Glioblastoma Multiforme (GBM) subtype classification using SUPREME

Sriram Kandadai

1 Abstract

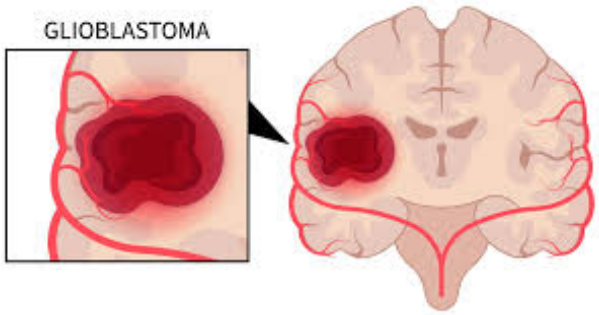

Figure 1: Glioblastma Multiforme

To advance precise cancer diagnosis of Gliblastoma multiforme (GBM), we have employed a framework called SUPREME which uses Graph convolutional neural networks for cancer sub- type prediction. The biology of a cancer varies in each molecular biomarker such as transcriptome profiles and DNA methylation. So to capture the relationships of both the molecular profiles with cancer subtypes, SUPREME model integrates embeddings of both the modalities to consider all the characteristics and biological signals of the GBM for subtype prediction. Where previous works like [8] haven't utilized multi-omics integration, where SUPREME makes use of it for greater performance.

Initially patient similarity networks are generated from each data modality based on pearson's correlation coefficient, then each modality is trained on GCN to generate the embeddings, where it uses all multi-omic features for embedding generation. On top of that, SUPREME also integrates all the raw multi-omic features with embeddings of all combinations for subtype prediction task, which exhibits an increase in accuracy. Thereby experimenting will all combinations of embeddings SUPREME is insightful about each omic features explaining their contribution in distinguishing the subtype. We used two different datasets from The Cancer Genome Atlas (TCGA) namely gene expression and DNA methylation. SUPREME performs well on the combination of both embeddings with raw features as complementary signals integrated to it .

State of the art (SOTA) model architectures like Transformers with attention mechanism are good at capturing global dependencies and dynamic relationships, we have tweaked the architecture of SUPREME by replacing GCN with Graph attention networks (GATs) which provides greater flexibility by enabling the network to focus on most relevant neighbors. SUPREME-GAT performs on par with SUPREME-GCN in macro F1 score but outperforms the SUPREME-GCN on accuracy and weighted F1 score.

2 Introduction

2.1 Motivation

Glioblastoma multiforme (GBM) is one of the highly invasive types of brain cancer, which is known for its volatile progression and complex response to treatment. Due to its intricacies and ambiguous behavior of the cancer evolution it perplexes the subtype prediction of the GBM. The advances in research have made progression to large-scale projects, like The Cancer Genome Atlas (TCGA) [2] which provides datasets like gene expression and DNA methylation profiles and additional omics data. Where using multiple types of omics data together can reveal insightful discoveries in cancer's biology. With Graph Neural networks [6] which can model the complex relationships and dependencies between biological entities such as genes ,proteins and patients. GNN is well suited for capturing these intricate relationships by treating the multi-omics as nodes and their relationship strength as edges enabling the rich structured representations.

2.2 Previous work

The traditional approaches to cancer subtype prediction only utilized the single data modalities for the subtype classification. Whereas SUPREME framework [3] includes the raw features as node features when training patient similarity networks to generate the embeddings i.e for each data type separately and finally integrates the patient embeddings with raw multi-omic features train it on MLP for subtype prediction enabling the capture of complementary biological signals. In [5] a graph convolutional network (GCN) was used to predict cancer types across 33 cancer and non-cancer categories, including normal samples from each cancer type. The network was built using gene co-expression or protein-protein interaction data, but the convolutional processing was limited to the gene expression data alone. As a result, this approach overlooked valuable information from other data modali-

ties. MOGONET [8], a supervised multi-omics framework which integrates mRNA expression, DNA methylation, and microRNA data using distinct GCN models for each modality to generate patient predictions. But it only creates datatype-specific networks and embeddings, considering only the predictions from individual models without integrating features across all the modalities.

3 Materials and Methods

	Gene expression	DNA Methylation
Total Features	5985	16244
samples	348	
Network (Nodes, Edges)	(348, 6960)	

Table 1: Summary of each datatype

3.1 Data Pre-processing

3.1.1 Gene expression

Gene expression data is retrieved from firehose broad GDAC repository which has legacy datasets of microarray-based expression analysis, which are already normalized and ready for downstream analysis. As there are many genes which are redundant we excluded the ones which has expression value ≤ 1, leaving only the genes which have meaningful expression value. Then we performed differential expression analysis between tumor and normal samples using limma package [4], where there are only 10 normal samples out of 538 samples.

We excluded the samples which has p-value $>$ 0.01 to only include the tumor related genes. Since there were no available subtype labels from the repository, we retrieved them by filtering the overlapping patient barcodes with DNA methylation dataset which has the subtype labels. We considered the first 12 characters of the barcode strings of gene expression and DNA methylation datasets to find the overlapping patients, for which we got 358 overlapping patient samples between the both data types.

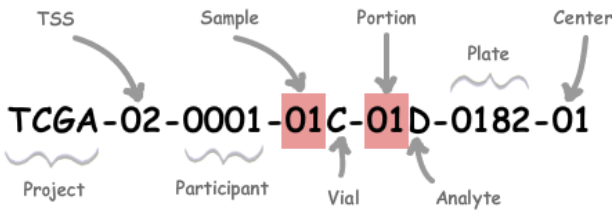

Figure 2: Patient barcode

3.1.2 DNA methylation

In TCGA repository of GBM's project, DNA methylation is available in two platforms specifically HumanMethylation450 and HumanMethylation27 has the probe level data. We concatenated both datasets into one with common genes. As mentioned section 2.1.1, we found 358 overlapping patient samples between the both data modalities. Methylation data is already normalized, so it didn't require any further pre-processing.

3.1.3 Subtype labels for GBM

Subtype labels for GBM are available as an array in the colData object of Methylation data. The subtype labels are Classical, Neural, Proneural, Mesenchymal and G-CIMP. We excluded the label values of patient samples that were filtered out during the preprocessing of the methylation and gene expression datasets. So we extracted that 358 patient labels which has 10 missing values which are omitted resulting with 348 patient labels. Finally index order of the labels are organized according to the patient samples' indexes from methylation data.

3.2 Building Patient Similarity networks

To build similarity networks for each data type we calculated pearson correlation coefficient for DNA methylation and gene expression data types. Where each node in the network is a patient sample and edges indicates correlation strength between two nodes(patients). we selected 6960 edges for both DNA methylation and gene expression data modalities, We selected the top 20 edges with the highest correlation for each node (patient). The raw omic features (DNA methylation and Gene expression) are used as node features for the corresponding datatype specific network which is served as an input to the GCN model in SUPREME framework [3], for patient embedding generation. Later these raw features are integrated with the embeddings for subtype prediction task using a ML model.

3.3 Embedding generation in SUPREME

The patient networks with raw features as the node features is given as input to the GCN model to train and generate patient embeddings. The GCN model in SUPREME generates patient embeddings by leveraging both multi-omics features linked to individual patient nodes and the structural information from their local neighborhoods. To obtain the best results, the hyper parameters like hidden layers, learning rate are tuned with 50 iterations of evaluation metrics and 500 epochs within each iteration of evaluation metric for each hyperparameter combination. The hyperparameter is finalized based on the highest median macro F1 score to select the foremost model. The patient embeddings are generated by making use of the finalized model.

This process is repeated for the other dataset by training it with a different GCN model. After gener-

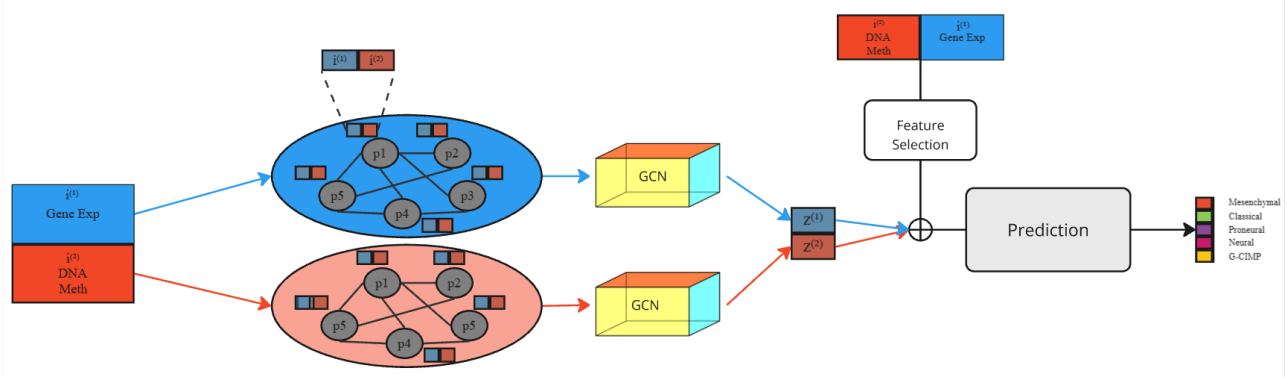

Figure 3: SUPREME for GBM

ating embeddings SUPREME will integrate them with raw features and trains a ML model with the fused data for GBM subtype prediction. There is an optional feature selection for raw embeddings before we concatenate it to the embeddings

3.4 GBM Subtype prediction

The generated embeddings of both modalities are concatenated with raw features and are trained with the ML algorithms like Multi Layer Perceptron, Random Forest. We have optimized the hyperparameters for subtype prediction task following the same approach used for embedding generation using GCN, reiterating the ML prediction run for 10 times and selected the best hyperparmater combination based on F1-macro score. Then the model with optimized hyperparameters is evaluated on unseen data with metrics like accuracy, F1-weighted and F1-macro for total of 50 runs and their average scores for each metric is computed. The subtype prediction is performed on each combination of the embeddings with raw features concatenated to it.

3.5 Graph attention network for Embedding generation

To capture the intricate subtle correlations of epigenetics and transcriptomic profiles between the patient nodes we employed Graph attention replacing Graph convolution for neighborhood aggregation mechanism in GNN.

GCN uses normalized adjacency matrix to aggregate feature information uniformly from a node's neighbor which assumes that all neighbors contribute equally, whereas GAT [7] introduces a learnable attention coefficient which makes the model to assign different importance weights to each neighbor based on their correlation strength and feature similarity.

Patient Node Feature Update Rule is given by:

$$\mathbf{h}_i^{(l+1)} = \sigma\left(\sum_{j\in\mathcal{N}_i} \alpha_{ij} W \mathbf{h}_j^{(l)}\right) \tag{1}$$

And attention coefficient is computed as follows:

$$\alpha_{ij} = \frac{\exp\left(\text{LeakyReLU}\left(\mathbf{a}^\top \left[W\mathbf{h}_i \| W\mathbf{h}_j\right]\right)\right)}{\sum_{k\in\mathcal{N}_i} \exp\left(\text{LeakyReLU}\left(\mathbf{a}^\top \left[W\mathbf{h}_i \| W\mathbf{h}_k\right]\right)\right)} \tag{2}$$

- $\mathbf{h}_i^{(l+1)}$ is the updated feature vector of node i at layer $l+1$.
- α_{ij} is the normalized attention coefficient for nodes i and j.
- $\mathbf{a}$ is the trainable attention vector.
- W is the trainable weight matrix for the layer
- $W\mathbf{h}_i$ and $W\mathbf{h}_j$ are the transformed feature vectors of nodes i and j.

GAT studies the importance of relationship during the training phase which adapts to non-uniform correlations in data. It makes use of multiple attention head to interpret the diverse relationships in the data simultaneously, which is gene expression and DNA methylation in this case.

4 Results & Discussion

4.1 Analysis of Patient Similarity Networks

Using the graph visualization tool Gephi [1] we generated the patient similarity network graph for both gene expression and DNA methylation data types.

The clusters in the graph (Figure 4) indicate the GBM subtype of the patient node where each node is annotated with distinct colors that correspond to their GBM subtype. (The subtype names for the colors are shown in figure 6).

The gene expression network reveals that Mesenchymal and Classical forms large clusters specifying robust similarity within these subtypes based on gene expression values. There are some regions which are intermixed with subtypes which possibly suggests

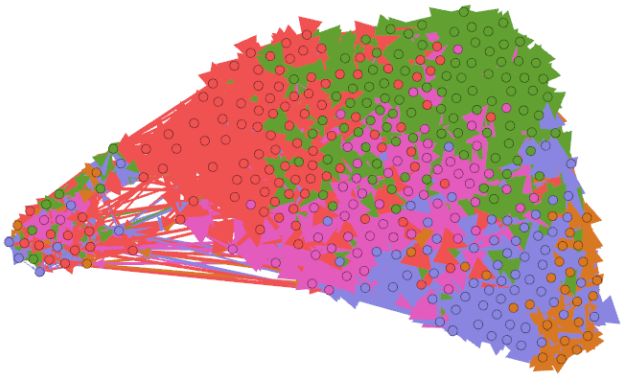

Figure 4: Similarity network of gene expression

that some patients indicate heterogeneity whose gene expression profiles overlap between these subtypes. whereas G-CIMP subtype has formed a small cluster away from the other subtypes which stipulates its unique gene expression values

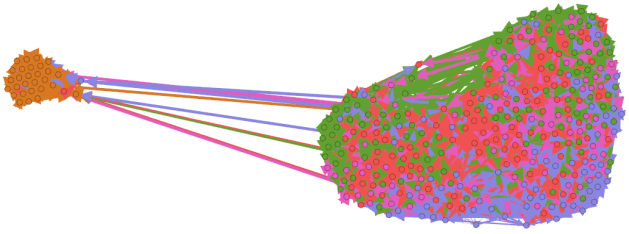

Figure 5: Similarity network of DNA Methylation

Contrasting with gene expression network, the G-CIMP subtype cluster in methylation network (Figure 5) is distinctively isolated away from all the other subtype clusters indicating that patients within this subtype cluster has methylation profiles utterly different from other patients.

Mesenchymal, Classical, Proneural and Neural subtype clusters displays that their methylation values has more overlapping between the patients compared to gene expression network, which suggests that methylation profiles are less distinct for these subtypes.

Ultimately the gene expression profiles are better at separating classical and Mesenchymal subtypes, whereas Methylation profile is more effective for isolating G-CIMP subtype.

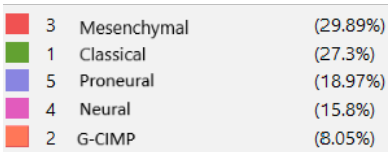

Figure 6: Subtype labels

4.2 Visualization of Embeddings

To determine if the embeddings accurately reflect the intrinsic patterns in the data, we used tSNE to visualize the embeddings in 2-D space.

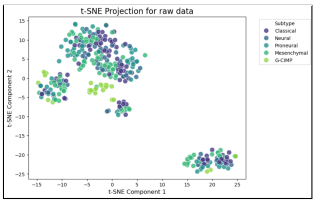

Figure 7: Raw feature

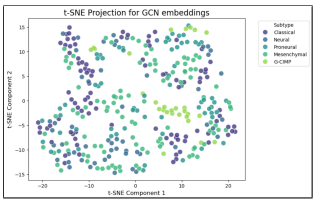

Figure 8: GCN embeddings

The tSNE plot of raw features (Figure 7) indicates that gene expression values of patients form the tight clusters, however the clusters contains intermixed subtype patients, which lacks to distinguish between patterns of different subtypes in GBM. Whereas the GCN embeddings (Figure 8) are more diffused in the tSNE plot but there is less overlapping compared to raw features, this indicates the GCN is able to learn the underlying patterns of subtypes but it couldn't form tight clusters of the subtypes, it appears to be random, but we could see subtypes like G-CIMP form short cluster expressing its varying attributes from other subtype samples.

4.3 Comparison of Performance

	SUPREME	**Random Forest**
Gene Exp	**0.818 ± 0.008**	0.771± 0.023
DNA Meth	**0.804 ± 0.023**	0.77 ± 0.025
(Gene Exp, DNA Meth)	**0.818 ± 0.014**	0.762 ± 0.026

Table 2: macro F1 score for each combination

To evaluate the efficacy of the GCN embeddings we ran a baseline method to contrast its performance with SUPREME. The baseline method Random Forest is trained on the raw data (Gene expression and DNA methylation) and SUPREME is trained on GCN embeddings along with raw features concatenated to it. We also wanted to test the SUPREME performance on each combination of the dataset, to discover the dataset's efficacy towards subtype prediction. The Table 2 display summarizes the results of each dataset combination performance of both the methods.

The SUPREME model surpasses the baseline method Random forest for all the dataset combination. It's best macro score is the combination of both gene expression and DNA methylation data modalities which includes GCN embeddings of both datasets concatenated with raw features. With GCN embeddings SUPREME model is able to produce better results with 4% difference with the baseline method for all dataset combinations. This suggests that GCN embeddings is able to capture underlying patterns of the data modalities which is associated with the subtypes. In case of single data modality combination Gene expression data exhibits better performance with 1% difference than DNA methylation, where this can be accounted in the similarity network of

gene expression data which displays distinct clusters and less overlapping compared to DNA methylation data which contained samples with overlapping gene expression values across various GBM subtypes.

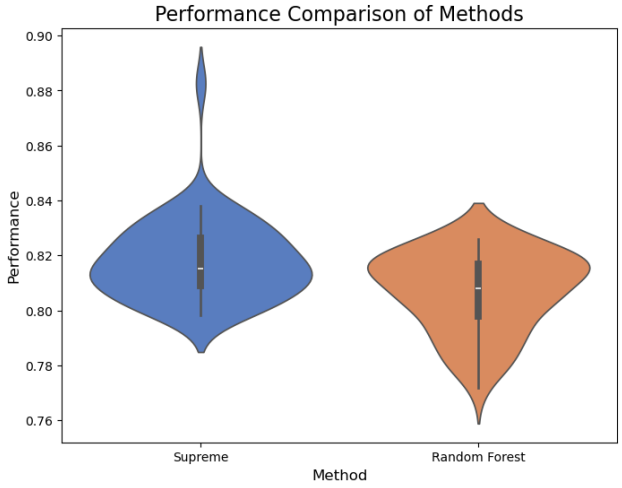

Figure 9: SUPREME vs. RF baseline (with raw features concatenated to embeddings)

In figure 9 we can observe that SUPREME's macro F1 score is higher in most of the run comparing to the baseline method Random Forest. While experimenting with different methods for ML prediction task in SUPREME framework, Random Forest produced optimal results than other methods, such as MLP and Support vector machine.

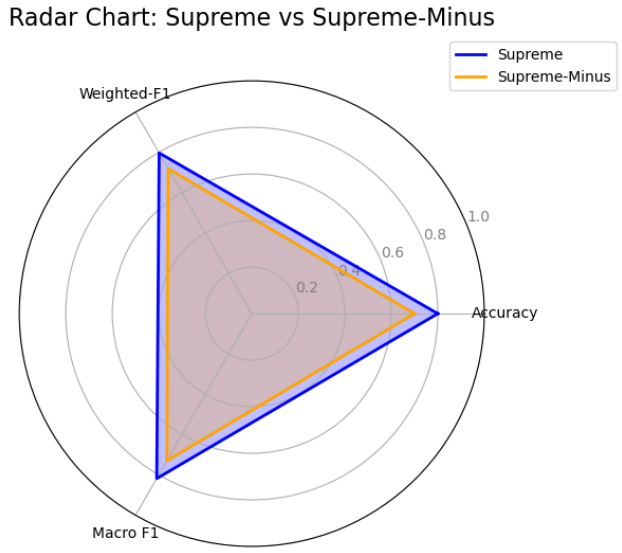

Figure 10: Supreme vs. Supreme-Minus(without feature integration)

We also did experiments to evaluate the raw features contribution towards subtype prediction. When we compared Supreme and Supreme-Minus(Supreme without feature integration) there is 20% difference in macro f1 scores and also other respective metrics, which can be observed in Figure 10. Which suggests the feature integration is very effective and stimulates the performance of the model for GBM subtype prediction.

4.4 SUPREME-GAT

We have replaced GCN with GAT in the SUPREME framework introducing attention mechanism which computes pairwise interactions explicitly, allowing complex dependencies between all features to be modeled. This is particularly useful for capturing relationships that are not spatially localized. Below tables displays the performance comparison between SUPREME-GAT and SUPREME across three metrics.

	SUPREME	**SUPREME-GAT**
Gene Exp	**0.818 ± 0.008**	0.817± 0.01
DNA Meth	**0.804 ± 0.023**	0.756 ± 0.024
(Gene Exp, DNA Meth)	**0.818 ± 0.014**	0.807 ± 0.022

Table 3: macro F1 score for each combination

	SUPREME	**SUPREME-GAT**
Gene Exp	0.8 ± 0.006	**0.843± 0.015**
DNA Meth	**0.786 ± 0.02**	0.786 ± 0.016
(Gene Exp, DNA Meth)	0.8 ± 0.013	**0.814 ± 0.019**

Table 4: accuracy for each combination

	SUPREME	**SUPREME-GAT**
Gene Exp	0.797 ± 0.007	**0.818± 0.014**
DNA Meth	**0.782 ± 0.022**	0.755 ± 0.022
(Gene Exp, DNA Meth)	0.797 ± 0.014	**0.801 ± 0.022**

Table 5: weighted F1 score for each combination

SUPREME-GAT outperformed SUPREME in accuracy and weighted F1 score and lost on macro F1 score which suggests that GAT may have focused more on enhancing the performance for dominant classes at the expense of minority classes.
We also tuned the average connectivity of the patient nodes in similarity networks with the combinations [3, 6, 11], to interpret how GAT performs across various neighborhood sizes. In [Table 6] the results indicate GAT performs well with increase in average connectivity which explains that it is able to leverage dense networks with rich pool of information. As GAT utilizes multi head attention which combined them to learn diverse patterns and interactions which makes it suitable for capturing the nuanced relationships present in high connectivity graphs

Avg Neighbor	**Gene**	**DNA meth**	**Fused**
3	0.708 ± 0.014	0.745 ± 0.018	0.716 ± 0.016
6	0.741 ± 0.01	0.738 ± 0.021	0.757 ± 0.012
11	0.744 ± 0.028	0.751 ± 0.021	0.731 ± 0.018

Table 6: SUPREME-GAT's Performance comparison for different average node connectivities .

5 Conclusion

In this project, we have utilized and enhanced SUPREME framework for Glioblastma Multiforme (GBM) subtype prediction leveraging multi omics integration through Graph Convolutional Networks(GCNs) and Graph Attention Networks (GATs). SUPREME successfully performs well with multi-omics integration which captures intricate relationships in gene expression and DNA methylation modalities to predict the GBM subtypes. By using patient similarity networks, we have shown that integrating raw multi-omic features and embeddings contributes to distinguish the subtypes, improving the performance of the model significantly compared to the traditional methods which only uses single omic type for subtype prediction.

Our study showed that GCN effectively captures the intricate biological relationships associated with GBM subtypes, outperforming traditional methods like Random Forest across different dataset combinations, notably integrating raw multiomic features to the GCN embeddings which provides complementary signals that enhances the model's predictability. Additionally our experiments with SUPREME-GAT displayed that it performs well with attention mechanism which provides considerable flexibility in centering around most relevant neighbors which outperforms SUPREME in terms of accuracy and weighted F1 score but under-performs at macro F1 score which is potentially due to its emphasis on dominant classes over minority classes.

Our experiments with different parameters of average node connectivity in similarity networks has revealed that GAT's performance improve with increased average node connectivity which signifies that it leverages rich, dense networks to capture complex dependencies in multi-omic datasets. In conclusion SUPREME framework which includes with GAT modification represents a robust approach for multi omic integration in cancer subtype prediction, which offers interesting insights into the molecular basis of GBM. This approach can be further extended to other cancers and multi-omic datasets which demonstrates precise and tailored cancer diagnosis.

References

[1] Mathieu Bastian, Sebastien Heymann, and Mathieu Jacomy. Gephi: An open source software for exploring and manipulating networks, 2009.

[2] Cameron Brennan, Roel Verhaak, Aaron McKenna, Benito Campos, Houtan Noushmehr, Sofie Salama, Siyuan Zheng, Debyani Chakravarty, J Sanborn, Samuel Berman, Rameen Beroukhim, Brady Bernard, Terrence Wu, Giannicola Genovese, Ilya Shmulevich, Jill Barnholtz-Sloan, Rahulsimham Vegesna, Sachet Shukla, and Lynda Chin. The somatic genomic landscape of glioblastoma. *Cell*, 155:462–477, 10 2013.

[3] Ziynet Nesibe Kesimoglu and Serdar Bozdag. Supreme: A cancer subtype prediction methodology integrating multiomics data using graph convolutional neural network. *bioRxiv*, 2022.

[4] Emslie D. Corcoran L. Oshlack, A. and G. K. Smyth. Normalization of boutique two-color microarrays with a high proportion of differentially expressed probes. *Genome Biology*, (7):265–273., 2007.

[5] Ricardo Ramirez, Yu-Chiao Chiu, Allen Hererra, Milad Mostavi, Joshua Ramirez, Yidong Chen, Yufei Huang, and Yu-Fang Jin. Classification of cancer types using graph convolutional neural networks. *Frontiers in Physics*, 8, 2020.

[6] Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. The graph neural network model. *IEEE Transactions on Neural Networks*, 20(1):61–80, 2009.

[7] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks, 2018.

[8] Tongxin Wang, Wei Shao, Zhi Huang, Haixu Tang, Jie Zhang, Zhengming Ding, and Kun Huang. Mogonet integrates multi-omics data using graph convolutional networks allowing patient classification and biomarker identification. *Nature Communications*, 12:3445, 06 2021.