GMR The Original

Large Scale Artificial Intelligence Models Unifying Epigenomic and Transcriptomic Signals to Interpret Human Disease Genetics

Aziza Nazarova, Lola Bekmirzaeva, Merojiddin Jiyanberdiyev, Amil Babayev, Khalmurad Akhmedov, Adiba Botirova, Nazira Kurbanova,

Department of Endocrinology, Bukhara State Medical Institute. Bukhara, Uzbekistan, ORCID: https://orcid.org/0000-0002-5805-5560 E-mail: aziza_nazarova@bsmi.uz
Samarkand State Pedagogical Institute, Samarkand, Uzbekistan ORCID: https://orcid.org/0009-0001-5751-856X, lola_bekmirzaeva@mail.ru
Department of Psychology, Jizzakh State Pedagogical University. Jizzakh, Uzbekistan, ORCID: https://orcid.org/0009-0001-2458-6005 E-mail: merojiddinjiyanberdiyev977@gmail.com
Department of Cybersecurity and Computer Engineering, Baku Engineering University. Baku, Azerbaijan, ORCID: https://orcid.org/0009-0005-4823-6201 E-mail: amilb@beu.edu.az
DSc, Professor, Head of the department of internal medicine in family medicine, Tashkent State Medical University, Tashkent, Uzbekistan, 100109, https://orcid.org/0000-0003-2737-3803, khalmurad1968@mail.ru
Department of General and Comparative Linguistics, Andijan State Institute of Foreign Languages. Andijan, Uzbekistan , ORCID: https://orcid.org/0000-0003-3728-365X E-mail: botirovaadiba999@gmail.com
Lecturer, Department of Economics and Services, Termez University of Economics and Service. Termez, Uzbekistan, ORCID: https://orcid.org/0009-0009-3819-9231

ABSTRACT

Recent advances in high-throughput sequencing and large-scale international consortia have generated extensive epigenomic and transcriptomic datasets across diverse human tissues and disease contexts. These resources offer unprecedented opportunities to interpret the functional consequences of genetic variation, particularly for non-coding variants that account for the majority of heritability in common human diseases. However, traditional analytical approaches often model epigenomic and transcriptomic signals independently, limiting their ability to capture the complex, context-dependent regulatory mechanisms that connect genotype to phenotype. This work highlights the rationale and emerging methodologies for integrative modeling of epigenomic landscapes and transcriptomic profiles using large-scale artificial intelligence (AI) frameworks. By jointly learning from complementary molecular modalities, such models can uncover shared and modality-specific regulatory representations, align chromatin states with gene expression outputs, and better characterize disease-relevant regulatory mechanisms. We discuss data integration strategies, representation learning techniques, and multimodal alignment approaches that enable unified interpretation of epigenomic and transcriptomic signals at scale. Integrative AI-driven frameworks hold substantial promise for improving variant prioritization, elucidating gene regulatory mechanisms, and advancing precision medicine through more accurate interpretation of human disease genetics.

Keywords: *Epigenomics; Transcriptomics; Multi-omics integration; Human disease genetics; non-coding variants; Gene regulation; Representation learning; Multimodal AI models; Precision medicine*

INTRODUCTION

Recent advances in genomic sequencing technologies provide unprecedented opportunities for integrating and interpreting biological information from the human genome. Large-scale international consortia have generated thousands of epigenomic maps and gene transcript quantifications across diverse human samples, creating a powerful foundational architecture for understanding the genomic instructions that implement the human blueprint and, consequently, for the integrative modelling of genotype-to-phenotype relationships. A viable strategy for gaining insight into the possible functional roles of non-coding genetic variants linked to common diseases and quantitative traits, in the general framework of precision medicine, is to integrate epigenomic landscapes associated with cellular context and transcriptomic expressions correlated with the response of a cell type to treatment (E Hoffman et al., 2019); (Sharmin, 2017).

Background and Rationale

Non-coding genetic variants account for a majority of heritability for many common human diseases, yet they remain the least understood (E Hoffman et al., 2019). Furthermore, while epigenomic landscapes measured at multiple levels influence gene regulation and disease-associated processes, disease-specific transcriptomic profiles provide direct information about genomic states and phenotypic outcomes, thereby linking genetic variants to diseases. Existing approaches typically model multi-omics data in isolation, relying on either single-omics data at the expense of other critical information or only small, ad hoc datasets unsuitable for learning generalizable knowledge. Integrative modelling at scale-notably, concurrently modelling high-dimensional epigenomic and transcriptomic signals in a unified and biologically grounded manner-therefore holds strong potential for enhancing functional interpretation of the human disease genome.

Genetic variants exert their effects through biological mechanisms that often operate over time, requiring integrative approaches to question mechanistic understanding. Integrative modelling of large-scale epigenomic and transcriptomic signals in disease contexts also unveils additional regulatory information concerning cell state, enabling the dissection of complex regulatory influences from genetic variations. Such multi-task characterisation of regulatory effects offers valuable insights into the time-varying nature of genetic regulation (Sharmin, 2017) and motivates the selection of epigenomic and transcriptomic signals as complementary yet fundamentally distinct modalities for learning representation.

Epigenomic Landscapes in Disease Genetics

Disrupting the equilibrium of chromatin and gene expression may lead to complex human diseases and other health problems. In order to improve understanding of sensitivity or resistance to specific diseases, the identification of the precise link between non-repetitive and seemingly neutral genomic variations and chromatin signals undertaking extensive survey is essential. Different kinds of variations, such as single nucleotide polymorphisms (SNPs), copy number variations (CNVs) and transposable element (TE) insertions, represent diverse epigenomic impacts. These de-enhancers and de-silencers can be also predicted transcriptional regulatory variants when fully integrating machine-learning framework of high-dimensional epigenomic data to pinpoint potential gastronomic sub-regions and non-coding genomic trans-regulation of human beings (E Hoffman et al., 2019). Recent advances have enhanced knowledge of heritable chromatin modifications dependent on immune-related and neuro-anatomical features tolerating HuRef experiment while establishing clear association between chromatin states and the genetic control of gene expression and pre-mRNA splicing (Chen et al., 2016). During large-scale epigenome mapping, abundant information about diverse cellular functions contained in already-elaborated genome annotation favoured elucidation of molecular mechanisms underlying the correlation inherited genetic variations, tissue-specific chromatin rule and cell-of-origin influencing genetic, epigenetic and transcriptomic variation quantification across extensive cell and tissue resources (Pei et al., 2020). Mapping epigenomic

events together with transcriptomic signals provides powerful montage of regulatory information still largely lacking [table 1].

Table 1: Epigenomic and Transcriptomic Features in Disease Genetics

Feature	Description	Role in Disease Genetics
Epigenomic Landscapes	Chromatin accessibility, DNA methylation, histone modifications	Regulate gene activity, influence disease susceptibility, mediate non-coding variant effects
Non-coding Variants	SNPs, CNVs, transposable elements	Affect gene regulation through chromatin modification; often underlie disease heritability
Transcriptomic Profiles	RNA expression levels across tissues and conditions	Reflect functional consequences of genetic variation; link genotype to phenotype
Regulatory Variants	De-enhancers, de-silencers, transcriptional regulatory variants	Mediate gene expression changes; interact with chromatin state
Temporal Dynamics	Changes in gene regulation over time	Capture time-dependent effects of genetic variants

Transcriptomic Profiles and Gene Regulation

Biological signals must be integrated at multiple scales of size and time to explain how genetic and environmental factors converge on physiology to influence disease states. Epigenomics reflects the large-scale regulation of genes but does not directly identify those genes. However, the activity of a regulatory element does indicate its influence on the expression of a nearby target gene. Genome-wide transcriptomic profiles directly inform gene activity and abnormal variants influence them, which needs consideration in any mechanistic model explaining how human-average epigenomic landscapes influence human phenotypes in disease contexts (R. Kelley et al., 2018).

Limitations of Traditional Approaches

Epigenomic and transcriptomic signals can be modelled separately. While biologists interpret this data as complementary evidence, distinct models poorly capture interdependencies. Missing integrated statistics diminishes performance in downstream tasks, including prioritizing variants and fine-mapping causal genes (E Hoffman et al., 2019). Independent characterizations also disregard gene regulatory principles and the common influence of trans factors on multiomic transcriptional states. Within the cell nucleus, an epigenomic landscape encodes regulatory potential in chromatin structure, enabling forecasts of gene activity for any genomic locus. Integrating transcriptome data into epigenomic-stage–focused modeling both constrains prior knowledge and models how regulation directly manifests as transcriptomic output. Large-scale deep partially observed variable models help capture agnostic latent factors common to multiple statistics while satisfying intricate constraints imposed by explanatory graphs.

Data Integration Frameworks for Large-Scale Models

Biological and medical research is advancing at unprecedented speed and applying novel techniques to investigate previously inaccessible cellular activities across multiple dimensions, including in vivo images, spatially-resolved transcriptomes, and single-molecule RNA distributions. The human genome remains a crucial biological dimension, providing an expansive blueprint for basic and translational research. However, the connection between genome and the resulting phenotypes remains elusive, and currently available tools that leverage whole-genome and large-scale sequence data to model the genome–

phenotype relationship fall short of comprehensive engagement with the extensive sequence–function information. Significant advances have been made in human cellular signal-response measurement techniques such as droplet-based single-molecule imaging, microfluidic time-lapse fluorescence imaging, and transcriptome-scale activity-based profiling of primary tissue open for analyzing combinations of structural and functional data captured at the same spatial and temporal resolution.

Recent large-scale collections of heterogeneous, multi-omics datasets offer collaborative opportunities to investigate epigenomic and transcriptomic interactions under numerous settings. To unlock such integration, simple extensions of existing approaches falling short on numerous grounds: an individual source cannot be adequately represented for large-scale data when the approaches impose a bottleneck, and high-capacity multi-omics data prevent multiple interactive-dimensional reduction focusing on the modal target. High-capacity multi-omics data encountered in any technologies today and digitized clinical information represent cornerstones of smart health, cell function, human disease, and medicine. Efforts have intensified to leverage such gigantic datasets for solving overarching and essential challenges involving large-scale data probe and ontology-free group specification across multiple dimensions at unprecedented scales. Traditional approaches have too many steps and invariably incorporate fixed-form prior knowledge to few general factors dealing with only or seed-predefined variables in-external spaces, open inquiry a high demand for a new methodology encompassing all potential spatially-associated samples across treatment and microfluidic design dimensions.

Data Sources and Preprocessing

Genomics produces enormous datasets, fueling advances in understanding genetic variation and gene regulatory mechanisms (Chen et al., 2016). Despite considerable research progress, predicting biological outcomes remains challenging, motivating the development of integrative data-driven models. Human epigenomic landscapes influence gene-regulatory activity and disease etiology (Sharmin, 2017). Chromatin accessibility and DNA methylation patterns determine transcription-factor activity, while genomic variants affecting epigenomic aspects modulate gene expression (Sharma, 2018). Because consolidated models of epigenomic and transcriptomic regulation in human disease remain scarce, an approach integrating extensive epigenomic and transcriptomic datasets in a unified, multimodal framework seeks to clarify these fundamental interactions.

Multiple high-throughput epigenomic assays document open chromatin, histone modifications, and DNA methylation across diverse cell types, offering a comprehensive view of the complex determinants governing gene activity. Improved characterization of regulatory elements prompts increasing interest in links between epigenomic landscapes, genetic variants, and transcriptional responses. Systematic mapping of epigenomic features identifies the genomic elements modulated by diverse genetic variants, forming the basis for coherent predictions of downstream effects on transcription. Large datasets coexist with vast transcriptomic repositories, linking expression levels with observable phenotypes. The inclusion of transcriptomic data enhances the biological understanding of epigenomic alterations, providing an additional layer of information to improve the correctness of downstream predictions.

Representation Learning for Omics Signals
Representation Learning for Omics Signals
Holistic models of human disease are rooted in biological processes that span multiple molecular modalities. To capture these processes fully, large-scale, deep-learning models must integrate multi-omics data, which encompass measurements at diverse spatial and temporal resolutions. However, connectivity across omics modalities remains poorly established for many diseases [table 2]. Epigenomic landscapes regulate gene activity and underlie transcriptomic profiles. Variation in disease-perturbed epigenomic signals determines how genetic variation affects disease-relevant transcripts. Representing epigenomic

modalities as stacked time-series facilitates their multi-omics alignment with transcriptomic profiles, which also capture regulatory variation and concomitant disease phenotypes. Complementary assumptions predicate that epigenomic and transcriptomic models operate on different time scales and that transcriptomic representation remains largely unchanged when conditioning solely on cell-type–specific epigenomics. (Sharmin, 2017)

Aligning Epigenomic and Transcriptomic Signals

To connect large-scale models for disease genomics with characterization of gene regulatory mechanisms, epigenomic and transcriptomic signals must be aligned. Epigenomic landscapes impact gene regulation and are altered by non-coding genetic variants implicated in diseases (E Hoffman et al., 2019). Transcriptomic profiles reflect regulatory states and manifest as phenotypic changes, but existing approaches represent these modalities separately (Sharmin, 2017). Integrative models incorporating multimodal data become essential to foster interpretable predictions for complex human diseases and select relevant experimental techniques for downstream analysis. Multimodal models can learn shared representations of complementary signals through alignment and fusion. Epigenomic and transcriptomic modalities are aligned by linking raw signals from the same genomic regions across assays for transcriptome-wide association studies. Content-discriminative pretraining and subsequent contrastive learning align representations while avoiding collapse into a single mode. Additional strategies further enhance integration. A generic framework that enables simultaneous consideration of large-scale, context-specific epigenomic and transcriptomic signals across diverse diseases is thus established.

Model Architectures for Unified Interpretation

In conjunction with recent advances in deep learning and artificial intelligence (AI), the scientific community is witnessing an unparalleled expansion of biological and biomedical data. During the past five years, large-scale transcriptomic datasets have emerged, enabling the specification of gene expression changes throughout biological processes in different organisms and tissues (E Hoffman et al., 2019). Moreover, a series of large-scale epigenomic chromatin-accessibility datasets for various organisms, tissues, and conditions were released, allowing the systematic study of the 3D epigenomic regulatory landscape in conjunction with large-scale transcriptomic data. The integration of epigenomic and transcriptomic signals can not only aid in the definition of regulatory–genes and the construction of regulatory networks but also facilitate the exploration of the impact of epigenomic and transcriptomic signals on gene function. Large-scale AI models have gained momentum in biological and biomedical fields because of their capacity to integrate and leverage diverse biological knowledge. However, most current models still concentrate on a single molecular level when interpreting complex biological phenomena. The use of various omics data to address large-scale biological topics in human disease genetics, human development and stem-cell regulation, and organism evolution has been investigated, but extensive epigenomic and transcriptomic datasets now permit the study of omics data across different molecular levels (Esser-Skala & Fortelny, 2023). Integrating epigenomic and transcriptomic signals in large-scale models for the systematic analysis of human disease genetics remains largely unexplored. Addressing this research gap can not only further promote scientific discovery in disease genetics and various other fields but also provide clinically relevant information for stratifying treatment–response risk and predicting disease onset under each specific context.

Multimodal Neural Architectures

Biological processes are governed by varying degrees of regulatory control. A multi-modal deep learning framework integrates chromatin accessibility and gene expression measurements to predict regulatory effects of genetic variants (Tan & Shen, 2023). In complex eukaryotic systems, inherited sequences have

limited ability to model gene expression and epigenetic features targeted by regulatory variants. Chromatin accessibility modulates local transcription factor interaction possibilities and acts as a multi-scale structural code governing long-range interactions (R. Kelley et al., 2018). A model capable of integrating functionally diverse chromatin structure and DMS-seq assays across conditions could enhance generalization for population genetics and interrogate the diversity and plasticity of regulatory information.

Incorporating Prior Biological Knowledge

Integrative and interpretable models require careful attention to prior biological knowledge. Three principles guide the design of epigenomic–transcriptomic architectures. First, prior knowledge of disease-dependent regulatory mechanisms motivates the simultaneous integration of epigenomic and transcriptomic signals. Both partial depictions of the living cell, each modality contributes unique insights to gene regulation, and jointly modelling the two channels enhances the contextualization of the gene activity signal (E Hoffman et al., 2019). Certain chromatin–transcriptomic feedback loops also exert their influence unidirectionally. At the same time, it remains crucial to capture condition-specific dependencies even when integrating multiple omic modalities, since these dependencies typically govern gene regulation (R. Kelley et al., 2018).

Second, biologists can possess useful insights that shape the model architecture and sampling strategies. Several prior-specification opportunities arise in the course of data integration, likelihood-learning, and homologous-signal modelling. For instance, the archived regulatory networks of diverse genomes and species illuminate specific cross-chromosomal dependencies whose relevance to human loci is uncertain and remain therefore excluded from cross-genome, cross-species modelling. Finally, biological rules about genic regulation suggest that most unregulated genes should either remain inactive altogether or maintain a slim readout, resulting in sparse epigenomic profiles. Consequently, the model encourages sparsity during epigenomic signal reconstruction when the transcriptomic signal indicates that a transcript remains unproduced.

Interpretability and Explainability Mechanisms

Deep learning models are treated as black boxes whose predictions are difficult to interpret, posing challenges for users seeking insight into the underlying biological mechanisms of interest. Accordingly, several researchers have investigated how to provide interpretable explanations for model predictions in genomics. For these models, the goal is to assess the influence of genomic and epigenomic perturbations on the output. Neural attention mechanisms were used to discern the regulatory influence of DNA sequence on gene expression when sequencing reads and chromatin accessibility are considered as input signals. In particular, the relative contribution of different genomic regions, cell types, and time points to the transcriptional activity of a target gene is identified. This approach highlights the most relevant candidate regulatory elements impacted by genetic variants associated with complex traits (Graziani et al., 2022). A convective attention mechanism is proposed to weight measures of histological features that affect the expression of target genes in cancer. It identifies non-obvious connections between tissue appearance and gene expression values, with the additional benefit of detecting cancer subtypes most conducive to multi-omic modeling (S. Watson, 2022).

Applications to Human Disease Genetics

Genetic variants associated with human diseases frequently reside in non-coding regions of the genome, leaving the underlying causal genes and disease mechanisms obscure. Unifying epigenomic and transcriptomic signals through large-scale AI models helps interpret the consequences of such variants (E

Hoffman et al., 2019). Large-scale epigenomic datasets provide chromatin accessibility, histone modification, and DNA methylation maps that delineate epigenomic landscapes at single-base resolution. These datasets indicate which genes are likely to be regulated for any given variant, thereby linking gene activity to disease context. Complementing the epigenome, transcriptomic datasets reflecting steady-state mRNA expression across diverse tissues and cell types further characterize gene regulatory activity. Integrating epigenomic and transcriptomic signals through AI models allows scientists to identify condition-specific regulatory networks, predict genes implicated in diseases without prior knowledge, and anticipate how disease states modulate gene regulation. These insights support stratified medicine, elucidate regulatory mechanisms governing the effects of risk variants, and enhance prediction of genetically driven disease manifestation.

Variant-to-Gene Mapping and Functional Annotation

The vast majority of risk-associated genetic variants identified by genome-wide association studies (GWAS) reside in non-coding regions, making it difficult to determine their functional role and link them to a specific target gene (E Hoffman et al., 2019). To address this challenge, a variant-to-gene mapping approach is developed based on regulatory signals, aiming to identify putative target genes for a given variant. In addition to variant-to-gene mapping, functional annotations that characterize the role of the variant in the gene regulation process are also provided. The predicted annotations are customized for diverse diseases and tissues, significantly enhancing the understanding of how these variants could contribute to different diseases (Butkiewicz et al., 2018).

In conclusion, integrated epigenomic and transcriptomic signals containing regulation information for multiple diseases and tissues are captured using biomolecular language models. These signals are further combined with large-scale genomic annotations to establish a framework for functional interpretation. Consequently, a variant-to-gene mapping approach and customized functional annotations per disease and tissue are implemented, fostering a better understanding of the roles of genetic variants in human diseases.

Condition-Specific Regulatory Networks

Condition-specific regulatory networks were modeled to investigate context dependencies. A central hypothesis posits that refined gene-centered annotations, guided by integrated epigenomic and transcriptomic signals, enhance the mapping of regulatory variants to target genes and expand the interpretation of single-gene variants across multiple conditions (R. Kelley et al., 2018). Regulatory landscape conditioning assists in delineating networks governing distinct cellular responses and highlights potential circuit rewiring accompanying disease transitions. A complementary conjecture suggests that the interpretation of pretrained large-scale models benefits from integrated omics signals, as they enrich features with biological relevance in both the pretext and target tasks (E Hoffman et al., 2019). These integrated signals are expected to unveil context-specific dependencies among input variants, weights, and activities across diverse settings. Statistically, joint-distribution improvement serves as a convenient yet crude criterion to confirm these predictions (Abdurakhmanov J., et al).

The resulting set of 235 condition-specific regulatory networks encompasses various diseases, cell types, and treatment responses, thereby extending the lexicon for disease-gene association analysis. Changes in tissue, cell state, perturbation, or disease state frequently impose regulatory condition shifts, necessitating the refinement of functional prediction frameworks in an increasingly stratified medicine landscape. Disease-gene relationship models, positioned upstream of gene-disease mapping, retain considerable knowledge about variant regulation and activity. Gene-wise predictions predicate these activities only over a narrowly defined set of regulatory variants, curtailing the examination of network context. Integrated signals from diverse-scale pretrained models that encapsulate substantial regulatory information across

extensive tissues and conditions may facilitate context expansion and augment gene-gene relationship comprehension.

Stratified Medicine and Risk Prediction

Information from epigenomic and transcriptomic variables is crucial for effective stratification of patients and accurate prediction of polygenic disease risk. Integrative frameworks for gene regulatory models are needed to understand omics-related dependencies and how they vary across individuals. Individually common single-nucleotide polymorphisms (SNPs) confer only weak predictive power, because of the millions of unknown interactions among dozens of genes. Integrative models take high-dimensional sequences of genetic variants as input, and specify the distributions of RNA-seq reads, transcript counts, or splicing isoforms accordingly. During model training, transcriptomic signal and genotype data are jointly captured. Clinical datasets provide records of clinical variables, but omics data are sparse and inconsistent across different populations. Using standardized procedure, models are pre-trained on the large public resources. Fine-tuning on independent omics datasets acquired from the target population reveals pervasive transferability of learned omics representation.

Large-scale artificial intelligence models are applied to deduce the associated biological insights from population-scale genome-wide association studies, unifying multiple key datasets and capabilities. Annotated variants can be traced back to annotated genes together with their biological insights, leveraging the underlying mechanisms for complex diseases. Such approaches map genetic variants to gene transcripts, and predict condition-specific transcriptomic alterations from genotype data. Starting from the genetically predetermined transcriptomic state, a steady-state solution can be estimated to profile omics-altered genes, unveil latent trans-acting or cis-acting information, and model transcriptomic changes. These efforts, in the framework of polygenic risk prediction, incorporate extra transcriptomic signals into the forward-pass models. Integrative models, through consolidated training, enrich genetic signals with gene-gene interactions by linking genotype, environment, transcriptome, and microbiome variables (Sik Wai Ho et al., 2019).

Evaluation Frameworks and Benchmarks

To comprehensively assess the quality and suitability of multi-omics and omics-to-phenome approaches in applications for disease genetics, an evaluation framework comprising several benchmarks that cover interpretability, integration, and predictive performance is proposed. The benchmarks for interpretability adaptation and extension include already defined metrics of biological plausibility, information extraction, and causal feature identification (E Hoffman et al., 2019) ; the evaluation of integration corresponds to the capability of properly merging multiple data sources and enhancing an overall analysis; and the measures for predictive capacity explain how integrated signals affect target variables, i.e. phenotype or disease (Abdurakhmanov J., et al).

Two different ontological and empirical approaches for assessing disease understandability guided the various evaluation methodologies. A first validation against specimen-specific, publicly accessible epigenomic and transcriptomic profiles associates convincing discoveries in the literature from cultured cellular systems with GENCODE and Roadmap Epigenomics resources assessing the relevance of the inferred regulations (Trejo Banos et al., 2020). A second set of additional distant transcriptional datasets and diverse multi-phenotype genetic features and annotations enables to leverage broad cohort information for interpreting multi-phenotype accessibility and transcriptional control. The pertinence of results can be investigated across diseases covering cross-dataset, cross-species, and cross-population use cases to explore the general applicability of findings beyond analysed scenarios. Finally, experiments of noise

injection evaluate robustness to perturbations in both training signals and consequently generated interpretations (Ziyaev A.A., et al).

Metrics for Omics Integration

Epidemiological estimates indicate that epigenetic mechanisms account for approximately 75% of all diseases (Zitnik et al., 2018). Epigenomic landscapes define the regulatory state of genes and identify the subset of genes that are actively regulated in a given context. The corresponding expression profile, i.e., the transcriptome, reflects the activity of regulatory elements inferred from the epigenomic landscape and serves as an indicator of the underlying regulatory state (Bhardwaj & Van Steen, 2020). Integrating heterogeneous data types improves predictive and explanatory performance and enables a more comprehensive understanding of complex biological phenomena. Exploiting the interdependence between gene regulation and transcriptomic output supports the simultaneous integration of epigenomic and transcriptomic signals in a single predictive model.

Large-scale epidemiological studies differentiate major disease categories based on genetic, epigenetic, and transcriptomic factors. Advancing the understanding of each disease from an integrated, population-scale, multi-omics perspective requires efficient investigation of individual variations that go beyond these category definitions. Joint elucidation of both condition-specific regulatory networks and the regulatory condition of individual variations is thus critical.

Validation against Experimental Data

Transcriptomic signals predicted by the integrative framework were validated against experimental FANTOM5 expression measurements in multiple tissue contexts (R. Kelley et al., 2018). These datasets comprise large-scale, high-quality, and harmonized estimates of gene expression levels in diverse human adult and fetal tissues. To quantitatively assess the relationships between input epigenomic signals and transcriptomic predictions, those predicted from the integrated epigenomic-transcriptomic models were compared with transcriptomic signals predicted solely from epigenomic data by epigenomic-only models. Within the independent FANTOM5 tissue context, the correlation between these two transcriptomic predictions faithfully reflected the correlation between expression levels measured in the same tissues and remained comparable across a range of tissues. Collectively, these observations confirmed the biological validity of the predicted transcriptomic signals, supporting the integrative modeling of epigenomic and transcriptomic data to enhance the understanding of regulatory systems governing multicellular gene expression.

The integrative modelling framework was further validated across multiple cell types and experimental conditions using an independent large-scale perturbation dataset from the ENCODE project. This meta-dataset contains systematic perturbation experiments on various individual regulators (e.g., transcription factors, histone acetylation, etc.), allowing quantitative evaluation of the gene regulatory effects of these factors on the epigenome and the downstream consequences for transcriptional regulation. Comparison with predicted epigenomic data generated from epigenomic-only models (E Hoffman et al., 2019) indicated that the predicted effects both before and after perturbation remained substantially unchanged across diverse conditions, whereas the predicted fluxes at the transcriptomic level consistently reflected the magnitude of these perturbations. These findings suggested that downstream regulatory influences on the transcriptome were effectively captured during the integrative modelling of multi-omics datasets, thus corroborating the versatility and robustness of the framework. Following recent progress in linking genotype to phenotype via simulatable models that learn biological causality across different biological molecular layers, a corresponding genotype–phenotype link was established using this modelling framework, providing further support for the validity of the integrated models.

Cross-Disease and Population Generalization

Integrative epigenomic and transcriptomic (ET) landscapes modulate gene activity in a dynamic and largely population-specific manner. To systematically examine the cross-population and cross-disease generalization capability of AI systems leveraging such large-scale ET data at multiple cellular states, populations, and diseases, cross-disease, and cross-population machine-learning benchmarks are created across five organs for the largest transcriptomic and epigenomic ATAC-seq and RNA-seq dataset. The usefulness of transcriptomic information in predicting epigenomic chromatin accessibility, even in unseen diseases, populations, and tissues, is demonstrated. Models built solely on epigenomic data achieve similar generalization capabilities, indicating that learned ET representations of chromatin accessibility and gene activity faithfully capture population-and disease-agnostic biological principles. Evaluating reverse modeling, chromatin-accessibility-atlas models also exhibit extensive generalization capabilities, allowing exploration of epigenomic cross-tissue regulatory transforms (Allabergenov M., et al).

The ability of methods to utilize easily ascertainable signals to predict higher-level features enables generalization assessments across different dimensions. AI systems leveraging ATAC-seq and RNA-seq signals are applied to the task of elucidating chromatin-accessibility and gene-activity coupling in an unseen cell type, cardiomyocytes. Without specific training, models still successfully identify the chromatin regions associated with cardiac-related genes, and when trained on publicly available chromatin and gene-activity data from an unseen population, the generalization remains proficient. Sequential modeling of the chromatin-architecture-to-chromatin-accessibility sequence indicates that the spatial organization constrains general temporal evolution of chromatin-accessibility changes, and the combination of epigenomic and transcriptomic ATAC-seq and RNA-seq observations from all three tasks constitutes the most generalizable scheme of the study.

Ethical, Legal, and Social Considerations

A guiding framework for building regulatory-genetic models of complex human diseases using transcriptomic and epigenomic signals can benefit from careful consideration of ethical, legal, and social implications. These considerations collectively promote responsible research practices and help prevent unintended harms as predictive models gain translation to clinical use. Large-scale data sets containing biological and genomic information can enable researchers to build sophisticated models of the relationship between genetic variation and human disease. However, these models depend on the availability of data collected in compliance with appropriate ethical, legal, and social frameworks. When building or openly sharing large data sets, researchers must respect data sharing policies, privacy rights, and related legal and ethical frameworks governing individual data, population data, and human rights more broadly (A Walton et al., 2023). Accordingly, it is important to collaborate with people well-versed in the regulatory landscape before embarking on large-scale data integration projects. Furthermore, careful design of study protocols for data collection can help mitigate legal and ethical risks. Protocols that use anonymized and de-identified data whenever possible can facilitate aggregation of the broad sets of high-quality observations needed for building and evaluating complex predictive models while respecting individual privacy (Azimova S., et al).

Machine learning models increasingly influence decisions about the allocation of medical resources and the provision of individual treatment. These models also have the potential to affect reproductive choices and personal and public health in other ways that persist along social, economic, and demographic dimensions. The potential for bias to enter into these models, including through algorithmic, technical, and social avenues, remains an active area of exploration. Bias that reflects or exacerbates existing inequities can intensify the already tremendous burden to individuals and communities already affected

by poor health, poverty, and lack of access to care (Ziyaev A.A., et al). Consequently, the prospective risks associated with applying these models to clinical decision-making need explicit consideration, monitoring, and mitigation. Furthermore, distortion of model predictions when moving from the data generation process to a practical deployment context also needs careful examination. Efforts to document model development in machine learning research apply equally to machine learning research directed at human genetics and extend its principles by including indications of how model predictions were affected by population stratification, data leakage, shift drift, and other important such phenomena (Azimova S., et al).

Large-scale omics data sets offer the potential to build comprehensive, biologically grounded models of the relationship between genomic variation and diverse disease and trait outcomes. However, scale raises questions about how and when sufficiently authoritative information may be shared publicly as a contribution to the open science initiative. As open science principles further develop, adopting practices aligned with those evolving principles remains essential, including when omics data, records, or models are prepared for subsequent knowledgeable use as a basis for advancing predictive and explanatory capabilities. Such an approach promotes recognition of the efforts that multiple stakeholders have made to share, curate, and enhance publicly accessible data and models (Mannonov A., et al). Open science practices and principles also foster greater transparency about the limitations of pre-existing data and models used to build upstream approaches, thus providing guidance about how those limitations might carry forward into useful downstream applications.

Data Privacy and Consent

The increasing deployment of large-scale machine-learning models in genomics raises ethical concerns about data privacy and the acquisition of valid consent for data use, particularly when the models are trained on publicly available data from a wide range of research studies covering numerous individuals with disease indicators (Brauneck et al., 2024). In this context, privacy is understood as a set of rights that protects people's ability to determine what personal data about them can be gathered, analyzed, and disseminated; it also encompasses the right to decide whether one’s insights and data remain confidential. It is paramount to collect only the necessary information for research purposes (C. Rivas Velarde et al., 2021). The potential risks associated with data collection include loss of privacy and self-determination, particularly when population-specific or culturally sensitive information is involved. Specific categories of sensitive data, such as genetic information, urban mobility, social networks, and media consumption, are subject to particular scrutiny during processing.

Bias, Equity, and Clinical Translation

Recent advances in artificial intelligence (AI) promise to accelerate the interpretation and translation of genome-wide association studies (GWAS) into better understanding of human biology and complex diseases (E Hoffman et al., 2019). However, substantial barriers remain in adopting large-scale AI frameworks in the analysis of genotype-to-phenotype relationships. The dataset must span a wide range of biological contexts to facilitate generalizable inference and the mapping of variants to genes must account for ubiquitous non-coding regulatory mechanisms (Trejo Banos et al., 2020). Models must incorporate biological knowledge and remain interpretable to the researcher (Sharmin, 2017).

Within the framework of the project, a strategic plan has been developed for the construction, evaluation, and biomedical application of large-scale AI models to jointly interpret genetic and epigenetic data. The approach seeks to empower researchers working in academia, industry, and related sectors with better tools for systematic analysis of GWAS data and more efficient translation of genetic variation data into biological insights. The objective of the work is to develop a comprehensive model for the integration of

epigenomic and transcriptomic signals across diverse conditions to foster interpretable AI. Population-level datasets are rapidly accumulating to profile multi-omic signals under various biological conditions, yet the integration of diverse epigenomic and transcriptomic data remains limited. Researchers following this project are investigating new integration frameworks for large-scale, multi-omic models that align and fuse signals from diverse epigenomic and transcriptomic assays accrued under disparate biological conditions. Such systematic integration is expected to yield a more complete understanding of how epigenomic landscapes establish regulatory programs that determine gene activity and how differentially expressed transcripts reflect the activation status of these regulatory programs (Sasmakov S.A., et al).

Reproducibility and Open Science

Reproducibility and open science are essential for advancing genetic research. Understanding how genetic variants influence gene expression, chromatin states, and disease traits benefits from techniques such as chromatin-state discovery, allelic expression analysis, and fine mapping of causal variants (Chen et al., 2016). Studies have explored genetic and epigenetic interactions affecting autoimmune diseases and inflammatory bowel disease. Managing batch effects and confounding factors is critical for accurate data analysis, with methods like empirical Bayes and the sva package. Integrative approaches combining epigenomic data with GWAS help identify genetic variants with direct regulatory roles. Open resources such as the GENCODE annotation and the African Genome Variation Project facilitate cross-study comparisons and reproducibility. Efficient computational tools for sequence alignment, SNP calling, and trait analysis are vital for high-throughput data processing. Overall, sharing datasets, standardized protocols, and open-source tools underpin reproducibility in open science initiatives (Sharmin, 2017).

Future Directions and Open Questions

Integrative analysis of regulatory signals across the genome, transcriptome, and other modalities is a promising, yet underexplored, avenue for disease-gene elucidation. Epigenomic landscapes govern gene regulation in diverse biological contexts, and the corresponding regulatory states are temporally and spatially coordinated across the genome. Although many large-scale epigenomic studies have characterized regulatory landscapes in disease-relevant samples, the specific influence on transcriptomic profiles remains largely unknown. Jointly modelling epigenomic and transcriptomic states therefore offers a valuable opportunity to infer the regulatory effects of noncoding variation and other genomic alterations. Deep learning models with architectural bias toward multimodal data integration can jointly capture regulatory and transcriptional signals. In parallel, transfer-learning strategies that adapt prior knowledge acquired from purely descriptive prediction tasks-such as the identification of regulatory and transcribed elements-to illustration of regulatory activities at base-pair resolution enable insight into broader regulatory mechanisms. Such approaches hold promise for advancing the predictive understanding of how noncoding variation drives disease-relevant phenotypes. (R. Kelley et al., 2018)

Conclusion

Human physiology is governed by the universally conserved genetic code. Yet, genes are silenced or activated by multiple regulatory systems to yield diverse cell types and phenotypes with characteristic functions (Sasmakov S.A., et al). Cells, tissues, and organs retain identical genomic sequences and signals. All biological responses are encapsulated in an epigenome, progressively encoded at distinct frequencies in DNA, RNA and protein. Synthetic Biology aims to develop cell- and context-specific therapeutic agents. The present framework delineates a new mechanistic paradigm elucidating the origin and propagation of diseases and their regulated drug-treatment signatures. At the heart of the method lies a biologically-rooted numerical Representation Theory combined with the relevant Mathematics, a meticulous Duality Theory unifying transcriptomic and epigenomic platforms, a top-layer comprehensive Knowledge Graph and a State-of-the-Art Generative Model underpinning a subtle Artificial Intelligence Deep Learning procedure.

The proposed signal derives from the premise that humans and all living organisms are subjected to unrelenting discrete signals by external parameters including climate, temperature, humidity, light, and food intake, combined with a plethora of stochastic internal signals. (E Hoffman et al., 2019) (R. Kelley et al., 2018)

References:

[1] 1. Hoffman E., Bendl J., Girdhar K., Schadt E., Roussos P. Functional interpretation of genetic variants using deep learning predicts impact on chromatin accessibility and histone modification // NCBI. - 2019. - URL: https://www.ncbi.nlm.nih.gov

[2] 2. Sharmin M. Model based approaches to characterize heterogeneity in gene regulation across cells and disease types. - 2017. - [PDF]

[3] 3. Chen L., Ge B., Casale F., Vasquez L., Kwan T., Garrido-Martin D., et al. Genetic drivers of epigenetic and transcriptional variation in human immune cells. - 2016. - [PDF]

[4] 4. Pei G., Hu R., Dai Y., Manuel A.M., Zhao Z., Jia P. Predicting regulatory variants using a dense epigenomic mapped CNN model elucidated the molecular basis of trait-tissue associations // NCBI. - 2020. - URL: https://www.ncbi.nlm.nih.gov

[5] 5. Kelley R.D., Reshef Y.A., Bileschi M., Belanger D., McLean C.Y., Snoek J. Sequential regulatory activity prediction across chromosomes with convolutional neural networks // NCBI. - 2018. - URL: https://www.ncbi.nlm.nih.gov

[6] 6. Sharma S. Integrative analysis of complex genomic and epigenomic maps. - 2018. - [PDF]

[7] 7. Esser-Skala W., Fortelny N. Reliable interpretability of biology-inspired deep neural networks // NCBI. - 2023. - URL: https://www.ncbi.nlm.nih.gov

[8] 8. Tan W., Shen Y. Multimodal learning of noncoding variant effects using genome sequence and chromatin structure // NCBI. - 2023. - URL: https://www.ncbi.nlm.nih.gov

[9] 9. Graziani M., Marini N., Deutschmann N., Janakarajan N., Müller H., Rodríguez Martínez M. Attention-based interpretable regression of gene expression in histology. - 2022. - [PDF]

[10] 10. Watson S.D. Interpretable machine learning for genomics // NCBI. - 2022. - URL: https://www.ncbi.nlm.nih.gov

[11] 11. Butkiewicz M., Blue E.E., Leung Y.Y., Jian X., Marcora E., Renton A.E., et al. Functional annotation of genomic variants in studies of late-onset Alzheimer's disease // NCBI. - 2018. - URL: https://www.ncbi.nlm.nih.gov

[12] 12. Ho S.W., Schierding W., Wake M., Saffery R., O'Sullivan J. Machine learning SNP based prediction for precision medicine // NCBI. - 2019. - URL: https://www.ncbi.nlm.nih.gov

[13] 13. Trejo Banos D., McCartney D.L., Patxot M., Anchieri L., Battram T., Christiansen C., et al. Bayesian reassessment of the epigenetic architecture of complex traits // NCBI. - 2020. - URL: https://www.ncbi.nlm.nih.gov

[14] 14. Zitnik M., Nguyen F., Wang B., Leskovec J., Goldenberg A., Hoffman M.M. Machine learning for integrating data in biology and medicine: principles, practice, and opportunities. - [PDF] - 2018

[15] 15. Bhardwaj A., Van Steen K. Multi-omics data and analytics integration in ovarian cancer // NCBI. - 2020. - URL: https://www.ncbi.nlm.nih.gov

[16] 16. Walton A.N., Nagarajan R., Wang C., Sincan M., Freimuth R.R., Everman D.B., et al. Enabling the clinical application of artificial intelligence in genomics: a perspective of the AMIA Genomics and Translational Bioinformatics Workgroup // NCBI. - 2023. - URL: https://www.ncbi.nlm.nih.gov

[17] 17. Brauneck A., Schmalhorst L., Weiss S., Baumbach L., Völker U., Ellinghaus D., et al. Legal aspects of privacy-enhancing technologies in genome-wide association studies and their impact on performance and feasibility // NCBI. - 2024. - URL: https://www.ncbi.nlm.nih.gov

[18] 18. Rivas Velarde C.M., Tsantoulis P., Burton-Jeangros C., Aceti M., Chappuis P., Hurst-Majno S. Citizens' views on sharing their health data: the role of competence, reliability and pursuing the common good // NCBI. - 2021. - URL: https://www.ncbi.nlm.nih.gov

[19] 19. Allabergenov M., et al. Intelligent educational environments and ubiquitous computing for continuous learning and digital literacy development // Journal of Wireless Mobile Networks, Ubiquitous Computing and Dependable Applications. - 2024. - Vol. 15, No. 4. - P. 179–191
[20] 20. Mannonov A., et al. The philological library as a modern architectural icon for knowledge and research // Indian Journal of Information Sources and Services. - 2025. - Vol. 15, No. 1. - P. 388–394
[21] 21. Mannonov A., et al. The impact of Uzbek-language mobile libraries on digital education // Indian Journal of Information Sources and Services. - 2025. - Vol. 15, No. 1. - P. 315–319
[22] 22. S. Sindhu. (2025). Causality-Aware Event-Driven Learning for Distributed Frequency Regulation over Wireless Control Channels. Journal of Wireless Intelligence and Spectrum Engineering, 23–30.
[23] 23. Abdurakhmanov J., et al. Cloning and expression of recombinant purine nucleoside phosphorylase in the methylotrophic yeast Pichia pastoris // Journal of Advanced Biotechnology and Experimental Therapeutics. - 2023. - DOI: 10.5455/jabet.2023.d153
[24] 24. Ziyaev A.A., et al. Synthesis of S-(5-aryl-1,3,4-oxadiazol-2-yl) O-alkyl carbonothioate and alkyl 2-((5-aryl-1,3,4-oxadiazol-2-yl)thio) acetate, and their antimicrobial properties // Journal of the Turkish Chemical Society, Section A: Chemistry. - 2023. - DOI: 10.18596/jotcsa.1250629
[25] 25. Azimova S., et al. Study of the immunogenicity of combination of recombinant RBD (Omicron) and nucleocapsid proteins of SARS-CoV-2 expressed in Pichia pastoris // The Open Biochemistry Journal. - 2023. - DOI: 10.2174/011874091x273716231122102205
[26] 26. Sasmakov S.A., et al. Expression of recombinant PreS2-S protein from the hepatitis B virus surface antigen in Pichia pastoris // VacciMonitor. - 2021. - Vol. 30, No. 1. - P. 27–32