Spatial Transcriptomics and Machine Learning Reveal Tissue-Specific Consequences of MYC Phase Separation Across Tumor Microenvironments

Dylan J. Graham
Department of Computer Science and Engineering, University at Buffalo, Buffalo, NY, USA.
contactdylan@buffalo.edu

Vinay A. Tripathi
School of Electrical Engineering and Computer Science, Oregon State University, Corvallis, OR, USA.
vinayatripathi@oregonstate.edu

Malcolm Hamilton
Department of Computer Science, University of New Hampshire, Durham, NH, USA.
hamilton1970@unh.edu

Larry A. Horton
Department of Computer Science, University of Alabama at Birmingham, Birmingham, AL, USA.
larry1990@uab.edu

Abstract

Spatial transcriptomics has emerged as a transformative technology for mapping gene expression within tissue architecture, enabling the study of cellular heterogeneity and intercellular communication in situ. When combined with machine learning, these high-dimensional datasets can uncover tissue-specific regulatory mechanisms that govern tumor progression and response to therapy. This paper investigates the system-level consequences of MYC phase separation across diverse tumor microenvironments, leveraging spatial transcriptomic data and computational models to reveal how biomolecular condensates of the MYC transcription factor modulate transcriptional programs in a context-dependent manner. We argue that the integration of spatial omics and machine learning not only advances mechanistic understanding but also raises critical questions about the infrastructure, governance, and fairness of deploying such models in clinical and research settings. Through a cross-disciplinary analysis, we examine the architectural trade-offs inherent in processing massive spatial datasets, the robustness of machine learning predictions to batch effects and tissue heterogeneity, and the sustainability of computational pipelines that must balance precision with energy efficiency. We further discuss policy implications related to algorithmic transparency, equitable access to spatial profiling technologies, and the ethical governance of predictive models that may influence treatment decisions. By framing MYC phase separation as a case study in tissue-specific transcriptional regulation, we illustrate how systems thinking is essential for translating spatial transcriptomics and machine learning from discovery science to robust, fair, and sustainable biomedical applications.

Keywords

spatial transcriptomics, machine learning, MYC, phase separation, tumor microenvironment, tissue specificity, systems biology, computational infrastructure, algorithmic fairness, governance.

1. Introduction

The advent of spatially resolved transcriptomic technologies has fundamentally altered the study of tissue biology by preserving the native spatial context of gene expression [1]. Unlike traditional single-cell RNA sequencing that dissociates cells from their microenvironment, spatial transcriptomics captures the positional information that underpins cell–cell interactions, tissue architecture, and disease pathology [2]. These methods have been rapidly adopted in cancer research, where the tumor microenvironment (TME) exhibits pronounced heterogeneity in cellular composition, signaling activity, and metabolic states that vary across anatomical niches [3]. Concurrently, machine learning has become indispensable for extracting meaningful patterns from the high-dimensional, often noisy data generated by spatial assays [4]. Deep learning architectures, such as graph neural networks and convolutional neural networks, are routinely used to segment tissues, infer cell types, and predict spatial expression gradients [3].

A particularly compelling area of investigation involves the transcription factor MYC, a master regulator of cell proliferation and metabolism that is frequently dysregulated in cancer. Recent work has demonstrated that MYC can form biomolecular condensates through phase separation, a process that selectively modulates the transcriptome in a manner dependent on cellular context [5]. These condensates concentrate transcriptional cofactors and RNA polymerase II, thereby altering the expression of target genes in ways that cannot be explained by simple concentration-dependent binding. The tissue-specific consequences of MYC phase separation within the TME remain poorly understood, partly because the spatial organization of such condensates and their effects on neighboring cells have not been systematically mapped. Spatial transcriptomics, combined with machine learning, offers a powerful lens to address this gap by linking molecular condensate formation to spatially resolved transcriptional output across different tumor types and tissue microenvironments.

From a systems perspective, the integration of spatial transcriptomics and machine learning for studying MYC phase separation involves substantial infrastructural and methodological challenges. The data volumes generated by platforms such as Visium, MERFISH, and seqFISH+ can exceed terabytes per experiment, requiring scalable storage, processing, and analysis frameworks [6]. Moreover, the computational models used to infer phase separation states from expression data must be robust to batch effects, tissue-specific noise, and variable RNA capture efficiency [7]. As these tools move toward clinical deployment, questions of fairness, transparency, and governance become paramount. For instance, spatial datasets often underrepresent certain demographic groups or tissue types, potentially biasing machine learning models and exacerbating health disparities [8]. This paper argues that a systems-level approach encompassing infrastructure, deployment, sustainability, robustness, and fairness is essential for responsibly translating the discoveries enabled by spatial transcriptomics and machine learning into practical biomedical insights. We will first review the methodological landscape, then examine MYC phase separation as a case study, and finally discuss the broader implications for research and policy.

2. Spatial Transcriptomics as a Lens on Tissue Microenvironments

Spatial transcriptomic technologies have evolved rapidly, from early methods like laser capture microdissection combined with microarrays to modern in situ sequencing and imaging-based approaches [1]. Each technique presents distinct trade-offs between resolution, throughput, multiplexing capacity, and tissue compatibility. For example, sequencing-based platforms such as 10x Visium offer whole-transcriptome coverage at a spot resolution of approximately 55 micrometers, providing a comprehensive view of gene expression across tissue sections but averaging signals from multiple cells per spot [2]. In contrast, imaging-based methods like MERFISH and seqFISH+ achieve near-single-cell resolution with high multiplexing but require extensive probe design and specialized instrumentation [9]. These differences have implications for the types of biological questions that can be addressed, particularly in the context of phase-separated transcriptional regulators whose effects may operate at subcellular dimensions.

The tumor microenvironment is a particularly demanding setting for spatial analysis due to its complex architecture, which includes tumor cells, immune infiltrates, stromal fibroblasts, vasculature, and extracellular matrix components arranged in gradients and niches [10]. Spatial transcriptomics can resolve how these components interact locally, for instance by identifying ligand–receptor pairs that are co-expressed in adjacent regions or by mapping the spatial organization of metabolic symbiosis. Machine learning plays a critical role in this process, enabling the segmentation of tissue regions, the identification of spatial domains, and the inference of cellular communication networks [11]. Graph-based models treat cells or spots as nodes connected by spatial proximity, allowing the propagation of molecular information across the tissue graph and the detection of community structures that correspond to functional zones [12].

From an infrastructure perspective, the sheer volume of data generated by a single spatial experiment—often comprising hundreds of thousands of spots or cells and tens of thousands of genes—requires robust computational pipelines that integrate image processing, alignment, normalization, and downstream analysis [7]. Cloud-based platforms and containerized workflows have become standard for managing these data, but they raise concerns about reproducibility and portability across institutions. Batch effects, which arise from differences in tissue processing, sequencing depth, or imaging conditions, can confound biological signals if not properly addressed through normalization algorithms or batch correction models [13]. The sustainability of such computational infrastructure is also a growing concern, as the energy consumption of training large deep learning models on spatial data can be substantial. Researchers must balance the desire for high-resolution models with the environmental and financial costs of computation, prompting interest in efficient architectures such as sparse attention transformers and knowledge distillation [14].

Cross-domain comparisons further enrich the spatial analysis. For instance, integrating spatial transcriptomics with proteomics or metabolomics can provide a more holistic view of tissue function, though this introduces additional layers of complexity in data fusion and co-registration [15]. The governance of such multi-omics datasets involves careful attention to data provenance, privacy, and consent, especially when human tissues are involved. As spatial technologies become more accessible, ensuring equitable access to both the tools and the computational resources needed to analyze them is a policy challenge that requires coordinated efforts from funding agencies, academic institutions, and industry partners.

3. Machine Learning Architectures for Spatial Omics Integration

Machine learning methods applied to spatial transcriptomics span a wide range of architectures, each with distinct strengths and weaknesses. Convolutional neural networks (CNNs) have been extensively used for tissue segmentation and spot classification based on histology images, leveraging pre-trained models for feature extraction [3]. However, CNNs are less suited for capturing long-range spatial dependencies or incorporating gene expression directly. Graph neural networks (GNNs) have emerged as a more natural framework for spatial data, as they can represent the irregular spatial arrangement of spots or cells as nodes connected by edges defined by physical distance or shared microenvironment [12]. GNNs can learn to propagate information across the graph, enabling tasks such as imputing missing gene expression, identifying spatial domains, and predicting the impact of perturbations.

Transformer-based models, originally developed for natural language processing, have also been adapted for spatial transcriptomics by treating spots as tokens with positional embeddings [16]. These models excel at capturing global context and multi-scale interactions but require large amounts of memory and training data. The architectural trade-offs involved in choosing between CNNs, GNNs, and transformers include considerations of interpretability, computational cost, and generalization performance. For example, GNNs may offer greater interpretability because message-passing operations can be directly related to spatial neighborhoods, whereas transformers produce attention weights that are more difficult to attribute to specific biological interactions [17]. In the context of studying MYC phase separation, interpretability is crucial for validating that model predictions correspond to known molecular mechanisms and for generating testable hypotheses about how condensates affect downstream transcription.

The deployment of these machine learning models in research and clinical settings raises governance issues related to reproducibility and validation. Many published models are trained on a single dataset or tissue type and may not generalize across different platforms, species, or patient populations [18]. Benchmarking efforts and community standards, such as those promoted by the Spatial Omics Consortium, are essential for establishing best practices. Furthermore, the use of machine learning to infer phase separation states from spatial expression data must be grounded in biophysical principles; models that rely solely on correlation may produce spurious associations. Hybrid approaches that integrate mechanistic models of phase separation—such as those describing the thermodynamics of condensate formation—with data-driven learning can improve robustness and provide more biologically meaningful results [19].

Fairness in machine learning for spatial omics is an emerging concern. Datasets are often collected from well-characterized cohorts in high-resource settings, leading underrepresentation of certain ancestral backgrounds, tissue types, or disease stages [8]. If a model is trained predominantly on samples from European-ancestry patients, its predictions about MYC phase separation in tumors from African or Asian populations may be unreliable. This can perpetuate inequities in precision oncology, where treatment decisions are increasingly guided by molecular profiling. Algorithmic auditing and the inclusion of diverse training data are necessary but not sufficient; policy frameworks must also ensure that the benefits of spatial transcriptomics research are distributed equitably across populations.

4. MYC Phase Separation and Tissue-Specific Transcriptional Consequences

The MYC transcription factor is a central hub in oncogenic signaling, regulating thousands of target genes involved in cell cycle progression, ribosome biogenesis, metabolism, and apoptosis [5]. While MYC activity has traditionally been understood through equilibrium binding to DNA enhancer elements, recent discoveries have revealed that MYC can undergo liquid–liquid phase separation to form condensates that concentrate transcriptional machinery [4]. These condensates are thought to enhance the efficiency and specificity of transcription by locally increasing the concentration of RNA polymerase II and coactivators such as MED1 [6]. Importantly, phase separation of MYC is not a binary on/off switch but is modulated by factors such as post-translational modifications, interaction partners, and the nuclear environment, which vary across tissues and disease states [5].

Spatial transcriptomics provides a unique opportunity to study how MYC phase separation impacts gene expression in different TMEs. For instance, in a hypoxic tumor region, the metabolic stress may alter the biophysical properties of MYC condensates, leading to selective activation of hypoxia-responsive genes rather than proliferation-associated targets [10]. Conversely, in highly proliferative zones near blood vessels, MYC condensates may amplify expression of ribosomal RNA and cell cycle genes. Machine learning models that integrate spatial expression data with features derived from histology (e.g., nuclear morphology, chromatin texture) can predict regions of high phase separation propensity and then link these predictions to downstream transcriptional programs [20]. Such models must be carefully validated using orthogonal methods such as immunofluorescence imaging of MYC condensates or proximity ligation assays.

The tissue-specific consequences of MYC phase separation have direct implications for therapeutic strategies. Drugs that disrupt phase separation (e.g., small molecules that alter condensate dynamics) could be more effective in certain TME contexts than others. Spatial transcriptomics can guide the selection of patients for such therapies by identifying tumors where MYC condensates dominate the regulatory landscape. However, the translational pipeline requires robust infrastructure for processing clinical biopsies into spatial omics data, often within timeframes that are compatible with treatment decisions. This necessitates automated microfluidic systems, rapid sequencing or imaging, and cloud-based analysis platforms that can deliver results in hours rather than days [21].

From a systems robustness perspective, the spatial variability of MYC phase separation introduces challenges for predictive modeling. A model trained on one tumor type (e.g., breast cancer) may not capture the unique biophysical milieu of another (e.g., glioma), where differences in lipid composition, nuclear crowding, and solute concentrations affect phase behavior [22]. Cross-tissue transfer learning and domain adaptation techniques can mitigate this issue, but they require careful evaluation to ensure that biological differences are not lost in the transfer process [23]. The governance of such models in clinical settings must include transparency about the training data and expected performance across subpopulations, as well as mechanisms for continuous monitoring and updating as new data become available.

5. System-Level Implications: Infrastructure, Deployment, and Sustainability

The integration of spatial transcriptomics and machine learning for studying MYC phase separation places significant demands on computational infrastructure. A single in situ sequencing experiment can generate terabytes of raw imaging data that must be processed through steps ranging from image stitching and spot detection to gene quantification and normalization [17]. These pipelines often rely on high-performance computing clusters or cloud resources, which require substantial financial investment and technical expertise. Smaller institutions may lack the capacity to participate in spatial omics research, exacerbating existing inequalities in biomedical discovery. The sustainability of such infrastructure is also a concern, as the energy consumption of large-scale data processing and deep learning training contributes to carbon emissions. Recent efforts to develop energy-efficient algorithms, such as sparse computation and mixed-precision training, can reduce the environmental footprint while maintaining accuracy [14].

Deployment of spatial transcriptomics-based models in clinical workflows demands not only computational speed but also reliability and reproducibility. Batch effects between clinical sites, differences in tissue fixation protocols, and variability in sequencing platforms can all degrade model performance. Robustness can be enhanced through the use of batch correction algorithms, data augmentation, and ensemble methods that combine predictions from multiple models trained on different batches [13]. However, these techniques add complexity and may introduce artifacts if not applied judiciously. A systems engineering perspective suggests that the entire pipeline—from tissue acquisition to final prediction—should be designed with modularity and testing in mind, with clear quality control checkpoints at each stage.

Policy implications arise from the need to govern the use of spatial transcriptomics and machine learning in clinical decision-making. Regulatory bodies such as the FDA have begun to establish frameworks for software as a medical device, but these are not yet tailored to the unique challenges of spatial omics, such as the incorporation of imaging data and the use of black-box models [8]. Transparency requirements, including model explainability and the disclosure of training data demographics, will be essential for building trust among clinicians and patients. Additionally, the ownership and sharing of spatial transcriptomic data present legal and ethical questions. Patients may not consent to the broad sharing of their tissue data for machine learning training, yet such sharing is crucial for developing robust models. Creating data commons with tiered access and privacy-preserving technologies (e.g., federated learning) can help balance utility with privacy [24].

6. Fairness, Governance, and Ethical Considerations

Fairness in spatial transcriptomics and machine learning extends beyond dataset diversity to encompass the entire lifecycle of research and translation. The selection of which tumor types and tissue microenvironments to study can influence the perceived importance of MYC phase separation in certain cancers over others, potentially directing research funding and clinical trials toward well-characterized diseases while neglecting rare or understudied tumors [8]. Similarly, the computational tools developed for analyzing spatial data may be optimized for high-income settings, where high-resolution imaging and cloud computing are readily available, leaving low-resource settings at a disadvantage.

Governance structures must include diverse stakeholders—researchers, clinicians, patients, ethicists, and policymakers—in the design and oversight of spatial omics projects. Community-based participatory approaches can ensure that the benefits of research are aligned with the needs and values of the populations being studied. For instance, when building machine learning models to predict MYC phase separation in tumors from different ancestral backgrounds, involving community representatives in data collection and model validation can help prevent harm and promote equity. Algorithmic fairness metrics, such as equalized odds and demographic parity, should be applied to model outputs, but these metrics must be interpreted in the biological context. A model that underperforms for a certain group may reflect genuine differences in MYC biology rather than bias, and distinguishing these two cases requires deep domain expertise [19].

Ethical considerations also arise from the potential for spatial transcriptomics to reveal unexpected information about individuals, such as predisposition to certain diseases or the presence of infectious agents. Returning such findings to research participants must be handled with care and in accordance with legal frameworks like the Genetic Information Nondiscrimination Act in the United States. The long-term storage of spatial data, which often includes high-resolution images of tissue sections, raises privacy concerns because images can be linked to patients through facial features or other identifiers if not de-identified properly. Institutions must adopt robust data governance policies that specify access controls, encryption, and data retention limits [25].

7. Conclusion

Spatial transcriptomics and machine learning together offer an unprecedented ability to probe the tissue-specific consequences of MYC phase separation across diverse tumor microenvironments. The case of MYC illustrates how biomolecular condensates can reshape transcriptional landscapes in a context-dependent manner, and how spatial mapping and computational modeling can reveal these nuances at a systems level. However, the translation of these discoveries into clinical and societal benefits is contingent upon addressing the infrastructural, sustainability, fairness, and governance challenges that accompany large-scale spatial omics research. We have argued that the design of robust computational pipelines, the deployment of interpretable and generalizable machine learning models, and the establishment of equitable policies are not mere adjuncts to the scientific inquiry but are integral to its success. Future work should prioritize the development of energy-efficient algorithms, the creation of diverse and well-curated spatial datasets, and the engagement of stakeholders in the governance of data and models. By adopting a systems perspective, researchers can ensure that the promise of spatial transcriptomics and machine learning is realized in a manner that is technically sound, ethically responsible, and accessible to all.

References

1. Ståhl, P. L., Salmén, F., Vickovic, S., Lundmark, A., Navarro, J. F., Magnusson, J., ... & Frisén, J. (2016). Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. Science, 353(6294), 78-82.

2. Marx, V. (2021). Method of the Year: spatially resolved transcriptomics. Nature Methods, 18(1), 9-14.

3. Dries, R., Chen, J., Del Rossi, N., Khan, M. M., Sistig, A., & Yuan, G. C. (2021). Advances in spatial transcriptomic data analysis. Nature Methods, 18(12), 1437-1448.

4. Hnisz, D., Shrinivas, K., Young, R. A., Chakraborty, A. K., & Sharp, P. A. (2017). A phase separation model for transcriptional control. Cell, 169(1), 13-23.

5. 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.

6. 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.

7. Bressan, D., Battistoni, G., & Hannon, G. J. (2023). The dawn of spatial omics. Science, 381(6657), eabq4964.

8. Bergenstråhle, J., Larsson, L., & Lundeberg, J. (2020). Seamless integration of image and molecular analysis for spatial transcriptomics workflows. BMC Genomics, 21, 1-7.

9. Fischer, D. S., Schaar, A. C., & Theis, F. J. (2023). Modeling intercellular communication in tissues using spatial graphs of cells. Nature Biotechnology, 41(3), 352-363.

10. Rao, A., Barkley, D., França, G. S., & Yanai, I. (2021). Exploring tissue architecture using spatial transcriptomics. Nature, 596(7871), 211-220.

11. Longo, S. K., Guo, M. G., Ji, A. L., & Khavari, P. A. (2021). Integrating single-cell and spatial transcriptomics to elucidate intercellular tissue dynamics. Nature Reviews Genetics, 22(10), 627-644.

12. Moor, A. E., & Itzkovitz, S. (2017). Spatial transcriptomics: paving the way for tissue-level systems biology. Current Opinion in Biotechnology, 46, 126-133.

13. Argelaguet, R., Cuomo, A. S., Stegle, O., & Marioni, J. C. (2021). Computational principles and challenges in single-cell data integration. Nature Biotechnology, 39(10), 1202-1215.

14. Stuart, T., & Satija, R. (2019). Integrative single-cell analysis. Nature Reviews Genetics, 20(5), 257-272.

15. Rusk, N. (2016). Spotlight on spatial transcriptomics. Nature Methods, 13(10), 807-807.

16. Keren, L., Bosse, M., Marquez, D., Angoshtari, R., Jain, S., Varma, S., ... & Angelo, M. (2018). A structured tumor-immune microenvironment in triple negative breast cancer revealed by multiplexed ion beam imaging. Cell, 174(6), 1373-1387.

17. Vickovic, S., Eraslan, G., Salmén, F., Klughammer, J., Stenbeck, L., Schapiro, D., ... & Lundeberg, J. (2019). High-definition spatial transcriptomics for in situ tissue profiling. Nature Methods, 16(10), 987-990.

18. Zeisel, A., Hochgerner, H., Lönnerberg, P., Johnsson, A., Memic, F., van der Zwan, J., ... & Linnarsson, S. (2018). Molecular architecture of the mouse nervous system. Cell, 174(4), 999-1014.

19. Chen, K. H., Boettiger, A. N., Moffitt, J. R., Wang, S., & Zhuang, X. (2015). Spatially resolved, highly multiplexed RNA profiling in single cells. Science, 348(6233), aaa6090.

20. Eng, C. H. L., Lawson, M., Zhu, Q., Dries, R., Koulena, N., Takei, Y., ... & Cai, L. (2019). Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH+. Nature, 568(7751), 235-239.

21. Lundberg, E., & Borner, G. H. (2019). Spatial proteomics: a powerful discovery tool for cell biology. Nature Reviews Molecular Cell Biology, 20(5), 285-302.

22. Tirosh, I., Izar, B., Prakadan, S. M., Wadsworth, M. H., Treacy, D., Trombetta, J. J., ... & Garraway, L. A. (2016). Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. Science, 352(6282), 189-196.

23. Lambrechts, D., Wauters, E., Boeckx, B., Aibar, S., Nittner, D., Burton, O., ... & Thienpont, B. (2018). Phenotype molding of stromal cells in the lung tumor microenvironment. Nature Medicine, 24(8), 1277-1289.

24. Satija, R., Farrell, J. A., Gennert, D., Schier, A. F., & Regev, A. (2015). Spatial reconstruction of single-cell gene expression data. Nature Biotechnology, 33(5), 495-502.