Mathematical Programming and Graph Neural Networks illuminate functional heterogeneity of pathways in disease

Research Article

Charalampos P. Triantafyllidis
Nuffield Department of Medicine,
University of Oxford, Oxford, OX3 7BN, UK
harry.triantafyllidis@ndm.ox.ac.uk

Abstract

We employ a computationally intensive framework that integrates mathematical programming and graph neural networks to elucidate functional phenotypic heterogeneity in disease by classifying entire pathways of interest under various conditions. Our approach combines two distinct yet seamlessly integrated mathematical modelling schemes: (i) we first leverage Prior-Knowledge gene regulatory Networks (PKNs) derived from comprehensive and established databases to reconstruct and optimize their topology using genomic and transcriptomic data and mathematical programming optimization, and (ii) we apply tailored Graph Convolutional Networks (GCNs) to classify each network as a single data point at *graph-level*, using Mode of Regulation (MoR) and gene activity profiles as node embeddings. These networks may vary in their biological or molecular annotations, which serves as a labelling scheme for their supervised classification. As a case study, we investigate the DNA damage and repair pathway using the *TP53* regulon in a pancancer study, classifying gene networks across different *TP53* mutation types. Our findings pave the way for a systematic, causality-driven approach to classify stimuli or stress conditions and their functional implications, thereby advancing drug discovery and precision medicine.

Keywords Graph Neural Networks, Optimization, Linear Programming, Gene Networks, *TP53*

Introduction

The challenge of causal interpretation in medical sciences Rizzi and Pedersen (1992); Bach (2019); Richens et al. (2020) and the necessity to utilize it in the light of recent advances in high-throughput genomic and transcriptomic data and at the same time the wider availability of computational resources and modelling approaches, marks a distinct gap in our understanding of disease Katuwal and Chen (2016); Gilpin et al. (2018); Carvalho et al. (2019); Vellido (2020). To address the gap, particularly in biological deregulation of genomic underpinnings, it is instrumental to elucidate how signals propagate across molecules from root(s) to target(s). Genes and their proteins, can be effectively captured in a computational model by an artificial *wiring* system, resulting in network representations or *graphs* (Smolen et al., 2000; Brazhnik et al., 2002; Bonneau, 2008; Karlebach and Shamir, 2008a; Althaus et al., 2009; Le Novère, 2015). These graphs contain nodes (genes in this context) and edges (interactions typically mapping an activating or inhibitory effect) indicating the Mode of Regulation (MoR). Genes do not only interact in an amphi-directional way but mainly

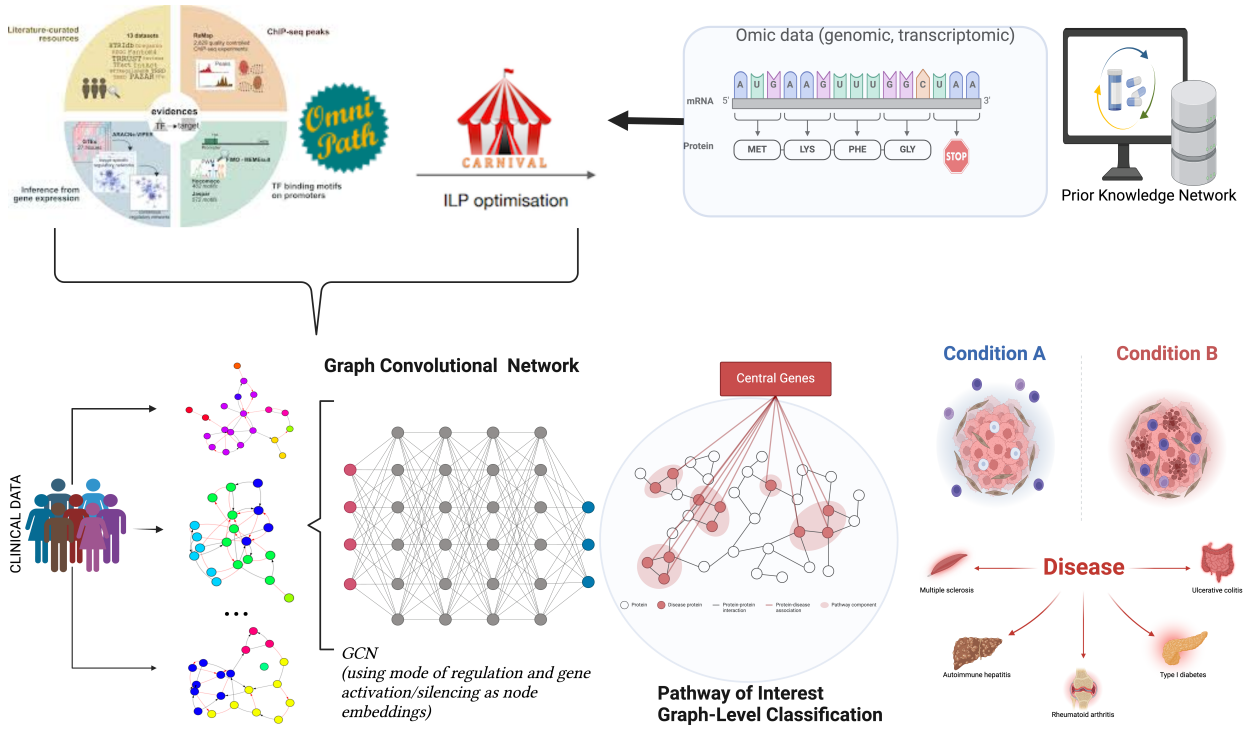

Figure 1: The developed framework workflow. Prior Knowledge Networks (PKNs) and omics-data are combined to provide the input for optimization models to reconstruct gene regulatory networks across pathways of interest, in this case the *TP53* pathway and its regulon. These optimized networks are then annotated across conditions of interests, in this example mutation types on the transcription factor *TP53*. We then analyse the networks in two different ways: first we extract unique signatures for each condition (mutation type) and compare them, using community detection and betweenness centrality. We then classify the optimized networks using Graph Convolutional Networks (GCNs) using as labels the condition of interest for each graph. Finally, the classifier can distinguish entire pathways as a single data-point (graph level classification) to denote diseases or states of diseases (onset, progression, severity etc.)

.

in symphony to express a particular phenotype, a cellular process carrying out specific biological tasks (such as DNA damage and repair, proliferation, apoptosis, etc.). The importance of this mathematical *language* to communicate gene interactions is rooted in its ability to provide causal interpretation of the shared effects (Pearl et al., 2016) by using directed acyclic graphs (DAG).

Thus, a systems biology approach, using high-throughput data and advanced computational methods is instrumental in the illumination of the functional implications of different conditions that can map genotypes and transcriptomic data to phenotypes in disease. In particular, graph-based models provide a natural representation of gene regulatory networks (GRNs) Karlebach and Shamir (2008b); Erwin and Davidson (2009), enabling the integration of multi-dimensional data to capture structural and therefore functional differences across layers such as gene activity profiles and regulatory direction. This line of work builds on two primary blocks, incorporating two distinct and yet seamlessly integrated modelling schemes. The first one is based on well-established Mixed-Integer Linear Programming (MILP) models Bertsimas and Tsitsiklis (1997) which we employ to i) map signalling amongst involved genes and consistency rules and

ii) reconstruct their topology and gene activity profiles for the pathway of interest as a whole. These models are deterministic and global optimal, hence offering advantages in the optimization of the networks. Furthermore, a plethora of network analysis metrics such as community detection, provide further insight into disease states.

The second modelling scheme builds on Graph Convolutional Networks (GCNs) (Kipf and Welling, 2016; Corso et al., 2024; Hamilton, 2020; Wu et al., 2020), representing a cutting-edge class of neural networks specifically derived for graph-based structured data, where other approaches of neural networks cannot be applied, making them ideal for interrogating GRNs. Graph Convolutional Networks (GCNs) have become a key approach in analysing graph-structured data, enabling the extraction of meaningful patterns from complex networks. In Lee et al. (2023) deep learning has been successfully used to identify resistance to therapy p53 pathway mediated mechanisms and related combinatorial targets for lung cancer. In Ling et al. (2013), a Recurrent Neural Network (RNN) has been employed to illuminate p53 temporal interaction dynamics. Furthermore, a Graph Neural Network approach has been utilized in Wang et al. (2023) to predict protein stability change upon mutation in genes.

We apply the developed framework in cancer, particularly because of its entwined causality of genomic nature, therefore making it ideal for demonstration. Cancer onset and its rapid progression is a result of pervasive biological deregulation taking place across a plethora of cellular processes Hanahan (2022); Lujambio and Lowe (2012); Hanahan and Weinberg (2000, 2011). As aberrant gene expression drives oncogenic processes and initiates a multitude of neoplastic diseases Nesbit et al. (1999), the master regulators of the genome, known as Transcription Factors (TF) Lambert et al. (2018); Yusuf et al. (2012); Papavassiliou (1995), dictate the gene activity profile of many downstream targets, known as their *regulon*. Across these approximately 1,600 DNA-binding TFs (Pratt et al., 2022), their intricate dynamics are paramount in furthering our understanding of gene expression programs, transcription and post-transcriptional regulation in disease. Their perturbations in cancer emanating from stress conditions such as mutations, irradiation, hypoxia Wilson and Hay (2011); Brown and Wilson (2004) or drug effects and other stimuli, are pivotal in determining a cascade of processes underlying disease onset, progression and prognosis. Particularly somatic mutations Tomasetti et al. (2017), can alter the functional form of proteins and their ability to dock and perform cellular tasks, leading to abnormal states. Among these transcription factors, *TP53* Soussi and Wiman (2015); Olivier et al. (2010); Petitjean et al. (2007) stands out due to its widespread impact in multiple types of cancer Hollstein et al. (1991); May and May (1999), influencing crucial pathways for cell cycle control, apoptosis, and DNA repair, thus widely known as the *Guardian of the Genome* Lane (1992).

Somatic mutations in *TP53*, manifest as heterogeneous effects on downstream gene expression Donehower et al. (2019); Kruse and Gu (2009) and can be used as annotated conditions of interest in the developed pipeline. Besides mutations, any external stimuli that can be described as a perturbing effect could be utilized, such as hypoxia and irradiation, as demonstrated in Triantafyllidis et al. (2023), or drug effects. *TP53* is a major transcription factor involved in cellular defences against cancer and is inactivated in nearly all tumours, with mutations in *TP53* occurring in approximately 50% of human cancers. Tumour suppressors like *TP53* typically experience loss-of-function mutations. Unlike other tumour suppressors, *TP53* mutations are predominantly missense variants concentrated in the DNA-binding domain (DBD). *TP53*'s diverse mutation landscape is a result of its structural fragility, evolutionary compromises, and intrinsic mutability. Most mutations lead to loss of tumour-suppressive function, with additional dominant-negative or aggregation-related effects exacerbating tumorigenesis. While some mutations exhibit neomorphic effects, these are often secondary to the primary loss of p53 function Joerger et al. (2024).

This functional heterogeneity poses a significant challenge for understanding mutation-specific effects, requiring the dissection of the molecular mechanisms driving distinct regulatory states. Analysis of the functional implications in *TP53* states support that $p53^{WT}$ enhances both innate and adaptive immunity by promoting pro-inflammatory cytokine secretion, activating the cGAS-STING pathway and regulating antigen presentation via MHC-I, influencing immune checkpoint pathways Efe et al. (2024). On the other hand, *TP53*

mutations can create an immunosuppressive tumour micro-environment (TME) with altered cytokine and chemokine production that recruits immunosuppressive cells like myeloid-derived suppressor cells (MDSCs) and M2 macrophages and induction of chronic inflammation via NF-κB and other signalling pathways, fostering tumour progression Efe et al. (2024). Additionally, further investigation on the functional effect of *TP53* mutations on its tumour suppressing ability shows that Pan et al. (2023) non-functional mutations are associated with destabilization of p53's DNA-binding domain (DBD) and oligomerization domain (OD) and at the same time, destabilization effects suggest a potential therapeutic strategy to restore p53 function by inducting again stabilization. Specifically missense *TP53* mutations can be predicted with extremely high accuracy Ben-Cohen et al. (2022), suggesting the heterogeneity across missense mutations themselves each contributing to varying functional effects on p53 activity, whilst Hotspot mutations were strongly associated with loss of function. These approaches, while providing inference in p53's functionality based on variant classifications, lack the ability to map this effect in the TFs downstream transcriptional regulation, uncover the causality and provide a holistic understanding of the regulatory landscape.

This study builds on these computational innovations to explore the regulatory impact of somatic mutations in TFs, focusing on *TP53* as a paradigm. By reconstructing and optimizing GRNs from the Cancer Cell Line Encyclopedia (CCLE) Tsherniak et al. (2017) dataset, we aim to elucidate mutation-driven differences in regulatory architecture. The integration of graph-based classification and network-level analyses allows us to answer if somatic mutations alter regulatory networks to a detectable degree, by finding a classification boundary. The insights gained not only advance our understanding of transcription factor biology but also hold potential for identifying novel therapeutic targets and biomarkers entwined to precision oncology, and expand further outside of our case study to disease in general.

Results

Models from CCLE were collected (n = 1618) incorporating either mutant/WT *TP53*. The DoroThEA Garcia-Alonso et al. (2019) regulon was extracted for this transcription factor using OmniPathR Turei et al. (2016) and accounting only for signed and directed edges. This amounted to approximately 200 downstream targets with all confidence levels included, to remain as less stringent as possible and allow the optimization prune as appropriate to minimize the mismatch between the expression matrix and the prior knowledge network in terms of its initial topological features and orientation. The CARNIVAL Liu et al. (2019) pipeline was used to reconstruct the networks using the normalized enrichment score and the deleterious function of mutation as a perturbing event on *TP53* being the master node in all networks, and the downstream targets following below. This mathematical programming optimization package allows for single sample reconstruction of networks and proves to be ideal in this line of work, as each network is depicting exactly one CCLE model carrying a single *TP53* mutation.

We first sought to explore any heterogeneity found in the expression profile of the regulon, and although this has been previously explored Triantafyllidis et al. (2023) in older versions of the dataset, for consistency we re-iterated the main exploratory angles. We extracted all cell line samples carrying a single *TP53* mutation and combined the expression levels of its regulon with the mutational profiles and other molecular information such as cancer type, protein change, hotspot and deleterious function of the mutation each time. We stratified then the expression levels of the regulon per mutation type and performed a dual hierarchical clustering of Z-standardized transformed expression levels to understand whether either cell-lines, or gene targets of *TP53* cluster together in a consistent pattern, Figure 6. No pattern was observed across either WT or mutated samples, emphasizing the transcriptional divergence of *TP53* and the potential functional implications being inseparable by isolated expression analysis. Transcription factor activation is known to be visible only in the gene activity profile of the regulon and not on its expression itself in isolation.

We utilize the deleterious function of *TP53* mutations as its perturbing effect in the network each time. As deleterious we consider the effect of rendering dysfunctional the resulting protein p53, and so, a deleterious

mutation would implicitly mean that *TP53* is deactivated , or in a similar fashion activated if it's not identified as deleterious. This perturbing effect is used as an input in the CARNIVAL pipeline. The computational analysis provides insights into the classification and expression patterns of *TP53* related mutations across various cell lines. Figure 10 shows the expression levels of *TP53* across mutant (and stratified across deleterious function of mutation) and WT models (cell-lines). Statistically significant ($p < 2.2e-16$, Wilcoxon and T-tests) higher expression is observed across deleterious and mutant models versus non-deleterious mutation cell-lines. WT samples exhibit similar expression levels to the high mutant deleterious models, again confirming that expression levels for *TP53* transcription factors are not representative of status (active/inactivated). Figure 11 shows the violin plot of *TP53* expression stratified by mutation type and WT samples included, and significant differences across the means of the variants.

Network Analysis

The backbone prior knowledge network as extracted from OmniPathR contains approximately 40k interactions. This graph serves as the building block for every CCLE model being re-optimized as a single gene regulatory network, each time using as input the expression of this single CCLE model. In essence, we reconstruct as many gene regulatory networks as cell-lines, or CCLE samples, contain expression and mutation data available to characterize *TP53* status, both in terms of expression but also in the deleterious function of the mutation, each time defining the perturbing effect.

All initial prior-knowledge networks extracted from OmniPathR including all downstream identified signed and directed interactions with *TP53*, have been reconstructed using the full expression matrix from CCLE and *TP53* perturbation from the deleterious function of mutation each time (deleterious:inactive). For WT samples we consider *TP53* functional and active. We used CARNIVAL to optimally reconstruct the networks, which then served as an input of training graphs for the GCN classifier. The network analysis we performed included two parts: i) community detection, and ii) centrality analysis per identified hub and extraction of unified signatures across each mutation type (See Methods). Figure 2 shows a hierarchical dual clustering with the regulon target genes on the x-axis and the mutation types plus WT samples as different categories on the y-axis. The genes can be absent (white), activated (blue) and inactive (red) , across each mutation type / WT samples. We can observe high dissimilarity between the extracted signatures, with WT and missense being closer to each other across the rest of the signatures. Enriched KEGG pathway analysis on the set difference between the signatures of WT and missense mutation in Figure 7 reveals ($p<0.05$, top ten pathways) transcriptional misregulation in cancer as a major pathway involved. Figure 8 depicts the intersections of the signatures ordered by frequency across all mutation types, demonstrating the high degree of heterogeneity across the signatures with minimum overlaps in the genes and their associated states.

GCN classification

The confusion matrix, Figure 9, reveals the classification performance for different mutation types, with strong accuracy in predicting WT (wild-type) and `missense_variant` classes, indicated by the high diagonal values (146 and 171, respectively). However, notable misclassification occurs for `stop_gained` and `frameshift_variant`, suggesting overlaps in their prediction boundaries. Figure 4 shows the Receiver Operating Characteristic (ROC) curve for the multi-label classifier across all variants and WT samples. We observe that generally the classifier performs sufficiently well to differentiate across samples with strong presence in the training data, whilst struggling to separate cases low in incidence. In CCLE, the vast majority of samples as also evident in Figure 6, belong to either missense or WT samples. For these two categories respectively the ROC is relatively quite strong (0.92 and 0.81). Scatter plots were generated for the true and predicted class Figure 5, and the classification accuracy (blue:correct, red:misclassified). Figure 3, shows an interactive plot generated by the platform. Here, the user can select each data point and bring up the

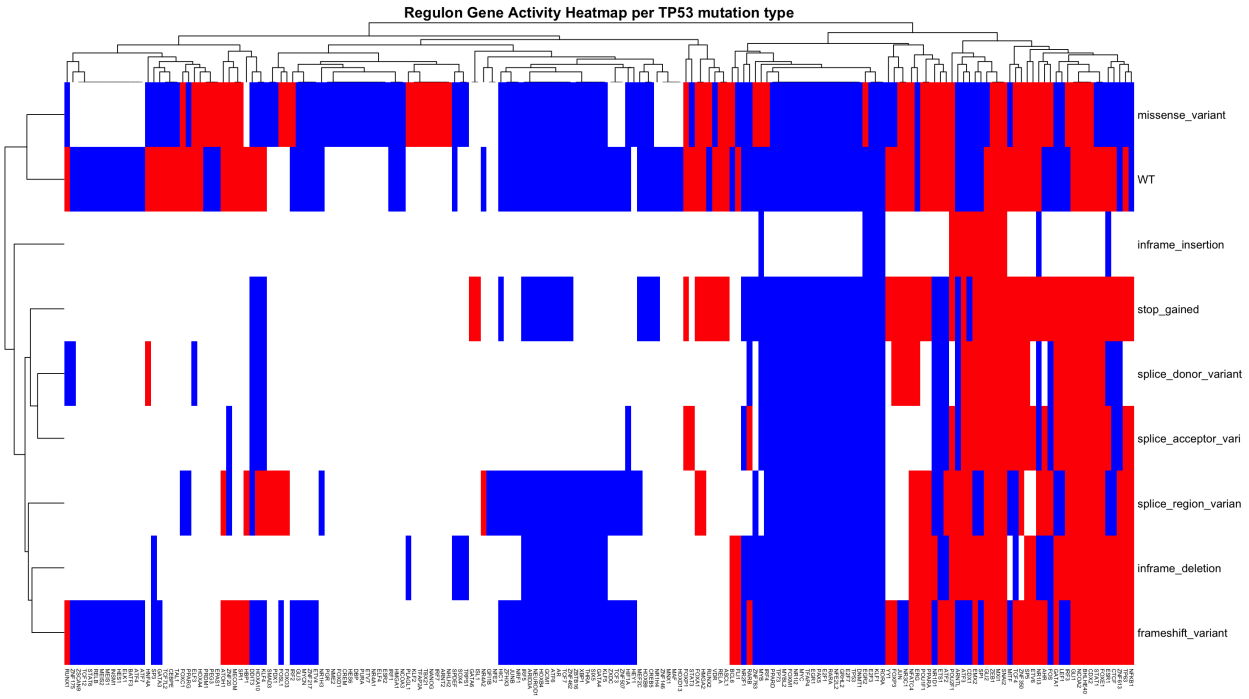

Figure 2: Plotting the gene signatures extracted from the community detection analysis across each mutation type and WT, and showing gene activity each time. All networks carrying the same *TP53* mutation where stratified into groups, and for each group, community detection was performed and centrality score for each hub identified one gene per community. The overlap of these sets across same mutation type or WT each time, is the set shown for each mutation type/WT here respectively. The heatmap shows clustered gene status, with genes on the y-axis and samples (or conditions) on the x-axis. Color Scheme: Blue: Indicates downregulated genes. Red: Indicates upregulated genes. White: Represents absence.

optimized network corresponding to this data point, with its graph representation, showing not only the directionality of the interactions but also the activity profile (genes up or down denoted by blue/red colour).

Discussion

This study combined various methods and models for the analysis of pathway heterogeneity and its association with disease states and conditions of interest. We combined a mathematical programming approach with a deep learning model to facilitate the generation and classification of gene regulatory networks that map to a specific biological mechanism, in this case the regulon of *TP53* as a pathway of transcriptional regulation in DNA repair and damage. Our first step was to collect a near optimal representation of the regulon, using DoRothEA, which included approximately 200 genes. Then, we extracted a basis gene regulatory Prior Knowledge Network (PKN) from OmniPathR to map the expression levels of the regulon and reconstruct its topology and gene activities using the mRNA sequencing from CCLE, for each sample/CCLE model that carried a single *TP53* mutation. This resulted into approximately 1,600 base networks which then we restructured via mathematical programming optimization using CARNIVAL. This suite allows us to optimize per sample, given perturbations of transcription factors and in this case we used the deleterious function of mutation to indicate *TP53* status, or WT being active. The optimized networks were then used as a training dataset for a Graph Convolutional Network which attempted to classify them across the type of *TP53* mutation. The trained classifier was able to differentiate mainly between WT and missense mutations, given the class imbalance in the dataset for the remaining MT variants.

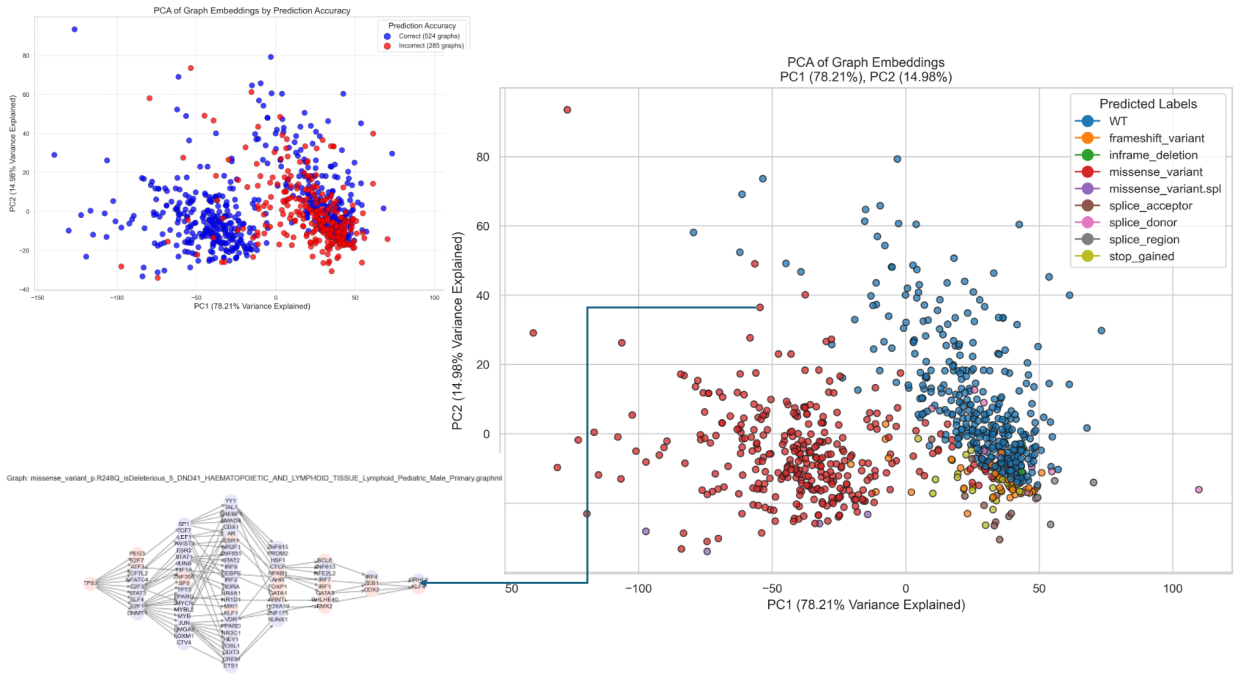

Figure 3: **Top left:** PCA Jolliffe and Cadima (2016) 2-D scatter plots of the node embeddings showing the prediction accuracy, correctly classified (blue) and misclassified (red) graphs. **Right:** PCA 2-D scatter plots of the node embeddings predictions with different colour annotation for the classified labels. The graph is interactive and the user can select in real-time each data point and bring forward as shown the actual graph behind each data-point with the gene activity profile mapped as different colours (lavender activated, light-rose inactive), along with all molecular features annotated such as true mutation type, cancer site, protein change, deleterious and hotspot.

Although this pipeline of methodological steps was demonstrated via the specific transcription factor and dataset, any different pathway (collection of interacting genes to perform a specific biological function of interest or suspected phenotypic diversity that links to disease states) can be used, assuming the availability of data (initial PKN, expression data) and the assumption that the condition of interest and the disease under investigation emanate from genomic causes primarily. The novelty of the approach lies in the fact that the resulting pathway variants are classified *as a whole*, or at graph-level, based on their genomic and topological characteristics : the GCN takes into account the Mode of Regulation (MoR), thus the type of interaction between the nodes (genes) as inhibition or activation but also the gene activity profile of the pathway (the gene statuses, up or down, of all nodes in the pathway). Although the proposed framework was demonstrated in an application for cancer, similar approaches can be used in other diseases too, to investigate a genotype-to-phenotype relationship of disease states and molecular data , thus not compromising the general utility and applicability of the proposed framework.

Additionally, given that networks exhibit significant flexibility of data representation and allow the incorporation of further genomic and transcriptomic or other (imaging) data, this framework paves the way for further expansion and inclusion of multi-modal datasets, assembling more detailed networks by the embedding of additional features as annotated states. In this case, features such as protein change, tissue origin, cancer type, gender, age, hotspot or deleterious function of mutation have been annotated on each network, as seen in figure 3 network, as a single identifier of the network itself. This simple approach is quite

effective in expanding the array of features (clinical, molecular, etc.) used to annotate the networks, allowing straightforward change of the classification labeling scheme used in the GCN step. For instance, one can easily generate labels for each network based on different delimiters (e.g. protein change) and then re-train the classifier based on the new labeling scheme. The code and the implementation itself have been tailored exactly for this type of experimentation in a modular way, allowing the GCN training and classification steps to be completely detached from the optimization step, so that it does not require repetition of GRN reconstruction.

However, as with any modelling approach, this line of work also comes with limitations. The size of the pathways that can be re-optimized per condition may not scale yet to thousands of genes (we used a few hundred in this use case in CCLE), and it might also pose questions about the validity of incorporating so many genes in a pathway without a *bona fide* knowledge of the underlying biological ground truth. Additionally, many methods exist in constructing GRNS from omics data, so future work could include a comparison of GCN classification by altering the method that reconstructs and optimizes the input GRN dataset.

To conclude, this line of work provides a computational basis of causal inference in the endeavour of identifying phenotypic diversity directly linked to disease states (for example severity, onset, progression, metastasis for cancer etc.) thus proving valuable in our attempt to disentangle the complex molecular and biological underpinnings of disease using networks and machine learning/optimization, successfully applied herein in cancer as a demonstrating paradigm.

Methods

The pipeline of the platform demonstrating the array of methods, software and packages utilized can be seen in Figure 1. Data was collected from DepMap CCLE and mutation and expression profiles were extracted for *TP53*.

Network Reconstruction Framework

A regulatory network is a collection of genes and their interactions; a gene can up-regulate or down-regulate other genes, and this interaction is depicted using graph theory and edges connecting nodes in a network. Each GRN was represented as an undirected weighted graph $G = (V, E)$, where V represents genes (nodes) and E represents regulatory interactions (edges). Node attributes included gene expression and mutation status, while edge weights corresponded to the strength of regulatory interactions.

Reconstructing a network amounts to the change of topology that corresponds to a particular sample (usually a full transcriptomic sequence, or else, the expression of all genes for a particular patient) so that the gene activity profiles represented by the network better fit the experimental data, than the prior knowledge. This Prior Knowledge Network (PKN) can be obtained in many different ways by combining information from multiple sources.

Given the large number of the reconstructed networks (~ 1600), it proved challenging to find an optimal way of *visualising* in a succinct way the structural differences of these networks, rendering the need for a *compression* method imperative. Our approach hence employed community detection and betweenness centrality measures to map each of these networks to a signature. We first applied an InfoMap community detection algorithm Smiljanić et al. (2023); Rosvall and Bergstrom (2008) to split each network into tightly connected hubs/communities of nodes. We then applied betweeness centrality for each of these hubs separately, and the highest scoring genes were collected to a single signature. As these signatures differ even across networks carrying the same type of *TP53* mutation, we took their intersection to form a unique for the mutation type signature. For each different mutation type a new signature hence was derived. This approach differs from Triantafyllidis et al. (2023) not only in the community detection algorithm used (InfoMap versus

Louvain) but also since we now included the states of the genes (up/down) and incorporated the information in the signatures. The resulting heatmap showing each signature per mutation type, and the states of the included genes per signature can be seen in Figure 2. Collectively, these signatures convey a clear message: only missense mutations exhibit high similarity with WT samples. The rest of the variants project similarities across themselves but not with either missense, or WT samples.

- **Input Data:** Gene expression data and mutation profiles were merged and obtained from the Cancer Cell Line Encyclopedia (CCLE), alongside prior knowledge of TF-target interactions from public databases jointly processed and combined in DoRothEA.
- **Optimization Model:** GRN reconstruction was formulated as a mixed-integer linear programming (MILP) problem as defined in Liu et al. (2019). Mixed-Integer Linear Programming (MILP) is an NP-hard special case of the general Linear Programming (LP) model, where one imposes integrality constraints for a subset of the variables:

$$\begin{aligned} \min \quad & c^T x \\ \text{subject to} \quad & \mathcal{A}x = b \\ & x_{\mathcal{G}} \in \mathcal{R}^+, x_{\mathcal{I}} \in \mathcal{Z}^+ \end{aligned} \tag{MILP}$$

  where $\mathcal{A} \in \mathbb{R}^{m \times n}$, $(c, x) \in \mathbb{R}^n$, $b \in \mathbb{R}^m$, T denotes transposition and $x = \begin{bmatrix} x_{\mathcal{I}} \\ x_{\mathcal{G}} \end{bmatrix}$. The integrality of the constraints serves as a mathematical way of mapping the prediction of link activation or removal as a binary variable. The objective function minimizes the mismatch between the experimental data (mRNA expression) and what the PKN implies for the status of each gene in the network, thus restructuring its topology, composure and directionality (MoR) to minimize loss. The main body of the constraints included encoding simple consistency rules, e.g. if A activates (via MoR) B and A is active then B should be active. For a more detailed description of the model see Liu et al. (2019).
- **Constraints:**
  - *Sparsity:* A sparsity constraint was imposed to limit the number of edges, ensuring biologically plausible networks.
  - *Consistency with Prior Knowledge:* Constraints ensured that reconstructed edges aligned with known TF-target interactions.
- **Optimization Solvers:** The MILP problem was solved using commercial solvers (IBM CPLEX).
- **Network Conversion:** Post-reconstruction, networks were converted from DOT format to graphml for further processing from R to Python.

Comparative Network Analysis

To identify mutation-specific differences in regulatory networks, comparative analyses were performed:

- Community Detection: Infomap algorithm was employed to identify functional modules, with module enrichment analyses revealing mutation-specific pathways and gene signatures.
- Centrality: Each identified hub was assessed for its key gene in terms of betweenness centrality score. All top centrality genes (per hub) with their associated state (up/down) were collected and intersected across all networks annoated as carrying the same *TP53* mutation type, thus generating a unique (master) signature per variant.

This integrative approach, combining machine learning, optimization, and GNNs, provides a powerful framework for uncovering the regulatory impact of somatic mutations and elucidating their functional implications in cancer biology.

Graph Neural Network Model

We implemented a Graph Convolutional Network (GCN) to classify graph data based on node features and structural information. The model was developed using the PyTorch Geometric framework and consisted of three `GraphConv` layers, followed by a global pooling layer and a fully connected linear layer for classification. Below, we describe the key components of the model and the data preparation pipeline. The GCN is defined mathematically as follows:

- **Input:** Each graph $G = (V, E)$ consists of nodes V and edges E. Node features are represented by a matrix $\mathbf{X} \in \mathbb{R}^{|V| \times d_{in}}$, where d_{in} is the dimension of the initial node embeddings.
- **Convolutional Layers:** The model employs three `GraphConv` layers, which iteratively update node embeddings. Let $\mathbf{H}^{(l)} = [\mathbf{h}_1^{(l)}, \mathbf{h}_2^{(l)}, \ldots, \mathbf{h}_N^{(l)}] \in \mathbb{R}^{N \times d_l}$ represent the node embeddings at layer l, where N is the number of nodes and d_l is the dimensionality of the embeddings at that layer. Each `GraphConv` layer updates the embeddings as:

  $$\mathbf{h}_v^{(l+1)} = \sigma\left(\sum_{u \in \mathcal{N}(v)} \frac{1}{\sqrt{d_v d_u}} \mathbf{W}^{(l)} \mathbf{h}_u^{(l)}\right),$$

  where:
  - $\mathcal{N}(v)$ is the set of neighbours of node v,
  - d_v and d_u are the degrees of nodes v and u, respectively,
  - $\mathbf{W}^{(l)}$ is the learnable weight matrix for layer l,
  - $\sigma(\cdot)$ is the ReLU activation function.

  This process is repeated for three layers to capture both local and higher-order structural information.
- **Global Pooling:** Node embeddings $\mathbf{H}^{(3)}$ are aggregated into a graph-level embedding $\mathbf{g}$ using global mean pooling:

  $$\mathbf{g} = \texttt{global_mean_pool}(\mathbf{H}^{(3)}, \texttt{batch}),$$

  where `batch` maps nodes to their respective graphs in batched inputs.
- **Fully Connected Layer:** The pooled embedding $\mathbf{g}$ is passed through a linear layer for classification:

  $$\mathbf{z} = \mathbf{W}_{\text{lin}} \mathbf{g} + \mathbf{b}_{\text{lin}},$$

  where $\mathbf{W}_{\text{lin}}$ and $\mathbf{b}_{\text{lin}}$ are learnable parameters.

The model is trained using a cross-entropy loss function:

$$\mathcal{L} = -\frac{1}{N} \sum_{i=1}^{N} \sum_{c=1}^{C} y_i^{(c)} \log \hat{y}_i^{(c)},$$

where:

- N is the number of training samples,
- C is the number of classes,
- $y_i^{(c)}$ is the true label for class c of sample i,
- $\hat{y}_i^{(c)}$ is the predicted probability for class c.

Graphs were represented using the GraphML format and processed using the `igraph` library. For each graph:

- **Node Features:** Node features were extracted from graph properties, including:
  1. Betweenness centrality,

2. Community membership identified using the Infomap algorithm,
3. In-degree and out-degree,
4. Node colors representing gene activation states
5. Edge type or, Mode of Regulation (MoR), -1,1 for inhibitory edges or activating each time

These features were standardized using `StandardScaler`.

- **Edge Representation:** Edges were represented as an adjacency list and converted to a `PyTorch Geometric` `edge_index` tensor.

The GNN was trained for a fixed number of epochs (`300`) using the Adam optimizer with a learning rate of `0.01`. Dropout regularization (`p = 0.5`) was applied to prevent overfitting, batch-size and hidden network dimesions were set to `16,32` respectively. Train-to-test ratio was 0.5.

Materials & Data

Code

All associated code with this line of work can be found in the repository https://github.com/harrytr/HarmonizeR, where we built a Shiny R application (*Harmonizer*) with user interface to facilitate reproducibility of our experiments.

Genomic and Transcriptomic Data

Data have been collected from DepMap 24Q2 https://depmap.org/portal/data_page/?tab=currentRelease, using

- OmicsExpressionProteinCodingGenesTPMLogp1BatchCorrected.csv
- OmicsSomaticMutations.csv
- OmicsCNGene.csv

Data from DoroThEA was collected using OmniPathR (using import of DoroThEA interactions command for human organism) and extracting only directed and signed edges.

References

Althaus, E., Klau, G.W., Kohlbacher, O., Lenhof, H.P., Reinert, K., 2009. Integer Linear Programming in Computational Biology. Springer Berlin Heidelberg, Berlin, Heidelberg. pp. 199–218. URL: https://doi.org/10.1007/978-3-642-03456-5_14, doi:doi:10.1007/978-3-642-03456-5_14.

Bach, J.F., 2019. Causality in medicine. Comptes Rendus. Biologies 342, 55–57.

Ben-Cohen, G., Doffe, F., Devir, M., Leroy, B., Soussi, T., Rosenberg, S., 2022. Tp53_prof: a machine learning model to predict impact of missense mutations in tp53. Briefings in Bioinformatics 23, bbab524. URL: https://doi.org/10.1093/bib/bbab524, doi:doi:10.1093/bib/bbab524, arXiv:https://academic.oup.com/bib/article-pdf/23/2/bbab524/42805631/supplementary_figure_s4a_and_b_bl

Bertsimas, D., Tsitsiklis, J., 1997. Introduction to linear optimization. Athena Scientific.

Bonneau, R., 2008. Learning biological networks: from modules to dynamics. Nature Chemical Biology 4, 658–664.

Brazhnik, P., de la Fuente, A., Mendes, P., 2002. Gene networks: how to put the function in genomics. Trends in Biotechnology 20, 467–472.

Brown, J.M., Wilson, W.R., 2004. Exploiting tumour hypoxia in cancer treatment. Nature Reviews Cancer 4, 437–447.

Carvalho, D.V., Pereira, E.M., Cardoso, J.S., 2019. Machine learning interpretability: A survey on methods and metrics. Electronics 8, 832.

Corso, G., Stark, H., Jegelka, S., Jaakkola, T., Barzilay, R., 2024. Graph neural networks. Nature Reviews Methods Primers 4, 17.

Donehower, L.A., Soussi, T., Korkut, A., Liu, Y., Schultz, A., Cardenas, M., Li, X., Babur, O., Hsu, T.K., Lichtarge, O., Weinstein, J.N., Akbani, R., Wheeler, D.A., 2019. Integrated analysis of tp53 gene and pathway alterations in the cancer genome atlas. Cell Reports 28, 1370 – 1384.e5. URL: http://www.sciencedirect.com/science/article/pii/S221112471930885X, doi:doi:https://doi.org/10.1016/j.celrep.2019.07.001.

Efe, G., Rustgi, A.K., Prives, C., 2024. p53 at the crossroads of tumor immunity. Nature Cancer 5, 983–995.

Erwin, D.H., Davidson, E.H., 2009. The evolution of hierarchical gene regulatory networks. Nature Reviews Genetics 10, 141–148.

Garcia-Alonso, L., Holland, C.H., Ibrahim, M.M., Turei, D., Saez-Rodriguez, J., 2019. Benchmark and integration of resources for the estimation of human transcription factor activities. Genome research 29, 1363–1375.

Gilpin, L.H., Bau, D., Yuan, B.Z., Bajwa, A., Specter, M., Kagal, L., 2018. Explaining explanations: An overview of interpretability of machine learning, in: 2018 IEEE 5th International Conference on data science and advanced analytics (DSAA), IEEE. pp. 80–89.

Hamilton, W.L., 2020. Graph representation learning. Morgan & Claypool Publishers.

Hanahan, D., 2022. Hallmarks of cancer: New dimensions. Cancer Discov 12, 31–46.

Hanahan, D., Weinberg, R.A., 2000. The hallmarks of cancer. cell 100, 57–70.

Hanahan, D., Weinberg, R.A., 2011. Hallmarks of cancer: the next generation. cell 144, 646–674.

Hollstein, M., Sidransky, D., Vogelstein, B., Harris, C.C., 1991. p53 mutations in human cancers. Science 253, 49–53.

Joerger, A.C., Stiewe, T., Soussi, T., 2024. Tp53: the unluckiest of genes? Cell Death & Differentiation , 1–6.

Jolliffe, I.T., Cadima, J., 2016. Principal component analysis: a review and recent developments. Philosophical transactions of the royal society A: Mathematical, Physical and Engineering Sciences 374, 20150202.

Karlebach, G., Shamir, R., 2008a. Modelling and analysis of gene regulatory networks. Nature Reviews Molecular Cell Biology 9, 770–780.

Karlebach, G., Shamir, R., 2008b. Modelling and analysis of gene regulatory networks. Nature reviews Molecular cell biology 9, 770–780.

Katuwal, G.J., Chen, R., 2016. Machine learning model interpretability for precision medicine. arXiv preprint arXiv:1610.09045 .

Kipf, T.N., Welling, M., 2016. Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907 .

Kruse, J.P., Gu, W., 2009. Modes of p53 regulation. Cell 137, 609–622.

Lambert, S.A., Jolma, A., Campitelli, L.F., Das, P.K., Yin, Y., Albu, M., Chen, X., Taipale, J., Hughes, T.R., Weirauch, M.T., 2018. The human transcription factors. Cell 172, 650–665.

Lane, D.P., 1992. p53, guardian of the genome. Nature 358.

Le Novère, N., 2015. Quantitative and logic modelling of molecular and gene networks. Nature Reviews Genetics 16, 146–158.

Lee, S.M., Han, Y., Cho, K.H., 2023. Deep learning untangles the resistance mechanism of p53 reactivator in lung cancer cells. iScience 26. URL: https://doi.org/10.1016/j.isci.2023.108377, doi:doi:10.1016/j.isci.2023.108377.

Ling, H., Samarasinghe, S., Kulasiri, D., 2013. Novel recurrent neural network for modelling biological networks: Oscillatory p53 interaction dynamics. Biosystems 114, 191–205. URL: https://www.sciencedirect.com/science/article/pii/S0303264713001925, doi:doi:https://doi.org/10.1016/j.biosystems.2013.08.004.

Liu, A., Trairatphisan, P., Gjerga, E., Didangelos, A., Barratt, J., Saez-Rodriguez, J., 2019. From expression footprints to causal pathways: contextualizing large signaling networks with carnival. npj Systems Biology and Applications 5, 40. URL: https://doi.org/10.1038/s41540-019-0118-z, doi:doi:10.1038/s41540-019-0118-z.

Lujambio, A., Lowe, S.W., 2012. The microcosmos of cancer. Nature 482, 347–355.

May, P., May, E., 1999. Twenty years of p53 research: structural and functional aspects of the p53 protein. Oncogene 18, 7621–7636.

Nesbit, C.E., Tersak, J.M., Prochownik, E.V., 1999. Myc oncogenes and human neoplastic disease. Oncogene 18, 3004–3016.

Olivier, M., Hollstein, M., Hainaut, P., 2010. Tp53 mutations in human cancers: origins, consequences, and clinical use. Cold Spring Harbor perspectives in biology 2, a001008.

Pan, Q., Portelli, S., Nguyen, T.B., Ascher, D.B., 2023. Characterization on the oncogenic effect of the missense mutations of p53 via machine learning. Briefings in Bioinformatics 25, bbad428. URL: https://doi.org/10.1093/bib/bbad428, doi:doi:10.1093/bib/bbad428, arXiv:https://academic.oup.com/bib/article-pdf/25/1/bbad428/53896419/bbad428.pdf.

Papavassiliou, A.G., 1995. Transcription factors. New England Journal of Medicine 332, 45–47.

Pearl, J., Glymour, M., Jewell, N., 2016. Causal Inference in Statistics: A Primer. Wiley. URL: https://books.google.co.uk/books?id=L3G-CgAAQBAJ.

Petitjean, A., Achatz, M., Borresen-Dale, A., Hainaut, P., Olivier, M., 2007. Tp53 mutations in human cancers: functional selection and impact on cancer prognosis and outcomes. Oncogene 26, 2157–2165.

Pratt, H.E., Andrews, G.R., Phalke, N., Purcaro, M.J., van der Velde, A., Moore, J.E., Weng, Z., 2022. Factorbook: an updated catalog of transcription factor motifs and candidate regulatory motif sites. Nucleic Acids Res 50, D141–D149.

Richens, J.G., Lee, C.M., Johri, S., 2020. Improving the accuracy of medical diagnosis with causal machine learning. Nature communications 11, 3923.

Rizzi, D.A., Pedersen, S.A., 1992. Causality in medicine: towards a theory and terminology. Theoretical Medicine 13, 233–254.

Rosvall, M., Bergstrom, C.T., 2008. Maps of random walks on complex networks reveal community structure. Proceedings of the National Academy of Sciences 105, 1118–1123.

Smiljanić, J., Blöcker, C., Holmgren, A., Edler, D., Neuman, M., Rosvall, M., 2023. Community detection with the map equation and infomap: Theory and applications. URL: https://arxiv.org/abs/2311.04036, arXiv:2311.04036.

Smolen, P., Baxter, D.A., Byrne, J.H., 2000. Mathematical modeling of gene networks. Neuron 26, 567–580.

Soussi, T., Wiman, K., 2015. Tp53: an oncogene in disguise. Cell Death & Differentiation 22, 1239–1249.

Tomasetti, C., Li, L., Vogelstein, B., 2017. Stem cell divisions, somatic mutations, cancer etiology, and cancer prevention. Science 355, 1330–1334. URL: https://www.science.org/doi/abs/10.1126/science.aaf9011, doi:doi:10.1126/science.aaf9011, arXiv:https://www.science.org/doi/pdf/10.1126/science.aaf9011.

Triantafyllidis, C.P., Barberis, A., Hartley, F., Cuervo, A.M., Gjerga, E., Charlton, P., van Bijsterveldt, L., Rodriguez, J.S., Buffa, F.M., 2023. A machine learning and directed network optimization approach to uncover tp53 regulatory patterns. iScience 26. URL: https://doi.org/10.1016/j.isci.2023.108291, doi:doi:10.1016/j.isci.2023.108291.

Tsherniak, A., Vazquez, F., Montgomery, P.G., Weir, B.A., Kryukov, G., Cowley, G.S., Gill, S., Harrington, W.F., Pantel, S., Krill-Burger, J.M., et al., 2017. Defining a cancer dependency map. Cell 170, 564–576.

Turei, D., Korcsmaros, T., Saez-Rodriguez, J., 2016. Omnipath: Guidelines and gateway for literature-curated signaling pathway resources. Nature Methods 13, 966–967. doi:doi:10.1038/nmeth.4077.

Vellido, A., 2020. The importance of interpretability and visualization in machine learning for applications in medicine and health care. Neural computing and applications 32, 18069–18083.

Wang, S., Tang, H., Shan, P., Wu, Z., Zuo, L., 2023. Pros-gnn: Predicting effects of mutations on protein stability using graph neural networks. Computational Biology and Chemistry 107, 107952. URL: https://www.sciencedirect.com/science/article/pii/S1476927123001433, doi:doi:https://doi.org/10.1016/j.compbiolchem.2023.107952.

Wilson, W.R., Hay, M.P., 2011. Targeting hypoxia in cancer therapy. Nature Reviews Cancer 11, 393–410.

Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C., Philip, S.Y., 2020. A comprehensive survey on graph neural networks. IEEE transactions on neural networks and learning systems 32, 4–24.

Yusuf, D., Butland, S.L., Swanson, M.I., Bolotin, E., Ticoll, A., Cheung, W.A., Cindy Zhang, X.Y., Dickman, C.T., Fulton, D.L., Lim, J.S., et al., 2012. The transcription factor encyclopedia. Genome biology 13, 1–25.

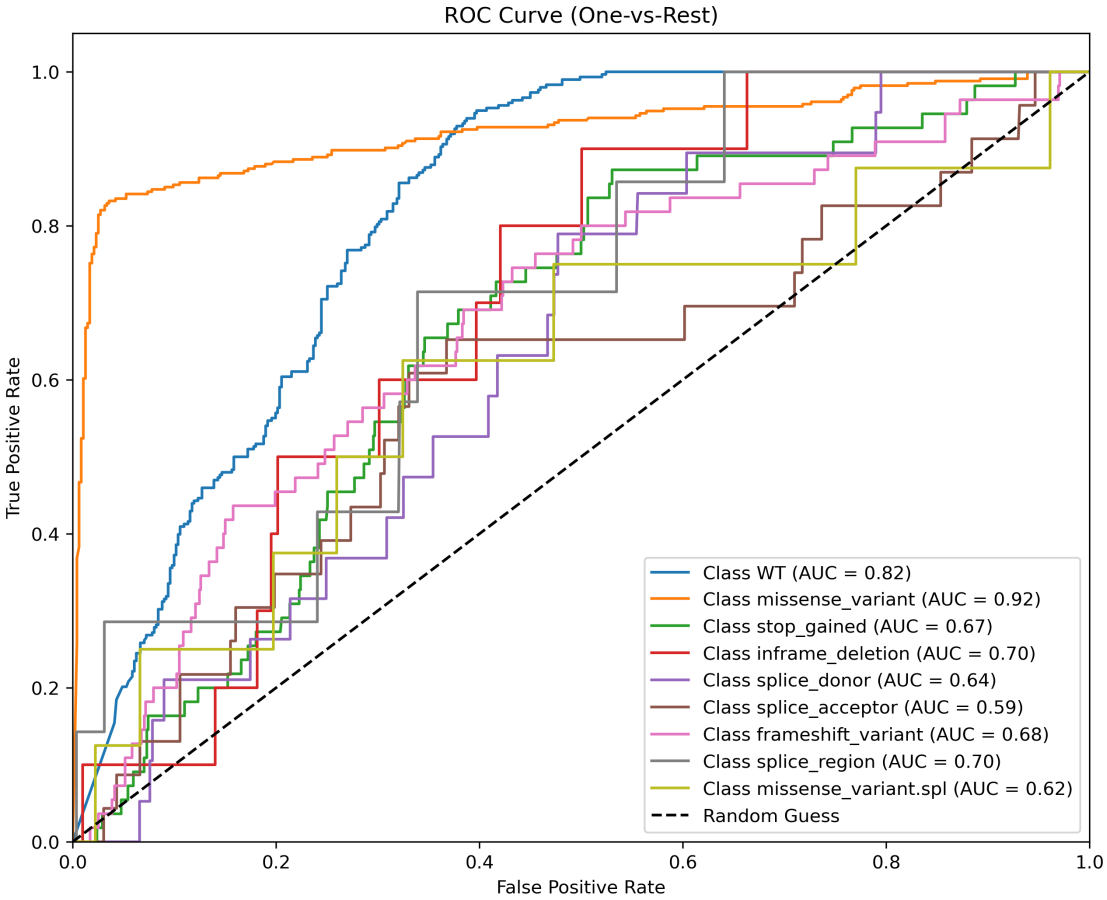

Figure 4: ROC curve for our graph-neural network classifier across all different classes each time (mutation type, or primary/metastatic), as a multinomial and binomial classification plot each time. We can see that the classifier performs very well on the majority of the classes involved, and that lower classification scores are probably due to the lack of training samples, or simply put class imbalance in the data. PCA plot of node embeddings from the Graph - Neural Network architecture used. Here we plot on 2-D , using the first two high variance explanation principal components the constellation of optimized regulatory networks, and annotate them with colours across mutation type. We can see that WT and missense mutations are vastly different in their internal structure which confirms our postulation of the regulon variability based on the variant classification of *TP53* mutation. The plot produced by HarmonizerR is interactive and allows us to further zoom as seen here in each of these data points and see the network behind it with its gene activity profile mapped in colours (light rose: deactivated, blue: active).

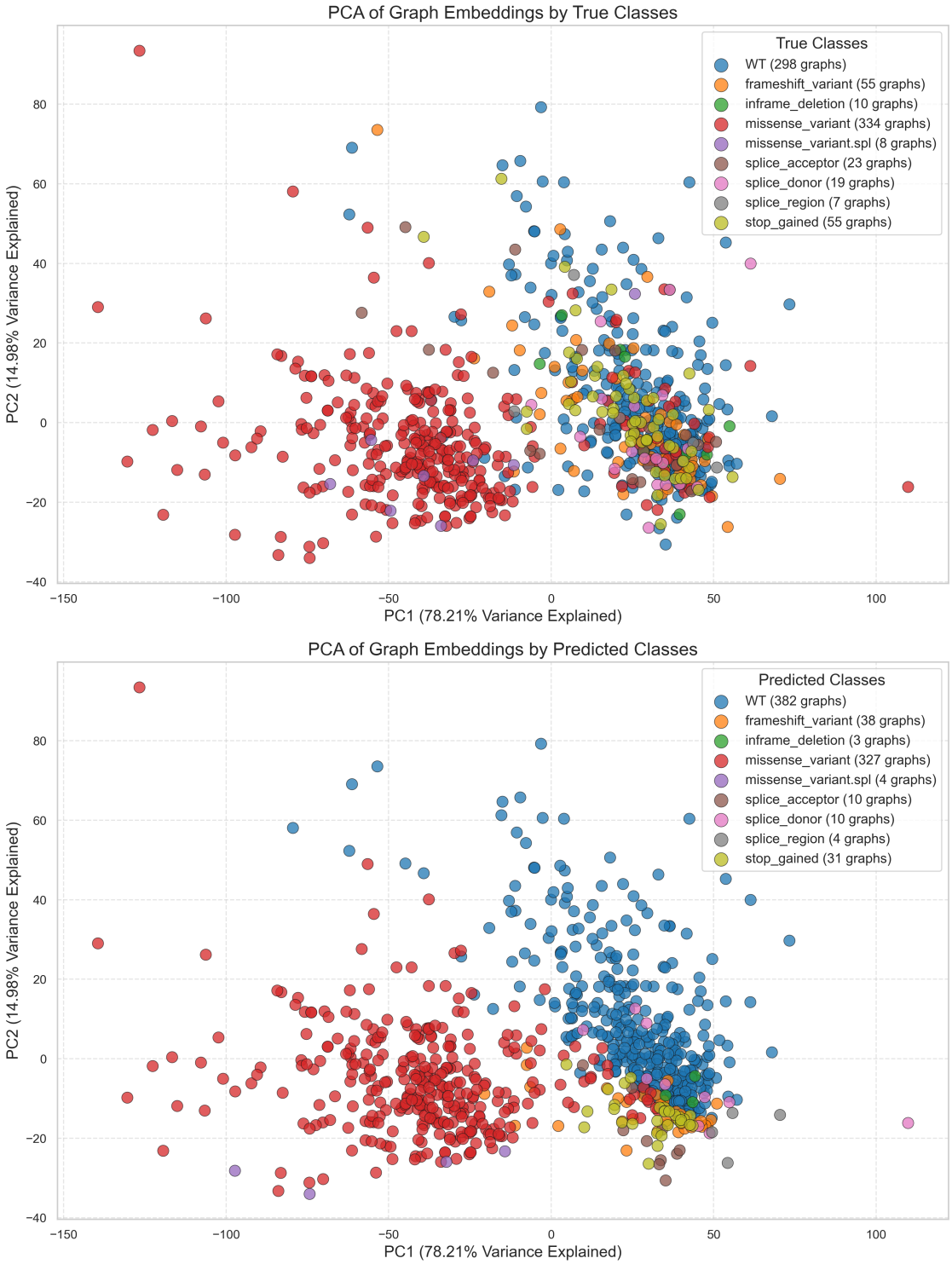

Figure 5: PCA 2-D scatter plots of the node embeddings with different colour annotation for the classified labels, correspondingly. The bottom figure shows the distribution of the embeddings from the graph neural network output as predicted classes and the top shows the true labels of the graphs. High loadings for the first two principal components can be justified by the small number of embeddings initially used (gene status, community id, in-out degrees, mode of regulation).

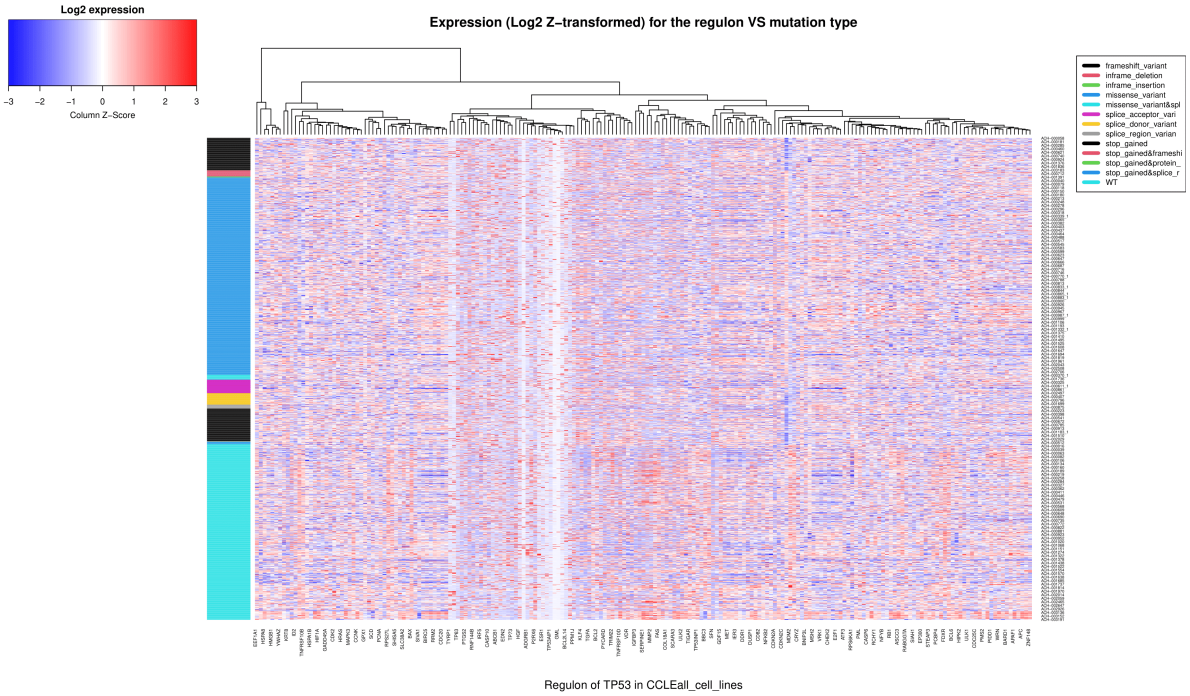

Figure 6: Z-normalized levels of expression for the regulon (set of target genes) of transcription factor *TP53*, as columns is depicted in this heatmap plot and samples (i.e. unique mutation of *TP53* per sample as rows. We stratify the samples/rows in the *TP53* mutation type using a coloured legend. We can see that missense mutations are prevalent, but also equally dominant are WT samples. Approximately 1600 samples are included in this figure, pancancer from CCLE 24Q2.

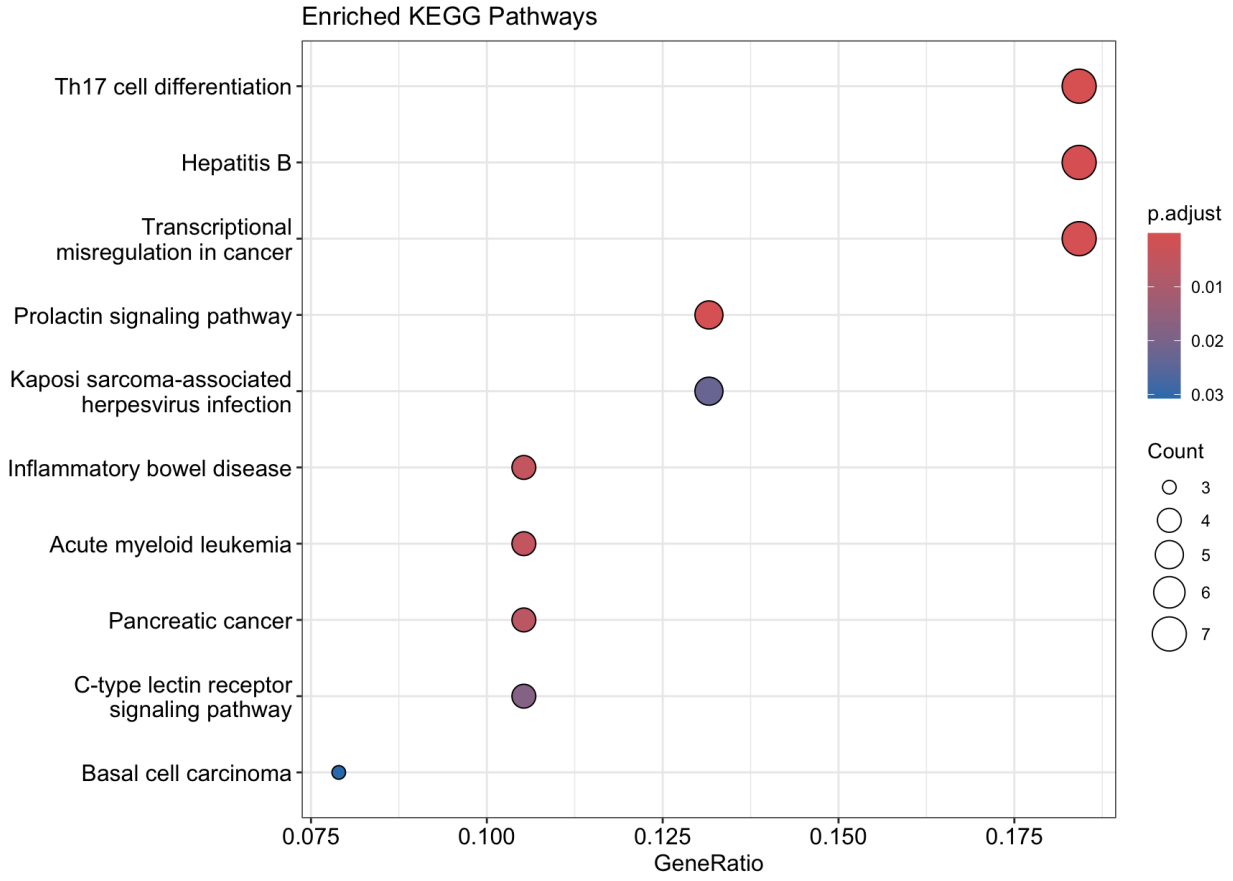

Figure 7: Enriched KEGG ontology analysis on the set difference between the WT signature and the Missense mutation signature reveals the top five major pathways involved. These signatures were derived from combined community detection and central gens per community applied on the optimized networks.

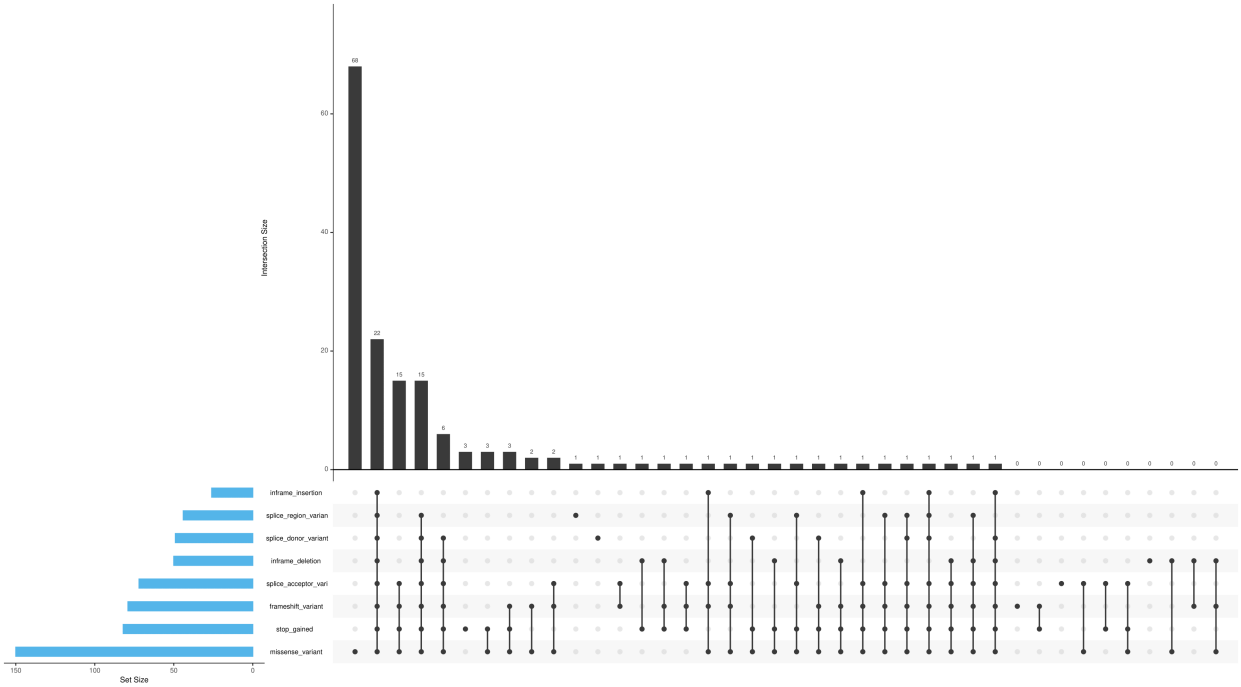

Figure 8: Description: The UpSet plot shows intersections between different mutation types, summarizing the overlap of mutations across categories. Components: Bottom Left: Horizontal bar plot shows the set size (number of mutations in each mutation type). For instance, "missense variant" is the largest set. Top Panel: Vertical bars indicate the intersection size—how many genes are shared among specific mutation combinations. Matrix (Bottom): Black dots and lines indicate which sets (mutation types) intersect. Example: A single black dot corresponds to genes unique to a specific type, while connected dots represent overlaps. Key Insights: "Missense variant" is the most frequent mutation category. Smaller intersections highlight rare combinations of mutations shared between multiple categories.

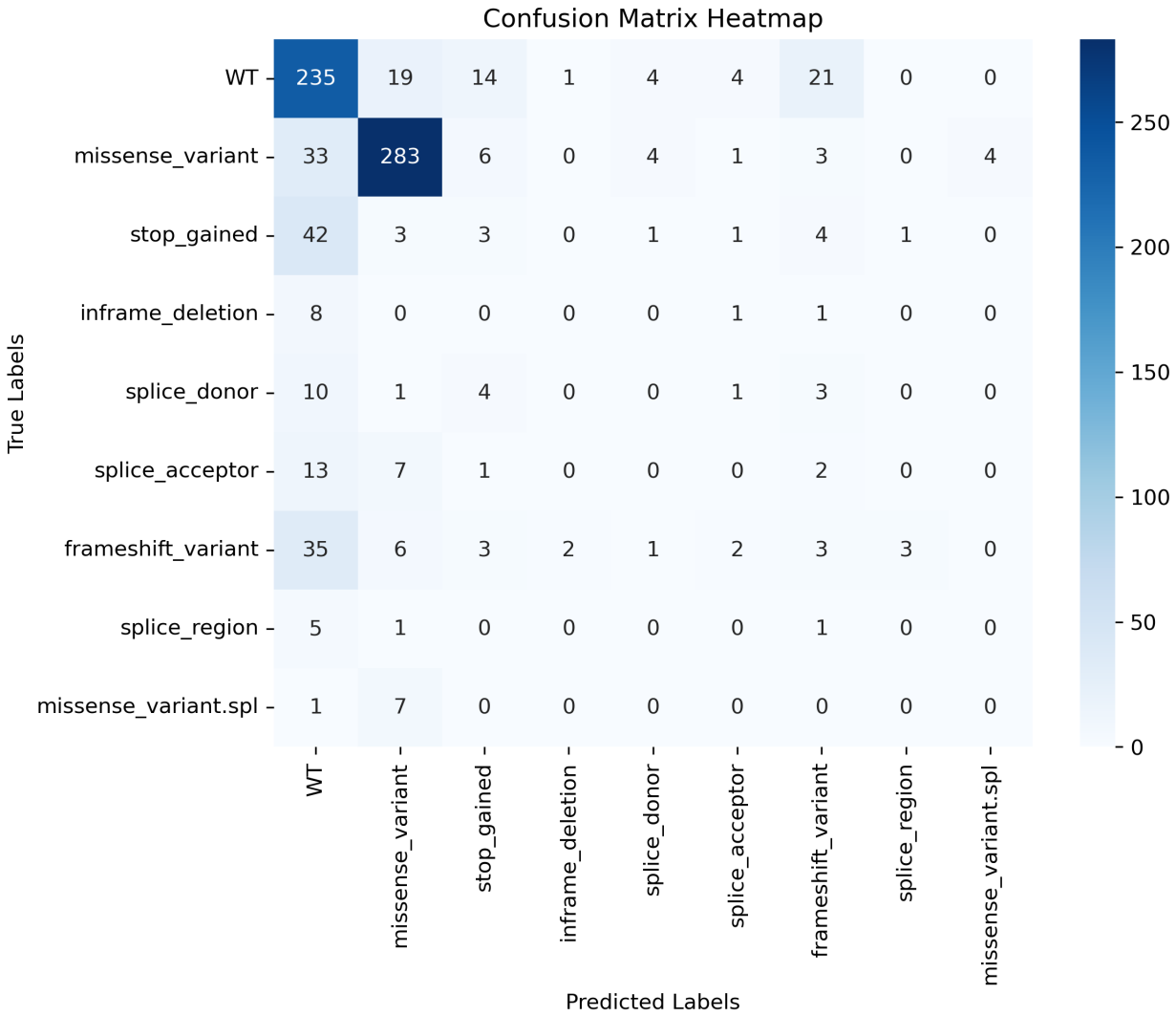

Figure 9: Confusion Matrix for the multi-label classifier built on Graph Convolutional Networks (GCNs). The classifier, as expected given the class imbalance and the availability of different types of mutations for *TP53* performs very well for missense and WT samples (the most prevalent ones, as also seen in 6) and poorly for other types such as `stop_gained`. However, the essence of the classification here is to provide evidence that GCNs can be used to classify pathways annotated by as specific event of interest, such as mutation types here.

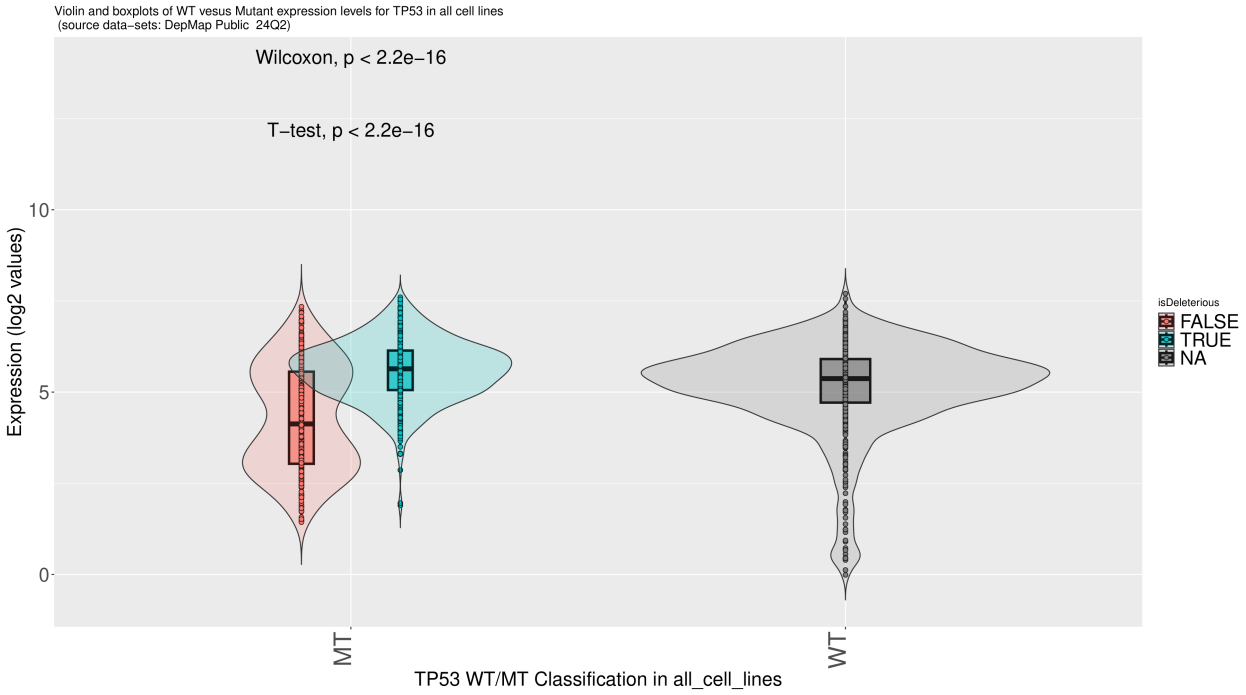

Figure 10: Deleterious function of mutation stratified across mutated samples and contrastede with controls (WT). We observe a statistically significant difference in the means of log2 normalized expression for *TP53* for the deleterious mutations against non-deleterious, and higher than WT expression.

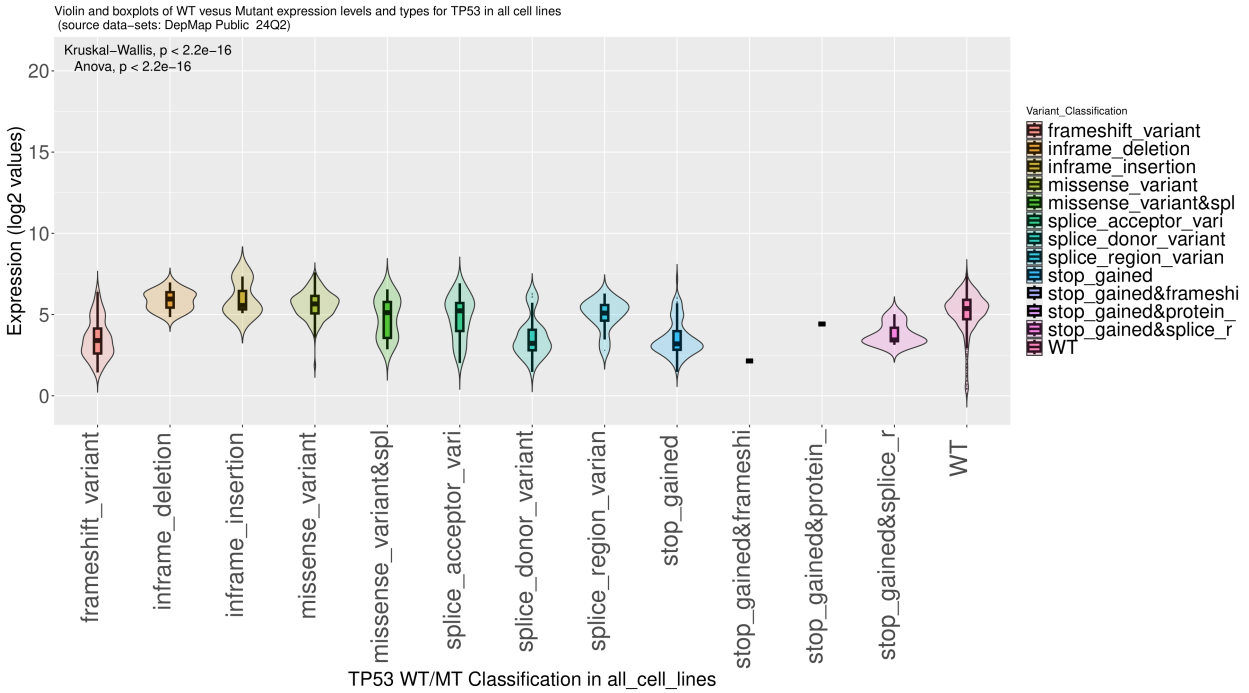

Figure 11: Description: A violin and boxplot analysis compares the expression levels of the TP53 gene across different mutation types in various cell lines. Axes: Y-axis: Expression values (log2 scale). X-axis: Mutation classifications (e.g., frameshift variant, missense variant, splice variants, WT). Key Observations: WT (wild-type) TP53 has a lower variance in expression. Mutant categories (e.g., "frameshift variant," "splice donor variant") show differing expression patterns. The Kruskal-Wallis and ANOVA tests indicate a highly significant difference in expression values across groups ($p < 2.2e-16$). Purpose: This plot helps assess how specific TP53 mutations influence gene expression and highlights significant expression differences between WT and mutant TP53 groups.