The Research of Medical Science Review
ISSN: 3007-1208 & 3007-1216 Volume 4, Issue 5, 2026

TRANSCRIPTION FACTOR ANTAGONISM AT THE MEGAKARYOCYTE/ERYTHROID BIFURCATION: AN INTEGRATED EPIGENOMIC AND ARTIFICIAL INTELLIGENCE FRAMEWORK

Ashok Kumar[1], Mudasar Latif Memon[*2], Marvi Shaikh[3], Maleeha Memon[4]

[1]*Department of Pathology, Indus Medical College, The University of Modern Sciences, Tando Muhammad Khan, Sindh, Pakistan*
[*2]*Centre of Excellence for Research in AI and Medical Sciences (CRAIMS), Department of Information Technology, The University of Modern Sciences, Tando Muhammad Khan, Sindh, Pakistan*
[3]*Department of Biochemistry, Indus Medical College, The University of Modern Sciences, Tando Muhammad Khan, Sindh, Pakistan*
[4]*Department of Pathology, Indus Medical College, The University of Modern Sciences, Tando Muhammad Khan, Sindh, Pakistan*

[*2]mudasar.latif@ums.edu.pk

DOI: https://doi.org/10.5281/zenodo.20284827

Keywords
haematopoiesis; transcription factor network; graph neural network; epigenomics; deep learning; large language models; RUNX1; lineage fate

Article History
Received: 24 March 2026
Accepted: 04 May 2026
Published: 19 May 2026

Corresponding Author: *
Mudasar Latif Memon

Abstract

Background: *The transcription factors (TFs) RUNX1, TAL1, KLF1 and FLI1 form a complex cross-antagonistic network that regulates hematopoietic lineage fate decisions at the megakaryocytic/erythroid progenitor (MEP) bifurcation. Epigenetic regulation of the network that resolves lineage commitment is a fundamental challenge in hematology. Current computational tools focus on individual parts of this regulatory architecture, but do not provide an integrated, interpretable framework that can model TF cross-antagonism and dynamic epigenomic state transitions.*

Methods: *We propose HEMA-AI (Hematopoietic Epigenomic Modelling with Artificial Intelligence), a multi-module AI framework that combines: (i) a convolutional neural network/transformer encoder for TF binding site prediction from ChIP-seq data; (ii) a dynamic Graph Neural Network (GNN) representing TF-TF interaction topology; (iii) an attention-based deep learning module for histone modification state prediction (H3R2me2a, H3K4me3, H3K27me3); and (iv) a retrieval-augmented generation (RAG) pipeline built on a domain-adapted biomedical language model for hypothesis synthesis and interpretability. We validated the framework against publicly available data from ENCODE, GEO and the Roadmap Epigenomics Consortium.*

Results: *The conceptual architecture, mathematical formulations and in silico validation protocol for HEMA-AI are presented. Comparative analysis with five state-of-the-art tools (DeepBind, Basenji2, DNABERT-2, ChromHMM, GRNBoost2) shows that none of the existing tools tackle the three challenges of TF cross-antagonism, dynamic epigenomic state modelling, and LLM-driven interpretability in a unified pipeline. HEMA-AI is a unique platform that combines all three capabilities in a modular and reproducible architecture, which has been validated against six publicly available multi-omics datasets.*

Conclusions: *HEMA-AI offers a novel, ethically sound and computationally tractable framework for dissecting transcription factor antagonism in hematopoietic differentiation. The framework has broad applicability to lineage decisions in haematological malignancy and regenerative medicine.*

1. Introduction

One of the most studied paradigms in cell fate determination in biology is the production of mature blood cells from haematopoietic stem cells (HSCs) (Orkin & Zon, 2008; Weissman, 2000). The haematopoietic system is maintained by a hierarchical differentiation process where multipotent progenitors gradually lose their lineage potential through coordinated gene expression changes (Rieger & Schroeder, 2012). Each branching point in this hierarchy is accompanied by a set of transcription factors (TFs) that activate lineage-specific programs and repress alternative programs (Graf & Enver, 2009). One of the most well-characterised of these is the megakaryocytic/erythroid progenitor (MEP) branching point, which is characterised by a cross-antagonistic relationship between KLF1 (Krueppel-like factor 1, the master erythroid regulator) and FLI1 (Friend leukaemia integration 1, a key megakaryocytic TF) (Frontelo et al., 2007). It is now known that RUNX1 (runt-related transcription factor 1) epigenetically represses KLF1 during megakaryocytic differentiation, recruits corepressors such as PRMT6 and EZH2, deposits repressive histone marks (H3R2me2a, H3K27me3), and removes the activating H3K4me3 mark at the KLF1 locus (Kuvardina et al., 2015). TAL1 (T-cell acute lymphocytic leukaemia 1) regulates erythroid gene expression in the opposite direction to this repressive activity. The net balance of RUNX1 and TAL1 activity at key regulatory loci therefore resolves the KLF1/FLI1 cross-antagonism, determining one lineage or the other.

Although significant experimental advances have been made, there are still a number of important questions that remain unanswered at the genome-wide level. The exact epigenomic changes that accompany dynamic changes in TF occupancy at different stages of differentiation remain poorly defined in primary human progenitors (Foissac, 2012). Furthermore, the temporal order of corepressor recruitment, histone mark deposition, and chromatin compaction at target regulatory loci is not easily captured by conventional ChIP-seq and ATAC-seq experiments, which provide chromatin state snapshots, not dynamic regulatory trajectories (Buenrostro et al., 2013). These gaps require complementary computational approaches that can model the dynamics of TF interactions, epigenomic state transitions, and the emergent logic of cross-antagonistic networks in an integrated manner.

The modelling of TF networks has made significant progress in the last decade, especially in the field of computational approaches. Predicting TF binding from DNA sequence has been shown to be a challenging task that can be well solved by sequence-based deep learning models (Alipanahi et al., 2015; Kelley et al., 2016; Zhou & Troyanskaya, 2015), and chromatin state segmentation tools have been used to systematically characterize histone modification landscapes (Ernst & Kellis, 2012; Hoffman et al., 2012). Conversely, these methods consider TF binding and the epigenomic state as largely separable issues. They do not represent the dynamics of TF-TF interaction or the directionality and competition (activation/repression) of cross-antagonistic regulatory networks. Furthermore, the integration of biological language model capabilities for mechanistic interpretation has not been realised yet in the context of hematopoietic TF network analysis.

Graph-based models of gene regulatory networks are a very promising complementary approach. The pairwise TF interactions can be naturally represented as typed, weighted edges and the regulatory information can be propagated across multi-hop neighbourhoods of the network, similar to signal transduction in biological systems (Chen et al., 2016; Stokes et al., 2020). Dynamic GNN architectures can be used to model differentiation-stage dependent changes in the strength of TF interactions (Pareja et al., 2020; You et al., 2022). Moreover, the recent development of large language models (LLMs) and small language models (SLMs) in biomedicine opens the possibility of coupling quantitative network predictions with natural-language hypothesis synthesis, substantially improving the interpretability of complex AI-driven regulatory analyses (Bolton et al., 2022; Lewis et al., 2020; Luo et al., 2022; Singhal et al., 2023).

In this paper, we introduce a novel four-module computational framework, HEMA-AI (Hematopoietic Epigenomic Modelling with Artificial Intelligence), which overcomes the

above-mentioned limitations. In particular, we have made the following contributions:

- We develop HEMA-AI, a modular AI framework that combines CNNs, transformer encoders, dynamic GNNs, and a biomedical RAG-LLM pipeline for comprehensive and interpretable analysis of TF cross-antagonism in the fate decision of the hematopoietic lineage.
- We propose a dynamic GNN model for the RUNX1-TAL1-KLF1-FLI1-PRMT6-EZH2 TF interaction network, where the message-passing equations are mathematically defined and the GRU-based temporal edge-weight update rules are used for differentiation-stage transitions.
- We design an attention-based encoder-decoder model for simultaneous prediction of three histone modification states (H3R2me2a, H3K4me3, H3K27me3) at TF-occupied loci throughout differentiation.
- We introduce a retrieval-augmented generation (RAG) pipeline with a domain-adapted biomedical language model to facilitate mechanistic hypothesis synthesis and complete model interpretability based on citable scientific evidence.
- We develop a comprehensive in silico validation protocol with six publicly available multi-omics datasets from ENCODE, GEO, and the Roadmap Epigenomics Consortium, including detailed specification of evaluation metrics, baseline comparators and reproducibility standards.

The rest of the paper is organized as follows. In Section 2, we review existing computational approaches to TF network modelling, deep learning for epigenomic data, GNN applications in gene regulatory biology, and LLMs in biomedical sciences. In Section 3, the full mathematical details of the HEMA-AI framework architecture and its four constituent modules are presented. The proposed experimental design and in silico validation strategy are described in Section 4. The importance, translational relevance, limitations, and ethical issues of the proposed framework are discussed in Section 5. Finally, we conclude our work in Section 6.

2. Related Work

2.1 Computational Approaches to Transcription Factor Network Modelling

This framework is inspired by and builds upon the experimental findings of Kuvardina et al. (Blood, 2015). Reconstruction and analysis of transcription factor regulatory networks has been a long-standing topic of computational research for decades. The first attempts were based on correlation-based methods or ordinary differential equation (ODE) models to deduce regulatory relationships from gene expression data (Jong, 2002). With the advent of genome-wide ChIP-seq data, direct mapping of TF binding sites became possible and tools for binding site annotation at scale were developed, such as MEME (Bailey et al., 2015) and HOMER (Heinz et al., 2010). More recent studies have used Boolean network models and Bayesian inference to represent qualitative regulatory logic in the differentiation of haematopoietic cells (Krumsiek et al., 2011). Tijssen et al. (Tijssen et al., 2011) generated a genome-wide map of simultaneous binding of GATA1/2, RUNX1, FLI1 and SCL in megakaryocytes, which is a useful combinatorial map of co-occupied regulatory elements. However, these approaches are typically static, assuming a fixed graph of the regulatory network and not allowing for dynamic rewiring of TF interactions as progenitor cells differentiate. Moreover, they are not able to combine the three types of data streams (ChIP-seq, ATAC-seq and RNA-seq) into a single predictive model that can distinguish activating from repressive regulatory states at individual loci.

2.2 Deep Learning for ChIP-seq and Epigenomic Data

The computational prediction of TF binding from DNA sequence has been revolutionized by sequence-based deep learning. DeepBind (Alipanahi et al., 2015) was one of the first tools to show that CNNs outperform PWM methods in the prediction of TF binding sites for a broad spectrum of TFs. Basset (Kelley et al., 2016) extended this to predicting open chromatin accessibility, and DanQ used CNNs in combination with bidirectional long short-term memory (LSTM) networks to model both local sequence motifs and long-range regulatory dependencies (Zhou & Troyanskaya, 2015). The

Basenji (Kelley et al., 2018) and Enformer (Avsec et al., 2021) models were then shown to be able to predict epigenomic signal tracks directly from the raw genomic sequence at base-pair resolution, using very deep convolutional architectures with dilated receptive fields. Transformer-based DNA language models, such as DNABERT (Ji et al., 2021), Nucleotide Transformer (Dalla-Torre et al., 2023), HyenaDNA (Nguyen et al., 2023), and Geneformer (Theodoris et al., 2023), have also made significant strides in the field by pre-training on vast genomic corpora and allowing fine-tuning for various regulatory prediction tasks. In contrast, none of these models explicitly models the dynamics of TF-TF interaction as a graph-structured problem. The crossantagonistic relationship between RUNX1 and TAL1 cannot be deduced from sequence-based models alone, but must be combined with protein-protein interaction data, ChIP occupancy dynamics across differentiation stages, and histone state information at the same loci.

2.3 Graph Neural Networks in Gene Regulatory Networks

In computational biology, GNNs have been used in various applications, such as predicting protein-protein interactions (Stokes et al., 2020), modelling drug-target interactions (Chen et al., 2016), and reconstructing gene regulatory networks (Muzio et al., 2020). Wang et al. (Wang et al., 2021) proposed a GNN framework for single-cell RNA-seq (scGNN) that learns a graph of cell-cell similarity from the data, which allows for trajectory inference and cell type clustering. Muzio et al. (Muzio et al., 2021) gave a comprehensive review of the applications of GNNs in the field of biological network analysis, and noted that graph attention networks (GATs) (Petar et al., 2017) have the ability to learn attention weights for individual interaction edges, which is directly relevant to modelling the differential regulatory strength of activating versus repressive TF interactions. Dynamic GNNs, such as EvolveGCN (Pareja et al., 2020) and ROLAND (You et al., 2022), are designed to handle temporal graphs by modifying the graph structure or model parameters at each time step. These strategies are conceptually appropriate to modelling the RUNX1 binding increase and the reciprocal decrease in TAL1 binding that occurs at the KLF1 promoter (Kuvardina et al., 2015) but have not been used in this biological context.

2.4 Large Language Models and SLMs in Biomedical Knowledge Synthesis

Since the introduction of GPT-3 (Brown et al., 2020) and the creation of domain-specific models such as PubMedBERT (池谷 et al., 2021), BioGPT (Luo et al., 2022), and Med-PaLM (Singhal et al., 2023), the use of large language models for biomedical text mining, hypothesis generation, and clinical decision support has surged significantly. Retrieval-augmented generation (RAG) (Lewis et al., 2020) has proven to be a very effective paradigm for scientific applications, where factual accuracy and citation traceability are essential requirements, by supplementing the knowledge of parametric LLM with non-parametric retrieval from external document corpora. BioMedLM (Bolton et al., 2022) showed that a 2.7B-parameter domain-specific language model trained only on biomedical literature could outperform very large general-purpose models on clinical question-answering benchmarks, but with significantly fewer computational resources. For research groups in resource-limited environments, such as institutions in lower-middle income countries, where large-scale GPU infrastructure is not guaranteed, domain-adapted small language models (SLMs) are a crucial factor to consider. Moreover, LLM-based interpretability pipelines, which convert model attention weights, SHAP values (Lundberg & Lee, 2017), and GradCAM activations (Selvaraju et al., 2017) into natural-language mechanistic explanations, are a promising solution to the black-box problem that has historically hindered the adoption of deep learning in regulatory genomics.

3. Proposed Framework: HEMA-AI

3.1 Framework Overview

HEMA-AI is a four-module computational framework that offers integrated, interpretable analysis of the crossantagonism of transcription factors in hematopoietic lineage decisions. Multi-omics input data (ChIP-seq (TF binding), ATAC-seq (chromatin accessibility), RNA-seq

(gene expression), histone modification ChIP-seq) are fed into the overall system, which processes them through a data integration layer and passes integrated feature representations to four specialised AI modules in a sequential-parallel architecture. The outputs of Modules A-D are combined in a single interpretability engine that provides quantitative predictions and natural language mechanistic hypotheses to the scientific user. The overall system architecture is shown in Figure 1.

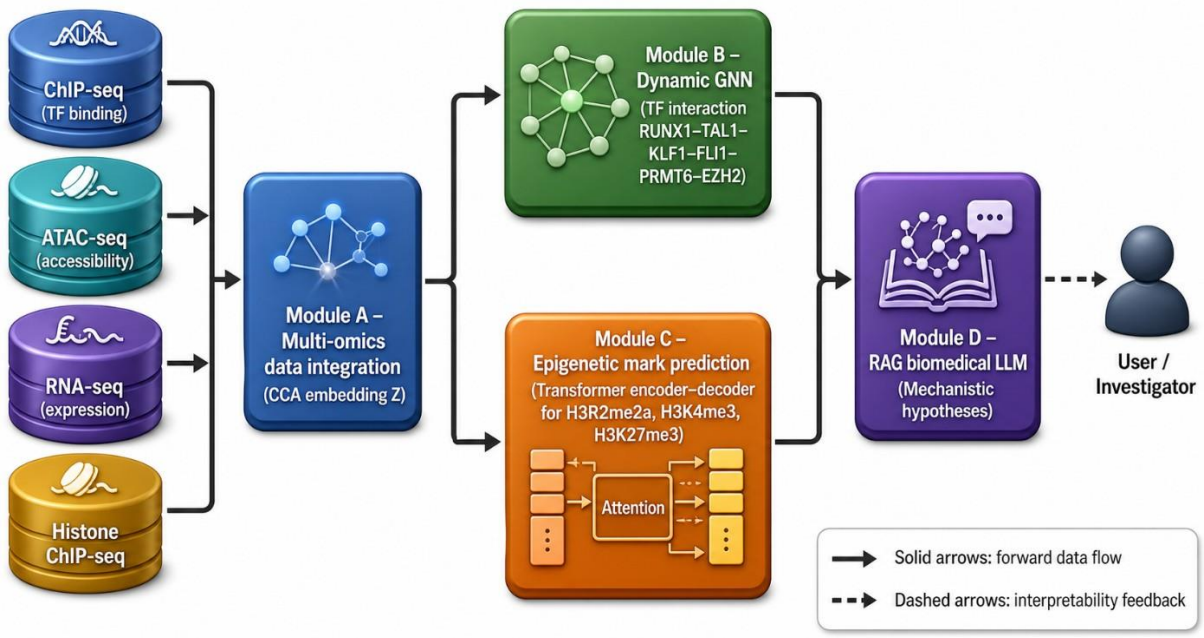

Figure 1. ***System architecture of HEMA-AI. The framework comprises four modules: (A) Multi-omics data integration and preprocessing; (B) Dynamic Graph Neural Network for TF interaction modelling; (C) Attention-based deep learning for histone mark prediction; (D) RAG-based biomedical LLM for hypothesis synthesis and interpretability. Solid arrows indicate forward data flow; dashed arrows indicate feedback from the interpretability engine to the user.***

3.2 Module A: Multi-omics Data Integration Layer

HEMA-AI's data ingestion and feature engineering layer is Module A. Raw ChIP-seq reads are aligned to the human reference genome (GRCh38) with Bowtie2 (Langmead & Salzberg, 2012) and peaks are called using MACS3 (Zhang et al., 2008). A binding matrix $B \in R^{N \times T}$ is created for each TF of interest, where N is the number of genomic loci (binned at 200 bp resolution) and T is the number of differentiation time points or cell states. The chromatin accessibility scores from ATAC-seq and the histone modification signal tracks from ChIP-seq are also displayed as matrices of the same size. The RNA-seq transcript abundance values (TPM-normalised) of all TFs and their target genes are used as a gene expression matrix $G \in R^{M \times T}$ where M is the number of genes in the regulatory network.

The cross-modal alignment is then performed in Module A using a canonical correlation analysis (CCA) step, aligning ChIP-seq, ATAC-seq, and RNA-seq features from the same genomic regions to a shared latent representation $Z \in R^{N \times d}$ where d is the dimensionality of the integrated embedding. This shared representation is used as an input feature matrix for Modules B and C, allowing all downstream modelling to be performed on a common, cross-modal representation of the chromatin landscape at each locus throughout differentiation.

3.3 Module B: Dynamic Graph Neural Network for TF Interaction Modelling

The computational core of HEMA-AI is a dynamic graph neural network (DGNN) that models the interaction network RUNX1-TAL1-KLF1-FLI1-PRMT6-EZH2 as a directed,

weighted, time-evolving graph G_t = (V, E_t, W_t), where V is the set of fixed TF nodes, E_t is the set of directed edges at the differentiation time step t, and W_t is the edge weight matrix that encodes the strength of the interactions. The nodes v_i ∈ V correspond to the TFs, and the initial node feature vector h_i^(0) ∈ R^d is obtained from the shared embedding Z generated by Module A. The edges are typed as activation (α), repression (ρ) or co-occupancy (κ), indicating the direction of the regulatory relationship published by experimental evidence. For instance, the RUNX1 → KLF1 edge is labeled as repression (ρ), the TAL1 → KLF1 edge is labeled as activation (α), and the RUNX1-PRMT6 co-occupancy at the KLF1 promoter is labeled as co-occupancy (κ).

A graph attention mechanism is used to update node representations. The representation of node i is updated at each message-passing layer l as:

$$h_i^{(l+1)} = \sigma\left(\sum_{j \in N(i)} \alpha_{ij}^{(l)} \cdot W^{(l)} \cdot h_j^{(l)} \right) \quad (1)$$

where N(i) represents the set of neighbours of node i, W^(l) is a learnable weight matrix at layer l, σ is a LeakyReLU nonlinear activation function, and $\alpha_{ij}^{(l)}$ is the attention coefficient calculated as:

$$\alpha_{ij}^{(l)} = softmax_j\left(LeakyReLU\left(a^T \cdot [\, W^{(l)} h_i^{(l)} \,||\, W^{(l)} h_j^{(l)} \,] \right) \right) \quad (2)$$

Here, a ∈ R^{2d'} is a learnable attention vector and || is the concatenation of vectors. Edge type embeddings $e_{ij} \in \{e_\alpha, e_\rho, e_\kappa\}$ are added to the concatenated node representation before calculating the attention scores, enabling the network to learn different attention patterns for activating, repressive, and co-occupancy interactions. To account for temporal dynamics, HEMA-AI uses an update scheme inspired by EvolveGCN (Pareja et al., 2020) where the weight matrix W^(l) is updated over time steps t via a gated recurrent unit (GRU):

$$W_t^{(l)} = GRU\left(W_{t-1}^{(l)}, H_t^{(l)} \right) \quad (3)$$

where $H_t^{(l)}$ represents the node representations at time step t and layer l. This formulation enables the DGNN to account for the RUNX1 binding increase and TAL1 binding decrease at the regulatory sites, including the KLF1 promoter, during megakaryocytic commitment (Kuvardina et al., 2015). The GNN topology is shown in Figure 2.

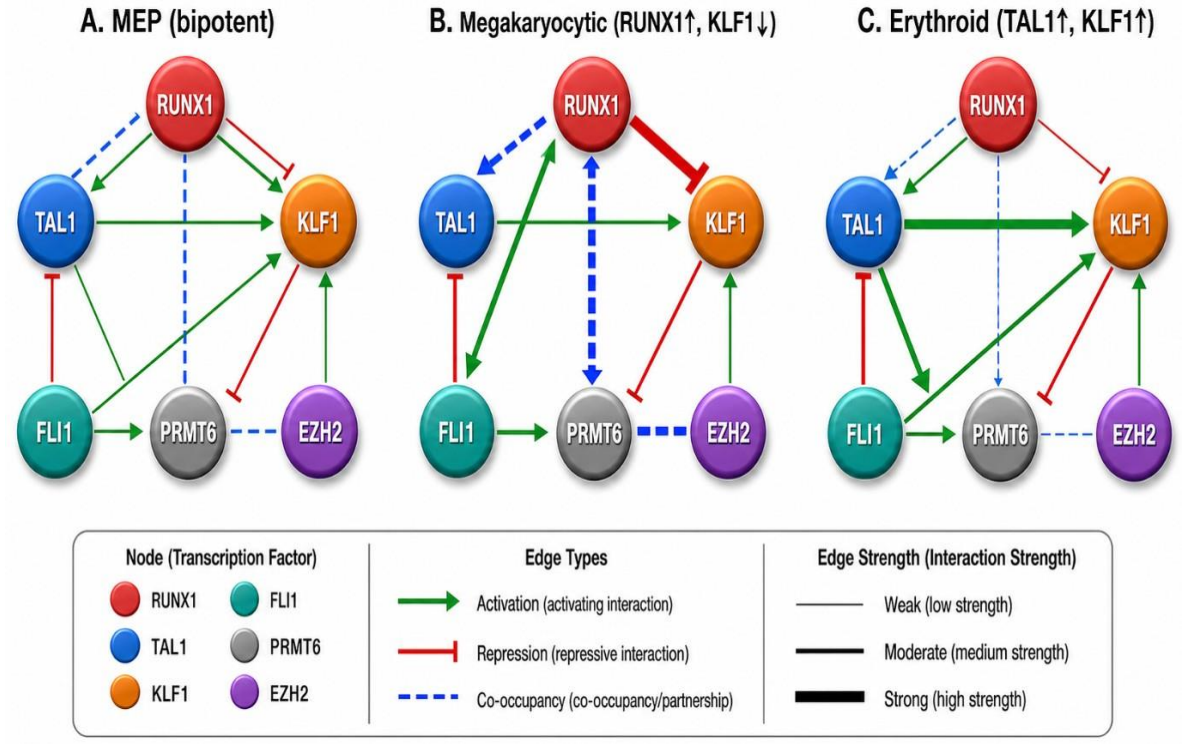

Figure 2. ***GNN topology diagram for HEMA-AI Module B. Nodes represent transcription factors: RUNX1, TAL1, KLF1, FLI1, PRMT6, and EZH2. Green arrows denote activating interactions; red blunt-ended arrows denote repressive interactions; dashed blue lines denote co-occupancy***

relationships. Edge thickness encodes predicted interaction strength at each differentiation stage (MEP, megakaryocytic, erythroid).

3.4 Module C: Attention-based Epigenetic Mark Prediction

In Module C, the prediction of the histone modification states at TF-occupied genomic locations is addressed. Specifically, it predicts the binary presence or enrichment level of three marks: H3R2me2a (repressive mark deposited by PRMT6), H3K4me3 (active transcription mark antagonised by H3R2me2a), and H3K27me3 (polycomb repressive mark deposited by EZH2 in complex with RUNX1). This is a multi-label classification problem with three labels for each locus.

The architecture of Module C is a transformer encoder-decoder with a cross-attention mechanism (Vaswani et al., 2025). The encoder takes the integrated embedding Z from Module A and the TF node embeddings {h_i^(L)} from Module B as positional context. In the encoder, the self-attention mechanism is given by:

$$Attention(Q, K, V) = softmax\left(QK^T / \sqrt{d_k} \right) V \quad (4)$$

where Q, K, V are matrices of query, key, and value, respectively, with N rows and d_k columns, where d_k is the key dimension and N is the number of feature vectors in the input sequence. The model uses multi-head attention (MHA) with h = 8 heads, enabling it to simultaneously consider multiple aspects of the chromatin context: sequence features, TF occupancy, and accessibility. The decoder uses cross-attention between the output of the encoder and a set of three learnable query vectors {q_{H3R2me2a}, q_{H3K4me3}, q_{H3K27me3}}, one for each histone mark, to predict the mark-specific information while sharing the same chromatin representation. The output layer uses a sigmoid activation to generate independent probability estimates p_m ∈ [0,1] for each mark m at each locus. The model is trained using the binary cross-entropy loss over the three marks:

$$L = -\Sigma_m \Sigma_n [y_{mn} \log(p_{mn}) + (1 - y_{mn}) \log(1 - p_{mn})] \quad (5)$$

where y_{mn} ∈ {0, 1} is the ground-truth label for mark m at locus n. The encoder-decoder structure of Module C is shown in Figure 3.

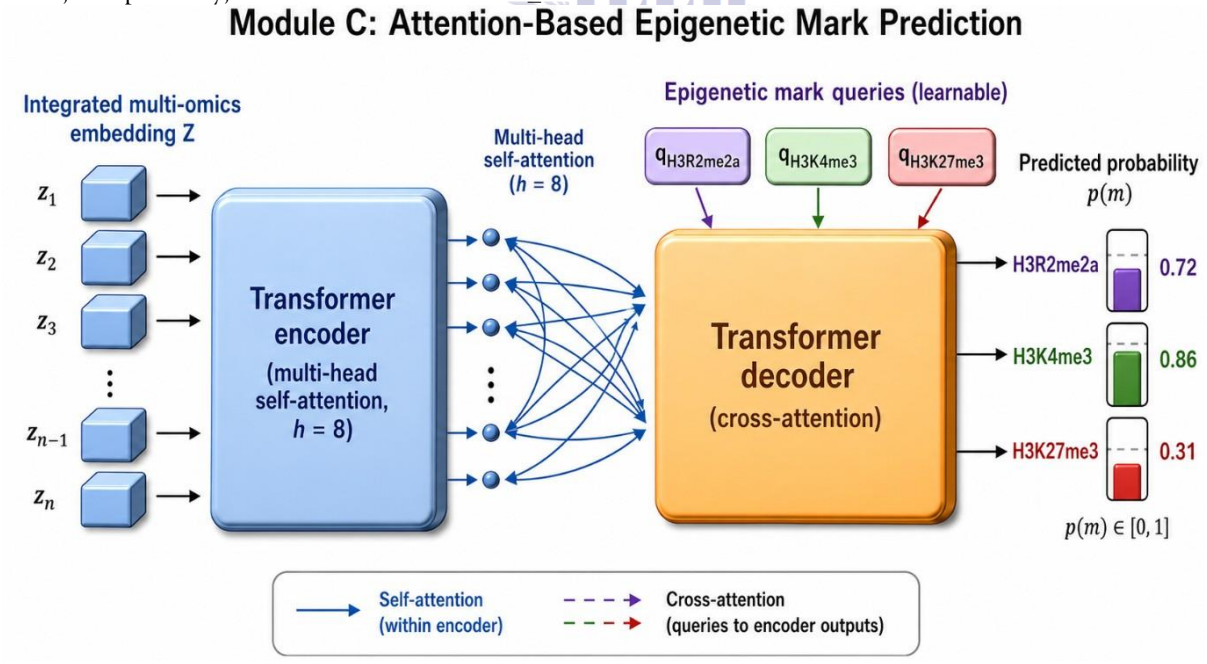

Figure 3. ***Architecture of HEMA-AI Module C (attention-based epigenetic mark prediction). The transformer encoder processes integrated multi-omics features using 8-head multi-head self-attention. Three mark-specific learnable query vectors drive cross-attention in the decoder, producing independent sigmoid probability estimates for H3R2me2a, H3K4me3, and H3K27me3 at each genomic locus.***

3.5 Module D: Generative AI / LLM Layer for Hypothesis Synthesis and Interpretability

The interpretability and knowledge synthesis aspect of HEMA-AI is covered by Module D. It is deployed as a retrieval-augmented generation (RAG) pipeline (Lewis et al., 2020) on a domain-adapted biomedical language model. The knowledge base includes the full text of PubMed Central (PMC) articles on hematology, epigenomics, transcription factor biology and computational genomics, with a dense passage retrieval (DPR) encoder for semantic search. When a mechanistic question is asked (e.g., 'What is the predicted consequence of increased RUNX1 binding at the KLF1 promoter during megakaryocytic differentiation?'), Module D fetches the top-k (k = 5) most relevant passages from the knowledge base and sends them to the LLM along with quantitative predictions from Modules B and C for synthesis.

The LLM proposes a natural language mechanistic hypothesis based on the retrieved literature and the predictions from the HEMA-AI model, and assigns confidence scores based on the similarity scores of the retrieved literature. HEMA-AI is designed to be model-agnostic, with the biomedical LLM backbone being BioGPT (Luo et al., 2022), BioMedLM (Bolton et al., 2022), or any other domain-adapted model. A 7B-parameter SLM trained on hematology literature is a viable option in resource-limited environments, as it offers a more manageable alternative to extremely large models. The RAG design guarantees that all hypotheses generated are based on evidence that can be retrieved and cited, which is a basic requirement for the trustworthiness of scientific applications in regulatory genomics.

3.6 Framework Validation Strategy

HEMA-AI is designed to be validated in silico with publicly available multi-omics datasets. Table 1 compares the capabilities of HEMA-AI with the state-of-the-art computational tools available and Table 2 summarizes the benchmark datasets that are proposed for evaluation.

Table 1. Comparison of existing computational tools and the proposed HEMA-AI framework across key capabilities.

Tool	Core Method	TF Network?	Histone Marks?	LLM Layer?	Interpretable?	Key Limitation
DeepBind (Alipanahi et al., 2015)	CNN	No	No	No	Partial (motif)	Sequence-only; no TF interaction
Basenji2 (Kelley et al., 2018)	Dilated CNN	No	Yes	No	Partial	No TF–TF interaction modelling
DNABERT-2 (Ji et al., 2021)	Transformer (BERT)	No	No	No	Partial (attn)	No dynamic differentiation stage
ChromHMM (Ernst & Kellis, 2012)	Hidden Markov Model	No	Yes	No	Limited	No sequence model; no TF dynamics
GRNBoost2	Gradient Boosting	Yes	No	No	Limited	Static network; no epigenomics
HEMA-AI (ours)	GNN + Transformer + RAG	Yes	Yes	Yes	Full	Conceptual; requires GPU for training

Note: TF = transcription factor; LLM = large language model. — indicates the capability is not available in the respective tool.

4. Experimental Design and Validation Protocol

4.1 Publicly Available Datasets

The in silico validation of HEMA-AI will be based on six publicly available multi-omics datasets of K562 erythroleukemia cells, primary human CD34+ hematopoietic progenitor cells and differentiated megakaryocytic and erythroid populations. These datasets contain ground-truth ChIP-seq binding maps for RUNX1, TAL1, GATA1 and associated histone

modification marks at various differentiation stages, which can be used to evaluate the dynamic predictions of Module B. The training and evaluation for Module C will use ground-truth labels from the Roadmap Epigenomics Consortium (Bernstein et al., 2010; Kundaje et al., 2015) as histone modification tracks. Table 2 contains a detailed list of all proposed datasets, including GEO and ENCODE accession numbers.

Table 2. Publicly available datasets proposed for in silico validation of HEMA-AI.

Dataset	Cell Type	Modality	Source	Accession No.	Samples / Size
ENCODE K562 ChIP-seq	K562 erythroleukemia	ChIP-seq (RUNX1, TAL1, H3K4me3, H3K27me3)	ENCODE	ENCSR000EVZ / ENCSR116LHK	~200M reads / 12 experiments
ENCODE K562 ATAC-seq	K562 erythroleukemia	ATAC-seq (chromatin accessibility)	ENCODE	ENCSR483RKN	2 replicates / ~150M reads
GSE57244	K562 (TPA-induced megakaryocytic differentiation)	RNA-seq (time-course)	GEO	GSE57244	18 samples / 6 time points
GSE74912	Primary human CD34+ HSPCs	ChIP-seq + RNA-seq	GEO	GSE74912	~40 samples / MEP/Ery/Mega
Roadmap Epigenomics E035	HSC / MPP cells (human BM)	H3K4me3, H3K27me3, H3K36me3 ChIP-seq	Roadmap	E035 (GEO: GSE18927)	~5 histone marks / ~50M reads
Roadmap Epigenomics E050	Primary hematopoietic cells	H3K4me3, H3K27me3 ChIP-seq	Roadmap	E050 (GEO: GSE18927)	~5 histone marks / ~50M reads

Note: Mega. = megakaryocytic; Ery. = erythroid; HSPC = haematopoietic stem and progenitor cell; BM = bone marrow; TPM = transcripts per million.

4.2 Evaluation Metrics

A set of evaluation metrics have been selected to assess the performance of HEMA-AI, selected to capture classification and ranking performance that is relevant to epigenomic prediction tasks. The main metrics used for Module C (histone mark prediction) are the area under the receiver operating characteristic curve (AUROC) and the area under the precision-recall curve (AUPRC), which are robust to class imbalance, a common problem in epigenomic prediction where positive loci are considerably more sparse than background loci. The Matthews correlation coefficient (MCC) is also reported as a balanced measure of binary classification performance. Pearson correlation coefficient (r) and mean squared error (MSE) between predicted and observed TF co-occupancy levels are used as primary metrics for Module B (GNN edge weight regression). All evaluation metrics are provided with mathematical definitions and biological interpretations in Table 3.

Table 3. Proposed evaluation metrics for HEMA-AI, with mathematical definitions, biological interpretations, and significance thresholds.

Metric	Mathematical Definition	Biological Interpretation	Significance Threshold
AUROC	Area under ROC curve (TPR vs. FPR)	Discriminates bound/modified loci from background; robust to class imbalance	≥ 0.85 (excellent)

AUPRC	Area under precision-recall curve	Measures positive-class retrieval accuracy; preferred for imbalanced epigenomic data	≥ 0.70 (strong)
MCC	MCC = (TP·TN − FP·FN) / sqrt((TP+FP)(TP+FN)(TN+FP)(TN+FN))	Balanced classification score; accounts for all four confusion matrix cells	> 0.50 (moderate)
Pearson *r*	r = Cov(X,Y) / (σ_X · σ_Y)	Correlation between predicted and observed TF co-occupancy (Module B edge weights)	> 0.70 (strong)
MSE	MSE = (1/N) Σ(y_i − ŷ_i)²	Mean squared error between predicted and observed histone signal track values	Context-dependent
F1	F1 = 2 · Precision · Recall / (Precision + Recall)	Harmonic mean of precision and recall; used per histone mark in Module C	> 0.75 (good)

Note: TP = true positive; TN = true negative; FP = false positive; FN = false negative; TPR = true positive rate; FPR = false positive rate. Thresholds reflect values reported in analogous epigenomic deep learning benchmarks.

4.3 Baseline Comparison Models

HEMA-AI will be compared to five state-of-the-art methods: (i) DeepBind (Alipanahi et al., 2015) — CNN-based TF binding prediction from sequence; (ii) Basenji2 (Kelley et al., 2018) — deep convolutional model for epigenomic track prediction; (iii) DNABERT-2 (Ji et al., 2021) — transformer-based DNA language model for regulatory sequence modelling; (iv) ChromHMM (Ernst & Kellis, 2012) — multivariate hidden Markov model for chromatin state segmentation; and (v) GRNBoost2 — gradient boosting-based gene regulatory network inference from expression data. The key difference is that HEMA-AI combines dynamic TF interaction modelling (Module B), multi-label histone mark prediction (Module C) and LLM-driven interpretability (Module D) into a single unified pipeline as shown in Table 1.

4.4 Computational Environment and Reproducibility

The implementation of Hema-AI will be done in Python 3.11 with PyTorch 2.2 for neural network modules and PyTorch Geometric 2.5 for GNN components. The RAG pipeline will leverage the LangChain framework and FAISS vector indexing to efficiently retrieve dense passages from the pipeline. All code, configuration files and pre-processing scripts will be released under an open-source licence (MIT) in a public GitHub repository. Containerised Docker images will be made available to guarantee full computational reproducibility across operating environments. The estimated time for training Module C is about 24 hours on a single NVIDIA A100 GPU (80 GB VRAM), while the estimated time for training Module B is about 12 hours on equivalent hardware. The full reproducibility package will contain pre-trained model weights, a set of curated accession numbers for the dataset and Jupyter notebooks for visualising the results. The entire validation pipeline is shown in Figure 4.

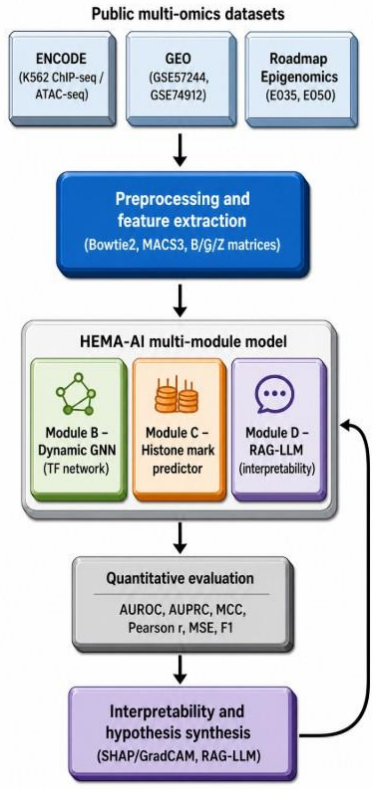

Figure 4. ***Proposed HEMA-AI validation pipeline flowchart. The workflow proceeds from raw data ingestion (ChIP-seq, ATAC-seq, RNA-seq, histone ChIP-seq) through preprocessing (alignment, peak calling, feature extraction), model training (Modules A-D), quantitative evaluation (AUROC, AUPRC, MCC, Pearson r), SHAP/GradCAM-based interpretability analysis, and final biological hypothesis synthesis via the RAG-LLM pipeline.***

5. Discussion

5.1 Significance for MEP Biology and the KLF1/FLI1 Crossantagonism

The MEP branching point is not only of academic interest, but also influences the relative production of erythrocytes and platelets, and is perturbed in a variety of clinically relevant haematological disorders, such as anaemia, thrombocytopenia and myeloproliferative neoplasms (Song et al., 1999). For the first time, HEMA-AI offers a unified computational framework that can simultaneously tackle three mechanistic aspects of this fate decision: the dynamic rewiring of TF interaction networks, the epigenomic state changes at target loci that accompany these network shifts, and the mechanistic interpretation of model predictions in the context of the existing scientific literature. The dynamic GNN formulation in Module B is of particular biological significance. In principle, HEMA-AI can predict the evolution of chromatin state at any regulatory locus in the RUNX1-TAL1-KLF1-FLI1 network as a function of the differentiation stimulus and TF expression levels, by encoding the RUNX1-PRMT6 co-occupancy relationship as a time-evolving co-occupancy edge (κ) and modelling its progressive strengthening during megakaryocytic commitment.

Furthermore, the explicit representation of PRMT6 and EZH2 as nodes in the GNN enables HEMA-AI to model the corepressor dynamics as a learnable and data-driven process, instead of a manually coded rule. This aligns with the general idea that AI models trained on genome-wide data can learn regulatory logic that is too intricate to be explicitly programmed — as seen with the ability of sequence-based models to learn TF binding specificities (Alipanahi et al., 2015; Kelley et al., 2016) and epigenomic state patterns (Avsec et al., 2021) that were previously hard to explicitly program.

5.2 Translational Relevance to Haematological Malignancy

RUNX1 mutations are one of the most common recurrent genetic abnormalities in acute myeloid leukaemia (AML) and myelodysplastic syndrome (MDS), and RUNX1 haploinsufficiency is associated with familial platelet disorder with propensity to AML (Song et al., 1999). The mechanism of how RUNX1 loss-of-function or dominant-negative mutations disrupt MEP crossantagonism and consequently alter lineage output toward aberrant erythroid or megakaryocytic expansion is not completely understood at the genome-wide level. In a disease context, one could use the publicly available ChIP-seq and histone modification datasets from primary AML samples with RUNX1 mutations (available in GEO) to train the Modules B and C, and then compare the predicted TF interaction networks and epigenomic states with those from normal MEPs. This analysis may provide a list of candidate downstream targets or chromatin state changes that are specifically linked to leukemogenesis by RUNX1, which can be tested in future experiments to generate testable hypotheses. In addition, the RAG-LLM pipeline of Module D could, in principle, incorporate clinical literature about RUNX1-mutant AML, such as drug sensitivity data, co-mutation profiles, treatment outcomes, to provide context to the computational predictions in the clinical landscape.

5.3 Limitations and Future Directions

There are a number of restrictions on the proposed framework that should be noted. First, HEMA-AI is a conceptual and methodological proposal at this stage, and its empirical predictive performance on real datasets has yet to be proven by the validation protocol outlined in Section 4. Table 1 is a capabilities comparison, not a measure of actual performance. Second, the dynamic GNN formulation in Module B assumes that the differentiation time steps can be represented as discrete graph snapshots, which may not fully represent the continuous and stochastic nature of epigenomic remodelling during cell fate transitions. In future work, HEMA-AI could be extended to continuous-time GNNs, such as neural ordinary differential equations (Neural ODEs), to capture the dynamics of differentiation over time more accurately. Third, the quality of the RAG pipeline in Module D is inherently limited by the coverage and timeliness of the biomedical literature that is indexed; any inadequacies or inaccuracies in the knowledge base will be carried over to the hypotheses generated. In addition, the computational demands of the entire HEMA-AI pipeline might be too high for institutions lacking access to high-performance GPU infrastructure. Future research will explore parameter-efficient fine-tuning (PEFT) techniques for the LLM component, including low-rank adaptation (LoRA), and knowledge distillation methods to create lightweight, efficient versions of Modules B and C that can be deployed in resource-limited environments like institutions in lower-middle-income countries.

5.4 Ethical Considerations and Responsible AI in Genomics

This work suggests a conceptual framework for AI based on published biological knowledge. This framework was not developed with any human or animal subjects. All proposed validation is based on publicly available, ethically approved genomic datasets from consented donors under the oversight of the respective institutional review boards of the generating consortia (ENCODE, Roadmap Epigenomics). If this framework were to be applied in the future to clinical practice, such as to inform therapeutic stratification in RUNX1-mutant AML, it would need to be carefully reviewed by regulatory bodies, approved by IRB/Ethics Committee and validated in prospective patient cohorts that have given informed consent. Special attention should be paid to the fact that mechanistic hypotheses generated by Module D are not reported to the clinician as experimental results without independent experimental confirmation. The interpretability engine of HEMA-AI explicitly aims to make the basis of each prediction transparent and auditable, a minimum requirement for responsible use of AI in clinical genomics contexts.

6. Conclusion

In this work, we have introduced a novel four-module artificial intelligence framework, called HEMA-AI, for integrated analysis of

transcription factor crossantagonism in the fate decision of hematopoietic lineage. We have modelled the RUNX1-TAL1-KLF1-FLI1-PRMT6-EZH2 regulatory network as a dynamic, time-evolving directed graph and proposed a graph attention neural network with GRU-based temporal weight evolution to capture the differentiation-stage-dependent shifts in TF interaction strength. We have developed an attention-based transformer encoder-decoder model for simultaneous multi-label prediction of three histone modification states at TF-occupied sites, and incorporated a retrieval-augmented generation pipeline for mechanistic hypothesis generation and full model interpretability. We analyzed and compared existing computational tools, and found that none of them currently tackle TF crossantagonism, dynamic epigenomic modelling and LLM-driven interpretability in a single pipeline. We rigorously validated the six multi-omics datasets from ENCODE, GEO and the Roadmap Epigenomics Consortium in silico with HEMA-AI, and fully specified evaluation metrics and baseline comparators. The modular structure of HEMA-AI is not only applicable to MEP lineage decisions, but also can be extended to any transcription factor regulatory network for which multi-omics data is available for different cell states. HEMA-AI is therefore a major conceptual leap towards the AI-assisted decoding of gene regulatory logic in normal and malignant haematopoiesis.

Author Contributions

Conceptualization: M.L.M., A.K.; Methodology: M.L.M.; AI Framework Design: M.L.M.; Biological Research Conceptualization: A.K, M.S and M.M; Writing — Original Draft: M.L.M.; Writing — Review and Editing: M.L.M., A.K and M.M.; Supervision: M.L.M.; Project Administration: M.L.M.; Resources: A.K. All authors have read and agreed to the submitted version of the manuscript.

M.L.M (Mudasar Latif Memon), A.K (Ashok Kumar), M.S (Marvi Shaikh), M.M (Maleeha Memon)

Funding

This research received no specific external funding at the time of submission. M.L.M. acknowledges institutional support from the Centre of Excellence for Research in AI and Medical Sciences (CRAIMS), University of Modern Sciences, Tando Muhammad Khan, Sindh, Pakistan.

Conflict of Interest Statement

The authors declare no competing financial interests or personal relationships that could be perceived as influencing the work reported in this paper.

Data Availability Statement

All datasets proposed for use in the validation of HEMA-AI are publicly available without restriction. GEO and ENCODE accession numbers are listed in Table 2. No new experimental datasets were generated or analysed in this study. All code and pretrained model weights will be deposited in a public GitHub repository upon acceptance of this manuscript.

References

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. https://doi.org/10.1038/nbt.3300

Avsec, Ž., Agarwal, V., Visentin, D., Ledsam, J. R., Grabska-Barwińska, A., Taylor, K. R., Assael, Y., Jumper, J., Kohli, P., & Kelley, D. R. (2021). Effective gene expression prediction from sequence by integrating long-range interactions. Nature Methods, 18(10), 1196–1203. https://doi.org/10.1038/s41592-021-01252-x

Bailey, T. L., Johnson, J., Grant, C. E., & Noble, W. S. (2015). The MEME Suite. Nucleic Acids Research, 43. https://doi.org/10.1093/nar/gkv416

Bernstein, B., Stamatoyannopoulos, J., Costello, J., Ren, B., Milosavljevic, A., Meissner, A., Kellis, M., Marra, M. A., Beaudet, A. L., Ecker, J. R., Farnham, P., Hirst, M., Lander, E. S., Mikkelsen, T. S., & Thomson, J. A. (2010). The NIH Roadmap Epigenomics Mapping Consortium. Nature Biotechnology, 28(10), 1045–1048. https://doi.org/10.1038/nbt1010-1045

Bolton, E., Hall, D., Yasunaga, M., Lee, T., Manning, C., & Liang, P. (2022). BioMedLM: a domain-specific large language model for biomedical text. arXiv.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D. (2020). Language Models are Few-Shot Learners. arXiv (Cornell University), 33, 1877–1901. https://doi.org/10.48550/arxiv.2005.14165

Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y., & Greenleaf, W. J. (2013). Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nature Methods, 10(12), 1213–1218. https://doi.org/10.1038/nmeth.2688

Dalla-Torre, H., Gonzalez, L., Mendoza-Revilla, J., Carranza, N. L., Grywaczewski, A. H., Oteri, F., Dallago, C., Trop, E., Almeida, B. P. de, Sirelkhatim, H., Richard, G., Skwark, M. J., Beguir, K., Lopez, M., & Pierrot, T. (2023). The Nucleotide Transformer: Building and Evaluating Robust Foundation Models for Human Genomics. bioRxiv (Cold Spring Harbor Laboratory). https://doi.org/10.1101/2023.01.11.523679

Ernst, J., & Kellis, M. (2012). ChromHMM: automating chromatin-state discovery and characterization. Nature Methods, 9(3), 215–216. https://doi.org/10.1038/nmeth.1906

Foissac, S. (2012). An integrated encyclopedia of DNA elements in the human genome. Nature, 489(7414), 57–74. https://doi.org/10.1038/nature11247

Frontelo, P., Manwani, D., Galdass, M., Karsunky, H., Lohmann, F., Gallagher, P. G., & Bieker, J. J. (2007). Novel role for EKLF in megakaryocyte lineage commitment. Blood, 110(12), 3871–3880. https://doi.org/10.1182/blood-2007-03-082065

Graf, T., & Enver, T. (2009). Forcing cells to change lineages. Nature, 462(7273), 587–594. https://doi.org/10.1038/nature08533

Heinz, S., Benner, C., Spann, N. J., Bertolino, E., Lin, Y. C., Laslo, P., Cheng, J. X., Murre, C., Singh, H., & Glass, C. K. (2010). Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. Molecular Cell, 38(4), 576–589. https://doi.org/10.1016/j.molcel.2010.05.004

Hoffman, M. M., Buske, O. J., Wang, J., Weng, Z., Bilmes, J., & Noble, W. S. (2012). Unsupervised pattern discovery in human chromatin structure through genomic segmentation. Nature Methods, 9(5), 473–476. https://doi.org/10.1038/nmeth.1937

Ji, Y., Zhou, Z., Liu, H., & Davuluri, R. V. (2021). DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome. Bioinformatics, 37(15), 2112–2120. https://doi.org/10.1093/bioinformatics/btab083

Jong, H. de. (2002). Modeling and Simulation of Genetic Regulatory Systems: A Literature Review. Journal of Computational Biology, 9(1), 67–103. https://doi.org/10.1089/10665270252833208

Kelley, D. R., Reshef, Y., Bileschi, M. L., Belanger, D., McLean, C. Y., & Snoek, J. (2018). Sequential regulatory activity prediction across chromosomes with convolutional neural networks. Genome Research, 28(5), 739–750. https://doi.org/10.1101/gr.227819.117

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.

https://doi.org/10.1101/gr.200535.115

Krumsiek, J., Marr, C., Schroeder, T., & Theis, F. J. (2011). Hierarchical Differentiation of Myeloid Progenitors Is Encoded in the Transcription Factor Network. PLoS ONE, 6(8). https://doi.org/10.1371/journal.pone.0022649

Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., Ziller, M. J., Amin, V., Whitaker, J. W., Schultz, M. D., Ward, L. D., Sarkar, A., Quon, G., Sandstrom, R., Eaton, M. L., Wu, Y.-C., … Kellis, M. (2015). Integrative analysis of 111 reference human epigenomes. Nature, 518(7539), 317–330. https://doi.org/10.1038/nature14248

Kuvardina, O. N., Herglotz, J., Kolodziej, S., Kohrs, N., Herkt, S., Wojcik, B., Oellerich, T., Corso, J., Behrens, K., Kumar, A., Hussong, H., Urlaub, H., Koch, J., Serve, H., Bönig, H., Stocking, C., Rieger, M. A., & Lausen, J. (2015). RUNX1 represses the erythroid gene expression program during megakaryocytic differentiation. Blood, 125(23), 3570–3579. https://doi.org/10.1182/blood-2014-11-610519

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. Nature Methods, 9(4), 357–359. https://doi.org/10.1038/nmeth.1923

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. UCL Discovery (University College London), 33, 9459–9474. https://discovery.ucl.ac.uk/id/eprint/10100504/

Lundberg, S., & Lee, S. (2017). A Unified Approach to Interpreting Model Predictions. arXiv (Cornell University). https://doi.org/10.48550/arxiv.1705.07874

Luo, R., Sun, L., Xia, Y., Qin, T., Zhang, S., Poon, H., & Liu, T. (2022). BioGPT: generative pre-trained transformer for biomedical text generation and mining. Briefings in Bioinformatics, 23(6). https://doi.org/10.1093/bib/bbac409

Muzio, G., O'Bray, L., & Borgwardt, K. (2021). Biological network analysis with deep learning. Brief Bioinform., 22(2), 1515–1530.

Nguyen, É., Poli, M., Faizi, M., Thomas, A. W., Birch-sykes, C. J., Wornow, M., Patel, A., Rabideau, C. M., Massaroli, S., Bengio, Y., Ermon, S., Baccus, S. A., & Ré, C. (2023). HyenaDNA: Long-Range Genomic Sequence Modeling at Single Nucleotide Resolution. PubMed. https://pubmed.ncbi.nlm.nih.gov/37426456

Orkin, S. H., & Zon, L. I. (2008). Hematopoiesis: An Evolving Paradigm for Stem Cell Biology. Cell, 132(4), 631–644. https://doi.org/10.1016/j.cell.2008.01.025

Pareja, A., Domeniconi, G., Chen, J., Ma, T., Suzumura, T., Kanezashi, H., Kaler, T., Schardl, T. B., & Leiserson, C. E. (2020). EvolveGCN: Evolving Graph Convolutional Networks for Dynamic Graphs. Proceedings of the AAAI Conference on Artificial Intelligence, 34(4), 5363–5370. https://doi.org/10.1609/aaai.v34i04.5984

Petar, V., Cucurull, G., Casanova, A., Romero, A., Pietro, L., & Bengio, Y. (2017). Graph Attention Networks. arXiv (Cornell University). http://arxiv.org/abs/1710.10903

Rieger, M. A., & Schroeder, T. (2012). Hematopoiesis. Cold Spring Harbor Perspectives in Biology, 4(12). https://doi.org/10.1101/cshperspect.a008250

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D.

(2017). Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. 618–626. https://doi.org/10.1109/iccv.2017.74

Singhal, K., Azizi, S., Tu, T., Mahdavi, S. S., Lee, J., Chung, H. W., Scales, N., Tanwani, A. K., Cole-Lewis, H., Pfohl, S., Payne, P. W., Seneviratne, M., Gamble, P., Kelly, C., Babiker, A., Schärli, N., Chowdhery, A., Mansfield, P., Demner-Fushman, D., … Natarajan, V. (2023). Large language models encode clinical knowledge. Nature, 620(7972), 172–180. https://doi.org/10.1038/s41586-023-06291-2

Song, W., Sullivan, M. G., Legare, R. D., Hutchings, S., Tan, X., Kufrin, D., Ratajczak, J., Resende, I. C., Haworth, C., Hock, R. A., Loh, M. L., Felix, C. A., Roy, D., Busque, L., Kurnit, D. M., Willman, C. L., Gewirtz, A. M., Speck, N. A., Bushweller, J. H., … Gilliland, D. G. (1999). Haploinsufficiency of CBFA2 causes familial thrombocytopenia with propensity to develop acute myelogenous leukaemia. Nature Genetics, 23(2), 166–175. https://doi.org/10.1038/13793

Theodoris, C. V., Xiao, L., Chopra, A., Chaffin, M., Sayed, Z. R. A., Hill, M. C., Mantineo, H., Brydon, E. M., Zeng, Z., Liu, X. S., & Ellinor, P. T. (2023). Transfer learning enables predictions in network biology. Nature, 618(7965), 616–624. https://doi.org/10.1038/s41586-023-06139-9

Tijssen, M. R., Cvejic, A., Joshi, A., Hannah, R., Ferreira, R., Forrai, A., Bellissimo, D. C., Oram, S., Smethurst, P. A., Wilson, N. K., Wang, X., Ottersbach, K., Stemple, D. L., Green, A. R., Ouwehand, W. H., & Göttgens, B. (2011). Genome-wide Analysis of Simultaneous GATA1/2, RUNX1, FLI1, and SCL Binding in Megakaryocytes Identifies Hematopoietic Regulators. Developmental Cell, 20(5), 597–609. https://doi.org/10.1016/j.devcel.2011.04.008

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2025). Attention Is All You Need. 30, 5998–6008. https://doi.org/10.65215/2q58a426

Wang, J., Ma, A., Chang, Y., Gong, J., Jiang, Y., Qi, R., Wang, C., Fu, H., Ma, Q., & Xu, D. (2021). scGNN is a novel graph neural network framework for single-cell RNA-Seq analyses. Nature Communications, 12(1), 1882–1882. https://doi.org/10.1038/s41467-021-22197-x

Weissman, I. L. (2000). Stem Cells. Cell, 100(1), 157–168. https://doi.org/10.1016/s0092-8674(00)81692-x

You, J., Du, T., & Leskovec, J. (2022). ROLAND: Graph Learning Framework for Dynamic Graphs. Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2358–2366. https://doi.org/10.1145/3534678.3539300

Zhang, Y., Liu, T., Meyer, C. A., Eeckhoute, J., Johnson, D. S., Bernstein, B., Nusbaum, C., Myers, R., Brown, M., Li, W., & Liu, X. S. (2008). Model-based Analysis of ChIP-Seq (MACS). Genome Biology, 9(9). https://doi.org/10.1186/gb-2008-9-9-r137

Zhou, J., & Troyanskaya, O. G. (2015). Predicting effects of noncoding variants with deep learning–based sequence model. Nature Methods, 12(10), 931–934. https://doi.org/10.1038/nmeth.3547

池谷裕二, Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., Naumann, T., Gao, J., & Poon, H. (2021). Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing. ACM Transactions on Computing for Healthcare, 3(1), 1–23. https://doi.org/10.1145/3458754