AI-Enabled Multi-Omics Modeling of Aberrant Gene Regulatory Programs in Tumor Development

Nikhil Nair
Department of Electrical Engineering and Computer Science, University of Kansas, Lawrence, KS, USA.
nnair@ku.edu

Sanjay Saxena
School of Information Technology, University of Cincinnati, Cincinnati, OH, USA.
sanjay.saxena872@uc.edu

Abstract

The integration of artificial intelligence with multi-omics data has opened transformative avenues for deciphering the complex gene regulatory programs that drive tumor development. This paper presents a systems-level examination of AI-enabled multi-omics modeling, focusing on how machine learning architectures can capture aberrant regulatory mechanisms from genomics, transcriptomics, proteomics, and epigenomics data. We argue that the central challenge lies not only in predictive accuracy but in the structural trade-offs between model interpretability, robustness, fairness, and scalability within socio-technical infrastructures. The paper systematically dissects the architectural choices for integrating heterogeneous omics layers, including graph-based and transformer models, and evaluates their capacity to represent non-linear regulatory interactions such as phase separation and chromatin remodeling. We then discuss the critical governance and policy implications, including data sovereignty, algorithmic bias, and equitable deployment across global health systems. By situating technical modeling within broader infrastructural and ethical considerations, we provide a comprehensive framework for deploying AI-omics systems in translational oncology. The analysis draws on recent advances in deep learning, regulatory genomics, and health informatics, and concludes with forward-looking perspectives on sustainable, fair, and robust AI-driven discovery in cancer biology.

Keywords

multi-omics integration, gene regulatory networks, artificial intelligence, cancer genomics, phase separation, algorithmic fairness, health infrastructure, robustness, governance, precision oncology.

1. Introduction

Cancer arises from the accumulation of somatic mutations that disrupt the finely tuned regulatory programs controlling cell proliferation, differentiation, and death. These disruptions manifest across multiple molecular layers – DNA mutations, altered RNA expression, protein abundance changes, and epigenetic modifications – creating a deeply interconnected system of aberrations. The advent of high-throughput omics technologies has made it possible to profile tumors at unprecedented resolution, yet the sheer dimensionality and heterogeneity of these data pose fundamental challenges for interpretation. Artificial intelligence, particularly deep learning, has emerged as a powerful tool to model the non-linear, high-dimensional relationships inherent in multi-omics datasets. However, the

deployment of AI in this domain is not merely a technical exercise; it involves complex structural trade-offs among model capacity, interpretability, data privacy, and clinical utility. This paper adopts a systems perspective to examine how AI-enabled multi-omics modeling can uncover aberrant gene regulatory programs in tumors, while critically assessing the architectural, infrastructural, and governance dimensions that determine real-world impact.

The urgency of this inquiry is underscored by the rapid accumulation of multi-omics cancer data from large consortia such as The Cancer Genome Atlas (TCGA) [1] and the International Cancer Genome Consortium [2]. These resources have provided a foundational landscape of molecular alterations, but translating these observations into mechanistic understanding of regulatory dysregulation remains a bottleneck. Traditional statistical methods often fail to capture the combinatorial and context-dependent nature of gene regulation, where transcription factors, chromatin states, and non-coding RNAs interact in dynamic, often phase-separated compartments [3]. AI models, especially those based on neural networks, have demonstrated remarkable ability to learn representations that approximate these interactions, yet they also introduce concerns about overfitting, reproducibility, and fairness when applied to diverse patient populations [4]. Therefore, a systematic evaluation of architectural choices, validation strategies, and deployment frameworks is essential for responsible innovation.

2. Multi-Omics Data Integration and AI Architectures

The first major challenge in modeling aberrant gene regulatory programs is the effective integration of disparate omics layers. Genomic data provide static DNA variants, transcriptomic data capture RNA expression levels, proteomic data measure protein abundance and post-translational modifications, and epigenomic data profile DNA methylation, histone modifications, and chromatin accessibility. Each layer has distinct noise characteristics, missingness patterns, and measurement scales. AI architectures must therefore be designed to fuse these heterogeneous inputs while preserving biological interpretability. Early approaches used concatenation of features followed by standard neural networks, but such methods ignore the structured relationships among omics layers [5]. More recently, graph neural networks (GNNs) have been applied to model the multi-layered network of molecular interactions, where nodes represent genes, proteins, or regulatory elements, and edges represent known or inferred relationships [6]. These models can propagate information across layers, capturing how a mutation in a coding region may influence protein expression or how an epigenetic change can alter transcription factor binding.

Another promising architecture is the transformer model, originally developed for natural language processing, which has been adapted to genomic sequences and expression profiles [7]. Transformers use self-attention mechanisms to capture long-range dependencies, making them suitable for modeling regulatory elements that are far apart in the linear genome or that interact across three-dimensional chromatin loops [8]. Hybrid models that combine convolutional layers for local sequence features with attention mechanisms for global context have shown improved performance in predicting enhancer-promoter interactions and transcription factor binding sites [9]. However, a critical trade-off exists: more expressive models often require larger datasets and more computational resources, raising questions about their sustainability in resource-constrained research environments. Moreover, the black-box nature of deep transformers complicates the biological interpretation of learned representations, which is essential for generating testable hypotheses about regulatory mechanisms.

Beyond architecture, the integration strategy itself must account for batch effects, platform differences, and data quality. Multi-omics integration methods such as MOFA (Multi-Omics Factor Analysis) and scVI (single-cell variational inference) offer principled ways to learn latent representations that factorize variation across omics layers [10, 11]. When combined with AI-driven imputation and normalization pipelines, these methods can enhance robustness but also introduce dependencies on prior assumptions about data distributions. The choice between early, intermediate, and late fusion strategies has practical implications for model generalization and interpretability. Early fusion builds a joint representation from raw features, intermediate fusion learns separate representations for each omics layer before combining them, and late fusion aggregates predictions from independent models. Each approach has distinct advantages: early fusion captures cross-omics interactions directly but may suffer from incompatible scales; intermediate fusion preserves layer-specific structure but requires careful alignment; late fusion simplifies training but may miss synergistic signals [12]. From a systems perspective, the optimal fusion strategy depends on the specific biological question and the available data infrastructure, highlighting the need for modular pipelines that can be adapted to different use cases.

3. Modeling Aberrant Gene Regulatory Programs

Gene regulatory programs in tumors are characterized by the rewiring of transcription factor networks, loss of tumor suppressor functions, and activation of oncogenic pathways. AI models can learn these programs by mapping multi-omics inputs to regulatory outputs, such as expression levels of target genes, promoter activity, or chromatin state. One influential approach is to treat gene regulation as a regression or classification problem, where deep neural networks predict expression from genomic sequence features and epigenetic marks [13]. These models have successfully identified key regulatory mutations that alter transcription factor binding sites, but they often fail to capture the dynamic, often stochastic nature of regulatory interactions. Recent advances in single-cell multi-omics have added a temporal and spatial dimension, revealing that regulatory programs are highly cell-type specific and exhibit plasticity during tumor progression [14]. Modeling such complexity requires AI systems that can handle sparse, high-dimensional single-cell data while accounting for cellular heterogeneity and technical noise.

A particularly intriguing aspect of aberrant regulation is the role of phase separation, wherein transcription factors and co-activators form liquid-liquid phase-separated condensates at super-enhancers to drive high-level expression of oncogenes [15]. This mechanism adds a layer of regulatory control that is not easily captured by conventional sequence-based models. Experimental studies have shown that phase separation of the YAP-MAML2 fusion protein differentially regulates the transcriptome in a context-dependent manner [16]. For AI models to incorporate such phenomena, they must learn representations of protein biophysical properties, local concentration gradients, and spatial organization within the nucleus. Graph neural networks with attention mechanisms can, in principle, model short-range and long-range interactions that mimic condensate formation, but the training data needed to learn these patterns remain scarce. Future integration of imaging-based spatial transcriptomics with molecular omics could provide the necessary spatial resolution, yet this poses new challenges in data fusion and computational scalability [17].

The inference of causal regulatory relationships is another frontier. Association-based models can identify correlations between a mutation and expression change, but they do not distinguish cause from effect. Causal inference frameworks, when combined with AI, can

leverage interventional data from CRISPR screens or perturbation experiments to infer directed regulatory edges [18]. Such approaches require careful experimental design and large-scale perturbation data, which are not yet available for most tumor types. Nevertheless, the potential to model causal regulatory programs opens the door to identifying driver events that are actionable for therapy. From an infrastructural perspective, building causally-aware AI systems demands tight integration with experimental platforms, data repositories that capture perturbations, and governance mechanisms to ensure reproducibility and data provenance.

4. Structural Trade-offs and Robustness

Every AI-driven multi-omics model faces fundamental trade-offs between predictive performance, interpretability, robustness, and fairness. A highly accurate deep learning model may rely on spurious correlations that do not generalize across patient cohorts or experimental conditions. For example, models trained on data from predominantly Caucasian populations may fail to capture regulatory variants common in African or Asian ancestries, leading to biased predictions and exacerbating health disparities [19]. Robustness to batch effects, sequencing depth, and library preparation protocols is also a major concern. Techniques such as data augmentation, adversarial training, and domain adaptation can improve out-of-distribution generalization but often at the cost of reduced performance on the primary task or increased computational overhead [20].

Interpretability is especially critical in oncology, where clinicians and researchers need to understand why a model identifies a particular gene regulatory program as aberrant. Post-hoc explanation methods such as SHAP (SHapley Additive exPlanations) and integrated gradients can highlight important features, but they provide only local approximations and may be inconsistent across model architectures [21]. Architectures that are inherently interpretable, such as sparse linear models or rule-based systems, are less capable of capturing the non-linear interactions characteristic of gene regulation. This trade-off suggests that no single model is optimal for all use cases. Instead, a portfolio of models with different levels of interpretability and accuracy should be deployed, with appropriate validation thresholds depending on the downstream decision – whether for basic discovery, biomarker development, or clinical decision support.

Robustness also encompasses model stability under distribution shifts caused by evolving tumor heterogeneity over time. Longitudinal multi-omics studies are rare, but emerging evidence indicates that regulatory programs can change dramatically during treatment and metastasis [22]. AI models trained on static snapshots may therefore be brittle when applied to progressive disease. Ensemble methods and continual learning techniques offer partial solutions, but they require careful management of model versions and data pipelines. From a socio-technical perspective, the robustness of AI-omics systems depends not only on algorithmic choices but also on the quality and diversity of the underlying data infrastructure. Investments in multi-institutional data sharing, standardized metadata, and federated learning frameworks can enhance robustness while respecting privacy regulations [23].

5. Governance, Fairness, and Policy Implications

The deployment of AI-enabled multi-omics modeling in cancer research and clinical practice raises profound governance and fairness questions. Genomic data are inherently sensitive, containing information about ancestry, disease risk, and family relationships. The use of AI to infer aberrant gene regulatory programs can inadvertently reveal latent information about

individuals or populations, even when data are de-identified [24]. Robust data governance frameworks are needed to ensure informed consent, data sovereignty, and equitable benefit sharing. Federated learning, which trains models across distributed data without centralizing raw data, is a promising technical approach, but it introduces new challenges in model auditing and bias detection across sites with varying data distributions [25].

Algorithmic fairness in multi-omics AI requires careful attention to representation in training datasets. Underrepresented populations may have distinct regulatory architectures that are not captured by models biased toward common variants in well-studied groups. As a result, predictions of gene regulatory aberrations could be systematically less accurate for minority patients, potentially leading to misdiagnosis or inappropriate treatment recommendations [19]. Policy interventions, such as inclusion mandates for diverse cohorts in large-scale omics projects and algorithmic impact assessments before clinical deployment, are necessary to mitigate these risks. Additionally, the interpretability of AI models must be accessible to clinicians, regulators, and patients, requiring plain-language explanations and audit trails that document how predictions were generated.

International governance also plays a role, as multi-omics data often cross borders through collaborative research. Different countries have varying regulations on data sharing, privacy (e.g., GDPR in Europe, HIPAA in the U.S.), and secondary use of biological samples. Harmonizing these frameworks to enable global AI training without compromising ethical standards is a monumental task. Technical infrastructure for privacy-preserving computation, such as differential privacy and secure multi-party computation, must be integrated into multi-omics pipelines [26]. Yet, these techniques often degrade model accuracy or increase computational cost, creating another trade-off that must be navigated through policy-driven resource allocation. Ultimately, the responsible advancement of AI-enabled multi-omics modeling requires a tripartite commitment from researchers, funders, and regulators to embed equity and accountability into every layer of the system.

6. Deployment and Sustainability in Clinical and Research Infrastructures

Translating AI-omics models from the research bench to clinical and research infrastructures demands careful consideration of deployment sustainability. Many current models are developed on powerful GPU clusters with ample memory, but clinical settings often have limited computing resources, strict latency requirements, and security constraints that preclude cloud-based inference. Edge computing solutions, where models run on local servers or even on instrument-integrated processors, can reduce latency and protect data privacy, but they may also limit model complexity due to hardware constraints [27]. Model compression techniques, such as quantization, pruning, and knowledge distillation, can reduce the footprint of deep learning models while preserving acceptable performance. However, these compression methods must be validated on multi-omics data to ensure that biologically relevant features are not lost.

Sustainability also encompasses the long-term maintenance of AI systems. Omics technologies evolve rapidly; a model trained on RNA-seq data from one sequencing platform may not generalize to data from newer long-read sequencers. Continuous retraining and monitoring are required, but these activities demand dedicated personnel, funding, and infrastructure management. The research community must develop standardized benchmarks and leaderboards for multi-omics AI to facilitate reproducibility and fair comparison. Open-source platforms and reproducibility checklists can lower barriers for new institutions, but

they require active community governance to prevent the proliferation of poorly validated models [28].

From an infrastructural perspective, the integration of AI-omics pipelines into electronic health records (EHRs) and clinical decision support systems (CDSS) represents a high-impact opportunity. However, interoperability standards between omics databases and EHRs remain underdeveloped. Projects such as the GA4GH (Global Alliance for Genomics and Health) have established data exchange standards, but adoption is uneven across health systems [29]. Furthermore, clinicians must be trained to interpret AI-generated regulatory insights, which often require molecular biology expertise beyond typical medical education. Certification and continuing education programs, coupled with user-friendly visualization tools, are necessary to bridge the gap. The deployment of AI-enabled multi-omics modeling should be viewed as a socio-technical transformation, not just a technological upgrade, requiring investments in human capital, institutional policies, and collaborative networks.

7. Conclusion

AI-enabled multi-omics modeling holds immense promise for uncovering the aberrant gene regulatory programs that drive tumor development. By integrating heterogeneous molecular data through advanced architectures such as graph neural networks and transformers, researchers can capture complex, non-linear interactions spanning DNA, RNA, protein, and epigenetic layers. However, this paper has emphasized that the pathway to reliable and equitable impact is fraught with structural trade-offs. Model interpretability must be balanced against predictive capacity; robustness against distribution shifts and batch effects requires careful validation strategies; fairness demands inclusive data representation and algorithmic auditing; and governance must ensure privacy, consent, and equitable access. The deployment of these systems in clinical and research infrastructures further highlights the need for sustainable, interoperable, and user-centered designs. As the field moves forward, interdisciplinary collaboration among computational scientists, biologists, clinicians, ethicists, and policymakers will be essential to build systems that are not only technically powerful but also socially responsible. The next generation of AI-omics research should prioritize the development of causally-aware, privacy-preserving, and globally validated models, grounded in open standards and equitable governance frameworks. Only then can the full potential of multi-omics data be harnessed to improve cancer understanding and patient outcomes across diverse populations.

References

1. Cancer Genome Atlas Research Network. (2013). The Cancer Genome Atlas Pan-Cancer analysis project. Nature Genetics, 45(10), 1113–1120.

2. International Cancer Genome Consortium. (2010). International network of cancer genome projects. Nature, 464(7291), 993–998.

3. 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.

4. Rajkomar, A., Dean, J., & Kohane, I. (2019). Machine learning in medicine. New England Journal of Medicine, 380(14), 1347–1358.

5. Rappoport, N., & Shamir, R. (2018). Multi-omic and multi-view clustering algorithms: review and cancer benchmark. Nucleic Acids Research, 46(20), 10546–10562.

6. Zhang, J., Luo, Y., & Peng, J. (2019). Graph neural networks for gene expression and drug response prediction. Bioinformatics, 35(14), 2474–2482.

7. 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.

8. Rao, S. S., Huntley, M. H., Durand, N. C., Stamenova, E. K., Bochkov, I. D., Robinson, J. T., ... & Aiden, E. L. (2014). A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell, 159(7), 1665–1680.

9. Kelley, D. R., Snoek, J., & Rinn, J. L. (2016). Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. Genome Research, 26(7), 990–999.

10. Argelaguet, R., Velten, B., Arnol, D., Dietrich, S., Zenz, T., Marioni, J. C., ... & Stegle, O. (2018). Multi-Omics Factor Analysis—a framework for unsupervised integration of multi-omics data sets. Molecular Systems Biology, 14(6), e8124.

11. Lopez, R., Regier, J., Cole, M. B., Jordan, M. I., & Yosef, N. (2018). Deep generative modeling for single-cell transcriptomics. Nature Methods, 15(12), 1053–1058.

12. Zitnik, M., Nguyen, F., Wang, B., Leskovec, J., Goldenberg, A., & Hoffman, M. M. (2019). Machine learning for integrating data in biology and medicine: Principles, practice, and opportunities. Information Fusion, 50, 71–91.

13. Zhou, J., & Troyanskaya, O. G. (2015). Predicting effects of noncoding variants with deep learning–based sequence model. Nature Methods, 12(10), 931–934.

14. Trapnell, C. (2015). Defining cell types and states with single-cell genomics. Genome Research, 25(10), 1491–1498.

15. 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.

16. Chung, C. I., Yang, J., Yang, X., Liu, H., Ma, Z., Szulzewsky, F., ... & Shu, X. (2024). Phase separation of YAP-MAML2 differentially regulates the transcriptome. Proceedings of the National Academy of Sciences, 121(7), e2310430121.

17. 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.

18. Dixit, A., Parnas, O., Li, B., Chen, J., Fulco, C. P., Jerby-Arnon, L., ... & Regev, A. (2016). Perturb-Seq: dissecting molecular circuits with scalable single-cell RNA profiling of pooled genetic screens. Cell, 167(7), 1853–1866.

19. Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. Science, 366(6464), 447–453.

20. Geirhos, R., Temme, C. R., Rauber, J., Schütt, H. H., Bethge, M., & Wichmann, F. A. (2018). Generalisation in humans and deep neural networks. Advances in Neural Information Processing Systems, 31.

21. Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30.

22. Dagogo-Jack, I., & Shaw, A. T. (2018). Tumour heterogeneity and resistance to cancer therapies. Nature Reviews Clinical Oncology, 15(2), 81–94.

23. Rieke, N., Hancox, J., Li, W., Milletari, F., Roth, H. R., Albarqouni, S., ... & Cardoso, M. J. (2020). The future of digital health with federated learning. NPJ Digital Medicine, 3(1), 119.

24. Gymrek, M., McGuire, A. L., Golan, D., Halperin, E., & Erlich, Y. (2013). Identifying personal genomes by surname inference. Science, 339(6117), 321–324.

25. 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.

26. Dwork, C., & Roth, A. (2014). The algorithmic foundations of differential privacy. Foundations and Trends in Theoretical Computer Science, 9(3–4), 211–407.

27. Chen, J., & Ran, X. (2019). Deep learning with edge computing: A review. Proceedings of the IEEE, 107(8), 1655–1674.

28. Beaulieu-Jones, B. K., & Greene, C. S. (2017). Reproducibility of computational workflows is automated using continuous analysis. Nature Biotechnology, 35(4), 342–346.

29. Global Alliance for Genomics and Health. (2016). A federated ecosystem for sharing genomic, clinical, and phenotypic data. Science, 352(6291), 1278–1280.