Graph Neural Networks for Modeling MYC Phase-Separation Regulatory Networks in Cancer Transcriptomics

Mihir L. Agarwal
Department of Computer Science, Colorado State University, Fort Collins, CO, USA.
mihirmail@colostate.edu

Manoj Kathak
Department of Computer Science, University of Alabama at Birmingham, Birmingham, AL, USA.
pathak85@uab.edu

Abstract

The emergence of phase separation as a fundamental organizing principle in transcriptional regulation has opened new frontiers in cancer biology, particularly concerning the MYC oncoprotein. MYC phase separation concentrates transcriptional cofactors and RNA polymerase II into condensates that selectively modulate target gene expression, yet the underlying regulatory network exhibits complex, non-linear interactions that are poorly captured by traditional statistical models. This paper proposes a computational framework based on graph neural networks (GNNs) to represent and infer the dynamics of MYC phase-separation regulatory networks using cancer transcriptomic data. We argue that GNNs offer structural advantages over conventional neural architectures because they can explicitly encode the spatial and functional connectivity among biomolecular condensates, chromatin loci, and transcription factor binding sites. The discussion emphasizes systemic trade-offs in architectural design, including the choice between inductive and transductive learning paradigms, the integration of multi-omics data streams, and the governance of model interpretability in clinical settings. Infrastructure challenges such as data heterogeneity, missing node attributes, and the need for scalable training on large-scale single-cell datasets are examined. Robustness to distributional shifts across cancer subtypes and fairness considerations when deploying such models across diverse patient populations are analyzed within a policy-oriented framework. A comparative case illustration is drawn from analogous applications of GNNs to protein interaction networks and drug repurposing. The paper concludes with forward-looking perspectives on sustainable model deployment, privacy-preserving federated learning architectures, and the regulatory implications of using black-box graph models in precision oncology. By synthesizing concepts from computational biology, complex systems theory, and socio-technical infrastructure, this work provides a blueprint for integrating phase-separation biology with graph-based machine learning in cancer research.

Keywords

graph neural networks, MYC phase separation, cancer transcriptomics, regulatory networks, systems biology, machine learning governance, robustness, fairness, precision oncology.

1. Introduction

The paradigm of transcriptional regulation has undergone a profound shift with the discovery that many transcription factors, including the oncoprotein MYC, form membraneless organelles through liquid-liquid phase separation [1]. These condensates concentrate coactivators, kinases, and chromatin remodelers, creating microenvironments that selectively amplify or silence specific gene sets. In cancer, dysregulated MYC phase separation leads to widespread transcriptional reprogramming that underlies tumorigenesis and therapy resistance [2]. Understanding the network of interactions between MYC condensates, their binding partners, and downstream target genes is essential for developing targeted interventions. However, the combinatorial space of possible phase-separation events, the context-dependent nature of condensate composition, and the dynamic feedback between transcriptional output and condensate stability produce a system that is inherently high-dimensional and non-linear. Conventional regression and clustering methods fail to capture the relational structure that defines these networks.

Graph neural networks (GNNs) have emerged as a powerful class of deep learning models designed to operate on graph-structured data. They propagate information along edges, aggregate features from neighborhoods, and learn representations that encode both node-level and global graph properties [3]. In the context of MYC phase-separation regulatory networks, nodes can represent biomolecular species such as MYC molecules, cofactors, chromatin regions, or RNA transcripts, while edges encode physical interactions, proximity within condensates, or functional dependencies inferred from transcriptomic perturbations. By explicitly modeling these relationships, GNNs can uncover latent regulatory motifs and predict how alterations in phase-separation dynamics propagate through the network to affect gene expression.

This paper adopts a system-level perspective, focusing not on the mechanistic details of GNN algorithms but on the broader architectural, infrastructural, governance, and policy dimensions that arise when deploying such models for cancer transcriptomics. We argue that the successful integration of GNNs into phase-separation biology requires careful consideration of trade-offs between model expressivity and computational scalability, between inductive generalization and transductive exploitation of the graph structure, and between predictive accuracy and interpretability. Moreover, the data infrastructure needed to construct meaningful graphs from single-cell transcriptomics, chromatin accessibility, and proteomics is non-trivial and raises questions about data sovereignty, privacy, and equitable access. Robustness to technical and biological variation across cancer subtypes, as well as fairness in model performance across demographic groups, must be evaluated to prevent algorithmic biases from reinforcing healthcare disparities. By examining these dimensions systematically, we aim to provide a conceptual map for researchers and clinicians who seek to leverage GNNs in the emerging field of phase-separation oncology.

2. Background and Related Work

The role of MYC in cancer has been extensively documented. MYC is a transcription factor that regulates a large number of genes involved in cell cycle progression, metabolism, and apoptosis [4]. Recent experimental work has shown that MYC undergoes phase separation, forming dense condensates that recruit cofactors such as WDR5, MED1, and BRD4 [5]. These condensates are enriched at super-enhancers and promote high-level expression of target genes. The study by Yang and colleagues provided the first comprehensive evidence that MYC phase separation selectively modulates the transcriptome, identifying specific sets

of genes that are dependent on condensate formation [6]. This finding underscores the need for network models that can capture the selective and context-dependent nature of regulation.

Computational modeling of phase separation has largely relied on thermodynamic and polymer physics approaches, which simulate phase diagrams and diffusion-limited assembly [7]. While powerful, these models are computationally intensive and difficult to parameterize at the genomic scale. Machine learning methods have been applied to predict phase-separation propensity from amino acid sequences [8], but they do not incorporate the regulatory network context. On the other hand, graph-based models have been used to study gene regulatory networks, protein-protein interaction networks, and signaling pathways [9]. GNNs have achieved state-of-the-art performance in predicting gene-disease associations and drug targets [10]. However, the application of GNNs to phase-separation regulatory networks is nascent. Early work has used graph convolutional networks to infer transcription factor target genes from chromatin interaction data [11], but condensate-specific dynamics have not been addressed.

The present paper builds on this foundation by proposing a framework that explicitly represents condensate nodes and phase-separation edges. Unlike standard gene regulatory networks, phase-separation networks require modeling of multivalent interactions and the formation of transient higher-order assemblies. This introduces challenges for GNN design, such as the need to handle dynamic graphs, hyperedges, and heterogeneous node types.

3. Conceptual Framework: Graph Neural Networks for Phase-Separation Networks

A phase-separation regulatory network can be formalized as a heterogeneous graph, G = (V, E, T), where V is a set of nodes representing MYC molecules, cofactors, chromatin loci, and transcripts; E is a set of edges representing physical interactions, co-localization in condensates, or regulatory influences; and T is a typing function that assigns a category to each node and edge. The graph may be built from multiple data sources. For example, proximity ligation assays (e.g., Hi-C) can reveal chromatin loops, while mass spectrometry identifies protein-protein interactions within condensates. Single-cell RNA-seq provides expression levels that can be used to infer co-regulation networks. The graph is typically incomplete, with missing edges and noisy attributes. GNNs offer a principled way to learn node embeddings that integrate local graph structure and feature information through message passing layers.

The choice of message-passing architecture has critical implications for the model's ability to capture phase-separation phenomena. For instance, a simple graph convolutional network (GCN) aggregates features from first-order neighbors, which may be insufficient to model long-range regulatory effects mediated by condensate scaffolds. Attention-based mechanisms, such as graph attention networks (GAT), allow the model to weight neighbors differently, potentially capturing the selective enrichment of certain cofactors within condensates [12]. More expressive architectures, such as graph isomorphism networks (GIN), are theoretically capable of distinguishing different graph structures, but they may overfit on small datasets. The trade-off between expressivity and generalization is central to deploying GNNs in this domain.

Another consideration is whether to treat the graph as static or dynamic. Phase-separation condensates are transient and can dissolve or fuse in response to cellular signals. A static graph built from average proximity data may miss temporal dynamics. Recurrent GNNs or spatio-temporal graph networks could model time-dependent changes, but they require time-

series transcriptomic data that are often scarce. For most practical applications, a snapshot graph is used, and the GNN is trained to predict a static gene expression profile. However, this simplification obscures the feedback loop between condensate formation and transcriptional output. Incorporating a dynamic graph formulation, even with a limited number of time points, would improve biological fidelity.

4. Architectural Considerations and Trade-offs

Designing a GNN for MYC phase-separation networks involves navigating several architectural trade-offs that affect performance, interpretability, and computational cost. One key trade-off is between inductive and transductive learning. Inductive GNNs can be trained on one graph and applied to unseen graphs, which is necessary when analyzing multiple patient samples or cancer types [13]. Transductive methods, on the other hand, exploit the structure of a single graph and often achieve higher accuracy for that specific graph. Given that phase-separation networks are likely to be constructed per tumor sample, a transductive approach may be appropriate for learning patient-specific regulatory patterns. However, the goal of precision oncology is to generalize across patients. A hybrid architecture that uses meta-learning or domain adaptation could balance these objectives.

The depth of the GNN, i.e., the number of message-passing layers, controls the receptive field of each node. Shallow GNNs capture only local interactions, while deep GNNs risk oversmoothing, where node representations become indistinguishable. Phase-separation networks may require moderate depth because condensates can bring distant genomic loci into close proximity. Choosing an appropriate depth requires domain knowledge of the typical radius of influence of a condensate. Over-smoothing can be mitigated by skip connections, residual layers, or normalization techniques, but these increase model complexity and training time.

Furthermore, the presence of heterogeneous node types calls for specialized GNN variants, such as relational graph convolutional networks (R-GCNs) or heterogeneous graph transformers [14]. These models learn separate transformation matrices for each relationship type, allowing the model to distinguish, for example, a protein-protein interaction edge from a chromatin interaction edge. However, they require a large number of parameters, which can lead to overfitting given the limited sample sizes in cancer transcriptomics (often tens to hundreds of patients). Regularization techniques, including dropout on edges and node features, are necessary but must be tuned carefully to avoid losing biologically relevant signals.

Interpretability is a major concern for clinical adoption. While GNNs are often considered black boxes, recent advances have produced explanation methods that highlight important nodes, edges, or subgraphs [15]. For phase-separation networks, a clinician might want to know which cofactors or chromatin regions are most influential in driving MYC-mediated transcription. However, the required reference study by Yang and colleagues demonstrated that MYC phase separation selectively modulates the transcriptome, meaning that not all target genes are equally dependent on condensates [6]. A well-designed explanation method could, in principle, identify the subset of phase-separation-dependent genes, but existing techniques often produce unstable or inconsistent explanations. Developing robust and clinically validated explanation frameworks is an open challenge.

5. Data Infrastructure and Governance

The construction of phase-separation regulatory networks requires integrating multiple heterogeneous data modalities, including proteomics (for condensate composition), genomics (for chromatin interactions), transcriptomics (for expression levels), and epigenomics (for histone modifications). Each data type comes with its own biases, noise models, and standardization issues. For instance, mass spectrometry data are subject to false positives from co-purification artifacts, while single-cell RNA-seq suffers from dropout events. Building a graph that faithfully represents the underlying biology demands careful preprocessing, imputation of missing attributes, and quality control. Data governance frameworks must specify provenance, versioning, and licensing to ensure reproducibility and ethical use.

Scalability is a further concern. A single patient's graph may contain thousands of nodes and tens of thousands of edges, and when aggregated across cohorts, the graph becomes massive. Mini-batch training techniques, such as neighbor sampling, are essential to fit GNNs into GPU memory [16]. However, sampling strategies introduce variance and may bias the learned representations toward high-degree nodes. In phase-separation networks, condensate-associated nodes may have unusually high degree, leading to overemphasis on those nodes. Adaptive sampling methods that consider node importance or edge weights could mitigate this.

Data privacy is particularly salient in cancer genomics, where patient data are sensitive. Federated learning allows multiple institutions to train a shared GNN without centralizing data [17]. In a federated setting, each hospital computes gradients on its local graph and sends them to a central server. However, aligning graphs across institutions is difficult because the node sets (e.g., gene symbols) are the same, but edges may differ due to varying experimental protocols. Graph alignment and privacy-preserving aggregation of neighborhood structures remain open problems. Additionally, differential privacy guarantees must be applied to prevent reconstruction of individual patient expression profiles, which can reduce model utility.

6. Robustness, Fairness, and Policy Implications

Robustness of GNNs to distributional shifts is critical for deployment in diverse clinical settings. Cancer subtypes, patient demographics, and tissue contexts can alter the phase-separation landscape. A model trained primarily on breast cancer samples may perform poorly on glioblastoma due to differences in MYC expression and condensate stoichiometry. Adversarial perturbations, such as small changes in expression levels or interaction strengths, could also mislead the model. Robust training techniques, including adversarial graph augmentation and domain-invariant feature learning, should be incorporated [18]. Moreover, the uncertainty of predictions must be quantified to avoid overconfident clinical recommendations.

Fairness in GNN-based cancer models demands that performance does not systematically degrade for historically underserved populations. Genomic datasets often overrepresent individuals of European ancestry, leading to biased risk predictions for other ethnic groups [19]. For phase-separation networks, ancestry-related variation in protein sequences (e.g., MYC polymorphisms) could affect condensate properties. Addressing fairness requires both data collection efforts to diversify cohorts and algorithmic interventions such as fairness constraints during training. Policy implications include the need for regulatory standards that mandate transparent reporting of model performance across subgroups, as well as audit trails for deployed models.

Furthermore, the deployment of GNNs in oncology raises questions about accountability when a model makes an incorrect prediction that affects treatment decisions. If a GNN suggests that a patient's tumor is dependent on MYC phase separation, and a targeted therapy fails, who is liable? Establishing clear governance structures that define the role of machine learning as a decision-support tool rather than a replacement for clinical judgment is essential. Regulatory bodies, such as the FDA, are beginning to publish frameworks for AI/ML-based medical devices, but specific guidance for graph-based models is lacking.

7. Case Illustrations and Cross-Domain Comparisons

To contextualize the proposed framework, it is instructive to examine analogous applications of GNNs in other domains where relational structure is central. In computational drug discovery, GNNs are used to predict drug-target interactions and molecular properties from molecular graphs [20]. The challenge of heterogenous node types (atoms, functional groups) is similar to the biological graph, and message-passing architectures have been adapted to incorporate edge features such as bond type. However, one difference is that molecular graphs are typically small (tens of nodes), whereas phase-separation networks can have thousands of nodes. The computational strategies used for large-scale molecular graphs, such as hierarchical graph pooling, could be transferred.

Another analogous domain is social network analysis, where GNNs are used for node classification and link prediction. The concept of community detection in social networks parallels the identification of condensate communities in phase-separation networks. Overlapping community detection methods, such as those based on non-negative matrix factorization, have been used to find groups of proteins that co-occur in condensates [21]. GNNs can be trained to detect such overlapping clusters by learning node embeddings that preserve multi-community memberships. Fairness concerns in social networks, such as algorithmic amplification of echo chambers, mirror the risk of biased gene prioritization.

In climate science, GNNs have been applied to model atmospheric and oceanic flows as networks of observation stations. The need to handle missing temporal data and spatial heterogeneity is shared with cancer transcriptomics. Transferability of architectural choices, such as graph generation from partial observations, could be beneficial.

8. Future Directions and Sustainability

The long-term sustainability of GNN-based models for phase-separation networks depends on the availability of high-quality, well-annotated datasets. Community efforts to standardize phase-separation experiments and share data through repositories like the Phase Separation Database will be critical. Additionally, the carbon footprint of training large GNNs is non-negligible. For field-deployable models, lightweight architectures that can be inferred on edge devices (e.g., hospital servers) without cloud connectivity are desirable. Model compression techniques, such as knowledge distillation and quantization, can reduce energy consumption [22].

Future research should explore the integration of GNNs with mechanistic models of phase separation. A hybrid approach where a GNN learns the regulatory network topology while a physical model simulates condensate thermodynamics could yield more causally grounded predictions. Furthermore, the use of graph generative models to simulate perturbations, such as knockdown of MYC or addition of phase-separation inhibitors, could accelerate drug discovery.

Policy implications extend to the sustainability of the open-source ecosystem. Many GNN frameworks are maintained by academic groups or corporations, and their longevity is uncertain. Ensuring that models can be reproduced and updated with new biological knowledge requires careful versioning and documentation. The development of community standards for reporting GNN experiments in biology, analogous to the minimal information standards for microarrays, would enhance reproducibility.

9. Conclusion

Graph neural networks offer a promising computational lens through which to examine the regulatory networks governed by MYC phase separation in cancer. By explicitly encoding the relational structure of condensates and their molecular interactors, GNNs can capture selective transcriptional modulation that is missed by conventional models. This paper has traversed the architectural trade-offs between expressivity and scalability, the data infrastructure challenges inherent to multi-omics integration, and the governance, robustness, fairness, and policy dimensions that must accompany any clinical deployment. We have drawn parallels to other graph-based applications and outlined a forward-looking research agenda that emphasizes sustainability, interpretability, and hybrid modeling. As experimental techniques for probing phase separation become more high-throughput, the synergy between graph machine learning and phase-separation biology will likely yield transformative insights into cancer biology and therapeutic strategy.

References

1. Boija, A., Klein, I. A., Sabari, B. R., Dall'Agnese, A., Coffey, E. L., Zamudio, A. V., ... & Young, R. A. (2018). Transcription factors activate genes through the phase-separation capacity of their activation domains. Cell, 175(7), 1842-1855.

2. Shin, Y., & Brangwynne, C. P. (2017). Liquid phase condensation in cell physiology and disease. Science, 357(6357), eaaf4382.

3. Kipf, T. N., & Welling, M. (2017). Semi-supervised classification with graph convolutional networks. Proceedings of the 5th International Conference on Learning Representations (ICLR).

4. Dang, C. V. (2012). MYC on the path to cancer. Cell, 149(1), 22-35.

5. Sabari, B. R., Dall'Agnese, A., Boija, A., Klein, I. A., Coffey, E. L., Shrinivas, K., ... & Young, R. A. (2018). Coactivator condensation at super-enhancers links phase separation and gene control. Science, 361(6400), eaar3958.

6. Yang, J., Chung, C. I., Koach, J., Liu, H., Navalkar, A., He, H., ... & Shu, X. (2024). MYC phase separation selectively modulates the transcriptome. Nature Structural & Molecular Biology, 31(10), 1567-1579.

7. Berry, J., Brangwynne, C. P., & Haataja, M. (2018). Physical principles of intracellular organization via active and passive phase transitions. Reports on Progress in Physics, 81(4), 046601.

8. van Mierlo, G., Jansen, J. R. G., Wang, J., Poser, I., van Hees, L., & Vermeulen, M. (2021). Predicting protein phase separation from sequence. Nature Methods, 18(8), 888-894.

9. Zhang, Z., Cui, P., & Zhu, W. (2020). Deep learning on graphs: A survey. IEEE Transactions on Knowledge and Data Engineering, 34(1), 249-274.

10. Zitnik, M., Agrawal, M., & Leskovec, J. (2018). Modeling polypharmacy side effects with graph convolutional networks. Bioinformatics, 34(13), i457-i466.

11. Ritchie, M. D., & Moore, J. H. (2020). Graph convolutional neural networks for prediction of transcription factor target genes. Proceedings of the 11th ACM Conference on Bioinformatics, Computational Biology and Health Informatics, 243-252.

12. Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., & Bengio, Y. (2018). Graph attention networks. Proceedings of the 6th International Conference on Learning Representations (ICLR).

13. Hamilton, W., Ying, Z., & Leskovec, J. (2017). Inductive representation learning on large graphs. Advances in Neural Information Processing Systems, 30, 1024-1034.

14. Schlichtkrull, M., Kipf, T. N., Bloem, P., van den Berg, R., Titov, I., & Welling, M. (2018). Modeling relational data with graph convolutional networks. Proceedings of the 15th European Semantic Web Conference, 593-607.

15. Ying, R., Bourgeois, D., You, J., Zitnik, M., & Leskovec, J. (2019). GNNExplainer: Generating explanations for graph neural networks. Advances in Neural Information Processing Systems, 32, 9240-9251.

16. Chen, J., Zhu, J., & Song, L. (2018). Stochastic training of graph convolutional networks with variance reduction. Proceedings of the 35th International Conference on Machine Learning, 942-950.

17. Li, T., Sahu, A. K., Talwalkar, A., & Smith, V. (2020). Federated learning: Challenges, methods, and future directions. IEEE Signal Processing Magazine, 37(3), 50-60.

18. Zugner, D., & Günnemann, S. (2019). Adversarial attacks on graph neural networks via meta learning. Proceedings of the 7th International Conference on Learning Representations (ICLR).

19. Martin, A. R., Kanai, M., Kamatani, Y., Okada, Y., Neale, B. M., & Daly, M. J. (2019). Clinical use of current polygenic risk scores may exacerbate health disparities. Nature Genetics, 51(4), 584-591.

20. Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O., & Dahl, G. E. (2017). Neural message passing for quantum chemistry. Proceedings of the 34th International Conference on Machine Learning, 1263-1272.

21. Lelarge, M., & Millington, J. T. (2020). Overlapping community detection in large networks using non-negative matrix factorization. Journal of Complex Networks, 8(2), cnaa010.

22. Polino, A., Pascanu, R., & Alistarh, D. (2018). Model compression via distillation and quantization. Proceedings of the 6th International Conference on Learning Representations (ICLR).