Learning Causal Gene Relationships in Biological Pathways with Graph Attention Networks (GATs)

Anonymous Author(s)
Affiliation
Address
email

Abstract

Biological pathways are natural causal graphs mapping gene-gene interactions that govern human processes. Despite their importance, most ML models treat genes as unstructured tokens, discarding causal structure. The latest pathway-informed models capture pathway-pathway interactions, but still treat each pathway as a "bag of genes" via MLPs, discarding its topology and gene-gene interactions. We propose a Graph Attention Network (GAT) framework that encodes gene-level pathway priors. We show that GATs generalize much better than MLPs, achieving an 81% reduction in MSE when predicting pathway dynamics under unseen treatment conditions. We further validate the correctness of our biological prior by encoding drug mechanisms as causal graph modifications, improving robustness. Finally, trained without a prior, we show that our GAT model correctly rediscovers all five gene-gene interactions in the canonical TP53-MDM2-MDM4 feedback loop from raw time-series mRNA data, demonstrating the ability to learn causal gene relationships and generate novel biological insights directly from experimental data. [All code will be released upon publication.]

1 Introduction

Biological pathways encode the logic of human processes, and can guide models to learn true biological relationships. However, existing ML models often treat genes as unstructured tokens, or at most encode interactions at the pathway level. To improve on this, we make the following contributions:

1. We propose a novel method to explicitly encode biological pathways at the gene level using Graph Attention Networks (GATs).
2. We show that encoding known biological pathways as a mechanistic prior allows models to learn a more robust, interpretable, and generalizable set of pathway dynamics.
3. We suggest the potential of our GAT formulation to discover new biological insights, such as candidate pathways and novel gene-gene interactions.

2 Related Work

The latest approaches to incorporating biological pathway knowledge into ML models include: (1) encoding gene-pathway membership via sparse neural networks [1; 2]; and (2) encoding pathway-pathway interactions using attention biases or graph priors [3; 4; 5]. While these approaches have achieved superior results on many downstream tasks, they remain limited by only encoding interactions at the pathway level, discarding the structured gene-gene interactions that define a pathway. For example, in the canonical p53 pathway, the gene TP53 activates MDM2, while MDM2 in turn

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

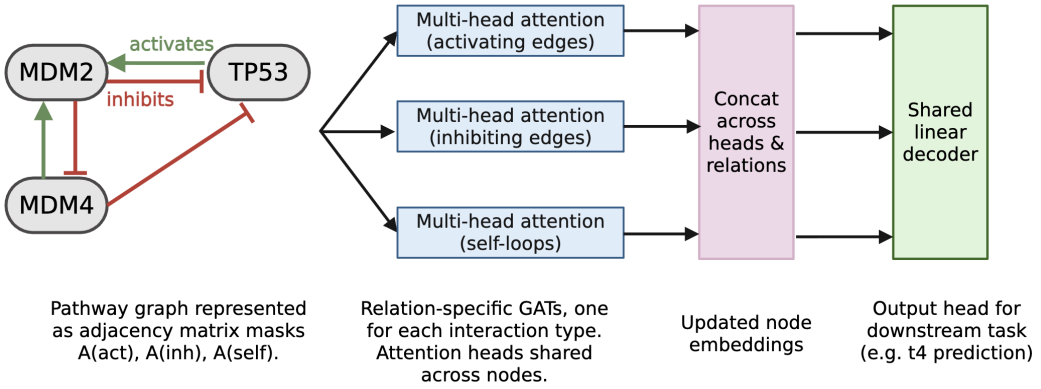


Figure 1: Model schematic

inhibits TP53, forming a negative feedback loop that is critical to regulating p53. Treating TP53 and MDM2 as independent tokens aggregated into a single "p53 pathway" node loses this mechanistic detail. We address this by proposing a Graph Attention Network (GAT) method to encode the natural graph structure of a biological pathway. We encode genes as nodes, and use multiple adjacency matrices for different interaction types (e.g. activatory/inhibitory). We show: (1) better generalization to unseen treatment conditions versus MLP; (2) correctness of our biological prior via edge interventions reflecting drug mechanisms; and (3) ability to rediscover known biology.

3 Methods

3.1 Data

We use our framework to model the core feedback loop of the p53 pathway, composed of the genes TP53, MDM2, and MDM4 (Figure 1). TP53 is a tumor suppressor whose activity is tightly regulated by MDM2 and MDM4 via negative feedback. This feedback loop generates oscillatory dynamics, presenting a richer challenge than linear signaling pathways.

Dataset. We use time-series mRNA expression for TP53, MDM2, and MDM4 from Hafner et al. [6] (GEO: GSE100099). Experiments cover three treatment conditions: **Wild-type (WT)** with no intervention; **TP53-sh**, a knock-down that dampens TP53 expression; and **Nutlin**, a drug that blocks the MDM2-TP53 interaction. For each condition there are two independent trajectories; within a trajectory, the three genes are measured at 6-12 time points over 0-24 hours.

Problem formulation. Given measurements at t_1, t_2, t_3 plus metadata (elapsed time between measurements $\Delta t_{1:3}$ and treatment indicators), predict the expression at t_4 for the three genes.

Evaluation protocol (LOCO). To test model robustness, we adopt Leave-One-Condition-Out (LOCO) validation. Each fold trains on two treatment conditions, and is then evaluated on the held-out condition. This assesses the model's ability to learn pathway dynamics that generalize to unseen treatment conditions. Unless stated otherwise, we report mean squared error (MSE) on the standardized target space averaged over 10 random seeds (mean $\pm$ std) and compare against an MLP baseline that reflects the "bag-of-genes" approach used in the literature.

3.2 Model Architecture

We build a pathway graph with N nodes (one per gene). Node i receives the feature vector:

$$\mathbf{x}_i = [\, \mathbf{y}_i(t_{1:3}) \,||\, \Delta\mathbf{t}_{1:3} \,||\, \mathbf{u} \,] \tag{1}$$

where $\mathbf{y}_i(t_{1:3}) \in \mathbb{R}^3$ are the first three expression measurements of gene i in the current window, $\Delta\mathbf{t}_{1:3} \in \mathbb{R}^3$ are the corresponding inter-measurement time gaps (identical for all nodes), and $\mathbf{u} \in \{0,1\}^K$ is the treatment indicator vector (also identical for all nodes).

Next, we encode pathway structure with a set of relation types $\mathcal{R}$ (here, $\mathcal{R}$ = {*activatory*, *inhibitory*, *self-loops*}). For each $r \in \mathcal{R}$, we define a binary adjacency $\mathbf{A}^{(r)} \in \{0,1\}^{N \times N}$, where $\mathbf{A}^{(r)}_{ij} = 1$ means that, under relation r, gene j can attend to gene i. For example, the relation "TP53 activates MDM2" is represented as $\mathbf{A}^{(\textit{activatory})}_{\textit{MDM2,TP53}} = 1$. We use raw node features as embeddings: $\mathbf{h}_i \leftarrow \mathbf{x}_i$.

For each relation $r \in \mathcal{R}$, we run a GAT block with H heads. Each head h learns its own projection matrix $\mathbf{W} \in \mathbb{R}^{F \times D}$ and node-level attention matrix $\mathbf{a} \in \mathbb{R}^{2D}$. We compute attention scores as:

$$\alpha_{ij} = \textit{softmax}_{j \in \mathcal{N}_i^{(r)}}(\textit{LeakyReLU}(\mathbf{a}^T[\mathbf{W}\mathbf{h}_i \,||\, \mathbf{W}\mathbf{h}_j])) \tag{2}$$

where the softmax function is applied over all permitted neighbors of node i under relation r, $\mathcal{N}_i^{(r)} = \{j : A_{ij}^{(r)} = 1\}$. The output feature of node i is thus:

$$\mathbf{h}_i' = \sum_{j \in \mathcal{N}_i^{(r)}} \alpha_{ij} \mathbf{W}\mathbf{h}_j \tag{3}$$

We concatenate outputs across the heads of each relation, then concatenate across relations, generating our aggregated node embedding $\mathbf{z}_i \in \mathbb{R}^{|\mathcal{R}|HD}$. A linear readout $\mathbf{W}_{dec} \in \mathbb{R}^{|\mathcal{R}|HD \times 1}$ then maps each node's aggregated embedding to its predicted t_4 expression. Dropout is applied on features and attention weights, and all $\mathbf{W}^{(r,h)}$ and $\mathbf{a}^{(r,h)}$ are randomly initialized using the Xavier method.

4 Results

4.1 GAT model generalizes to unseen treatment conditions much better

Table 1: **LOCO performance, GAT vs MLP** (10 seeds)

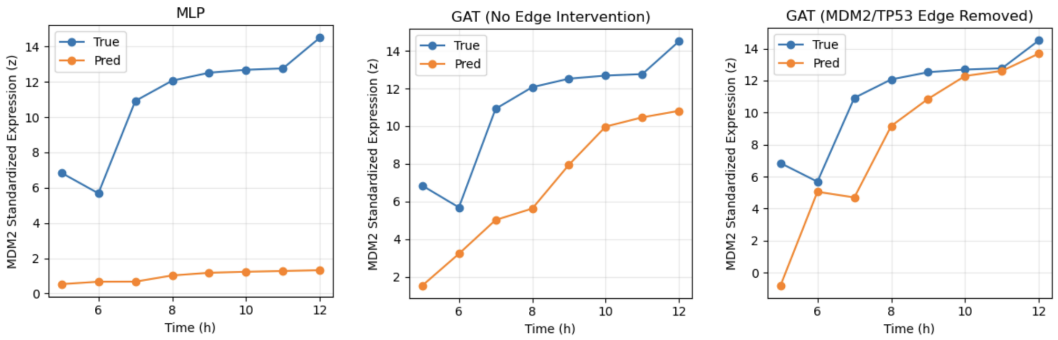
Fold (hold-out)	MLP MSE (mean±std)	GAT MSE (mean±std)
1 (WT)	1.07 ± 0.02	0.44 ± 0.13
2 (TP53-SH)	4.92 ± 0.13	0.31 ± 0.18
3 (Nutlin)	50.4 ± 1.6	10.0 ± 4.1
Overall mean	$\mathbf{18.8 \pm 0.6}$	$\mathbf{3.57 \pm 1.46}$

Under LOCO cross-validation, the GAT model achieved 81% lower overall MSE than the MLP baseline (Table 1), showing that the pathway prior enables much stronger generalization to unseen treatment conditions. The Nutlin case illustrates this best: Nutlin blocks MDM2 from degrading the TP53 protein; elevated TP53 then drives continuous MDM2 transcription, producing a monotonic rise up to 15 standard deviations above normal expression seen in training data (Fig. 2, blue). Without a pathway prior, the MLP fails entirely at predicting this behavior (Fig. 2a). In contrast, the GAT recognizes the feedback loop has been disrupted, and captures the rising trajectory (Fig. 2b-c).

4.2 Encoding the drug mechanism via edge intervention improves performance

Table 2: **LOCO performance, Drug Mechanism Encoding** (10 seeds)

Fold (hold-out)	MSE (mean ± std)	
	Unmodified Pathway	Edge Intervention
1 (WT)	0.52 ± 0.13	0.44 ± 0.13
2 (TP53-SH)	0.37 ± 0.17	0.31 ± 0.18
3 (Nutlin)	11.0 ± 6.8	10.0 ± 4.1
Overall mean	$\mathbf{3.97 \pm 2.24}$	$\mathbf{3.57 \pm 1.46}$



(a) MLP (b) GAT, unmodified pathway (c) GAT, drug target edge removed

Figure 2: True vs Predicted MDM2 expression under Nutlin

The correctness of the biological prior is further shown via an edge intervention. In Fig. 2b, the prior is unchanged; in Fig. 2c, we explicitly model Nutlin's mechanism of blocking the MDM2-TP53 interaction by setting $\mathbf{A}^{inhibitory}_{TP53,\,MDM2} = 0$. We report a further 11% improvement in prediction accuracy, demonstrating the GAT model's ability to explicitly incorporate known drug mechanisms.

4.3 Given no prior, GAT model rediscovers p53 pathway

Finally, we trained a GAT with a single, fully-connected adjacency matrix, using the $tanh()$ activation function to allow negative attention scores. From just raw time-series mRNA data, the model correctly recovered the signs of all 5 gene-gene interactions (Fig. 3). Their relative magnitudes also match known biology (e.g. TP53 being the main activator of MDM2; MDM2 being the main inhibitor of TP53). This suggests the GAT model can also be used to generate novel biological hypotheses (e.g. new pathways/interactions) when trained on raw experimental data without a prior.

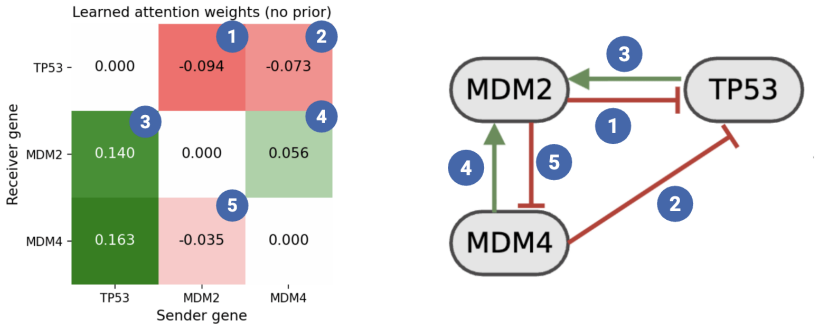


Figure 3: Learned attention weights (no prior) vs Ground truth pathway graph

Limitations. Our study focuses on a single pathway with few genes, and relies on a relatively limited dataset. Nonetheless, our results strongly suggest that encoding pathways at the gene level with GATs yields substantial improvements over MLP baselines, warranting further exploration.

5 Conclusion

We introduced a Graph Attention Network (GAT) framework that encodes biological pathways at the gene level, serving as a mechanistic prior. Our approach generalizes substantially better than existing MLP methods, and offers interpretability to encode known drug mechanisms and generate novel biological insights. Extensions include encoding temporal dynamics (e.g. via a GRU), and building a multi-pathway, hierarchical framework that uses one GAT layer to map gene data to pathways; and a further GAT layer to map pathway-pathway interactions – moving towards a foundation model covering all biological pathways that can be broadly applied to many life science tasks.

References

[1] Jiajing Xie, Ying Chen, Shijie Luo, Wenxian Yang, Yuxiang Lin, Liansheng Wang, Xin Ding, Mengsha Tong, and Rongshan Yu. Tracing unknown tumor origins with a biological-pathway-based transformer model. *Cell Reports Methods*, 4(6), 2024.

[2] Zhaoxiang Cai, Rebecca C Poulos, Adel Aref, Phillip J Robinson, Roger R Reddel, and Qing Zhong. Deepathnet: A transformer-based deep learning model integrating multiomic data with cancer pathways. *Cancer Research Communications*, 4(12):3151–3164, 2024.

[3] Xiaofan Liu, Yuhuan Tao, Zilin Cai, Pengfei Bao, Hongli Ma, Kexing Li, Mengtao Li, Yunping Zhu, and Zhi John Lu. Pathformer: a biological pathway informed transformer for disease diagnosis and prognosis using multi-omics data. *Bioinformatics*, 40(5):btae316, 2024.

[4] Zehao Dong, Qihang Zhao, Philip RO Payne, Michael A Province, Carlos Cruchaga, Muhan Zhang, Tianyu Zhao, Yixin Chen, and Fuhai Li. Highly accurate disease diagnosis and highly reproducible biomarker identification with pathformer. *Research Square*, pages rs–3, 2023.

[5] Teng Ma and Jianxin Wang. Graphpath: a graph attention model for molecular stratification with interpretability based on the pathway–pathway interaction network. *Bioinformatics*, 40(4): btae165, 2024.

[6] Antonina Hafner, Jacob Stewart-Ornstein, Jeremy E Purvis, William C Forrester, Martha L Bulyk, and Galit Lahav. p53 pulses lead to distinct patterns of gene expression albeit similar dna-binding dynamics. *Nature structural & molecular biology*, 24(10):840–847, 2017.