END-TO-END INTERPRETABLE GRAPH LEARNING FOR PATIENT CLASSIFICATION

Maria Boulougouri, Mohan Vamsi Nallapareddy & Pierre Vandergheynst
LTS2 Signal Processing Laboratory, IEL, STI
École Polytechnique Fédérale de Lausanne
Lausanne 1015, Vaud, Switzerland
{maria.boulougouri,vamsi.nallapareddy,pierre.vandergheynst}@epfl.ch

ABSTRACT

Genes don't operate in a vacuum - they operate in the form of complex networks. Traditional gene expression data analysis often includes the analysis of co-expression patterns to understand these interactions; however, most machine learning methodologies don't properly account for context-dependent relationships between input features. Here, we propose a novel latent graph learning framework, titled Learnable Graph Interaction Module (LGIM), that employs a differentiable graph module to learn interactions between genes. We conduct a pilot study of our model on seven TCGA cancer datasets, where it either outperforms or performs comparably to the baseline models while learning meaningful gene representations. Conducting an interpretability analysis on the learned gene interaction graph for breast cancer, we notice that the extracted nodes and edges of higher importance correspond to being more predictive, and to known protein-protein interactions respectively. Furthermore, the clusters in the learned graph corroborate with relevant biological pathways.

1 INTRODUCTION

1.1 GENES IN CONTEXT - LEARNING CO-EXPRESSION GRAPHS

Bulk RNA sequencing is one of the most abundant types of biological data, where transcriptomic changes between conditions (e.g. healthy vs disease) are used to identify perturbation-related genes. While approaches such as differential expression analysis seek to isolate individual genes, complementary functional analysis methods try to detect consistent patterns in the transcriptomic landscape Conesa et al. (2016). The aim of this is to uncover the underlying network through which the gene functions are exerted; implying that expression alterations can spread through the links to neighboring components Barabási et al. (2011). Data-driven methods group genes with similar co-expression patterns into gene modules, with the underlying assumption that functionally-related genes have more similar expression profiles than unrelated ones Li et al. (2022). The widespread application of co-expression network analysis methods, such as WGCNA Zhang & Horvath (2005), has led to the identification of potential prognostic genes across different cancer types Yang et al. (2014); Tang et al. (2018).

1.2 GRAPH NEURAL NETWORKS FOR EXPRESSION DATA

Unlike traditional models that treat data points independently, graph neural networks (GNNs) explicitly incorporate topological information into the training process. In the case of gene expression, gene connectivity is often interpreted as a network of interactions between the gene products. While different sources have been used in the literature - ConsensusPath DB in Schulte-Sasse et al. (2019), OmniPath in Hwang et al. (2020), the Human Protein Reference Database Chereda et al. (2021), Reactome Liang et al. (2022), and, most often STRING, as in Zhuang et al. (2023), the use of signaling pathways or Protein-Protein Interaction (PPI) networks as adjacency matrices is very common in GNN models. However, these are not context-specific, and may fail to capture all gene regulatory relationships. We note in Ramirez et al. (2021), that a Graph Convolutional Neural Network (GCN)

that uses a co-expression based input graph performed better at cancer type prediction compared to the same model using a PPI based graph, which could imply that the co-expression based graphs are better at capturing complex gene-to-gene relationships. In addition to gene co-expression, matrix factorization Han et al. (2019), and topological analysis Mandal et al. (2020); Nicolau et al. (2011) methods have been used directly on the expression data to extract meaningful relations.

1.3 Learning connections for expression data

We hypothesize that learning the underlying latent graph of gene interactions would help us build better predictive models using gene expression data, in addition to uncovering novel interactions between genes. Considering that gene interactions take the form of a graph, we can employ a variety of algorithms for learning them. These algorithms can be broadly divided into three categories, attention based, dynamic graph based, and graph learning based. Attention based algorithms involve learning an attention vector pertaining to the edges and the nodes, which can be used to reconstruct the latent graph Liu et al. (2018); Zhang et al. (2018); Abu-El-Haija et al. (2018). The primary drawback of these methods is that they require an initial interaction matrix, which is often absent, or potentially quite sparse in the case of gene interactions making them difficult to implement. The dynamic graph based algorithms such as the Dynamic Graph CNNs (DGCNNs) Wang et al. (2018), and the PGC-DGCNN Tran et al. (2018) can be employed without a-priori knowledge about the connections, but these methods use the k-Nearest Neighbors (kNN) operation to design the graph which is neither optimal nor differentiable. The third category of graph learning-based algorithms design the graph in different ways such as modeling the connectivities as a learnable hyperparameter Franceschi et al. (2019), or by implementing differentiable graph pooling modules Ying et al. (2018). Recent advancements implement graph kernel neural networks Cosmo et al. (2021) that utilize local filters on the graph to learn the connections.

1.4 Learnable Gene Interaction Module

The outlined methods for graph learning have been scarcely used to study gene interaction networks, and in our study, we implement and modify the state-of-the-art Differentiable Graph Module (DGM) Kazi et al. (2023) algorithm to learn gene interactions in various cancers. The DGM algorithm was chosen because of the end-to-end nature of its training, in addition to being able to learn larger graphs more efficiently. The proposed model, titled Learnable Gene Interactions Module (LGIM), modifies the DGM to perform graph-level classification of patients for clinical metadata in different cancer datasets available in TCGA. In addition to this, the graph generation process is modified to allow for adjacency matrices with unequal node degree distribution. We give an overview of the contributions of our study below:

Novel latent graph learning for gene interactions. To the best of our knowledge, we are the first to apply a latent graph learning approach to learn gene interactions from gene expression data. We conduct a pilot study where we apply the proposed LGIM model to seven different cancer datasets, and compare it to five baseline models where it has either better or comparable performance.

Ablations to design an optimal graph learning setting. We conduct three different ablation studies to identify the best adjacency matrix initialization method, and graph message passing algorithm. For the LGIM, we test four initialization methods (Spearman, Full, Empty, and Random) with different edge probability binarization thresholds, and two message-passing algorithms (GCN, and SAGE). Additionally, for the GNN baseline models we test two initialization methods (Spearman, and PPI). Although these initialization methods have been compared in previous studies, the best approach remains inconclusive. Here, we test different methods, and provide implementation details.

Interpretability studies on the learned gene interaction graph. The proposed LGIM model learns a latent graph that represents the predicted interactions between the different genes in the input dataset. The learned connectivity matrix for the biggest cancer dataset, BRCA, was studied using standard gradient-based interpretability methods to identify genes and interactions that are considered by the model to be important. Through this analysis, we corroborate that the identified genes are more predictive in nature, and that the identified edges relate to known PPIs. Additionally, we conduct a preliminary clustering analysis which highlights that the final adjacency matrix contains genes placed in biologically meaningful groups.

The repository with the code for the data processing, model training, and interpretability analysis can be found here [1].

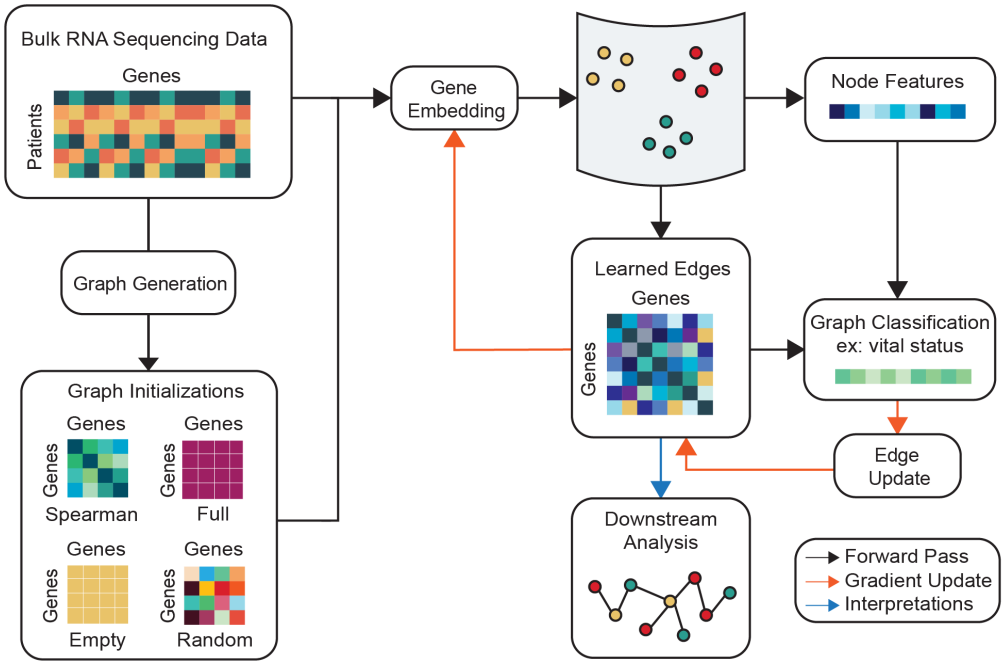

Figure 1: LGIM model overview. To generate the input graph, the genes are assigned as nodes and the gene expression values are assigned as node features. The edges are either initialized using a transformation of the input data (Spearman) or without additional information (Empty, Full, and Random). Using a convolution-based GNN, the input genes are embedded on a latent space, that is subsequently used to update the node features and learn new graph edges, which are used to build the adjacency matrix for the next training step. The model training is conducted end-to-end and the new node features and edge connections are learned from the patient classification task. The final learned graph can be used for further downstream analysis to learn more about the interactions between different genes.

2 METHODS

2.1 DATASET PROCESSING

Datasets generated by the TCGA Research Network [2] with sufficient sample sizes ($>$ 300 patients) were obtained via the Firebrowse [3] portal (accessed 21.06.2024) for the breast invasive carcinoma (BRCA), colorectal adenocarcinoma (COADREAD), glioblastoma multiforme and lower grade glioma (GBMLGG), head and neck squamous cell carcinoma (HNSC), pan-kidney carcinoma (KIPAN), lung adenocarcinoma (LUAD), and stomach and esophageal carcinoma (STES) cohorts. The clinical metadata, specifically patient vital status at the latest time of follow-up (alive or dead), and RNA-Seq data from tumor sites were filtered for patients that have both types of information available. The downloaded Illumina HiSeq RNA-Seq expression data was normalized at the gene level, and processed by the RSEM pipeline Li & Dewey (2011). Subsequently, we filtered out genes absent in less than 10% of the patients, genes with little variance across samples (threshold set at 0.1), as well as non-protein coding genes using the gene product names in the STRING database as reference Szklarczyk et al. (2023), and then the expression values were $log2(x + 1)$ transformed to

[1] https://github.com/mariaboulougouri/Gene_gr_inf/tree/main
[2] https://www.cancer.gov/tcga/
[3] http://firebrowse.org/

re-scale them. This pre-processed dataset was divided into training, and testing sets with an 80-20 stratified split, and 10% of the training set was separated randomly to compile the validation set used for training. The dataset statistics for all the seven cancers can be found in table A1.

2.2 GRAPH NEURAL NETWORKS

Graph Neural Networks (GNNs) are a variant of neural networks designed to better process graph-based inputs. A graph (G) is defined by a node set (V) consisting of n nodes, and edge set (E) consisting of m edges. The connectivity of the graph is represented by the adjacency matrix $A \in \mathbb{R}^{n \times n}$, and the node features by $X \in \mathbb{R}^{n \times d}$, where d represents the size of the node embedding. Message Passing Neural Networks (MPNNs) utilize A and X to learn new node-level features by applying a combination of aggregator (AGG) and combination (COM) functions. The node feature update after the k^{th} pass (refer to eq. (1)) is conducted by aggregating features from the neighborhood nodes, and combining them with the features of the current node. Here, $(AGG^{(k)})$, and $(COM^{(k)})$ refer to the aggregation and combination functions of the l^{th} layer, whereas x_i^k refers to the node features of the node i obtained from layer l.

$$\begin{aligned} m_i^k &= AGG^{(k)}\left(\left\{\left(x_i^{k-1}, x_j^{k-1}\right) \colon (i,j) \in E\right\}\right) \\ x_i^k &= COM^{(k)}\left(x_i^{k-1}, m_i^k\right) \end{aligned} \tag{1}$$

2.3 MODEL & BASELINES

The proposed Learnable Graph-Interaction Module (LGIM) uses the processed bulk RNA-sequencing data from each patient sample as a graph input, where the nodes are represented by the genes, and the edges are represented by their interactions. This model builds upon the algorithm presented by Kazi et al. (2023), which consists of two components: the Differentiable Graph Module (DGM), and the Diffusion Module. The DGM learns an optimal underlying latent graph from the input data for a given classification task, whereas the Diffusion Module uses the connectivity matrix learned by the DGM and the node features to obtain the new set of node features. This entire pipeline is trained in an end-to-end fashion, with the edge probability between the nodes defined by the relationship between their features. The probability that an edge exists between nodes i and j is explained in eq. (2), where t is a tunable parameter, $\Delta(.,.)$ is the euclidean distance, and f_Θ is the parametric function to obtain the node features. Two versions of this model have been presented and their differences lie in the way that the edges of the latent graph are sampled, either in a discrete or continuous fashion which results in the discrete DGM (dDGM), and continuous DGM (cDGM) models. The LGIM model builds on top of the dDGM model, and was modified to better fit to the task at hand. The Gumbel Top-K trick Kool et al. (2019) that was used for the discrete sampling of the edges was replaced by a global percentage threshold applied to the euclidian distances of the latent space to retain only the most important edges, resulting in a binary unweighted adjacency matrix. Furthermore, it was modified to conduct the classification at a graph-level instead of the node-level, thus allowing for sample-wise classification.

$$p_{ij}(\mathbf{X}; \Theta, t) = e^{-t\Delta(\hat{x}_i, \hat{x}_j)^2} = e^{-t\Delta(f_\Theta(x_i), f_\Theta(x_j))^2} \tag{2}$$

For the LGIM model to learn the edge connectivity of the input graph, one could optionally provide an initial adjacency matrix. The training was conducted with four different initial adjacency matrices, which were the zero matrix, ones matrix, random matrix, and a matrix designed from Spearman correlations between genes; these are referred to as the Empty, Full, Random, and Spearman initializations respectively. We note that the initial adjacency matrix is shared across all patients. In the case of the Spearman initialization, the correlation matrix was binarized by setting all the values between the (mean ± std * coef) to zeros, and the rest to ones. The value of "coef" is designed to be a tunable hyperparameter. The random matrix was generated in a similar fashion, where an array was populated by values randomly sampled from a normal distribution $(mean = 0.0, std = 1)$. For each comparison, the same "coef" was used to binarize the Spearman and Random initialized matrices, to ensure the same sparsity. This LGIM model was tested with the GCNConv Kipf &

Welling (2017) and SAGEConv Hamilton et al. (2017) message-passing algorithms (referred to as LGIM-GCN, and LGIM-SAGE hereafter respectively) to decide on the ideal framework.

The performance of the proposed model was compared against four simple baselines, the Logistic Regression (LogReg), Linear Support Vector Classifier (LSVC), XGBoost Chen & Guestrin (2016), and Multi-Layer Perceptron (MLP), in addition to an out-of-the-box Graph Neural Network (GNN) model. We tested the GNN benchmark with the GCN, and SAGE algorithms (referred to as GCN, and SAGE models hereafter). For the two GNN models, we tested two different input adjacency matrices, one based on the Spearman correlation, and the other designed using the PPI co-expression scores from the STRING database. All of these models were trained with a stratified k-fold ($k = 5$) pipeline, and to compare their performances we used the weighted F1 (WF1) score and the balanced accuracy (BAcc) metrics.

2.4 Interpretability Analysis

The proposed LGIM model uses information from the learned gene features, in conjunction with the learned interactions between the genes to conduct the classification task. To identify which of the genes and the interactions are the most important for the classification task in the context of different cancers, we have employed a standard gradient based interpretability technique called "Integrated Gradients" Sundararajan et al. (2017) using the implementation provided by Captum Kokhlikyan et al. (2020). Using this algorithm, we perturbed all the genes, and the interactions in the learned graph to obtain node and edge level attributions for each of the samples in the datasets. These attributions would highlight the importance that LGIM ascribes to individual genes and their interactions during training.

For the best performing model, the final adjacency matrix was combined with the edge attributions, and the resulting matrix was hierarchically clustered using the maximum cluster number as criterion. The maximum cluster number was estimated using matrix factorization on the Laplacian of the matrix; we included the minimum number of eigenvectors after which the eigenvalues plateau. Overrepresentation analysis Subramanian et al. (2005) was performed on each cluster using the GSEApy package Fang et al. (2022), with the total number of genes used for training set as background. The genesets tested were GO_Biological_Process_2021 Ashburner et al. (2000); The Gene Ontology Consortium et al. (2023), KEGG_2021_Human Kanehisa et al. (2025), Reactome_Pathways_2024 Milacic et al. (2024), WikiPathways_2024_Human Agrawal et al. (2024), Human_Phenotype_Ontology Gargano et al. (2024), MSigDB_Hallmark_2020 Liberzon et al. (2015), MSigDB_Oncogenic_Signatures Liberzon et al. (2015), Transcription_Factor_PPIs Xie et al. (2021). Adjusted p values below 0.1 were considered significant.

3 Results

3.1 LGIM initialized with prior knowledge attains comparable performance against the baselines across all cancers

The Learnable Gene Interaction Module (LGIM) was designed to use gene expression data for patient-level classification tasks. In this study, we build an initial graph considering the genes to be nodes, and the interactions between them to be the edges. We use this graph to predict the vital status of the patient across seven different cancers. We test the LGIM-GCN model with different initializations, such as full, empty, and random matrices, as well as a Spearman correlation-based matrix, whose optimal sparsity was tuned as a hyperparameter (refer table A2). We noticed that the model with the Spearman correlation-based initialization matrix consistently resulted in better performances (refer tables A3 and A4), and the final adjacency matrices from this model were among the most similar to each other across the different training folds (refer fig. 2 (H)). The model with the random initialization was just as stable across the folds, but the full and empty initialized models have reduced stability. All of these were compared with a set of random matrices as negative control. The higher performance in the models with Spearman initialization highlights the importance of using prior information in order to help LGIM-GCN to attain a greater performance.

Additionally, we tested two variants of the proposed model with the Spearman initialization, LGIM-GCN, and LGIM-SAGE, to understand which message-passing algorithm is ideal for this use-case. We noticed that LGIM-GCN has comparable performance as that of LGIM-SAGE in terms of the F1

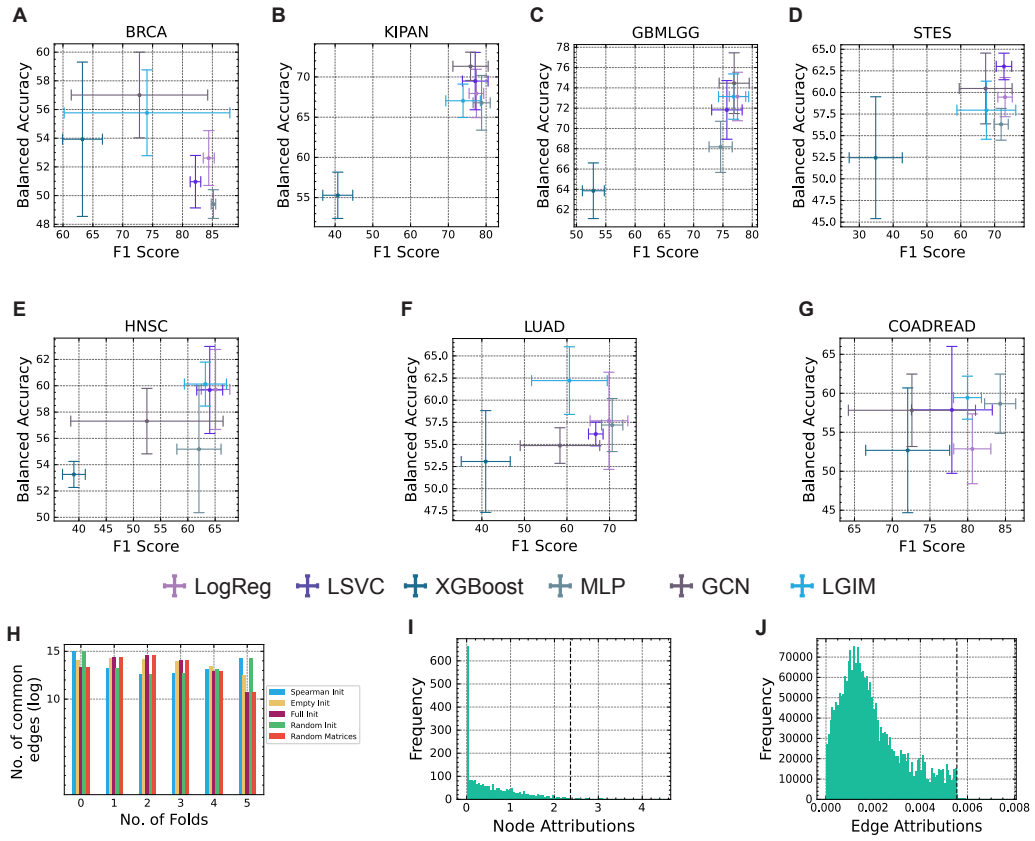

Figure 2: Model performance and interpretability. **(A-G)** Performance comparisons of LGIM and baselines for BRCA, KIPAN, GBMLGG, STES, HNSC, LUAD, and COADREAD respectively. **(H)** Comparison of the edge stability across the 5 training folds for the LGIM models with four adjacency matrix initializations (Spearman, Full, Empty, and Random) with Random matrices provided for control. **(I-J)** Distributions of integrated gradients based on graph attributions extracted for the LGIM model for all the nodes and the edges respectively.

score and the balanced accuracy, but throughout the training the LGIM-GCN model is consistently more stable, which is reflected in the standard deviations for the performances (refer tables A5 and A6). Considering these two ablation studies, we chose the proposed model to be the LGIM variant that uses the Spearman initialization, and the GCN message passing algorithm. This would hereafter be referred to simply as LGIM.

The proposed LGIM model was compared against five different baselines which were LogReg, LSVC, XGBoost, MLP, and a GNN. To choose the best GNN baseline, we compared GCN and SAGE models with a Spearman and PPI co-expression adjacency matrix. Between the GCN and SAGE models, we noticed that they have comparable performance, although the GCN is more stable during the different training folds, similar to what we notice with LGIM. Considering the two initializations, the models based on the Spearman initialization consistently attained a good performance, and were more stable across the folds (refer tables A7 and A8). Hence, we chose the GCN model with the Spearman initialization to be the GNN baseline; we further tune this model to adjust the Spearman correlation threshold (refer table A2). Among the six models trained in this study, we noticed that LGIM outperforms all the baselines in terms of the balanced accuracy for HNSC, LUAD, and COADREAD, whereas on BRCA, KIPAN, GBMLGG, and STES it attains a comparable performance (refer table A10 and fig. 2 (A-G)). Also in terms of the F1 score, LGIM either outperforms some baselines or reaches a comparable performance on all seven cancer datasets (refer table A9 and fig. 2 (A-G)).

3.2 LGIM LEARNS POTENTIAL GENE INTERACTION NETWORKS PRESENT IN BREAST CANCER

We conducted an interpretability study on the BRCA dataset, considering that this is the biggest among the seven cancer datasets. We chose the best performing LGIM model (WF1: 74.06 ± 13.87, & BAcc: 55.77 ± 2.99) and studied the learned latent graph of gene connections from fold 1 of its training, as it was the best performing one. We employed an "Integrated Gradients"-based graph explanation method to extract the node and edge attributions. These indicated the importance that the model assigns to each of the genes and the connections during training. Applying the knee locator method Satopaa et al. (2011), we identified a cutoff that extracted 77 of the most important genes from the node attribution distribution (refer fig. 2 (I)). Similarly, we extracted 77 of the least important genes using the node attribution distribution. In order to understand, if the genes with the higher attribution values are more predictive, we implemented a simple LogReg model that predicted the vital status in BRCA using only the expression values of the genes with high or low attributions. We noticed that the LogReg model using only the high attribution genes had about 6% higher performance in terms of balanced accuracy compared to the LogReg model that used only the low attribution genes (BAcc: 55.28 ± 4.13 vs. BAcc: 49.22 ± 3.3). In order to corroborate this even further, we obtained the mutual information values for all genes using the gene expression data and vital status metadata as the label, we then binarized the scores into high and low mutual information. We conducted a Fisher's exact test which confirmed our hypothesis that high attribution genes also have high mutual information (p-value: 0.015). We corroborated our findings with previous studies conducted on the TCGA BRCA dataset Center (2016) which identify differentially expressed marker genes for seven expression-based subtypes that significantly correlate with clinical phenotypes. 75 out of the 77 high attribution genes, and none of the low attribution genes were present in this set of marker genes. These analyses highlighted that LGIM is capable of picking up on genes that are more relevant to clinical metadata.

Furthermore, we analyzed the edge attributions in conjunction with the PPI co-expression values to study if the genes connected by our model are known to have similar co-expression profiles, and therefore a predicted interaction of their gene products. We conducted two tests, one on the global level and the other only on the highest and lowest attribution edges. The highest attribution edges were chosen in a similar fashion as we did for the nodes using the knee locator method, and were 6612 in number (refer fig. 2 (J)). The 6612 lowest attribution edges were then also isolated. We observed that there isn't a clear correlation on the global level between the edge attributions, and the corresponding PPI co-expression values, however, we noticed that there is a mild difference between the frequency of the presence of co-expression values for the highest and the lowest attribution edges (198 vs. 168). These tests indicated that LGIM may be capable of assigning relative importance values in accordance with known interactions for the genes.

This final connectivity matrix of the best performing model (fold 1, WF1: 86.4, BAcc: 51.62) was supplemented with the edge attributions. The Laplacian of this matrix was calculated and the optimal number of clusters was estimated at 14, after which the eigenvalues of the Laplacian eigenvectors plateau. After hierarchical clustering of this matrix, we performed overrepresentation analysis on each cluster to identify associated functionalities. We find cluster 7 to be significantly associated with transcriptomic signatures related to immune response, both in terms of B cell receptor signaling pathways (KEGG, Reactome, WikiPathways) and phenotypic aberrations of the lymphoid lineage (Human Phenotype Ontology). Interestingly, it is also significantly associated with the transcription factors TP63 Bankhead et al. (2020) and EP300 Gronkowska & Robaszkiewicz (2024) which have been associated with immune regulatory gene expression in breast cancer, as well as with other transcription factors that have separate known roles in breast cancer and immunity, namely ELF1 Gerloff et al. (2011); Seifert et al. (2019), CHD1 Zhao et al. (2017; 2020), and JUNB Ren et al. (2023). Additional significant associations are found with transcription factors with known roles in tumor prognosis ZNF263 Zhou et al. (2018), HNF4G Chen et al. (2022), TAF7 Zhang et al. (2025), and IRF3 Tian et al. (2020). Other examples of clusters with consistent biological patterns include cluster 8, which contains genes associated with Notch signaling and lipid metabolism (Hallmarks, Reactome), and cluster 12, which contains genes associated with apoptosis (Hallmarks, GO Biological Processes, KEGG). The findings for clusters 7, 8, and 12 have been mentioned in detail in tables A11 to A13 respectively.

4 DISCUSSION

In this study, we present LGIM, which is to the best of our knowledge, the first end-to-end trained latent graph learning algorithm to learn gene interactions from gene expression data. The aim of this study was to set up a pipeline that goes beyond static gene connectivity matrices and learns more about gene networks as a whole. We compare our proposed model against 5 baselines, including simple models and more complex graph neural networks on seven different cancer datasets. We either outperform the baselines or attain a comparable performance on all the benchmarks. We conduct an optimization study to discover the optimal adjacency matrix for initialization and message-passing algorithm to employ, in an effort to standardize the application of graph learning algorithms on gene expression data.

The learned graph from the LGIM model can be studied to learn more about specific genes, their interactions, and their role. We conduct one such analysis on the learned graph from the BRCA vital status classification task, where we used a gradient based interpretability method to extract important nodes and edges. The important nodes correlated strongly with having higher predictive capabilities, higher mutual information, and were indicated to be highly differentially expressed. Additionally, the learned edges seem to overlap with known PPIs. In terms of the learned graph, we identified gene clusters that seem to reflect biologically meaningful processes in breast cancer. These findings highlight the exploratory potential of the learned gene connectivity matrix to discover more about the task at hand.

The LGIM model was trained on several datasets obtained from the TCGA database, with sample sizes ranging from 377 (COADREAD) to $1,093$ (BRCA). This small number of patient samples in each of the datasets presents a wall in terms of the complexity of the models that can be employed. This results in higher instability during the training process, which can be reflected in the standard deviations for the performance of all the trained models. Further approaches could potentially focus on designing more parameter-efficient models that can work robustly in a low data setting. These could also be augmented with data from other modalities such as epigenetics, and genetics. Although, a potential challenge for such a task would be the fact that not all patients would be associated with all modalities. Overcoming these limitations could help us learn a more reliable latent graph of gene interactions, even for cancers where the data is sparse. Future studies could include the expansion of the datasets to include additional clinical metadata, such as neoplasm for binary classification, and pathological staging for multi-class classification.

Overall, we exploit the potential of gene expression data and propose a standardized approach to learn novel gene interactions for different cancers. Our preliminary analyses of the model findings indicate its ability to pick up on essential information regarding the breast cancer dataset. Still, we acknowledge that further validation on the clinical implications of this work, including exploration of potential novel biomarkers compared to current approaches in the field, should be explored in future work. Nonetheless, we believe that the proposed model is, through learning of the latent graph, offering a previously-unexplored approach to provide a holistic view of the biological processes in different tumors that drive each clinical outcome.

5 MEANINGFULNESS STATEMENT

Genes in our body interact in complex, context-specific networks, which have been widely studied as meaningful representations of the molecular mechanisms behind our phenotypic variations, such as predisposition to disease, or response to different therapeutic treatments. In this study, we propose a novel interpretable graph learning pipeline, titled Learnable Gene-Interaction Module (LGIM), designed to learn such networks. We report preliminary results on seven different cancer types, and illustrate that we are able to extract relevant biological patterns related to breast carcinoma, that we corroborate with findings from previous studies.

REFERENCES

Sami Abu-El-Haija, Bryan Perozzi, Rami Al-Rfou, and Alexander A Alemi. Watch your step: Learning node embeddings via graph attention. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Ad-*

vances in Neural Information Processing Systems, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/8a94ecfa54dcb88a2fa993bfa6388f9e-Paper.pdf.

Ayushi Agrawal, Hasan Balcı, Kristina Hanspers, Susan L Coort, Marvin Martens, Denise N Slenter, Friederike Ehrhart, Daniela Digles, Andra Waagmeester, Isabel Wassink, Tooba Abbassi-Daloii, Elisson N Lopes, Aishwarya Iyer, Javier Millán Acosta, Lars G Willighagen, Kozo Nishida, Anders Riutta, Helena Basaric, Chris T Evelo, Egon L Willighagen, Martina Kutmon, and Alexander R Pico. WikiPathways 2024: Next generation pathway database. *Nucleic Acids Research*, 52 (D1):D679–D689, January 2024. ISSN 0305-1048. doi: 10.1093/nar/gkad960.

Michael Ashburner, Catherine A. Ball, Judith A. Blake, David Botstein, Heather Butler, J. Michael Cherry, Allan P. Davis, Kara Dolinski, Selina S. Dwight, Janan T. Eppig, Midori A. Harris, David P. Hill, Laurie Issel-Tarver, Andrew Kasarskis, Suzanna Lewis, John C. Matese, Joel E. Richardson, Martin Ringwald, Gerald M. Rubin, and Gavin Sherlock. Gene Ontology: Tool for the unification of biology. *Nature Genetics*, 25(1):25–29, May 2000. ISSN 1546-1718. doi: 10.1038/75556.

Armand Bankhead, Thomas McMaster, Yin Wang, Philip S. Boonstra, and Phillip L. Palmbos. TP63 isoform expression is linked with distinct clinical outcomes in cancer. *EBioMedicine*, 51:102561, January 2020. ISSN 2352-3964. doi: 10.1016/j.ebiom.2019.11.022.

Albert-László Barabási, Natali Gulbahce, and Joseph Loscalzo. Network medicine: a network-based approach to human disease. *Nature Reviews Genetics*, 12(1):56–68, January 2011. ISSN 1471-0064. doi: 10.1038/nrg2918. URL https://www.nature.com/articles/nrg2918. Publisher: Nature Publishing Group.

Broad Institute TCGA Genome Data Analysis Center. *Clustering of mRNAseq gene expression: consensus NMF*. 2016. doi: 10.7908/C1RB73ZS. URL http://gdac.broadinstitute.org/runs/analyses__2016_01_28/reports/cancer/BRCA-TP/mRNAseq_Clustering_CNMF/nozzle.html.

Lu Chen, Huanying Shi, Xinhai Wang, Tianxiao Wang, Yingjie Wang, Zimei Wu, Wenxin Zhang, Haifei Chen, Mingkang Zhong, Xiang Mao, Xiaojin Shi, and Qunyi Li. Hepatocyte nuclear factor 4 gamma (HNF4G) is correlated with poor prognosis and promotes tumor cell growth by inhibiting caspase-dependent intrinsic apoptosis in colorectal cancer. *European Journal of Pharmacology*, 916:174727, February 2022. ISSN 1879-0712. doi: 10.1016/j.ejphar.2021.174727.

Tianqi Chen and Carlos Guestrin. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794, San Francisco California USA, August 2016. ACM. ISBN 978-1-4503-4232-2. doi: 10.1145/2939672.2939785.

Hryhorii Chereda, Annalen Bleckmann, Kerstin Menck, Júlia Perera-Bel, Philip Stegmaier, Florian Auer, Frank Kramer, Andreas Leha, and Tim Beißbarth. Explaining decisions of graph convolutional neural networks: Patient-specific molecular subnetworks responsible for metastasis prediction in breast cancer. *Genome Medicine*, 13(1):42, March 2021. ISSN 1756-994X. doi: 10.1186/s13073-021-00845-7.

Ana Conesa, Pedro Madrigal, Sonia Tarazona, David Gomez-Cabrero, Alejandra Cervera, Andrew McPherson, Michał Wojciech Szcześniak, Daniel J. Gaffney, Laura L. Elo, Xuegong Zhang, and Ali Mortazavi. A survey of best practices for RNA-seq data analysis. *Genome Biology*, 17(1):13, January 2016. ISSN 1474-760X. doi: 10.1186/s13059-016-0881-8.

Luca Cosmo, Giorgia Minello, Alessandro Bicciato, Michael M. Bronstein, Emanuele Rodolà, Luca Rossi, and Andrea Torsello. Graph kernel neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–14, 2021. ISSN 2162-2388. doi: 10.1109/tnnls.2024.3400850. URL http://dx.doi.org/10.1109/TNNLS.2024.3400850.

Zhuoqing Fang, Xinyuan Liu, and Gary Peltz. Gseapy: a comprehensive package for performing gene set enrichment analysis in python. *Bioinformatics*, 39(1):btac757, 11 2022. ISSN 1367-4811. doi: 10.1093/bioinformatics/btac757. URL https://doi.org/10.1093/bioinformatics/btac757.

Luca Franceschi, Mathias Niepert, Massimiliano Pontil, and Xiao He. Learning discrete structures for graph neural networks. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 1972–1982. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/franceschi19a.html.

Michael A. Gargano, Nicolas Matentzoglu, Ben Coleman, Eunice B. Addo-Lartey, Anna V. Anagnostopoulos, Joel Anderton, Paul Avillach, Anita M. Bagley, Eduard Bakštein, James P. Balhoff, Gareth Baynam, Susan M. Bello, Michael Berk, Holli Bertram, Somer Bishop, Hannah Blau, David F. Bodenstein, Pablo Botas, Kaan Boztug, Jolana Čady, Tiffany J. Callahan, Rhiannon Cameron, Seth J. Carbon, Francisco Castellanos, J. Harry Caufield, Lauren E. Chan, Christopher G. Chute, Jaime Cruz-Rojo, Noémi Dahan-Oliel, Jon R. Davids, Maud de Dieuleveult, Vinicius de Souza, Bert B. A. de Vries, Esther de Vries, J. Raymond DePaulo, Beata Derfalvi, Ferdinand Dhombres, Claudia Diaz-Byrd, Alexander J. M. Dingemans, Bruno Donadille, Michael Duyzend, Reem Elfeky, Shahim Essaid, Carolina Fabrizzi, Giovanna Fico, Helen V. Firth, Yun Freudenberg-Hua, Janice M. Fullerton, Davera L. Gabriel, Kimberly Gilmour, Jessica Giordano, Fernando S. Goes, Rachel Gore Moses, Ian Green, Matthias Griese, Tudor Groza, Weihong Gu, Julia Guthrie, Benjamin Gyori, Ada Hamosh, Marc Hanauer, Kateřina Hanušová, Yongqun Oliver He, Harshad Hegde, Ingo Helbig, Kateřina Holasová, Charles Tapley Hoyt, Shangzhi Huang, Eric Hurwitz, Julius O. B. Jacobsen, Xiaofeng Jiang, Lisa Joseph, Kamyar Keramatian, Bryan King, Katrin Knoflach, David A. Koolen, Megan L. Kraus, Carlo Kroll, Maaike Kusters, Markus S. Ladewig, David Lagorce, Meng-Chuan Lai, Pablo Lapunzina, Bryan Laraway, David Lewis-Smith, Xiarong Li, Caterina Lucano, Marzieh Majd, Mary L. Marazita, Victor Martinez-Glez, Toby H. McHenry, Melvin G. McInnis, Julie A. McMurry, Michaela Mihulová, Caitlin E. Millett, Philip B. Mitchell, Veronika Moslerová, Kenji Narutomi, Shahrzad Nematollahi, Julian Nevado, Andrew A. Nierenberg, Nikola Novák Čajbiková, John I. Nurnberger, Soichi Ogishima, Daniel Olson, Abigail Ortiz, Harry Pachajoa, Guiomar Perez de Nanclares, Amy Peters, Tim Putman, Christina K. Rapp, Ana Rath, Justin Reese, Lauren Rekerle, Angharad M. Roberts, Suzy Roy, Stephan J. Sanders, Catharina Schuetz, Eva C. Schulte, Thomas G. Schulze, Martin Schwarz, Katie Scott, Dominik Seelow, Berthold Seitz, Yiping Shen, Morgan N. Similuk, Eric S. Simon, Balwinder Singh, Damian Smedley, Cynthia L. Smith, Jake T. Smolinsky, Sarah Sperry, Elizabeth Stafford, Ray Stefancsik, Robin Steinhaus, Rebecca Strawbridge, Jagadish Chandrabose Sundaramurthi, Polina Talapova, Jair A. Tenorio Castano, Pavel Tesner, Rhys H. Thomas, Audrey Thurm, Marek Turnovec, Marielle E. van Gijn, Nicole A. Vasilevsky, Markéta Vlčková, Anita Walden, Kai Wang, Ron Wapner, James S. Ware, Addo A. Wiafe, Samuel A. Wiafe, Lisa D. Wiggins, Andrew E. Williams, Chen Wu, Margot J. Wyrwoll, Hui Xiong, Nefize Yalin, Yasunori Yamamoto, Lakshmi N. Yatham, Anastasia K. Yocum, Allan H. Young, Zafer Yüksel, Peter P. Zandi, Andreas Zankl, Ignacio Zarante, Miroslav Zvolský, Sabrina Toro, Leigh C. Carmody, Nomi L. Harris, Monica C. Munoz-Torres, Daniel Danis, Christopher J. Mungall, Sebastian Köhler, Melissa A. Haendel, and Peter N. Robinson. The Human Phenotype Ontology in 2024: Phenotypes around the world. *Nucleic Acids Research*, 52(D1):D1333–D1346, January 2024. ISSN 1362-4962. doi: 10.1093/nar/gkad1005.

Alice Gerloff, Angela Dittmer, Ilka Oerlecke, Hans-Jürgen Holzhausen, and Jürgen Dittmer. Protein expression of the Ets transcription factor Elf-1 in breast cancer cells is negatively correlated with histological grading, but not with clinical outcome. *Oncology Reports*, 26(5):1121–1125, November 2011. ISSN 1021-335X. doi: 10.3892/or.2011.1409.

Karolina Gronkowska and Agnieszka Robaszkiewicz. Genetic dysregulation of EP300 in cancers in light of cancer epigenome control – targeting of p300-proficient and -deficient cancers. *Molecular Therapy: Oncology*, 32(4):200871, December 2024. ISSN 2950-3299. doi: 10.1016/j.omton.2024.200871.

Will Hamilton, Zhitao Ying, and Jure Leskovec. Inductive representation learning on large graphs. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/5dd9db5e033da9c6fb5ba83c7a7ebea9-Paper.pdf.

Peng Han, Peng Yang, Peilin Zhao, Shuo Shang, Yong Liu, Jiayu Zhou, Xin Gao, and Panos Kalnis. GCN-MF: Disease-Gene Association Identification By Graph Convolutional Networks and

Matrix Factorization. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 705–713, Anchorage AK USA, July 2019. ACM. ISBN 978-1-4503-6201-6. doi: 10.1145/3292500.3330912.

Doyeong Hwang, Minji Jeon, and Jaewoo Kang. A Drug-Induced Liver Injury Prediction Model using Transcriptional Response Data with Graph Neural Network. In *2020 IEEE International Conference on Big Data and Smart Computing (BigComp)*, pp. 323–329, February 2020. doi: 10.1109/BigComp48618.2020.00-54.

Minoru Kanehisa, Miho Furumichi, Yoko Sato, Yuriko Matsuura, and Mari Ishiguro-Watanabe. KEGG: Biological systems database as a model of the real world. *Nucleic Acids Research*, 53 (D1):D672–D677, January 2025. ISSN 1362-4962. doi: 10.1093/nar/gkae909.

Anees Kazi, Luca Cosmo, Seyed-Ahmad Ahmadi, Nassir Navab, and Michael M. Bronstein. Differentiable graph module (dgm) for graph convolutional networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(2):1606–1617, 2023. doi: 10.1109/TPAMI.2022.3170249.

Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=SJU4ayYgl.

Narine Kokhlikyan, Vivek Miglani, Miguel Martin, Edward Wang, Bilal Alsallakh, Jonathan Reynolds, Alexander Melnikov, Natalia Kliushkina, Carlos Araya, Siqi Yan, and Orion Reblitz-Richardson. Captum: A unified and generic model interpretability library for pytorch, 2020.

Wouter Kool, Herke Van Hoof, and Max Welling. Stochastic beams and where to find them: The Gumbel-top-k trick for sampling sequences without replacement. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3499–3508. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/kool19a.html.

Bo Li and Colin N. Dewey. RSEM: Accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*, 12(1):323, August 2011. ISSN 1471-2105. doi: 10.1186/1471-2105-12-323.

Michelle M. Li, Kexin Huang, and Marinka Zitnik. Graph Representation Learning in Biomedicine and Healthcare. *Nature biomedical engineering*, 6(12):1353, October 2022. doi: 10.1038/s41551-022-00942-x.

Bilin Liang, Haifan Gong, Lu Lu, and Jie Xu. Risk stratification and pathway analysis based on graph neural network and interpretable algorithm. *BMC Bioinformatics*, 23(1):394, September 2022. ISSN 1471-2105. doi: 10.1186/s12859-022-04950-1.

Arthur Liberzon, Chet Birger, Helga Thorvaldsdóttir, Mahmoud Ghandi, Jill P. Mesirov, and Pablo Tamayo. The Molecular Signatures Database Hallmark Gene Set Collection. *Cell Systems*, 1(6): 417–425, December 2015. ISSN 2405-4712. doi: 10.1016/j.cels.2015.12.004.

Ziqi Liu, Chaochao Chen, Longfei Li, Jun Zhou, Xiaolong Li, and Le Song. Geniepath: Graph neural networks with adaptive receptive paths. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33, 02 2018. doi: 10.1609/aaai.v33i01.33014424.

Sayan Mandal, Aldo Guzmán-Sáenz, Niina Haiminen, Saugata Basu, and Laxmi Parida. A Topological Data Analysis Approach on Predicting Phenotypes from Gene Expression Data. *Algorithms for Computational Biology*, 12099:178–187, February 2020. doi: 10.1007/978-3-030-42266-0_14.

Marija Milacic, Deidre Beavers, Patrick Conley, Chuqiao Gong, Marc Gillespie, Johannes Griss, Robin Haw, Bijay Jassal, Lisa Matthews, Bruce May, Robert Petryszak, Eliot Ragueneau, Karen Rothfels, Cristoffer Sevilla, Veronica Shamovsky, Ralf Stephan, Krishna Tiwari, Thawfeek Varusai, Joel Weiser, Adam Wright, Guanming Wu, Lincoln Stein, Henning Hermjakob, and Peter D'Eustachio. The Reactome Pathway Knowledgebase 2024. *Nucleic Acids Research*, 52(D1): D672–D678, January 2024. ISSN 0305-1048. doi: 10.1093/nar/gkad1025.

Monica Nicolau, Arnold J. Levine, and Gunnar Carlsson. Topology based data analysis identifies a subgroup of breast cancers with a unique mutational profile and excellent survival. *Proceedings of the National Academy of Sciences*, 108(17):7265–7270, April 2011. doi: 10.1073/pnas.1102826108.

Ricardo Ramirez, Yu-Chiao Chiu, SongYao Zhang, Joshua Ramirez, Yidong Chen, Yufei Huang, and Yu-Fang Jin. Prediction and Interpretation of Cancer Survival Using Graph Convolution Neural Networks. *Methods (San Diego, Calif.)*, 192:120–130, August 2021. ISSN 1046-2023. doi: 10.1016/j.ymeth.2021.01.004.

Fu-jia Ren, Xiao-yu Cai, Yao Yao, and Guo-ying Fang. JunB: A paradigm for Jun family in immune response and cancer. *Frontiers in Cellular and Infection Microbiology*, 13:1222265, September 2023. ISSN 2235-2988. doi: 10.3389/fcimb.2023.1222265.

Ville Satopaa, Jeannie Albrecht, David Irwin, and Barath Raghavan. Finding a "kneedle" in a haystack: Detecting knee points in system behavior. In *31st International Conference on Distributed Computing Systems Workshops*, pp. 166–171, 2011. doi: 10.1109/ICDCSW.2011.20.

Roman Schulte-Sasse, Stefan Budach, Denes Hnisz, and Annalisa Marsico. Graph Convolutional Networks Improve the Prediction of Cancer Driver Genes. In Igor V. Tetko, Věra Kůrková, Pavel Karpov, and Fabian Theis (eds.), *Artificial Neural Networks and Machine Learning – ICANN 2019: Workshop and Special Sessions*, pp. 658–668, Cham, 2019. Springer International Publishing. ISBN 978-3-030-30493-5. doi: 10.1007/978-3-030-30493-5_60.

Leon Louis Seifert, Clara Si, Debjani Saha, Mohammad Sadic, Maren de Vries, Sarah Ballentine, Aaron Briley, Guojun Wang, Ana M. Valero-Jimenez, Adil Mohamed, Uwe Schaefer, Hong M. Moulton, Adolfo García-Sastre, Shashank Tripathi, Brad R. Rosenberg, and Meike Dittmann. The ETS transcription factor ELF1 regulates a broadly antiviral program distinct from the type I interferon response. *PLoS Pathogens*, 15(11):e1007634, November 2019. ISSN 1553-7366. doi: 10.1371/journal.ppat.1007634.

Aravind Subramanian, Pablo Tamayo, Vamsi K. Mootha, Sayan Mukherjee, Benjamin L. Ebert, Michael A. Gillette, Amanda Paulovich, Scott L. Pomeroy, Todd R. Golub, Eric S. Lander, and Jill P. Mesirov. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102(43): 15545–15550, October 2005. doi: 10.1073/pnas.0506580102.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, pp. 3319–3328. JMLR.org, 2017.

Damian Szklarczyk, Rebecca Kirsch, Mikaela Koutrouli, Katerina Nastou, Farrokh Mehryary, Radja Hachilif, Annika L. Gable, Tao Fang, Nadezhda T. Doncheva, Sampo Pyysalo, Peer Bork, Lars J. Jensen, and Christian von Mering. The STRING database in 2023: Protein-protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic Acids Research*, 51(D1):D638–D646, January 2023. ISSN 1362-4962. doi: 10.1093/nar/gkac1000.

Jianing Tang, Deguang Kong, Qiuxia Cui, Kun Wang, Dan Zhang, Yan Gong, and Gaosong Wu. Prognostic Genes of Breast Cancer Identified by Gene Co-expression Network Analysis. *Frontiers in Oncology*, 8, September 2018. ISSN 2234-943X. doi: 10.3389/fonc.2018.00374.

The Gene Ontology Consortium, Suzi A Aleksander, James Balhoff, Seth Carbon, J Michael Cherry, Harold J Drabkin, Dustin Ebert, Marc Feuermann, Pascale Gaudet, Nomi L Harris, David P Hill, Raymond Lee, Huaiyu Mi, Sierra Moxon, Christopher J Mungall, Anushya Muruganugan, Tremayne Mushayahama, Paul W Sternberg, Paul D Thomas, Kimberly Van Auken, Jolene Ramsey, Deborah A Siegele, Rex L Chisholm, Petra Fey, Maria Cristina Aspromonte, Maria Victoria Nugnes, Federica Quaglia, Silvio Tosatto, Michelle Giglio, Suvarna Nadendla, Giulia Antonazzo, Helen Attrill, Gil dos Santos, Steven Marygold, Victor Strelets, Christopher J Tabone, Jim Thurmond, Pinglei Zhou, Saadullah H Ahmed, Praoparn Asanitthong, Diana Luna Buitrago, Meltem N Erdol, Matthew C Gage, Mohamed Ali Kadhum, Kan Yan Chloe Li, Miao Long, Aleksandra Michalak, Angeline Pesala, Armalya Pritazahra, Shirin C C Saverimuttu, Renzhi Su, Kate E Thurlow, Ruth C Lovering, Colin Logie, Snezhana Oliferenko, Judith Blake, Karen Christie,

Lori Corbani, Mary E Dolan, Harold J Drabkin, David P Hill, Li Ni, Dmitry Sitnikov, Cynthia Smith, Alayne Cuzick, James Seager, Laurel Cooper, Justin Elser, Pankaj Jaiswal, Parul Gupta, Pankaj Jaiswal, Sushma Naithani, Manuel Lera-Ramirez, Kim Rutherford, Valerie Wood, Jeffrey L De Pons, Melinda R Dwinell, G Thomas Hayman, Mary L Kaldunski, Anne E Kwitek, Stanley J F Laulederkind, Marek A Tutaj, Mahima Vedi, Shur-Jen Wang, Peter D'Eustachio, Lucila Aimo, Kristian Axelsen, Alan Bridge, Nevila Hyka-Nouspikel, Anne Morgat, Suzi A Aleksander, J Michael Cherry, Stacia R Engel, Kalpana Karra, Stuart R Miyasato, Robert S Nash, Marek S Skrzypek, Shuai Weng, Edith D Wong, Erika Bakker, Tanya Z Berardini, Leonore Reiser, Andrea Auchincloss, Kristian Axelsen, Ghislaine Argoud-Puy, Marie-Claude Blatter, Emmanuel Boutet, Lionel Breuza, Alan Bridge, Cristina Casals-Casas, Elisabeth Coudert, Anne Estreicher, Maria Livia Famiglietti, Marc Feuermann, Arnaud Gos, Nadine Gruaz-Gumowski, Chantal Hulo, Nevila Hyka-Nouspikel, Florence Jungo, Philippe Le Mercier, Damien Lieberherr, Patrick Masson, Anne Morgat, Ivo Pedruzzi, Lucille Pourcel, Sylvain Poux, Catherine Rivoire, Shyamala Sundaram, Alex Bateman, Emily Bowler-Barnett, Hema Bye-A-Jee, Paul Denny, Alexandr Ignatchenko, Rizwan Ishtiaq, Antonia Lock, Yvonne Lussi, Michele Magrane, Maria J Martin, Sandra Orchard, Pedro Raposo, Elena Speretta, Nidhi Tyagi, Kate Warner, Rossana Zaru, Alexander D Diehl, Raymond Lee, Juancarlos Chan, Stavros Diamantakis, Daniela Raciti, Magdalena Zarowiecki, Malcolm Fisher, Christina James-Zorn, Virgilio Ponferrada, Aaron Zorn, Sridhar Ramachandran, Leyla Ruzicka, and Monte Westerfield. The Gene Ontology knowledgebase in 2023. *Genetics*, 224(1):iyad031, May 2023. ISSN 1943-2631. doi: 10.1093/genetics/iyad031.

Miao Tian, Xiumei Wang, Jihong Sun, Wenlong Lin, Lumin Chen, Shengduo Liu, Ximei Wu, Liyun Shi, Pinglong Xu, Xiujun Cai, and Xiaojian Wang. IRF3 prevents colorectal tumorigenesis via inhibiting the nuclear translocation of β-catenin. *Nature Communications*, 11(1):5762, November 2020. ISSN 2041-1723. doi: 10.1038/s41467-020-19627-7.

Dinh Van Tran, Nicoló Navarin, and Alessandro Sperduti. On filter size in graph convolutional networks. *2018 IEEE Symposium Series on Computational Intelligence (SSCI)*, pp. 1534–1541, 2018. URL https://api.semanticscholar.org/CorpusID:53749157.

Yue Wang, Yongbin Sun, Ziwei Liu, Sanjay Sarma, Michael Bronstein, and Justin Solomon. Dynamic graph cnn for learning on point clouds. *ACM Transactions on Graphics*, 38, 01 2018. doi: 10.1145/3326362.

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*, 1(3):e90, 2021. ISSN 2691-1299. doi: 10.1002/cpz1.90.

Yang Yang, Leng Han, Yuan Yuan, Jun Li, Nainan Hei, and Han Liang. Gene co-expression network analysis reveals common system-level properties of prognostic genes across cancer types. *Nature communications*, 5:3231, 2014. ISSN 2041-1723. doi: 10.1038/ncomms4231.

Zhitao Ying, Jiaxuan You, Christopher Morris, Xiang Ren, Will Hamilton, and Jure Leskovec. Hierarchical graph representation learning with differentiable pooling. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/e77dbaf6759253c7c6d0efc5690369c7-Paper.pdf.

Bin Zhang and Steve Horvath. A general framework for weighted gene co-expression network analysis. *Statistical Applications in Genetics and Molecular Biology*, 4:Article17, 2005. ISSN 1544-6115. doi: 10.2202/1544-6115.1128.

Jiani Zhang, Xingjian Shi, Junyuan Xie, Hao Ma, Irwin King, and Dit-Yan Yeung. Gaan: Gated attention networks for learning on large and spatiotemporal graphs, 2018. URL https://arxiv.org/abs/1803.07294.

Wanjun Zhang, Jun Wang, Hanning Li, Xue Zhang, Dunjie Yao, Huimin Zhang, Xinhong Zhou, Jiaqi Nie, Tongxing Lai, Haichuan Zhu, Yiping Gong, Yoshimasa Tanaka, Xingrui Li, Xinghua Liao, and Li Su. TAF7 Directly Targets SAA1 to Enhance Triple-Negative Breast Cancer Metastasis via Phosphorylating E-Cadherin and N-Cadherin. *iScience*, 0(0), February 2025. ISSN 2589-0042. doi: 10.1016/j.isci.2025.111989.

Di Zhao, Xin Lu, Guocan Wang, Zhengdao Lan, Wenting Liao, Jun Li, Xin Liang, Jasper Robin Chen, Sagar Shah, Xiaoying Shang, Ming Tang, Pingna Deng, Prasenjit Dey, Deepavali Chakravarti, Peiwen Chen, Denise J. Spring, Nora M. Navone, Patricia Troncoso, Jianhua Zhang, Y. Alan Wang, and Ronald A. DePinho. Synthetic essentiality of chromatin remodelling factor CHD1 in PTEN-deficient cancer. *Nature*, 542(7642):484–488, February 2017. ISSN 1476-4687. doi: 10.1038/nature21357.

Di Zhao, Li Cai, Xin Lu, Xin Liang, Jiexi Li, Peiwen Chen, Michael Ittmann, Xiaoying Shang, Shan Jiang, Haoyan Li, Chenling Meng, Ivonne Flores, Jian H. Song, James W. Horner, Zhengdao Lan, Chang-Jiun Wu, Jun Li, Qing Chang, Ko-Chien Chen, Guocan Wang, Pingna Deng, Denise J. Spring, Y. Alan Wang, and Ronald A. DePinho. Chromatin Regulator CHD1 Remodels the Immunosuppressive Tumor Microenvironment in PTEN-Deficient Prostate Cancer. *Cancer Discovery*, 10(9):1374–1387, September 2020. ISSN 2159-8290. doi: 10.1158/2159-8290.CD-19-1352.

Xin Zhou, Zhibin Chen, and Xiaodong Cai. Identification of epigenetic modulators in human breast cancer by integrated analysis of DNA methylation and RNA-Seq data. *Epigenetics*, 13(5):473–489, August 2018. ISSN 1559-2294. doi: 10.1080/15592294.2018.1469894.

Yonghua Zhuang, Fuyong Xing, Debashis Ghosh, Brian D. Hobbs, Craig P. Hersh, Farnoush Banaei-Kashani, Russell P. Bowler, and Katerina Kechris. Deep learning on graphs for multi-omics classification of COPD. *PLOS ONE*, 18(4):e0284563, April 2023. ISSN 1932-6203. doi: 10.1371/journal.pone.0284563.

A Appendix

Statistic	**BRCA**	**KIPAN**	**GBMLGG**	**STES**	**HNSC**	**LUAD**	**COADREAD**
No. of patients	1093	889	666	599	520	515	377
No. of genes	2527	2725	2051	3042	2621	2461	2356

Table A1: Statistics regarding the number of patient samples, and the genes for the seven TCGA cancer datasets after conducting the data processing pipeline.

Dataset	Hyperparameter	Value
BRCA	Spearman Correlation Threshold	$[0.25, 0.5, 0.75, \mathbf{1.0}, 1.25, 1.5]$
	Global Edge Threshold	$[6.25, 12.5, 25, \mathbf{37.5}, 50, 75]$
	Batch Size	$[1, 2, \mathbf{4}, 8, 16]$
	Number of Nodes	$[4, 8, 16, 32, \mathbf{64}]$
	Learning Rate	$[1e-3, \mathbf{1e-4}, 1e-5]$
KIPAN	Spearman Correlation Threshold	$[0.25, 0.5, 0.75, 1.0, \mathbf{1.25}, 1.5]$
	Global Edge Threshold	$[6.25, 12.5, \mathbf{25}, 37.5, 50, 75]$
	Batch Size	$[1, 2, \mathbf{4}, 8, 16]$
	Number of Nodes	$[4, 8, 16, 32, \mathbf{64}]$
	Learning Rate	$[1e-3, \mathbf{1e-4}, 1e-5]$
GBMLGG	Spearman Correlation Threshold	$[0.25, 0.5, 0.75, 1.0, \mathbf{1.25}, 1.5]$
	Global Edge Threshold	$[6.25, \mathbf{12.5}, 25, 37.5, 50, 75]$
	Batch Size	$[1, 2, \mathbf{4}, 8, 16]$
	Number of Nodes	$[4, 8, 16, 32, \mathbf{64}]$
	Learning Rate	$[1e-3, \mathbf{1e-4}, 1e-5]$
STES	Spearman Correlation Threshold	$[0.25, 0.5, 0.75, 1.0, \mathbf{1.25}, 1.5]$
	Global Edge Threshold	$[6.25, 12.5, 25, \mathbf{37.5}, 50, 75]$
	Batch Size	$[1, 2, \mathbf{4}, 8, 16]$
	Number of Nodes	$[4, 8, 16, 32, \mathbf{64}]$
	Learning Rate	$[1e-3, \mathbf{1e-4}, 1e-5]$
HNSC	Spearman Correlation Threshold	$[0.25, 0.5, 0.75, \mathbf{1.0}, 1.25, 1.5]$
	Global Edge Threshold	$[6.25, 12.5, 25, \mathbf{37.5}, 50, 75]$
	Batch Size	$[1, 2, \mathbf{4}, 8, 16]$
	Number of Nodes	$[4, 8, 16, 32, \mathbf{64}]$
	Learning Rate	$[1e-3, \mathbf{1e-4}, 1e-5]$
LUAD	Spearman Correlation Threshold	$[0.25, 0.5, 0.75, \mathbf{1.0}, 1.25, 1.5]$
	Global Edge Threshold	$[6.25, 12.5, \mathbf{25}, 37.5, 50, 75]$
	Batch Size	$[1, 2, \mathbf{4}, 8, 16]$
	Number of Nodes	$[4, 8, 16, 32, \mathbf{64}]$
	Learning Rate	$[1e-3, \mathbf{1e-4}, 1e-5]$
COADREAD	Spearman Correlation Threshold	$[0.25, 0.5, 0.75, 1.0, \mathbf{1.25}, 1.5]$
	Global Edge Threshold	$[6.25, 12.5, \mathbf{25}, 37.5, 50, 75]$
	Batch Size	$[1, 2, \mathbf{4}, 8, 16]$
	Number of Nodes	$[4, 8, 16, 32, \mathbf{64}]$
	Learning Rate	$[1e-3, \mathbf{1e-4}, 1e-5]$

Table A2: Hyperparameter tuning for the proposed LGIM model conducted for the seven cancer datasets.

Init	BRCA	KIPAN	HNSC	LUAD	GBMLGG	STES	COADREAD
Spearman	74.06 ± 13.87	73.9 ± 4.57	63.22 ± 3.86	60.57 ± 8.89	76.82 ± 2.58	67.56 ± 8.69	79.97 ± 1.85
Full	57.04 ± 9.71	73.84 ± 4.64	52.31 ± 14.89	64.57 ± 3.64	69.84 ± 11.06	59.6 ± 13.22	70.25 ± 11.13
Empty	58.1 ± 16.07	75.12 ± 2.23	61.47 ± 8.58	54.77 ± 13.47	68.69 ± 5.73	59.47 ± 12.2	69.81 ± 24.29
Random	55.43 ± 18.99	69.09 ±10.51	56.51 ± 9.08	57.10 ± 14.39	72.13 ± 5.52	66.03 ± 12.09	62.15 ± 16.71

Table A3: Ablation study to understand the effect of four different adjacency matrix initializations, Spearman, Full, Empty and Random, on the LGIM model. Performances reported are for the weighted F1 score.

Init	BRCA	KIPAN	HNSC	LUAD	GBMLGG	STES	COADREAD
Spearman	55.77 ± 2.99	67.04 ± 2.08	60.13 ± 1.67	62.21 ± 3.82	73.14 ± 2.23	57.95 ± 3.36	59.44 ± 2.76
Full	58.59 ± 2.52	67.8 ± 4.52	57.2 ± 4.3	58.9 ± 2.44	66.77 ± 3.63	58.48 ± 2.73	61.21 ± 3.98
Empty	59.03 ± 1.73	68.46 ± 3.07	61.08 ± 3.53	59.6 ± 2.69	70.94 ± 4.68	55.75 ± 3.05	65.28 ± 7.46
Random	58.36 ± 1.55	67.33 ± 2.60	58.09 ± 1.38	56.48 ± 1.35	67.56 ± 5.42	56.25 ± 1.29	60.89 ± 5.23

Table A4: Ablation study to understand the effect of four different adjacency matrix initializations, Spearman, Full, Empty and Random, on the LGIM model. Performances reported are for the balanced accuracy score.

Model	BRCA	KIPAN	GBMLGG	STES	HNSC	LUAD	COADREAD
GCN	72.8 ± 11.42	75.88 ± 4.66	76.91 ± 2.57	67.34 ± 7.81	52.48 ± 13.98	58.35 ± 9.35	72.61 ± 8.42
SAGE	61.9 ± 19.74	77.93 ± 1.35	78.94 ± 1.31	71.32 ± 6.50	63.12 ± 5.79	68.36 ± 2.39	77.37 ± 4.9
LGIM	74.06 ± 13.87	73.9 ± 4.57	76.82 ± 2.58	67.56 ± 8.69	63.22 ± 3.86	60.57 ± 8.89	79.97 ± 1.85
LGIM-SAGE	72.28 ± 9.65	74.98 ± 3.65	77.42 ± 3.32	65.4 ± 5.94	65.14 ± 2.94	68.56 ± 11.96	76.69 ± 5.04

Table A5: Ablation study to understand the effect of GCN, and SAGE message passing algorithms on the GNN baseline, and the LGIM model. Performances reported are for the weighted F1 score.

Model	BRCA	KIPAN	GBMLGG	STES	HNSC	LUAD	COADREAD
GCN	57.01 ± 2.99	71.33 ± 1.76	74.46 ± 2.99	60.46 ± 4.10	57.31 ± 2.49	54.87 ± 2.01	57.82 ± 4.65
SAGE	61.48 ± 2.84	73.39 ± 1.85	75.44 ± 2.93	65.21 ± 2.16	64.18 ± 1.89	62.16 ± 3.82	66.95 ± 4.28
LGIM	55.77 ± 2.99	67.04 ± 2.08	73.14 ± 2.23	57.95 ± 3.36	60.13 ± 1.67	62.21 ± 3.82	59.44 ± 2.76
LGIM-SAGE	55.89 ± 3.44	73.86 ± 2.38	75.18 ± 4.21	58.89 ± 4.72	62.39 ± 3.84	61.12 ± 4.15	64.74 ± 3.99

Table A6: Ablation study to understand the effect of GCN, and SAGE message passing algorithms on the GNN baseline, and the LGIM model. Performances reported are for the balanced accuracy score.

Model	Init	BRCA	KIPAN	GBMLGG	STES	HNSC	LUAD	COADREAD
GCN	Spearman	72.8 ± 11.42	75.88 ± 4.66	76.91 ± 2.57	67.34 ± 7.81	52.48 ± 13.98	58.35 ± 9.35	72.61 ± 8.42
GCN	PPI	66.61 ± 8.47	76.6 ± 4.83	77.82 ± 3.6	63.71 ± 6.1	52.52 ± 7.35	58.33 ± 5.26	73.41 ± 11.39
SAGE	Spearman	61.9 ± 19.74	77.93 ± 1.35	78.94 ± 1.31	71.32 ± 6.50	63.12 ± 5.79	68.36 ± 2.39	77.37 ± 4.9
SAGE	PPI	64.12 ± 12.87	76.25 ± 2.64	77.95 ± 3.18	65.98 ± 8.22	63.08 ± 4.61	65.94 ± 4.96	77.09 ± 4.68

Table A7: Ablation study to understand the effect of two different adjacency matrix initializations, Spearman and PPI, on the GCN and SAGE models. Performances reported are for the weighted F1 score.

Model	Init	BRCA	KIPAN	GBMLGG	STES	HNSC	LUAD	COADREAD
GCN	Spearman	57.01 ± 2.99	71.33 ± 1.76	74.46 ± 2.99	60.46 ± 4.10	57.31 ± 2.49	54.87 ± 2.01	57.82 ± 4.65
GCN	PPI	59.2 ± 3.65	72.19 ± 1.83	75.17 ± 4.35	62.86 ± 5.16	58.7 ± 2.49	60.69 ± 2.81	65.25 ± 7.92
SAGE	Spearman	61.48 ± 2.84	73.39 ± 1.85	75.44 ± 2.93	65.21 ± 2.16	64.18 ± 1.89	62.16 ± 3.82	66.95 ± 4.28
SAGE	PPI	61.52 ± 3.59	73.25 ± 2.21	75.75 ± 2.83	64.62 ± 2.13	65.38 ± 2.39	63.21 ± 3.96	68.20 ± 5.7

Table A8: Ablation study to understand the effect of two different adjacency matrix initializations, Spearman and PPI, on the GCN and SAGE models. Performances reported are for the balanced accuracy score.

Model	BRCA	KIPAN	GBMLGG	STES	HNSC	LUAD	COADREAD
LogReg	84.39 ± 0.92	77.43 ± 1.94	77.48 ± 1.41	73.09 ± 2.13	65.09 ± 2.63	69.86 ± 4.44	80.61 ± 2.46
LSVC	82.17 ± 0.89	77.12 ± 3.39	75.67 ± 2.60	72.76 ± 2.29	64.02 ± 2.38	66.76 ± 1.71	77.90 ± 5.35
XGB	63.26 ± 3.34	40.72 ± 3.97	52.87 ± 1.87	34.85 ± 7.85	39.07 ± 2.09	40.90 ± 5.76	72.05 ± 5.56
MLP	85.18 ± 0.38	78.81 ± 2.27	74.60 ± 1.98	71.92 ± 2.10	62.04 ± 4.05	70.63 ± 2.50	84.30 ± 2.06
GCN	72.8 ± 11.42	75.88 ± 4.66	76.91 ± 2.57	67.34 ± 7.81	52.48 ± 13.98	58.35 ± 9.35	72.61 ± 8.42
LGIM	74.06 ± 13.87	73.9 ± 4.57	76.82 ± 2.58	67.56 ± 8.69	63.22 ± 3.86	60.57 ± 8.89	79.97 ± 1.85

Table A9: Performance of the 5 baseline models, and the proposed LGIM model on the seven cancer datasets in terms of the weighted F1 score.

Model	BRCA	KIPAN	GBMLGG	STES	HNSC	LUAD	COADREAD
LogReg	52.62 ± 1.91	67.95 ± 3.01	73.15 ± 2.39	59.46 ± 2.27	59.72 ± 3.04	57.67 ± 5.49	52.87 ± 4.48
LSVC	50.97 ± 1.83	69.49 ± 3.57	71.83 ± 2.89	63.01 ± 1.54	59.68 ± 3.31	56.18 ± 1.34	57.87 ± 8.14
XGB	53.93 ± 5.38	55.29 ± 2.88	63.86 ± 2.74	52.45 ± 7.05	53.26 ± 0.99	53.07 ± 5.75	52.68 ± 8.00
MLP	49.41 ± 1.00	66.79 ± 3.40	68.19 ± 2.52	56.32 ± 1.84	55.18 ± 4.83	57.19 ± 3.00	58.66 ± 3.81
GCN	57.01 ± 2.99	71.33 ± 1.76	74.46 ± 2.99	60.46 ± 4.10	57.31 ± 2.49	54.87 ± 2.01	57.82 ± 4.65
LGIM	55.77 ± 2.99	67.04 ± 2.08	73.14 ± 2.23	57.95 ± 3.36	60.13 ± 1.67	62.21 ± 3.82	59.44 ± 2.76

Table A10: Performance of the 5 baseline models, and the proposed LGIM model on the seven cancer datasets in terms of the balanced accuracy score.

Term	Geneset	Adj P Val
Lymphoma (HP:0002665)	Human_Phenotype_Ontology	0.022
Abnormality of cells of the lymphoid lineage (HP:0012140)	Human_Phenotype_Ontology	0.032
Lymphopenia (HP:0001888)	Human_Phenotype_Ontology	0.032
Leukopenia (HP:0001882)	Human_Phenotype_Ontology	0.032
B Cell Receptor Signaling WP23	WikiPathways_2024_Human	0.001
B cell receptor signaling pathway	KEGG_2021_Human	0.078
Antigen Activates B Cell Receptor (BCR) Leading to Generation of Second Messengers	Reactome_Pathways_2024	0.035
Signaling by the B Cell Receptor (BCR)	Reactome_Pathways_2024	0.035
TP63	Transcription_Factor_PPIs	0.075
EP300	Transcription_Factor_PPIs	0.075
ELF1	Transcription_Factor_PPIs	0.075
CHD1	Transcription_Factor_PPIs	0.075
JUNB	Transcription_Factor_PPIs	0.075
ZNF263	Transcription_Factor_PPIs	0.075
HNF4G	Transcription_Factor_PPIs	0.099
TAF7	Transcription_Factor_PPIs	0.099
IRF3	Transcription_Factor_PPIs	0.099

Table A11: Overrepresentation analysis for cluster 7 from the best performing LGIM model fold on the BRCA dataset. Selected terms are shown.

Term	Geneset	Adj P Val
Adipogenesis	MSigDB_Hallmark_2020	0.088
Pperoxisome	MSigDB_Hallmark_2020	0.088
Drug ADME	Reactome_Pathways_2024	0.047
Biological Oxidations	Reactome_Pathways_2024	0.051
Metabolism	Reactome_Pathways_2024	0.051
Metabolism of Lipids	Reactome_Pathways_2024	0.051
NOTCH3 Intracellular Domain Regulates Transcription	Reactome_Pathways_2024	0.051
Signaling by NOTCH3	Reactome_Pathways_2024	0.051
Triglyceride Metabolism	Reactome_Pathways_2024	0.075
Fatty Acids	Reactome_Pathways_2024	0.085
Metabolism of Steroid Hormones	Reactome_Pathways_2024	0.085

Table A12: Overrepresentation analysis for cluster 8 from the best performing LGIM model fold on the BRCA dataset. Selected terms are shown.

Term	Geneset	Adj P Val
TNF-alpha Signaling via NF-kB	MSigDB_Hallmark_2020	0.014
Inflammatory Response	MSigDB_Hallmark_2020	0.014
negative regulation of intrinsic apoptotic signaling pathway (GO:2001243)	GO_Biological_Process_2021	0.087
negative regulation of apoptotic signaling pathway (GO:2001234)	GO_Biological_Process_2021	0.087
TBX5	GO_Biological_Process_2021	0.041
JARID2	GO_Biological_Process_2021	0.041
ESR1	GO_Biological_Process_2021	0.060
GATA4	GO_Biological_Process_2021	0.060
NF-kappa B signaling pathway	KEGG_2021_Human	0.070
IL-17 signaling pathway	KEGG_2021_Human	0.070
Apoptosis	KEGG_2021_Human	0.078

Table A13: Overrepresentation analysis for cluster 12 from the best performing LGIM model fold on the BRCA dataset. Selected terms are shown.