Transformer-Based Prediction of Context-Dependent Transcriptional Regulation in Cancer Biology

Chetan Subramanian
Department of Computer Science, Colorado State University, Fort Collins, CO, USA.
chetanmail@colostate.edu

Nathan Douglas
Department of Computer Science, University of North Texas, Denton, TX, USA.
hellonathan@unt.edu

Abstract

Transcriptional regulation is a highly context-dependent process that governs gene expression programs in health and disease. In cancer, aberrant regulation arises from mutations in transcription factors, epigenetic alterations, and changes in chromatin accessibility that vary across cell types, developmental stages, and microenvironments. Accurate prediction of transcription factor binding and downstream gene expression in such heterogeneous contexts remains a fundamental challenge. Transformer-based deep learning architectures, originally developed for natural language processing, have recently been adapted to model long-range dependencies in genomic sequences and epigenomic signals. This paper presents a comprehensive system-level analysis of transformer-based approaches for predicting context-dependent transcriptional regulation in cancer biology. We examine the architectural trade-offs between self-attention mechanisms, positional encodings, and multi-scale feature integration. We discuss the infrastructure requirements for training models on large-scale cancer genomics datasets, including data provenance, quality control, and computational scalability. The paper explores robustness and fairness considerations, particularly regarding representation of underrepresented populations and cancer subtypes. Deployment challenges in clinical and research settings are analyzed, with emphasis on interpretability, uncertainty quantification, and integration with existing bioinformatics pipelines. We also address governance and policy implications, including data sharing standards, model validation frameworks, and sustainability of computational resources. Through cross-domain comparisons with applications in regulatory genomics, drug response prediction, and single-cell analysis, we highlight the potential of transformer models to unify disparate data modalities and enable precision oncology. The analysis underscores the need for careful architectural design, rigorous benchmarking, and ethical deployment to ensure that these models translate into equitable and robust clinical tools.

Keywords

transformer models, transcriptional regulation, cancer biology, context dependence, deep learning, regulatory genomics, precision oncology, data infrastructure, model robustness, governance.

1. Introduction

Transcriptional regulation is orchestrated by the dynamic interplay of transcription factors, cofactors, chromatin remodelers, and non-coding RNAs, all operating within a three-dimensional nuclear environment. In cancer, this regulatory circuitry is frequently rewired

through mutations in transcription factor genes, copy number alterations, and epigenetic reprogramming, leading to aberrant gene expression that drives tumorigenesis and metastasis. The context-dependent nature of these regulatory interactions means that a given transcription factor may bind different genomic loci in different cell types or under different environmental conditions. Predicting such context-specific binding and its functional consequences is essential for understanding cancer mechanisms and developing targeted therapies.

Deep learning has emerged as a powerful tool for modeling regulatory genomics, with convolutional neural networks and recurrent architectures achieving notable success in tasks such as predicting transcription factor binding, chromatin accessibility, and enhancer-promoter interactions. However, these models often struggle to capture long-range dependencies that span hundreds of kilobases and to integrate multiple sources of contextual information, including histone modifications, DNA methylation, and three-dimensional chromatin conformation. Transformer architectures, which rely on self-attention mechanisms to process sequences in a permutation-invariant manner, offer a natural solution to these challenges. By treating a genomic region as a sequence of tokens and learning attention weights across all positions, transformers can model interactions between distal regulatory elements and condition their predictions on the surrounding genomic and epigenomic context.

This paper provides a system-level perspective on the use of transformer-based models for predicting context-dependent transcriptional regulation in cancer. We do not focus on specific model architectures or benchmark results but instead examine the broader implications of deploying these models within the socio-technical infrastructure of cancer genomics research. We discuss architectural design choices, data and compute requirements, robustness and fairness considerations, and the policy and governance frameworks needed for responsible translation. The goal is to offer a holistic analysis that informs both researchers and practitioners about the opportunities and challenges inherent in this rapidly evolving field.

2. Background and Related Work

The study of transcriptional regulation has been transformed by high-throughput sequencing technologies that profile transcription factor binding, chromatin accessibility, and gene expression across diverse cell states. Consortia such as the ENCODE Project and the Roadmap Epigenomics Mapping Consortium have generated extensive catalogues of regulatory elements and their activities in hundreds of cell types [1,2]. In cancer, The Cancer Genome Atlas and International Cancer Genome Consortium have provided multi-omic data across thousands of tumors [3,4]. Despite this wealth of data, computational models that predict regulatory outcomes from genomic sequence and epigenomic signals have historically been limited by their inability to handle long-range interactions and cell-type-specific contexts.

Early machine learning approaches for predicting transcription factor binding relied on position weight matrices and k-mer-based feature representations. These were supplanted by deep learning models, such as DeepBind and DeepSEA, which used convolutional neural networks to learn sequence motifs and their spatial arrangements [5,6]. More recent architectures, including Basenji and Enformer, incorporated dilated convolutions and attention layers to process sequences up to 200 kilobases in length, achieving state-of-the-art performance in predicting chromatin states and gene expression [7,8]. However, these models still face limitations in generalizing across cell types and capturing subtle context-dependent effects.

Transformer networks, introduced for machine translation, have been adapted to biological sequences by treating DNA or RNA as a one-dimensional language [9]. The self-attention mechanism enables each position to attend to every other position, providing a flexible way to model dependencies of arbitrary length. In regulatory genomics, models such as DNABERT, LOGO, and various adaptations of the transformer architecture have demonstrated improvements in predicting transcription factor binding and regulatory variants [10,11,12]. These models can incorporate position-specific embeddings, strand information, and epigenetic tracks as additional input channels. Furthermore, the ability of transformers to handle multiple modalities simultaneously makes them attractive for integrating chromatin accessibility, histone modification, and gene expression data within a single framework.

In cancer biology, context-dependent regulation is particularly pronounced due to tumor heterogeneity, clonal evolution, and microenvironmental influences. For example, the transcription factor MYC, a master regulator of cell growth, exhibits context-specific binding patterns that are modulated by phase separation and cofactor availability [17]. Predicting such behavior requires models that can capture not only primary sequence features but also higher-order chromatin organization and post-translational modifications. Transformer-based approaches are well-suited for this task because they can jointly model sequence context, epigenetic signals, and spatial chromatin interactions through attention mechanisms that learn which features are predictive in a given context.

3. Transformer Architecture for Regulatory Prediction

Designing a transformer architecture for context-dependent transcriptional regulation involves several architectural choices that trade off expressivity, computational cost, and interpretability. The core component is the self-attention layer, which computes a weighted sum of value vectors based on pairwise similarity of query and key vectors derived from each position in the input sequence. In genomic applications, the input sequence typically consists of DNA bases represented as one-hot encoded vectors, possibly concatenated with continuous epigenomic tracks. The attention mechanism allows the model to learn which genomic positions are most relevant for predicting the regulatory outcome at a particular locus, effectively performing a soft alignment between sequence features and regulatory functions.

One critical trade-off is between the number of attention heads and the dimensionality of each head. More heads allow the model to attend to different types of relationships, such as sequence motifs, distal enhancer contacts, and epigenetic marks, but increase the number of parameters and memory footprint. The choice of positional encoding is another important design decision. Transformers are permutation-invariant without explicit position information; thus, positional encodings must be added to the input tokens. Sinusoidal encodings, learned absolute positions, and relative position biases have each been explored [9,13]. Relative position encodings have proven beneficial in biological sequences because they capture distance-dependent interactions, such as the tendency for enhancers to act within a certain distance window.

Multi-scale feature integration is essential for modeling the hierarchical nature of genome regulation. A single transformer layer processes the entire sequence at the same resolution, but regulatory elements operate at scales from individual base pairs to megabase-scale topologically associating domains. Some architectures employ hierarchical transformers that downsample the sequence through pooling or stride-based attention, allowing the model to capture local motifs at higher resolution and long-range interactions at lower resolution. Alternatively, cross-attention between multiple resolutions can be used to combine

information from different scales. The choice of training objective also influences architectural design. Sequence-level objectives, such as masked language modeling or next-token prediction, can be used for self-supervised pretraining on large genomic corpora, followed by fine-tuning on specific regulatory prediction tasks [11]. This approach leverages the vast amount of unlabeled genomic data to learn general regulatory grammar before adapting to cancer-specific contexts.

Another architectural consideration is the integration of multiple data modalities. In cancer biology, regulatory predictions often require inputs from RNA sequencing, chromatin immunoprecipitation sequencing, assay for transposase-accessible chromatin sequencing, and genome-wide association studies. Transformers can be designed to accept multiple input streams, either by concatenating them along the channel dimension or by using separate encoders with cross-attention layers. The latter approach, known as a multimodal transformer, allows each modality to be processed with its own attention patterns before fusion. This is particularly useful when the modalities have different resolutions or noise characteristics. For example, gene expression data may be available at the gene level, while chromatin accessibility is measured at nucleotide resolution. A multimodal architecture can align these representations through learnable projections.

The computational cost of training transformer models on genomic sequences is substantial, especially when dealing with long contexts (e.g., 100,000 base pairs). The self-attention mechanism has quadratic complexity with respect to sequence length, which limits the feasible input size. Sparse attention patterns, such as local windows combined with global tokens, or linear attention approximations, have been developed to mitigate this issue [14]. These approximations trade off some modeling capacity for scalability, and careful evaluation is needed to ensure that they do not degrade prediction performance for long-range regulatory interactions. Additionally, memory-efficient implementations using gradient checkpointing and mixed-precision training are often necessary to fit models into GPU memory.

4. Context-Dependent Mechanisms and Cancer Biology

Cancer is fundamentally a disease of context. The same oncogenic mutation can produce dramatically different effects depending on the cell type, developmental stage, and microenvironment. For example, mutations in the tumor suppressor p53 are common in many cancers, but the downstream transcriptional programs they disrupt vary widely across tissues. Similarly, transcription factors such as MYC, which is frequently amplified in cancer, bind to thousands of genomic sites, but their functional impact depends on the availability of cofactors and the local chromatin state [17]. Indeed, recent work has shown that MYC undergoes phase separation at super-enhancers, selectively modulating the transcription of genes involved in ribogenesis and metabolism [17]. This context-dependent behavior cannot be captured by models that treat transcription factor binding as a simple function of sequence motifs.

Transformer models offer a way to incorporate contextual information by conditioning predictions on the surrounding genomic and epigenomic environment. The attention mechanism can learn to weight inputs from different regions based on the cell type-specific signals present in the data. For instance, when predicting the effect of a mutation in a regulatory region, a transformer can attend to nearby histone modification marks that indicate active enhancers, as well as to the expression levels of transcription factors in the given cell type. This conditional approach moves beyond static motif analysis to capture the dynamic regulatory logic that underlies cancer plasticity.

The ability to model long-range dependencies is particularly relevant for cancer because many regulatory variants associated with cancer risk lie in non-coding regions that are far from the genes they regulate. Genome-wide association studies have identified thousands of cancer risk loci, most of which are located in intronic or intergenic regions. Understanding how these variants disrupt regulatory circuits requires linking them to target genes through chromatin loops and enhancer-promoter interactions. Transformer models that incorporate three-dimensional chromatin contact maps, either as input features or through attention biases, can learn to propagate regulatory influence across large genomic distances. This capability is critical for interpreting non-coding mutations and identifying potential therapeutic targets.

Tumor heterogeneity presents another layer of context dependence. Within a single tumor, different subclones may exhibit distinct regulatory programs, driven by genetic, epigenetic, and microenvironmental variations. Single-cell sequencing technologies now allow profiling of transcriptomes and epigenomes at unprecedented resolution, revealing cell-state-specific regulatory patterns. Transformer architectures can be extended to handle single-cell data by treating each cell as a separate sequence or by using attention over a cell embedding to condition predictions on cell state. This enables the model to learn regulatory rules that are shared across cells while also capturing cell-type-specific deviations.

The deployment of transformer models for context-dependent regulation in cancer also requires careful consideration of the training data composition. If the model is trained primarily on common cancer cell lines or bulk tumor samples, its predictions may not generalize to rare cancer types, pediatric cancers, or tumors from patients of diverse ancestries. Data imbalance can lead to biased predictions that reinforce existing disparities in cancer research and treatment. Therefore, developing context-dependent models must be accompanied by strategies for data augmentation, transfer learning, and uncertainty estimation to ensure that predictions are robust across the full spectrum of cancer contexts.

5. Data Infrastructure and Training Considerations

Building transformer models for regulatory prediction in cancer requires a robust data infrastructure that spans data generation, storage, preprocessing, and versioning. The scale of modern genomics data is enormous; a single whole-genome sequencing experiment can produce hundreds of gigabytes of raw data, and a typical cancer genomics project may involve thousands of samples. Aggregating these data into a unified training set requires standardized file formats, such as bigWig for continuous tracks and BED for genomic intervals, as well as consistent quality control pipelines to remove artifacts and batch effects.

Data provenance is critical for reproducibility and trust. Each training example should be traceable to its original experiment, sample metadata, and processing steps. In the cancer domain, patient privacy regulations, such as HIPAA and GDPR, impose additional constraints on data sharing and storage. Models trained on controlled-access data may not be publicly distributable, complicating validation and comparison. Federated learning frameworks offer a potential solution, allowing models to be trained across multiple institutions without centralizing sensitive patient data. However, the communication overhead and heterogeneity of local data distributions pose technical challenges for transformer-based models with large parameter counts.

Preprocessing genomic sequences for transformer input involves tokenization. While DNA can be tokenized at the nucleotide level (A, C, G, T) or as k-mers, the choice affects model capacity and computational cost. Nucleotide-level tokenization yields a vocabulary of size 4

(or 5 including unknown), but sequences must be long to capture regulatory information. K-mer tokenization reduces sequence length but increases vocabulary size, potentially requiring larger embedding matrices. Some models use byte-pair encoding or sentencepiece tokenization learned from genomic sequences to create an optimal vocabulary [15]. These decisions must be grounded in the downstream task; for transcription factor binding, motif-level tokens may be more informative, while for gene expression, gene-level tokens are natural.

Data augmentation can improve model robustness to context shifts. For example, introducing simulated mutations, reversing complement sequences, or swapping epigenetic profiles between cell types can help the model learn invariant features. However, augmentation must be carefully designed to avoid introducing biologically implausible patterns. In the cancer context, augmentations that mimic tumor-specific mutations or copy number alterations may be beneficial for improving generalization to unseen genomic rearrangements.

Training large transformer models demands substantial computational resources. The state-of-the-art models in regulatory genomics, such as Enformer, require hundreds of GPU-days to train. This raises sustainability concerns, as the energy consumption of deep learning has significant environmental impact. Efficient training strategies, including mixed-precision arithmetic, gradient accumulation, and distributed training across multiple nodes, are essential. Moreover, the carbon footprint of training can be mitigated by using cloud-based instances powered by renewable energy or by sharing pre-trained models to reduce redundant training. The development of open-access model repositories and standardized benchmarks would lower the barrier for entry for researchers in resource-constrained settings.

6. Model Robustness, Fairness, and Deployment Challenges

Deploying transformer models for clinical or translational applications in cancer requires rigorous assessment of robustness and fairness. Robustness refers to the model's ability to maintain predictive accuracy under distributional shifts, such as differences in sequencing platform, library preparation protocol, patient ancestry, or tumor evolution. A model trained on datasets from a single sequencing center may fail when applied to data from another center due to technical artifacts. Adversarial perturbations, such as small sequence changes that do not affect biological function, can also fool models into making incorrect predictions. Stress-testing with synthetic and real-world perturbations is necessary to quantify robustness.

Fairness in machine learning for cancer genomics is a multidimensional issue. Models may exhibit disparate performance across racial and ethnic groups if the training data are predominantly from individuals of European ancestry. For example, transcription factor binding models may have higher error rates for variants that are more common in African populations because those variants are underrepresented in the training set. Similarly, cancer subtypes that are rare or have different molecular profiles (e.g., triple-negative breast cancer vs. luminal A) may be poorly predicted. To mitigate these biases, training data should be strategically sampled to ensure diversity, and evaluation metrics should be disaggregated by relevant subgroups. Additionally, interpretability methods can help identify whether the model relies on proxies for ancestry or other sensitive attributes, enabling corrective interventions.

Interpretability is a crucial requirement for clinical adoption. Regulatory predictions need to be explainable to clinicians and researchers to build trust and facilitate mechanistic insights. Attention weights can provide a window into which genomic regions the model considers

important, but they do not constitute a causal explanation. Methods such as integrated gradients, SHAP, and counterfactual analysis can attribute predictions to specific input features, but they are computationally expensive for long sequences [16]. Developing efficient and faithful interpretability tools specific to transformer architectures in genomics remains an active area of research.

Deployment pipelines must integrate transformer models with existing bioinformatics workflows. Many clinical laboratories already use pipelines for variant calling, annotation, and reporting. Adding a deep learning component requires careful versioning of model weights, environment dependencies, and input preprocessing. Real-time inference on whole genomes is challenging; a single forward pass of a transformer over a 100,000-base-pair window takes seconds, but scanning the entire genome entails tens of thousands of windows, leading to hours of computation. Optimization techniques such as sliding window caching, early stopping, and model distillation can reduce inference time. Moreover, prediction uncertainty should be quantified, for instance through Monte Carlo dropout or ensemble methods, to flag low-confidence predictions that may require further validation.

7. Governance, Policy, and Sustainability Implications

The adoption of transformer-based models in cancer research and clinical practice raises significant governance and policy questions. Data governance frameworks must balance the benefits of large-scale data aggregation for model training with the privacy rights of patients. Consent models that allow secondary use of genomic data, while respecting individual autonomy, are needed. The use of synthetic data or differentially private training could provide additional privacy guarantees, but they may degrade model performance and require careful validation.

Model governance involves establishing standards for validation, benchmarking, and certification before deployment. Regulatory agencies such as the FDA have begun to issue guidelines for artificial intelligence as a medical device, but specific guidance for deep learning models that predict transcriptional regulation is still nascent. A consensus on acceptable performance metrics, such as area under the receiver operating characteristic curve, precision-recall curves, and calibration error, needs to be developed. Furthermore, models should be subject to continuous monitoring for drift as new data become available.

Sustainability of computational infrastructure is a growing concern. Training and hosting large transformer models consume substantial energy, contributing to carbon emissions. Institutions and funding agencies should encourage the use of efficient model architectures, hardware accelerators designed for low power, and shared computing resources. Open-source models and pre-trained weights can reduce redundant computation across groups. Additionally, the choice of model should consider the trade-off between performance and energy cost; for many clinical use cases, a smaller model that is easier to deploy may be preferable to a marginally more accurate but resource-intensive one.

Intellectual property and access also intersect with governance. Patents on transformer architectures or training methods could restrict the use of models in resource-limited settings. Promoting open science through permissive licenses and community-driven initiatives, such as the Genomics and Health Data Commons, can ensure equitable access. International collaborations, particularly with researchers in low- and middle-income countries, are essential to ensure that models reflect global genomic diversity and address cancer burdens worldwide.

8. Conclusion

Transformer-based architectures represent a significant advance in the prediction of context-dependent transcriptional regulation, offering the ability to model long-range genomic interactions, integrate multiple data modalities, and condition predictions on cell state and microenvironment. In cancer biology, these capabilities are particularly valuable for deciphering the complex regulatory rewiring that underlies tumor heterogeneity, metastasis, and drug resistance. However, the successful deployment of these models requires careful attention to architectural design, data infrastructure, robustness, fairness, and governance. The trade-offs between model complexity and computational cost, between generalization and specialization, and between interpretability and accuracy must be navigated with a systems perspective. As the field moves toward clinical translation, interdisciplinary collaboration among computational scientists, biologists, clinicians, ethicists, and policymakers will be essential to ensure that these tools are robust, equitable, and sustainable. Future research should focus on developing sparse and efficient transformers, improving uncertainty quantification, and creating benchmarks that reflect real-world clinical scenarios. By addressing these challenges, transformer-based models can become a cornerstone of precision oncology, enabling personalized predictions of regulatory dysfunction and guiding therapeutic intervention.

References

1. ENCODE Project Consortium. (2012). An integrated encyclopedia of DNA elements in the human genome. Nature, 489(7414), 57-74.

2. Roadmap Epigenomics Mapping Consortium. (2015). Integrative analysis of 111 reference human epigenomes. Nature, 518(7539), 317-330.

3. The Cancer Genome Atlas Research Network. (2013). The cancer genome atlas pan-cancer analysis project. Nature Genetics, 45(10), 1113-1120.

4. Hudson, T. J., Anderson, W., Aretz, A., Barker, A. D., Bell, C., Bernabé, R. R., ... & International Cancer Genome Consortium. (2010). International network of cancer genome projects. Nature, 464(7291), 993-998.

5. Alipanahi, B., Delong, A., Weirauch, M. T., & Frey, B. J. (2015). Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. Nature Biotechnology, 33(8), 831-838.

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

7. 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.

8. Avsec, Ž., Agarwal, V., Visentin, D., Ledsam, J. R., Grabska-Barwinska, A., Taylor, K. R., ... & Gagneur, J. (2021). Effective gene expression prediction from sequence by integrating long-range interactions. Nature Methods, 18(10), 1196-1203.

9. 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.

10. Ji, Y., Zhou, Z., Liu, H., & Davuluri, R. V. (2021). DNABERT: pre-trained bidirectional encoder representations from transformers for DNA language in genome. Bioinformatics, 37(15), 2112-2120.

11. Luo, Y., Tang, J., & Kellis, M. (2022). Genomic language model using transformers and contrastive learning for regulatory element prediction. Nature Machine Intelligence, 4(11), 1030-1042.

12. Sanabria, M., Hirsch, J. D., & Hoogendoorn, M. (2023). LOGO: a transformer-based architecture for learning regulatory genomic embeddings. PLOS Computational Biology, 19(2), e1010906.

13. Shaw, P., Uszkoreit, J., & Vaswani, A. (2018). Self-attention with relative position representations. Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics, 464-468.

14. Beltagy, I., Peters, M. E., & Cohan, A. (2020). Longformer: the long-document transformer. arXiv preprint arXiv:2004.05150.

15. Kudo, T., & Richardson, J. (2018). SentencePiece: a simple and language independent subword tokenizer and detokenizer for neural text processing. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 66-71.

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

17. 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.

18. Avsec, Ž., Weilert, M., Shrikumar, A., Krueger, S., Alexandari, A., Dalal, K., ... & Kundaje, A. (2021). Base-resolution models of transcription-factor binding reveal soft motif grammar. Nature Genetics, 53(3), 354-366.

19. Chen, S., Zhou, Y., Chen, Y., & Gu, J. (2018). fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics, 34(17), i884-i890.

20. Subramanian, I., Verma, S., Kumar, S., & Jere, A. (2020). Multi-omics data integration, interpretation, and its application. BioData Mining, 13(1), 12.

21. Kumar, S., & Buckner, J. (2022). Federated learning for genomics: principles, challenges, and opportunities. Annual Review of Biomedical Data Science, 5, 37-61.