AI-Enabled Multi-Omics Integration for Characterizing Dynamic Gene Expression Programs in Tumor Cells

Shane Sanders
Department of Computer Science and Engineering, University at Buffalo, Buffalo, NY, USA.
shane2002@buffalo.edu

Ross Gregory
Department of Computer Science, Colorado State University, Fort Collins, CO, USA.
rossgregory92@colostate.edu

Vishal L. Pandey
Department of Computer Science and Engineering, University of Nevada, Reno, Reno, NV, USA.
vishallpandey492@unr.edu

Varun Subramanian
Department of Computer Science, Binghamton University, Binghamton, NY, USA.
vsubramanian@binghamton.edu

Abstract

The rapid accumulation of multi-omics data from tumor samples has created an unprecedented opportunity to understand the dynamic gene expression programs that drive cancer progression, metastasis, and treatment resistance. However, the high dimensionality, heterogeneity, and temporal sparsity of such data present fundamental computational challenges that conventional statistical methods cannot address. This paper presents a comprehensive systems-level examination of how artificial intelligence, particularly deep learning architectures, can be harnessed to integrate multi-omics layers for the characterization of dynamic gene expression programs in tumor cells. We explore the architectural design space of integrative models, including autoencoders, graph neural networks, and transformer-based approaches, and analyze the structural trade-offs between predictive accuracy, interpretability, and computational cost. Infrastructure considerations such as distributed computing, data storage, and energy consumption are discussed in the context of sustainability and scalability. We further examine critical issues of robustness and fairness, focusing on how biases in training data, model calibration across heterogeneous patient populations, and adversarial vulnerabilities can undermine clinical translation. Governance and policy implications are addressed through the lens of regulatory frameworks for AI-driven diagnostics and the ethical deployment of omics-level predictions. By synthesizing methodological advances with socio-technical challenges, this paper provides a roadmap for the responsible integration of AI-enabled multi-omics systems into precision oncology.

Keywords

multi-omics integration, artificial intelligence, dynamic gene expression, tumor transcriptomics, deep learning architecture, systems biology, precision oncology, fairness, sustainability, governance.

1. Introduction

Cancer is fundamentally a disease of dysregulated gene expression, where alterations at the genomic, epigenomic, transcriptomic, proteomic, and metabolomic levels collectively drive aberrant cellular programs. The advent of high-throughput sequencing technologies has enabled the generation of multi-omics profiles from individual tumor samples, offering a multidimensional snapshot of the molecular state of a cell. Yet the dynamic nature of gene expression programs, which evolve over developmental stages, in response to therapy, and during metastatic dissemination, cannot be captured by static measurements alone. Characterizing these temporal trajectories requires the integration of data across omics layers and time points, a task that has proven intractable with traditional biostatistical approaches due to the combinatorial explosion of interactions and the sparsity of longitudinal cohorts [1].

Artificial intelligence, particularly deep learning, has emerged as a transformative paradigm for modeling complex, high-dimensional biological data. Neural network architectures can learn hierarchical representations that capture nonlinear dependencies across genomics, transcriptomics, epigenomics, and proteomics, enabling the reconstruction of regulatory networks and the prediction of expression dynamics under perturbations [2]. However, the deployment of AI in multi-omics integration is not merely a technical optimization problem. It raises profound questions about the structural design of models, the sustainability of computational infrastructures, the fairness of predictions across diverse populations, and the governance frameworks required to ensure safe clinical translation [3]. This paper adopts a systems-level perspective, examining the interplay between algorithmic innovation, infrastructure constraints, and socio-technical governance in the context of AI-enabled multi-omics integration for dynamic gene expression characterization.

The remainder of this paper is organized as follows. Section 2 reviews the landscape of multi-omics data and the specific challenges of integrating heterogeneous, sparse, and temporally misaligned measurements. Section 3 surveys AI architectures that have been applied to this problem, highlighting representational strengths and limitations. Section 4 analyzes structural trade-offs in model design, including accuracy versus interpretability, generality versus specialization, and batch size versus memory footprint. Section 5 addresses infrastructure and sustainability considerations, such as energy consumption, data federation, and cloud deployment. Section 6 turns to robustness, fairness, and governance, exploring algorithmic bias, distribution shift, and regulatory policy. Section 7 discusses future directions and policy frameworks, and Section 8 concludes with a synthesis of key findings.

2. Multi-Omics Data Landscapes and Integration Challenges

The molecular characterization of tumor cells now routinely encompasses whole-genome sequencing, RNA sequencing, chromatin immunoprecipitation sequencing, assay for transposase-accessible chromatin sequencing, proteomic mass spectrometry, and metabolomic profiling. Each omics layer captures a distinct aspect of cellular function, yet their interplay determines the emergent behavior of gene expression programs. For instance, copy number alterations can influence transcript abundance, while epigenetic modifications modulate chromatin accessibility and thus transcriptional activity. Integrating these layers into a coherent model requires addressing several deep challenges: data heterogeneity in scale and

distribution, missingness due to sample constraints, batch effects across experiments, and the lack of temporal alignment when samples are collected at different disease stages [4].

Gene expression programs are inherently dynamic, exhibiting oscillatory, switch-like, or stochastic patterns that traditional bulk sequencing methods average over populations of cells. Single-cell RNA sequencing has partially resolved this issue, but it introduces additional noise, dropout events, and sparse coverage of the transcriptome. Temporal multi-omics studies, where serial biopsies are obtained from the same patient, remain rare due to clinical impracticality, leading to severe underdetermination of the underlying dynamical system [5]. Furthermore, the integration of static omics measurements with time-stamped clinical outcomes, such as progression-free survival, requires methods that can infer latent trajectories from cross-sectional or partially observed longitudinal data. AI models offer the capacity to learn such latent dynamics by imposing regularities, such as conservation of regulatory motifs or coupling between omics layers, but they also risk encoding spurious correlations if not carefully validated [6].

3. AI Architectures for Dynamic Gene Expression Modeling

A diverse set of neural network architectures has been adapted for multi-omics integration. Variational autoencoders (VAEs) have been widely used to learn a low-dimensional latent representation that captures the joint variability across omics layers, enabling the imputation of missing modalities and the generation of synthetic samples [7]. Graph neural networks (GNNs) model the regulatory relationships between genes, proteins, and metabolites as a heterogeneous graph, where nodes represent molecular entities and edges denote known or inferred interactions. GNNs can propagate information across the graph to predict gene expression levels from genomic and epigenomic features, and they can incorporate temporal edges to model dynamics [8].

Transformer-based architectures, originally developed for natural language processing, have been adapted to model sequential dependencies in gene expression time series. By treating each time point as a token and using self-attention mechanisms, transformers can capture long-range temporal correlations and interactions between omics layers without the computational bottlenecks of recurrent networks. Recent work has demonstrated that transformers pre-trained on large corpora of bulk and single-cell transcriptomes can be fine-tuned to predict expression dynamics under drug perturbations, offering a transfer-learning paradigm that reduces the need for extensive temporal training data [9]. However, the attention matrices in such models scale quadratically with the number of features, posing memory constraints that necessitate careful architectural pruning.

4. Structural Trade-Offs in Model Design and Deployment

The design of AI systems for multi-omics integration involves multiple trade-offs that have direct implications for their practical utility. A primary trade-off exists between model capacity and interpretability. Deep, densely connected networks can achieve high predictive accuracy but operate as black boxes, making it difficult to attribute predictions to specific molecular features or to validate biological mechanisms. Conversely, sparser models with explicit regulatory modules, such as those based on graph neural networks or attention mechanisms, offer greater interpretability but may underperform on complex nonlinear interactions [10]. This tension is especially acute in clinical settings, where regulatory agencies require transparency and the ability to audit model decisions.

Another critical trade-off is between generality and specialization. Models trained on large pan-cancer cohorts can capture broad patterns of gene expression dysregulation but may fail to capture tumor-type-specific dynamics. Specialized models trained on a single cancer type or even a single patient cohort achieve higher local accuracy but lack transferability, requiring retraining or fine-tuning for each new application. The choice between these strategies depends on the deployment context: population-level screening tools favor generality, while personalized treatment planning demands specialization [11].

Computational resource trade-offs also arise. Modern deep learning models require substantial GPU memory and energy, particularly when training on multi-omics datasets with millions of features. Batch normalization, gradient checkpointing, and mixed-precision training can mitigate some of these demands, but they introduce additional hyperparameters and can affect convergence stability. The trade-off between model accuracy and energy efficiency is increasingly relevant as sustainability becomes a priority in large-scale biomedical computing [12].

5. Infrastructure, Scalability, and Sustainability Considerations

Deploying AI-enabled multi-omics integration at scale demands a robust computational infrastructure capable of handling data ingestion, preprocessing, model training, inference, and versioning. Cloud platforms offer elastic resources, but the movement of large genomic datasets across networks raises bandwidth and privacy concerns. Federated learning has been proposed as a paradigm to train models across multiple institutions without centralizing raw data, preserving patient privacy while still benefiting from diverse cohorts. However, federated learning introduces communication overhead and statistical heterogeneity across nodes, which can degrade model performance and require sophisticated aggregation strategies [13]. Notably, recent work on phase separation in MYC-driven transcription underscores the importance of understanding how intrinsically disordered regions of transcription factors can influence gene expression at the level of condensate formation, a phenomenon that may require integration of structural biology data with omics layers in future modeling frameworks [13].

Sustainability is an often-overlooked dimension of AI deployment in genomics. Training a single large transformer model can emit several tons of carbon dioxide equivalent, and the cumulative environmental impact of thousands of models trained for various cancer studies is non-trivial. Life cycle assessments of computational pipelines are rarely performed, yet as AI becomes embedded in clinical workflows, the energy cost per prediction will become a factor in hospital budgets and carbon reduction targets. Hardware innovations such as neuromorphic chips or analog computing may eventually reduce energy footprints, but near-term strategies include model compression, quantization, and early stopping based on validation curves [14].

6. Robustness, Fairness, and Governance Implications

Robustness of AI models to distributional shift is a major concern for clinical translation. Tumor samples from different laboratories, sequencing platforms, or patient demographics may exhibit systematic biases that are not present in training data. Models that rely on subtle batch effects rather than true biological signals can fail catastrophically when applied to new cohorts. Domain adaptation techniques, such as adversarial training and normalization layers that correct for batch effects, have been developed but often require a subset of labeled target data for calibration [15]. Furthermore, adversarial attacks that perturb input omics data by

small amounts, undetectable by human annotators, can cause large prediction errors, raising safety concerns for AI-driven treatment recommendations [16].

Fairness in multi-omics AI systems demands that predictions are equally accurate across racial, ethnic, and socioeconomic groups. Genomic databases are disproportionately populated with samples from individuals of European ancestry, leading to models that perform poorly for underrepresented populations. This disparity can exacerbate existing health inequities in cancer care. Mitigation strategies include stratified cross-validation, reweighting of training examples, and the development of ancestry-aware models that explicitly account for population structure [17]. However, these technical fixes are insufficient without policy frameworks that mandate diversity in training data and require continuous monitoring of model performance across demographic subgroups.

Governance structures for AI in oncology are still nascent. The U.S. Food and Drug Administration has issued guidance on adaptive algorithms, but the dynamic nature of multi-omics models—which may be retrained as new data accrue—poses regulatory challenges. A model that performs well at deployment may degrade over time due to shifts in laboratory protocols or patient populations. Continuous validation and algorithmic auditing are necessary, yet the computational cost of frequent re-evaluation can be prohibitive. Transparent reporting standards, such as model cards and datasheets, have been advocated to document training data composition, intended use, and known limitations [18]. These governance tools must be integrated into the development lifecycle of AI-enabled multi-omics systems.

7. Future Directions and Policy Frameworks

The next generation of AI models for dynamic gene expression integration will likely incorporate temporal graph neural networks that can propagate information across both spatial (regulatory network) and temporal axes, enabling the simulation of perturbation experiments in silico. Such models could predict how a tumor's expression program evolves under a given therapy, guiding adaptive treatment strategies. However, the validation of these predictions will require prospective clinical trials that collect longitudinal multi-omics samples, a significant logistical and financial undertaking. Policy incentives, such as funding mandates from national cancer institutes for data sharing and integration, can accelerate this process [19].

Another promising direction is the use of generative AI to create synthetic temporal multi-omics data, augmenting sparse real cohorts. Large language models trained on biomedical literature could also be used to extract prior knowledge about gene regulatory interactions, serving as a structured prior for neural networks. The hallucination risk of such models must be carefully managed through fact-checking and expert review [20].

From a governance standpoint, the adoption of international standards for multi-omics data formats, metadata, and model interoperability will be critical. Without common data models, integrative AI systems cannot be easily transferred across institutions or countries. The Global Alliance for Genomics and Health has advanced some standards, but implementation remains uneven [21]. Regulatory bodies should work with academic consortia and industry to develop certification programs for AI-enabled omics tools, similar to the Clinical Laboratory Improvement Amendments for laboratory tests.

8. Conclusion

AI-enabled multi-omics integration holds immense promise for characterizing the dynamic gene expression programs that underlie tumor biology. By leveraging deep learning architectures that can model nonlinear interactions across genomic, transcriptomic, epigenomic, and proteomic layers, researchers can infer latent temporal trajectories and predict cellular responses to perturbations. However, the successful translation of these methods into clinical practice depends on navigating complex structural trade-offs in model design, ensuring the robustness and fairness of predictions, building sustainable computational infrastructure, and establishing governance frameworks that promote transparency and equity. This systems-level examination reveals that the path forward is as much about socio-technical innovation as it is about algorithmic advancement. Interdisciplinary collaboration among computer scientists, biologists, clinicians, ethicists, and policymakers is essential to realize the full potential of AI in dynamic multi-omics characterization of cancer.

References

1. 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.

2. Eraslan, G., Avsec, Ž., Gagneur, J., & Theis, F. J. (2019). Deep learning: new computational modelling techniques for genomics. Nature Reviews Genetics, 20(7), 389–403.

3. Vayena, E., Blasimme, A., & Cohen, I. G. (2018). Machine learning in medicine: Addressing ethical challenges. PLOS Medicine, 15(11), e1002689.

4. Hawe, J. S., Theis, F. J., & Heinig, M. (2022). Inferring interaction networks from multi-omics data. Frontiers in Genetics, 13, 856245.

5. Weinreb, C., Wolock, S., Tusi, B. K., Socolovsky, M., & Klein, A. M. (2018). Fundamental limits on dynamic inference from single-cell snapshots. Proceedings of the National Academy of Sciences, 115(10), E2467–E2476.

6. Lotfollahi, M., Wolf, F. A., & Theis, F. J. (2019). scGen predicts single-cell perturbation responses. Nature Methods, 16(8), 715–721.

7. Gayoso, A., Steier, Z., Lopez, R., Regier, J., Nazor, K. L., Streets, A., & Yosef, N. (2022). A Python library for probabilistic analysis of single-cell omics data. Nature Biotechnology, 40(2), 163–166.

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

9. Theodoris, C. V., Xiao, L., Chopra, A., Chaffin, M. D., Al Sayed, Z. R., Hill, M. C., ... & Ellinor, P. T. (2023). Transfer learning enables predictions in network biology. Nature, 618(7965), 616–624.

10. Carvalho, D. V., Pereira, E. M., & Cardoso, J. S. (2019). Machine learning interpretability: A survey on methods and metrics. Electronics, 8(8), 832.

11. Wang, T., Huang, J., Yu, T., Xu, J., & Li, S. (2023). Generalizable and transferable gene expression predictions using a transformer model. Nature Communications, 14, 5563.

12. 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.

13. 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.

14. Patterson, D., Gonzalez, J., Holzle, U., & Le, Q. (2021). The carbon footprint of machine learning training will plateau, then shrink. Computer, 54(8), 18–28.

15. Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., ... & Lempitsky, V. (2016). Domain-adversarial training of neural networks. Journal of Machine Learning Research, 17(59), 1–35.

16. Finlayson, S. G., Bowers, J. D., Ito, J., Zittrain, J. L., Beam, A. L., & Kohane, I. S. (2019). Adversarial attacks on medical machine learning. Science, 363(6433), 1287–1289.

17. 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.

18. Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., ... & Gebru, T. (2019). Model cards for model reporting. Proceedings of the Conference on Fairness, Accountability, and Transparency, 220–229.

19. Hinkson, I. V., Davidsen, T., Hoadley, K. A., Morin, G. B., Mesirov, J. P., & Mills, G. B. (2017). A comprehensive genomic characterization of the human tumor microenvironment. Cell, 170(5), 916–929.

20. Singhal, K., Azizi, S., Tu, T., Mahdavi, S. S., Wei, J., Chung, H. W., ... & Natarajan, V. (2023). Large language models encode clinical knowledge. Nature, 620(7972), 172–180.

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