likely to be low (**Fig. 2A**). Thus, the corresponding loss function on unlabeled samples in this group is

$$\mathcal{L}_{neu^+} = \sum_{i \in neu^+} \left\|\hat{y}_{i,pro}\right\|_2^2 + \left\|\hat{y}_{i,inf}\right\|_2^2 ,$$

where $\| \quad \|_2^2$ is the squared L2 norm. For the 'low' group, we have the knowledge that recGBM_proliferative and recGBM_inflammatory are likely to be negatively correlated (**Fig. 2A**). Thus, the corresponding loss function on unlabeled samples in this group is

$\mathcal{L}_{neu^-} = \sum_{i \in neu^-} \left(1 - \hat{y}_{i,pro}\,,\ \hat{y}_{i,pro}\right)\left(1 - \hat{y}_{i,inf}\,,\ \hat{y}_{i,inf}\right)^T.$

Combining the two losses, the knowledge regularization (KR) loss is

$\mathcal{L}_{KR} = \mathcal{L}_{neu^-} + \gamma \mathcal{L}_{neu^+}.$

The reconstruction loss of the AAEs on both unlabeled and biopsy/labeled samples is

$\mathcal{L}_{recon} = \sum_{i \in neu^+ \cup neu^- \cup \{l\}} \left\| x_i - \hat{x}_{i,pro} \right\|_2^2 + \left\| x_i - \hat{x}_{i,inf} \right\|_2^2.$

Additionally, we connected the representation of the AAE with a classification network for each gene module. The prediction loss of the classification networks on biopsy/labeled samples is

$\mathcal{L}_{pred} = \sum_{i=1}^{l} \left\| y_{i,pro} - \hat{y}_{i,pro} \right\|_1 + \left\| y_{i,inf} - \hat{y}_{i,inf} \right\|_1,$

where $\| \quad \|_1$ is the L1 norm.

In summary, BioNet_pro and BioNet_inf are trained together to minimize an overall loss function that combines the KR loss, reconstruction loss, and prediction loss, i.e.,

$\mathcal{L} = \mathcal{L}_{pred} + \alpha \mathcal{L}_{recon} + \beta \mathcal{L}_{KR}.$

The models were trained by a three-phase iterative process: (1) knowledge regularization phase, (2) adversarial regularization phase, and (3) classification phase.

(1) In the knowledge regularization phase, we update the encoder En, decoder De, and classifier Cl for each parallel NN based on biopsy and unlabeled samples. The training aims to minimize the overall loss $\mathcal{L} = \mathcal{L}_{pred} + \alpha \mathcal{L}_{recon} + \beta \mathcal{L}_{KR}$.

(2) In the adversarial regularization phase, we update the encoder En and discriminator Di for each parallel NN based on labeled and unlabeled samples. The adversarial regularization is formulated as a min-max problem, i.e.,
$\min_{En} \max_{Di} \sum_{k \in neu^+ \cup neu^- \cup \{l\}} \left(Di(r_k) - Di\left(En(x_i)\right)\right),$
where r_k denotes a sample generated from a Gaussian-prior distribution through a learned smooth layer. To separate the prior-generated samples r_k from the encoder-generated samples $En(x_i)$, the discriminator Di aims to maximize the Wasserstein distance between the two sets of samples, while the encoder En aims to minimize this distance to fool the discriminator. En and Di are trained iteratively.

(3) In the classification phase, we update the classifier Cl for each parallel NN based on labeled samples. The training aims to minimize the absolute prediction error $\mathcal{L}_{pred} = \sum_{i=1}^{l} \left\| y_{i,pro} - \hat{y}_{i,pro} \right\|_1 + \left\| y_{i,inf} - \hat{y}_{i,inf} \right\|_1$.

The three phases run iteratively to optimize the entire model. When the losses in the three phases converge, the training process is complete.

The above three-phase training process was repeated for each patient under leave-one-patient-out (LOPO) cross-validation (CV). Specifically, the biopsy/labeled and unlabeled samples of all but one patient were used to train BioNet_pro and BioNet_inf. Then the trained models were used to make predictions for the left-out patient. This is to mimic the clinical workflow when the models are in use, i.e., the models will be used patient by patient, not sample by sample.

To select a subset of unlabeled samples to include in training, we ordered the unlabeled samples according to their uncertainty of prediction by BioNet_neu, where the uncertainty is measured by predictive entropy. We selected a subset of unlabeled samples with predictive entropy < 0.44. About 7×10^4 unlabeled samples survived this threshold across the experiments of different patients.

Details about testing of BioNet_neu, BioNet_pro and BioNet_inf are provided in Section 6 of SI Appendix.

ACKNOWLEDGMENTS. This work was supported by NIH grant 1U01CA250481-01A1, NSF grant DMS-2053170, NIH/NINDS R01NS103473, NIH/NCI R01CA161404 and the NIH/NCI Cancer Center Support Grant P30CA013696. We would like to acknowledge the Herbert Irving Comprehensive Cancer Center Molecular and Pathology Core.

References

1. C. I. Ene, H. A. Fine, Many Tumors in one: A daunting therapeutic prospect. *Cancer Cell* **20**, 695–697 (2011).
2. L. S. Hu, A. Hawkins-Daarud, L. Wang, J. Li, K. R. Swanson, Imaging of intratumoral heterogeneity in high-grade glioma. *Cancer Lett* **477**, 97–106 (2020).
3. M. A. Qazi, *et al.*, Intratumoral heterogeneity: pathways to treatment resistance and relapse in human glioblastoma. *Annals of Oncology* **28**, 1448–1456 (2017).
4. N. Renee Parker, *et al.*, Molecular heterogeneity in glioblastoma: potential clinical implications. *Front Oncol* **5**, 55 (2015).
5. A. F. Kazerooni, S. Bakas, H. S. Rad, C. Davatzikos, Imaging Signatures of Glioblastoma Molecular Characteristics: A Radiogenomics Review. *Journal of Magnetic Resonance Imaging* **52**, 54–69 (2020).
6. G. Singh, *et al.*, Radiomics and radiogenomics in gliomas: a contemporary update. *Br J Cancer* **125**, 641–657 (2021).
7. L. S. Hu, A. Hawkins-Daarud, L. Wang, J. Li, K. R. Swanson, Imaging of intratumoral heterogeneity in high-grade glioma. *Cancer Lett* **477**, 97–106 (2020).
8. B. AL, *et al.*, Patient-specific metrics of invasiveness reveal significant prognostic benefit of resection in a predictable subset of gliomas. *PLoS One* **9**, E17–E18 (2014).
9. M. L. Neal, *et al.*, Discriminating survival outcomes in patients with glioblastoma using a simulation-based, patient-specific response metric. *PLoS One* **8** (2013).
10. R. C. Rockne, *et al.*, A patient-specific computational model of hypoxia-modulated radiation resistance in glioblastoma using 18F-FMISO-PET. *J R Soc Interface* **12** (2015).
11. H. Akbari, *et al.*, In vivo evaluation of EGFRvIII mutation in primary glioblastoma patients via complex multiparametric MRI signature. *Neuro Oncol* **20**, 1068–1079 (2018).
12. H. Chen, *et al.*, Deep Learning Radiomics to Predict PTEN Mutation Status From Magnetic Resonance Imaging in Patients With Glioma. *Front Oncol* **11** (2021).
13. P. Kickingereder, *et al.*, Radiogenomics of glioblastoma: Machine learning-based classification of molecular characteristics by using multiparametric and multiregional MR imaging features. *Radiology* **281**, 907–918 (2016).
14. E. S. Tykocinski, *et al.*, Use of magnetic perfusion-weighted imaging to determine epidermal growth factor receptor variant III expression in glioblastoma. *Neuro Oncol* **14**, 613–623 (2012).
15. X. Zhang, *et al.*, IDH mutation assessment of glioma using texture features of multimodal MR images. *Medical Imaging 2017: Computer-Aided Diagnosis* **10134**, 462–469 (2017).
16. B. Zhang, *et al.*, Multimodal MRI features predict isocitrate dehydrogenase genotype in high-grade gliomas. *Neuro Oncol* **19**, 109–117 (2017).
17. S. E. Combs, *et al.*, Prognostic significance of IDH-1 and MGMT in patients with glioblastoma: One step forward, and one step back? *Radiation Oncology* **6**, 1–5 (2011).
18. X. P. Chang, *et al.*, Deep-Learning Convolutional Neural Networks Accurately Classify Genetic Mutations in Gliomas. *American Journal of Neuroradiology* **39**, 1201--1207 (2018).
19. A. L. Baldock, *et al.*, Invasion and proliferation kinetics in enhancing gliomas predict IDH1 mutation status. *Neuro Oncol* **16**, 779–786 (2014).

20. J. C. Urcuyo, *et al.*, Image-localized Biopsy Mapping of Brain Tumor Heterogeneity: A Single-Center Study Protocol. *medRxiv*, 2022.11.14.22282304 (2022).
21. B. J. Gill, *et al.*, MRI-localized biopsies reveal subtype-specific differences in molecular and cellular composition at the margins of glioblastoma. *Proc Natl Acad Sci U S A* **111**, 12550–12555 (2014).
22. L. S. Hu, *et al.*, Multi-parametric MRI and texture analysis to visualize spatial histologic heterogeneity and tumor extent in glioblastoma. *PLoS One* **10**, e0141506 (2015).
23. N. Gaw, *et al.*, Integration of machine learning and mechanistic models accurately predicts variation in cell density of glioblastoma using multiparametric MRI. *Sci Rep* **9**, 1–9 (2019).
24. L. Wang, A. Hawkins-Daarud, K. R. Swanson, L. S. Hu, J. Li, Knowledge-Infused Global-Local Data Fusion for Spatial Predictive Modeling in Precision Medicine. *IEEE Transactions on Automation Science and Engineering* (2021) https:/doi.org/10.1109/TASE.2021.3076117.
25. P. D. Chang, *et al.*, A Multiparametric Model for Mapping Cellularity in Glioblastoma Using Radiographically Localized Biopsies. *AJNR Am J Neuroradiol* **38**, 890–898 (2017).
26. C. H. Wang, *et al.*, Prognostic significance of growth kinetics in newly diagnosed glioblastomas revealed by combining serial imaging with a novel biomathematical model. *Cancer Res* **69**, 9133–9140 (2009).
27. K. R. Swanson, C. Bridge, J. D. Murray, E. C. Alvord, Virtual and real brain tumors: using mathematical modeling to quantify glioma growth and invasion. *J Neurol Sci* **216**, 1–10 (2003).
28. L. S. Hu, *et al.*, Radiogenomics to characterize regional genetic heterogeneity in glioblastoma. *Neuro Oncol* **19**, 128–137 (2017).
29. L. S. Hu, *et al.*, Uncertainty quantification in the radiogenomics modeling of EGFR amplification in glioblastoma. *Sci Rep* **11**, 1–14 (2021).
30. P. Y. Wen, *et al.*, Response Assessment in Neuro-Oncology Clinical Trials. *Journal of Clinical Oncology* **35**, 2439–2449 (2017).
31. H. A. Elmarakeby, *et al.*, Biologically informed deep neural network for prostate cancer discovery. *Nature* **598**, 348–352 (2021).
32. Z. L. Azher, L. J. Vaickus, L. A. Salas, B. C. Christensen, J. J. Levy, Development of Biologically Interpretable Multimodal Deep Learning Model for Cancer Prognosis Prediction. *Proceedings of the 37th ACM/SIGAPP Symposium on Applied Computing*, 636–644 (2022).
33. G. Erion, J. D. Janizek, P. Sturmfels, S. M. Lundberg, S. I. Lee, Improving performance of deep learning models with axiomatic attribution priors and expected gradients. *Nat Mach Intell* **3**, 620–631 (2021).
34. T. M. Deist, *et al.*, Simulation-assisted machine learning. *Bioinformatics* **35**, 4072–4080 (2019).
35. R. K. Sevakula, V. Singh, N. K. Verma, C. Kumar, Y. Cui, Transfer Learning for Molecular Cancer Classification Using Deep Neural Networks. *IEEE/ACM Trans Comput Biol Bioinform* **16**, 2089–2100 (2019).
36. J. Amin, *et al.*, IMAGE & SIGNAL PROCESSING A New Approach for Brain Tumor Segmentation and Classification Based on Score Level Fusion Using Transfer Learning. *J Med Syst* **43**, 1–16 (2019).
37. R. Zoetmulder, E. Gavves, M. Caan, H. Marquering, Domain- and task-specific transfer learning for medical segmentation tasks. *Comput Methods Programs Biomed* **214**, 106539 (2022).
38. C. C. Vu, *et al.*, Deep convolutional neural networks for automatic segmentation of thoracic organs-at-risk in radiation oncology – use of non-domain transfer learning. *J Appl Clin Med Phys* **21**, 108–113 (2020).
39. J. Yosinski, J. Clune, Y. Bengio, H. Lipson, How transferable are features in deep neural networks? *Adv Neural Inf Process Syst* **27** (2014).
40. N. Srivastava, G. Hinton, A. Krizhevsky, R. Salakhutdinov, Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research* **15**, 1929–1958 (2014).

41. A. Kendall, Y. Gal, What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision? *Adv Neural Inf Process Syst* **30** (2017).
42. Y. Gal, Z. Ghahramani, Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning in *International Conference on Machine Learning*, (2016), pp. 1050–1059.
43. Y. Ouali, C. Hudelot, M. Tami, An Overview of Deep Semi-Supervised Learning. *arXiv preprint arXiv:2006.05278* (2020) https:/doi.org/10.48550/arxiv.2006.05278 (November 13, 2022).
44. S. Ruder, An Overview of Multi-Task Learning in Deep Neural Networks. *arXiv preprint arXiv:1706.05098* (2017) https:/doi.org/10.48550/arxiv.1706.05098 (November 13, 2022).
45. A. Makhzani, J. Shlens, N. Jaitly, I. Goodfellow, B. Frey, Adversarial Autoencoders. *arXiv preprint arXiv:1511.05644* (2015).
46. Y. Ganin, V. Lempitsky, Unsupervised Domain Adaptation by Backpropagation. *arXiv preprint arXiv:1409.7495* (2014).
47. B. M. Ellingson, P. Y. Wen, T. F. Cloughesy, Modified Criteria for Radiographic Response Assessment in Glioblastoma Clinical Trials. *Neurotherapeutics* **14**, 307–320 (2017).
48. M. Holdhoff, *et al.*, The consistency of neuropathological diagnoses in patients undergoing surgery for suspected recurrence of glioblastoma. *J Neurooncol* **141**, 347 (2019).
49. O. Al-Dalahmah, *et al.*, Re-convolving the compositional landscape of primary and recurrent glioblastoma reveals prognostic and targetable tissue states. *bioRxiv* (2022) https:/doi.org/10.1101/2021.07.06.451295 (November 13, 2022).
50. E. F. Spinazzi, *et al.*, Chronic convection-enhanced delivery of topotecan for patients with recurrent glioblastoma: a first-in-patient, single-centre, single-arm, phase 1b trial. *Lancet Oncol* **23**, 1409–1418 (2022).
51. C. N. Kersch, P. Ambady, B. E. Hamilton, R. F. Barajas, MRI and PET of Brain Tumor Neuroinflammation in the Era of Immunotherapy, From the AJR Special Series on Inflammation HHS Public Access. *AJR Am J Roentgenol* **218**, 582–596 (2022).
52. D. P. Kingma, D. J. Rezende, S. Mohamed, M. Welling, Semi-Supervised Learning with Deep Generative Models. *Adv Neural Inf Process Syst* **27** (2014).

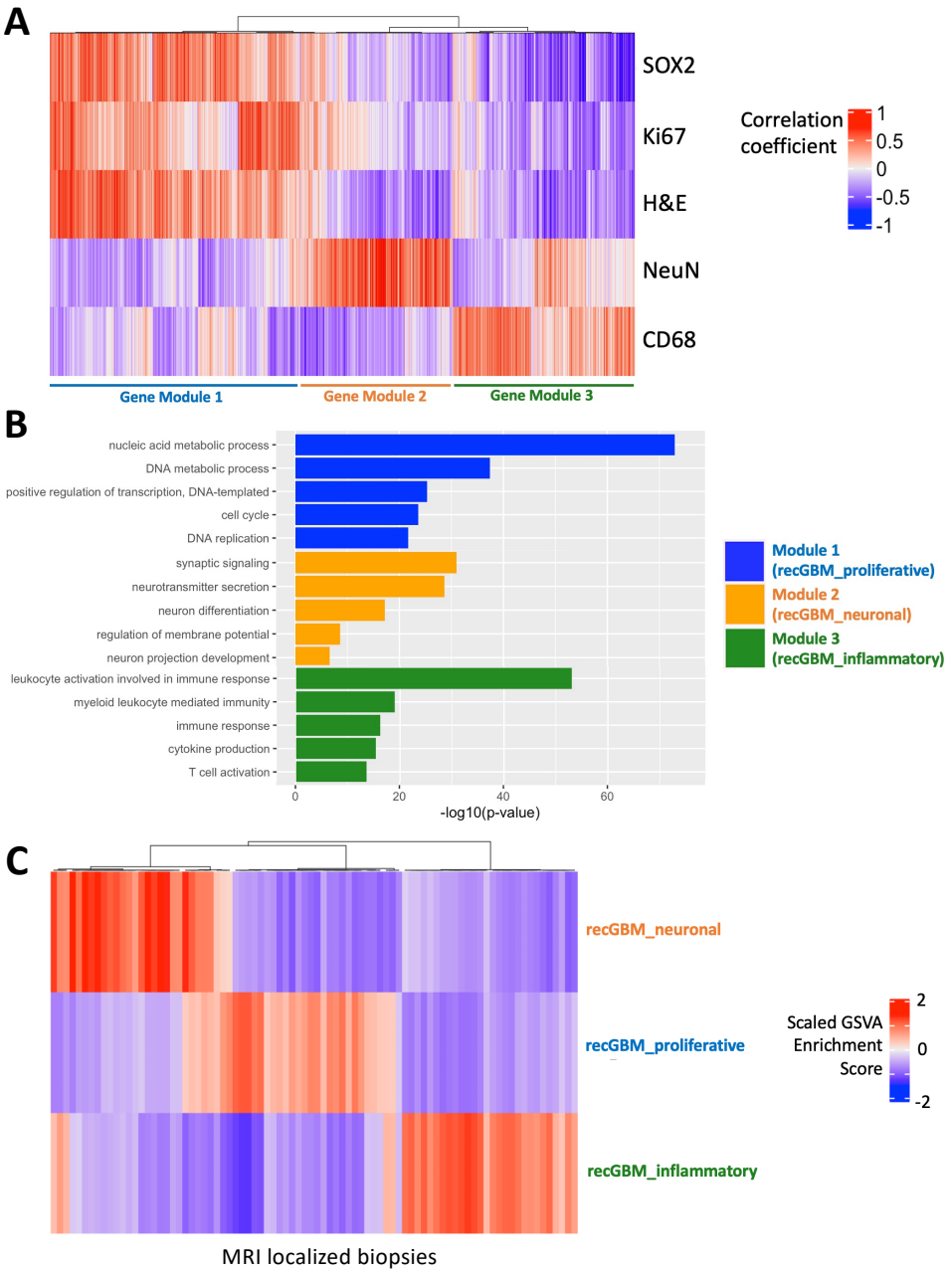


Figure 1: Defining tissue-specific gene modules to connect with key immunohistochemical features. **(A)** Heatmap depicting correlation between normalized gene expression and immunohistochemical labeling indices, with subsequent hierarchical clustering revealed three orthogonal tissue-specific gene sets. Module 1 consists of 3688 genes significantly positively correlated with SOX2/Ki67/H&E; Module 2 consists of 1673 genes correlated with NeuN; Module 3 consists of 2418 genes correlated with CD68. **(B)** Bar plot depicting top significant gene ontologies enriched in each of the three tissue-specific gene sets derived from the IHC-RNAseq correlation analysis. X-axis is $-\log_{10}$(p-value) of each ontology. Module 1 is enriched in genes involved in proliferation (recGBM_proliferative), Module 2 in neuronal-specific genes (recGBM_neuronal), and Module 3 in genes involved in immune infiltration (recGBM_inflammatory). **(C)** Heatmap depicting single-sample GSVA analysis for each of 84 MRI-localized biopsies for each of the three tissue-specific gene sets. Color gradient represents magnitude/direction of tissue-specific enrichment for each biopsy.

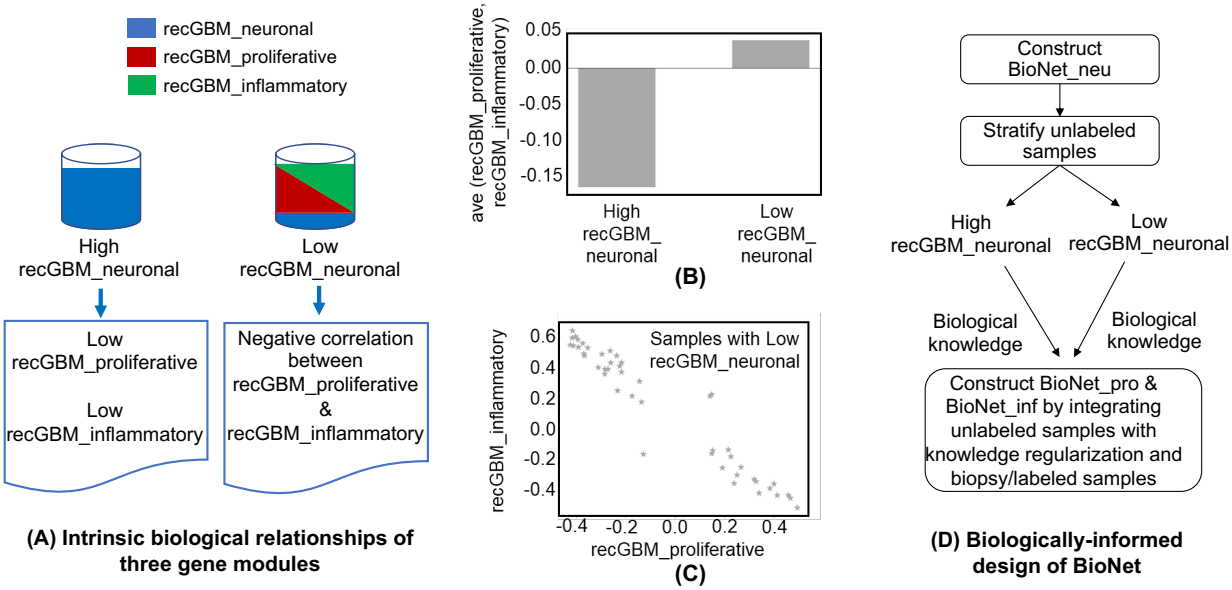


Fig 2. Defining biological relationships to include in regularization of the DL framework. **(A)** Intrinsic biological relationships of three gene modules (cartoon demonstration). **(B)** Bar chart for average score of recGBM_proliferative and recGBM_inflammatory for groups with high and low recGBM_neuronal. **(C)** Scatter plot for recGBM_proliferative and recGBM_inflammatory for the group with low recGBM_neuronal. **(D)** Hierarchical design of BioNet informed by biological knowledge about the intrinsic relationships of three gene modules shown in (A).

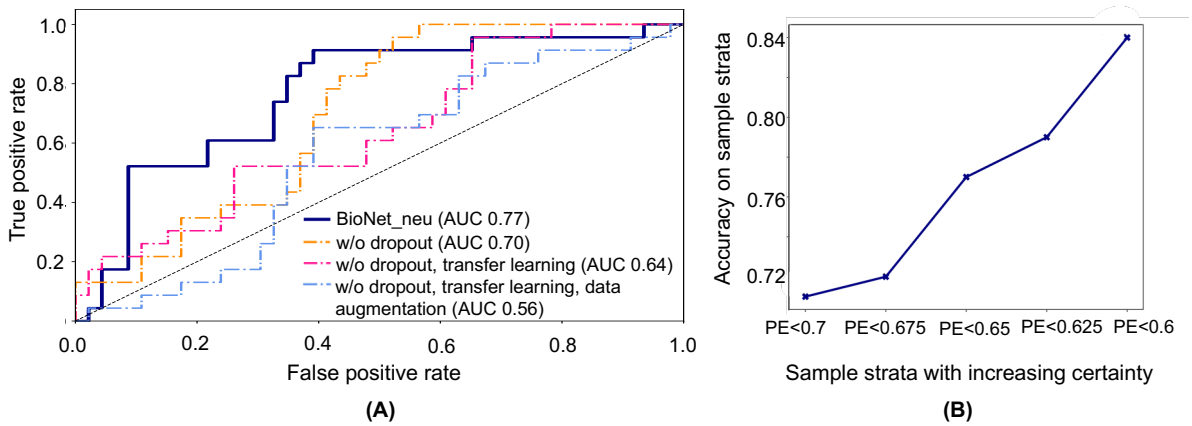


Fig 3. Comparison of model accuracy and uncertainty across key parametric settings. **(A)** ROCs of BioNet_neu vs. BioNet_neu without various strategies of augmenting learning capability; **(B)** Accuracy of BioNet_neu evaluated on sample strata with increasing certainty.

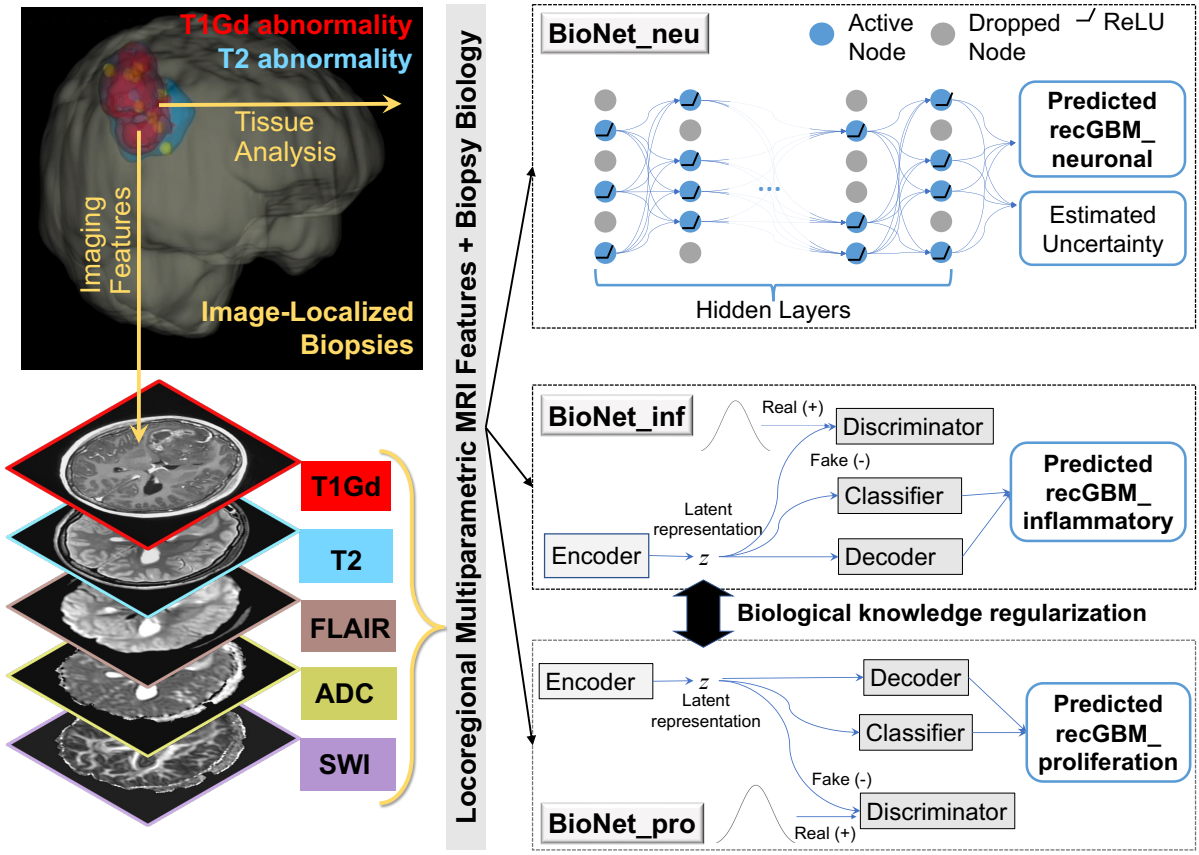
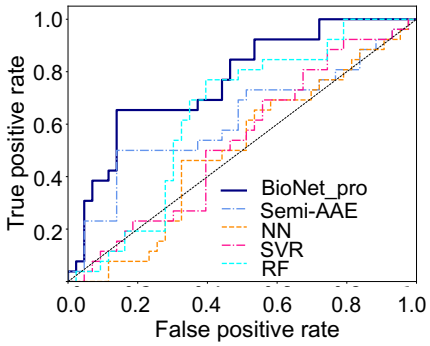


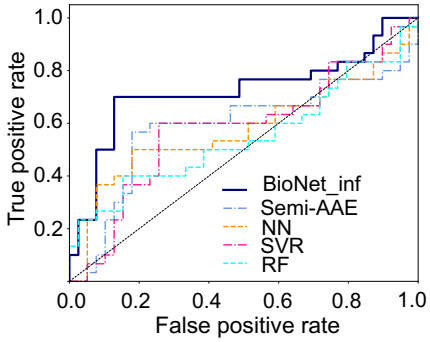
Fig.4. Overall architecture of three NNs that take as input image-localizes biopsies aligned to the multi-parametric MR imaging. The BioNet_neu is a feedforward NN with dropout. The BioNet_inf and BioNet_pro are two parallel NNs each includes an AAE reconstruction network (including an encoder, a decoder, a discriminator), and a classification network (i.e., classifier). The two NNs are coupled together by sharing an overall loss function that combines the knowledge-regularization loss, reconstruction loss, and prediction loss.

	Evaluation metrics		
	AUC	ACC	MAE
BioNet_pro	**0.77**	**0.78**	0.09
Semi-AAE	0.62	0.71	**0.08**
NN	0.50	0.59	0.12
SVR	0.53	0.54	0.12
RF	0.64	0.61	0.13

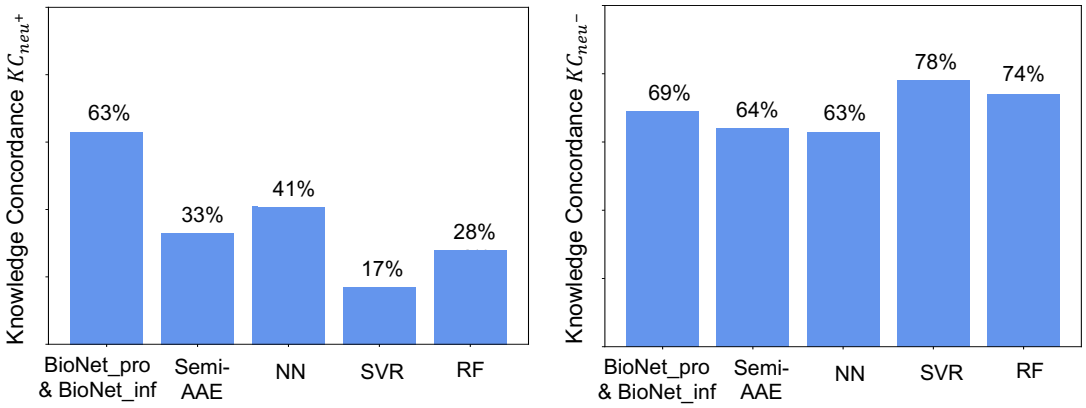


(A) Prediction performance for recGBM_proliferative

	Evaluation metrics		
	AUC	ACC	MAE
BioNet_inf	**0.72**	**0.80**	**0.11**
Semi-AAE	0.59	0.71	0.12
NN	0.58	0.68	0.14
SVR	0.58	0.65	0.17
RF	0.55	0.54	0.16



(B) Prediction performance for recGBM_inflammatory



(C) Knowledge concordance

Fig.5. Performance of BioNet on Cohort A. Comparison of prediction performance for (**A**) BioNet_pro and (**B**) BioNet_inf with other methods, and (**C**) knowledge concordance, based on leave-one-patient-out (LOPO) cross-validation (CV).

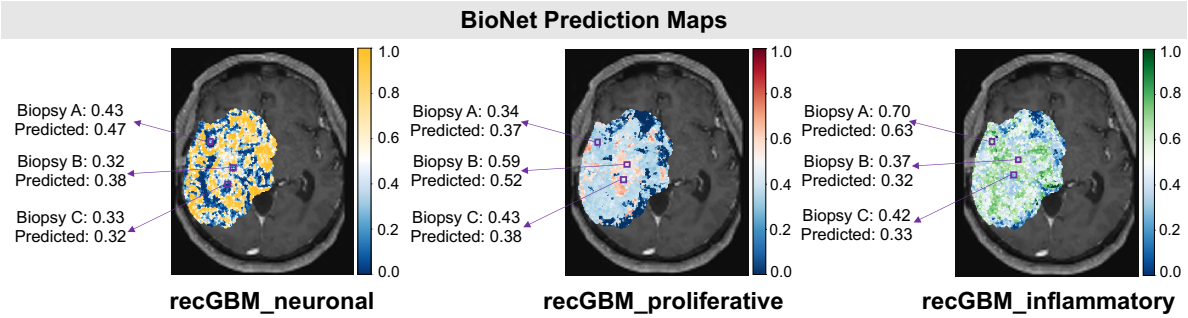
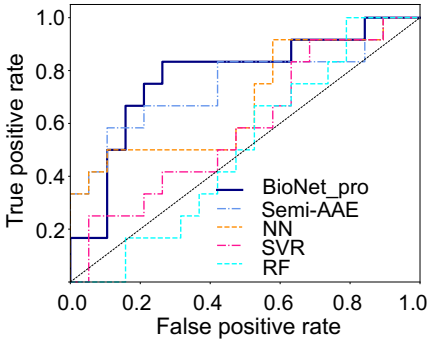


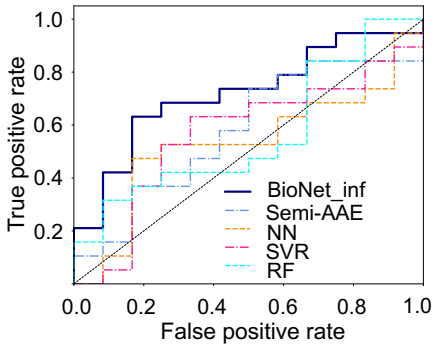
Fig.6. Mapping intratumoral heteorgeniety in proliferative and inflamtoory gene modules for test patients. Prediction maps for recGBM_neuronal (left), recGBM_proliferative (middle) and recGBM_inflammatory (right) by BioNet_neu, BioNet_pro and BioNet_inf for a patient from Cohort A with three biopsy samples A, B, C. Predicted scores are consistent with the true scores for all three samples.

	Evaluation metrics		
	AUC	ACC	MAE
BioNet_pro	**0.78**	**0.74**	0.11
Semi-AAE	0.75	0.71	**0.10**
NN	0.69	0.65	0.15
SVR	0.59	0.52	0.13
RF	0.51	0.52	0.14

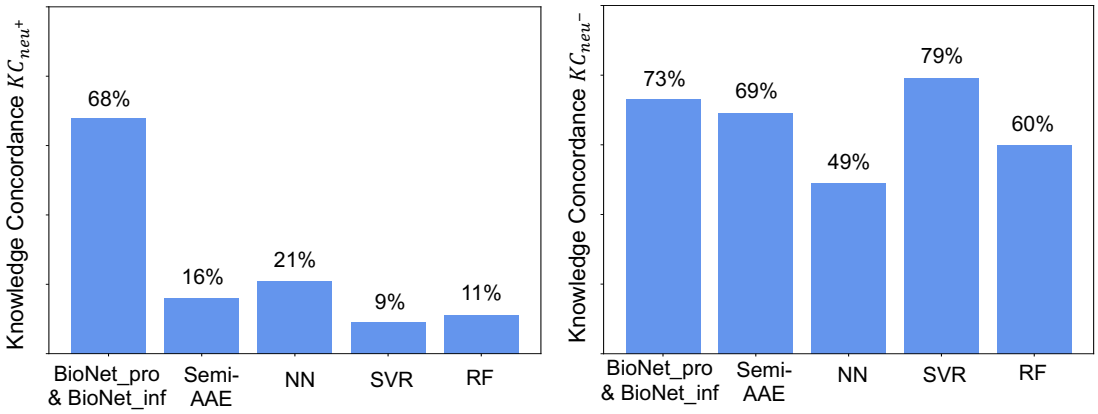


(A) Prediction performance for recGBM_proliferative

	Evaluation metrics		
	AUC	ACC	MAE
BioNet_inf	**0.72**	**0.71**	0.18
Semi-AAE	0.58	0.65	0.17
NN	0.53	0.58	0.18
SVR	0.57	0.58	0.17
RF	0.57	0.52	0.17



(B) Prediction performance for recGBM_inflammatory



(C) Knowledge concordance

Fig.7. Validation of BioNet performance in independent cohort B. Comparison of prediction performance for (**A**) BioNet_pro and (**B**) Bio_Net_inf with other methods, and (**C**) knowledge concordance, testing on Cohort B.

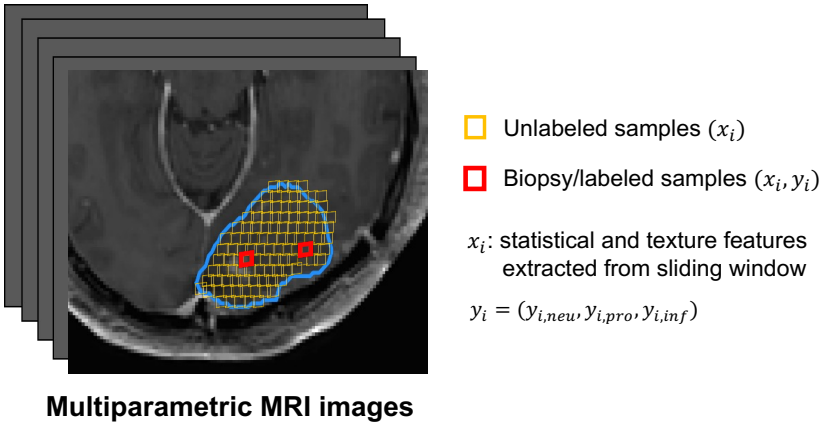


Figure 8: Within the ROI of each patient (blue outline), sliding windows created local regions (small squares). Statistical and texture features (x_i) were computed based on multiparametric MRI within the sliding windows at biopsy locations (red, a few) and remaining locations (yellow, many). BioNet aims to accurately predict the GSVA enrichment scores of three gene modules $(y_{i,neu}, y_{i,pro}, y_{i,inf})$ in the remaining locations in order to map out the intratumoral heterogeneity for each patient.