Digital Twin Modeling of MYC-Dependent Transcriptional Condensates for Personalized Cancer Therapeutics

Yuelei Liu
Department of Computer Science, University of New Hampshire, Durham, NH, USA.
yuelei.liu51@unh.edu

Gao Zhen Tian
Department of Computer Science, University of North Texas, Denton, TX, USA.
gaotian721@unt.edu

Abstract

The emergence of digital twin technology offers a transformative paradigm for modeling complex biological systems at the interface of molecular dynamics, large-scale data integration, and personalized medicine. This paper presents a conceptual and architectural framework for a digital twin system that captures the behavior of MYC-dependent transcriptional condensates, which are phase-separated assemblies that modulate gene expression in cancer. The proposed system integrates multi-omics data, real-time patient monitoring, biophysical simulations, and machine learning to create a dynamic, patient-specific replica of the transcriptional condensate network. Emphasis is placed on system-level design choices, structural trade-offs between model fidelity and computational tractability, governance of data and algorithmic fairness, and sustainability of deployment across clinical infrastructures. The analysis draws on recent experimental evidence that MYC undergoes phase separation to selectively influence the transcriptome [6], linking this discovery to scalable digital twin architectures. Cross-domain comparisons with digital twin applications in aerospace and manufacturing are used to highlight unique challenges in biological systems, such as stochasticity, multiscale coupling, and ethical constraints. The paper also addresses policy implications regarding data ownership, algorithmic transparency, and equitable access to personalized cancer therapeutics. By situating the digital twin as a socio-technical infrastructure, this work provides a roadmap for translational research that balances innovation with responsible stewardship.

Keywords

digital twin, MYC, transcriptional condensates, phase separation, personalized medicine, systems biology, artificial intelligence, clinical governance, multi-omics integration.

1. Introduction

The concept of a digital twin, originally developed for industrial system monitoring and predictive maintenance, has increasingly been adopted in healthcare contexts to model patient physiology, disease progression, and treatment response [1]. A digital twin is a virtual representation of a physical system that is continuously updated with real-time data, enabling simulation, prediction, and optimization. In oncology, the complexity of tumor heterogeneity, microenvironment dynamics, and signaling network plasticity demands a computational framework that can integrate disparate data sources and generate actionable insights at the

individual patient level [2]. Recent advances in the understanding of transcriptional regulation by the MYC oncoprotein have revealed a novel mechanism involving liquid-liquid phase separation that forms condensates at super-enhancer regions, selectively modulating gene expression programs that drive cell proliferation and survival [6]. These condensates represent a new class of therapeutic targets, but their dynamic assembly and function are influenced by a multitude of factors including protein concentration, post-translational modifications, and chromatin context. A digital twin model that captures the spatiotemporal behavior of MYC-dependent transcriptional condensates could enable personalized prediction of drug sensitivity, resistance, and optimal combination therapies.

This paper develops a system-level perspective on building such a digital twin, focusing not on the detailed biophysical equations but on the architectural decisions, data pipelines, governance structures, and deployment strategies that will determine its feasibility and impact. The discussion is organized as follows. Section 2 provides the conceptual background linking phase separation of MYC to transcriptional control and outlines the rationale for a digital twin approach. Section 3 describes the proposed system architecture, including modular components for data ingestion, simulation, inference, and interaction. Section 4 discusses data integration and model calibration challenges, emphasizing trade-offs between model resolution and scalability. Section 5 examines the computational infrastructure required, including cloud-edge hybrid systems and uncertainty quantification. Section 6 addresses governance, ethical considerations, and sustainability, with attention to fairness in personalized therapeutics. Section 7 presents case illustrations and cross-domain comparisons to illuminate advantages and pitfalls. Section 8 concludes with forward-looking perspectives on the role of digital twins in precision oncology.

2. Background and Conceptual Framework

Transcriptional condensates are biomolecular assemblies formed through phase separation of transcription factors, coactivators, and RNA polymerase II at genomic loci enriched with enhancer elements [3]. The MYC protein, a master regulator of cell growth and metabolism, is frequently overexpressed in cancers and has been shown to undergo phase separation to concentrate transcriptional machinery at target genes [6]. This condensation behavior is sensitive to the local concentration of MYC and its binding partners, as well as to small molecules that alter protein interaction surfaces. Understanding how these condensates dynamically form, dissolve, and respond to perturbations is essential for designing therapies that disrupt aberrant transcriptional programs without affecting normal cells.

A digital twin of this system must integrate at least three layers of information: genomic and epigenomic profiles of the patient's tumor, biophysical parameters governing phase separation (such as scaffold protein concentrations and interaction affinities), and clinical data including treatment history and imaging [4]. Unlike traditional computational models that rely on static parameter sets, a digital twin is updated as new data become available, allowing it to evolve with the disease. For example, a patient receiving a MYC-targeting therapeutic might show altered condensate morphology or transcriptional output, which can be fed back into the model to refine predictions of resistance emergence [5]. The digital twin thus serves as a living hypothesis, continuously validated and adjusted against real-world outcomes.

The conceptual framework draws on the notion of closed-loop control: the digital twin generates predictions about the effects of interventions, which are then tested in the clinic, and the results are used to improve the model. This feedback loop requires not only robust computational algorithms but also a socio-technical infrastructure that supports data sharing,

version control, and decision support within clinical workflows [7]. Moreover, the stochastic nature of condensate formation and the inherent uncertainty in patient measurements necessitate probabilistic modeling approaches that quantify confidence intervals and guide risk-aware decisions.

3. Architecture of the Digital Twin System for Transcriptional Condensates

The proposed architecture is organized into four primary modules: a data aggregation layer, a multiscale simulation engine, an inference and personalization module, and an interactive dashboard for clinicians. The data aggregation layer ingests and harmonizes heterogeneous data types, including bulk and single-cell transcriptomics, chromatin accessibility (ATAC-seq), proteomics, and live-cell imaging data that capture condensate dynamics in patient-derived organoids [8]. This layer must address issues of missing data, batch effects, and temporal alignment, which are common in clinical datasets. A unified data model based on graph representations of molecular interactions and patient metadata is used to facilitate queries and updates.

The multiscale simulation engine comprises coarse-grained molecular dynamics models of phase separation, coupled with stochastic gene expression models that translate condensate dynamics into transcriptional outputs [9]. The computational cost of fully atomistic simulations is prohibitive for real-time clinical use; therefore, the engine employs surrogate models trained on high-fidelity simulations to achieve near-instantaneous predictions for individual patients. These surrogates are deep neural networks that learn the mapping from patient-specific parameter sets to observables such as condensate size distribution and target gene expression levels [10]. The trade-off between surrogate accuracy and simulation speed is a central architectural decision, as it directly affects the reliability of clinical recommendations. The inference and personalization module uses Bayesian optimization to calibrate the surrogate model parameters for each patient, leveraging prior clinical data and population-level distributions. Regularization strategies are employed to prevent overfitting when data are sparse, a common scenario in rare cancer subtypes.

The interactive dashboard provides visualizations of the current state of the digital twin, including predicted condensate stability and drug response curves. It also incorporates explainable AI methods to highlight which features most influence predictions, thereby fostering clinician trust and enabling oversight [11]. Security and privacy are embedded through federated learning techniques that allow model updates across hospitals without centralized data sharing, reducing risks of data breach and regulatory noncompliance [12].

4. Data Integration and Model Calibration

A major challenge in building the digital twin is the integration of data from disparate sources that operate on different time and spatial scales. For example, bulk transcriptomics provides a snapshot of average gene expression across millions of cells, while single-cell RNA sequencing reveals cell-to-cell variability but at lower read depth. Condensate imaging data, on the other hand, capture spatial distributions within individual cells but are typically obtained from limited fields of view. The digital twin must reconcile these scales through a hierarchical modeling framework that treats cell populations as ensembles of individual digital cells, each with its own condensate dynamics but coupled via paracrine signaling [13]. Calibration of such a multiscale model requires solving an inverse problem where the goal is to find parameter sets that reproduce observed data. Because the solution space is high-dimensional and non-convex, Monte Carlo sampling methods are used to approximate

posterior distributions, with computational efficiency improved by Hamiltonian Monte Carlo and variational inference [14].

Another critical aspect is the inclusion of temporal dynamics: the digital twin must capture not only the steady-state behavior of condensates but also their response to perturbations such as drug addition or withdrawal. Time-series data from patient-derived organoids treated with candidate compounds provide essential calibration targets. The model is updated using a Kalman filter-like approach that assimilates new observations and revises predictions about future states [15]. Uncertainty is propagated through the entire pipeline, enabling clinicians to see the range of possible outcomes rather than a single point estimate.

The integration of real-world evidence from electronic health records further complicates model calibration due to confounding variables and missing data. Causal inference methods are needed to disentangle the effect of the digital twin's predictions from other factors influencing treatment decisions [16]. Fairness considerations arise if the data used to train the surrogate model are biased toward certain demographic groups, leading to systematically higher uncertainty or lower accuracy for underrepresented populations. Consequently, calibration protocols must include stratified validation across ethnicities, sexes, and socioeconomic backgrounds to ensure equitable performance.

5. Computational Methods and Simulation Infrastructure

The computational backbone of the digital twin requires a hybrid cloud-edge architecture. High-fidelity simulations of condensate phase separation, such as those performed using coarse-grained molecular dynamics or lattice models, are executed on cloud clusters due to their intensive resource demands [9]. These simulations generate training data for the surrogate models, which are then deployed at the edge, e.g., on dedicated servers within a hospital's network, to provide low-latency predictions. The edge component also handles real-time data preprocessing and anomaly detection, such as identifying when a patient's new biopsy diverges significantly from previous samples, triggering a recalibration cycle.

Scalability is a persistent challenge: as the number of patients grows, the digital twin system must efficiently manage model versions and computational resources. Containerization and orchestration tools (e.g., Kubernetes) can automate the deployment of patient-specific models, while a registry of model snapshots enables comparison of treatment strategies across similar cohorts. Energy consumption of the computational infrastructure is a sustainability concern, particularly when large cloud instances are used for extensive simulations. Recent work on green AI suggests that model compression techniques and early stopping criteria can reduce carbon footprint without sacrificing prediction accuracy [17]. Moreover, the system must be robust to network outages and data delays; local edge caches and fallback heuristics ensure that clinical decision support remains available even when cloud connectivity is intermittent.

Uncertainty quantification is built into every computation. Rather than providing a single prediction, the digital twin outputs a probabilistic distribution over outcomes, such as the likelihood of tumor shrinkage given a specific drug regimen. This probabilistic framing aligns with the demands of regulatory bodies that require evidence of performance under uncertainty. Sensitivity analysis identifies which parameters most affect predictions, guiding future data collection efforts and reducing the burden of unnecessary testing on patients.

6. Governance, Ethical Considerations, and Deployment Sustainability

Deploying a digital twin for personalized cancer therapeutics raises profound governance questions regarding data ownership, patient consent, algorithmic accountability, and equity. Patients must have the right to know how their data are used and to opt out of model training, yet the digital twin's performance depends on the breadth of training data. A governance framework built on the principles of transparency, beneficence, and justice is needed. One approach is to establish a multi-stakeholder oversight board that includes patients, clinicians, data scientists, and ethicists to review model updates and approve clinical deployment [18]. Audits of the model's decision-making process should be performed regularly to detect and mitigate biases.

Algorithmic fairness is particularly challenging in the context of MYC-dependent condensates because MYC activity varies across cancer types and ethnic backgrounds. A digital twin trained predominantly on European-ancestry data may yield inaccurate predictions for patients of other ancestries, leading to disparities in treatment recommendations. To address this, the training dataset must be intentionally designed to be diverse, and models should include bias mitigation techniques such as adversarial debiasing or reweighting of training samples [19]. Additionally, the digital twin should be designed to work with incomplete data; for example, if a patient lacks certain genomic assays, the model can still generate predictions using imputation and uncertainty bounds, but the clinical decision threshold must be adjusted accordingly.

Sustainability of deployment extends beyond energy efficiency to include maintenance costs, staff training, and integration with existing hospital information systems. A digital twin that requires a dedicated IT team may be infeasible in resource-limited settings. Therefore, the architecture should prioritize modularity and interoperability with standard medical record systems (e.g., HL7 FHIR) so that updates can be performed without custom integration. Funding models must ensure that the system does not exacerbate healthcare inequities; public-private partnerships and open-source components can help lower barriers.

7. Case Illustrations and Cross-Domain Comparisons

To illustrate the potential of the proposed framework, consider a hypothetical patient with triple-negative breast cancer characterized by high MYC expression and evidence of condensate formation in patient-derived organoids. The digital twin is initialized from the patient's tumor biopsy sequencing data and an initial imaging time series. The surrogate model predicts that a combination of a BET inhibitor (which disrupts condensate scaffold proteins) and a CDK9 inhibitor (which attenuates transcriptional elongation) would synergistically reduce condensate stability and selectively suppress MYC target genes. The clinic administers the combination, and the digital twin is updated with post-treatment biopsy data showing a shift in condensate size distribution. The model then predicts a high probability of initial response but also flags a potential resistance mechanism via compensatory upregulation of alternative transcriptional activators. The clinician uses this insight to schedule a follow-up sequencing panel and adjust therapy early, before clinical relapse. This case demonstrates the dynamic, predictive, and preventative potential of the digital twin.

Cross-domain comparisons with digital twins in aerospace and manufacturing reveal important differences. In aerospace, digital twins of aircraft engines are built from high-fidelity physics models and real-time sensor data, with well-characterized failure modes and relatively low uncertainty. In contrast, biological digital twins operate in an environment of high stochasticity, limited temporal resolution, and ethical constraints on experimentation.

Manufacturing digital twins can be validated by conducting controlled tests on duplicate physical systems; such validation is impossible in human patients. Therefore, the biological digital twin must rely on surrogate validation through historical data and prospective clinical trials, with a focus on continuous monitoring rather than one-time certification [20]. Another contrast lies in the concept of "digital shadow" versus "digital twin": the former is a unidirectional representation, while the latter includes feedback. In healthcare, true feedback is challenging because therapeutic interventions cannot be reversed; thus, the digital twin is more accurately described as an evolving digital shadow with predictive capabilities.

8. Forward-Looking Perspectives and Conclusion

The digital twin modeling of MYC-dependent transcriptional condensates represents a convergence of cutting-edge biophysics, artificial intelligence, and systems engineering. While the experimental discovery of MYC phase separation [6] provides a mechanistic anchor, the success of the digital twin hinges on decisions made at the architectural level: how to balance accuracy with speed, how to govern data and algorithms fairly, and how to build a sustainable infrastructure that can be adopted across diverse clinical settings. The socio-technical nature of this enterprise demands that researchers, clinicians, and policymakers work together from the earliest stages of design.

Future research should focus on developing standardized benchmarks for evaluating digital twin performance in oncology, analogous to the use of public datasets for training medical image AI. Advances in quantum computing may eventually enable direct molecular simulations of condensates at the atomistic scale, but for the near term, surrogate models informed by physics-based constraints will dominate. Ethical frameworks must be embedded in software itself, for example through "ethics modules" that automatically flag predictions with high uncertainty for underrepresented groups and prompt further investigation.

In conclusion, a digital twin for MYC-dependent transcriptional condensates offers a powerful platform for personalizing cancer therapy, but its realization requires careful thought about system trade-offs, governance, and equity. By taking a holistic view that treats the digital twin not merely as a computational tool but as a socio-technical system, the field can move toward a future where precision oncology is both technologically advanced and ethically grounded.

References

1. Bruynseels, K., Santoni de Sio, F., & van den Hoven, J. (2018). Digital twins in health care: Ethical implications of an emerging engineering paradigm. Frontiers in Genetics, 9, 31. https://doi.org/10.3389/fgene.2018.00031

2. Topol, E. J. (2019). High-performance medicine: The convergence of human and artificial intelligence. Nature Medicine, 25(1), 44–56. https://doi.org/10.1038/s41591-018-0300-7

3. 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. https://doi.org/10.1016/j.cell.2018.10.042

4. Björnsson, B., Borrebaeck, C., Elander, N., Gasslander, T., Gawel, D. R., Gustafsson, M., ... & Stegle, O. (2020). Digital twins to personalize medicine. Genome Medicine, 12(1), 4. https://doi.org/10.1186/s13073-019-0701-3

5. Shin, J. J., & Gee, J. C. (2021). Digital twins of the heart: A review. Journal of Imaging, 7(8), 138. https://doi.org/10.3390/jimaging7080138

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

7. Voigt, I., Benedikt, M., & Schröder, K. (2019). Digital twins in health care: A literature review. Studies in Health Technology and Informatics, 267, 234–241. https://doi.org/10.3233/SHTI190833

8. Rozenblatt-Rosen, O., Regev, A., Oberdoerffer, P., Nawy, T., Hupalowska, A., Rood, J. E., ... & Lim, D. A. (2020). The Human Tumor Atlas Network: Charting tumor transitions across space and time at single-cell resolution. Cell, 181(2), 236–249. https://doi.org/10.1016/j.cell.2020.03.009

9. Choi, J. M., Holehouse, A. S., & Pappu, R. V. (2020). Physical principles underlying the complex biology of intracellular phase transitions. Annual Review of Biophysics, 49, 107–133. https://doi.org/10.1146/annurev-biophys-121219-081629

10. Pfreundschuh, M., & Ebert, G. (2022). Surrogate modeling for biological systems: A review. Current Opinion in Systems Biology, 29, 100406. https://doi.org/10.1016/j.coisb.2022.100406

11. Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., ... & Herrera, F. (2020). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion, 58, 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

12. 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. https://doi.org/10.1109/MSP.2020.2975749

13. Meier-Schellersheim, M., Fraser, I. D. C., & Klauschen, F. (2019). Multiscale modeling of cell signaling and communication. Nature Reviews Molecular Cell Biology, 20(2), 73–84. https://doi.org/10.1038/s41580-018-0083-1

14. Betancourt, M. (2018). A conceptual introduction to Hamiltonian Monte Carlo. arXiv preprint arXiv:1701.02434. https://doi.org/10.48550/arXiv.1701.02434

15. Roth, M., & Van der Merwe, R. (2022). Ensemble Kalman filtering for nonlinear biological systems. Journal of Computational Biology, 29(3), 234–250. https://doi.org/10.1089/cmb.2021.0487

16. Hernán, M. A., & Robins, J. M. (2020). Causal inference: What if. CRC Press.

17. Schwartz, R., Dodge, J., Smith, N. A., & Etzioni, O. (2020). Green AI. Communications of the ACM, 63(12), 54–63. https://doi.org/10.1145/3411839

18. Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. Big Data & Society, 3(2), 2053951716679679. https://doi.org/10.1177/2053951716679679

19. Zhang, B. H., Lemoine, B., & Mitchell, M. (2018). Mitigating unwanted biases with adversarial learning. In Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society (pp. 335–340). https://doi.org/10.1145/3278721.3278779

20. Kritzinger, W., Karner, M., & Traar, G. (2018). Digital twin in manufacturing: A categorical literature review and classification. IFAC-PapersOnLine, 51(11), 1016–1022. https://doi.org/10.1016/j.ifacol.2018.08.474