Using Graph Neural Networks to Model the Glioblastoma Free Energy Landscape

Nitya Thakkar
Advisor: Ritambhara Singh, Reader: Stephen Bach

A thesis submitted in partial fulfillment of the requirements for the degree of Bachelor of Science with Honors in Computer Science at Brown University

Brown University
Providence, RI
May 2023

1 Abstract

Glioblastoma (GBM) is an aggressive cancer in the brain that has a high associated mortality rate. Understanding the cellular interactions within the GBM tumor environment can allow researchers to develop new therapies to more effectively treat this cancer. We developed a graph convolutional neural network (GCN) trained on gene expression data from patients with GBM. The cells from these patients were treated with Chi3l1, a protein highly expressed in GBM that makes the cells less responsive to therapeutics, and then the surprisal analysis framework was used to extract information about cell state energy from the data. Using this data, we trained a GCN to predict the cell state energy given the gene expression data, and to learn the underlying graph structure of the gene-gene interactions. Our model received consistent testing AUPRC scores across many runs. We visually modeled the gene-gene interactions to understand how the underlying graph structure changed after training, and analyzed the top genes relevant to the model predictions. Furthermore, we conducted a gene set enrichment analysis on the most important genes, and found correlations between the pathways and the dataset. We aim to further analyze the robustness of the model and apply existing interpretation methods to better understand the predictions. Ultimately, these findings can better inform biological research and contribute to developing novel GBM therapies.

2 Introduction

Glioblastoma (GBM) is a fast-growing and aggressive cancer that can occur in the brain or spinal cord. GBM accounts for nearly 50 percent of all malignant brain tumors, making it the most common primary malignant brain tumor in adults. Currently, the average length of survival for GBM patients is estimated at around 8 months and the five-year survival rate for GBM patients is less than 7% [17]. Standard treatment of GBM is currently surgical resection followed by radiation. However, due to the high presence of chemotherapy-resistant glioma stem cells (GSCs), tumor recurrence is near universal [18]. Research into understanding the cellular mechanisms and interactions within the GBM tumor environment and among GSCs is imperative to develop new therapies to effectively treat this cancer.

Gene expression data, specifically single cell RNA sequencing (scRNA-seq) data, provides a crucial avenue to understanding cellular responses [9]. This data enables determining gene expression levels in individual cells. In addition to gene expression data, understanding the cellular states within the cancerous environment is important to reveal cellular dynamics and interactions. The free energy of a cell indicates how much potential it has in transitioning between different states, which reveals information regarding cell dynamics. This can lead to further insights into biological phenotypes, such as understanding the spatial distributions of GSCs, the mechanisms of cellular interactions, and cell responses to drug treatments [13].

Collaborators in Dr. Nikos Tapinos's lab worked to create a novel data source from four patients with GBM. The data contains information about gene expression levels, as well as unique information about cell state energy obtained using surprisal analysis. No prior work has been conducted using this data, providing a rich data source to understand cellular and genetic interactions.

In this work, we propose a novel method of understanding and modeling the GBM cellular environment using graph convolutional neural networks (GCN) and free energy landscapes. We trained a GCN to learn the underlying graph structure of the gene interactions, and then learned the gene-gene interactions most important to predicting the cells energy state.

In summary, we make the following contributions: (1) We propose a GCN that is able to learn the edge weights through training; (2) we perform a robustness analysis to evaluate the certainty of the predictions; (3) we analyze the genes that appear to be most relevant, as well as their downstream processes. The code is publicly available at https://github.com/rsinghlab/GBM-GNN.git.

Ultimately, this project has the potential to impact the field of personalized medicine as well as cancer research. By developing a method that uses a patient's unique cellular environment to understand the specific interactions of their cells and the energetic states, researchers understand the specific genes and mechanisms responsible for cancerous proliferation. This can enable targeted therapies downstream, with the potential to impact how clinicians approach treating GBM.

3 Background

3.1 Surprisal Analysis

The dataset used in this paper contains information about gene expression levels, as well as unique information about cell state energy. It was obtained using surprisal analysis, defined as:

$$\text{surpisal} = \ln\left(\frac{X(t)}{X^0(t)}\right) = \sum_{\alpha} \lambda_\alpha G_\alpha(t)$$

Where $X^0(t)$ is the balanced state or prior probability, and $X(t)$ is the mutated state. This expression is equivalent to the sum over the constraints, which is a measure of the deviation from the balanced state, ranked by the Lagrange multipliers. $G_\alpha(t)$ is the value of the constraint α for the event t [16]. This measurement is useful when measuring the expression level in cancerous vs normal tissues. Furthermore, surprisal analysis allows us to look at patient-specific information in relation to the free energy landscape. From this analysis, two values are derived: lambda and g values. The lambda values indicate the importance of a cell in a certain process state (from steady state to high entropy), and the g values indicate the importance of a gene in a process. For the purposes of the experiments run in the paper, we only consider steady state (process 0) and processes 1-2. Process 3 and 4 are ignored due to a high class imbalance.

4 Related Works

As this specific GBM data is novel, there is no existing literature on using it within a machine learning framework. Furthermore, there has been no prior work using machine learning methods to make predictions based off of surprisal analysis.

4.1 Surprisal Analysis of GBM data

Prior work in revealing the role of cell state energy in understanding biological phenomenon has provided a rich source of new data, yet there has been limited work in the area.

Data about cellular energetic states can lead to further insights into their biological phenotypes, such as understanding the spatial distributions of GSCs, the mechanisms of cellular interactions, and cell responses to drug treatments. Kravchenko-Balasha, 2020 found that an aggressively growing tumor for patients with GBM is not in a balanced state. Surprisal analysis techniques consider the environmental and genomics constraints that prevent this cellular environment from reaching the balanced state. Therefore, the surprisal analysis method takes as input the gene expression levels of the cancerous cells and extracts the expected distribution of the expression levels in the balanced state. It then compares this expected distribution to the observed distribution to understand the deviation [13].

Additional work using surprisal analysis to understand GBM found that, when comparing eight GBM tumors, each had a different distribution of cellular processes in the unbalanced state. The authors proposed that the identification of specific constraints on each unique GBM tumor suggested tumor-specific therapeutics may be needed to return the tumors to their balanced state [14].

More generally, surprisal analysis can allow for advancements in the field of personalized medicine. Inter-tumor heterogeneity negatively impacts the effectiveness of treatment by clinicians as patients' unique cellular environments react differently. Therefore, using surprisal analysis to understand a patients individual cellular environment can allow for more effective therapeutic treatments and therefore improved treatments for GBM [13].

Alkhatib et al., 2022 tested this approach in the lab. They used surprisal analysis techniques to create an individualized assessment of tumor composition to create a customized treatment approach. They were able to locate distinct subpopulations that were evolving within the tumor in response to therapeutics, and map them. Using these mappings, they created a guided drug cocktail and found that it significantly enhanced tumor response to therapeutics and lowered occurrence of regrowth [1].

These findings motivated the strong potential of surprisal analysis to drive cancer therapeutics at a personalized level.

4.2 GNNs to Model Genetic Data

Recent works have studied the relevance of GNNs and graph convolutional network (GCNs) to model genetic datasets. Wang et al., 2021 developed a GNN (titled scGNN) that formulated the cell-cell relationships in a graphical structured to model heterogeneous gene expression patterns. Then, they integrated multi-modal autoencoders for gene imputation and cell clustering. Gene imputation aims to solve the issue that arises in scRNA-seq data where the expression values for many genes are marked as zero by recovering (guessing) the correct values [19].

Further work by Yuan and Bar-Joseph, 2020 created a GCN (titled GCNG) to encode the spatial information of cells as a graph and combine it with expression data. Their model is able to propose pairs of extracellular interacting genes, as well as be used for downstream analysis in areas such as functional gene assignment [25].

Similar to the scGNN paper, additional researchers developed sigGCN for cell classification that combines a GCN and a neural network to reveal gene interaction networks. They found that combining prior knowledge about gene-gene interactions with gene expression data using a GCN allowed them to extract many important features and improve upon existing SOTA methods for cell classification performance [20].

Motivated by these results, we modeled our approach as a GCN to efficiently extract features from the gene-gene interactions and gene expression data for cell energy state classification.

5 Methods

5.1 Graph Convolutional Network

A graph convolutional network (GCN) is an approach for semi-supervised learning on graph-structured data [11]. GCNs are advantageous as they can utilize the power of convolution even for cases with sparse spatial relationships. Instead of encoding data as a 2D matrix or 1D vector, GCNs encode the data as a graph structure to extrapolate relationships between the samples [22].

This graph structure, which is represented as a normalized interaction matrix, is convolved with the node features in the graph. Therefore, the GCN can utilize information inherent to each node as well as the relationships among the nodes.

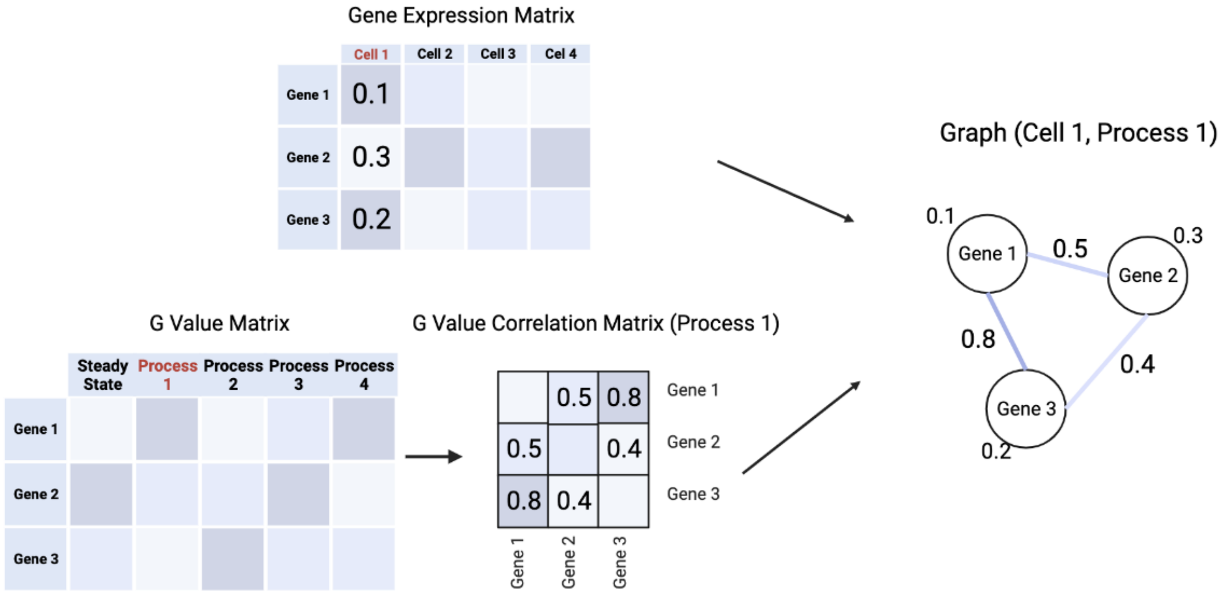


Figure 1: This is an overview figure of the graph initialization process. A graph was created for each cell, and each node in the graph is a gene where the node features are the gene expression value for that cell. The edges are initialized using the g value correlation matrix.

We formulate the graph as follows. There is one graph generated for each cell, where within each graph every gene is a node. Each node has a corresponding node feature of the gene expression value for that specific gene and cell. Each process is treated as a separate model. For a given process, the G value matrix is subsetted and constructed into a cosine similarity matrix. It is then normalized and bounded between -1 and 1, thus becoming a correlation matrix. This correlation matrix becomes the bases for the edge connections and edge weights. In this regard, the initial graph is a fully connected graph, where the weight between two nodes is the g value correlation value. Lastly, the graph-wise label is the lambda value for that specific cell, given the current process. The label is binarized to 0 or 1 depending on whether a cell is in the process or not. See Figure 1 for a visual overview of the graph initialization process.

The goal of training was to learn how to predict the cell state given the graph structure. Additionally, the model learned the edge weights as it trained, revealing the most important gene-gene interactions resulting in the predictions.

5.2 Graph Visualization

To understand the most important gene-gene interactions driving the cell process prediction, we plotted a subgraph before and after training. Using the Python NetworkX package, we first extracted the 100 highest edge weights from the initialized graph. These edge weights are initially based on the g value correlation matrix. The top 100 weights were then normalized between 0-1 and then the top 50 edge weights were taken. Edge weights are colored by value, with a darker edge color representing a higher edge weight and therefore more importance. The graph nodes, representing genes, are sized according to their degree.

5.3 Model Architecture

The GCN consisted of two Graph Convolutional layers from the Pytorch Geometric codebase [5]. In addition to passing in the graph structure to the layers, the edge weights - with a sigmoid activation function applied to them - were also passed in. The edge weights, initialized to the g value correlation matrix, were passed into the model as a trainable parameter. Following the two convolution layers, a global mean pool was applied as we are doing graph-wise classification (rather than node-wise classification). Finally, a dense layer with softmax was applied.

The loss function used was Cross Entropy loss, which is summarized by the following

$$L_{CE} = -\sum_{i=1}^{n} t_i \log(p_i), \text{ for n classes}$$

where t_i is the truth label, and p_i is the softmax probability for the i^{th} class. To report training metrics, I computed the area under Precision-Recall (PR) curve (AUPRC), which is a useful performance metric for imbalanced data. The data we used was imbalanced, as for a given process there is not guaranteed to be – and often not – an equal number of cells in the process compared to not in the process.

The code was written using the Pytorch framework. The graphical representation was constructed using Pytorch Geometric. The models were run using the computational resources and services at the Center for Computation and Visualization at Brown University.

6 Experiments

6.1 Data

The data used for this project was collected by collaborators in Dr. Nikos Tapinos's lab. The data is from four patients with GBM. The glioma stem cells in these patients were treated with Chi3l1, a secreted protein highly expressed in GBM. This protein is a modulator of GSC cellular states and reduces transition probabilities of GSCs towards terminal cellular states [6]. The data contained single cell gene expression levels for all four patients, and data regarding cell state energy was obtained using surprisal analysis. The data was post-processed to remove all mitochondrial and ribosomal genes.

Not all of the cells were treated, resulting in two datasets: control and treated. Therefore, in total we trained six models, one for all combinations of each process (0-2) and control vs treated.

6.2 Model Performance

The experiments outlined in the following sections are based off of these hyperparameter tuning results. All experiments were run on one patient's data, gb2. Each process was ran as a separate model. In total, we trained six models.

Given these results, we used a learning rate of 5e-4 and a hidden layer size of 20 for all models in process 0 and 1. For the for the process 2 control model, the AUPRC score for a hidden size of 100 with a learning rate of 5e-4 was 0.6806, which is marginally different than the best AUPRC reported above. Similarly, for the process 2 treated model, when hyperparameter tuning was run the testing AUPRC for a hidden size of 100 with a learning rate of 5e-4 was 0.6977. Therefore,

Process	Model Type	Testing AUPRC	Hidden Layer Size	Learning Rate
0	Control	1.0	20	0.0005
0	Treated	1.0	20	0.0005
1	Control	0.8238	20	0.0005
1	Treated	0.8585	20	0.0005
2	Control	0.6820	100	0.0001
2	Treated	0.6992	100	0.001

Table 1: Hyperparameter tuning results processes 0-2 for both control and treated models. ”Hidden Layer Size” refers to the GCN convolution layer’s hidden size. For all models, an optimizer of Adam was used.

for the process 2 models we used a learning rate of 5e-4 and a hidden layer size of 100. We report average testing AUPRC across 10 different starting seeds in figure 2.

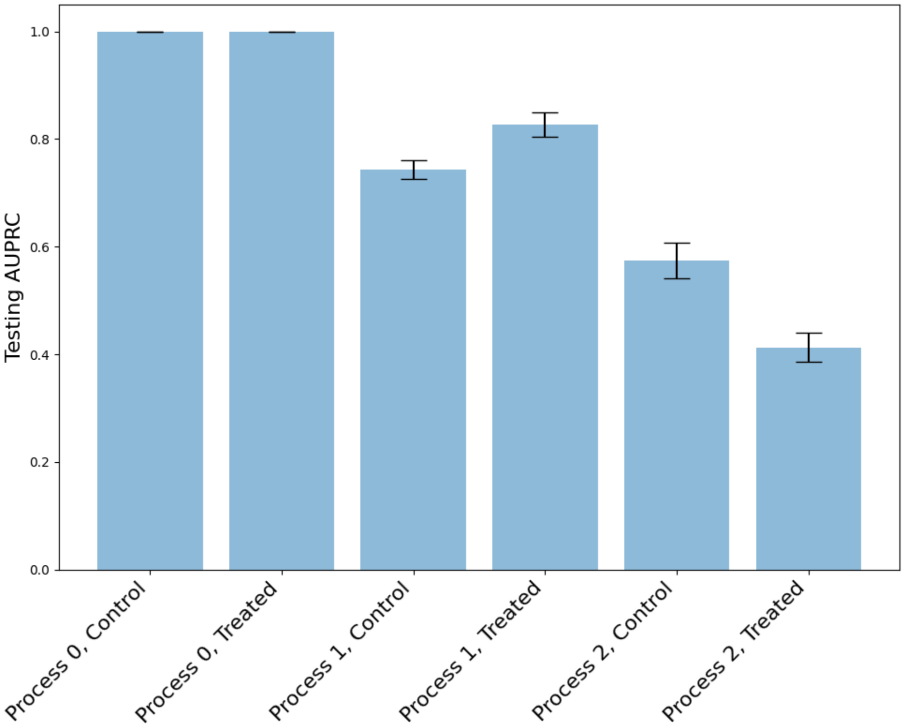


Figure 2: The average testing AUPRC across 10 different starting seeds, with standard deviation reported. All models were run using the hyperparameters described above.

6.3 Gene-Gene Interactions

We plotted the 50 most important edges within the subgraph before and after training the model. Below are the results for all six models. For all graphs, the node size is relative to the node degree, that is how many edges the node has. The edges are colored based on weight value, with a darker color indicating a higher weight and therefore importance. In the following panels, the subgraph on the left is the 50 most important edges from the g value correlation matrix initialization, and the subgraph on the right is the 50 most important edges in the graph structure after training. See Figures 4, 5, 6, 7, 8, 9.

Based on this analysis, we sought to understand if the most important genes were solely a result of having the greatest gene expression value. We plotted the average expression value for all genes in the data. The results are summarized in 3 and they indicate that gene expression values do drive the importance for some genes.

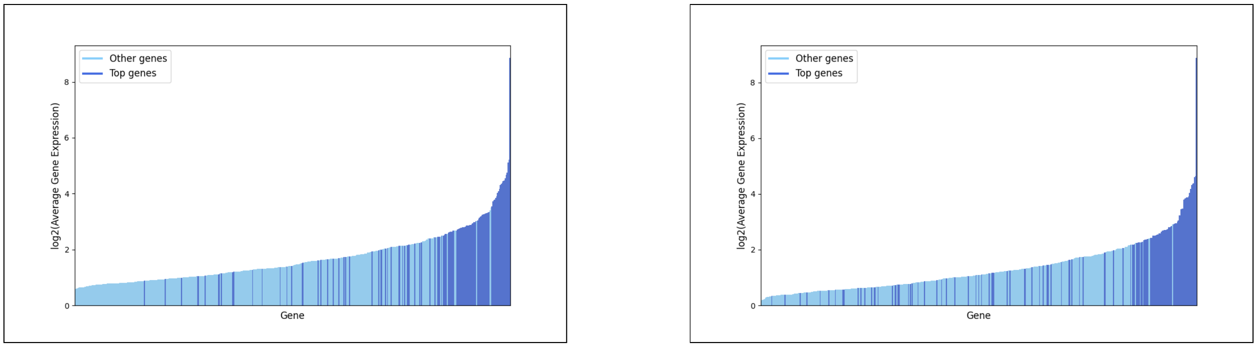


Figure 3: Left: control expression values, Right: treated expression values - treated genes are in dark blue. These plots indicate that gene expression value do drive the importance for some genes.

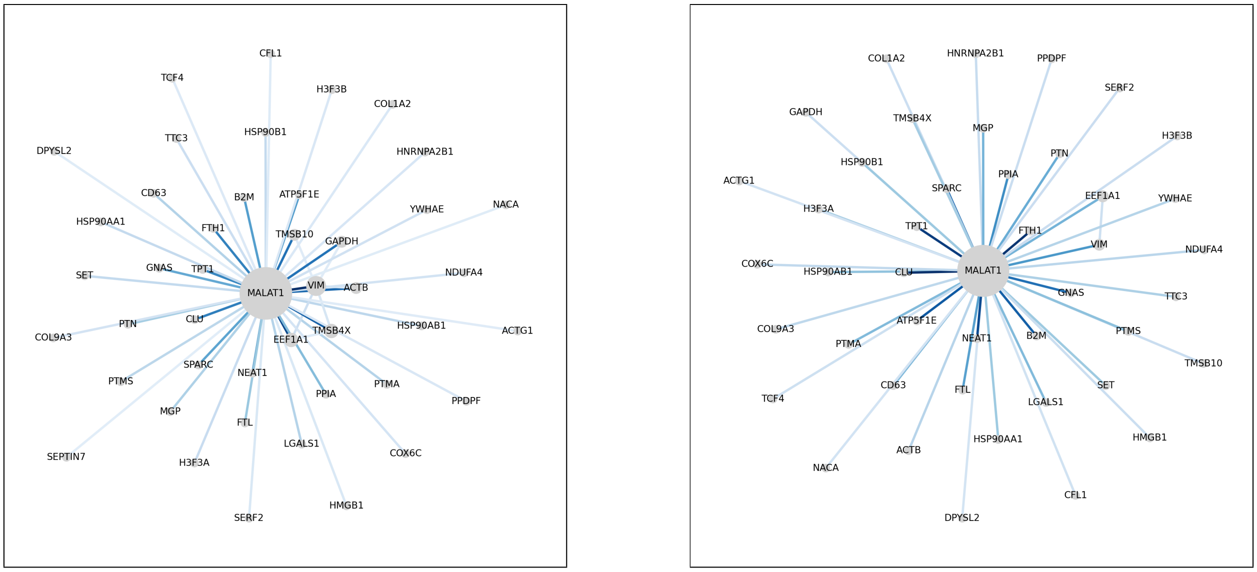


Figure 4: Process 0 control subgraph results indicate that the MALAT1 gene plays a vital role in predicting cell state from gene expression values. Left: 50 most important edges in graph from g value correlation matrix initialization. Right: 50 most important edges in graph after training.

6.4 Top Gene Analysis

Based on the gene-gene interaction subgraphs produced, we sought to analyze the central genes after training to understand their role in GBM. Central to process 0 - both control and treated - was MALAT1 (metastasis-associated lung adenocarcinoma transcript 1) (see Figure 4, 5). MALAT1 is

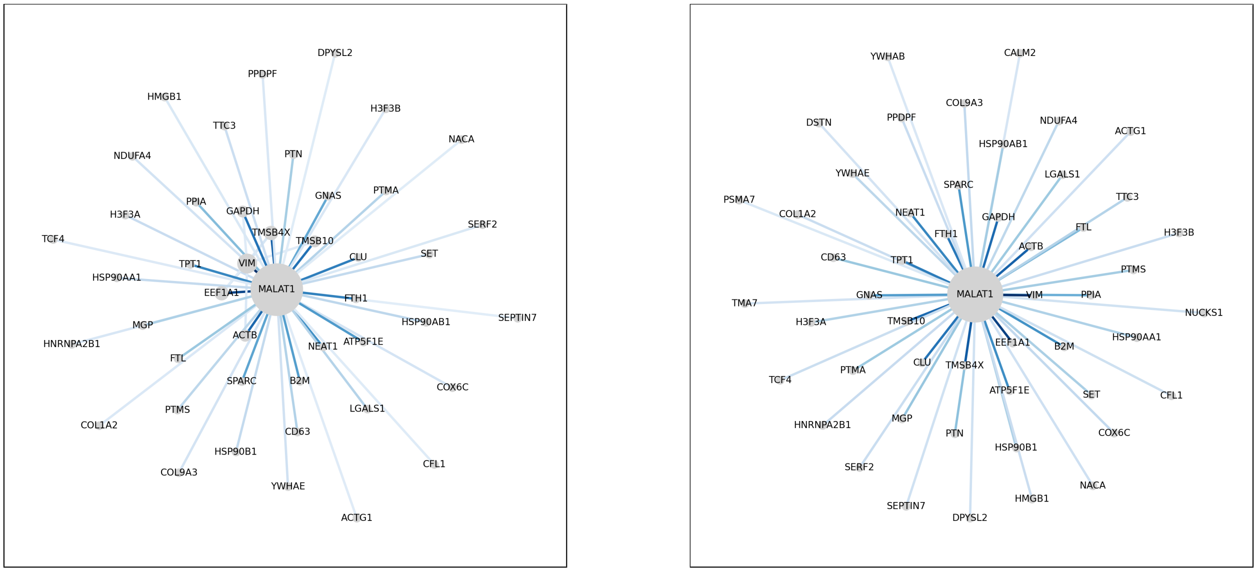


Figure 5: Process 0 treated subgraph results indicate that the MALAT1 gene plays a vital role in predicting cell state from gene expression values. Left: 50 most important edges in graph from g value correlation matrix initialization. Right: 50 most important edges in graph after training.

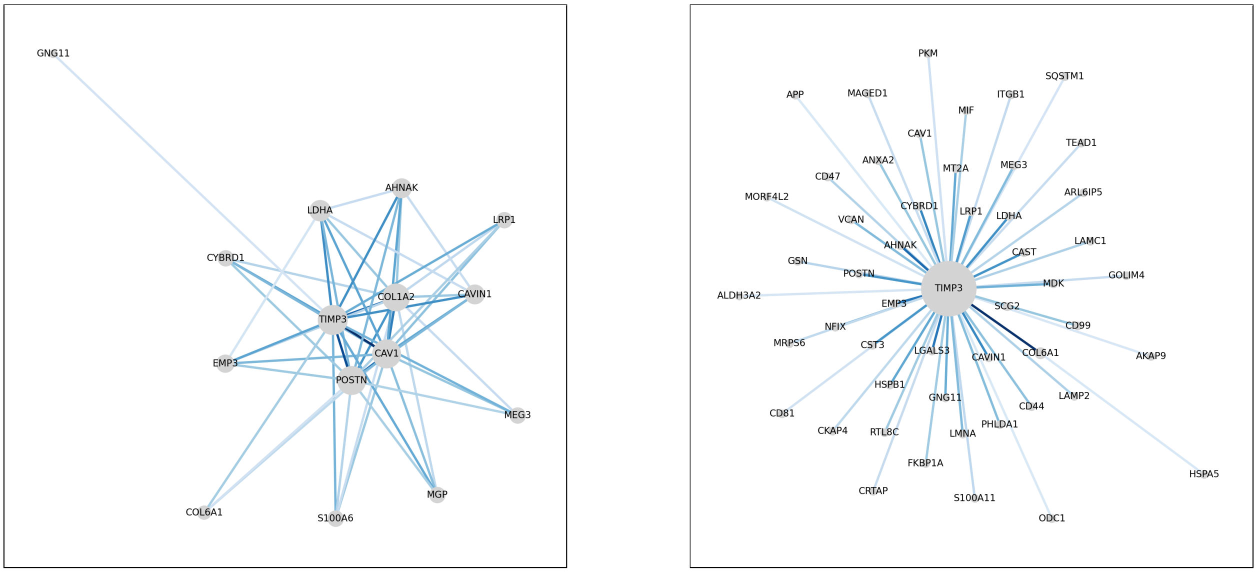


Figure 6: Process 1 control subgraph results indicate that the TIMP3 gene plays a vital role in predicting cell state from gene expression values. Left: 50 most important edges in graph from g value correlation matrix initialization. Right: 50 most important edges in graph after training.

a tumor promoter[2]. Previous studies have found that high levels of MALAT1 expression confers a poor therapeutic efficacy, and inhibiting MALAT1 levels could lead to improved therapeutic results in GBM patients [4].

Central to the process 1 control model was the TIMP3 (tissue inhibitor of metalloproteinase-3) gene (see Figure 6). Higher TIMP3 expression levels are correlated with better overall survival and prognosis in GBM patients. TIMP genes regulate many cellular processes by controlling extracellular matrix degradation [8].

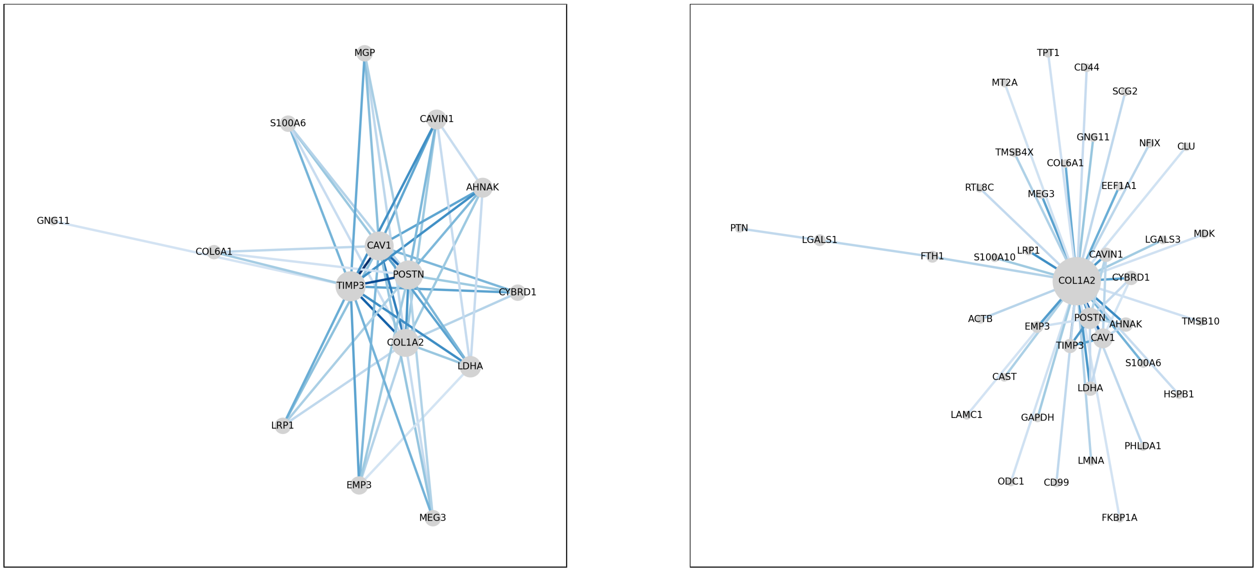


Figure 7: Process 1 treated subgraph results indicate that the COL1A2 gene plays a vital role in predicting cell state from gene expression values. Left: 50 most important edges in graph from g value correlation matrix initialization. Right: 50 most important edges in graph after training.

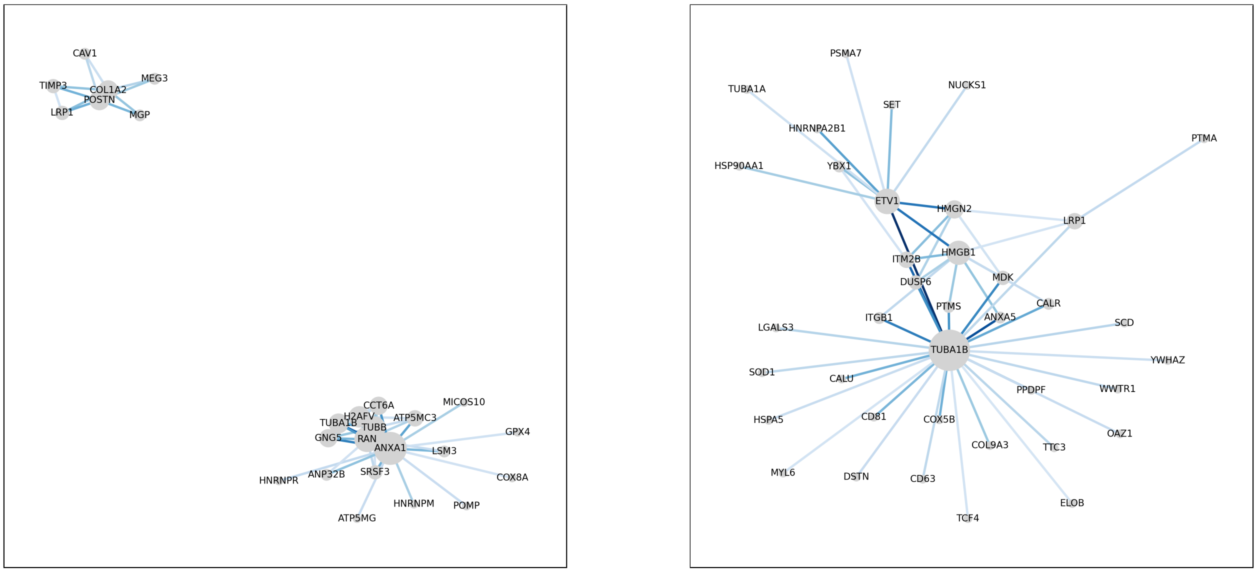


Figure 8: Process 2 control subgraph results indicate that the TUBA1B gene, along with the ETV1 and HMGB1 genes, play a vital role in predicting cell state from gene expression values. While the initialized figure indicates there are two disconnected subgraphs, after training one connected subgraph is produced with the most important edges. Left: 50 most important edges in graph from g value correlation matrix initialization. Right: 50 most important edges in graph after training.

For the process 1 treated model, a different gene is at the center of its predictions: COL1A2 (collagen alpha-2(I) chain) (see Figure 7). This gene provides instructions for making part of a large molecule called type I collagen. An extracellular matrix, such as one made of collagen, is an essential part of the tumor microenvironment. Previous studies have found that COL1A2 is significantly higher in the blood of patients with GBM than healthy patients. Furthermore,

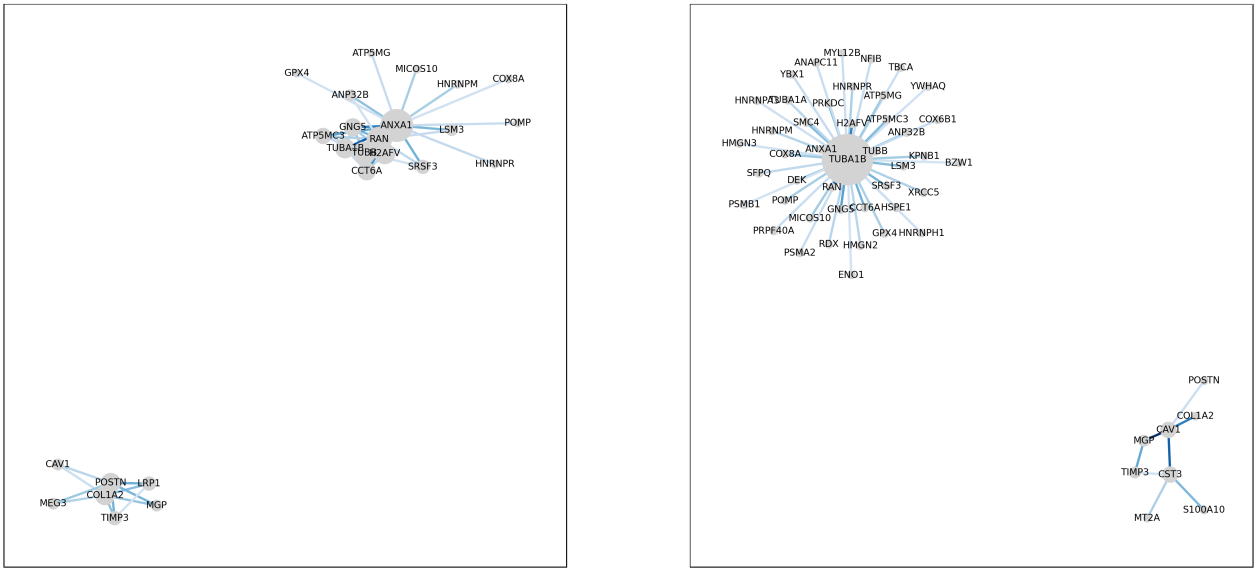


Figure 9: Process 2 treated subgraph results indicate that the TUBA1B gene plays a vital role in predicting cell state from gene expression values. Unlike process 2 control, there are still two disconnected subgraphs with the highest edge weights after training. It is important to note that, in the post-training figure, the smaller subgraph in the bottom right contains the highest edge weights (indicated by darker edge color). Therefore, the CAV1, MGP, COL1A2, and TIMP3 genes are considered important to the process 2 treated model. Left: 50 most important edges in graph from g value correlation matrix initialization. Right: 50 most important edges in graph after training.

inhibition of COL1A2 suppresses GSC proliferation [21].

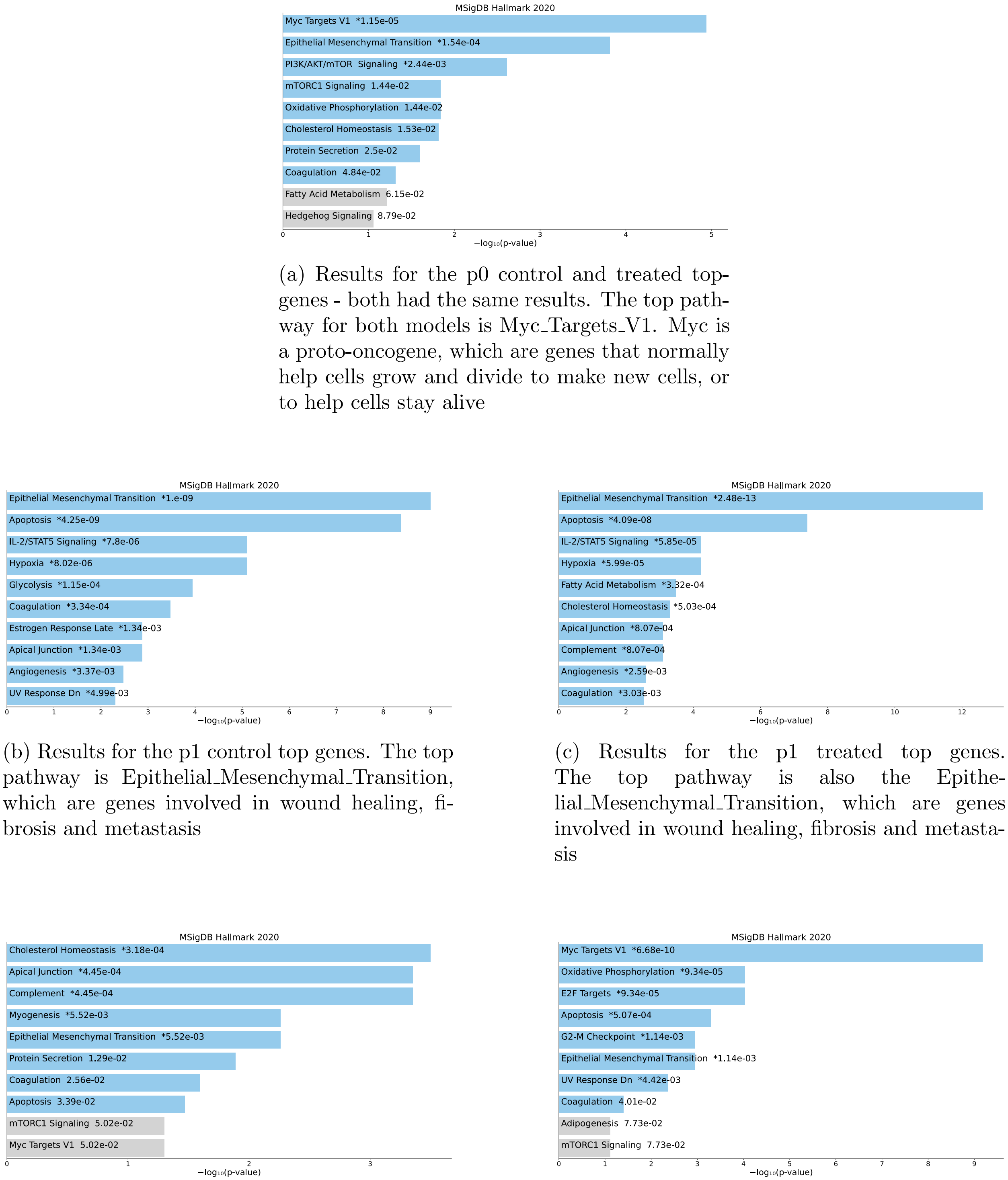
Lastly, central to both the process 2 control and treated models is the gene TUBA1B (tubulin alpha-1B) (see Figure 8, 9). TUBA1B has been shown to have increased expression in patients with GBM, and the expression increases with GBM grade. Previous studies have found TUBA1B to be a significant predictor of poor overall survival and prognosis in patients with GBM. [7]

6.5 Downstream Pathway Analysis

Model Robustness To assess the robustness of the models, we ran each of the six models above 10 times with a different starting seed. We recorded the highest 50 edge weights and corresponding edges, and determined how often that edge appeared in each of the 10 runs. For the downstream pathway analysis, we used the genes which corresponded to edges that appeared as a top 50 edge with 70% certainty. In other words, the edges that appeared in 7 out of the 10 runs.

Analysis We conducted downstream pathway analysis for all six models separately. To construct the gene set, we first determined the top 50 edge weights after training that appeared with a certainty greater than 70%. We did this using the robustness analysis detailed above. Then, for those edges we selected the genes the edges connected and used those genes to construct the gene set. We used Enrichr tool, which is an online interactive tool for gene list enrichment analysis [3, 15, 23]. We used the Hallmark gene set for this analysis, which was created by researchers at The Broad Institute and represents and summarizes well-defined biological states and processes.

The results from this analysis are summarized in Figure 10. There were no significant differences in the results comparing the process 0 control and treated models, and also comparing the process 1 control and treated models. However, for process 2 we found that genes driving homeostasis are dominant in the process 2 control top genes, and genes driving cell proliferation are dominant in the in process 2 treated top genes. These findings are significant as they indicate the model is learning the genetic differences between the control and treated genes for process 2.



(a) Results for the p0 control and treated top-genes - both had the same results. The top pathway for both models is Myc_Targets_V1. Myc is a proto-oncogene, which are genes that normally help cells grow and divide to make new cells, or to help cells stay alive

(b) Results for the p1 control top genes. The top pathway is Epithelial_Mesenchymal_Transition, which are genes involved in wound healing, fibrosis and metastasis

MSigDB Hallmark 2020
Epithelial Mesenchymal Transition *2.48e-13
Apoptosis *4.09e-08
IL-2/STAT5 Signaling *5.85e-05
Hypoxia *5.99e-05
Fatty Acid Metabolism *3.32e-04
Cholesterol Homeostasis *5.03e-04
Apical Junction *8.07e-04
Complement *8.07e-04
Angiogenesis *2.59e-03
Coagulation *3.03e-03
0
2
4
6
8
10
12
−log₁₀(p-value)

(c) Results for the p1 treated top genes. The top pathway is also the Epithelial_Mesenchymal_Transition, which are genes involved in wound healing, fibrosis and metastasis

MSigDB Hallmark 2020
Cholesterol Homeostasis *3.18e-04
Apical Junction *4.45e-04
Complement *4.45e-04
Myogenesis *5.52e-03
Epithelial Mesenchymal Transition *5.52e-03
Protein Secretion 1.29e-02
Coagulation 2.56e-02
Apoptosis 3.39e-02
mTORC1 Signaling 5.02e-02
Myc Targets V1 5.02e-02
0
1
2
3
−log₁₀(p-value)

(d) Results for the p2 control top genes. The top pathway is Cholesterol_Homeostasis, which are genes involved in cholesterol homeostasis

(e) Results for the p2 treated top genes. The top pathway is Myc_Targets_V1. Myc is a proto-oncogene, which are genes that normally help cells grow and divide to make new cells, or to help cells stay alive

Figure 10: Gene set enrichment analysis results for all six models. Genes driving homeostasis are dominant in the process 2 control top genes, and genes driving cell proliferation are dominant in the in process 2 treated top genes.

7 Conclusions

We developed a GNN to predict cell state energy given gene expression data. We analyzed the gene-gene interactions most important to the model predictions, and conducted a downstream pathway analysis to understand their functions. Our findings of the most important genes can help inform future therapeutics for GBM.

We plan to test interpretation methods on the GNN in the future. As GNNs are graph-based, we are able to view exactly which nodes are most important to the predictions, compared to other black-box model architectures. Therefore, we plan to validate our findings of the most important genes by using state-of-the-art methods for GNN interpretation, such as GNNExplainer and Global Counterfactual Explainer for GNNs [24, 12]. This will further help validate which genes are most important for biologists to study.

In the future, we hope to incorporate more robust uncertainty modeling methods into the predictions. One method of doing so is by using Bayesian inference (BI), a mathematical model that allows us to determine the uncertainty in our predictions. BI is well suited for models that are trained on small datasets, which applies in this case as the data is only from four patients. Furthermore, BI reduces the risk of overconfident predictions [10]. These false predictions could have detrimental impacts in the areas of medical diagnoses and treatments.

Ultimately, this project has implications for biologists to reduce the time, money, and labor spent manually labeling this cell state energy data. By training on a variety of different patient's data, the model can also learn generalizability which can allow it to overcome the batch effect, which is a phenomenon in biology when non-biological factors in an experiment cause changes in the data produced by the experiment. This can have large potential in expanding the amount of labeled data available so researchers can better understand GBM and discover novel therapies to impact cancer treatment.

8 Acknowledgements

I would like to thank Professor Ritambhara Singh and Hyeyeon Hwang for all their mentorship, support, and guidance throughout this project. When I joined Professor Singh's lab freshman year, I didn't even know how to navigate the command line. She taught me how to be persistent and inquisitive, and I'm so grateful to have had her as my mentor. I would also like to thank my family and friends for their constant support throughout my thesis. I'm excited to continue developing my passion for research and the fields of machine learning and computational biology as I start my Ph.D. in computer science at Stanford University this fall.

References

[1] Heba Alkhatib, Ariel M. Rubinstein, Swetha Vasudevan, Efrat Flashner-Abramson, Shira Stefansky, Sangita Roy Chowdhury, Solomon Oguche, Tamar Peretz-Yablonsky, Avital Granit, Zvi Granot, Ittai Ben-Porath, Kim Sheva, Jon Feldman, Noa E. Cohen, Amichay Meirovitz, and Nataly Kravchenko-Balasha. Computational quantification and characterization of independently evolving cellular subpopulations within tumors is critical to inhibit anti-cancer therapy resistance. *Genome Medicine*, 14(120), 2022.

[2] Yucel Baspinar, Ilhan Elmaci, Aysel Ozpinar, and Meric A Altinoz. Long non-coding rna malat1 as a key target in pathogenesis of glioblastoma. janus faces or achilles' heal? *Gene*, 739, 2020.

[3] Edward Y Chen, Christopher M Tan, Yan Kou, Qiaonan Duan, Zichen Wang, Gabriela Vaz Meirelles, Neil R Clark, and Avi Ma'ayan. Enrichr: interactive and collaborative html5 gene list enrichment analysis tool. *BMC Bioinformatics*, 128(14), 2013.

[4] Wei Chen, Xin-Ke Xu, Jun-Liang Li, Kuan-Kei Kong, Hui Li, Cheng Chen, Jing He, Fangyu Wang, Ping Li, Xiao-Song Ge, and Fang-Cheng Li. Malat1 is a prognostic factor in glioblastoma multiforme and induces chemoresistance to temozolomide through suppressing mir-203 and promoting thymidylate synthase expression. *Oncotarget*, 8(14):22783–22799, 2017.

[5] Matthias Fey and Jan E. Lenssen. Fast graph representation learning with PyTorch Geometric. In *ICLR Workshop on Representation Learning on Graphs and Manifolds*, 2019.

[6] Charlotte Guetta-Terrier, David Karambizi, Bedia Akosman, Jia-Shu Chen, Suchitra Kamle, Eduardo Fajardo, Andras Fiser, Ritambhara Singh, Steven A. Toms, Chun Geun Lee, Jack A. Elias, and Nikos Tapinos. Chi3l1 is a modulator of glioma stem cell states and a therapeutic vulnerability for glioblastoma. *bioRxiv*, 2021.

[7] Liu H.-J., Wang L., Kang L., Du J., Li S., and Cui H.-X. Sulforaphane-n-acetyl-cysteine induces autophagy through activation of erk1/2 in u87mg and u373mg cells. *Cell Physiol Biochem*, 51, 2018.

[8] Jinkun Han, Yajun Jing, Fubing Han, and Peng Sun. Comprehensive analysis of expression, prognosis and immune infiltration for timps in glioblastoma. *BMC Neurology*, 2021.

[9] Ashraful Haque, Jessica Engel, Sarah A. Teichmann, and Tapio Lönnberg. A practical guide to single-cell rna-sequencing for biomedical research and clinical applications. *Genome Medicine*, 9(75), 2017.

[10] Laurent Valentin Jospin, Hamid Laga, Farid Boussaid, Wray Buntine, and Mohammed Bennamoun. Hands-on bayesian neural networks—a tutorial for deep learning users. *IEEE Computational Intelligence Magazine*, 17(2):29–48, may 2022.

[11] Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks, 2017.

[12] Mert Kosan, Zexi Huang, Sourav Medya, Sayan Ranu, and Ambuj Singh. Global counterfactual explainer for graph neural networks, 2022.

[13] Nataly Kravchenko-Balasha. Toward deciphering of cancer imbalances: Using information-theoretic surprisal analysis for understanding of cancer systems. In *Advances in Info-Metrics: Information and Information Processing across Disciplines*. Oxford University Press, 10 2020.

[14] Nataly Kravchenko-Balasha, Hannah Johnson, Forest M. White, James R. Heath, and R. D. Levine. A thermodynamic-based interpretation of protein expression heterogeneity in different glioblastoma multiforme tumors identifies tumor-specific unbalanced processes. *The Journal of Physical Chemistry*, 120(26):5990–5997, 2016.

[15] Maxim V Kuleshov, Matthew R Jones, Andrew D Rouillard, Nicolas F Fernandez, Qiaonan Duan, Zichen Wang, Simon Koplev, Sherry L Jenkins, Kathleen M Jagodnik, Alexander Lachmann, Michael G McDermott, Caroline D Monteiro, Gregory W Gundersen, and Avi Ma'ayan. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Research*, 2016.

[16] F. Remacle, Nataly Kravchenko-Balasha, Alexander Levitzki, and R. D. Levine. Information-theoretic analysis of phenotype changes in early stages of carcinogenesis. *Proceedings of the National Academy of Sciences*, 107(22):10324–10329, 2010.

[17] National Brain Tumor Society. About glioblastoma.

[18] Priyanka Soni, Sumaira Qayoom, Nuzhat Husain, Praveen Kumar, Anil Chandra, Bal Krishan Ojha, and Rakesh Kumar Gupta. Cd24 and nanog expression in stem cells in glioblastoma: Correlation with response to chemoradiation and overall survival. *Asian Pacific Journal of Cancer Prevention*, 18(8):2215–2219, August 2017.

[19] Juexin Wang, Anjun Ma, Yuzhou Chang, Jianting Gong, Yuexu Jiang, Ren Qi, Cankun Wang, Hongjun Fu, Qin Ma, and Dong Xu. scgnn is a novel graph neural network framework for single-cell rna-seq analyses. *Nature Communications*, 12(1882), 2021.

[20] Tianyu Wang, Jun Bai, and Sheida Nabavi. Single-cell classification using graph convolutional networks. *Nature Communications*, 22(364), 2021.

[21] Yi Wang, Maki Sakaguchi, Hemragul Sabit, Sho Tamai, Toshiya Ichinose, Shingo Tanaka, Masashi Kinoshita, Yasuo Uchida, Sumio Ohtsuki, and Mitsutoshi Nakada. Col1a2 inhibition suppresses glioblastoma cell proliferation and invasion. *Journal of Neurosurgery*, 138(3):639 – 648, 2023.

[22] Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and Philip S. Yu. A comprehensive survey on graph neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, 32(1):4–24, jan 2021.

[23] Zhuorui Xie, Allison Bailey, Maxim V Kuleshov, Daniel J B Clarke, John E Evangelista, Sherry L Jenkins, Alexander Lachmann, Megan L Wojciechowicz, Eryk Kropiwnicki, Kathleen M Jagodnik, Minji Jeon, and Avi Ma'ayan. Gene set knowledge discovery with enrichr. *Current Protocols*, 2021.

[24] Rex Ying, Dylan Bourgeois, Jiaxuan You, Marinka Zitnik, and Jure Leskovec. Gnnexplainer: Generating explanations for graph neural networks, 2019.

[25] Ye Yuan and Ziv Bar-Joseph. Gcng: graph convolutional networks for inferring gene interaction from spatial transcriptomics data. *Genome Biology*, 21(300), 2020.