cancers

MDPI

Article

Machine Learning-Based Analysis of Glioma Grades Reveals Co-Enrichment

Mateusz Garbulowski [1,2,*], Karolina Smolinska [1], Uğur Çabuk [1,3,4], Sara A. Yones [1], Ludovica Celli [1,5,6], Esma Nur Yaz [1,7], Fredrik Barrenäs [1,8], Klev Diamanti [1,9], Claes Wadelius [9] and Jan Komorowski [1,8,10,11,*]

1 Department of Cell and Molecular Biology, Uppsala University, 752 37 Uppsala, Sweden; karolina.smolinska@icm.uu.se (K.S.); ugur.cabuk@awi.de (U.Ç.); sara.younes@icm.uu.se (S.A.Y.); ludovica.celli@igm.cnr.it (L.C.); esma.yaz@std.medipol.edu.tr (E.N.Y.); fredrik.barrenas@icm.uu.se (F.B.); klev.diamanti@igp.uu.se (K.D.)
2 Science for Life Laboratory, Department of Biochemistry and Biophysics, Stockholm University, 106 91 Solna, Sweden
3 Polar Terrestrial Environmental Systems, Alfred Wegener Institute Helmholtz Centre for Polar and Marine Research, 14473 Potsdam, Germany
4 Institute of Biochemistry and Biology, University of Potsdam, 14469 Potsdam, Germany
5 Institute of Molecular Genetics Luigi Luca Cavalli-Sforza, National Research Council, 27100 Pavia, Italy
6 Department of Biology and Biotechnology, University of Pavia, 27100 Pavia, Italy
7 Department of Biomedical Engineering and Bioinformatics, The Graduate School of Engineering and Natural Sciences, Istanbul Medipol University, Istanbul 34810, Turkey
8 Washington National Primate Research Center, Seattle, WA 98195, USA
9 Department of Immunology, Genetics and Pathology, Uppsala University, 751 85 Uppsala, Sweden; claes.wadelius@igp.uu.se
10 Swedish Collegium for Advanced Study, 752 38 Uppsala, Sweden
11 Institute of Computer Science, Polish Academy of Sciences, 01-248 Warsaw, Poland
* Correspondence: mateuszgarbulowski@gmail.com (M.G.); jan.komorowski@icm.uu.se (J.K.)

Simple Summary: Gliomas are heterogenous types of cancer, therefore the therapy should be personalized and targeted toward specific pathways. We developed a methodology that corrected strong batch effects from The Cancer Genome Atlas datasets and estimated glioma grade-specific co-enrichment mechanisms using machine learning. Our findings created hypotheses for annotations, e.g., pathways, that should be considered as therapeutic targets.

Abstract: Gliomas develop and grow in the brain and central nervous system. Examining glioma grading processes is valuable for improving therapeutic challenges. One of the most extensive repositories storing transcriptomics data for gliomas is The Cancer Genome Atlas (TCGA). However, such big cohorts should be processed with caution and evaluated thoroughly as they can contain batch and other effects. Furthermore, biological mechanisms of cancer contain interactions among biomarkers. Thus, we applied an interpretable machine learning approach to discover such relationships. This type of transparent learning provides not only good predictability, but also reveals co-predictive mechanisms among features. In this study, we corrected the strong and confounded batch effect in the TCGA glioma data. We further used the corrected datasets to perform comprehensive machine learning analysis applied on single-sample gene set enrichment scores using collections from the Molecular Signature Database. Furthermore, using rule-based classifiers, we displayed networks of co-enrichment related to glioma grades. Moreover, we validated our results using the external glioma cohorts. We believe that utilizing corrected glioma cohorts from TCGA may improve the application and validation of any future studies. Finally, the co-enrichment and survival analysis provided detailed explanations for glioma progression and consequently, it should support the targeted treatment.

Keywords: glioma; machine learning; batch effect; TCGA; co-enrichment; rough sets

Citation: Garbulowski, M.; Smolinska, K.; Çabuk, U.; Yones, S.A.; Celli, L.; Yaz, E.N.; Barrenäs, F.; Diamanti, K.; Wadelius, C.; Komorowski, J. Machine Learning-Based Analysis of Glioma Grades Reveals Co-Enrichment. *Cancers* **2022**, *14*, 1014. https://doi.org/10.3390/cancers14041014

Academic Editor: Daniela Lötsch

Received: 31 December 2021
Accepted: 14 February 2022
Published: 17 February 2022

1. Introduction

Gliomas are heterogeneous brain and spinal cord tumors [1]. The expected survival of patients with glioma is extremely poor. In recent years, it was one of the leading cancer-related causes of death among most sex and age groups in adolescents and young adults [2]. The World Health Organization (WHO) in 2007 used cell types to classify gliomas into subtypes (astrocytoma, oligodendroglioma, oligoastrocytoma or ependymoma) or grades from I to IV [3]. In 2016, the subtyping system was updated according to molecular parameters such as the presence of a mutation in the *IDH1* gene [4]. However, the subtyping system update did not influence the grading system that is based on the histological criteria derived from a biological behavior of neoplasm. Specifically, WHO discerns four glioma grades that are defined as follows: grade I (GI or G1) with low proliferative potential, grade II (GII or G2) with low-level proliferative activity, grade III (GIII or G3) histological evidence of malignancy and grade IV (GIV or G4) cytologically malignant that is the most malignant form of glioma [5]. Here, the grading system was adopted from The Cancer Genome Atlas (TCGA) which classifies gliomas into lower-grade gliomas (LGG) including GII and GIII [6] and glioblastoma multiforme (GBM) including GIV.

Biological mechanisms behind any tumor progression, including glioma, are robust and affect many crucial signaling pathways. Numerous studies have identified alterations in the genome and characterized core pathways that are dysregulated. The study by [7] concluded that NF-κB participates in glioma angiogenesis that increases its malignancy. Interestingly, NF-κB is a critical factor that regulates immune response and the development of inflammatory diseases and cancer [8]. Furthermore, it has been established that the following pathways are disrupted in GBMs: (1) growth factor downstream signaling via phosphatidylinositol 3-kinase (PI3K) pathway; (2) apoptosis regulation via p53 signaling; (3) cell cycle regulation via cyclin-dependent kinases and retinoblastoma 1 signaling (RB1) pathway [9,10].

One of the most extensive data resources of transcriptomics datasets for gliomas is TCGA [11]. TCGA hosts a broad collection of samples sequenced with an RNA-seq, as well as other omics techniques. Recent studies have reported that various decision-unrelated sources of bias, i.e., batch effects, could occur among cohorts obtained from different sequencing facilities [12,13]. Importantly, a batch effect may influence the downstream analyses, especially when confounded with the outcome of the analysis, such as TCGA-LGG and TCGA-GBM. Furthermore, the impact of batch effect correction on datasets may remove biologically-relevant information that drastically affects the statistical analysis and thus, it shall be applied with great care [14]. To enhance reproducibility and limit variation, researchers created projects aiming to precompute and unify public cohorts such as recount2 or University of California, Santa Cruz (UCSC) Xena [15,16]. However, if the batch effect is highly confounded with an outcome of interest, a novel methodology needs to be designed and employed for correction.

In recent years, machine learning (ML) has been applied successfully in many scientific areas, including life sciences [17–19]. This popular field has been shown to effectively support knowledge mining and patterns recognition in big biological data [20]. For example, the prediction of a cancer outcome using ML techniques led to the detection of biomarkers and the exploration of novel ways of treatment [21]. An accurate prediction of the disease condition is a substantial challenge. However, the interpretation of ML models is also extremely challenging and increasingly fascinating [22]. Furthermore, the effect of batch on ML has also been studied [23]. The study showed that the bias is carried through the ML process and, thus it can affect the results and conclusions.

In this study, we corrected the strong batch effect between LGG and GBM cohorts from TCGA. For the corrected data, we provided a comprehensive ML analysis for two models of glioma grading: (1) GII vs. GIII and (2) LGG vs. GBM. To decrease variation from the batch effect, we used unified TCGA cohorts from the UCSC Xena repository [16] that were recomputed under the UCSC Toil RNA-seq pipeline. The analysis was divided into two stages. First, we focused on analyzing differentially expressed genes (DEGs). Second, we

performed a single-sample gene set enrichment analysis (ssGSEA) that was followed by a comprehensive ML evaluation and analysis. We applied ssGSEA to detect the most accurate Molecular Signatures Database (MSigDB) [24–26] collections that discern between glioma grades. Next, topmost collections were used to determine dependencies among annotations that revealed co-enrichment for glioma grading. Finally, we validated our results and evaluated the survival of the glioma patients for the most co-enriched annotations.

2. Materials and Methods

This study focused on developing a methodology for ML-based analysis of glioma cohorts from TCGA (Figure 1). First, we aimed at correcting the strong batch effect that biased the ML analysis performance. After successfully correcting the datasets, we performed two separate analyses: (1) Differential gene expression analysis; (2) ML analysis on ssGSEA scores based on MSigDB collections. Finally, two cohorts from the Chinese Glioma Genome Atlas (CGGA) [27] were used for validating the results. Below, we included detailed descriptions of the applied methods.

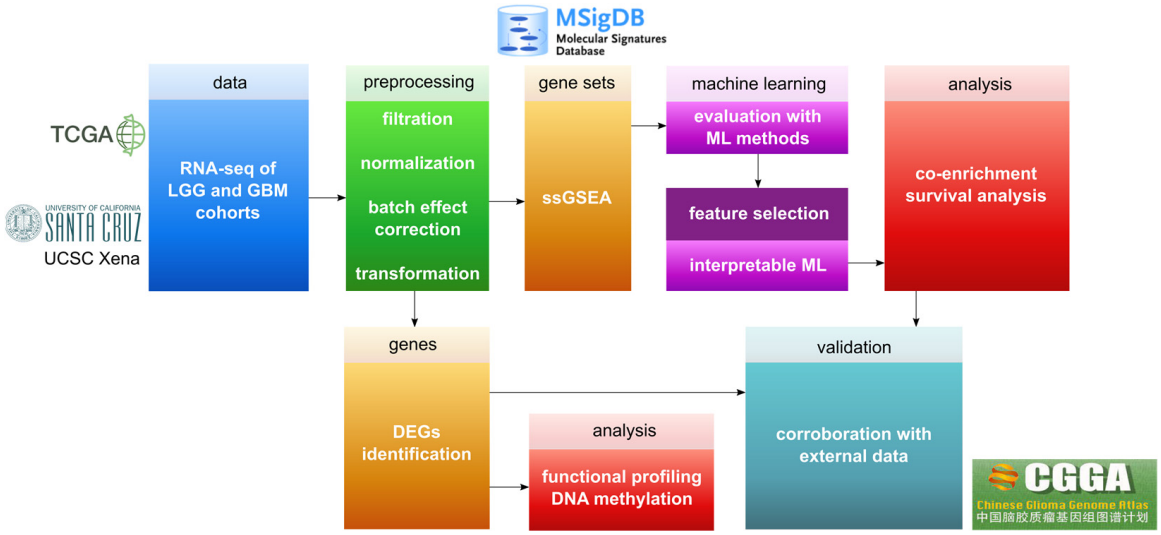

Figure 1. Overview of the pipeline applied in this work. Following preprocessing were two separate analyses: The lower tier of the pipeline illustrates the steps employed for identification and basic analysis of DEGs, while the upper tier demonstrates the steps for ssGSEA analysis based on ML approaches using MSigDB collections. The final step illustrates the validation of the results.

2.1. Preprocessing of Gene Expression Datasets

Initially, we analyzed raw transcriptomics data from the TCGA repository. To perform the analysis, we collected RNA-seq datasets from Genomic Data Commons (GDC; [28], including TCGA-LGG and TCGA-GBM cohorts. Using principal component analysis (PCA), we observed a strong variability between GDC-based cohorts (Figure S1A). Therefore, we used a unified and transcript per million-normalized dataset from UCSC Xena Toil RNA-seq recompute data hub, which contains merged cohorts TCGA, TARGET and GTEx [29]. We chose this cohort to allow future expansion of the analysis as it also includes TARGET and GTEx. However, even for the unified cohort, strong variability was still visible (Figure S1B). Therefore, we attempted to correct for unknown sources of batch effects (Figure S1C,D, Table S1) [30]. Using an ML evaluation and Student's *t*-test, we examined how bias influences classification between particular grades groups (Table S1). We found that classifiers result in very high quality for randomly chosen genes (Figure S2A–D) and an unusually high fraction of these genes being DEGs (Figure S2E).

In the next step, we examined principal components (PCs) of the unified cohort and we retrieved Gaussian mixtures (GMs) from the first PC (PC1) (Figure S3, Table S1). This allowed us to detect mixtures that corresponded to hidden groups of samples. Based on PC1,

two GMs could be distinguished (Figure S3A–C). These are GM1 that contained LGG and GBM samples, and GM2 that contained the vast majority of LGG samples. Therefore, we divided the unified TCGA dataset according to GMs (Figure S3D–I, Table S2). Specifically, the final GM1-based dataset contained 151 GBM and 108 LGG samples, and the GM2-based dataset collected 231 GII and 168 GIII samples (Table 1). We also evaluated which samples were separated based on GMs. Here, we provided global PCA and local t-SNE approaches (Figure S4). To show the potential source of the batch effect, we visualized tissue source sites (TSSs) for all samples (Figure S4C,D).

We first evaluated the GM1-based dataset and observed that classifiers built with randomly selected genes are closer to the accuracy of permutation tests (Figure S5). In addition, as expected the fraction of DEGs decreased (Figure S5F). Next, we ran surrogate batch effect analysis on GM1-based samples using two methods, namely "leek" and "be" (Figure S5D,E). Furthermore, after applying batch effect correction, we selected protein-coding genes (Figures S3G and S5E) using a reference file from HUGO Gene Nomenclature Committee [31]. As a result, datasets contained 19,028 protein-coding genes. To evaluate if the biological information was not affected by batch effect correction, we again performed the *t*-test on the GM1-based dataset that resulted in 255 significant genes with *p* value adjusted for false discovery rate (FDR) less than 0.001. Next, GM2-based data were processed in a similar fashion. We examined the variability between GII and GIII (Figure S6), where no batch effect was visible and the correction step was omitted. Finally, we performed a *t*-test between GII and GIII and found 439 significant genes (FDR-adjusted *p*-value < 0.001). For the list of significant genes, we run functional profiling with *gProfiler*, using all available databases, that revealed sets of significant (FDR-adjusted *p*-value < 0.05) pathways. In addition, we checked the influence of sex and age on GM1- and GM2-based datasets (Figure S7).

Table 1. A summary of sample amounts used in the analysis to obtain and validate results. In total, 1671 publicly available samples were used in this analysis.

TCGA				CGGA					
GM1		GM2		Batch 1			Batch 2		
GII	GIII	LGG	GBM	GII	GIII	GBM	GII	GIII	GBM
231	168	108	151	188	255	249	103	79	139

2.2. DNA Methylation Data

The DNA methylation data were used to examine epigenetic changes in DEGs for samples corresponding to GM1 and GM2. The DNA methylation profiling was based on the Illumina Infinium HumanMethylation450 platform for 685 samples corresponding to samples in transcriptomics analysis. The dataset (GBM-LGG) was downloaded from the UCSC Xena browser. CpG sites with no recorded beta values were filtered out prior to downstream processing. After filtering, 364,859 CpG sites remained. To check for possible batch effects, we visualized the dataset using PCA. We annotated CpG sites with their associated genes (Table S1). The average beta values across different groups were compared using a non-parametric Wilcoxon test.

2.3. ssGSEA Analysis

To further analyze GM1- and GM2-based datasets, we employed ssGSEA, a single-sample extension of GSEA. Using the ssGSEA approach, we transformed all the variables from gene expression values to the degree of enrichment. The single-sample approach decreases the variability of datasets suffering from confounding factors. To perform ssGSEA, we used a method proposed by [32] (Table S1). We first selected all 20 collections from MSigDB v7.4 and then ran ssGSEA on glioma cohorts. These collections included the following gene sets: hallmark, positional (PG), chemical and genetic perturbations

(CGP), BioCarta, Kyoto Encyclopedia of Genes and Genomes (KEGG), Pathway Interaction Database (PID), Reactome, WikiPathways (WP), microRNA targets (MIR), transcription factor targets (TFT), cancer gene neighborhoods (CGN), cancer modules (CM), gene ontology cellular component, biological process and molecular function (GOCC, GOBP and GOMF, respectively), human phenotype ontology (HPO), oncogenic signatures (Onco), ImmuneSigDB (Immuno), vaccine response gene sets (VAX) and cell type (CT).

Some ML methods, such as rule-based learning, require discrete variables to perform learning. In ssGSEA, enrichment scores describe the activity degree of a given gene set. Each score represents the enrichment degree to which the genes are simultaneously down- or up-regulated for a single sample. The ssGSEA scores were discretized with equal frequency, as it was done in rule-based modeling, into three levels of the degree: low, medium and high. For instance, a high ssGSEA degree means that there are many down- or up-regulated genes from the given gene set in a particular glioma sample. We believe that such simplification of the ssGSEA degree could lead to improved interpretability.

2.4. ML Evaluation

To evaluate the MSigDB collections, we performed an ML analysis before and after applying Monte Carlo feature selection (MCFS) [33] that is a decision tree-based non-linear method. According to the ML evaluation (Table S1), we selected the top three most accurate collections for discerning grades. In order to evaluate the classification abilities of collections, we selected five different and well-established ML approaches [34], namely: sequential minimal optimization (SMO) for training support vector classifiers [35], instance-based learning algorithms (IBk) that is an extension of *k*-nearest neighbors algorithm [36], bagging predictors [37], J48 that generates C4.5-based decision trees [38] and repeated incremental pruning to produce error reduction (JRip) that creates classifiers by rule learning algorithm [39]. We based our choice criterion on mixing several well-known black-box and interpretable ML methods. As the datasets included an uneven distribution of decision classes, we applied undersampling of the majority class to match the size of the minority class. For instance, the number of balanced classes in the case of GII vs. GIII was equal to the total number of GIII samples, i.e., 168 samples (Table 1). The undersampling was performed 20-times in order to obtain balanced datasets. Next, the ML modeling was performed with 10-fold cross-validation (CV). In addition, we performed a permutation test for each model. The permutation test has been performed by randomly shuffling the decision classes. The test was included within an undersampling loop and performed with 10-fold CV.

We employed two well-known classification quality measures in this work, namely accuracy (ACC) and area under the ROC curve (AUC). The ACC was used for the ML evaluation of datasets before applying ssGSEA analysis. After applying ssGSEA, we used the AUC measure for evaluation. We used undersampling in all experiments, we believe these metrics can be used interchangeably.

2.5. Interpretable ML

To find dependencies between annotations and provide interpretable classifiers, we generated rule-based models (RBMs) with *R.ROSETTA* [40]. The method uses a rough sets theory for producing a set of IF-THEN rules that constitute an RBM [41]. The set of rules was initially created using a Boolean reasoning approach. However, since a Boolean reasoning approach is a non-deterministic polynomial hard problem, several algorithms called reducers have been developed to tackle this dilemma. Here, we used the Johnson reducer method that produces a high fraction of significant rules and does not overestimate their total amount [40]. Importantly, we have created RBMs only for the topmost MSigDB collections selected based on the AUC value. The rules were further filtered according to their *p*-value (FDR-adjusted *p*-value < 0.01). Notably, such rules are directly interpretable and reflect co-predictive mechanisms among features. As in the case of well-known co-expression analysis, such dependencies may reflect biological interactions. However, rules

characterize non-linear, local and supervised dependencies of features. As in the case of previous ML evaluations, we applied undersampling and 10-fold CV for obtaining RBMs. Importantly, equal frequency discretization of ssGSEA scores was performed within the CV loop.

2.6. Rule-Based Networks of Co-Enrichment

Here, we presented co-predictive mechanisms as a co-enrichment that is defined as two or more annotations being simultaneously enriched for a specific group of samples regarding the decision class, i.e., glioma grade. Usually, annotations are treated independently, but we assumed that annotations might have complementary functions [42]. Such annotation-annotation dependencies have been successfully investigated for evaluating drug effects [43]. In general, co-enrichment has been shown as an interesting concept for analyzing data in the form of a network [44]. Thus, we are aware of its high importance in analyzing complex diseases such as glioma.

RBMs were further visualized using a rule-based network approach with *VisuNet* [45]. This approach transforms a set of rules into a network. Here, the network represents annotations and their values as nodes and rule-derived connections as edges. We used the decision coverage value to define the size of nodes. Furthermore, the feature enrichment score degree corresponds to the color of nodes and the adjusted connection strength between two nodes from a rule defines the color and width of edges. We adjusted connection values on the network to normalize the co-enrichment that may occur due to the overlapping gene sets. To obtain normalized connection values on networks and total correlation values [46](Table S1) on heatmaps, we used the following formula:

$$v_{norm} = v * (1 - \alpha), \quad (1)$$

where v is the connection value between two nodes or total correlation value and α is the degree of overlapping genes between two gene sets. For instance, if there are no overlapping genes between gene sets, then $\alpha = 0$ and $v_{norm} = v$. All networks presented in the paper were created for the 20 most connected nodes of the top 10% rules ranked by the connection value of rules. Finally, as another level of visualization, we used a concept of arc diagrams for displaying particular nodes, i.e., nodes of interest (NOI), from networks.

3. Results

3.1. Data Correction

As a result of comprehensive data preprocessing, we detected two subsets of samples within TCGA glioma cohorts. We assumed that these subsets correspond to the hidden, i.e., unknown, batch effect. However, we suspect that this batch effect is related to TSS (Figure S4C,D) as it is clearly visible that the majority of GBM-related TSSs are visualized as a separate cluster. In other words, source sites (hospitals, universities, etc.) are highly confounded with the decision class LGG vs. GBM. Thus, we further used the surrogate variable analysis on the GM-based subset that assisted in removing the batch effect from LGG vs. GBM data (Table S1). Therefore, we enclosed a table (Table S2) that may help in future studies of TCGA glioma datasets for more accurate analysis, which includes TCGA sample IDs, GM groups and grade information. We believe that GM modeling, together with PCA, can be applied in similar situations to correct highly confounded batch effects.

3.2. DEGs Evaluation

First, we identified lists of highly significant DEGs (FDR-adjusted p-value < 0.001) for GII vs. GIII (Table S3) and LGG vs. GBM (Table S4) that we used to perform functional profiling (Tables S5 and S6). Based on the results, we examined the most significant and interesting functional annotations for discerning glioma grades. For GII vs. GIII (Table S5), we observed that the cell cycle, p53, DNA replication and Fanconi anemia were among the most significant pathways of KEGG. In addition, from GOBP, we noticed that the cell cycle is highly enriched. Furthermore, GOCC suggested that the list of DEGs is highly

enriched for chromosome-related annotations. Interestingly, Reactome and WP also pointed towards cell cycle-related annotations. In addition, several cancer-related pathways from WP were enriched.

For LGG vs. GBM (Table S6), we observed that nonsense-mediated decay (NMD) was highly significant in the Reactome database. This finding corroborates a recent discovery that showed modulation of NMD promoting the growth of GBM in humans [47]. Based on the Reactome, the metabolism of RNA is significantly important for GBM grading processes. In addition, rRNA and mRNA processing signaling pathways were significantly enriched. All significant GOMF annotations pointed towards binding processes, e.g., RNA or nucleic acids binding. In both grading-related cases, no brain-related tissues were detected among human protein atlas annotations, highlighting no tissue-specific DEGs. In the next step, using CGGA data, we validated sets of DEGs. For validation, we used preprocessed and normalized RNA-seq CGGA datasets. While using CGGA, we assumed that LGGs are samples marked as GII or GIII, as well as it was done by TCGA. Here, we examined how many TCGA-based DEGs were in two CGGA cohorts (batch 1 and 2) (Table 2). We observed a good overlap of DEGs between these independent two cohorts.

Table 2. The results of DEG list validation with CGGA cohorts. *p* values in validation cohorts were FDR-adjusted.

CGGA Batch 1				CGGA Batch 2			
GII vs. GIII		**LGG vs. GBM**		**GII vs. GIII**		**LGG vs. GBM**	
$p < 0.001$	$p < 0.05$	$p < 0.001$	$p < 0.05$	$p < 0.001$	$p < 0.05$	$p < 0.001$	$p < 0.05$
62%	88%	27%	44%	85%	96%	52%	71%

Finally, we intersected both gene lists, i.e., GII vs. GIII and LGG vs. GBM, and found 6 DEGs in common: *IGIP*, *NSMCE2*, *CNIH4*, *NONO*, *CKLF* and *RAN* (Figure S8). We then validated the expression of these genes using CGGA batch 1 and batch 2. Both batches confirmed that these 6 DEGs were differentially expressed in the validation sets for discerning GII vs. GIII and LGG vs. GBM (Figures S9 and S10). In addition, we examined DNA methylation data corresponding to GM1- and GM2-based samples (Figure S11). We further checked DNA methylation profiles of the 6 shared DEGs (Figures S12 and S13). We found several differentially methylated regions (DMRs). Interestingly, there were more DMRs in 6 common DEGs for LGG vs. GBM, which may suggest that more robust epigenetic changes are visible while progressing to a higher grade.

3.3. ML for ssGSEA

We evaluated MSigDB collections using an ML approach (Figure 2). We applied learning to each collection separately to choose the best collections discerning glioma grades. To further assess the MSigDB collections, we performed feature selection on each collection and selected a balanced number of important features (Figure S14, Table S1). Afterward, we selected the three top collections for GII vs. GII CGP, BioCarta and PID, and for LGG vs. GBM GOCC, GOBP and WP.

Interestingly, all three selected collections for GII vs. GIII were curated gene sets, while two out of three collections for LGG vs. GBM were ontology gene sets. This may suggest that more biological processes on the cellular structure level are disrupted for progression to a higher glioma grade. Furthermore, we observed that cancer-related collections were also highly predictive in both cases (Figures 2 and 3). As expected, differences between LGG and GBM are more prominent than between GII and GIII. In addition, feature selection improved the quality of all models. Thanks to feature selection, we received fewer features that in turn enhanced the interpretability of the models. In addition, MCFS application was necessary to balance the number of features across compared collections (Figure S14). From MCFS, we presented the most important annotations discerning between grades.

For instance, "Fanconi", "cell cycle" and "Spermatocyte" [48] were annotations with the highest relative importance (RI) values discerning GII from GIII.

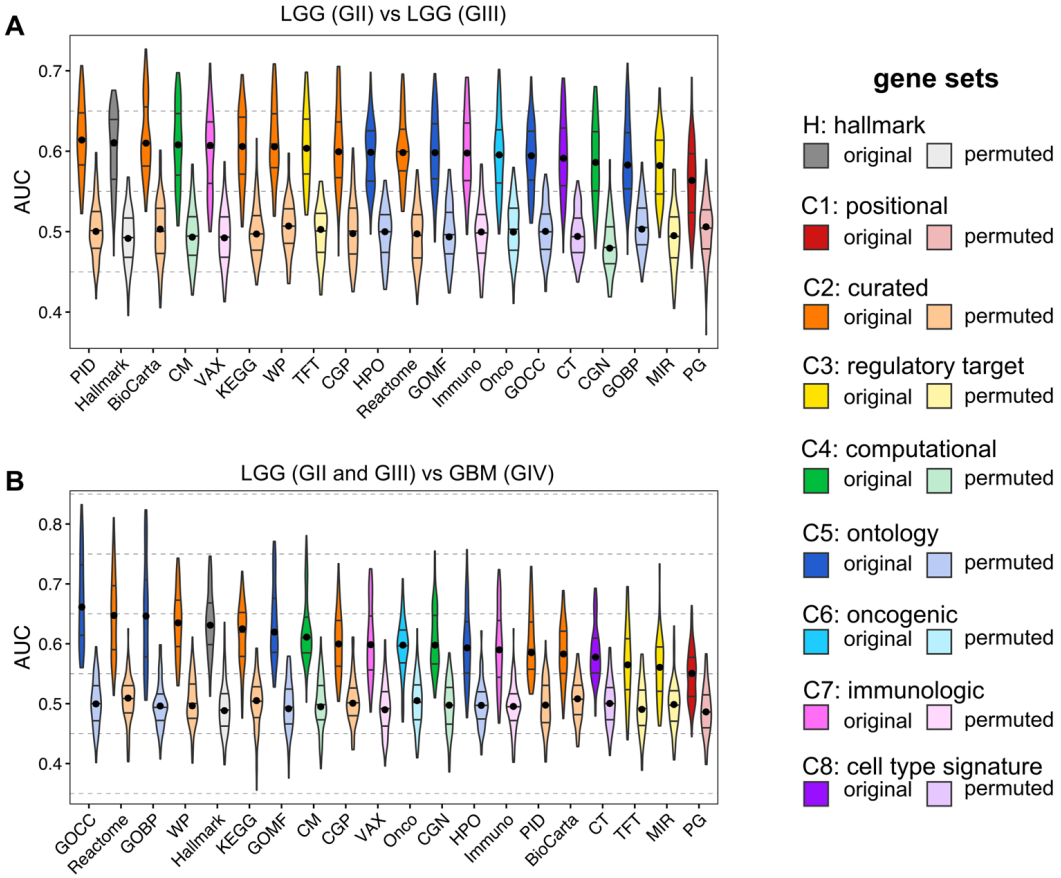

Figure 2. Evaluation of MSigDB collections using ML models for discerning glioma grades using ssGSEA scores for classifying (**A**) GII vs. GIII and (**B**) LGG vs. GBM. Five different ML approaches were used: SMO, IBk, Bagging, J48 and JRip. Each ML method was undersampled 20-times with 10-fold CV. The median was marked with a black dot on each violin.

To create rule-based models, we used the R.ROSETTA method that resulted in highly accurate classifiers (Figure S15). The best collection for discerning GII from GIII was CGP, which resulted in 0.71 AUC, and the best collection for discerning LGG from GBM was GOCC, which resulted in 0.84 AUC. For each of the two comparisons, we created a joint rule-based model by merging the features from the top three collections. However, AUCs for joint models were not better than the best AUC of single models (Figure S15). Finally, we validated rule-based models by classifying CGGA batch 1 and batch 2 gene expression datasets. For the validation, we used TCGA-derived annotations that classified grades with similar performance (Figures S16 and S17). In addition, we estimated total correlation values among all annotation pairs for the topmost collections. We calculated this correlation for ssGSEA scores discretized into three levels with equal frequency binning.

3.4. Glioma Co-Enrichment

We visualized RBMs as networks to evaluate co-enrichment mechanisms among annotations for topmost MSigDB collections (Figure 4, Figures S18 and S19). As networks, displayed for the entire set of rules, were unbalanced with respect to decision classes, we generated balanced networks for each decision class separately (Figure 4, Figures S18 and S19). We discovered co-enrichment mechanisms among the most significant rules (p-value < 0.01). We included sets of significant rules in Tables S7 and S8, for the networks Figure 4A,B, respectively. Notably, these mechanisms are local and correspond to a group of patients, i.e.,

the rule is supported by a set of samples that fulfill its conditions. Connections values on all networks were normalized according to the number of genes shared between gene sets.

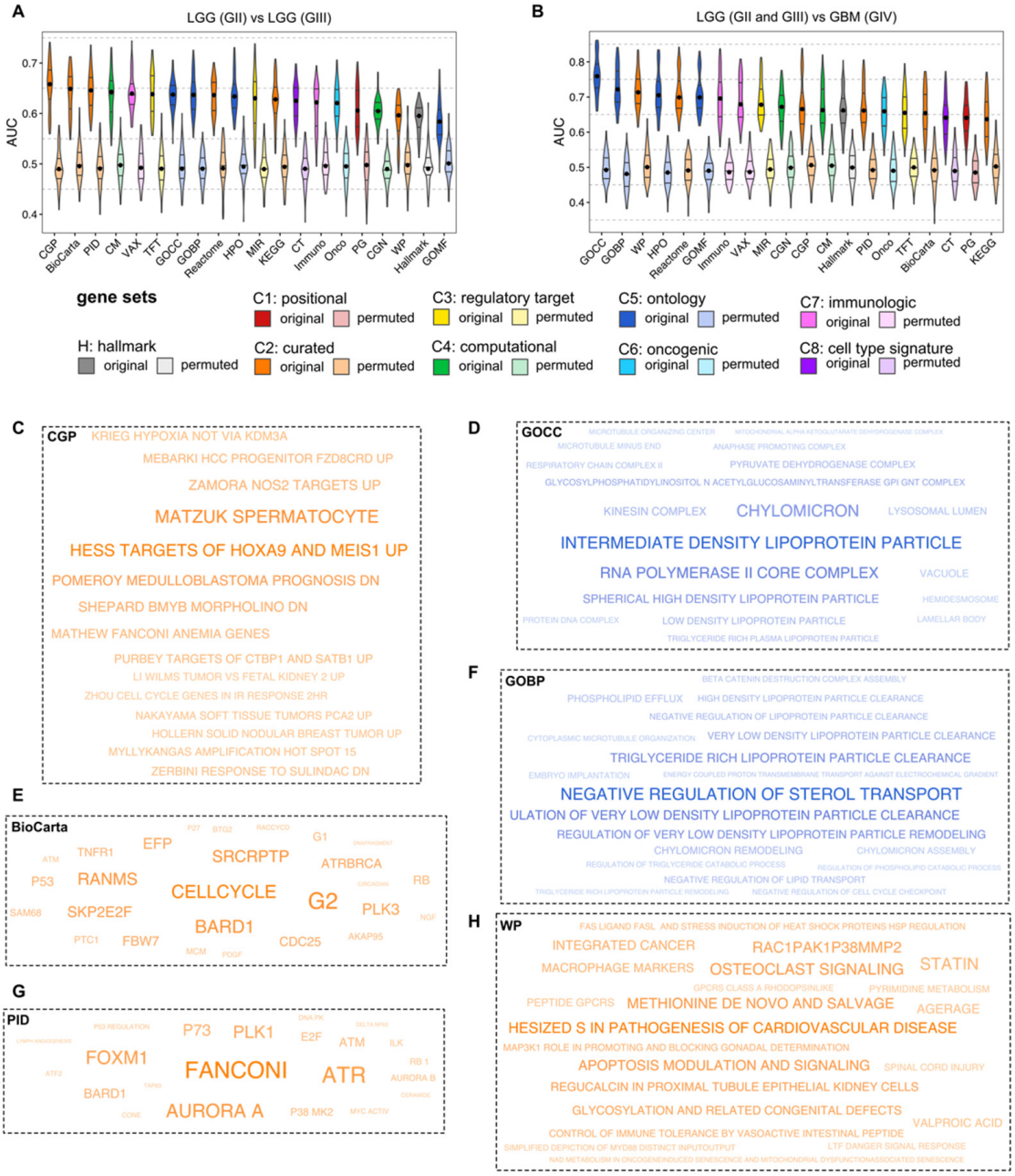

Figure 3. (**A**,**B**) Evaluation of MSigDB collections using ML models with feature selection for discerning glioma grades using ssGSEA scores. Five different ML approaches were used: SMO, IBk, Bagging, J48 and JRip. Each ML method was undersampled 20-times with 10-fold CV. The median was marked with a black dot on each violin. Panels (**C**,**E**,**G**) represent MCFS results for the top three collections for GII vs. GIII. Size of annotations represents RI values from MCFS. Panels (**D**,**F**,**H**) represent MCFS results for the top three collections for GII vs. GIII. Size of annotations represents RI values from MCFS.

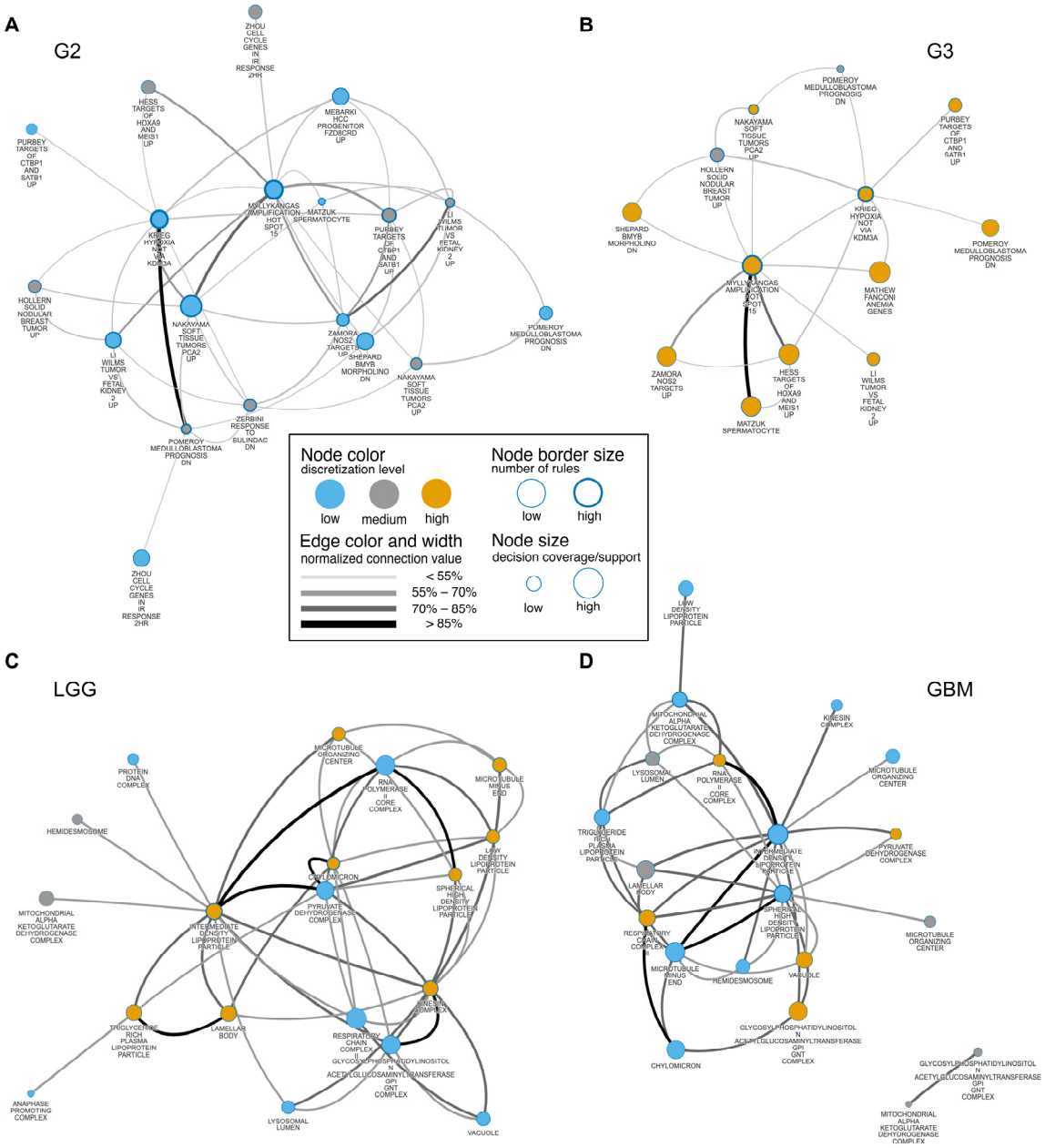

Figure 4. Rule-based network displaying the most relevant co-enrichments of annotations obtained from (**A**,**B**) the CGP collection for the GII vs. GIII model (Table S7) and from (**C**,**D**) the GOCC collection for the LGG vs. GBM model (Table S8). The networks show the 20 most connected nodes obtained from the top 10% of significant rules (FDR-adjusted p-value < 0.01) based on the rule connection. Connection values of nodes and edges represent a strength of co-enrichment from the classifier. Subnetworks were generated separately with respect to the decision class for each RBM.

By analyzing networks and investigating NOIs, several findings can be described. Here, the main hub in the GII and GIII networks (Figure 4A,B) was "amplification hot spot 15" [49]. In both subnetworks (Figure 4A,B), this node is connected to several annotations, among others "Wilms tumor vs. fetal kidney 2 up" [50] and "Soft tissue tumors PCA2 up" [51]. The latter indicated that these pathways may be linked to activating the human set of specific oncogenes during progression from a lower to a higher grade. We could also observe more generic annotations, for instance, "cell cycle" or "G2 phase" (Figures 3 and S18A,B). Here, the "G2 phase"-related gene set is more relevant than "G1 phase" for GIII (Figure 3E). However, several authors have reported changes in cell cycle

regulation under various circumstances [9,52–54]. Interestingly, a high degree of enrichment for the "Aurora A" pathway played a major role in grading (Figures 3G and S19A,B). The "Mitotic Aurora A kinase" (AurA) pathway is an essential factor in the survival, radioresistance, self-renewal and proliferation of glioblastoma cells [55]. Its potential therapeutic abilities have also been investigated [56]. Here, we found that the enrichment of the AurA pathway differed between GII and GIII. More specifically, we found that "Aurora A" was a highly co-enriched pathway for GII vs. GIII (Figure S19A,B). Interestingly, "Aurora B" appeared to be highly interactive in validation cohorts (Figure S22A,C). Importantly, both Aurora pathways were investigated for treating cancer [57]. Furthermore, we found that the Fanconi anemia pathway played a crucial role in glioma grading (Figures 3G, S19B and S22A,C). Recently, its potential therapeutic role has been described [58]. From other topmost MSigDB collections discerning between GII vs. GIII, we found evidence of a local co-enrichment between "P53" and "PDGF" pathways, i.e., strong connection in a network (Figure S18B). It is well-known that alterations in the *P53* gene promote tumor development, malignancy and resistance to radio and drug therapy [59,60]. Furthermore, the *PDGF* gene is one of several growth factors participating in glioma angiogenesis [61]. Here, we provided a hypothesis that these two pathways may be co-dependent.

Next, we investigated results obtained for grading to GBM (Figures 4C,D, S18C,D and S19C,D) that were further validated (Figures S20–S22). Intermediate density lipoprotein (IDL) was NOI in LGG and GBM networks (Figure 4C,D). This annotation was also highly co-correlated in both validations sets (Figure S20B,D). A high connection between IDL and microtubule-related annotations was visible in all cases. The role of microtubules in the degradation of lipoproteins has been previously identified [62] and this result might provide additional evidence of its potential role as a therapeutic target for GBM treatment. Chylomicron and IDL particle detected from GOCC collection (Figures 3D, 4C,D and S20A) and sterol transport-related annotations from GOBP (Figures 3F, S18C,D and S21D) may suggest a link to cholesterol-related mechanisms and glioma grading. Recent studies have elaborated that cholesterol metabolism may be a potential therapeutic target in glioma [63,64]. In this study, we found that the degree of enrichment of cholesterol-related pathways strongly differs between LGG and GBM. Interestingly, we found that cholesterol-related pathways were highly co-enriched and may affect microtubule organization in the case of GBM (Figure S19C,D). Thus, annotations related to cholesterol could be further investigated for their therapeutic potential in GBM patients. Moreover, we found a high activation degree of "methionine de novo and salvage" pathway (Figures 3H and S19C,D). The survival and proliferation of cancer cells were shown to be dependent on methionine levels [65]. Finally, we found that microtubule organization differed between LGG and GBM (Figures 3D and 4C,D). The inhibition of microtubule dynamics was explored previously for its potential in GBM treatment [66].

3.5. Survival Analysis

We performed a survival analysis [67,68] for selected NOIs of the RBM networks (Figures 4, S18 and S19, Table S1). Here, we obtained overall survival and its status from cBioPortal [69,70]. The survival analysis was done to determine the degree of a given annotation enrichment corresponding to discrete levels obtained with equal frequency discretization. The high degree of enrichment for both NOIs from CGP-based networks, i.e., "Amplification hot spot 15" and "Hypoxia not via KDM3A", contributed to poor survival for LGGs (Figure 5A,B). We observed that a low degree of enrichment, i.e., low activity, for "IDL particle", "negative regulation of sterol transport" and "Valproic acid" corresponded to poor survival in samples with higher glioma grade. Interestingly, "Valproic acid" (Figure 5C,G,K) has been recently proposed as a promising therapeutic target for gliomas [71]. We also noticed that microtubule-related annotations did not influence the overall survival (Figure 5D,H). For LGGs, we observed that a high degree of enrichment with "G2 phase" (Figure 5F) played a role in overall survival, while the impact of the "G1

phase" was insignificant (Figure 5E). Furthermore, a high degree of AurA and P38/MK2 enrichment was significantly associated with poor survival in LGGs (Figure 5I,J).

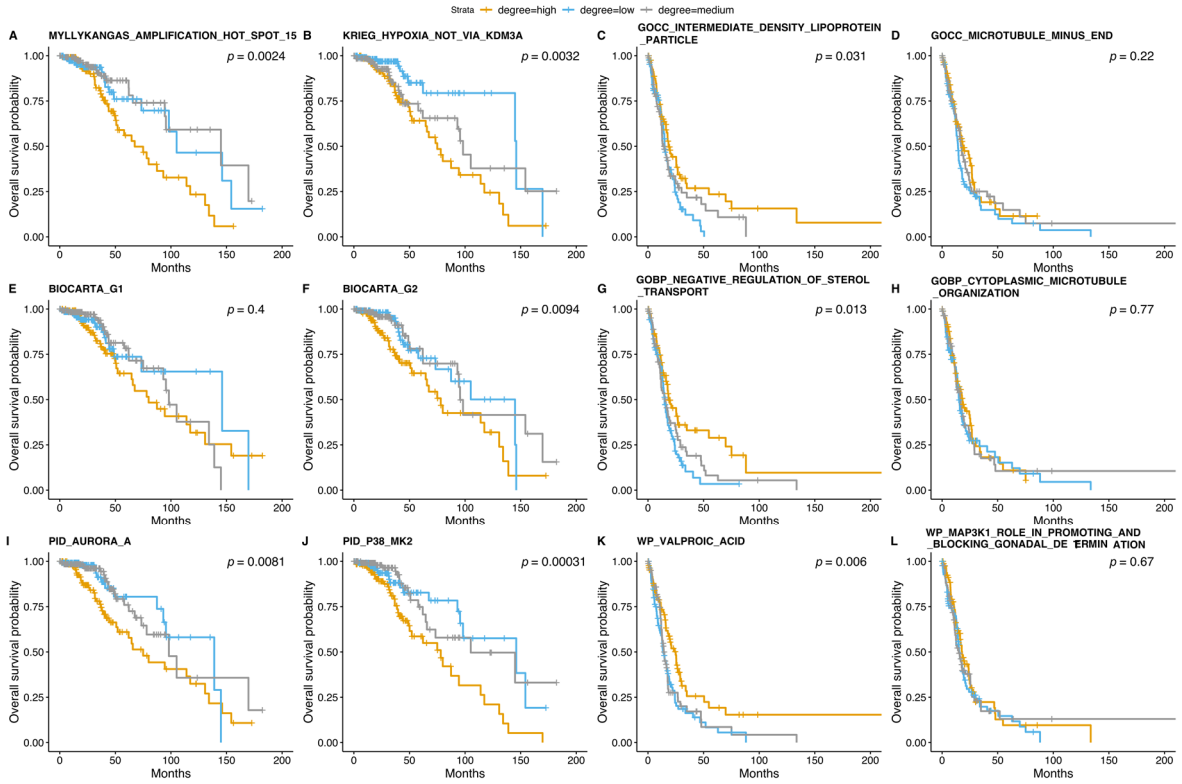

Figure 5. Survival curves of several NOIs characterized based on rule networks. We investigated NOIs for the topmost predictive three MSigDB collections discerning glioma grades: (**A**–**D**) CGP and GOCC, (**E**–**H**) BioCarta and GOBP, and (**I**–**L**) PID and WP. Each plot displays a *p*-value that was estimated with the default set of parameters while constructing the curves (Table S1).

4. Discussion

This study provides hypotheses of co-enrichment between glioma grade-related annotations regarding their high predictivity. There are several advantages of discovering co-enrichment between annotations. From the study by [72], we know that interaction may occur between two perturbed pathways that can lead either to increased or decreased disease risk. Several studies [73–75] have shown that discovering an interaction among pathways may improve therapy for treating cancer. Thus, we believe that the findings of this analysis may provide insights for future research and aid in novel ways of clinical treatment.

It is imperative to provide non-biased analyses in bioinformatics. Thus, we aimed at comprehensive preprocessing to perform unbiased ML analysis and retrieve biologically meaningful results in this work. As we aimed at performing ML analysis, we focused on the global structure of data, i.e., PCA, in order to investigate a low number of clusters enriched with a high number of samples. We showed that the local structure of the data, i.e., t-SNE analysis, generated very similar clusters (Figure S4). Thus, we concluded that in this study, PCA and t-SNE approaches were comparable. Moreover, we provided a thorough benchmarking of several ML methods that revealed specific MSigDB collections corresponding to glioma that may guide future research. However, the predictivity of specific collections can be disease-specific, so it shall be estimated separately while analyzing other types of cancers or other diseases.

In this work, we transformed gene expression into annotations, as we believe that pathways are more universal than genes. As biology is robust and diverse, it is more reasonable to perform analysis based on pathways. For instance, assuming that expression alterations of Gene A and Gene B lead to a change in Pathway X. In such case, Patient 1 that

has a change in expression of Gene A and Patient 2 that has a change in expression of Gene B would be merged into a common group of patients that have a change in Pathway X. Thus, the variability is decreased and a common unit, i.e., a pathway, is established. There are several advantages [76] of applying pathway enrichment methods, such as aggregating information, reducing data dimensionality, enhancing the interpretation of results, identifying drug targets, providing better comparability within the same omics technology or between various omics technologies. Despite the wide range of advantages, there are also limitations [76,77]. Among others, effectiveness is determined by the strength of signals from multiple genes, databases are biased toward well-studied genes and pathways, and interactions among genes are neglected, i.e., gene independence is assumed. Moreover, here we used ssGSEA and thus, we are limiting the data space to a single-sample approach. The reason for that was to follow the idea proposed by [78] to provide an additional layer that prevented the strong batch effect that occurred for this particular dataset. Furthermore, we used data on a discrete scale as it is necessary for our rule-based approach [79]. Thanks to this, we improved the interpretability of results and reduced the influence of noise in the data. However, the discretization process reduced the information and neglected the continuous nature of the data. Lastly, we performed the basic survival analysis for relevant annotations. To explore various survival tasks in a more comprehensive way, we encourage using a more advanced and recent approach such as DeepPAM [80,81].

We observed that most of the co-enrichment mechanisms detected via networks (Figures 4, S18 and S19) for LGG vs. GBM were also observed in validation analyses of the CGGA batches (Figures S20–S22). In contrast, fewer co-enrichment mechanisms describing GII vs. GIII networks could be validated with the CGGA batches. Thus, it is highly possible that in the case of GII vs. GIII, ML models with low accuracy could have affected the analysis and disrupted the validation. In general, we could observe that the quality for discerning GII and GIII is not very high. On the other hand, the quality of interpretable rule-based models is above median compared to other ML techniques (cf. Figures 3A,B and S15). Thus, we provided evidence that interpretable learning is not only legible but also produces high-quality models [82]. Importantly, this work demonstrates the results of non-linear dependencies of features. Thus, it may be also interesting to investigate linear dependencies of features in the future. For instance, by using other approaches such as importance-based sequential procedure [83].

We provided a simple normalization method to avoid false positives due to overlapping gene sets for co-enrichment. While normalizing networks or validating heatmaps, using Equation (1), we observed that overlaps of genes between gene sets are minor. This may be due to comparing gene sets within separate MSigDB collections. Thus, the normalization adjusted the final results slightly. Here, validation represented global co-enrichment with respect to decision classes, while networks showed local co-enrichment. Thus, the validation of co-enrichment mechanisms provided a general overview of our findings.

Taken together, this study provided a methodology that not only demonstrates how to perform batch effect removal towards ML analysis but also reveals potential interactions among pathways using an interpretable ML approach. We supported the analysis with statistical measures and tests. We believe that our findings can serve as potential therapeutic targets that could improve glioma treatment on various grade levels. The major future perspective is that these hypotheses can be validated experimentally to ensure our findings and incorporate them into glioma treatment.

5. Conclusions

A key challenge in bioinformatics is to perform the analysis in an unbiased, repetitive and accurate way. This study demonstrated how to remove a strong batch effect from TCGA glioma datasets and perform comprehensive ML analysis. Herein, LGG and GBM cohorts included a strong batch effect confounded with outcome classes. In such cases, it is essential to correct the batch effect, but it has to be done carefully in order to keep the biological information included in the data. Furthermore, this work describes co-

enrichment mechanisms that reflect robust processes for glioma progression. Notably, the proposed methodology is generic and can be used on any problematic data. To the best of our knowledge, this is the first co-enrichment analysis of glioma grades using rule-based learning.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/cancers14041014/s1, Figure S1: Principal component analysis (PCA) to investigate the variance for The Cancer Genome Atlas (TCGA) cohorts for Lower Grade Glioma (LGG) and Glioblastoma Multiforme (GBM). (A) Original data from the GDC repository. (B) Unified cohorts from the UCSC Xena repository. (C) Batch effect correction of unified cohorts for the leek parameter with 532 surrogate variables. (D) Batch effect correction of unified cohorts for the be parameter with 48 surrogate variables; Figure S2: Machine learning analysis was performed on 100 randomly selected genes with 10 various seed values to investigate the bias for TCGA cohorts of LGGs and GBMs. All five machine learning methods were applied, viz. SMO, IBk, Bagging, J48 and JRip (see section ML evaluation). (A) Original data from the GDC repository. (B) Unified cohorts from the UCSC Xena repository. (C) Batch effect correction of unified cohorts for the leek parameter with 532 surrogate variables. (D) Batch effect correction of unified cohorts for the be parameter with 48 surrogate variables. (E) Student's t-test results for particular cases a–d corresponding to subplots (A–D); Figure S3: Data evaluation to localize and remove bias from the glioma cohorts (A) PCA for unified cohorts from the UCSC Xena repository. (B) Clustering of 1–10 PCs of unified cohorts from the UCSC Xena repository. (C) Gaussian mixtures (GM) were detected from PC1. In GM2-based sample sets, two GBMs were included in the mixture and excluded in further analyses. (D) PCA was performed on LGG and GBM samples selected based on GM1. (E) Batch effect correction performed on LGG and GBM samples selected based on GM1 for the leek parameter with 232 surrogate variables. (F) Batch effect correction for the be parameter with 30 surrogate variables performed on samples selected based on GM1. (G) The final data set was based on GM1 after batch effect correction and filtration for protein-coding genes (H–I) Evaluation of LGG samples included in GM1 based data set; Figure S4: The global and local analysis of the data structure for TCGA glioma cohorts. (A) PCA and (B) t-SNE for decision class selected based on GMs. (C) PCA and (D) t-SNE for tissue source sites (TSSs). t-SNE plots were created for seed 1. Remaining TSSs included groups of samples equal to or less than 10; Figure S5: Machine learning analysis was performed on 100 randomly selected genes with 10 various seed values to investigate the bias for TCGA cohorts of LGGs and GBMs. All five machine learning methods were applied, viz. SMO, IBk, Bagging, J48 and JRip (see section ML evaluation). (A) unified cohorts from the UCSC Xena repository. (B) GM2 samples from unified cohorts from the UCSC Xena repository. (C) Batch effect correction of 259 GM2-based samples for the leek parameter with 232 surrogate variables. (D) Batch effect correction of 259 GM2-based samples for the be parameter with 30 surrogate variables. (E) The final data set is based on GM1 after batch effect correction and filtration for protein-coding genes. (F) Student's t-test results for particular cases a-e corresponding to subplots (A–E); Figure S6: The evaluation of samples in GM2 consisting of the vast majority of LGGs. (A) PCA for LGGs based on GM2 samples. (B) PCA for LGGs based on GM2 samples for protein-coding genes only. (C) Machine learning evaluation of GM2 samples for discerning between GII and GIII. (D) Machine learning evaluation of GM2 samples for discerning GII and GIII for protein-coding genes only. (E) Evaluation of fraction of statistically significant genes discerning between GII and GIII with all genes (case a) and protein-coding genes only (case b). (F) The proportion of grades within a GM2-based data set for LGGs. For machine learning evaluation (C,D) all five methods were applied, viz. SMO, IBk, Bagging, J48 and JRip (see section ML evaluation); Figure S7: The evaluation of other clinical factors within GM1- and GM2-based data sets. (A) PCA for GM1 for LGGs and their sex information. (B) PCA for GM2 for LGGs and GBMs and their sex information. (C) The variation of the age of LGG samples is based on GM1. (D) The variation of the age of LGG and GBM samples is based on GM2. Pearson correlation (r) value is marked in the plot caption; Figure S8: Expression profiles for six common differentially expressed genes (DEGs) were selected based on the intersection of the DEGs list between GII vs. GIII and LGG vs. GBM. Gene expression profiles were generated based on TCGA cohorts. *P* values were marked on boxplots as ns ($p > 0.05$), * ($p \leq 0.05$), ** ($p \leq 0.01$), and *** ($p \leq 0.001$); Figure S9: Expression profiles for six common DEGs were selected based on the intersection of DEGs list between GII vs. GIII and LGG vs. GBM. Gene expression profiles were generated based on the Chinese Glioma Genome Atlas (CGGA) batch 1 cohort. *P* values were marked on boxplots as ns ($p > 0.05$), * ($p \leq 0.05$), ** ($p \leq 0.01$), and

*** ($p \leq 0.001$); Figure S10: Expression profiles for six common DEGs were selected based on the intersection of DEGs list between GII vs. GIII and LGG vs. GBM. Gene expression profiles were generated based on the CGGA batch 2 cohort. *P* values were marked on boxplots as ns ($p > 0.05$), * ($p \leq 0.05$), ** ($p \leq 0.01$), and *** ($p \leq 0.001$); Figure S11: Variability in DNA methylation data from TCGA based on samples selected upon GM modeling for (A) GM2 and (B) GM1. Explained variation is given in parenthesis; Figure S12: DNA methylation status of six common DEGs for GII vs. GIII. *P* values were marked on boxplots as ns ($p > 0.05$), * ($p \leq 0.05$), ** ($p \leq 0.01$), and *** ($p \leq 0.001$); Figure S13: DNA methylation status of six common DEGs for LGG vs. GBM. *p* values were marked on boxplots as ns ($p > 0.05$), * ($p \leq 0.05$), ** ($p \leq 0.01$), and *** ($p \leq 0.001$); Figure S14: The Monte Carlo feature selection (MCFS) evaluation of thresholds was performed for choosing a threshold for selecting top features. (A) All 20 MSigDB collections for GII vs. GIII. (B) All 20 MSigDB collections for LGG vs. GBM; Figure S15: Evaluation of MSigDB collections with R.ROSETTA rule-based learning for classifying glioma grades using ssGSEA scores based on TCGA cohorts. (A) Top three MSigDB collections for classifying GII vs. GIII. (B) Top three MSigDB collections for classifying LGG vs. GBM. (C) Merged top three MSigDB collections for two RBMs classifying GII vs. GIII (right bar), and LGG vs. GBM (left bar); Figure S16: Evaluation of MSigDB collections with R.ROSETTA rule-based learning for classifying glioma grades using ssGSEA scores based on CGGA batch 1. (A) Top three MSigDB collections for classifying GII vs. GIII. (B) Top three MSigDB collections for classifying LGG vs. GBM. (C) Merged top three MSigDB collections for two RBMs classifying GII vs. GIII (right bar), and LGG vs. GBM (left bar); Figure S17: Evaluation of MSigDB collections with R.ROSETTA rule-based learning for classifying glioma grades using ssGSEA scores based on CGGA batch 2. (A) Top three MSigDB collections for classifying GII vs. GIII. (B) Top three MSigDB collections for classifying LGG vs. GBM. (C) Merged top three MSigDB collections for two RBMs classifying GII vs. GIII (right bar), and LGG vs. GBM (left bar); Figure S18: Rule-based network displaying the most relevant co-enrichments of annotations obtained from (A,B) the BioCarta collection for the GII vs. GIII model and (C,D) the GOBP collection for the LGG vs. GBM model. The networks show 20 most connected nodes obtained from the top 10% based on the rule connection from a set of significant rules (FDR-adjusted p value < 0.01). Connection values of nodes and edges represent a strength of co-enrichment from the classifier. Subnetworks were generated separately with respect to the decision class for each RBM; Figure S19: Rule-based network displaying the most relevant co-enrichments of annotations obtained from (A,B) PID for the GII vs. GIII model and (C,D) WP collections for the LGG vs. GBM model. The networks show 20 most connected nodes obtained from the top 10% based on the rule connection from a set of significant rules (FDR-adjusted p value < 0.01). Connection values of nodes and edges represent a strength of co-enrichment from the classifier. Subnetworks were generated separately with respect to the decision class for each RBM; Figure S20: Validation of top MSigDB collections: CGP for GII vs. GIII and GOCC for LGG vs. GBM. Heatmaps were generated based on cohort (A,B) CGGA batch 1 and (C,D) CGGA batch 2. Values of total correlation among two variables and a decision class given in nats; Figure S21: Validation of top MSigDB collections: BioCarta pathways for GII vs. GIII and GOBP for LGG vs. GBM. Heatmaps were generated based on cohort (A,B) CGGA batch 1 and (C,D) CGGA batch 2. Values of total correlation among two variables and a decision class given in nats; Figure S22: Validation of top MSigDB collections: PID for GII vs. GIII and WP for LGG vs. GBM. Heatmaps were generated based on cohort (A,B) CGGA batch 1 and (C,D) CGGA batch 2. Values of total correlation among two variables and a decision class given in nats; Table S1: Computational methods and R packages that were used in this study; Table S2: GM1 and GM2 subsets retrieved from GM modelling of PC1; Table S3: A list of significant DEGs for G2 vs. G3 (FDR < 0.001); Table S4: A list of significant DEGs for LGG vs. GBM (FDR < 0.001); Table S5: Significant enrichment results (FDR < 0.05) for G2 vs. G3 DEGs from gProfiler; Table S6: Significant enrichment results (FDR < 0.05) for LGG vs. GBM DEGs from gProfiler; Table S7: A list of significant rules (FDR < 0.01) for G2 vs. G3 RBM built for the CGP collection; Table S8: A list of significant rules (FDR < 0.01) for LGG vs. GBM RBM built for the GOCC collection; and Supplementary References for Table S1.

Author Contributions: Conceptualization and methodology, M.G., K.S. and U.Ç.; data curation, M.G.; formal analysis M.G., K.S., U.Ç., S.A.Y., L.C. and E.N.Y.; DNA methylation analysis, S.A.Y.; network analysis, K.S.; writing original draft preparation, M.G. and K.S.; manuscript review and editing, M.G., K.S., U.Ç., S.A.Y., L.C., E.N.Y., K.D. and C.W.; supervision, F.B., K.D., C.W. and J.K.;

project administration, M.G., K.S. and J.K. All authors have read and agreed to the published version of the manuscript.

Funding: M.G. and K.S. were funded in by Uppsala University and the eSSence program. L.C. was supported by the University of Pavia fellowship. The salary of S.A.Y. was supported by a grant from the Knut and Alice Wallenberg Foundation held by Linda Holmfeldt (KAW 2013-0159). C.W. was supported by the Swedish Cancer Foundation (180765). J.K. was funded in by grants from the Polish National Science Centre (DEC-2015/16/W/NZ2/00314), The University of Washington, Seattle, The National Institute of Allergy and Infectious Diseases, Division of AIDS, National Institutes of Health (ABL Contract No. HHSN272201700010I) and the eSSence program.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data used in this study were acquired from public resources: https://xenabrowser.net/hub/ (accessed on 6 April 2021), http://www.cgga.org.cn/download.jsp (accessed on 10 February 2021) and http://www.gsea-msigdb.org/gsea/msigdb/ (accessed on 27 April 2021).

Acknowledgments: The results shown here are in whole or part based upon data generated by the TCGA Research Network: https://www.cancer.gov/tcga (accessed on 17 February 2021). The authors would like to thank two anonymous reviewers for their insightful suggestions and careful reading of the manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Chen, R.; Smith-Cohn, M.; Cohen, A.L.; Colman, H. Glioma subclassifications and their clinical significance. *Neurotherapeutics* **2017**, *14*, 284–297. [CrossRef] [PubMed]
2. Miller, K.D.; Fidler-Benaoudia, M.; Keegan, T.H.; Hipp, H.S.; Jemal, A.; Siegel, R.L. Cancer statistics for adolescents and young adults, 2020. *CA A Cancer J. Clin.* **2020**, *70*, 443–459. [CrossRef] [PubMed]
3. Louis, D.N.; Ohgaki, H.; Wiestler, O.D.; Cavenee, W.K.; Burger, P.C.; Jouvet, A.; Scheithauer, B.W.; Kleihues, P. The 2007 WHO classification of tumours of the central nervous system. *Acta Neuropathol.* **2007**, *114*, 97–109. [CrossRef] [PubMed]
4. Louis, D.N.; Perry, A.; Reifenberger, G.; Von Deimling, A.; Figarella-Branger, D.; Cavenee, W.K.; Ohgaki, H.; Wiestler, O.D.; Kleihues, P.; Ellison, D.W. The 2016 World Health Organization classification of tumors of the central nervous system: A summary. *Acta Neuropathol.* **2016**, *131*, 803–820. [CrossRef] [PubMed]
5. Kobayashi, K.; Miyake, M.; Takahashi, M.; Hamamoto, R. Observing deep radiomics for the classification of glioma grades. *Sci. Rep.* **2021**, *11*, 10942. [CrossRef]
6. Network, C.G.A.R. Comprehensive, integrative genomic analysis of diffuse lower-grade gliomas. *N. Engl. J. Med.* **2015**, *372*, 2481–2498.
7. Tu, J.; Fang, Y.; Han, D.; Tan, X.; Jiang, H.; Gong, X.; Wang, X.; Hong, W.; Wei, W. Activation of nuclear factor-κB in the angiogenesis of glioma: Insights into the associated molecular mechanisms and targeted therapies. *Cell Prolif.* **2021**, *54*, e12929. [CrossRef]
8. Hayden, M.S.; Ghosh, S. NF-κB in immunobiology. *Cell Res.* **2011**, *21*, 223–244. [CrossRef]
9. Cohen, A.L.; Colman, H. Glioma biology and molecular markers. *Curr. Underst. Treat. Gliomas* **2015**, *163*, 15–30. [CrossRef]
10. Network, C.G.A.R. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* **2008**, *455*, 1061.
11. Weinstein, J.N.; Collisson, E.A.; Mills, G.B.; Shaw, K.R.M.; Ozenberger, B.A.; Ellrott, K.; Shmulevich, I.; Sander, C.; Stuart, J.M. The cancer genome atlas pan-cancer analysis project. *Nat. Genet.* **2013**, *45*, 1113–1120. [CrossRef] [PubMed]
12. Rasnic, R.; Brandes, N.; Zuk, O.; Linial, M. Substantial batch effects in TCGA exome sequences undermine pan-cancer analysis of germline variants. *BMC Cancer* **2019**, *19*, 783. [CrossRef] [PubMed]
13. Ibing, S.; Michels, B.E.; Mosdzien, M.; Meyer, H.R.; Feuerbach, L.; Körner, C. On the impact of batch effect correction in TCGA isomiR expression data. *NAR Cancer* **2021**, *3*, zcab007. [CrossRef] [PubMed]
14. Leek, J.T.; Scharpf, R.B.; Bravo, H.C.; Simcha, D.; Langmead, B.; Johnson, W.E.; Geman, D.; Baggerly, K.; Irizarry, R.A. Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat. Rev. Genet.* **2010**, *11*, 733–739. [CrossRef] [PubMed]
15. Collado-Torres, L.; Nellore, A.; Kammers, K.; Ellis, S.E.; Taub, M.A.; Hansen, K.D.; Jaffe, A.E.; Langmead, B.; Leek, J.T. Reproducible RNA-seq analysis using recount2. *Nat. Biotechnol.* **2017**, *35*, 319–321. [CrossRef] [PubMed]
16. Goldman, M.J.; Craft, B.; Hastie, M.; Repečka, K.; McDade, F.; Kamath, A.; Banerjee, A.; Luo, Y.; Rogers, D.; Brooks, A.N. Visualizing and interpreting cancer genomics data via the Xena platform. *Nat. Biotechnol.* **2020**, *38*, 675–678. [CrossRef]
17. Deo, R.C. Machine learning in medicine. *Circulation* **2015**, *132*, 1920–1930. [CrossRef]

18. Alimadadi, A.; Aryal, S.; Manandhar, I.; Munroe, P.B.; Joe, B.; Cheng, X. Artificial intelligence and machine learning to fight COVID-19. *Physiol. Genom.* **2020**, *52*, 200–202. [CrossRef]
19. Serra, A.; Galdi, P.; Tagliaferri, R. Machine learning for bioinformatics and neuroimaging. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **2018**, *8*, e1248. [CrossRef]
20. Qiu, J.; Wu, Q.; Ding, G.; Xu, Y.; Feng, S. A survey of machine learning for big data processing. *EURASIP J. Adv. Signal Process.* **2016**, *2016*, 67. [CrossRef]
21. Kourou, K.; Exarchos, T.P.; Exarchos, K.P.; Karamouzis, M.V.; Fotiadis, D.I. Machine learning applications in cancer prognosis and prediction. *Comput. Struct. Biotechnol. J.* **2015**, *13*, 8–17. [CrossRef]
22. Murdoch, W.J.; Singh, C.; Kumbier, K.; Abbasi-Asl, R.; Yu, B. Definitions, methods, and applications in interpretable machine learning. *Proc. Natl. Acad. Sci. USA* **2019**, *116*, 22071–22080. [CrossRef] [PubMed]
23. Soneson, C.; Gerster, S.; Delorenzi, M. Batch effect confounding leads to strong bias in performance estimates obtained by cross-validation. *PLoS ONE* **2014**, *9*, e100335. [CrossRef] [PubMed]
24. Subramanian, A.; Tamayo, P.; Mootha, V.K.; Mukherjee, S.; Ebert, B.L.; Gillette, M.A.; Paulovich, A.; Pomeroy, S.L.; Golub, T.R.; Lander, E.S. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 15545–15550. [CrossRef] [PubMed]
25. Liberzon, A.; Birger, C.; Thorvaldsdóttir, H.; Ghandi, M.; Mesirov, J.P.; Tamayo, P. The molecular signatures database hallmark gene set collection. *Cell Syst.* **2015**, *1*, 417–425. [CrossRef] [PubMed]
26. Liberzon, A.; Subramanian, A.; Pinchback, R.; Thorvaldsdóttir, H.; Tamayo, P.; Mesirov, J.P. Molecular signatures database (MSigDB) 3.0. *Bioinformatics* **2011**, *27*, 1739–1740. [CrossRef] [PubMed]
27. Zhao, Z.; Zhang, K.-N.; Wang, Q.; Li, G.; Zeng, F.; Zhang, Y.; Wu, F.; Chai, R.; Wang, Z.; Zhang, C. Chinese Glioma Genome Atlas (CGGA): A comprehensive resource with functional genomic data from Chinese glioma patients. *Genom. Proteom. Bioinform.* **2021**, *19*, 1–12. [CrossRef] [PubMed]
28. Grossman, R.L.; Heath, A.P.; Ferretti, V.; Varmus, H.E.; Lowy, D.R.; Kibbe, W.A.; Staudt, L.M. Toward a shared vision for cancer genomic data. *N. Engl. J. Med.* **2016**, *375*, 1109–1112. [CrossRef]
29. Vivian, J.; Rao, A.A.; Nothaft, F.A.; Ketchum, C.; Armstrong, J.; Novak, A.; Pfeil, J.; Narkizian, J.; Deran, A.D.; Musselman-Brown, A. Toil enables reproducible, open source, big biomedical data analyses. *Nat. Biotechnol.* **2017**, *35*, 314–316. [CrossRef]
30. Leek, J.T.; Johnson, W.E.; Parker, H.S.; Jaffe, A.E.; Storey, J.D. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics* **2012**, *28*, 882–883. [CrossRef]
31. Povey, S.; Lovering, R.; Bruford, E.; Wright, M.; Lush, M.; Wain, H. The HUGO gene nomenclature committee (HGNC). *Hum. Genet.* **2001**, *109*, 678–680. [CrossRef] [PubMed]
32. Barbie, D.A.; Tamayo, P.; Boehm, J.S.; Kim, S.Y.; Moody, S.E.; Dunn, I.F.; Schinzel, A.C.; Sandy, P.; Meylan, E.; Scholl, C. Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1. *Nature* **2009**, *462*, 108–112. [CrossRef] [PubMed]
33. Dramiński, M.; Koronacki, J. rmcfs: An R package for Monte Carlo feature selection and interdependency discovery. *J. Stat. Softw.* **2018**, *85*, 1–28. [CrossRef]
34. Hornik, K.; Buchta, C.; Zeileis, A. Open-source machine learning: R meets Weka. *Comput. Stat.* **2009**, *24*, 225–232. [CrossRef]
35. Platt, J. *Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines*; Microsoft Research: Redmond, WA, USA, 1998; pp. 1–21.
36. Aha, D.W.; Kibler, D.; Albert, M.K. Instance-based learning algorithms. *Mach. Learn.* **1991**, *6*, 37–66. [CrossRef]
37. Breiman, L. Bagging predictors. *Mach. Learn.* **1996**, *24*, 123–140. [CrossRef]
38. Quinlan, J.R. *C4. 5: Programs for Machine Learning*; Elsevier: Amsterdam, The Netherlands, 2014.
39. Cohen, W.W. Fast effective rule induction. In *Machine Learning Proceedings 1995*; Elsevier: Amsterdam, The Netherlands, 1995; pp. 115–123.
40. Garbulowski, M.; Diamanti, K.; Smolińska, K.; Baltzer, N.; Stoll, P.; Bornelöv, S.; Øhrn, A.; Feuk, L.; Komorowski, J.R. ROSETTA: An interpretable machine learning framework. *BMC Bioinform.* **2021**, *22*, 110. [CrossRef]
41. Pawlak, Z. Rough sets. *Int. J. Comput. Inf. Sci.* **1982**, *11*, 341–356. [CrossRef]
42. Pita-Juárez, Y.; Altschuler, G.; Kariotis, S.; Wei, W.; Koler, K.; Green, C.; Tanzi, R.; Hide, W. The pathway Coexpression network: Revealing pathway relationships. *PLoS Comput. Biol.* **2018**, *14*, e1006042. [CrossRef]
43. Chen, D.; Zhang, H.; Lu, P.; Liu, X.; Cao, H. Synergy evaluation by a pathway–pathway interaction network: A new way to predict drug combination. *Mol. BioSystems* **2016**, *12*, 614–623. [CrossRef]
44. Dutkowski, J.; Kramer, M.; Surma, M.A.; Balakrishnan, R.; Cherry, J.M.; Krogan, N.J.; Ideker, T. A gene ontology inferred from molecular networks. *Nat. Biotechnol.* **2013**, *31*, 38–45. [CrossRef] [PubMed]
45. Smolinska, K.; Garbulowski, M.; Diamanti, K.; Davoy, X.; Anyango, S.O.O.; Barrenäs, F.; Bornelöv, S.; Komorowski, J. VisuNet: An interactive tool for rule network visualization of rule-based learning models. *Diva* **2021**, *2*, 141–166.
46. Meyer, P.E.; Meyer, M.P.E. Package 'infotheo'. In *R Package Version*; Citeseer: Princet, NJ, USA, 2009; Volume 1.
47. Li, F.; Yi, Y.; Miao, Y.; Long, W.; Long, T.; Chen, S.; Cheng, W.; Zou, C.; Zheng, Y.; Wu, X. N6-methyladenosine modulates nonsense-mediated mRNA decay in human glioblastoma. *Cancer Res.* **2019**, *79*, 5785–5798. [CrossRef]
48. Matzuk, M.M.; Lamb, D.J. The biology of infertility: Research advances and clinical challenges. *Nat. Med.* **2008**, *14*, 1197–1213. [CrossRef] [PubMed]

49. Myllykangas, S.; Himberg, J.; Böhling, T.; Nagy, B.; Hollmén, J.; Knuutila, S. DNA copy number amplification profiling of human neoplasms. *Oncogene* **2006**, *25*, 7324–7332. [CrossRef] [PubMed]
50. Li, W.; Kessler, P.; Williams, B.R. Transcript profiling of Wilms tumors reveals connections to kidney morphogenesis and expression patterns associated with anaplasia. *Oncogene* **2005**, *24*, 457–468. [CrossRef] [PubMed]
51. Nakayama, R.; Nemoto, T.; Takahashi, H.; Ohta, T.; Kawai, A.; Seki, K.; Yoshida, T.; Toyama, Y.; Ichikawa, H.; Hasegawa, T. Gene expression analysis of soft tissue sarcomas: Characterization and reclassification of malignant fibrous histiocytoma. *Mod. Pathol.* **2007**, *20*, 749–759. [CrossRef]
52. Xia, H.; Qi, Y.; Ng, S.S.; Chen, X.; Chen, S.; Fang, M.; Li, D.; Zhao, Y.; Ge, R.; Li, G. MicroRNA-15b regulates cell cycle progression by targeting cyclins in glioma cells. *Biochem. Biophys. Res. Commun.* **2009**, *380*, 205–210. [CrossRef]
53. Liu, E.; Wu, J.; Cao, W.; Zhang, J.; Liu, W.; Jiang, X.; Zhang, X. Curcumin induces G2/M cell cycle arrest in a p53-dependent manner and upregulates ING4 expression in human glioma. *J. Neuro-Oncol.* **2007**, *85*, 263–270. [CrossRef]
54. Doan, P.; Musa, A.; Candeias, N.R.; Emmert-Streib, F.; Yli-Harja, O.; Kandhavelu, M. Alkylaminophenol induces G1/S phase cell cycle arrest in glioblastoma cells through p53 and cyclin-dependent kinase signaling pathway. *Front. Pharmacol.* **2019**, *10*, 330. [CrossRef]
55. Willems, E.; Dedobbeleer, M.; Digregorio, M.; Lombard, A.; Goffart, N.; Lumapat, P.N.; Lambert, J.; Van den Ackerveken, P.; Szpakowska, M.; Chevigné, A. Aurora A plays a dual role in migration and survival of human glioblastoma cells according to the CXCL12 concentration. *Oncogene* **2019**, *38*, 73–87. [CrossRef] [PubMed]
56. Lehman, N.L.; O'Donnell, J.P.; Whiteley, L.J.; Stapp, R.T.; Lehman, T.D.; Roszka, K.M.; Schultz, L.R.; Williams, C.J.; Mikkelsen, T.; Brown, S.L. Aurora A is differentially expressed in gliomas, is associated with patient survival in glioblastoma and is a potential chemotherapeutic target in gliomas. *Cell Cycle* **2012**, *11*, 489–502. [CrossRef] [PubMed]
57. Warner, S.L.; Munoz, R.M.; Stafford, P.; Koller, E.; Hurley, L.H.; Von Hoff, D.D.; Han, H. Comparing Aurora A and Aurora B as molecular targets for growth inhibition of pancreatic cancer cells. *Mol. Cancer Ther.* **2006**, *5*, 2450–2458. [CrossRef]
58. Liu, W.; Palovcak, A.; Li, F.; Zafar, A.; Yuan, F.; Zhang, Y. Fanconi anemia pathway as a prospective target for cancer intervention. *Cell Biosci.* **2020**, *10*, 39. [CrossRef] [PubMed]
59. Squatrito, M.; Brennan, C.W.; Helmy, K.; Huse, J.T.; Petrini, J.H.; Holland, E.C. Loss of ATM/Chk2/p53 pathway components accelerates tumor development and contributes to radiation resistance in gliomas. *Cancer Cell* **2010**, *18*, 619–629. [CrossRef] [PubMed]
60. Zhang, Y.; Dube, C.; Gibert, M.; Cruickshanks, N.; Wang, B.; Coughlan, M.; Yang, Y.; Setiady, I.; Deveau, C.; Saoud, K. The p53 pathway in glioblastoma. *Cancers* **2018**, *10*, 297. [CrossRef] [PubMed]
61. Dunn, I.F.; Heese, O.; Black, P.M. Growth factors in glioma angiogenesis: FGFs, PDGF, EGF, and TGFs. *J. Neuro-Oncol.* **2000**, *50*, 121–137. [CrossRef] [PubMed]
62. Rajan, V.; Menon, K. Involvement of microtubules in lipoprotein degradation and utilization for steroidogenesis in cultured rat luteal cells. *Endocrinology* **1985**, *117*, 2408–2416. [CrossRef]
63. Ahmad, F.; Sun, Q.; Patel, D.; Stommel, J.M. Cholesterol metabolism: A potential therapeutic target in glioblastoma. *Cancers* **2019**, *11*, 146. [CrossRef]
64. Li, D.; Li, S.; Xue, A.Z.; Callahan, L.A.S.; Liu, Y. Expression of SREBP2 and cholesterol metabolism related genes in TCGA glioma cohorts. *Medicine* **2020**, *99*, e18815. [CrossRef]
65. Cavuoto, P.; Fenech, M.F. A review of methionine dependency and the role of methionine restriction in cancer growth control and life-span extension. *Cancer Treat. Rev.* **2012**, *38*, 726–736. [CrossRef] [PubMed]
66. Calinescu, A.-A.; Castro, M.G. Microtubule targeting agents in glioma. *Transl. Cancer Res.* **2016**, *5*, S54. [CrossRef] [PubMed]
67. Therneau, T.M.; Lumley, T. Package 'survival'. *R Top Doc* **2015**, *128*, 28–33.
68. Kassambara, A.; Kosinski, M.; Biecek, P.; Fabian, S. Package 'Survminer', CRAN: CRAN Repository. 2017. Available online: https://cran.microsoft.com/snapshot/2017-04-21/web/packages/survminer/survminer.pdf (accessed on 31 December 2021).
69. Cerami, E.; Gao, J.; Dogrusoz, U.; Gross, B.E.; Sumer, S.O.; Aksoy, B.A.; Jacobsen, A.; Byrne, C.J.; Heuer, M.L.; Larsson, E. The cBio cancer genomics portal: An open platform for exploring multidimensional cancer genomics data. *Cancer Discov.* **2012**, *2*, 401–404. [CrossRef] [PubMed]
70. Gao, J.; Aksoy, B.A.; Dogrusoz, U.; Dresdner, G.; Gross, B.; Sumer, S.O.; Sun, Y.; Jacobsen, A.; Sinha, R.; Larsson, E. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci. Signal.* **2013**, *6*, pl1. [CrossRef]
71. Han, W.; Guan, W. Valproic Acid: A Promising Therapeutic Agent in Glioma Treatment. *Front. Oncol.* **2021**, *11*, 687362. [CrossRef]
72. Fang, G.; Wang, W.; Paunic, V.; Heydari, H.; Costanzo, M.; Liu, X.; Liu, X.; VanderSluis, B.; Oately, B.; Steinbach, M. Discovering genetic interactions bridging pathways in genome-wide association studies. *Nat. Commun.* **2019**, *10*, 4274. [CrossRef]
73. Pisano, C.; Tucci, M.; Di Stefano, R.F.; Turco, F.; Scagliotti, G.V.; Di Maio, M.; Buttigliero, C. Interactions between androgen receptor signaling and other molecular pathways in prostate cancer progression: Current and future clinical implications. *Crit. Rev. Oncol./Hematol.* **2021**, *157*, 103185. [CrossRef]
74. Jeong, W.-J.; Ro, E.J.; Choi, K.-Y. Interaction between Wnt/β-catenin and RAS-ERK pathways and an anti-cancer strategy via degradations of β-catenin and RAS by targeting the Wnt/β-catenin pathway. *NPJ Precis. Oncol.* **2018**, *2*, 5. [CrossRef]
75. Liu, K.-Q.; Liu, Z.-P.; Hao, J.-K.; Chen, L.; Zhao, X.-M. Identifying dysregulated pathways in cancers from pathway interaction networks. *BMC Bioinform.* **2012**, *13*, 126. [CrossRef]

76. Reimand, J.; Isserlin, R.; Voisin, V.; Kucera, M.; Tannus-Lopes, C.; Rostamianfar, A.; Wadi, L.; Meyer, M.; Wong, J.; Xu, C. Pathway enrichment analysis and visualization of omics data using g: Profiler, GSEA, Cytoscape and EnrichmentMap. *Nat. Protoc.* **2019**, *14*, 482–517. [CrossRef] [PubMed]
77. Nguyen, T.-M.; Shafi, A.; Nguyen, T.; Draghici, S. Identifying significantly impacted pathways: A comprehensive review and assessment. *Genome Biol.* **2019**, *20*, 203. [CrossRef] [PubMed]
78. Gao, F.; Wang, W.; Tan, M.; Zhu, L.; Zhang, Y.; Fessler, E.; Vermeulen, L.; Wang, X. DeepCC: A novel deep learning-based framework for cancer molecular subtype classification. *Oncogenesis* **2019**, *8*, 44. [CrossRef] [PubMed]
79. Liu, H.; Hussain, F.; Tan, C.L.; Dash, M. Discretization: An enabling technique. *Data Min. Knowl. Discov.* **2002**, *6*, 393–423. [CrossRef]
80. Kopper, P.; Pölsterl, S.; Wachinger, C.; Bischl, B.; Bender, A.; Rügamer, D. Semi-structured deep piecewise exponential models. In Proceedings of the Survival Prediction-Algorithms, Challenges and Applications, Palo Alto, CA, USA, 22–24 March 2021; pp. 40–53.
81. Kopper, P.; Wiegrebe, S.; Bischl, B.; Bender, A.; Rügamer, D. DeepPAMM: Deep Piecewise Exponential Additive Mixed Models for Complex Hazard Structures in Survival Analysis. In Proceedings of the Advances in Knowledge Discovery and Data Mining (PAKDD '22), Jeju, Korea, 23–26 May 2022.
82. Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.* **2019**, *1*, 206–215. [CrossRef]
83. Au, Q.; Herbinger, J.; Stachl, C.; Bischl, B.; Casalicchio, G. Grouped feature importance and combined features effect plot. *Arxiv Prepr.* **2021**, arXiv:2104.11688.