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 under various conditions of interest. Our approach combines three distinct yet seamlessly integrated modelling schemes: i) we first leverage Prior-Knowledge Networks (PKNs) derived from comprehensive and established databases to reconstruct their topology using genomic and transcriptomic data via mathematical programming optimization, ii) we apply causal learning via Additive Noise Models (ANMs) to further prune the optimized networks, and iii) 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. We demonstrate the framework in the DNA damage and repair pathway using the *TP53* regulon in a pancancer study, classifying Gene Regulatory Networks (GRNs) across different *TP53* mutation types. This scalable approach enables the classification of diverse conditions while addressing the multifactorial nature of diseases. It disentangles their polygenic complexity and reveals new functional patterns through a causal representation.

Keywords Graph Neural Networks, Optimization, Mathematical Programming, Causal Learning, Gene Networks, *TP53*

Introduction

The challenge of causal interpretation in medical sciences [45, 2, 44] 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 [25, 15, 8, 55]. 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* [48, 6, 5, 23, 1, 30]. 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 bi-directional way but mainly 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

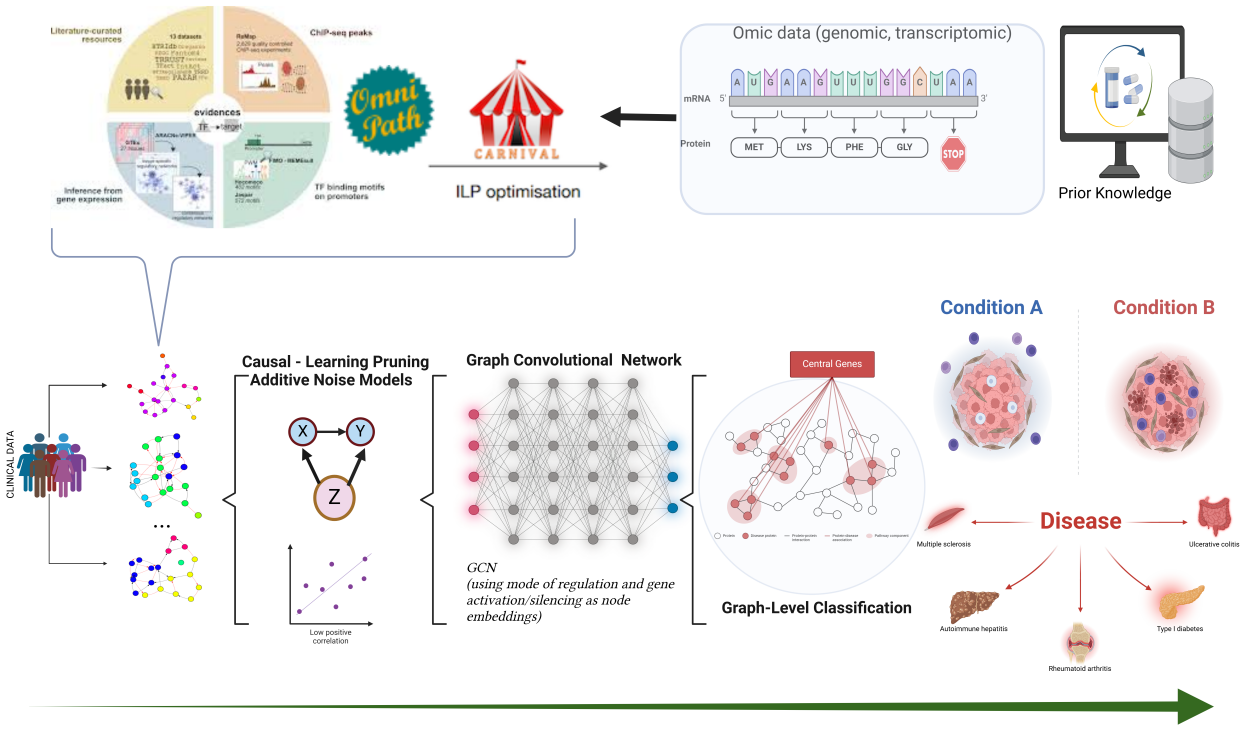

Figure 1: The developed framework workflow. Prior Knowledge Networks (PKNs) and omics-data are combined to provide the input for optimization models to reconstruct gene regulatory networks across pathways of interest, in this case the *TP53* pathway and its regulon. These optimized networks are then annotated across conditions of interest, 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 use ANMs to prune edges that represent no causal effect so that we can then classify the optimized networks using GCNs, leveraging enhanced structural learning. We use 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.).

interactions is rooted in its ability to provide causal interpretation of the shared effects [41] 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) [24, 12], 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 [4] 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.

When using the Prior Knowledge Networks (PKNs) from established sources, multiple levels of confidence can be used (from very stringent to less stringent, A,B,C,D,E), symbolizing how confident we are that the target genes used are actual targets or not. By using the less stringent approach, to include as many (initially) genes as possible, we then seek to minimize the number of edges in our network topology by using causal learning constraint methods, such as ANMs [60, 21]. These methods work independently and derive pruning based on simply the input of the full expression matrix and the network itself. This adds an additional layer of confidence that our learned GRNs have causal structural topologies. These pruned networks now serve as input to the subsequent deep learning module.

This modelling scheme based on GCNs [26, 9, 16, 58], represents 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. These models have become a key approach in analysing graph-structured data, enabling the extraction of meaningful patterns from complex networks. In [31] deep learning has been successfully used to identify resistance to therapy p53 pathway mediated mechanisms and related combinatorial targets for lung cancer. In [32], a Recurrent Neural Network (RNN) has been employed to illuminate p53 temporal interaction dynamics. Furthermore, a Graph Neural Network approach has been utilized in [56] 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 [17, 34, 18, 19]. As aberrant gene expression drives oncogenic processes and initiates a multitude of neoplastic diseases [36], the master regulators of the genome, known as Transcription Factors (TF) [28, 59, 40], dictate the gene activity profile of many downstream targets, known as their *regulon*. Across these approximately 1,600 DNA-binding TFs [43], 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 [57, 7] or drug effects and other stimuli, are pivotal in determining a cascade of processes underlying disease onset, progression and prognosis. Particularly somatic mutations [50], 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* [49, 38, 42] stands out due to its widespread impact in multiple types of cancer [20, 35], influencing crucial pathways for cell cycle control, apoptosis, and DNA repair, thus widely known as the *Guardian of the Genome* [29].

Somatic mutations in *TP53*, manifest as heterogeneous effects on downstream gene expression [10, 27] 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 [51], 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 [22].

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 Wild-Type (WT) status 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 [11]. 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 [11]. Additionally, further investigation of *TP53* mutations reveals their impact on tumour suppressing ability, and shows that [39] 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 [3], 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) [53] 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 [14] regulon was extracted for this transcription factor using OmniPathR [54] and accounting only for signed and directed edges. This amounted to approximately 200 downstream targets with all confidence levels included, to remain less stringent 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 [33] 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 heterogeneity found in the expression profile of the regulon, and although this has been previously explored [51] in previous versions of the CCLE dataset, we re-iterated for consistency. 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 then stratified 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 7. 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 well known to be observable only in the gene activity profile of the regulon, not in its own expression level.

We utilize the deleterious function of *TP53* mutations to model 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.

Optimized network analysis reveals functional differences across condition of interest

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. 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 learning shows classifiable at graph-level networks based on condition of interest (mutation type)

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 7, 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 optimized network corresponding to this data point, with its graph representation, showing both the directionality of interactions and the activity profile (indicating gene activity status, with blue denoting activation and red denoting inactivity).

Discussion

This study combines 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 response to DNA 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

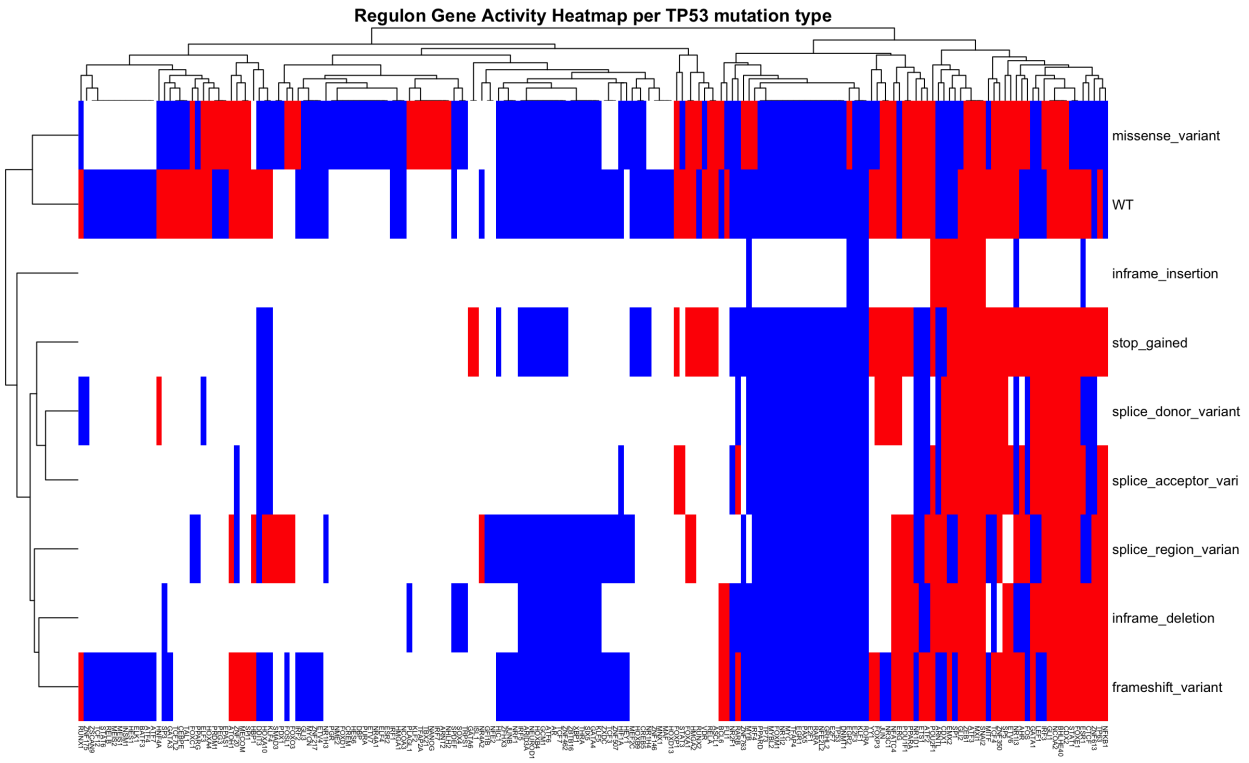

Figure 2: Plotting the gene signatures extracted from the community detection analysis across each mutation type and WT, and showing gene activity each time. All networks carrying the same *TP53* mutation were stratified into groups. Community detection was performed for each group, and a centrality (maximum) score was used to identify one representative gene per community. The overlap of these sets across the 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. Colour Scheme: Blue: Indicates upregulated genes. Red: Indicates downregulated genes. White: Represents absence.

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.

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

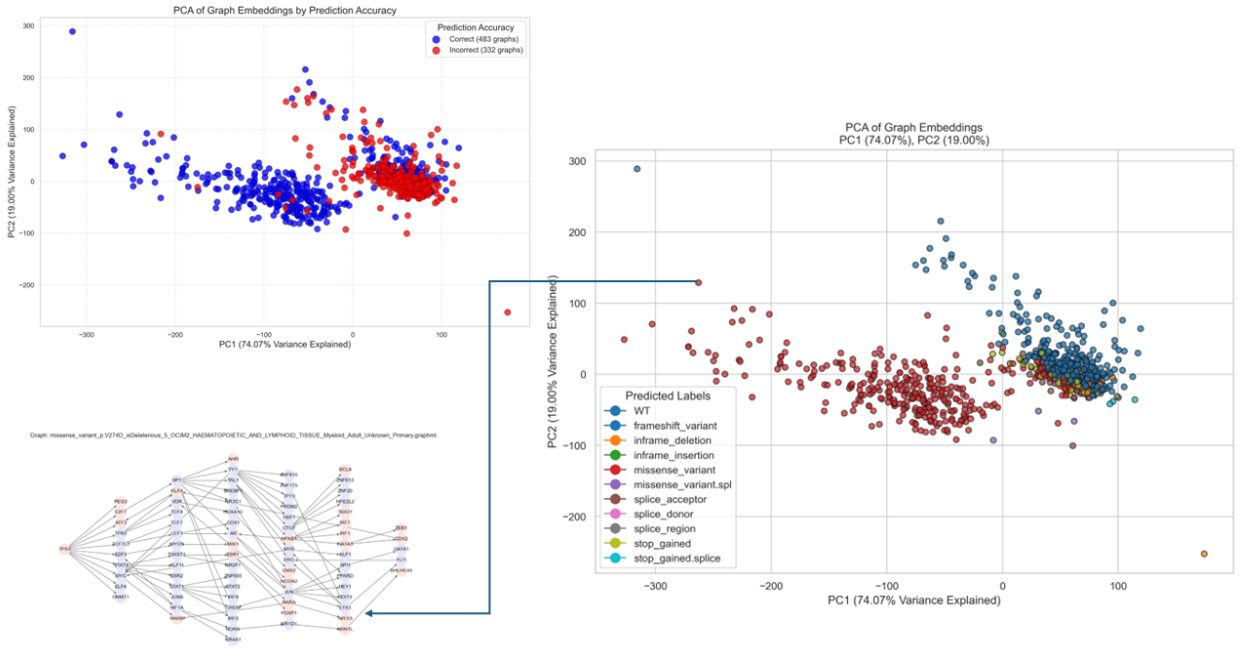

Figure 3: **Top left:** PCA scatter plot of the node embeddings depict prediction accuracy, with correctly classified graphs in blue and misclassified graphs in red. **Right:** PCA scatter plots of the node embeddings predictions with different colour annotation for the classified labels. The graph is 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.

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.

While our study demonstrates the potential of integrating mathematical programming and GCNs in pathway analysis, several limitations must be considered. 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 mutations, 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 ($\sim$ 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 [47, 46] 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 [51] not only in the community detection algorithm used, 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.

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. GRN reconstruction was formulated as a mixed-integer linear programming (MILP) problem as defined in [33]. Mixed-Integer Linear Programming (MILP) is an NP-hard special case of the general Linear Programming (LP) model [52], where one imposes integrality constraints for a subset of the variables:

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

where $\mathcal{A} \in \mathbb{R}^{m \times n}$, $(c, x) \in \mathbb{R}^n$, $b \in \mathbb{R}^m$, T denotes transposition and $x = \begin{bmatrix} x_{\mathcal{I}} \\ x_{\mathcal{J}} \end{bmatrix}$. 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 [33]. A sparsity constraint was imposed to limit the number of edges, ensuring biologically plausible networks.Constraints ensured that reconstructed edges aligned with known TF-target interactions. The MILP problem was solved using commercial solvers (IBM CPLEX).

To further refine the reconstructed networks, we employ ANMs to identify and remove non-causal edges. Given one node causally influences another, this influence can be modelled with an additive noise term. For a causal edge, the expression levels of the target are modelled as a function of the input plus some noise. If the noise is not independent of the input, it suggests the edge can be removed. We fit a simple linear regression model $Y = ax + b$, calculate the residuals $Y - \hat{Y}$, where $\hat{Y}$ is the predicted value from the regression and then calculate the correlation coefficient between the residuals and X. If the absolute correlation is greater than a threshold, the edge is removed.

Community Detection to extract mutation type gene signatures

Community detection [37] methods aim to disentangle the complexity and topology of networks, giving insights to tightly connected communities or *hubs*. Biologically, these hubs may give insight to phenotypes, identifying at the same time key node (gene) players in the network. To identify mutation-specific differences in the reconstructed regulatory networks, community detection analyses were performed using the *Edge Betweenness* clustering algorithm. Each identified hub was assessed for its key gene(s) in terms of (max) betweenness centrality score. All top centrality genes (per hub) with their associated state (up/down) were collected and intersected across all networks annotated as carrying the same *TP53* mutation type, thus generating a unique (master) signature per mutation type. Figure 2 shows the heatmap visualization of these master signatures and figure 6 shows both the optimized reconstructed GRN (B) for a given sample and its dissected community detection version (A). These signatures not only possess functional implications but also provide a way to summarize the topology of the networks carrying the same *TP53* mutation type.

Graph Convolutional Network Model to learn entire GRNs

We implemented GCNs to classify graph data based on node features and structural information. The model was developed using the PyTorch Geometric framework [13] 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 $\mathbf{X} \in \mathbb{R}^{|V| \times d_{\text{in}}}$, 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 `0.005`. 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.

Declaration of interests

The authors declare no competing interests.

References

[1] Ernst Althaus, Gunnar W. Klau, Oliver Kohlbacher, Hans-Peter Lenhof, and Knut Reinert. *Integer Linear Programming in Computational Biology*, pages 199–218. Springer Berlin Heidelberg, Berlin, Heidelberg, 2009.

[2] Jean-François Bach. Causality in medicine. *Comptes Rendus. Biologies*, 342(3-4):55–57, 2019.

[3] Gil Ben-Cohen, Flora Doffe, Michal Devir, Bernard Leroy, Thierry Soussi, and Shai Rosenberg. Tp53_prof: a machine learning model to predict impact of missense mutations in tp53. *Briefings in Bioinformatics*, 23(2):bbab524, 01 2022.

[4] D. Bertsimas and J.N. Tsitsiklis. *Introduction to linear optimization*. Athena Scientific, 1997.

[5] Richard Bonneau. Learning biological networks: from modules to dynamics. *Nature Chemical Biology*, 4(11):658–664, November 2008.

[6] Paul Brazhnik, Alberto de la Fuente, and Pedro Mendes. Gene networks: how to put the function in genomics. *Trends in Biotechnology*, 20(11):467–472, November 2002.

[7] J Martin Brown and William R Wilson. Exploiting tumour hypoxia in cancer treatment. *Nature Reviews Cancer*, 4(6):437–447, 2004.

[8] Diogo V Carvalho, Eduardo M Pereira, and Jaime S Cardoso. Machine learning interpretability: A survey on methods and metrics. *Electronics*, 8(8):832, 2019.

[9] Gabriele Corso, Hannes Stark, Stefanie Jegelka, Tommi Jaakkola, and Regina Barzilay. Graph neural networks. *Nature Reviews Methods Primers*, 4(1):17, 2024.

[10] Lawrence A. Donehower, Thierry Soussi, Anil Korkut, Yuexin Liu, Andre Schultz, Maria Cardenas, Xubin Li, Ozgun Babur, Teng-Kuei Hsu, Olivier Lichtarge, John N. Weinstein, Rehan Akbani, and David A. Wheeler. Integrated analysis of tp53 gene and pathway alterations in the cancer genome atlas. *Cell Reports*, 28(5):1370 – 1384.e5, 2019.

[11] Gizem Efe, Anil K Rustgi, and Carol Prives. p53 at the crossroads of tumor immunity. *Nature Cancer*, 5(7):983–995, July 2024.

[12] Douglas H Erwin and Eric H Davidson. The evolution of hierarchical gene regulatory networks. *Nature Reviews Genetics*, 10(2):141–148, 2009.

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

[14] Luz Garcia-Alonso, Christian H Holland, Mahmoud M Ibrahim, Denes Turei, and Julio Saez-Rodriguez. Benchmark and integration of resources for the estimation of human transcription factor activities. *Genome research*, 29(8):1363–1375, 2019.

[15] Leilani H Gilpin, David Bau, Ben Z Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. Explaining explanations: An overview of interpretability of machine learning. In *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*, pages 80–89. IEEE, 2018.

[16] William L Hamilton. *Graph representation learning*. Morgan & Claypool Publishers, 2020.

[17] Douglas Hanahan. Hallmarks of cancer: New dimensions. *Cancer Discov*, 12(1):31–46, January 2022.

[18] Douglas Hanahan and Robert A Weinberg. The hallmarks of cancer. *cell*, 100(1):57–70, 2000.

[19] Douglas Hanahan and Robert A Weinberg. Hallmarks of cancer: the next generation. *cell*, 144(5):646–674, 2011.

[20] Monica Hollstein, David Sidransky, Bert Vogelstein, and Curtis C Harris. p53 mutations in human cancers. *Science*, 253(5015):49–53, 1991.

[21] Patrik Hoyer, Dominik Janzing, Joris M Mooij, Jonas Peters, and Bernhard Schölkopf. Nonlinear causal discovery with additive noise models. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems*, volume 21. Curran Associates, Inc., 2008.

[22] Andreas C Joerger, Thorsten Stiewe, and Thierry Soussi. Tp53: the unluckiest of genes? *Cell Death & Differentiation*, pages 1–6, 2024.

[23] Guy Karlebach and Ron Shamir. Modelling and analysis of gene regulatory networks. *Nature Reviews Molecular Cell Biology*, 9(10):770–780, October 2008.

[24] Guy Karlebach and Ron Shamir. Modelling and analysis of gene regulatory networks. *Nature reviews Molecular cell biology*, 9(10):770–780, 2008.

[25] Gajendra Jung Katuwal and Robert Chen. Machine learning model interpretability for precision medicine. *arXiv preprint arXiv:1610.09045*, 2016.

[26] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

[27] Jan-Philipp Kruse and Wei Gu. Modes of p53 regulation. *Cell*, 137(4):609–622, 2009.

[28] Samuel A Lambert, Arttu Jolma, Laura F Campitelli, Pratyush K Das, Yimeng Yin, Mihai Albu, Xiaoting Chen, Jussi Taipale, Timothy R Hughes, and Matthew T Weirauch. The human transcription factors. *Cell*, 172(4):650–665, 2018.

[29] David P Lane. p53, guardian of the genome. *Nature*, 358(6381), 1992.

[30] Nicolas Le Novère. Quantitative and logic modelling of molecular and gene networks. *Nature Reviews Genetics*, 16(3):146–158, March 2015.

[31] Soo Min Lee, Younghyun Han, and Kwang-Hyun Cho. Deep learning untangles the resistance mechanism of p53 reactivator in lung cancer cells. *iScience*, 26(12), Dec 2023.

[32] Hong Ling, Sandhya Samarasinghe, and Don Kulasiri. Novel recurrent neural network for modelling biological networks: Oscillatory p53 interaction dynamics. *Biosystems*, 114(3):191–205, 2013.

[33] Anika Liu, Panuwat Trairatphisan, Enio Gjerga, Athanasios Didangelos, Jonathan Barratt, and Julio Saez-Rodriguez. From expression footprints to causal pathways: contextualizing large signaling networks with carnival. *npj Systems Biology and Applications*, 5(1):40, 2019.

[34] Amaia Lujambio and Scott W Lowe. The microcosmos of cancer. *Nature*, 482(7385):347–355, 2012.

[35] Pierre May and Evelyne May. Twenty years of p53 research: structural and functional aspects of the p53 protein. *Oncogene*, 18(53):7621–7636, 1999.

[36] Chadd E Nesbit, Jean M Tersak, and Edward V Prochownik. Myc oncogenes and human neoplastic disease. *Oncogene*, 18(19):3004–3016, 1999.

[37] M. E. J. Newman and M. Girvan. Finding and evaluating community structure in networks. *Phys. Rev. E*, 69:026113, Feb 2004.

[38] Magali Olivier, Monica Hollstein, and Pierre Hainaut. Tp53 mutations in human cancers: origins, consequences, and clinical use. *Cold Spring Harbor perspectives in biology*, 2(1):a001008, 2010.

[39] Qisheng Pan, Stephanie Portelli, Thanh Binh Nguyen, and David B Ascher. Characterization on the oncogenic effect of the missense mutations of p53 via machine learning. *Briefings in Bioinformatics*, 25(1):bbad428, 11 2023.

[40] Athanasios G Papavassiliou. Transcription factors. *New England Journal of Medicine*, 332(1):45–47, 1995.

[41] J. Pearl, M. Glymour, and N.P. Jewell. *Causal Inference in Statistics: A Primer*. Wiley, 2016.

[42] Achatz Petitjean, MIW Achatz, AL Borresen-Dale, P Hainaut, and M Olivier. Tp53 mutations in human cancers: functional selection and impact on cancer prognosis and outcomes. *Oncogene*, 26(15):2157–2165, 2007.

[43] Henry E Pratt, Gregory R Andrews, Nishigandha Phalke, Michael J Purcaro, Arjan van der Velde, Jill E Moore, and Zhiping Weng. Factorbook: an updated catalog of transcription factor motifs and candidate regulatory motif sites. *Nucleic Acids Res*, 50(D1):D141–D149, January 2022.

[44] Jonathan G Richens, Ciarán M Lee, and Saurabh Johri. Improving the accuracy of medical diagnosis with causal machine learning. *Nature communications*, 11(1):3923, 2020.

[45] Dominick A Rizzi and Stig Andur Pedersen. Causality in medicine: towards a theory and terminology. *Theoretical Medicine*, 13:233–254, 1992.

[46] Martin Rosvall and Carl T. Bergstrom. Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences*, 105(4):1118–1123, 2008.

[47] Jelena Smiljanić, Christopher Blöcker, Anton Holmgren, Daniel Edler, Magnus Neuman, and Martin Rosvall. Community detection with the map equation and infomap: Theory and applications, 2023.

[48] Paul Smolen, Douglas A Baxter, and John H Byrne. Mathematical modeling of gene networks. *Neuron*, 26(3):567–580, June 2000.

[49] T Soussi and KG Wiman. Tp53: an oncogene in disguise. *Cell Death & Differentiation*, 22(8):1239–1249, 2015.

[50] Cristian Tomasetti, Lu Li, and Bert Vogelstein. Stem cell divisions, somatic mutations, cancer etiology, and cancer prevention. *Science*, 355(6331):1330–1334, 2017.

[51] Charalampos P. Triantafyllidis, Alessandro Barberis, Fiona Hartley, Ana Miar Cuervo, Enio Gjerga, Philip Charlton, Linda van Bijsterveldt, Julio Saez Rodriguez, and Francesca M. Buffa. A machine learning and directed network optimization approach to uncover tp53 regulatory patterns. *iScience*, 26(12), Dec 2023.

[52] Charalampos P Triantafyllidis and Nikolaos Samaras. A new non-monotonic infeasible simplex-type algorithm for linear programming. *PeerJ Comput Sci*, 6:e265, March 2020.

[53] Aviad Tsherniak, Francisca Vazquez, Phil G Montgomery, Barbara A Weir, Gregory Kryukov, Glenn S Cowley, Stanley Gill, William F Harrington, Sasha Pantel, John M Krill-Burger, et al. Defining a cancer dependency map. *Cell*, 170(3):564–576, 2017.

[54] Denes Turei, Tamas Korcsmaros, and Julio Saez-Rodriguez. Omnipath: Guidelines and gateway for literature-curated signaling pathway resources. *Nature Methods*, 13:966–967, 11 2016.

[55] Alfredo Vellido. The importance of interpretability and visualization in machine learning for applications in medicine and health care. *Neural computing and applications*, 32(24):18069–18083, 2020.

[56] Shuyu Wang, Hongzhou Tang, Peng Shan, Zhaoxia Wu, and Lei Zuo. Pros-gnn: Predicting effects of mutations on protein stability using graph neural networks. *Computational Biology and Chemistry*, 107:107952, 2023.

[57] William R Wilson and Michael P Hay. Targeting hypoxia in cancer therapy. *Nature Reviews Cancer*, 11(6):393–410, 2011.

[58] Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and S Yu Philip. A comprehensive survey on graph neural networks. *IEEE transactions on neural networks and learning systems*, 32(1):4–24, 2020.

[59] Dimas Yusuf, Stefanie L Butland, Magdalena I Swanson, Eugene Bolotin, Amy Ticoll, Warren A Cheung, Xiao Yu Cindy Zhang, Christopher TD Dickman, Debra L Fulton, Jonathan S Lim, et al. The transcription factor encyclopedia. *Genome biology*, 13:1–25, 2012.

[60] Yujia Zheng, Biwei Huang, Wei Chen, Joseph Ramsey, Mingming Gong, Ruichu Cai, Shohei Shimizu, Peter Spirtes, and Kun Zhang. Causal-learn: Causal discovery in python. *Journal of Machine Learning Research*, 25(60):1–8, 2024.

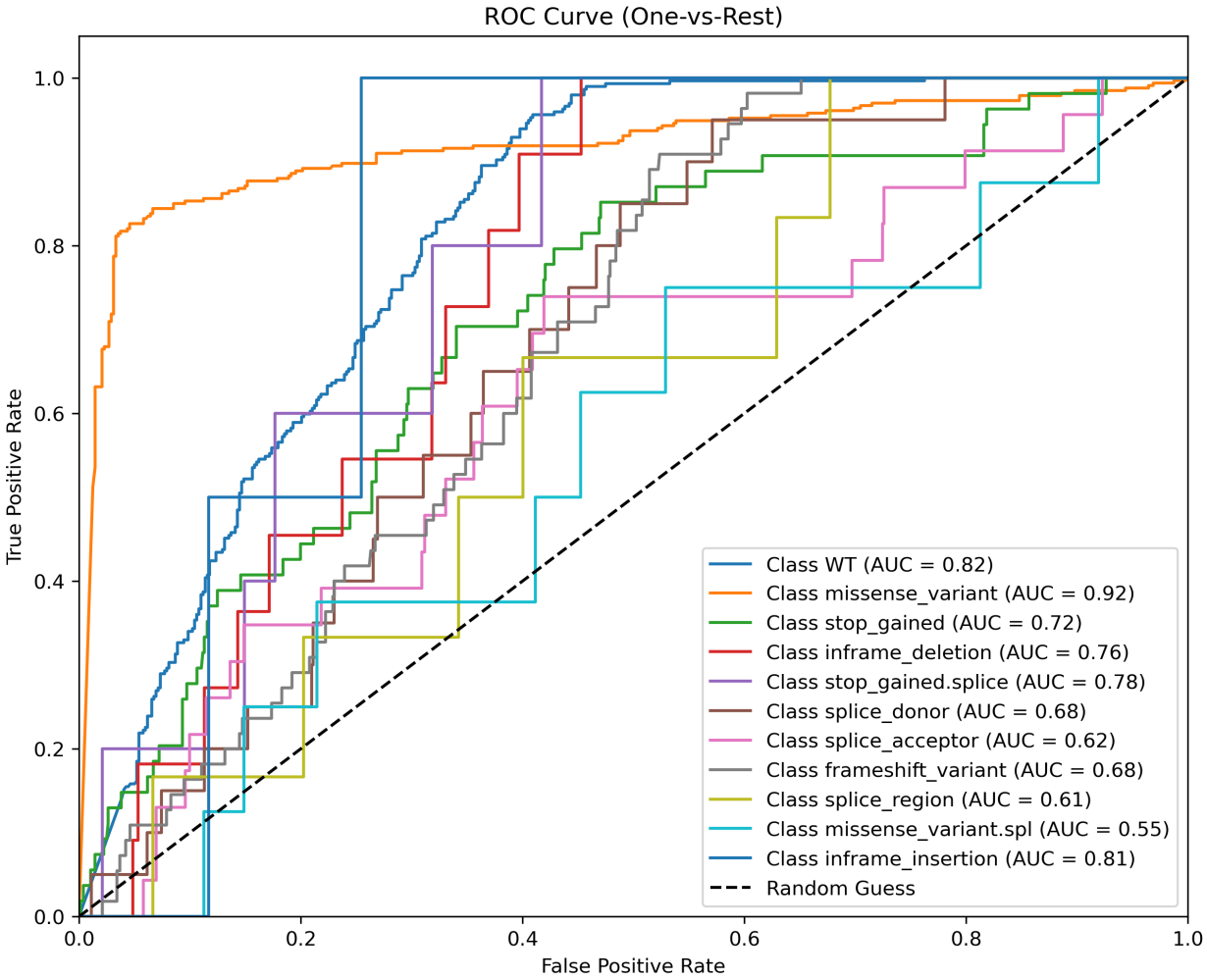

Figure 4: ROC curve for our Graph-Neural network classifier across all different classes each time. We can see that the classifier performs very well on the majority of the classes involved, specifically the WT and missense samples in abundance. The lower classification performance is attributed to the limited number of training samples for the less prevalent classes.

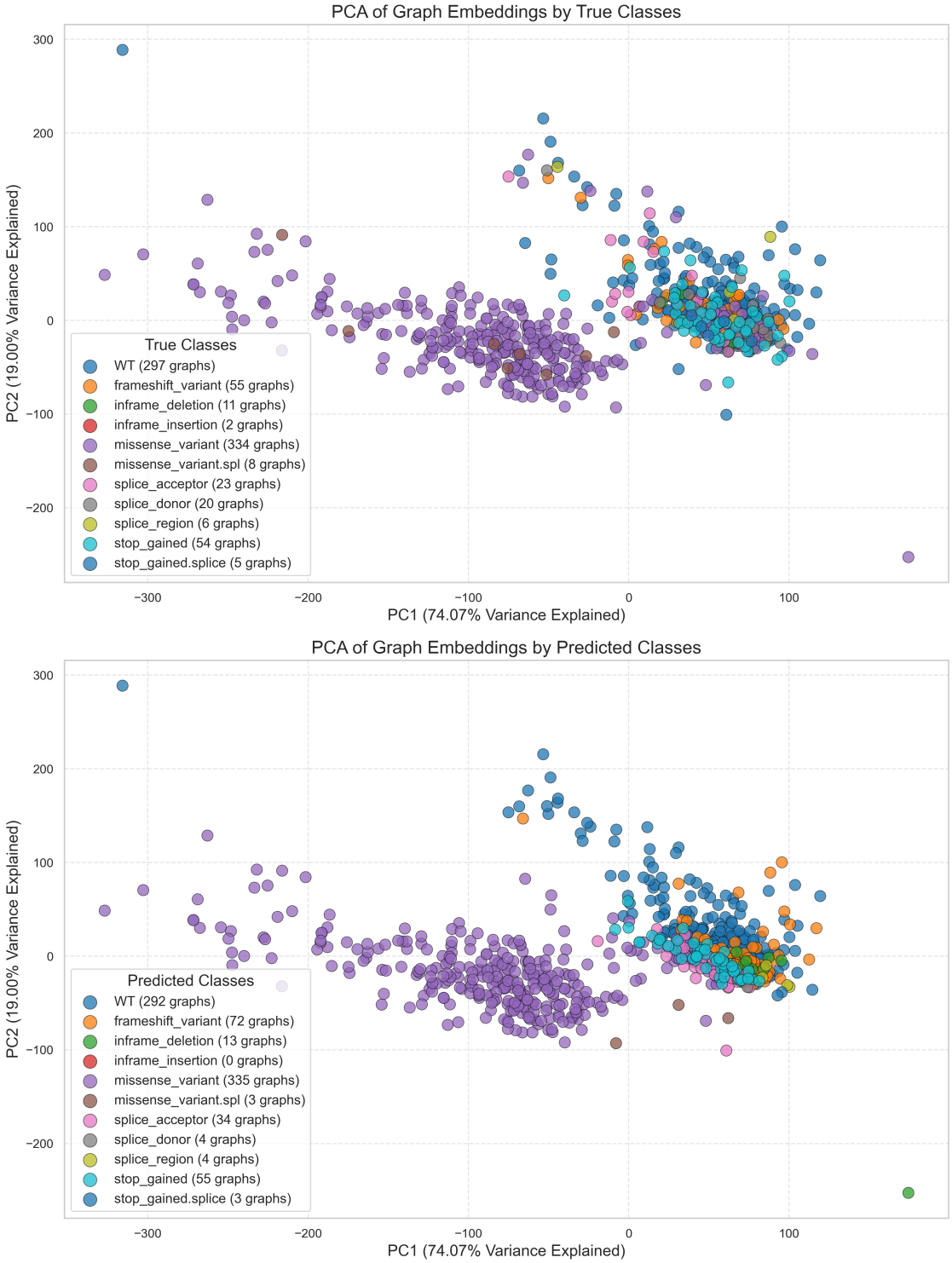

Figure 5: PCA 2-D scatter plots of the node embeddings with different colour annotation for the classified labels, correspondingly. The bottom figure shows the distribution of the embeddings from the graph neural network output as predicted classes and the top shows the true labels of the graphs. High 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).

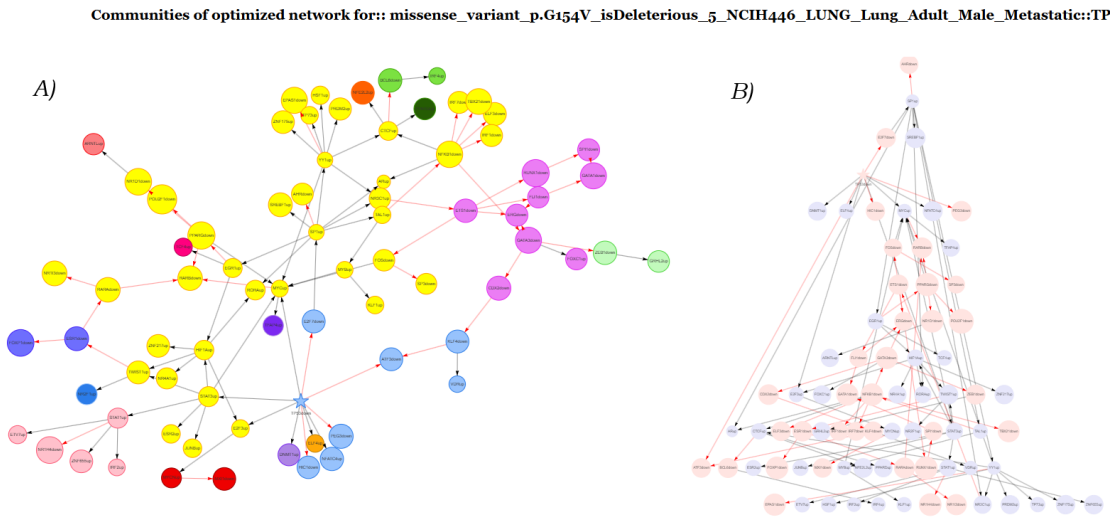

Figure 6: Community detection edge betweenness centrality (A-left) applied to the optimized GRN (B-right) to dissect the graph to hubs denoted by different colour. The annotation (top label) carries all molecular and mutation information to characterize this sample/network, such that it was a missense mutation in lung cancer from an adult male with deleterious effect. The GRN shows the transcription factor (star) connected with its regulon and genes are either up-regulated (*Lavender*) or down-regulated (*Light Rose*). The GCN *learns* this whole GRN as a single data point and is able to then, given as an input a graph likes this, predict the mutation type for the transcription factor, or any other feature of interest we can extract from the annotation of the sample.

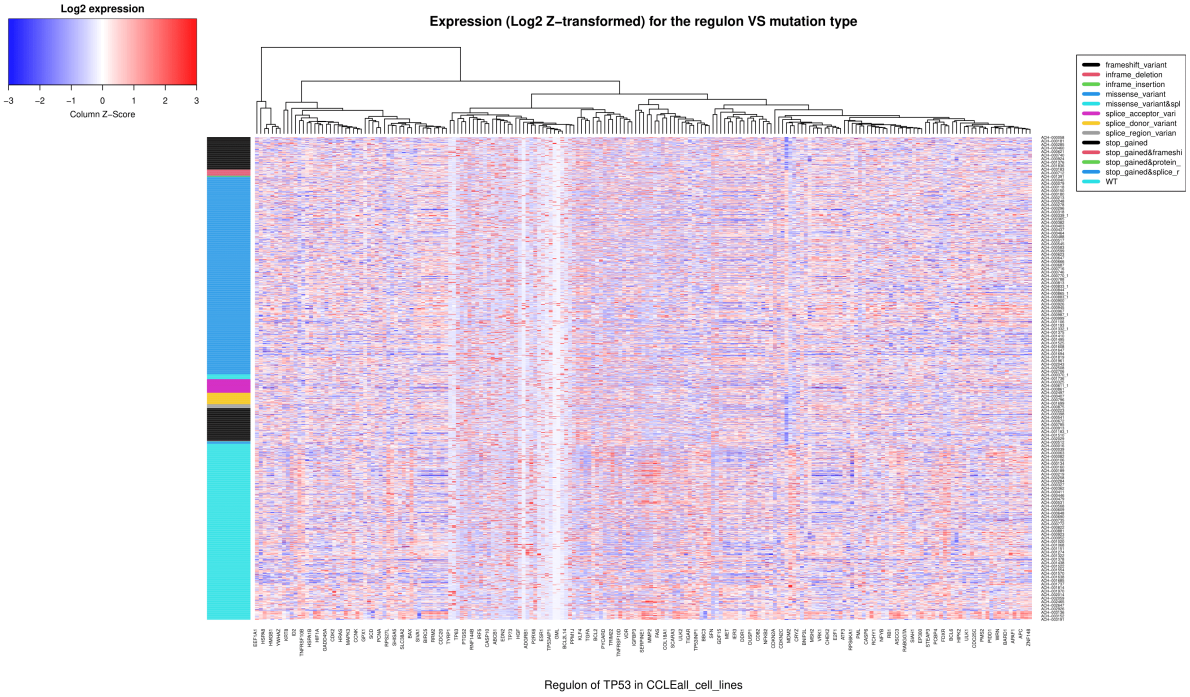

Figure 7: Z-normalized levels of expression for the regulon (set of target genes) of transcription factor *TP53*, as columns are depicted in this heatmap plot and samples (i.e. unique mutation of *TP53* per sample as rows. We stratify the samples/rows in the *TP53* mutation type using a coloured legend. We can see that missense mutations are prevalent, but WT samples are also dominant. Approximately 1600 samples are included in this figure, pancancer from CCLE 24Q2.

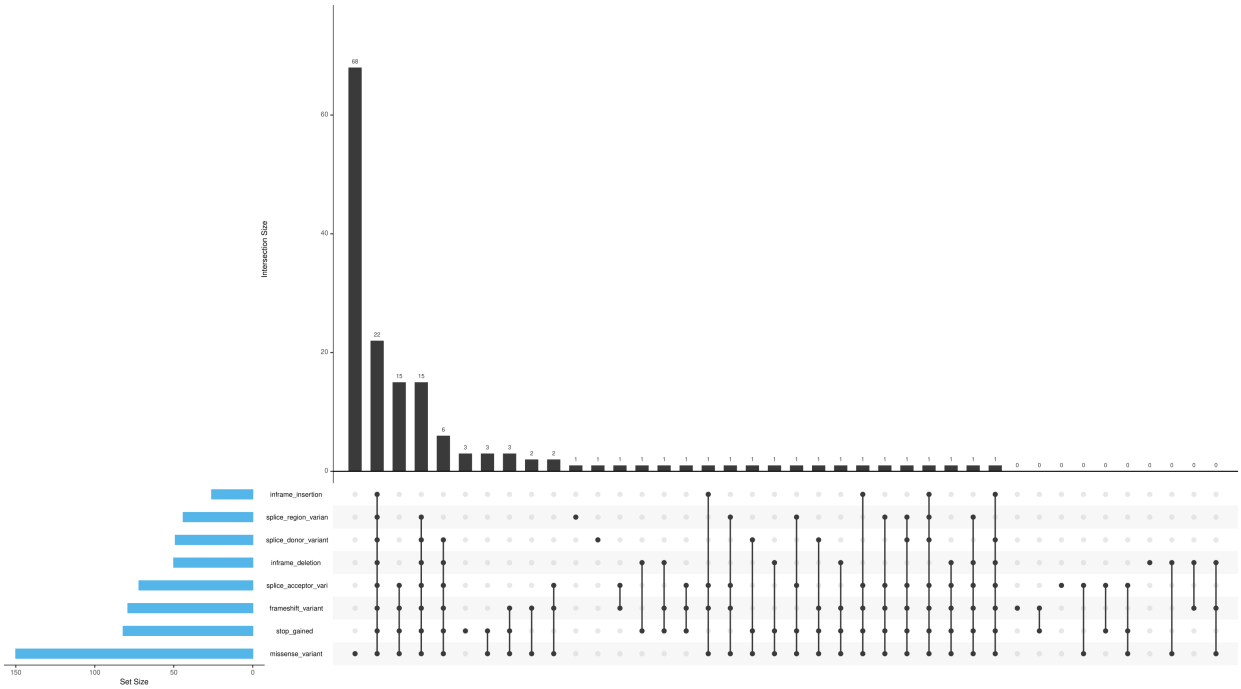

Figure 8: Description: The UpSet plot shows intersections between different mutation types, summarizing the overlap of these different mutation types. Components: Bottom Left: Horizontal bar plot shows the set size (number of mutations in each mutation type). For instance, “missense variant” is the largest set. Top Panel: Vertical bars indicate the intersection size; the number of genes shared across specific mutation combinations. Matrix (Bottom): Black dots and lines indicate which sets (mutation types) intersect. Key Insights: “Missense variant” is the most frequent mutation category. Smaller intersections highlight rare combinations of mutations shared between multiple types.

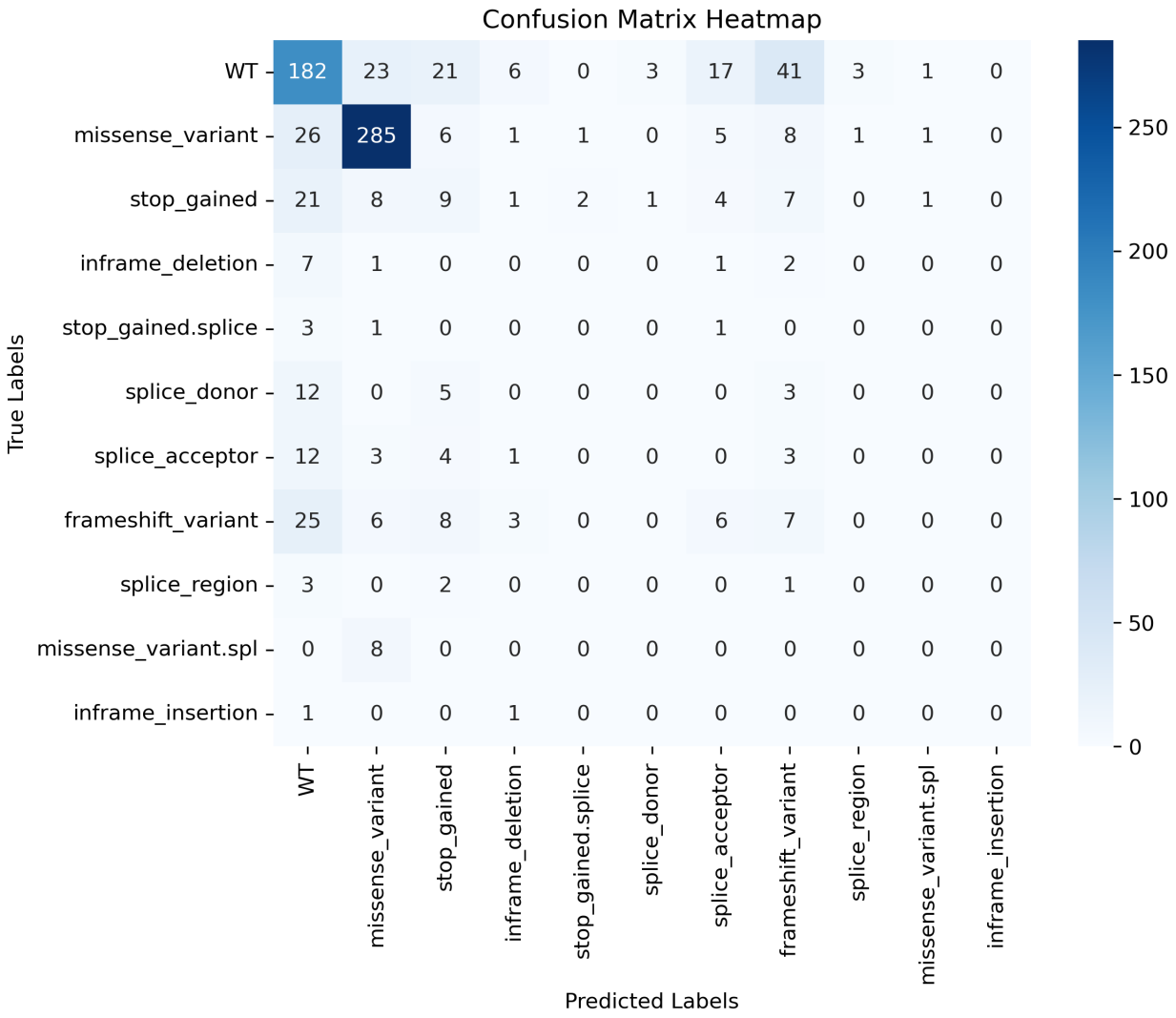

Figure 9: Confusion Matrix for the multi-label classifier built on GCNs. The classifier, as expected given the class imbalance and the availability of different types of mutations for *TP53* detects very well missense and WT samples (the most prevalent ones, as also seen in 7) and poorly for other types such as $stop_gained$.

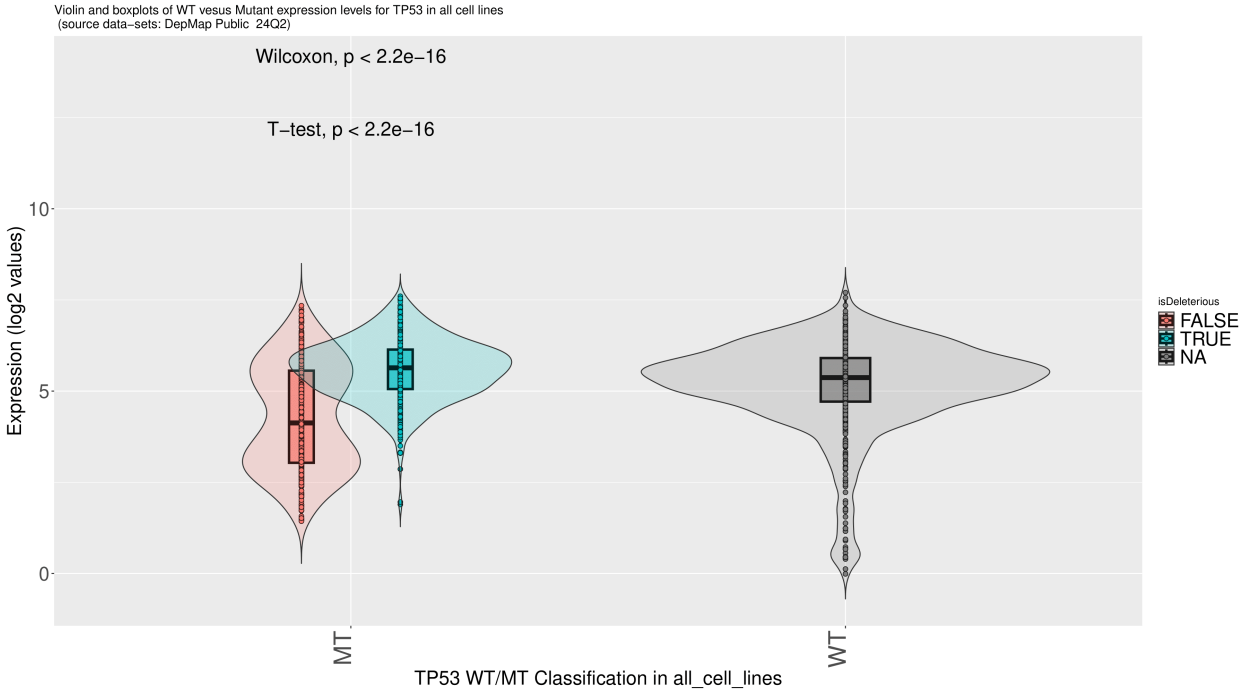

Figure 10: Deleterious function of mutation stratified across mutated samples and contrasted with controls (WT). We observe a statistically significant difference in the means of log2 normalized expression for *TP53* for the deleterious mutations against non-deleterious, and higher than WT expression.

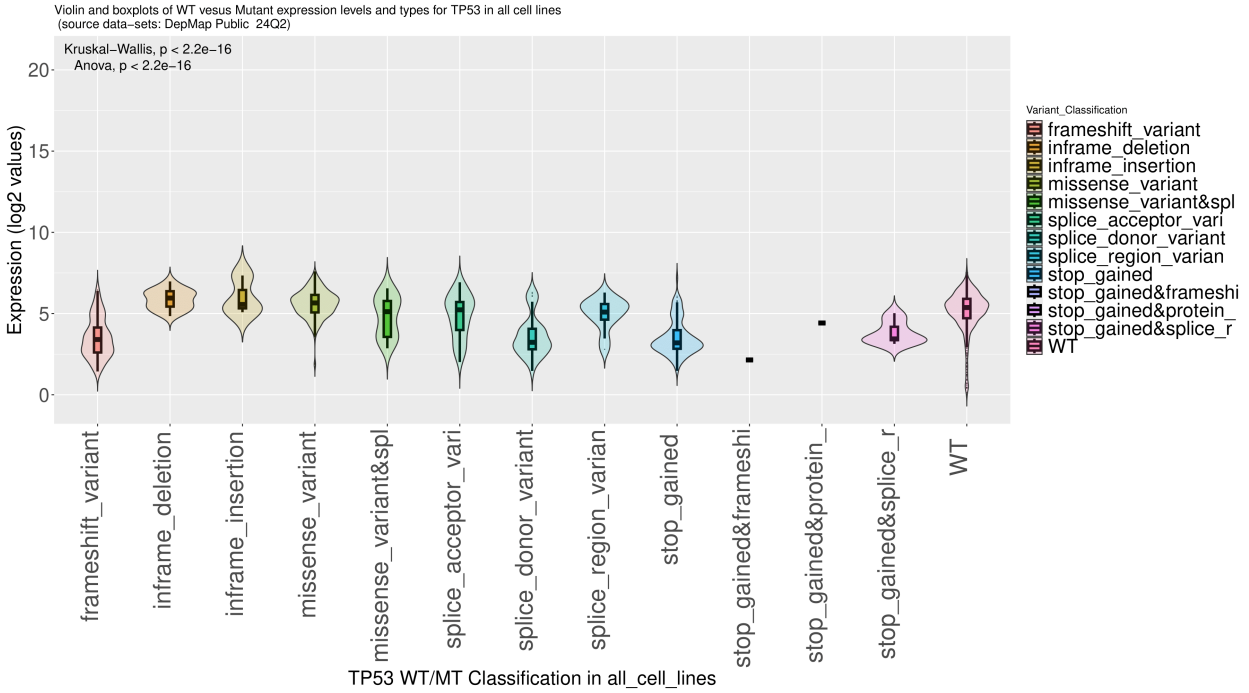

Figure 11: Description: A violin and boxplot analysis compares the expression levels of the TP53 gene across different mutation types in various cell lines. Axes: Y-axis: Expression values (log2 scale). X-axis: Mutation classifications (e.g., frameshift variant, missense variant, splice variants, WT). 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.