Pathway-informed deep learning model for survival analysis and pathological classification of gliomas

Gun Kaynar[1] Doruk Cakmakci[2] Caroline Bund[3,4,5] Julien Todeschi[6]
Izzie Jacques Namer[3,4,5] A. Ercument Cicek[1,7,*]

[1]Computer Engineering Department, Bilkent University, Ankara, Turkey 06800
[2]School of Computer Science, McGill University, Montreal, QC, Canada
[3]MNMS Platform, University Hospitals of Strasbourg, France
[4]ICube, University of Strasbourg / CNRS UMR 7357, Strasbourg, France
[5]Department of Nuclear Medicine and Molecular Imaging, ICANS, Strasbourg, France
[6]Department of Neurosurgery, Strasbourg University Hospital, Strasbourg, France
[7]Computational Biology Department, Carnegie Mellon University, Pittsburgh, PA 15213
[*]Correspondence: cicek@cs.bilkent.edu.tr

Abstract

Online assessment of tumor characteristics during surgery is important and has the potential to establish an intra-operative surgeon feedback mechanism. With the availability of such feedback, surgeons could decide to be more liberal or conservative regarding the resection of the tumor. While there are methods to perform metabolomics-based online tumor pathology prediction, their model complexity and, in turn, the predictive performance is limited by the small dataset sizes. Furthermore, the information conveyed by the feedback provided on the tumor tissue could be improved both in terms of content and accuracy. In this study, we propose a metabolic pathway-informed deep learning model, PiDeeL, to perform survival analysis and pathology assessment based on metabolite concentrations. We show that incorporating pathway information into the model architecture substantially reduces parameter complexity and achieves better survival analysis and pathological classification performance. With these design decisions, we show that PiDeeL improves tumor pathology prediction performance of the state-of-the-art in terms of the Area Under the ROC Curve (AUC-ROC) by 3.38% and the Area Under the Precision-Recall Curve (AUC-PR) by 4.06%. Similarly, with respect to the time-dependent concordance index (c-index), we observe that PiDeeL achieves better survival analysis performance (improvement up to 4.3%) when compared to the state-of-the-art. Moreover, we show that importance analyses performed on input metabolite features as well as pathway-specific hidden-layer neurons of PiDeel provide insights into tumor metabolism. We foresee that the use of this model in the surgery room will help surgeons adjust the surgery plan on the fly and will result in better prognosis estimates tailored to surgical procedures. The code is released at `https://github.com/ciceklab/PiDeeL`. The data used in this study is released at `https://zenodo.org/record/7228791`.

1 Introduction

Gliomas have uncertain patient mortality rates due to differences in suspected brain cell types and the variance observed in cancer progression dynamics [1]. 2021 version of central nervous system tumor taxonomy developed by the World Health Organization (WHO) [2] categorizes gliomas according to cell types (e.g. astrocytes, oligodendrocytes) as well as four pathologically different grades (i.e. grade I - IV). This taxonomy takes differential histopathologic, molecular, and genetic characteristics of gliomas into account and orders grades with respect to increasing malignancy and decreasing prognostic optimism. Considering the prevalence of gliomas among brain tumors diagnosed in adults [3], devising a surgical management strategy that aims to make personalized clinical decision-making plausible via an intraoperative or preoperative surgeon feedback mechanism becomes essential.

In order to amplify the percentage of tumor tissue resected from a patient's brain, various imaging [4], spectroscopy as well as mass spectrometry [5–13], and optical spectrometry [14–26] based techniques, were proposed in the literature. As a member of this cohort, ^{1}H High-Resolution Magic Angle Spinning Nuclear Magnetic Resonance (HRMAS NMR) spectroscopy measures metabolic activity in a given unprocessed tissue specimen [27] with a turnaround time of $\sim$ 20 mins and generates a time domain signal (Free Induction Decay - FID) in the form of the sum of decaying exponential functions. The FID signal is often preprocessed to obtain the HRMAS NMR spectrum defined on the frequency domain.

After preprocessing, the HRMAS NMR spectrum becomes well-suited to analyze the presence of left-over tumor tissues in the excision cavity and provide feedback to the surgeon about the nature of the tumor. Establishing this HRMAS NMR-guided intraoperative feedback loop requires a pathologist knowledgeable about tumor metabolism, and an NMR technician who detects an abundance of a few metabolites from the spectrum. However, the feedback resulting from this strategy is limited by (1) the availability of the experts during the surgery timeframe; and (2) the throughput of the metabolite quantification procedure manually performed by the technician one metabolite at a time for HRMAS NMR spectra obtained from each tissue specimen. Moreover, overlapping metabolite peaks may not be fully separated in the preprocessed ^{1}H spectra [28].

Due to the high-dimensional, complex, and structured nature of the preprocessed HRMAS NMR spectra, data-driven computational modeling (i.e. machine learning) has the promise to be a valuable companion to the above-mentioned manual feedback loop. To this end, various supervised machine learning models were explored for providing automated feedback on pathological classification [29, 30] of the resected glioma tissue specimens. The dataset used for these studies [31] consists of HRMAS NMR spectra of 568 glioma and healthy control tissue samples. It is relatively small despite being the largest of its kind which prohibited the use of complex models and forced the models built to be shallow. This suggests room for improvement in the performance of the feedback pipeline.

From a methodological perspective, the applicability of deep learning models in small data regimes can be improved by making informed architectural choices to simplify models. Incorporating biological prior information may limit the connectivity of the neurons in a model and simultaneously (1) alleviate overfitting when available dataset sizes are small, and (2) aid model interpretability regardless of the dataset size. While there is no application in the metabolomics domain, biologically informed deep learning architectures were successfully utilized in the context of metastasis prediction in patients with prostate cancer based on their genomic profiles [32], and chromatin accessibility prediction based on genome sequences while accounting for known transcription factor motifs during weight initialization step [33].

Furthermore, the pathological classification of tumors may not be the only determinant of multi-faceted and patient-specific cancer progression semantics. In the case of glioblastoma

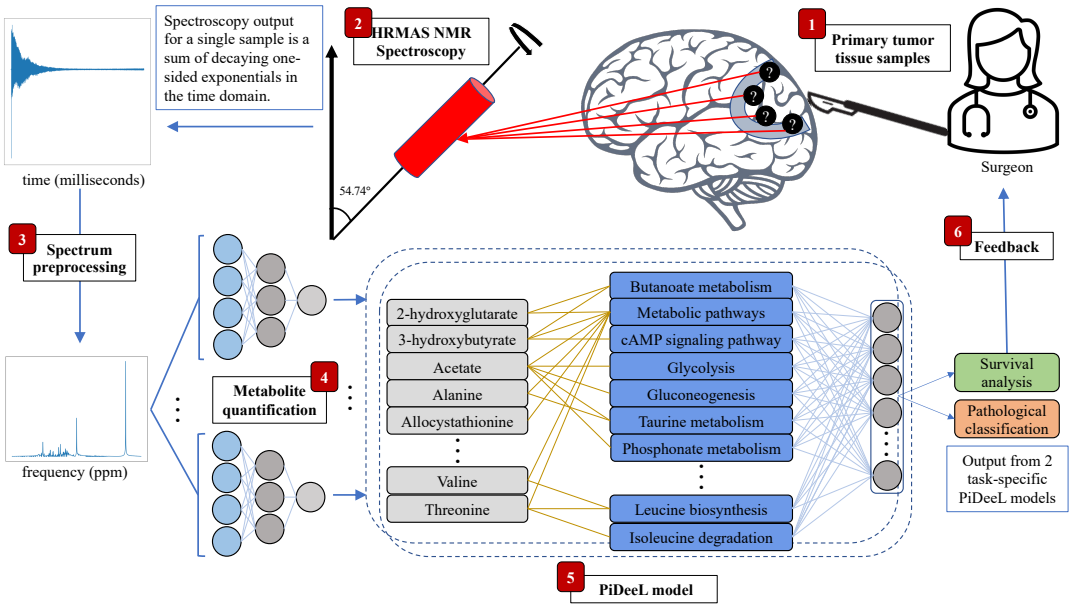

Figure 1: PiDeeL assisted intraoperative feedback loop. (Step 1) The surgeon removes primary tumor tissue from the patient's brain and prepares specimens from the resected tissue. (Step 2) Prepared samples are sent for HRMAS NMR spectroscopy analysis. (Step 3) Time domain signals resulting from the spectroscopy of each sample are preprocessed to uncover metabolite peaks harbored in the corresponding frequency domain spectrum. (Step 4) 37 metabolites are quantified from preprocessed spectra in the order of milliseconds via metabolite-specific 2-layer fully-connected neural networks. (Step 5) PiDeeL transforms metabolite concentrations to survival analysis and pathological classification predictions also in the order of milliseconds. (Step 6) Based on this feedback, the surgeon decides to preempt or continue performing the surgery, again with the guidance of PiDeeL.

multiforme (GBM), a grade IV astrocytic glioma type, a dismal prognosis is often expected due to the highly malignant and invasive nature of this class of tumors [1]. Despite their severity, prolonged and progression-free survival was observed in isocitrate dehydrogenase (IDH) wildtype GBM [34–36]. This finding promotes the inclusion of survival risk estimates in the feedback given to surgeons for more accurate surgical decision-making.

Here, we propose a deep neural network architecture (PiDeel) to predict the survival of the glioma patient and the pathological classification of the tumor using the HRMAS NMR signal to give feedback to the surgeons (Figure 1). While there were methods proposed for automated pathological classification, this is the first study to incorporate these two orthogonal sources of information to the surgeon as feedback at the same time. For the first time in the metabolomics domain, we incorporate prior biological information from the metabolic networks into the neural network architecture and increase the interpretability of the predictions which is also important for the surgeon to understand the underlying biology. We show that PiDeel achieves better performance compared to the state-of-the-art model in pathological classification and, provides survival prediction with a c-index of 68.7%. We also analyze the most important metabolites and metabolic pathways for the reasoning of the model and discuss their relevance in the context of gliomas.

2 Methods

2.1 Dataset

The dataset used in this study is a subset of the patient cohort released at `https://zenodo.org/record/5781769`, and it consists of 568 HRMAS NMR spectra acquired from 458 glioma (80.7% aggressive and 19.3% benign) and 110 control tissue specimens collected during respective surgeries of 513 patients performed at University Hospitals of Strasbourg. For the details

of tissue specimen collection, and ethics statement, please see Section 1.1 of the Supplementary Text.

In order to prepare the dataset for survival analysis, we retain the samples acquired from primary tumor tissue specimens of glioma origin whose record of the date of metabolomics-guided surgery and the date of decease or last control is present. We removed specimens resected from healthy control patients and excision cavities of glioma patients. The resulting set of 384 HRMAS NMR spectra contains one sample per glioma patient and is composed of 299 (77.9%) aggressive and 85 (22.1%) benign tumors whose subtype and grade information are provided in Supplementary Table 1. 70.1% of the patient cohort (269 patients) were reported deceased before the spectra acquisition from tissue specimens (i.e. until January 2021). The remaining 115 patients were considered to be censored as their survival status is not known. The distribution of patient age, time-to-event for deceased, and censored patients are provided in panels A, B, and C of Supplementary Figure S1, respectively. Furthermore, ERETIC-CPMG HRMAS NMR FID signals of these 384 glioma patients, whose spectra were acquired according to the procedure described in Section 1.1.3 of Supplementary Text, are released at `https://zenodo.org/record/7228791`.

2.2 Automated Metabolite Quantification

In this section, we describe the steps applied to the FID signals (output of HRMAS NMR Spectroscopy) to obtain metabolite concentrations, which were used as input to the survival analysis and pathological classification models presented throughout the work. We adopt the deep learning-based automated metabolite quantification method proposed in [30] in order to quantify 37 metabolites of interest from HRMAS NMR spectra. Briefly, we preprocess our samples according to the input specification of their pipeline and quantify each metabolite from the spectra in our sample set with their publicly available quantification models per metabolite from `https://github.com/gunkaynar/targeted_brain_tumor_margin_assessment`. For the complete list of 37 metabolites, details of automated metabolite quantification, and HRMAS NMR spectrum preprocessing, please see Section 1.2 of Supplementary Text.

Formally, let the predicted concentration of a given metabolite j for a sample set of size N be $\widehat{M_j}^{(1:N)} \in \mathbb{R}^{N\times 1}$. We denote the full output of the automated metabolite quantification pipeline for 37 metabolites and a sample set of size N as the concatenation of all predicted quantification vectors across metabolites:

$$\widehat{M}^{(1:N)} = [\widehat{M}_1^{(1:N)}, \widehat{M}_2^{(1:N)}, \cdots, \widehat{M}_{37}^{(1:N)}] \in \mathbb{R}^{N\times 37} \tag{1}$$

2.3 Problem Formulation

Here, we formulate pathological classification and survival analysis tasks based on the full output of the metabolite quantification pipeline described in Section 2.2. as well as tumor malignancy label, patient-specific survival event indicator (i.e. censored or deceased), and time-to-event variables. For a given sample i, we denote the feature vector, tumor malignancy label, survival event indicator, and time-to-event as $\widehat{M}^{(i)}$, $Y_m^{(i)}$, $Y_e^{(i)}$ and $Y_d^{(i)}$, respectively. The feature vector, $\widehat{M}^{(i)} \in \mathbb{R}^{37}$, is the 37-dimensional vector corresponding to the full output of the automated metabolite quantification pipeline. Survival event indicator, $Y_e^{(i)} \in \{0, 1\}$, is a binary variable that indicates whether the sample is from a deceased (i.e. 1) or a censored patient (i.e. 0). Time to survival event, $Y_d^{(i)} \in [0, \infty]$ is a non-negative continuous variable defined as the time (in days) between metabolomics-guided tumor removal surgery and the date of decease or last control for deceased and censored patients, respectively. Finally, the tumor malignancy label, $Y_m^{(i)} \in \{0, 1\}$, indicates whether the sample is from a tumor having aggressive (i.e. 1) or benign (i.e. 0) pathology, similar to the problem formulations presented in [29, 30].

For our formulation of the survival analysis task, we refer to the hazard function denoted as $\lambda(t, \widehat{M}^{(i)})$ for a given sample i where t is time-to-event (i.e. $Y_d^{(i)}$). Assuming proportional hazards from Cox's model [37], the hazard function factorizes [38, 39] to a time-dependent baseline hazard function, $h_0(t)$, and a sample-dependent risk function, $h(\widehat{M}^{(i)})$, as given in Equation 2.

$$\lambda(t, \widehat{M}^{(i)}) = h_0(t) \cdot e^{h(\widehat{M}^{(i)})} \tag{2}$$

We train survival analysis models based on this formulation of hazard function on a dataset of N samples to learn a mapping function $f : \mathbb{R}^{37} \rightarrow [-\infty, \infty]$ such that $f(\widehat{M}^{(1:N)}) = h(\widehat{M}^{(1:N)})$. Here, the goal of the trained models is to predict risk scores for glioma patients based on their metabolic profile. In order to infer these models, we make use of $Y_e^{(i)}$ and $Y_d^{(i)}$ as labels.

On the other hand, we train pathological classification models on a dataset of N samples to learn a mapping function $g : \mathbb{R}^{37} \rightarrow \{0, 1\}$ such that $g(\widehat{M}^{(1:N)}) = Y_m^{(1:N)}$ (a binary classification problem similar to [29, 30]).

2.4 Baseline Models

In this section, we provide the details of models benchmarked against PiDeel for survival analysis and pathological classification tasks formulated in Section 2.3.

2.4.1 Survival Analysis

For the survival analysis task, we consider Cox Proportional Hazards (Cox-PH), Random Survival Forest (RSF), and DeepSurv [39] models as baselines.

Cox-PH model is a linear model that assumes proportional hazards to factorize instantaneous hazard function as a product of a baseline hazard function and a sample-dependent risk function, $h(\widehat{M}^{(i)})$, as formulated in Section 2.3. Based on the formulation provided in [37], the risk function estimate for sample i, $\widehat{h}_\beta(\widehat{M}^{(i)})$, is provided by the linear model given in Equation 3.

$$\widehat{h}_\beta(\widehat{M}^{(i)}) = \beta^t \widehat{M}^{(i)} \tag{3}$$

where $\widehat{M}^{(i)}$ is the feature vector of metabolite concentrations for sample i, and β is the coefficient vector. The model is fit by maximizing Cox partial likelihood [37–39] given in Equation 4.

$$L(\beta) = \prod_{i:Y_e^{(i)}=1} \frac{exp(\widehat{h}_\beta(\widehat{M}^{(i)}))}{\sum_{j:Y_d^{(j)}>Y_d^{(i)}} exp(\widehat{h}_\beta(\widehat{M}^{(j)})))} \tag{4}$$

RSF [40] is a nonlinear ensemble method that constructs a predefined number of survival trees using a random subset of features and bootstraps of training data to estimate the cumulative hazard function, $\Lambda(t)$, defined in Equation 5. Internal nodes of survival trees are split in order to maximize log-rank test [41, 42] statistic between resulting child nodes.

$$\Lambda(t) = \int_0^t \lambda(u)\, du \tag{5}$$

Neural network-based approaches to survival analysis are introduced by Faraggi et al.,1995 [38]. Deeper variants of these models are more recently proposed as DeepSurv model [39] which is a multi-layer perceptron that models sample-specific risk function, $h(\widehat{M}^{(i)})$, as formulated in Section 2.3. The model has a single output neuron with linear activation and adopts a negative log of Cox partial likelihood given in Equation 6 as the loss function.

$$log(L(\beta)) = -\sum_{i:Y_e^{(i)}=1} (\widehat{h}_\beta(\widehat{M}^{(i)}) - log \sum_{j:Y_d^{(j)}>Y_d^{(i)}} e^{(\widehat{h}_\beta(\widehat{M}^{(j)}))}) \tag{6}$$

We consider 2, 3, and 4-layer DeepSurv as our baseline models. We use a single neuron in the last layer with linear activation to output the predicted risk scores of samples. The trained 3-layer DeepSurv model on our dataset of size N can be summarized as follows:

$$X_{hid}^{(1:N)} = ReLU(FC^{(64)}(\widehat{M}^{(1:N)})) \tag{7}$$

$$X_{hid}^{(1:N)} = ReLU(FC^{(64)}(X_{hid}^{(1:N)})) \tag{8}$$

$$\widehat{h}^{(1:N)} = FC^{(1)}(X_{hid}^{(1:N)}) \tag{9}$$

where X_{hid} represents the output of the first hidden layer, $\widehat{h}^{(i)}$ represents the predicted risk score, $FC^{(\cdot)}$ represents a fully-connected layer with $(\cdot)$ neurons, and $ReLU$ represents the rectified linear unit. In the case of 2 and 4-layer DeepSurv models, we remove or repeat the hidden layer defined in Equation 8.

2.4.2 Pathological Classification

We consider random forest (RF) from [30] and fully-connected neural networks as baseline models for the pathological classification task.

We consider neural network-based models for the pathological classification task as baseline models. We train 2, 3, and 4-layer fully-connected neural networks with class-weighted binary cross entropy loss. The trained 3-layer fully-connected networks on our dataset of size N can be summarized as follows:

$$X_{hid}^{(1:N)} = ReLU(FC^{(64)}(\widehat{M}^{(1:N)})) \tag{10}$$

$$X_{hid}^{(1:N)} = ReLU(FC^{(64)}(X_{hid}^{(1:N)})) \tag{11}$$

$$\widehat{Y_m}^{(1:N)} = Sigmoid(FC^{(1)}(X_{hid}^{(1:N)})) \tag{12}$$

where X_{hid} represents the output of the first hidden layer, $\widehat{Y_m}^{(1:N)}$ represents the predicted tumor pathology, $FC^{(\cdot)}$ represents a fully-connected layer with $(\cdot)$ neurons, and $ReLU$ represents the rectified linear unit. We also consider 2 and 4-layer fully-connected networks where we remove or repeat the hidden layer defined in Equation 11, respectively.

2.5 Pathway-informed Deep Learning Model (PiDeel)

In this section, we introduce PiDeeL, a pathway-informed multi-layer perceptron model designed to integrate biomarker metabolite concentrations provided as input at the metabolic pathway level. We use the Kyoto Encyclopedia of Genes and Genomes (KEGG) database to parse the metabolite-to-pathway mapping and construct a 37x138 matrix, hereafter PI_matrix, where the value in the ith row and j^{th} column of matrix PI_matrix is 1 if the metabolite i appears in the pathway j, and 0 otherwise. Similar to the baseline fully-connected neural network models for the survival analysis and pathological classification tasks, we consider 2, 3, and 4-layer PiDeeL models.

2.5.1 Model Architecture

Given a dataset of N samples, the architecture of the 3-layer PiDeeL model for the survival analysis can be summarized as follows:

$$PL^{(138)} = \left[\ FC^{(138)}weights\ \right]\left[PI_matrix\right] \tag{13}$$

$$PA^{(1:N)} = ReLU(PL^{(138)}(\widehat{M}^{(1:N)}) + b^{138}) \tag{14}$$

$$X_{hid}^{(1:N)} = ReLU(FC^{(64)}(PA^{(1:N)})) \tag{15}$$

$$\widehat{h}^{(1:N)} = FC^{(1)}(X_{hid}^{(1:N)}) \tag{16}$$

where $PL^{(138)}$ and b^{138} represent the weights and biases of the pathway-informed layer, $PA^{(1:N)}$ represents the activations of the first hidden layer (hereafter pathway-activation), X_{hid} represents the output of hidden layers, $\widehat{h}^{(i)}$ represents the predicted risk score, $FC^{(\cdot)}$ represents a fully-connected layer with $(\cdot)$ neurons and $ReLU$ represents the rectified linear unit. We incorporate the metabolic pathway information into PiDeeL by defining the first layer of the model using PI_matrix. This architectural decision makes the first layer of PiDeeL sparse since the connections in the first layer are only permitted to be formed between input metabolites and the pathways they are involved in. We use the negative log of Cox Partial Likelihood as the loss function, similar to the DeepSurv models formulated in Section 2.4.1, in order to train PiDeeL for the survival analysis task.

The architecture of PiDeeL used for survival analysis and pathological classification tasks differs only in the output layer (i.e. Equation 16). In order to adapt PiDeeL for pathological classification we use the sigmoid function as the output activation function by replacing Equation 16 with Equation 17. Additionally, we use class-weighted binary cross entropy as the loss function.

$$\widehat{Y_m}^{(1:N)} = Sigmoid(FC^{(1)}(X_{hid}^{(1:N)})) \tag{17}$$

2.5.2 Interpretability and Importance Analyses

We perform SHapley Additive exPlanations (SHAP) analysis on our feature vector ($\widehat{M}^{(1:N)}$), and pathway activations ($PA^{(1:N)}$) to assess the importance of each metabolite and pathway on the prediction of PiDeeL. For the details of these analyses, please see Section 1.4.3 of the Supplementary Text.

2.5.3 Ablation Studies for Pathway-informed Architecture

In order to verify that the improvements achieved by the PiDeeL model over baseline models are due to the pathway-informed architecture, we conduct a series of ablation tests as follows:

- We train fully-connected models with 138 neurons in the first hidden layer to check whether the number of neurons in the first layer (138) of PiDeeL contributes to the performance.
- We train fully-connected models with different dropout rates $(0.5, 0.6, 0.7, 0.8, 0.9)$ in the first hidden layer to check whether the pathway-informed layer has a benefit over this widely used technique to regularize neural networks.
- We train PiDeeL models using randomly connected $PI{-}matrices$ instead of $PI{-}matrices$ that are constructed from the KEGG database to check whether the pathway-informed connections has a benefit over same number of randomly-connected edges.
- We train PiDeeL models using shuffled $PI - matrices$. This is similar to randomly connected $PI - matrices$ except it preserves the in-degrees of pathways.

Please refer to Section 1.4.4 of the Supplementary Text for the details of ablation studies for pathway-informed architecture based on survival analysis and pathological classification tasks.

3 Results

3.1 Experimental Setup

In this section, we describe the experimental setup used for evaluating PiDeeL as well as baseline models formulated in Section 2.4 in the context of survival analysis and pathological classification scenarios. We measure the performance of survival analysis models in the unit of time-dependent concordance index (c-index) described in [43]. The original concordance index [44] is computed with the individually predicted patient risk scores, whereas, with a time-dependent definition, individual risk scores are calculated with the predicted risk function and the baseline hazard function that is computed on the sample set [43, 45].

We assess the performance of the pathological classification models with respect to the Area Under the ROC Curve (AUC-ROC) and the Area Under the Precision-Recall Curve (AUC-PR) metrics. For each task and model type, we adopt a 5-fold cross-validation routine and repeat it 3 times with different seeds (i.e. 15 iterations). Before each iteration, we shuffle and divide our dataset into training, validation, and test set such that no sample or patient is shared across folds. The results discussed hereafter are with respect to the predictive capability of models on the test split. During iteration i of the cross-validation setup, Cox-PH and RSF models were tested on the fold i, validated on fold $(i+1) mod 5$, and trained on the remaining 3 folds. We re-train the selected configuration on the union of training and validation splits to get stronger baseline models. For all deep learning models, re-training steps are not performed due to the presence of early-stopping dynamics. Please see Section 1.3 of Supplementary Text for hyperparameter space and selection details.

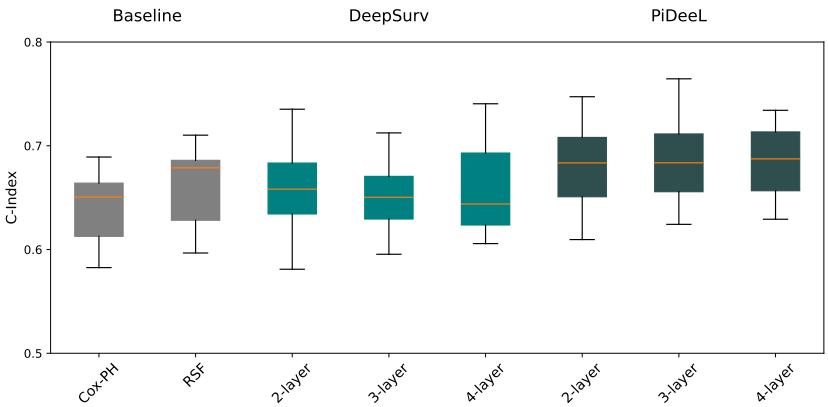

Figure 2: Comparison of baseline models and PiDeeL on survival analysis in terms of c-index. This evaluation is based on 5-fold cross-validation that is repeated 3 times. (i.e., 15 iterations).

3.2 Survival Analysis

Please see Figure 2 for the survival analysis performance of Cox-PH and RSF (gray boxplots) as well as DeepSurv (cyan boxplots) and PiDeeL (navy blue boxplots) with respect to the c-index. We observe that 4-layer PiDeeL outperforms baselines and DeepSurv models on survival prediction and achieves a median c-index of 68.7%. We also see that the performance slightly increases when the depth of PiDeeL is increased from 2 to 4 (improvement up to median c-index of 0.4%). On the other hand, the performance of DeepSurv models diminishes with increasing model depth (deterioration up to a median c-index of 1.5%). Furthermore, when DeepSurv and PiDeeL with the same number of layers are compared, we see the performance gap increase with respect to the number of layers in the model. Here, we quantitatively observe the benefit of incorporating biological information in modeling choices. For 2, 3, and 4-layered networks, PiDeeL improves upon DeepSurv with the same model depth by 2.5%, 3.4%, and 4.3% according to the median c-index, respectively. This supports the robustness of sparse architectures in small data regimes for the survival analysis task. When the performance of baseline models is considered in this setup, RSF appears to be a particularly strong baseline with a median c-index of 67.8%. This is also expected since the power of random forest-based models on HRMAS-NMR data has been demonstrated previously on a cohort of similar size [29,30]. Still, PiDeeL achieves better performance than RSF also providing more interpretability by enabling pathway-activation analysis. Finally, the Cox-PH model performs substantially worse than RSF, and PiDeeL. Please see Section 1.4.1 of the Supplementary Text for the results of the pathological classification.

3.3 Interpretability and Importance Analyses

When combined with downstream feature importance analysis methods, (SHAP [46]), the pathway-informed architecture of PiDeeL enables us to jointly analyze the importance of metabolites and pathways for the model output. In addition to the quantitative analyses presented above, we perform feature importance analyses on the 4-layer PiDeeL to empirically pinpoint metabolites and pathways candidate for being implicated in gliomas, their progression, and associated patient survival. We calculate SHAP values on the input features (i.e. in silico predicted metabolite concentrations - $\widehat{M}^{(1:N)}$) as well as pathway-informed layer activations ($PA^{(1:N)}$).

Please see Supplementary Figures S4 and S5 where we show PiDeel's metabolite importance plots for the survival prediction and tumor pathological classification tasks. We highlight the top 3 metabolites having the highest mean SHAP value for the survival analysis among all 37 metabolites according to our analyses in the decreasing order of importance *alanine*, *glutamine*, and *glutamate*. We also pinpoint the top 3 pathways having the highest SHAP values among the 138 pathways used for pathway-informed layer, in the decreasing order of importance as: *Mineral absorption*, *mTOR signaling pathway* and *Alanine - aspartate and glutamate metabolism*.

Through a case-control study, alanine was reported to be a potential biomarker for malignant gliomas [47]. Similarly, glutamate was also found to play a central role in malignant gliomas through multiple mechanisms [48]. Since malignant gliomas usually have a poor prognosis and survival, the high ranking of alanine and glutamate in the list are expected. Furthermore, glutamine is the primary precursor of glutamate and plays a critical role in brain function [49]. Glutamine metabolism can regulate multiple pathways including energy production and redox homeostasis in brain cancer cells [49]. Glutamine itself was pinpointed as a candidate biomarker for glioma progression and response to treatment [50]. Thus, the rank of glutamine in the list is also expected.

We biologically motivate the connection between the top 3 pathways and gliomas. Minerals and their absorption are fundamentally important to sustain life [51]. For example, calcium plays a wide variety of roles in the human body. Moreover, various natural dietary components were reported to have potential benefits in the prevention and treatment of gliomas [52]. Calcium was reported to be one of the dietary components having a potential for glioma prevention by affecting apoptosis and DNA repair [53]. Thus, the rank of *Mineral absorption* pathway according to calculated importance is biologically grounded. Furthermore, the mammalian target of rapamycin (mTOR), is reported to have a key role in integrating signal transduction and metabolic pathways in glioblastoma [54]. Since $\sim$ 46% of the cohort analyzed in this work were diagnosed with glioblastoma, the importance rank of *mTOR signaling pathway* is viable. Finally, two of the three metabolites implicated in *Alanine— aspartate and glutamate metabolism* pathway, were also detected among the top 3 important metabolites with respect to metabolite-level importance scores, as discussed above. This qualitatively reinforces the predictive capabilities of PiDeeL and showcases its interpretability in both metabolite and pathway levels.

3.4 Ablation Studies for Pathway-informed Architecture

In this section, we present the results of ablation studies discussed in Section 2.5.3. In Figure 3, each panel demonstrates the result of the respective itemized ablation test.

Panel A of Figure 3 shows the survival prediction performance attained by 2, 3, or 4-layer DeepSurv (cyan boxplots) models with 138 neurons in the first layer, and respective PiDeeL models (navy blue boxplots). We observe that PiDeeL improves the median c-index over DeepSurv with 138 neurons in the first layer by 3.6%, 2.6%, and 3.6% for 2, 3, and 4-layered networks, respectively. We also see that using PiDeeL over DeepSurv, decreases the variance by an average of 29.7%. Note that the original DeepSurv model has 64 neurons in the first layer.

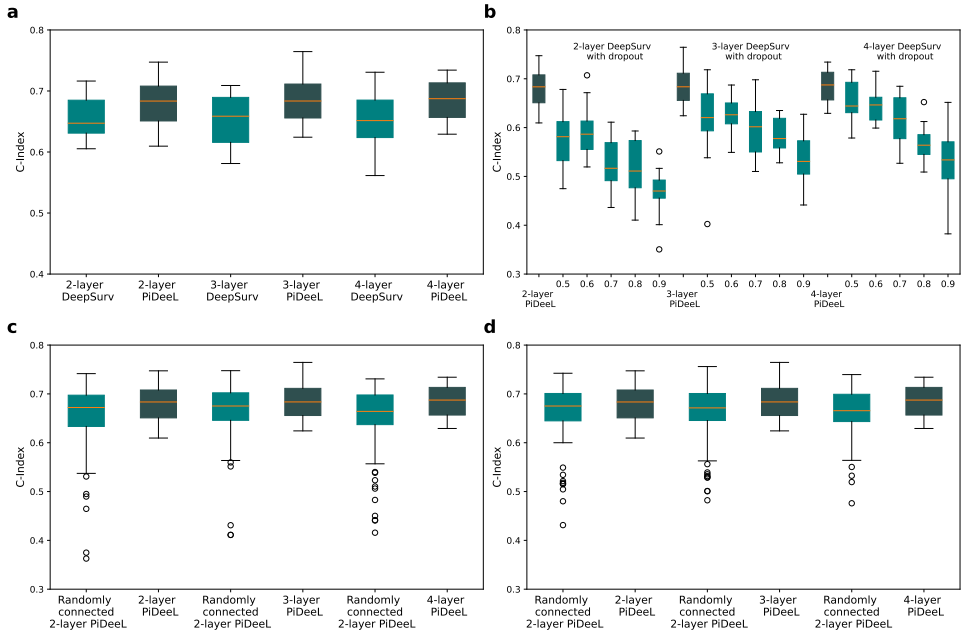

Figure 3: Performance comparison of PiDeeL on survival analysis task against (a) DeepSurv with 138 neurons in the first hidden layer, (b) DeepSurv with different dropout rates, (c) PiDeeL with randomly connected PI_matrix, and (d) PiDeeL with shuffled PI_matrix, with respect to c-index.

This shows that not the larger number of neurons in the first layer but the pathway information leads to the performance improvement over DeepSurv.

Panel B of Figure 3 shows the survival prediction performance comparison of PiDeel and DeepSurv with various dropout rates and PiDeeL. We observe that PiDeeL substantially outperforms DeepSurv-with-dropout regardless of the rate. We also see that DeepSurv's performance deteriorates when the dropout rate is increased. The 2-layer PiDeeL model achieves a median c-index of 68.3%, in contrast, the DeepSurv model attains median c-index scores of 58.1%, 58.6%, 51.6%, 51.1% and 47.0% for 0.5, 0.6, 0.7, 0.8 and 0.9 dropout rates, respectively. We observe the same trend as we increase the number of layers for both models. This indicates that sparsity provided by the dropout does not contribute to the performance of the fully-connected model for survival prediction, whereas PiDeeL's pathway-informed sparsity with PI_matrix improves the performance.

In Panel C of Figure 3, we compare the performance of PiDeeL with PI_matrix against PiDeeL with $pseudo_PI_matrices$. In this setup, we randomly connect our metabolite nodes (input vector) and pathway activation layer neurons. For this, we constructed randomly connected 37x138 matrices preserving the total number of connections in the original PI_matrix (468). While the improvements are not statistically significant with respect to the Kolmogorov-Smirnov test (p-value 0.101), we observe that PiDeeL with PI_matrix outperforms PiDeeL with $pseudo_PI_matrices$ using 2, 3, and 4-layer networks. Using PI_matrix improves the performance attained by using $pseudo_PI_matrices$ in terms of median c-index by 1.44% on average. We also notice a substantial increase in the variance of the results when $pseudo_PI_matrices$ are used as opposed to PI_matrix by an average of 36.2%. The minimum c-index values obtained using $pseudo_PI_matrices$ are below and often close to 0.4 which means the performance is random or worse than random, respectively. Whereas the minimum c-index values attained by PiDeel are always above 0.6. This shows that the utilization of PI_matrix in the model serves to enhance its robustness and simultaneously decrease the variance.

In Panel D of Figure 3, we show the performance comparison of PiDeeL with PI_matrix

against PiDeeL with *pseudo_PI_matrices*. However, this time we construct *PI_matrix* and shuffle its rows to preserve each metabolite-to-pathway profile. We observe a similar trend as in Panel D. This time, *PI_matrix* improves the performance attained by using *pseudo_PI_matrices* on average in terms of median c-index by 1.43%. Additionally, using *PI_matrix* decreases the variance by an average of 28.1%.

For the ablation studies for pathway-informed architecture performed on the pathological classification task, please refer to Section 1.4.4 of the Supplementary Text.

4 Discussion

1H HRMAS NMR spectrum has been a reliable resource for distinguishing malign and healthy samples taken from the excision cavity during surgery. It is a good fit because it is non-destructive and can analyze small samples of raw tissue specimens [27]. Machine learning techniques that learn from the NMR signal to distinguish healthy and tumor tissue, as well as benign and aggressive tumors, have been an important tool. Yet, the performance was limited due to small training set sizes [29, 30] which prohibit using complex architectures like deep neural networks learning a hierarchical composition of complex features. In this study we are able to use such a hierarchically complex model for distinguishing benign and aggressive tumors for the first time with comparable performance. Despite having a deep model, we are able to decrease the number of trainable parameters in the model by composing the architecture with prior biological information obtained from metabolic networks.

The pathology of the tumor is an important cue for the surgeon who might want to resect more tumor tissue and risk the patient's well-being if the tumor is likely aggressive, or who might risk leaving some residual tumor tissue if the tumor is likely benign. Yet, this does not one-to-one determine the prognosis. Thus, it is also important to estimate the survival of a patient, which is a proxy for prognosis, and inform the surgeon. For this, we train a PiDeel model to predict the survival of patients and show that we attain high performance. An example showing the usefulness of the survival prediction along with pathological classification is as follows: Our model is able to correctly predict long survival for a patient with aggressive glioma, who lived 5,656 days after surgery. The surgeon can be more conservative in this case compared to being informed with only a prediction of an aggressive tumor. Similarly, our model correctly predicts low survival for a patient with a benign tumor who lived only 43 days after surgery.

We also investigate the use of a multi-task learning architecture that combines two architectures and predicts both labels simultaneously. Our results demonstrate that this approach only slightly increases the performance for both tasks. For this reason, we ultimately decided to proceed with the utilization of two task-specific single-task PiDeeL models.

One limitation of our study is that it is bottlenecked by the automated metabolite quantification model which has been shown to have promising performance and generalizability. Yet, this component consists of individual multi-layered perceptrons for 37 metabolites for which a training dataset exists. We think our model would benefit from a more comprehensive spectrum of metabolites to predict the survival of a patient which is a complex outcome to approximate. The human metabolome roughly consists of 3,000 metabolites and we are currently limited to using only 1% of this information source.

Acknowledgements

This work was supported by grants from BPI France (ExtempoRMN Project), Hôpitaux Universitaires de Strasbourg, Bruker BioSpin, Univ. de Strasbourg and the Centre National de la Recherche Scientifique; also by TUBA GEBIP, Bilim Akademisi BAGEP and TUSEB Research Incentive awards to AEC.

References

[1] Omuro, A. & DeAngelis, L. M. Glioblastoma and other malignant gliomas: a clinical review. *Jama* **310**, 1842–1850 (2013).

[2] Osborn, A., Louis, D., Poussaint, T., Linscott, L. & Salzman, K. The 2021 world health organization classification of tumors of the central nervous system: What neuroradiologists need to know. *American Journal of Neuroradiology* **43**, 928–937 (2022).

[3] Ostrom, Q. T. *et al.* Cbtrus statistical report: primary brain and other central nervous system tumors diagnosed in the united states in 2010–2014. *Neuro-oncology* **19**, v1–v88 (2017).

[4] Stummer, W., van den Bent, M. J. & Westphal, M. Cytoreductive surgery of glioblastoma as the key to successful adjuvant therapies: new arguments in an old discussion. *Acta neurochirurgica* **153**, 1211–1218 (2011).

[5] Brown, M. V. *et al.* Cancer detection and biopsy classification using concurrent histopathological and metabolomic analysis of core biopsies. *Genome medicine* **4**, 33 (2012).

[6] Schafer, K.-C. *et al.* Real time analysis of brain tissue by direct combination of ultrasonic surgical aspiration and sonic spray mass spectrometry. *Analytical Chemistry* **83**, 7729–7735 (2011).

[7] Eberlin, L. S. *et al.* Ambient mass spectrometry for the intraoperative molecular diagnosis of human brain tumors. *Proceedings of the National Academy of Sciences* **110**, 1611–1616 (2013).

[8] Pirro, V. *et al.* Intraoperative assessment of tumor margins during glioma resection by desorption electrospray ionization-mass spectrometry. *Proceedings of the National Academy of Sciences* **114**, 6700–6705 (2017).

[9] Santagata, S. *et al.* Intraoperative mass spectrometry mapping of an onco-metabolite to guide brain tumor surgery. *Proceedings of the National Academy of Sciences* **111**, 11121–11126 (2014).

[10] Calligaris, D. *et al.* Molecular typing of meningiomas by desorption electrospray ionization mass spectrometry imaging for surgical decision-making. *International journal of mass spectrometry* **377**, 690–698 (2015).

[11] Calligaris, D. *et al.* Maldi mass spectrometry imaging analysis of pituitary adenomas for near-real-time tumor delineation. *Proceedings of the National Academy of Sciences* **112**, 9978–9983 (2015).

[12] Jarmusch, A. K. *et al.* Lipid and metabolite profiles of human brain tumors by desorption electrospray ionization-ms. *Proceedings of the National Academy of Sciences* **113**, 1486–1491 (2016).

[13] Fatou, B. *et al.* In vivo real-time mass spectrometry for guided surgery application. *Scientific reports* **6**, 1–14 (2016).

[14] Stummer, W. *et al.* Fluorescence-guided surgery with 5-aminolevulinic acid for resection of malignant glioma: a randomised controlled multicentre phase iii trial. *The lancet oncology* **7**, 392–401 (2006).

[15] Tsugu, A. *et al.* Impact of the combination of 5-aminolevulinic acid–induced fluorescence with intraoperative magnetic resonance imaging–guided surgery for glioma. *World neurosurgery* **76**, 120–127 (2011).

[16] Colditz, M. J. & Jeffree, R. L. Aminolevulinic acid (ala)–protoporphyrin ix fluorescence guided tumour resection. part 1: Clinical, radiological and pathological studies. *Journal of Clinical Neuroscience* **19**, 1471–1474 (2012).

[17] Montcel, B., Mahieu-Williame, L., Armoiry, X., Meyronet, D. & Guyotat, J. Two-peaked 5-ala-induced ppix fluorescence emission spectrum distinguishes glioblastomas from low grade gliomas and infiltrative component of glioblastomas. *Biomedical optics express* **4**, 548–558 (2013).

[18] Li, Y., Rey-Dios, R., Roberts, D. W., Valdés, P. A. & Cohen-Gadol, A. A. Intraoperative fluorescence-guided resection of high-grade gliomas: a comparison of the present techniques and evolution of future strategies. *World neurosurgery* **82**, 175–185 (2014).

[19] Lu, F.-K. *et al.* Label-free neurosurgical pathology with stimulated raman imaging. *Cancer research* **76**, 3451–3462 (2016).

[20] Jermyn, M. *et al.* Highly accurate detection of cancer in situ with intraoperative, label-free, multimodal optical spectroscopy. *Cancer research* **77**, 3942–3950 (2017).

[21] Jermyn, M. *et al.* Raman spectroscopy detects distant invasive brain cancer cells centimeters beyond mri capability in humans. *Biomedical optics express* **7**, 5129–5137 (2016).

[22] Chan, D. T. M., Sonia, H. Y.-P. & Poon, W. S. 5-aminolevulinic acid fluorescence guided resection of malignant glioma: Hong kong experience. *Asian journal of surgery* **41**, 467–472 (2018).

[23] Orringer, D. A. *et al.* Rapid intraoperative histology of unprocessed surgical specimens via fibre-laser-based stimulated raman scattering microscopy. *Nature biomedical engineering* **1**, 0027 (2017).

[24] Poulon, F. *et al.* Optical properties, spectral, and lifetime measurements of central nervous system tumors in humans. *Scientific reports* **7**, 1–8 (2017).

[25] Hollon, T. C. *et al.* Rapid intraoperative diagnosis of pediatric brain tumors using stimulated raman histology. *Cancer research* **78**, 278–289 (2018).

[26] Xue, Z. *et al.* Fluorescein-guided surgery for pediatric brainstem gliomas: Preliminary study and technical notes. *Journal of Neurological Surgery Part B: Skull Base* **79**, S340–S346 (2018).

[27] Gogiashvili, M. *et al.* Hr-mas nmr based quantitative metabolomics in breast cancer. *Metabolites* **9**, 19 (2019).

[28] Karakaslar, E. O., Coskun, B., Outilaft, H., Namer, I. J. & Cicek, A. E. Predicting carbon spectrum in heteronuclear single quantum coherence spectroscopy for online feedback during surgery. *IEEE/ACM transactions on computational biology and bioinformatics* **17**, 719–725 (2019).

[29] Cakmakci, D. *et al.* Machine learning assisted intraoperative assessment of brain tumor margins using hrmas nmr spectroscopy. *PLoS Computational Biology* **16**, e1008184 (2020).

[30] Cakmakci, D. *et al.* Targeted metabolomics analyses for brain tumor margin assessment during surgery. *Bioinformatics* (2022).

[31] Cakmakci, D. *et al.* 1h hrmas nmr dataset of glioma and control samples (2020). URL `https://doi.org/10.5281/zenodo.3951448`.

[32] Elmarakeby, H. A. *et al.* Biologically informed deep neural network for prostate cancer discovery. *Nature* **598**, 348–352 (2021).

[33] Novakovsky, G., Dexter, N., Libbrecht, M. W., Wasserman, W. W. & Mostafavi, S. Obtaining genetics insights from deep learning via explainable artificial intelligence. *Nature Reviews Genetics* 1–13 (2022).

[34] Poon, M. T., Sudlow, C. L., Figueroa, J. D. & Brennan, P. M. Longer-term (2 years) survival in patients with glioblastoma in population-based studies pre-and post-2005: a systematic review and meta-analysis. *Scientific reports* **10**, 1–10 (2020).

[35] Burgenske, D. M. *et al.* Molecular profiling of long-term idh-wildtype glioblastoma survivors. *Neuro-oncology* **21**, 1458–1469 (2019).

[36] Wong, D. & Yip, S. Finding a four-leaf clover—identifying long-term survivors in idh-wildtype glioblastoma. *Neuro-oncology* **21**, 1352 (2019).

[37] Cox, D. R. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)* **34**, 187–202 (1972).

[38] Faraggi, D. & Simon, R. A neural network model for survival data. *Statistics in medicine* **14**, 73–82 (1995).

[39] Katzman, J. L. *et al.* Deepsurv: personalized treatment recommender system using a cox proportional hazards deep neural network. *BMC medical research methodology* **18**, 1–12 (2018).

[40] Ishwaran, H., Kogalur, U. B., Blackstone, E. H. & Lauer, M. S. Random survival forests. *The annals of applied statistics* **2**, 841–860 (2008).

[41] Segal, M. R. Regression trees for censored data. *Biometrics* 35–47 (1988).

[42] LeBlanc, M. & Crowley, J. Relative risk trees for censored survival data. *Biometrics* 411–425 (1992).

[43] Antolini, L., Boracchi, P. & Biganzoli, E. A time-dependent discrimination index for survival data. *Statistics in medicine* **24**, 3927–3944 (2005).

[44] Harrell, F. E., Califf, R. M., Pryor, D. B., Lee, K. L. & Rosati, R. A. Evaluating the yield of medical tests. *Jama* **247**, 2543–2546 (1982).

[45] Gerds, T. A., Kattan, M. W., Schumacher, M. & Yu, C. Estimating a time-dependent concordance index for survival prediction models with covariate dependent censoring. *Statistics in medicine* **32**, 2173–2184 (2013).

[46] Lundberg, S. M. & Lee, S.-I. A unified approach to interpreting model predictions. In Guyon, I. *et al.* (eds.) *Advances in Neural Information Processing Systems 30*, 4765–4774 (Curran Associates, Inc., 2017). URL `http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf`.

[47] Firdous, S. *et al.* Dysregulated alanine as a potential predictive marker of glioma—an insight from untargeted hrmas-nmr and machine learning data. *Metabolites* **11**, 507 (2021).

[48] de Groot, J. & Sontheimer, H. Glutamate and the biology of gliomas. *Glia* **59**, 1181–1189 (2011).

[49] Natarajan, S. K. & Venneti, S. Glutamine metabolism in brain tumors. *Cancers* **11**, 1628 (2019).

[50] Ekici, S., Risk, B. B., Neill, S. G., Shu, H.-K. & Fleischer, C. C. Characterization of dysregulated glutamine metabolism in human glioma tissue with 1h nmr. *Scientific reports* **10**, 1–10 (2020).

[51] Kegg pathway: Mineral absorption - reference pathway (ko) (usp.br). `https://www.genome.jp/pathway/hsa04978`. Accessed: 2022-10-16.

[52] Kyritsis, A. P., Bondy, M. L. & Levin, V. A. Modulation of glioma risk and progression by dietary nutrients and antiinflammatory agents. *Nutrition and cancer* **63**, 174–184 (2011).

[53] Tedeschi-Blok, N., Schwartzbaum, J., Lee, M., Miike, R. & Wrensch, M. Dietary calcium consumption and astrocytic glioma: the san francisco bay area adult glioma study, 1991-1995. *Nutrition and cancer* **39**, 196–203 (2001).

[54] Akhavan, D., Cloughesy, T. F. & Mischel, P. S. mtor signaling in glioblastoma: lessons learned from bench to bedside. *Neuro-oncology* **12**, 882–889 (2010).