Multi-Omics Foundation Models for Deciphering Phase Separation–Mediated Transcriptional Control in Precision Oncology

Niklas L. Butler
School of Computing, Clemson University, Clemson, SC, USA.
butler197@clemson.edu

Anton Chandra
Department of Computer Science and Engineering, University of Nevada, Reno, Reno, NV, USA.
antonc@unr.edu

Wiktor Dastro
Department of Computer Science, University of New Hampshire, Durham, NH, USA.
castroviktor@unh.edu

Abstract

The emergence of liquid-liquid phase separation as a fundamental principle in gene regulation has reshaped our understanding of transcriptional control, particularly in cancer. Concurrent advancements in artificial intelligence and multi-omics profiling have created unprecedented opportunities to model these complex molecular processes at scale. This paper proposes a systems-level framework for integrating multi-omics foundation models with the mechanistic biology of phase-separated transcriptional condensates to advance precision oncology. We examine the architectural requirements for such models, including the representation of dynamic, non-equilibrium biomolecular assemblies, the fusion of heterogeneous data modalities, and the handling of spatial and temporal heterogeneity in tumor microenvironments. The discussion emphasizes structural trade-offs between model interpretability and predictive power, the robustness of learned representations across diverse patient populations, and the sustainability of deploying large-scale models in clinical workflows. Governance challenges, including data privacy, algorithmic fairness, and regulatory oversight, are critically assessed in the context of model-driven therapeutic decision-making. Forward-looking perspectives on federated learning, dynamic model updating, and policy frameworks for responsible innovation are provided. By bridging the gap between biophysical principles and deep learning architectures, this work outlines a roadmap for building trustworthy, scalable, and biologically grounded foundation models that can translate phase separation dynamics into actionable insights for cancer diagnosis, prognosis, and treatment.

Keywords

foundation models, multi-omics, phase separation, transcriptional regulation, precision oncology, systems biology, artificial intelligence, governance, robustness, sustainability.

1. Introduction

Precision oncology has long sought to decode the molecular heterogeneity of tumors and tailor therapeutic interventions accordingly. While genomic sequencing and transcriptomic profiling have yielded transformative insights, the mechanistic drivers of aberrant gene expression in cancer remain only partially understood. A growing body of evidence implicates liquid-liquid phase separation as a key organizing principle in the assembly of transcriptional machinery at super-enhancers and other regulatory loci [1,2]. These phase-separated condensates concentrate transcription factors, coactivators, and RNA polymerase II to enable sharp, switch-like responses in gene expression. In cancer, mutations or overexpression of oncoproteins such as MYC, BRD4, and MED1 can alter phase separation dynamics, leading to dysregulation of entire transcriptional programs [3,4]. However, the direct observation and modeling of phase separation in living cells remains technically challenging, and computational approaches that integrate multi-omics data with biophysical principles are urgently needed.

Foundation models, which are large-scale neural networks pre-trained on vast and diverse datasets, have demonstrated remarkable capacity for learning generalizable representations across domains ranging from natural language to protein structures [5,6,7]. In biomedicine, foundation models trained on genomic sequences, transcriptomic profiles, and proteomic data have begun to enable zero-shot and few-shot predictions for tasks such as variant effect prediction, drug response, and cell type annotation [8,9]. Yet these models largely ignore the mechanistic underpinnings of gene regulation, such as the role of phase separation in controlling transcriptional burst kinetics and chromatin accessibility. Incorporating phase separation biology into foundation model architectures could significantly improve their ability to predict the impact of oncogenic alterations on gene expression and to identify novel therapeutic vulnerabilities.

This paper presents a comprehensive systems-level analysis of the design, deployment, and governance of multi-omics foundation models that explicitly incorporate phase separation–mediated transcriptional control. We argue that such models must move beyond static molecular snapshots to represent the dynamic, multivalent interactions that drive condensate formation and dissolution. We examine the architectural trade-offs between model complexity and interpretability, the robustness of learned representations across genetically and epigenetically diverse patient cohorts, and the infrastructure required to scale these models from research laboratories to clinical settings. Critical governance issues, including data sovereignty, algorithmic bias, and regulatory approval pathways, are analyzed through the lens of responsible AI in healthcare. Finally, we outline a forward-looking agenda for integrating mechanistic biology with foundation model research to accelerate the realization of precision oncology.

2. The Biological Basis of Phase Separation in Transcriptional Control

Liquid-liquid phase separation (LLPS) describes the condensation of biomolecules into membraneless organelles that concentrate specific proteins and nucleic acids. In the nucleus, phase-separated condensates known as transcriptional hubs or super-enhancer condensates bring together transcription factors, coactivators, and the Mediator complex to facilitate efficient transcription [2,11]. The formation of these condensates is driven by intrinsically disordered regions (IDRs) that mediate multivalent, weak interactions, resulting in a dynamic, liquid-like state that can rapidly respond to cellular signals. Phase separation imparts several functional advantages to transcriptional regulation: it enables the co-localization of multiple transcription factors at sub-saturating concentrations, it sharpens the regulatory response to

input signals through thresholding mechanisms, and it allows for the spatial separation of distinct transcriptional programs within the same nucleus [12].

In cancer, the dysregulation of phase separation can arise from mutations in IDRs, amplification of master transcription factors, or alterations in chromatin state. For example, the oncoprotein MYC, which is frequently overexpressed in a wide range of cancers, has been shown to undergo phase separation mediated by its IDR, and this process selectively modulates the expression of a subset of target genes involved in proliferation and metabolism [10]. The study by Yang et al. directly demonstrated that MYC phase separation enhances the transcription of growth-promoting genes while having minimal effect on housekeeping genes, providing a mechanism for the selective transcriptomic reprogramming characteristic of MYC-driven tumors [10]. Similarly, fusions involving the IDR of the transcription factor DUX4 have been shown to induce ectopic phase separation and drive rhabdomyosarcoma [13]. These findings underscore the necessity of modeling phase separation as a critical regulatory layer that cannot be inferred solely from linear sequence or static expression data.

The dynamic nature of phase separation poses unique challenges for computational modeling. Condensates are non-equilibrium structures whose composition, size, and lifetime are regulated by post-translational modifications, RNA abundance, and ATP-dependent processes such as chromatin remodeling and transcription itself [14]. Traditional machine learning approaches that treat gene expression as a static output fail to capture the temporal coupling between condensate assembly and transcriptional bursts. Multi-omics technologies now provide measurements of chromatin accessibility, histone modifications, nascent transcription, and protein-protein interactions at increasing resolution and throughput. Integrating these data types into a unified model that respects the biophysics of phase separation requires novel architectural innovations that go beyond the standard transformer or graph neural network paradigms.

3. Multi-Omics Data Integration and Foundation Models

Foundation models in biomedicine have been predominantly developed for single-modality tasks, such as predicting protein structure from sequence or classifying disease subtypes from bulk RNA-seq [7,8]. Multi-omics foundation models, however, aim to learn a joint representation across multiple data types including genomics, epigenomics, transcriptomics, proteomics, and metabolomics. The fundamental premise is that the interplay between these layers contains synergistic information that is lost when each modality is analyzed in isolation. For phase separation–mediated transcriptional control, the relevant modalities include chromatin conformation data (Hi-C), transcription factor binding profiles (ChIP-seq), nascent transcription (PRO-seq), and quantitative proteomics of condensate components. Several recent efforts have demonstrated the feasibility of cross-modal pretraining using masked modeling objectives, where the model learns to reconstruct masked modalities from the remaining ones [15,16]. However, these approaches have not been explicitly designed to incorporate physical principles such as multivalent binding, liquid-liquid phase behavior, or the spatial organization of the nucleus.

One promising direction is to condition the foundation model on biophysical priors derived from coarse-grained simulations of phase separation. For example, a neural network could be equipped with a latent variable representing the "condensate propensity" of a given genomic locus, which is updated based on the local concentration of factors known to drive phase separation. This latent variable would then modulate the predicted transcriptional output. Such an architecture would allow the model to transfer knowledge from in vitro phase

separation assays to in vivo tumor contexts. Another approach is to use graph neural networks that represent the three-dimensional chromatin architecture and the spatial distribution of molecular condensates, leveraging recent advances in imaging-based omics such as MERFISH and seqFISH [17]. By treating the nucleus as a polymer network with phase-separated domains, the model could learn to predict how perturbations in IDR sequences or expression levels affect transcriptional coordination across distant loci.

A critical design consideration is the scale and diversity of the training data. Foundation models typically require massive, heterogeneous datasets to learn generalizable features. For phase separation, the available data are still sparse and often limited to a small number of cell lines or contexts. Transfer learning from larger, related datasets (e.g., general chromatin state maps) may mitigate this problem, but it risks diluting the mechanistic signal. Active learning and experimental design strategies that prioritize the collection of phase separation–specific measurements based on model uncertainty could accelerate data generation. Furthermore, the incorporation of synthetic data from molecular dynamics simulations or droplet assays could provide an infinite supply of realistic training examples for the phase separation module of the model [18]. Ensuring that synthetic distributions align with real biological distributions remains a major challenge.

4. Architectural Considerations for Foundation Models in Precision Oncology

Building a multi-omics foundation model that captures phase separation dynamics requires careful architectural choices that balance expressivity, tractability, and interpretability. The transformer architecture has become the de facto backbone for many foundation models due to its ability to capture long-range dependencies in sequence data. In the context of genomics, DNA transformers such as Enformer and Nucleotide Transformer have shown strong performance in predicting expression and regulatory effects from sequence alone [6,19]. However, these models operate on linear genomic sequences and do not explicitly model the three-dimensional genome or the physics of phase separation. To incorporate phase separation, one might extend the attention mechanism to operate on a graph where nodes represent genomic loci and edges represent spatial proximity or co-condensation as inferred from Hi-C and imaging data. This graph must be dynamic, as condensate boundaries shift with cellular state.

Another architectural innovation is the use of neural ordinary differential equations (ODEs) to model the temporal evolution of condensate composition and transcriptional activity. Instead of predicting a static expression level, a neural ODE could simulate the trajectory of key molecular species (e.g., transcription factor concentrations, RNA polymerase II occupancy) over time, with phase separation represented as a nonlinear function of these concentrations. The model would then be trained on time-series data from live-cell imaging or nascent transcription assays. The computational cost of such models is high, but they offer the advantage of mechanistic interpretability: the learned parameters can be mapped to biophysical quantities such as binding affinities and diffusion coefficients. Moreover, neural ODEs can be regularized with physical constraints, such as mass conservation and thermodynamic consistency, to improve generalization under data-scarce regimes.

Interpretability is a paramount concern when deploying foundation models in clinical oncology. A black-box model that predicts drug response from omics data may achieve high accuracy, but it will not engender trust among clinicians or regulatory agencies unless its reasoning can be explained. For models incorporating phase separation, interpretability can be enhanced by designing modules that output intermediate predictions such as the probability of

a given locus forming a condensate, the key driver molecules involved, and the expected transcriptional burst frequency. Visualization tools that map these predictions onto three-dimensional nuclear simulations could provide intuitive explanations. Additionally, feature attribution methods like integrated gradients or SHAP can be adapted to highlight which input modalities (e.g., a specific histone mark or a transcription factor's IDR sequence) are most influential for a given prediction. However, there is a fundamental tension between model complexity and interpretability: more expressive models that simulate nonlinear dynamics may be harder to explain. A hybrid approach that combines a simple interpretable core (e.g., a logistic regression on key biophysical features) with a deep representation learner for feature extraction might strike an acceptable balance.

5. Structural Trade-Offs and System Robustness

Deploying a foundation model that integrates phase separation biology into precision oncology workflows introduces several structural trade-offs that must be carefully managed. One major trade-off is between model specificity and generalizability. By explicitly modeling phase separation, the model may become highly specialized for diseases where LLPS is known to play a role, such as MYC-driven lymphomas or NUT midline carcinoma involving BRD4. However, this specialization could reduce the model's performance on tumors where phase separation is less relevant, potentially leading to misdiagnosis or inappropriate treatment recommendations. To address this, the model architecture should include a gating mechanism that determines whether phase separation modeling is applicable for a given sample, based on the expression patterns of key LLPS-associated proteins. Such a mechanism would allow the model to fallback to a general multi-omics predictor when phase separation signals are weak, thereby maintaining overall robustness.

Another trade-off concerns the reliance on high-resolution multi-omics data that may not be routinely available in clinical settings. While research-grade datasets often include Hi-C, ChIP-seq, and proteomics, standard clinical workflows typically only produce whole-exome sequencing and bulk RNA-seq. A foundation model that requires all omics modalities for inference will have limited translational utility. A practical solution is to train the model to handle missing modalities through imputation or by learning to reconstruct missing data from available inputs. For instance, given only RNA-seq, the model could predict chromatin accessibility using a learned decoder, and then use that predicted accessibility to infer phase separation dynamics. This approach, known as cross-modal imputation, has been successful in natural language processing and could be adapted here. However, imputation introduces additional uncertainty that must be quantified and communicated to clinicians.

Robustness also encompasses the model's ability to maintain performance across diverse patient populations. Phase separation dynamics can be influenced by genetic ancestry, germline variation in IDR sequences, and environmental exposures that alter the cellular context. If a foundation model is trained predominantly on European-ancestry cell lines, it may underperform in patients of African or Asian ancestry, exacerbating health disparities. To ensure fairness, training datasets must be intentionally diverse, and the model should be evaluated on subgroup-specific metrics during validation. Domain adaptation techniques, such as adversarial debiasing or invariant risk minimization, can help the model learn representations that are predictive but not dependent on sensitive attributes [20]. Moreover, continuous monitoring of model performance across demographic groups after deployment is essential for early detection of drift.

6. Governance, Deployment, and Sustainability

The integration of large AI models into clinical oncology raises profound governance questions. Data privacy is the foremost concern: multi-omics datasets contain highly sensitive genetic and health information that could be re-identified. Federated learning offers a promising framework where models are trained across multiple institutions without centralizing raw data [21]. In this paradigm, each hospital retains its own omics data, computes local model updates, and shares only the gradients (or parameter updates) with a central server. However, federated learning is not immune to privacy attacks, and differential privacy mechanisms must be added to guarantee that no individual patient's data can be inferred from the aggregated updates. The computational overhead of differential privacy can degrade model accuracy, creating another trade-off between privacy and performance.

Regulatory approval of AI-based medical devices is a complex process involving agencies such as the FDA and EMA. Foundation models that are continuously updated with new data present a challenge to traditional pre-market approval pathways, which assume a static model. The concept of a "locked" model versus a "continuously learning" model is debated. For phase separation–aware foundation models, it is plausible that the model will be retrained periodically as new biological discoveries (e.g., new IDR-mediated condensates) are validated. One governance approach is to define a "baseline" version of the model that is submitted for regulatory clearance, while subsequent updates are treated as minor modifications subject to change control protocols. Establishing trust in the model's safety and efficacy requires rigorous prospective clinical validation in prospective trials, where the model's predictions are used to guide treatment decisions (e.g., recommending a BRD4 inhibitor for a patient whose tumor is predicted to rely on phase-separated MYC condensates).

Sustainability of large foundation models is an increasing concern due to their energy consumption and carbon footprint. Training a transformer with billions of parameters can emit as much carbon as several transatlantic flights [22]. In the healthcare domain, where models may need to be retrained frequently on updated data, the environmental impact cannot be ignored. Strategies for sustainable AI include model pruning, quantization, and knowledge distillation to create smaller, energy-efficient versions that retain most of the predictive performance. Additionally, using specialized hardware such as tensor processing units from renewable energy sources can mitigate some environmental costs. From a policy perspective, funding agencies and institutions should require carbon budgeting and reporting for large AI projects in biomedicine.

7. Future Directions and Policy Implications

Looking forward, the convergence of foundation models with mechanistic biology of phase separation offers a pathway toward truly predictive and prescriptive oncology. One promising direction is the development of "digital twins" of tumor cells that incorporate a foundation model as the core inference engine, coupled with a mechanistic simulator of condensate dynamics. Such digital twins could be personalized for each patient using their multi-omics data and used to simulate the effects of various drug combinations in silico before administration. Early efforts in this direction have been demonstrated for signaling pathways, but extending them to phase separation requires further breakthroughs in multiscale simulation and uncertainty quantification.

Policy implications extend beyond model regulation. As foundation models become more capable, they may be used to identify new drug targets by predicting which phase separation processes are essential for tumor survival. Intellectual property rights for AI-discovered targets are currently ambiguous. Should the model itself be considered an inventor? Legal

frameworks need to be updated to address the role of AI in drug discovery. Furthermore, the use of these models in low-resource settings is hindered by the need for high-performance computing infrastructure. International collaborations and open-source model initiatives could democratize access, but careful consideration must be given to data sovereignty and the potential for exploitation of genomic data from under-represented populations.

Finally, the education of future researchers and clinicians must evolve to include training in computational biophysics and AI ethics. The complexity of models that combine deep learning with phase separation theory demands a workforce that can critically evaluate model outputs and understand their limitations. Interdisciplinary training programs that bridge molecular biology, computer science, and health policy will be essential to realize the promise of multi-omics foundation models in precision oncology.

8. Conclusion

Multi-omics foundation models hold immense potential to decode the role of phase separation in transcriptional control and translate that knowledge into precision oncology. However, building such models requires surmounting significant challenges in data integration, architectural design, interpretability, robustness, and governance. By explicitly incorporating the biophysics of liquid-liquid phase separation, these models can provide mechanistic insights that today's sequence-based transformers cannot. This paper has outlined a systems-level framework that addresses structural trade-offs, deployment sustainability, and ethical governance. The path forward demands a collaborative effort among biologists, computer scientists, clinicians, and policymakers to ensure that the resulting tools are both powerful and responsible. With careful design and thoughtful regulation, these models can become a cornerstone of next-generation cancer medicine.

References

1. Shin, Y., Berry, J., Pannucci, N., Haataja, M. P., Toettcher, J. E., & Brangwynne, C. P. (2017). Spatiotemporal control of intracellular phase transitions using light-activated optoDroplets. Cell, 168(1-2), 159-171.

2. 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.

3. 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.

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. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30.

6. Avsec, Ž., Agarwal, V., Visentin, D., Lian, J., Pjanic, M., David, E., ... & Kelley, D. R. (2021). Effective gene expression prediction from sequence by integrating long-range interactions. Nature Methods, 18(10), 1196-1203.

7. Rives, A., Meier, J., Sercu, T., Goyal, S., Lin, Z., Liu, J., ... & Fergus, R. (2021). Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. Proceedings of the National Academy of Sciences, 118(15), e2016239118.

8. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., ... & Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. Nature, 596(7873), 583-589.

9. Dalla-Torre, H., Gonzalez, L., Mendoza-Revilla, J., Carranza, N. L., Grl, M., Constantinou, A., ... & Pierro, M. D. (2023). The nucleotide transformer: Building and evaluating robust foundation models for human genomics. bioRxiv, 2023-01.

10. 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.

11. Cho, W. K., Spille, J. H., Hecht, M., Lee, C., Li, C., Grube, V., & Cisse, I. I. (2018). Mediator and RNA polymerase II clusters associate in transcription-dependent condensates. Science, 361(6400), 412-415.

12. Chong, S., Dugast-Darzacq, C., Liu, Z., Dong, P., Dailey, G. M., Cattoglio, C., ... & Tjian, R. (2018). Imaging dynamic and selective low-complexity domain interactions that control gene transcription. Science, 361(6400), eaar2555.

13. Cihlarova, M., & Vagnerova, M. (2023). The role of DUX4 phase separation in facioscapulohumeral muscular dystrophy and beyond. Nature Reviews Molecular Cell Biology, 24(4), 245-256.

14. Wei, M. T., Elbaum-Garfinkle, S., Holehouse, A. S., Chen, C. C., Feric, M., Arnold, C. B., ... & Brangwynne, C. P. (2017). Phase behaviour of disordered proteins underlying cell signaling. Nature Chemical Biology, 13(9), 958-965.

15. Lotfollahi, M., Naghipourfar, M., Luecken, M. D., Khajavi, M., Büttner, M., Wagenstetter, M., ... & Theis, F. J. (2022). Mapping single-cell data to reference atlases by transfer learning. Nature Biotechnology, 40(1), 121-130.

16. Cao, Z. J., & Gao, G. (2022). Multi-omics single-cell data integration and regulatory inference with graph-linked embedding. Nature Biotechnology, 40(10), 1458-1466.

17. 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.

18. Regy, R. M., Thompson, J., Kim, Y. C., & Mittal, J. (2021). Improved coarse-grained model for studying sequence dependent phase separation of disordered proteins. Protein Science, 30(7), 1371-1382.

19. Kelly, D. R., Snoek, J., & Adams, H. C. (2018). Dynamic genome-scale modeling of transcription factor networks. Cell Systems, 6(4), 461-473.

20. Arjovsky, M., Bottou, L., Gulrajani, I., & Lopez-Paz, D. (2019). Invariant risk minimization. arXiv preprint arXiv:1907.02893.

21. McMahan, B., Moore, E., Ramage, D., Hampson, S., & y Arcas, B. A. (2017). Communication-efficient learning of deep networks from decentralized data.

Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, 54, 1273-1282.

22. Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and policy considerations for deep learning in NLP. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 3645-3650.