Developing a Biophysically Grounded Deep Learning Model for Gene Expression Prediction

A thesis
Submitted towards the partial fullfilment of
BS-MS dual degree programme
by

Prarabdh Shivhare

Date:27/03/2025

under the guidance of

Dr. Rosa Martinez Corral and Dr. Lars Velten

Barcelona Collaboratorium for Modelling and Predictive Biology, Centre for Genomic Regulation, Barcelona

from May 2024 to Mar 2025

Indian Institute of Science Education and Research Pune

Declaration

I, hereby declare that the matter embodied in the report titled "Developing Biophysically Grounded Deep Learning Models to Investigate Non-Monotonic Responses" is the results of the investigations carried out by me at the "Barcelona Collaboratorium for Modelling and Predictive Biology and Centre for Genomic Regulation, Barcelona" from the period 01-08-2024 to 31-03-2025 under the supervision of Dr. Rosa Martinez Corral for mathematical and biophysical modelling and Dr. Lars Velten for deep learning based modelling and the same has not been submitted elsewhere for any other degree.

Prarabdh Shivhare
20201147
BS-MS
IISER Pune

Supervisor:
Dr. Rosa Martinez Corral
Independent Fellow
Barcelona Collaboratorium for Modelling and Predictive Biology

Date:
27/03/2025

Date:
27/03/2025

Certificate

This is to certify that this dissertation entitled "Developing Biophysically Grounded Deep Learning Models to Investigate Non-Monotonic Responses" submitted towards the partial fulfillment of the BS-MS degree at the Indian Institute of Science Education and Research, Pune represents original research carried out by "Prarabdh Shivhare" at "Barcelona Collaboratorium for Modelling and Predictive Biology and Centre for Genomic Regulation", under the supervision of "Dr. Rosa Martinez Corral" during academic year May 2024 to March 2025.

Supervisor:
Dr. Rosa Martinez
Corral
Independent Fellow
Barcelona
Collaboratorium for
Modelling and
Predictive Biology

Prarabdh Shivhare
20201147
BS-MS
IISER Pune

Date: 27/03/2025

1 Preface

During my time at IISER, I was drawn to a wide range of subjects. I often told people that I liked both physics and biology, among other things, but I would often hesitate to identify 'biophysics' as one of my interests. At that time, my exposure to biophysics was largely limited to molecular-level studies, whereas my curiosity leaned more toward systems and cellular-level modeling. Physics research often felt inaccessible as an undergraduate, while biology, on the other hand, seemed too messy to model rigorously.

That changed when I came across two formative textbooks: Physical Biology of the Cell by Phillips, Kondev, Theriot, and Garcia, and William Bialek's Biophysics: Searching for Principles. I feel incredibly fortunate that my thesis work now touches upon areas of research pursued by the very authors of these books, authors whose work significantly shaped my interests.

By 2023, I began to realize that while theoretical approaches drawing from the physical, mathematical, and information sciences would remain indispensable in biology, understanding complex biological systems would increasingly require data.I was looking for opportunities to fulfilll these interests and ones where I could employ myinterdisciplinary undergraduate training, and thankfully, one came along. Interpretable machine learning and high-throughput synthetic biology, I believe, will transform the way biological research is practiced, and I hope to be able to make meaningful contributions to this transformation.

2 Abstract

In the past 2 decades of literature dealing with modeling complex systems, there has been a balance, or rather, a tension between the predictive power and the interpretability of machine learning models using vast amounts of data. Biological complex systems are no different. The past decade has seen an astonishing increase in the amount of publicly available functional genomics data. While the adoption of deep learning techniques to determine the sequence patterns, syntax and grammar in DNA sequence elements that govern gene regulatory activity has been a natural consequence, most of these investigations have adopted a 'black box' approach, with model predictions that are hard to interpret mechanistically. Multiple attribution strategies, which seek to extract meaningful post-hoc interpretations from neural networks have been proposed for addressing this problem. However, there remains a substantial gap in the literature between the outputs of such post-hoc methods and fully mechanistic models, specifically in the context of gene regulation. This problem can be at least partially overcome by including some level of mechanistic detail in the internal structure of deep learning algorithms. This can enable us to better understand the predictions of the model to obtain mechanistic insight. Here, we use a cell-state specific Massively Paraellel Reporter Assay dataset from hemotopoeitic stem cells to model gene regulation using deep learning to predict transcription factor (TF) binding on DNA sequence employing cell-state specific Chip-Seq data and graph-based representations of markov processes to model effects of bound TFs on different rate-limiting steps in the transcriptional cycle. Our model assumptions are grounded in recent biophysical findings in literature.

3 Introduction

Understanding regulation of transcription has been one of the central aims of fundamental biological research. If classical genetics and proteins dominated research headlines in the early and mid 20th century respectively, gene regulation has been one of our major focuses ever since Jakob and Monod first elucidated the working of the lac operon in 1961 (Jacob F, Monod J. 1961 J. Mol. Biol), . This interest is unlikely to wither as progress in development and stem cell biology grows. This introduction begins by outlining the mechanistic basis of eukaryotic transcription—including the essential roles of chromatin accessibility, promoters, and transcription factors —then transitions to modeling transcriptional regulation via kinetic synergy and cofactor interactions, before examining how enhancers and the cis-regulatory code integrate genomic interactions to establish cell type–specific expression, and finally describes advances in synthetic enhancer design and deep learning methods for predicting regulatory activity .

3.1 Transcription iin Eukaryotes

During development, genetic information needs to be activated in a precise manner. This requires a highly controlled expression of genes, which to a large extent is regulated at the level of transcription. For transcription to occur, the chromatin needs to be accessible for the RNA polymerase II (RNA Pol II) (Knezetic Luse, 1986; Lorch et al., 1987). Promoters are the elements initiating transcription and active promoters are usually found in nucleosome-depleted regions (Schones et al., 2008). Only a small subset of Pol II-dependent promoters is active at any given time. Enhancers are other gene regulatory elements which can promote transcription, since TFs initially bind to enhancers and promoters in a sequence-specific manner and will help guiding the polymerase

to its target (Dynan Tjian, 1983). It is estimated, that a 200 base pair (bp) enhancer contains on average five transcription factor binding sites (TFBS), and that on average 6 enhancers are regulating a single gene (Vierstra et al., 2020). Most TFs bind free DNA, but some TFs have the ability to bind nucleosomal DNA, pioneering the access to chromatin (Soufi et al., 2015; Zaret Carroll, 2011). Eukaryotic transcription is a step-wise process. After TFs guide the DNA-dependent RNA polymerase Il closer to its target, the promoter sequence needs be recognized. Transcription initiation factors recognize conserved DNA sequence elements in the promoter and assemble the pre-initiation complex (PIC) (Haberle Stark, 2018). Key components of the PIC contain general class II initiation factors (Roeder, 1996), like the TATA box-binding protein TBP or TFIIB which is necessary for bridging promoter DNA and Pol Il. In the next step, the PIC is opening up the DNA due to DNA translocase XBP, a subunit of the TFIH complex (Egly Coin, 2011). After having access to the DNA the Mediator co-activator complex regulates Pol II initiation. The Mediator complex consists out of two core modules, the "head" contacts Pol II with initiation factors, whereas the "tail" binds activating TFs. The Mediator complex also helps facilitating the transition to the elongation phase by stimulating CDK7, which phosphorylates the C-terminal domain (Kornberg, 2005). Specific DNA sequences can lead to transcriptional pausing, which if not resolved can lead to arrest and termination (Landick, 2006). TFIIs help overcome this blockade in the arrested complex by inducing cleavage of the RNA and restarting transcription (Cheung Cramer, 2011). Other types of promoter-proximal pausing exists (Eick Borkamm, 1986). The paused complex gets stabilized by the factors DSIF and NELF.

Pol II pausing hass been shown to be imperative in various gene regulatory processes and pathologies, such as, stem cell differentiation and developmental biology, and up to 40 percent of coding genes undergo this regulatory mechanism during transcription. ren ma wang 2024CDK9, a subunit of the positive transcription elongation factor b, phosphorylates DSIF, NELF and the Pol II C-terminal domain. Upon phosphorylation, the formation of an active elongating complex is triggered. During elongation, the complex is then interacting with factors involved in co-transcriptional histone and chromatin modifications, as well as RNA processing of the nascent chain. Termination of transcription is the last step in the cycle, in which after recognizing the poly-adenylation sequence, a "Torpedo" nuclease complex cleaves off the RNA and the transcript is released. Taken together, transcription by RNA polymerases depends on multiple factors.

3.2 Transcription factors and assays for measuring TF-DNA interactions

Transcription factors are proteins crucially regulating the expression of genes. By binding to specific DNA sequences via their DNA-binding domain (DBD), they influence the RNA polymerase efficiency, positively or negatively. This interaction between TF and polymerase can be through direct binding or via indirect interactions, like the recruiting of cofactors. For closely related TFs, i.e. factors belonging to the same TF family, the DBDs prefer to bind to a very similar DNA sequence (Weirauch et al., 2014). Intrinsic disordered regions (IDRs), polypeptide segments without an inherent three dimensional structure, of the TFs can facilitate another level of binding and cooperativity with their surroundings. Since it is conceptually hard to study disordered structures, IDRs have been under-studied and overlooked for their interaction interface. Nonetheless, IDRs generate specificity and allow cooperativity and TFs are heavily enriched in them, as they offer a high interaction potential with diverse factors (Brodsky et al., 2020; Liu et al., 2006). TFs are essential in developmental processes, as well as in their response to environmental signals. So far, more than 1,600 TFs have been catalogued in humans with similar numbers in mouse using various experimental techniques. Transcription factors can directly influence the transcriptional cycle. For example, the binding of MYC promotes Pol I pause release (Rahl et al., 2010). An increased transcriptional burst can result from the frequent turnover of molecules by transient TF interactions (Pomp et al., 2024). Moreover, specific TFs have the ability to open compacted chromatin. Pioneer factors can target naive chromatin sites and bind nucleosomal DNA independently by DBD recognition and preceding the binding of other factors (Soufi et al., 2015; Zaret Caroll, 2011). In early embryonic development and during zygotic genome activation, pioneering TF activity is necessary for facilitating a change of the chromatin program (Schulz et al., 2015). While pioneering TFs facilitate chromatin accessibility during early development, the binding dynamics of TFs are influenced by other factors. Jolma et al., 2013 emphasize that the spatial arrangement of TF binding sites, particularly their orientation and spacing, also has a more prevalent role in TF-DNA interactions, offering additional insights into the regulation of gene expression. A follow up work from the same group of authors highlights, that cooperativity of TFs on DNA is achieved if the two TFs originate from diverse structural families. In addition, most TF pairs have a large overlap of their TEBS and substantially differ from the individual TF motif (Jolma et al., 2015). Most human TFs are only binding a small genomic subset of their potential sites based on motif preference and different subsets are bound in different cell types. Cell-specific binding of sites is predicted not only by the DNA sequence (co-occurrence of diverse motifs), but also by chromatin accessibility and DNA methylation status. (Gertz et al., 2013).

ChIP-Seq (Chromatin Immunoprecipitation followed by Sequencing) is a powerful technique used to map protein–DNA interactions in living cells, providing insights into transcription factor binding sites and the chromatin landscape across the genome. In a typical ChIP-Seq experiment, cells are first treated with a crosslink-

ing agent, usually formaldehyde, which covalently binds proteins to DNA. The chromatin is then fragmented by sonication or enzymatic digestion, breaking the DNA into manageable pieces. An antibody specific to a protein of interest—such as a transcription factor or a modified histone—is used to immunoprecipitate the protein–DNA complexes, enriching for the genomic regions bound by that protein. Following immunoprecipitation, the crosslinks are reversed to release the DNA, which is then purified and prepared into a sequencing library. (Nakato et al 2021) High-throughput sequencing generates millions of short reads that are subsequently aligned to a reference genome. The resulting data are analyzed to identify peaks, which represent regions with significant enrichment of binding events, indicating where the protein interacts with the genome. ChIP-Seq offers several advantages, including the ability to capture in vivo binding under physiological conditions and to provide a genome-wide view of binding patterns. However, the technique also faces challenges such as antibody specificity, resolution limits imposed by chromatin fragmentation, and the complexity of data analysis, which require robust computational pipelines to distinguish true binding sites from background noise. (Peter J. Park 2009)

Other assays for transcription factor binding complement ChIP-Seq by offering different advantages. For example, HT-SELEX (High-Throughput Systematic Evolution of Ligands by Exponential Enrichment) is an in vitro method that systematically profiles binding preferences and generates quantitative motifs. Additionally, techniques such as DAP-Seq (DNA Affinity Purification Sequencing) and protein-binding microarrays (PBMs) provide alternative platforms to characterize TF-DNA interactions on a genome-wide scale. (Inukai et al 2017)

3.3 Modelling Transcriptional Regulation

Modeling transcriptional regulation has evolved from traditional thermodynamic state ensemble models—originally developed for bacterial gene regulation—to approaches that incorporate multiple rate-limiting steps in the transcriptional cycle. Early models typically considered a single bottleneck (e.g., RNA polymerase recruitment) and were often employed to retrospectively fit experimental data (Bintu et al 2004).

A breif summary of the main ideas of these statistical mechanics based models is provided in the appendix. The probability of RNA polymerase (RNAP) binding to a promoter, p_{bound}, is evaluated by summing the Boltzmann weights over all possible microscopic states of P RNAP molecules distributed on the DNA, which comprises numerous non-specific sites (NNS) and a single promoter.

However, transcription, as outlined above, is recognized as a complex cycle in which several sequential steps, including the assembly of macromolecular complexes, pause release, polymerase elongation rate and chromatin modifications, are subject to regulation.

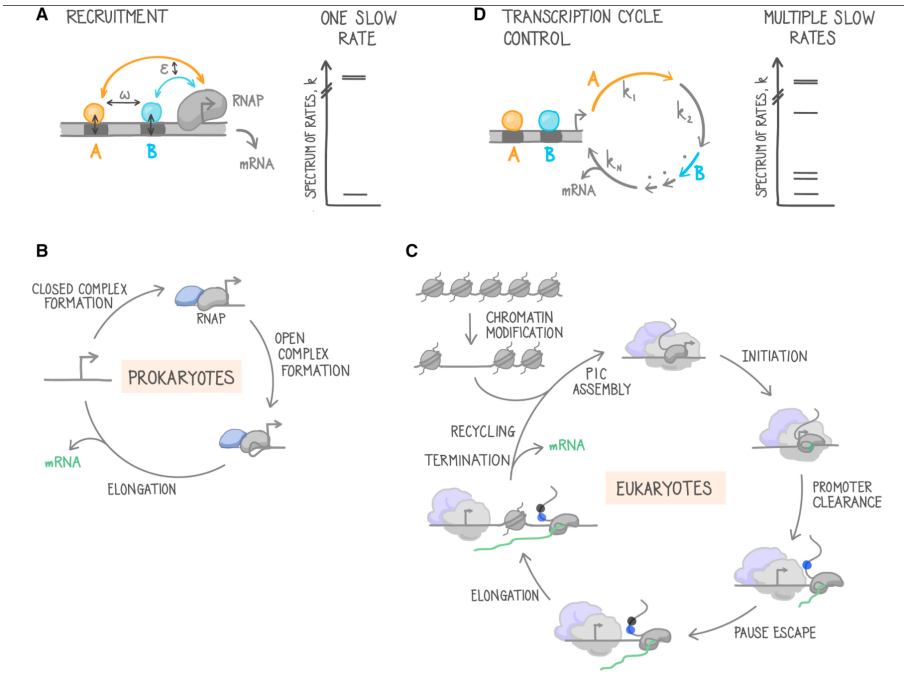

Figure 1: Figure from C Scholes et al. 2017. often modelling of transcriptional regulation in prokaryotes is limited tot= the regulated recruitment view, which does not yield accurate prredictions in eukaryotic contexts, since eukaryotic transcription is limited at multiple rate-limiting steps. A) w and e represent iinteractiions between TFs themselves and with RNAP. In eukaryotes, transcription factors communicate with the promoter via co-regulatory complexes—such as the mediator complex (light purple)—to assemble the preinitiation complex (PIC), which includes RNAP (dark gray) and general transcription factors (light gray). After the DNA is unwound to form the open complex, initiation occurs, and RNAP must then escape the promoter to enter elongation; this escape is accompanied by phosphorylation of its C-terminal domain (black and blue circles). In higher eukaryotes, RNAP typically pauses downstream of the promoter and must be actively released into the gene body. D) Gene expression can be combinatorially controlled through kinetic modulation of the transcription cycle, where transcription factors act on different slow steps. In a hypothetical transcription cycle, the rate diagram (right) shows that TF A acts to enhance the rate $k_1, while TF B influences a subsequent step. This dual regulation allows TF s to either accelerate or decelerate specific transiti$

For instance consider a kinetic model of transcription in which the cycle is represented by at least two steps, regulated by different transcription factors (TFs) influencing distinct steps to varying extents. When TF A predominantly modulates one step and TF B another, their combined effects can yield emergent regulatory behaviors such as logical AND gates—demonstrating that the system's information processing capability can arise from the differential kinetic contributions rather than from just mutually exclusive functions. This notion of kinetic synergy was indeed proposed first in a paper from 1993(Hershlog Johnson, 1993). translates these ideas into stochastic chemical kinetics by establishing a kinetic rate matrix for the transcription cycle and solving for its steady state.On the other hand, Martinez 2023 cell systems employ a synethic experiment tto identify if kinetic signatures like synergy can emerge from functional regulation of distinct steps in the cycle.

This minimal two-step model illustrates that the interplay between TFs on separate steps can generate a wide range of regulatory outcomes. It reveals that kinetic synergy is rooted in the coordinated regulation of multiple sequential processes.

3.4 The contribution of cofactors on transcription

Enhancer activation usually requires the combination of TFs, but cofactors play a crucial role in further specializing the spatiotemporal regulation in specific cell types. Cofactors are proteins that modulate the effects of TFs. This can mean that distinct combinations of TFs will only cooperate together and create specificity in a cofactor-dependent manner. Stampfel et al., 2015 show in Drosophila how cofactor dependency influences different classes of TFs. In this study, similar TFs were able to substitute each other ultimately enabling enhancer re-engineering (Stampfel et al., 2015). But what happens to enhancers and transcription factors if one erases cofactors? A rapid cofactor depletion assay in human HCTI16 cells shows that enhancer react differently to selective cofactor removal, which is related to their sequence, as well as to their chromatin properties (Neumayr et al., 2022). A brief case study on modelling co-factor interactions, following Janssens 2006 can be found in the appendix.

3.5 Enhancers

Enhancers are non-coding DNA elements that upregulate transcription by recruiting transcriptional machinery, either via genomic proximity or 3D chromatin interactions (van Steensel Furlong, 2019). They form dynamic three-dimensional hubs—aggregates of Mediator complexes, TFs, and RNA Pol II—that bridge distant enhancers to promoters (Chong et al., 2018; Sabari et al., 2018). Enhancers can often be hundreds of thousands of base pairs away from their promoter, and still influence gene regulation. One of the pioneering transformer based models, Enformer, Ziga Avsec 2021 Nature genetics) effectively predict gene expression , incorporating these long range interactions, taking in an input sequence of 198,000 base pairs.

Enhancer activity is governed by combinatorial TF binding: clusters of suboptimal binding sites can confer higher specificity than perfect sites (Smith et al., 2013; Farley et al., 2015), and short tandem repeats further enhance local TF density via multiple weak interactions (Horton et al., 2023). Active enhancers can also be transcribed bidirectionally to produce eRNAs that influence transcription (Carillo et al., 2020). Chromatin accessibility is an important precursor for TFs to be able to bind an enhancer, and chromatin states are defined by a complex interplay of nucleosome positioning, histone modifications, and DNA methylation (Ernst Kellis, 2017). For example, the poised enhancer mark H3K4me1 is associated with increased accessibility (Lara-Astiaso et al., 2014)

3.6 Deciphering the cis-regulatory code

The engagement of all these components highlights the presence of a "cis-regulatory code" that ties together the different levels of regulation. The cis-regulatory code describes how cells interpret DNA sequences to determine, where, when and how much of what genes should be expressed. In simpler terms, cis-regulation is a quantitative process reducing the complexity of DNA sequences into gene expression level (de Boer Taipale, 2024).

the four different layers of the cis-regulatory code are depicted. In the most inner layer TFs bind DNA. This is done with the recruitment of cofactors, which build the second layer. In the following layer, we observe longer genomic interactions, including enhancer-promoter communications. Ultimately, the diverse cis-regulatory elements interact with each other on subnuclear level (S. Kim Wysocka, 2023).

Gene regulatory networks (GRNs) are one way of understanding this code, as they model the interaction between genes, TFs, and other regulatory elements. In this network view, GRNs consists of nodes, representing genes and regulatory elements, and edges, representing their interaction between them. The GRN of cellular differentiation is organized in a flat hierarchy and only a few TFs control many downstream target (Graf Enver, 2009). Reorganization of developmental programs is achieved through regulatory alterations. Depending on the hierarchical position of these changes in the GRN, alterations in top-level factors tend to have more severe consequences (Erwin Davidson, 2009). In diseased and perturbed states the GRN re-wires, causing that usually exclusive regulatory programs run simultaneously. This mixed regulome causes aberrant mixed lineage states, a hallmark of disease formation and progression (Corces et al., 2016). A key limitation of using natural genomic sequences is their sequence homology and limited genetic diversity. A recent opinion piece by de Boer Taipale, 2024 argues that the natural genome is too small to learn all the rules by which it is decoded. The authors state that any of the approximately 1,639 human TFs could potentially could interact with any other, one would need to test around 220 million unknowns to capture all TF-TF cooperative interactions. Even if a genome of transcription factor interaction sequences were constructed, the lack of multiple independent examples would make it hard to distinguish these effects from those of the surrounding sequence context. Essentially, the natural genome might be too short to encompass all the parameters needed for a complete description of cis-regulation (de Boer Taipale, 2024). The authors propose that training models on designed DNA sequences offers a promising alternative. Such sequences can be tailored to test specific hypotheses or achieve precise expression goals, providing greater flexibility and accuracy. Since the same biochemical principles apply to both natural and synthetic DNA, models trained on designed sequences often predict genomic activity with higher precision than those based solely on natural genomic sequences. To tackle this problem, one needs models and model systems which are simplistic, but yet complex enough. The hematopoietic stem cell system gives unique opportunities for differentiation, manipulation, perturbation, as well as ex vivo or in vivo culturing options.

3.7 Dissecting the regulatory logic of synthetic GREs with Deep Learning

A key challenge in understanding gene regulation is that it's a complex, combinatorial, and multi-layered process with an immense number of variables. In genomic measurements, effects from all this variables mix together, and its challenging to extract the contribution from individual variables. The sequence space needed to thoroughly explore the combinations of these variables is vast and surpasses the diversity found in the native genome, particularly due to evolutionary constraints. For the same reasons, capitulating naturally occurring molecular interactions leading to transcription would result in a model with a wide range of parameters, with little generalisability.

To address this challenge, it is crucial to study these processes within a more controlled variable space. Ideally, an experimental system should allow for the selective perturbation of individual variables while keeping all others constant, enabling a more precise understanding of gene regulation. One approach to address this issue is the use of synthetic DNA libraries, which allows to test thousands of sequence while controlling for specific variables. By utilizing engineered, artificially designed sequences, these assays enable the measurement of biological outputs such as TF binding or transcription levels while systematically altering one genetic factor at a time such as number of TFBSs and motif strength.

One of the first studies focused on the 12 liver-specific TFs and assessed how homotypic and heterotypic enhancers are able to generate expression. The authors showed that a flexible organization model supports their data best. For the 12 TFs the authors were able to find distinct TF synergies, as well as a distinct TF interference of two TFs (Smith et al., 2013). In further detail and with a higher throughput, the same set of authors were able to expand their setup to measure 209,440 sequences for 18 TFs, measured in pairs and triplets, in the same cell type. For this set of TEs the impact of strand asymmetry of TFBSs, distance from the TSS, motif orientation/copy number and TF order were quantified for each pair/triplet. In their analysis TFBS orientation is a major driver of activity. What is surprising is that heterotypic TFBS optimal grammar seems to be independent from the number of copies (Georgakopoulos-Soares et al., 2023). One common component of all the described approaches is scaling up even further the throughput and efficiency. A pivotal example of this is the gene expression quantification of more than 100 million reporter constructs in yeast, which allows to build deep learning models of transcriptional regulation with very high accuracy (de Boer et al., 2020).

Synthetic enhancers offer several advantages over natural genomic sequences: they allow for the precise manipulation of individual variables, enable the exploration of sequence space without evolutionary constraints, and can be designed to test specific hypotheses about transcriptional regulation. By controlling factors such as motif number, strength, and arrangement, synthetic sequences provide a clearer view of how individual elements contribute to regulatory activity.

3.7.1 Employing Deep Learning to decode cis-regulatory rules

In this section, I will discuss how DNA sequence alone can be used to predict function and how sequences with desired regulatory functions can be designed.

DNA sequence activity prediction and in silico enhancer design One of the earliest successful models at prediction expression from sequence was given by Beer and colleagues(Beer Tavazoie, 2004). They employed a Bayesian network which maps sequence features (xi) to patterns of expression (ei) by encoding the probability P(ei—xi) that genes with these sequence features will participate in a certain expression pattern. Starting in 2015, even before contemporary deep learning libraries like tensorflow and pytorch were available, a deep learning model for TF binding prediction was developed called DEEPBind deepbind citation Nowadays, a slew of models for different data modalities and multimodal data are available. Model accuracy as well as the throughput and quality of data is increasing every year, taking us closer to a predictive underrating of biological systems. Deep neural networks excel at a wide ranging tasks from predicting Tf binding at base pair resolution (Avsec, Agarwal, et al., 2021) to, in stark contrast, incorporating long range interactions between distal enhancers (Avsec, John jumper 2021, deep mind). Deep learning models based on encoder-decoder architectures also excel at generative tasks. Afer training on millions of artificially designed 3' untranslated regions (UTRs), deep models were not only able to predict variant effect and usage of alternative polyadenylation sites, but they were also capable of accurately engineering polyadenylation signals with desired properties (Bogard et al., 2019). Whole new generative architecture structures, coordinated with the predictive components, generate sequences in silico in a high-throughput manner with a user-defined target (Linder et al., 2020). With all these advancements, it is now possible to build synthetic enhancers with cell type specific expression in Drosophila (de Almeida et al., 2024; Taskiran et al., 2024). State of the art neural networks are able to build synthetic enhancers for discrete cell types (Barbadilla-Martínez et al., 2024; Gosai et al., 2023).

3.7.2 A primer on Deep Learning in Biology for Sequence-Based Tasks

I will briefly focus on an introduction for deep learning methods applied to sequence-based tasks, focusing on Deep Neural Networks , Convolutional (CNNs), Recurrent (RNNs), and attention-based architectures.

Biological sequences, such as DNA, RNA, and protein sequences, encode complex information that governs cellular processes. Computational methods are crucial in deciphering these sequences, predicting structural and functional properties, and modeling gene regulation. Deep learning provides powerful tools to extract hierarchical features from raw sequences without requiring manual feature engineering.

Formally, a biological sequence can be represented as:

$$S = (s_1, s_2, \ldots, s_n), \tag{1}$$

where each s_i is a categorical element from an alphabet (e.g., $\{A, T, C, G\}$ for DNA).

Deep Neural Networks (DNNs) and Fully Connected Layers Deep Neural Networks (DNNs) consist of multiple layers of interconnected neurons that transform input data through a series of linear and non-linear operations. DNNs are commonly used for sequence-based classification tasks in biology, such as predicting enhancer activity or identifying disease-associated mutations. However, fully connected networks alone struggle to model spatial and temporal dependencies in sequences, necessitating the use of more specialized architectures like CNNs, RNNs, and attention mechanisms.

Convolutional Neural Networks (CNNs) for Sequence Tasks CNNs were originally designed for image data but have been successfully applied to sequence tasks by leveraging local dependencies.

A 1D convolutional layer applies a kernel W over a sequence S to extract features:

$$h_i = f\left(\sum_{j=1}^{k} W_j s_{i+j-1} + b\right), \tag{2}$$

where k is the kernel size, b is a bias term, and f is a non-linear activation function (e.g., ReLU).

A sequence of convolutional layers followed by pooling can capture hierarchical motifs in sequences, making CNNs effective for tasks like promoter prediction and protein function annotation.

Recurrent Neural Networks (RNNs) Unlike CNNs, RNNs explicitly model sequential dependencies using recurrent connections. The hidden state at time step t is updated as:

$$h_t = f(W_h h_{t-1} + W_x x_t + b). \tag{3}$$

Here, x_t represents the input at time step t, and h_t captures sequential dependencies.

However, RNNs suffer from vanishing gradients when modeling long-range dependencies, limiting their effectiveness for long biological sequences.

Attention Mechanisms and Transformer Architectures Transformers and self-attention mechanisms have recently outperformed RNN-based methods in sequence modeling tasks.

Self-attention computes context-aware embeddings for each sequence element by weighting all other elements based on learned attention scores:

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

where Q, K, V are the query and key and value matrices, and d_k is the dimensionality of keys.

Multi-head attention further enhances the model's ability to capture complex dependencies:

$$MultiHead(Q, K, V) = Concat(head_1, ..., head_h)W^O. \tag{5}$$

Transformers such as BERT and AlphaFold have demonstrated remarkable success in biological sequence analysis, including protein folding prediction and gene expression modeling.

3.8 Hematopoietic Stem Cells

Cell types are defined by a unique set of features—morphology, function, gene expression, and chromatin state—that are maintained by transcription factor complexes, enabling adaptability and the evolution of new cell types (Arendt et al., 2016). HSCs are multipotent and follow a continuous differentiation trajectory—as revealed by single-cell RNA sequencing (Velten et al., 2017; Giladi et al., 2018)—a concept that refines earlier established hierarchical models (Spangrude et al., 1988) and later detailed by Laurenti Göttgens (2018). This system exemplifies how a single stem cell population can produce a diverse array of cells via tightly regulated transcriptional programs and evolutionary diversification (Dubovik et al., 2024).

3.8.1 Transcription factors and their impact on lineage commitment

TFs can influence cellular lineage commitment through various mechanism. Their absence and presence can cause the cells to commit to the one or another lineage, functioning as a lineage-determining factor (Kulessa et al., 1995). These specific TF cross-antagonism circuits can explain how stable lineage choice decisions can be made (S. Huang et al., 2007).

In addition to these direct effects, fluctuations and noise of TFs can also be instrumental for lineage choice. Stochastic gene expression can alter the probability of committing towards a cellular fate (Chang et al., 2008; Raj van Oudenaarden, 2008).

3.8.2 Major TFs in hematopoietic lineage commitment

To ensure the long-term sustainability of the hematopoietic system, a stable pool of stem cells is preserved constantly, but about 50 billion new cells are produced per day Parslow TG, Stites DP, Terr AI, Imboden JB (1997). Medical Immunology (1 ed.). Appleton Lange. A heptad of TFs, consisting out of GATA1, Erg, Fli1, Gata2, Lyl1, Lmo2,and Runx1, are key for maintaining this stem cell reservoir (N. K. Wilson, et al., 2010). The HSC heptad forms a tightly interconnected transcriptional network that governs stem cell gene expression through extensive positive feedback and cross-regulation. For example, Gata2 is critical: its overexpression prevents HSC differentiation (Persons et al., 1999), while its deficiency halts precursor development and compromises HSC survival (Tsai Orkin, 1997). Moreover, Gata2 enhances Tal1 expression, which, together with Gata1, directs erythroid commitment (Swiers et al., 2006; Bresnick et al., 2010). In contrast, Runx1 stimulates Pu.1 and Cebpa expression to drive lymphoid and myeloid fates (Gu et al., 2014). The mutual antagonism between Gata1 and Pu.1 forms a bistable switch that biases cells toward either erythroid or lympho-myeloid fates (Arinobu et al., 2007), which has also been modelled mathematically Bokes et al 2009. Together, these transcription factors orchestrate lineage commitment by modulating target gene expression in a dosage- and context-dependent manner, with subtle shifts in their levels triggering specific differentiation programs (Hosokawa et al., 2018).

3.9 Assay for decoding cis-regulatory rules

Dissecting the regulatory logic of sequences and enhancers is key to understanding development, lineage specification and disease, as genome-wide association studies have revealed that much of the observed variation lies in regulatory elements. Manolio, T . A. et al. Massively Parallel Reporter Assays (MPRAs) enable the simultaneous testing of thousands of natural regulatory sequences by linking each enhancer or promoter to a reporter gene and a unique molecular barcode. This allows for quantitative measurement of gene expression, mapping of transcription factor binding sites, and assessment of motif strength. Variants in enhancers have been shown to alter activity, as demonstrated in studies by Kheradpour et al. (2013) and Patwardhan et al. (2012). Techniques such as lentiviral MPRA (lentiMPRA) extend these capabilities to cells that are hard to transfect, while STARR-seq leverages self-transcribing regulatory elements to directly predict enhancer activity from DNA sequence. Since a cell state can be defined arbitrarily to encompass a wide range of characteristics, Single-cell MPRAs (scMPRA) further refine these approaches by capturing cell-to-cell variability in enhancer activity, albeit with lower throughput. Zhao S, Hong CKY, Myers CA, Granas DM, White MA, Corbo JC, Cohen BA. A single-cell massively parallel reporter assay detects cell-type-specific gene regulation. Nat Genet. 2023 Feb;55(2):346-354. doi: 10.1038/s41588-022-01278-7. Epub 2023 Jan 12. PMID: 36635387; PMCID: PMC9931678. The data used for mathematical modeling and deep learning originates from Fromel et al. (2024). Data generation involved cloning libraries upstream of a minimal promoter within a lentiviral MPRA reporter vector (see Figure 1 for details). The Velten group employed 62,126 fully synthetic DNA sequences to investigate the differentiation of hematopoietic stem cells into seven myeloid lineages.

Library A focused on individual transcription factor (TF) binding sites within random DNA. It comprised candidate enhancer sequences featuring one to six binding sites for a single TF, systematically varying the spacing, arrangement, orientation, number, and affinity of the motifs.

In contrast, Libraries B and C explored the function of paired TFs by placing combinations of binding sites into random DNA. Library B examined all 45 pairwise combinations from a biologically important set of 10 TFs by positioning either one or three binding sites for each factor and varying the spacing, orientation, and arrangement of the motifs. Library C extended this analysis to cover a total of 861 distinct TF combinations. The resulting dataset had about a million data entries of interest (62,126 sequences and 15 columns of interest (cell state, design, sequence, expression). The datatset was first visualised and analysed independently using multiple strategies like Facet Grids, Pair Plots, Parallel coordinates plot, Heatmaps and correlation matrices, principal component analysis and Clustered Bar Plots, etc. Some interesting visualisations, which can be found in the appendix (Figures 2-4).

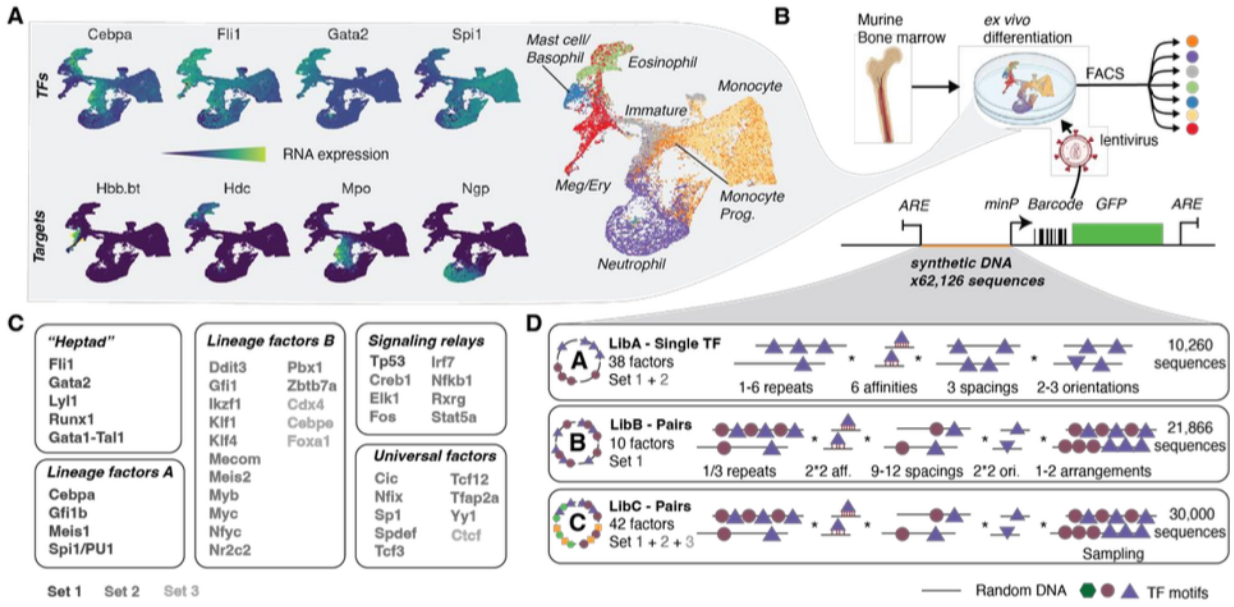

Figure 2: Description of the dataset. (Figure taken from Fromel et al 2024) A. Transcription factors are expressed with broad overlaps in expresssioin across multiple cells tatess. These states can be idetified with uMAPs of scRNA-seq data of undifferentiayed HSCs in culture highlighting gene expression. B. Experimental design. HSCs were placed into culture medium supporting pan-myeloid differentiation. Synthetic enhancer constructs were delivered using the lentiMPRA viral vector, and then subsequently cells were FACS-sorted into seven cell states post differentiation , and the activity of each synthetic enhancer in each cell state was measured. ARE stands for Antirepressor Element minP stands for minimal promoter. C. The various TFs who's motifs were implanted into the synthetic constructs. D) Design of candidate DNA libraries. TF motifs (depicted by circular or triangular figures) of single factors. or factor pairs (Library B/C) were embedded in random DNA at different binding site affinities, number of repeats, and spacing and orientation between motifs.

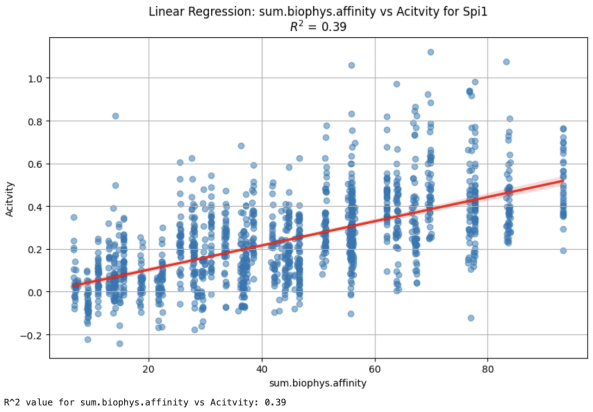

Figure 3: Activity (expression from synthetically deisgned sequence) vs biophysical affinity for implanted motifs with TF Spi1

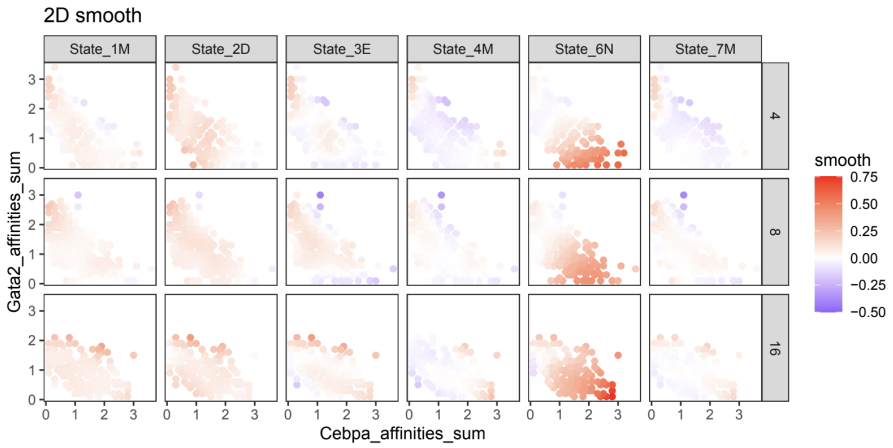

Figure 4: In cell state 4, both Cebpa and Gata2 are activators but repress when their tF binfind motifs are placed together. This behaviour will be recapitulated with mathematical models.

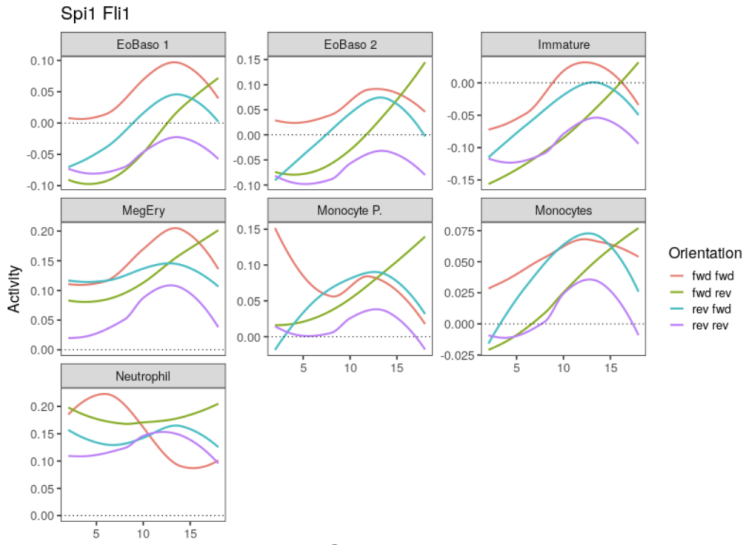

Figure 5: Spi1 and fli1 show spacing dependent increase or decrease in activity when their motifs are placed together.

3.10 Non Monotonic Responses

The following nonmonotonic responses were observed from the data for some transcription factors as a function of the number of affinities and the strength of implanted motifs (sum of affinities).

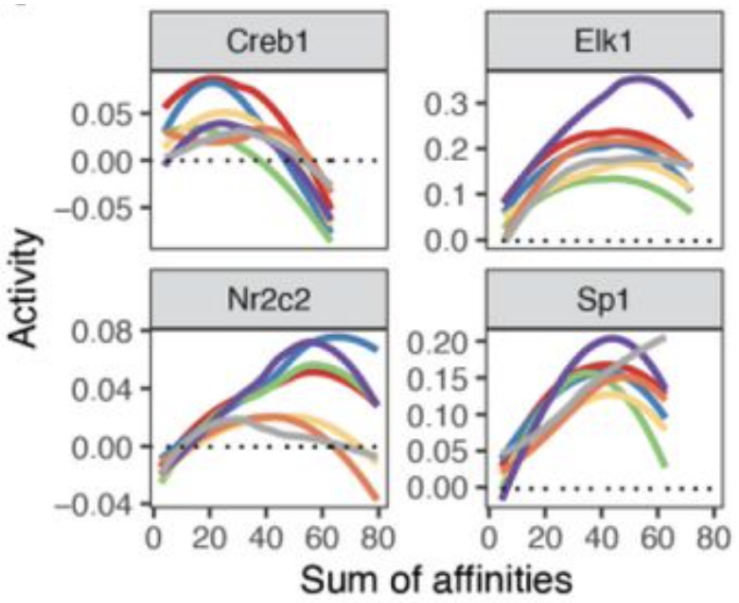

Figure 6: Enhancers can function as band-pass filters due to a combination of non-monotonicity and occupancy-dependent duality. The analysis, conducted using Library A, includes examples of non-monotonic behavior (Panel A). The line plots show smoothed mean activity values as a function of the total motif affinity in the sequences for Creb1, Elk1, Nr2c2, and Sp1. Different colors indicate various cell states.

Nonmonotonic responses have been theoretically explained in literature with the following two approaches given below, the latter being what the thesis builds upon. (Mahdavi et al 2024 PNAS, and Martinez Corral et al 2024 Biorxiv)

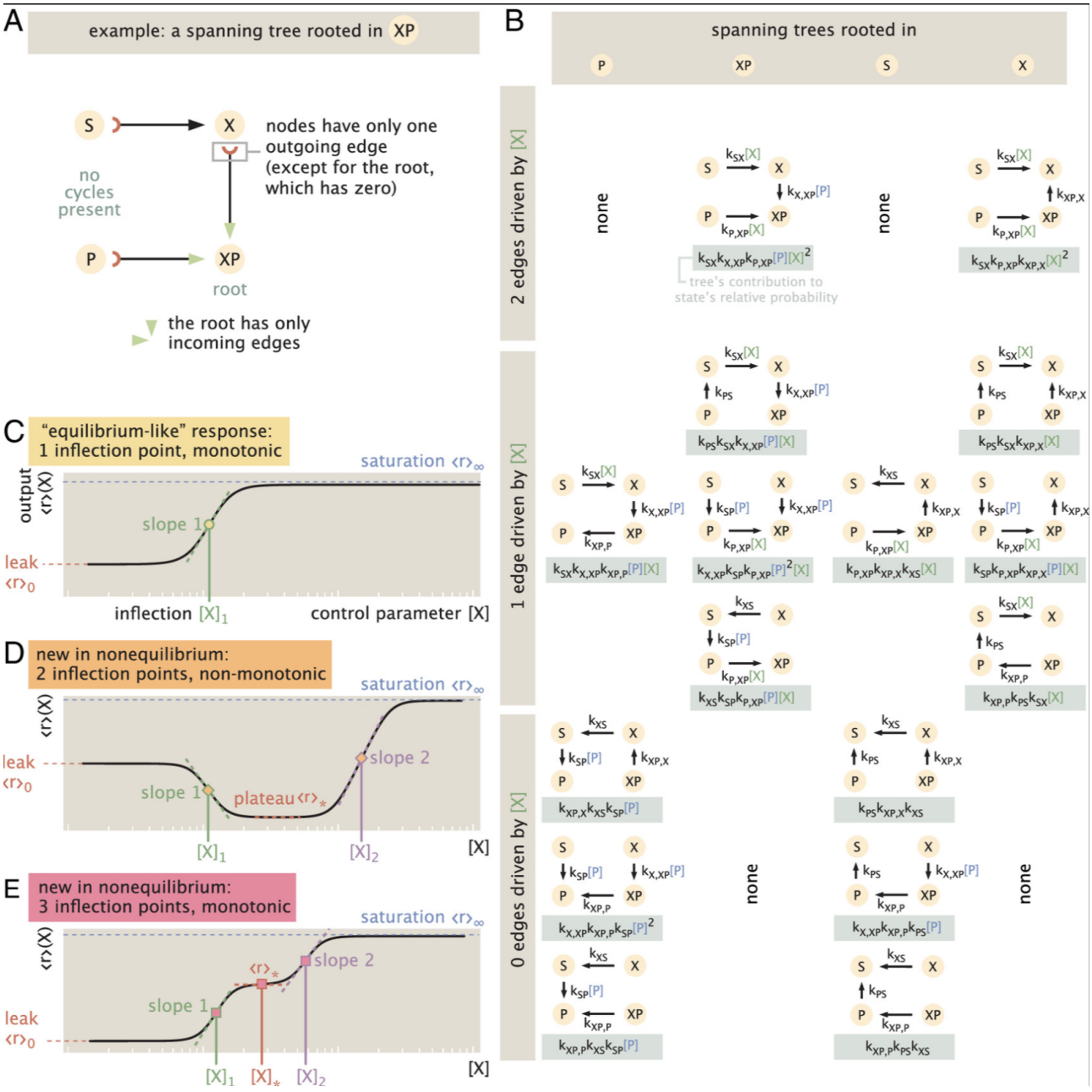

Figure 7: Non-equilibrium behavior of the four-state transcriptional motif. (A) Panel A shows an example of a spanning tree—rooted at state XP—similar to those used to compute steady-state probabilities via the Matrix Tree Theorem. (B) Panel B presents all 16 directed, rooted spanning trees corresponding to the four-state cycle depicted in Panel A. These trees are organized by their root state (columns) and by the number of edges affected by the control parameter (rows). According to the Matrix Tree Theorem, the steady-state probability of any state—whether at equilibrium or not—is determined by the sum of the weights of these spanning trees, which introduces up to a quadratic dependence in the output. (C–E) Panels C through E demonstrate three general output behaviors (regulatory shape phenotypes) emerging from this framework. A monotonic "equilibrium-like" output (Panel C) exhibits a Hill-like or MWC-like response, similar to what is seen in equilibrium thermodynamic models. In contrast, exclusively under non-equilibrium conditions, novel regulatory shapes with multiple inflection points can occur. For instance, Panel D shows outputs that vary non-monotonically with the control parameter, featuring two inflection points, while another scenario displays three inflection points with an overall monotonic trend.

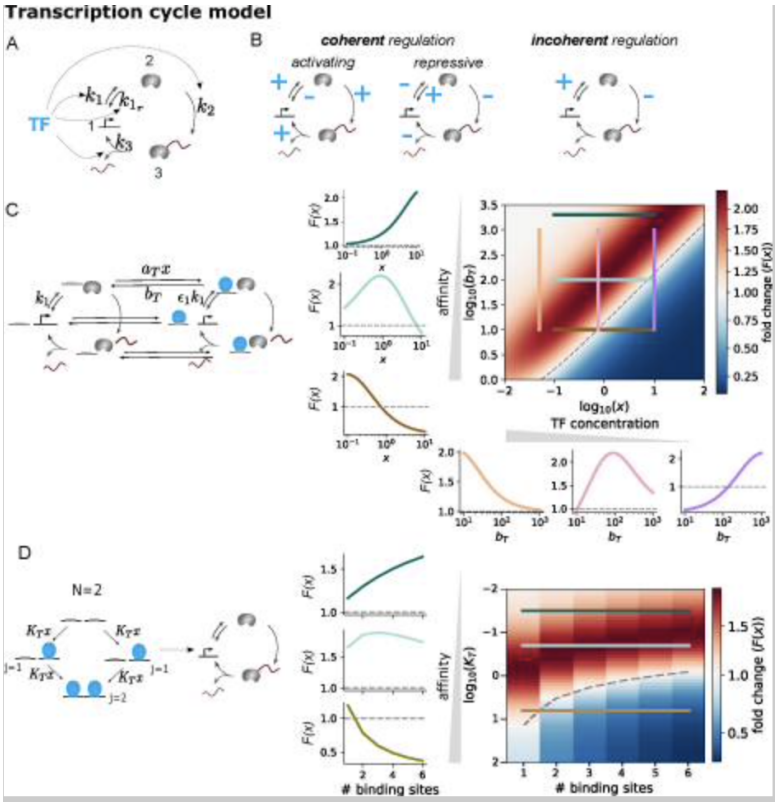

Figure 8: Duality in TF regulation of the RNA polymerase transcriptional cycle is illustrated as follows: A) A schematic of the model shows how a transcription factor (TF) can modulate one or several transitions within a three-state transcriptional cycle. B) Examples of regulatory effects are provided for both coherent and incoherent regulation. In this context, a plus sign indicates that the TF accelerates a transition, while a minus sign signifies deceleration. (See text for further details.) C) (Left) A graph is shown where the TF influences a rate while bound; only selected edge labels are displayed for clarity. (See text for further details.) D) A model is presented in which the transcription cycle rates are modulated by the equilibrium average number of bound TF molecules. The graph illustrates the case of independent binding to N = 2 sites. A colormap displays the fold change as a function of both affinity and number of binding sites, while additional plots on the left depict the fold change as a function of the number of sites for three different affinity values (KT = 0.03, 0.2, and 6.31).

4 Methods

4.1 Data Acquisition and Preprocessing

Gene expression datasets, particularly Massively Parallel Reporter Assay data, utilized in this research were sourced from the following preprint : R. Froemel et al 2024 Biorxiv). Visualizations were done using the column `mean.scaled.final` while for modeling the column of interest was `mean.norm.adj`.

ChIP-Seq data was downloaded from the GEO Accession Viewer link available in the Supplementary Information of S. Subramanian et al, BLOOD 2023. Pre-processing included converting BigWig files to Wig and then to Bed using BioConda. Peak files for 10 transcription factors of interest, in 4 hematopoietic stem cell states (after MACS2 processing), were also used. Sequences were obtained from the reference Hg38 genome in FASTA format. Genomic annotations for enhancer and promoter regions were taken from the following article by Zacher B et al. (2017) PLOS One and then cleaned to ensure the regions were selected for deep learning model training.

4.2 Biophysical Modeling

To elucidate gene regulatory mechanisms, a biophysical model was designed to represent the dynamic interactions between transcription factors (TFs), their effects on the transcriptional cycle, and DNA binding sites. Parameterization of the model was performed —- including TF-DNA binding affinities and reaction kinetics —-based on experimental literature (Martinez-Corral et al. 2023 Cell Systems). Model equations were solved

using computational tools implemented in Python (leveraging SciPy, NumPy, and Pandas). Sensitivity analyses were performed to evaluate the robustness and reliability of the parameter estimates.

4.3 Deep Learning Model Development

A comprehensive deep learning framework was constructed using PyTorch to predict gene regulatory outcomes derived from biophysical modeling inputs. Hyperparameter optimization, including tuning of the learning rate, layer depth, neuron density and activation functions, was executed using grid search techniques.

4.4 Deep Learning Training Pipeline

The deep learning training pipeline was implemented in Python using PyTorch and is designed to perform a comprehensive grid search over multiple file sets and hyperparameter combinations. The overall procedure is as follows:

1. **File Set Management:** A list of file sets is defined (each containing separate training, validation, and test files). For each file set, a dedicated output directory is created to store model weights and logs.
2. **Hyperparameter Grid Search:** The pipeline iterates over combinations of training batch sizes and learning rates. For each combination, the number of batches per epoch is calculated based on the size of the training dataset. The model components and data processors are reinitialized to ensure independent training runs.
3. **Training Loop:** For every experiment, the model is trained for a fixed number of epochs using a custom trainer module. After training, results (e.g., performance metrics or placeholder outputs) are stored for subsequent analysis.
4. **Reproducibility:** A PyTorch random generator is seeded to ensure consistent model initialization and training behavior across experiments.

The following subsections detail the three architectures integrated into this pipeline.

4.4.1 CNN-based Architecture

The CNN-based model leverages convolutional operations to capture local sequence motifs:

- **Feature Extraction:** The architecture starts with the `BHIFirstLayersBlock`, which applies convolutional filters (with kernel sizes of 9 and 15) to extract primary features from the 250 bp input sequences. Dropout is employed to mitigate overfitting.
- **Core Processing:** The extracted features are further processed by the `AutosomeCoreBlock`, which refines these representations through additional convolutional transformations.
- **Final Prediction:** The `AutosomeFinalLayersBlock` consolidates the processed features using convolutional and linear layers to output a scalar prediction that reflects the regulatory signal.

This design efficiently models spatial dependencies in the sequence data.

4.4.2 RNN-based Architecture

The RNN-based model extends the convolutional framework by incorporating sequential dynamics via recurrent layers:

- **Initial Feature Extraction:** Similar to the CNN-based model, the `BHIFirstLayersBlock` is used to capture local features from the input.
- **Recurrent Processing:** The core of the RNN model is implemented using the `BHICoreBlock`, which integrates an LSTM layer to capture long-range dependencies across the sequence. Supplementary convolutional operations and dropout layers further enhance feature extraction and robustness.
- **Final Layers:** As with the CNN architecture, the final layers (via the `AutosomeFinalLayersBlock`) aggregate the processed sequential features into a final prediction.

This hybrid approach is designed to capture both local patterns and temporal dependencies within the enhancer sequences.

4.4.3 Attention-based Architecture

The Attention-based model utilizes transformer-inspired mechanisms to focus dynamically on informative sequence regions:

- **Initial Processing:** The model begins with the `AutosomeFirstLayersBlock` for basic feature extraction via convolutional operations.
- **Attention Core:** The core processing is performed by the `UnlockDNACoreBlock`, which integrates multi-head attention mechanisms over the input. This module allows the model to weigh different regions of the sequence dynamically, thereby capturing long-range interactions and contextual dependencies more effectively.
- **Final Prediction:** Finally, the `AutosomeFinalLayersBlock` processes the attention-enhanced features and produces a scalar output corresponding to the predicted regulatory signal.

The attention mechanism provides a flexible way to learn which parts of the sequence contribute most significantly to transcription regulation.

4.5 Training and Validation

Datasets were systematically divided into training (80%), validation (10%), and testing (10%) subsets using stratified sampling methods to maintain the class distribution. Model training employed the Adam optimization algorithm with mean squared error (MSE) as the loss function for continuous prediction scenarios. Early stopping mechanisms based on monitoring validation loss were implemented to prevent overfitting. Performance evaluation metrics included Pearson and Spearman correlation coefficients (PCC and SCC).

4.6 Visualization and Interpretation

Visualization of results was achieved using Matplotlib and Seaborn libraries to provide representations of model predictions and underlying data distributions. To enhance interpretability, feature importance analysis and model explainability were performed using SHAP (SHapley Additive exPlanations), elucidating the contribution and significance of individual genomic features and TF binding sites on regulatory outcomes.

4.7 Computational Environment and Reproducibility

Analyses were carried out in Jupyter Notebook environments, facilitating reproducibility across computational platforms. Version control was managed using Git and GitHub. Dependencies were managed through Conda environments to ensure consistency and replicability of results. The following commands can be executed on a Linux system to reproduce the environment for deep learning model training:

```
conda create -n test python=3.10
conda activate test
conda install pytorch==2.5.0 torchvision==0.20.0 torchaudio==2.5.0-
-pytorch-cuda=11.8 -c pytorch -c nvidia
pip install jupyterlab
pip install pandas
pip install tqdm
pip install scikit-learn
pip install biopython
pip install torchinfo
```

4.8 Computational Analysis of the model for Transcription Regulation

The computational analysis was done for a model of transcription regulation that incorporates both cooperative and anti-cooperative mechanisms for all 4 kinetic parameters k1, k2, k3, km1 in figure . For simplicity, effects of transcription factor occupancy were initially only set to affect 2 of these kinetic parameters. In the cooperative branch, TF binding reduces the effective available time for transcription, whereas in the anti-cooperative branch, binding delays are additive. The final transcription rate is computed as a function of pathway-specific rate constants, which are derived from Hill functions applied to TF binding levels.

4.8.1 Monte Carlo Sampling and Constraint-Based Filtering

To explore the parameter space of the model, a Monte Carlo sampling approach was implemented:

- **Parameter Sampling:** Free parameters (e.g., Hill constants, delays, kinetic parameters, and the basal transcription rate) are sampled uniformly from predefined ranges. These ranges may be narrow, wide, or extra wide (see the sensitivity analysis section).
- **Output Evaluation:** For each sampled parameter set, the log fold change in transcription output is computed as

$$\log\left(\frac{x_{obs}}{x_{basal}}\right),$$

  where if x_{basal} is set to 1, x_{obs} is calculated via the two pathways:

  - **Cooperative Pathway:**

$$X_A = delay_{A,max} \times Hill(A, K_A, 3), \quad X_B = delay_{B,max} \times Hill(B, K_B, 3),$$

    with

$$k_1 = \frac{1}{T_1 - (X_A + X_B)}.$$

  - **Anti-Cooperative Pathway:**

$$X'_A = delay_{A,prime} \times Hill(A, K'_A, 3), \quad X'_B = delay_{B,prime} \times Hill(B, K'_B, 3),$$

    with

$$k_2 = \frac{1}{T_2 + (X'_A + X'_B)}.$$

- **Constraint Filtering:** The model can be evaluated at canonical regions of TF binding, for example:
  - Region 1: High A only ($A = 1$, $B = 0$)
  - Region 2: High B only ($A = 0$, $B = 1$)
  - Region 3: Both High ($A = 1$, $B = 1$)
  - Region 4: Both Low ($A = 0$, $B = 0$)

  And then to reflect expected biological behavior, some constraints can be imposed. For example, to observe non-monotonic behaviour:

  - Output in Regions 1 and 2 must exceed that in Region 4.
  - Output in Region 3 must be lower than that in Region 4.

  Parameter sets satisfying these inequalities are stored for further analysis. For these accepted sets, average free parameter values and average log fold change outputs per region are computed and histograms are generated to visualize the distributions.

4.8.2 Global Sensitivity Analysis

To assess the influence of individual parameters on the model's output, a global sensitivity analysis was performed using Pearson correlation:

- **Scenarios of Parameter Ranges:** Three different scenarios (narrow, wide, and extra wide ranges) were defined for the free parameters. For each scenario, 1,000 parameter sets were uniformly sampled.
- **Correlation Computation:** For each scenario and at multiple TF binding levels, the log fold change was computed. Pearson correlation coefficients (and corresponding p-values) between each free parameter and the output were calculated, identifying the most influential parameters.
- **Visualization:** Bar plots of the absolute Pearson correlation coefficients for each scenario were produced to provide an overview of parameter sensitivities.

4.8.3 Contour Mapping of Log Fold Change

To visualize the dependence of transcriptional output on TF binding levels, a contour map was generated:

- **Grid Construction:** Two-dimensional grids for TF binding levels A and B (ranging from 0 to 1) were created.
- **Model Evaluation:** For each grid point, the log fold change was computed using the established model functions, yielding a two-dimensional map of the predicted output.
- **Contour Plot:** The grid was visualized using contour plots (with 50 levels and a diverging colormap) to depict the variation in log fold change as a function of A and B. Additionally, the model was evaluated at predefined regions (e.g., high A only, high B only, both high, and both low) to obtain specific output values.

This computational framework, implemented in Python with NumPy, Matplotlib, and SciPy, allows exploration of the parameter space of our mechanistic transcription regulation model and quantification of the sensitivity of model outputs to variations in free parameters. The contour maps provide a visual interpretation of how TF binding levels affect transcriptional regulation.

4.9 Mathematical Framework for Cellular Information Processing

In this work, we adopt a linear framework for timescale separation and steady-state analysis of biochemical networks—a framework introduced by Dr. Gunawardena (2012 PLOS One) and extended by subsequent studies (Martinez Corral et al 2023). This approach decomposes complex biochemical systems into fast and slow components, with the fast subsystem represented as a directed, labeled graph.

4.10 Graph Representation and Laplacian Dynamics

The fast biochemical processes are modeled by a finite graph G where:

- **Vertices** represent the fast components (or microstates) of the system.
- **Directed edges** represent chemical reactions or transitions, with each edge labeled by a rate $\ell(i \to j)$.

Assuming mass action kinetics, the dynamics of the fast subsystem are governed by the Laplacian matrix $\mathcal{L}(G)$:

$$\frac{dx(t)}{dt} = \mathcal{L}(G) \cdot x(t),$$

where $x(t)$ is the vector of component concentrations. Conservation of mass is enforced by

$$x_1(t) + x_2(t) + \ldots + x_n(t) = x_{tot},$$

and equivalently,

$$\mathcal{L}(G) \cdot \mathbf{1} = 0.$$

This formulation is equivalent to the master equation for a continuous-time Markov process (with concentrations replaced by probabilities $p(t)$):

$$\frac{dp}{dt} = \mathcal{L}(G) \cdot p, \quad p_1 + p_2 + \ldots + p_n = 1.$$

4.11 Steady States and the Matrix-Tree Theorem

At steady state, the fast subsystem is characterized by a null vector of $\mathcal{L}(G)$. For a strongly connected graph, the kernel of $\mathcal{L}(G)$ is one-dimensional. The Matrix-Tree Theorem (MTT) provides an explicit expression for the steady-state distribution:

$$\rho_i(G) = \sum_{T \in \Theta_i(G)} \prod_{j \to k \in T} \ell(j \to k),$$

where $\Theta_i(G)$ is the set of all spanning trees of G rooted at i. The steady-state probability is then given by:

$$p_i^* = \frac{\rho_i(G)}{\sum_{j=1}^{n} \rho_j(G)}.$$

4.12 Equilibrium Conditions and Path Independence

Under thermodynamic equilibrium, the system satisfies reversibility, so that for any reversible transitions $i \leftrightarrow j$,

$$p_i^* \ell(i \to j) = p_j^* \ell(j \to i).$$

This detailed balance condition simplifies the steady-state solution via path independence. By choosing a reference vertex i_0 and considering any path $\mathcal{P}$ from i_0 to i, define:

$$\mu_i(G) = \prod_{(i_1 \to i_2) \in \mathcal{P}} \frac{\ell(i_1 \to i_2)}{\ell(i_2 \to i_1)}.$$

Due to detailed balance, $\mu_i(G)$ is independent of the path chosen, and the steady-state probabilities can be expressed as:

$$p_i^* = \frac{\mu_i(G)}{\sum_{j=1}^{n} \mu_j(G)}.$$

Furthermore, the ratio of transition rates encodes free energy differences:

$$\frac{\ell(i \to j)}{\ell(j \to i)} = \exp\left(\frac{\Delta F}{k_B T}\right),$$

which links this framework directly to equilibrium statistical mechanics. This formulation can be viewed as a generalized partition function valid even for systems driven far from equilibrium. The formula used for steady-state mRNA production from a 3 step transcriptional cycle with 1 reversible and 2 irreversible steps can be derived from the procedure outlined above.

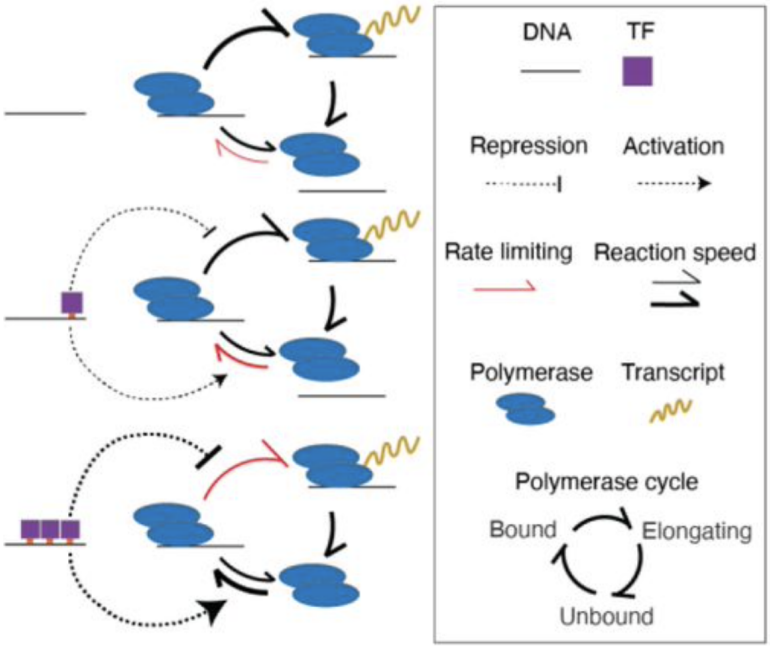

Figure 9: This model illustrates how transcription factors (TFs) can bind DNA and simultaneously activate one phase of the polymerase cycle while repressing another. Non-monotonic responses to TF binding site affinity may emerge when gene regulation is constrained at multiple steps and the TF exerts opposing effects—enhancing transcription at some steps and inhibiting it at others. To test whether this mechanism can explain the observed non-monotonicity, we employed a model in which the equilibrium average occupancy of TFs on the regulatory sequence modulates the rates of the polymerase cycle.

5 Results

5.1 Visualization and Analysis of Datasets

For a brief description of the MPRA dataset, please check the introduction. Chip-Seq data was taken from S. Subramanian et al. 2023 BLOOD, and data was available for 10 transcription factors across 4 cell states: HSC (undifferentiated), CMP (multipotent common myeloid progentior state for downstream differentiation of states GMP (granulocyte monocyte progrenitor) and MEP (megakaryocyte erythrocyte progenitor).

5.1.1 MPRA Analysis

Some of the visualisation can be found in the Supplementary information. However, only Fli1 is displayed which seems to have a saturation effect after the implantation of multiple high strength motifs.

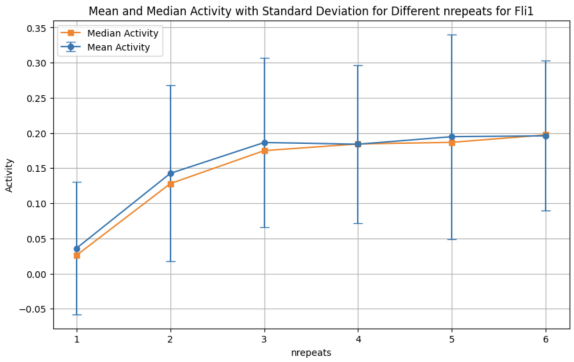

Figure 10:

5.1.2 ChIP-seq Analysis

More analysis and visualisations can be found in the appendix, especially with regards to the GENOSTAN annotations for enhancer and promoter regions, which offered around 6 per cent coverage from the whole genome. A fifth of the peaks lied outside these annottaions taken off the shelf, and hence it was decided against using them, besides other reasons mentioned below.

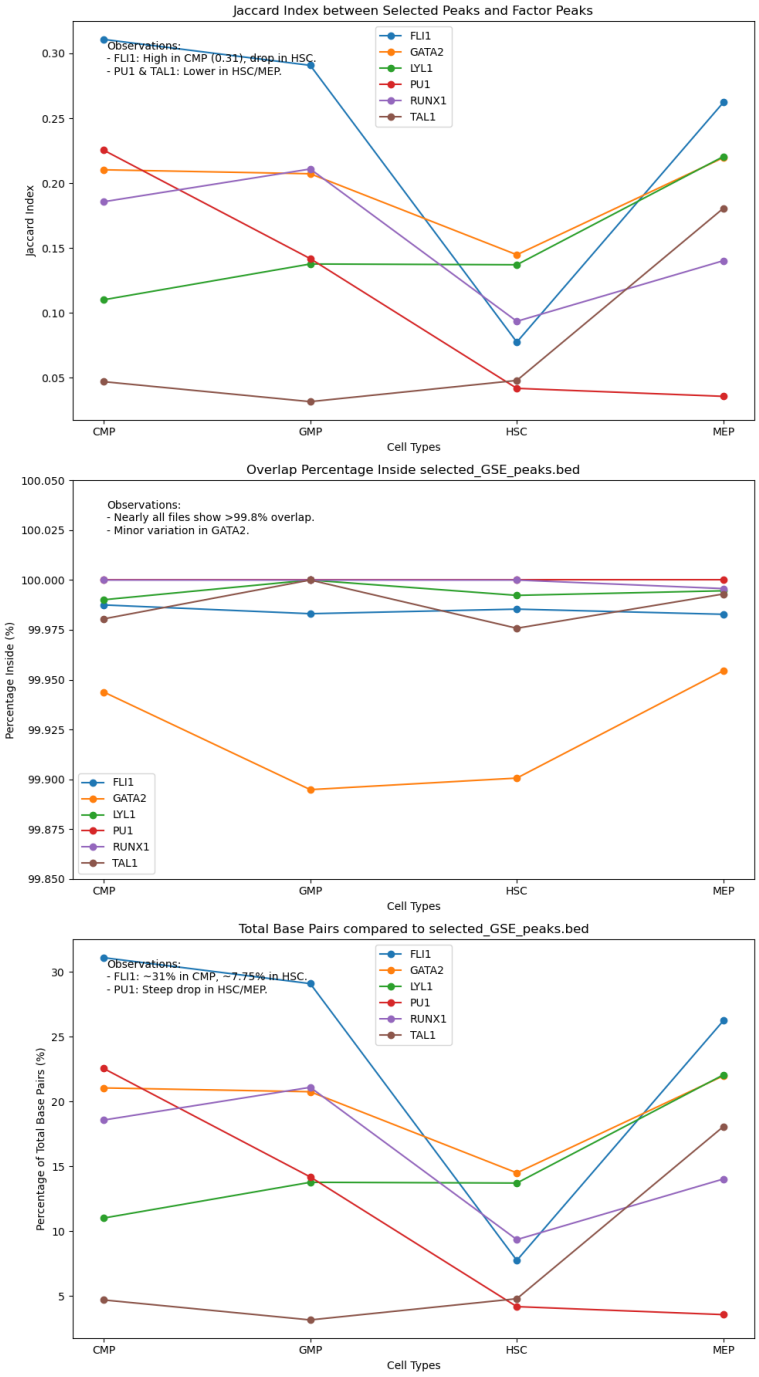

Figure 11: Comparison of Selected Peaks: (Top) Jaccard indices between selected peaks and factor peaks; (Middle) Percentage of peaks inside selected_GSE_peaks.bed; (Bottom) Total base pair coverage as a percentage of selected peaks.

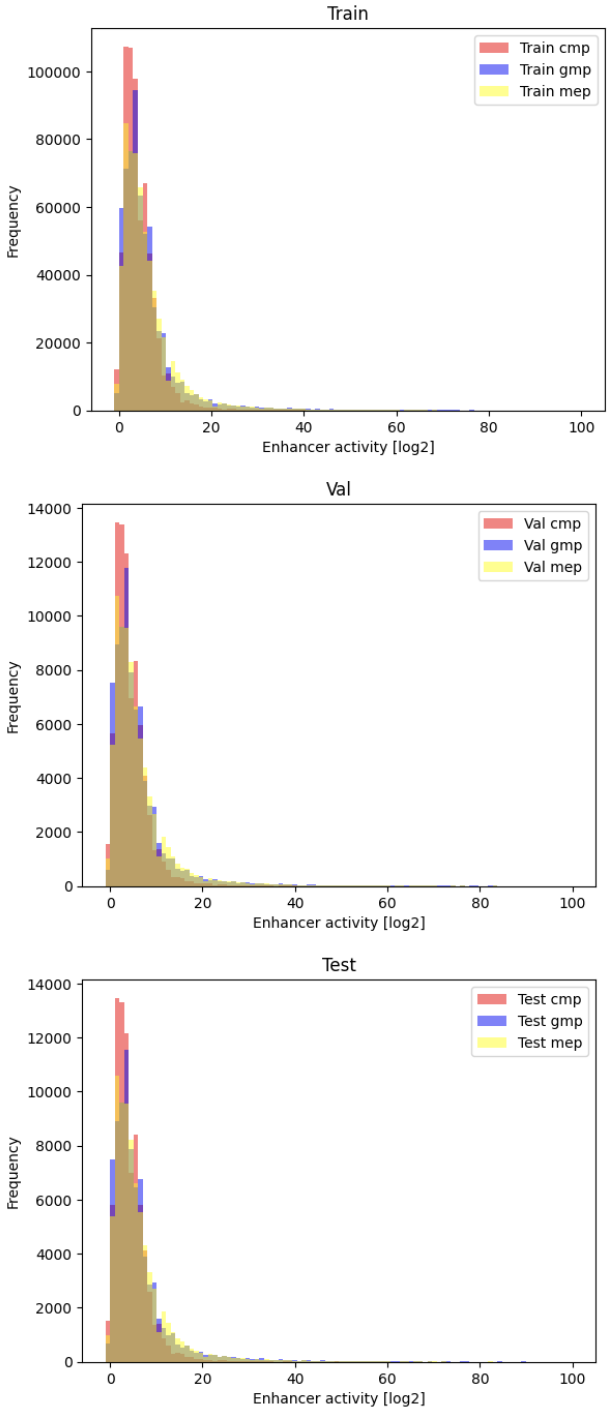

Figure 12: Train-Test-Val splits, before log-normalisation for the Chip-Seq data in 3 cell states

5.2 Detection of Cooperativity Between Heterotypic Transcription Factor Pairs

5.2.1 Methodology

An expectation maximization algorithm, developed by Datta et al. 2017 was used to infer cooperativity between heterotypic transcription factor pairs. A brief overview of the algorithm is given below, followed by plots for all pairs. of TFs with histograms for how probable it is that any identified pairs of peaks close together are cooperative or not. Transcription fcators seem to bind cooperatively to varying degrees in different cell states, as laid out in the Discussion.

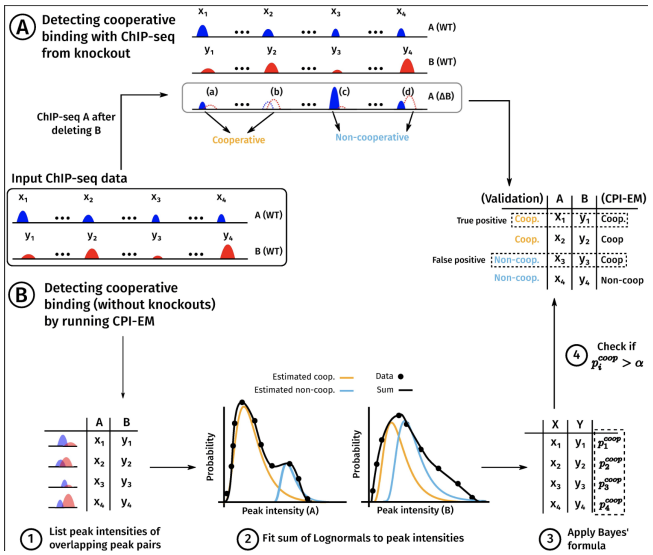

Figure 13: This schematic illustrates how the CPI-EM algorithm, is used to independently identify heterotypic TF pairs that bind cooperatively with method B. ChIP-seq experiments for two TFs, produce a list of genomic locations bound by both TFs, along with the corresponding peak intensities. From this dataset, two approaches are employed to determine regions of cooperative binding by A and B. The second method outlines the steps of the CPI-EM algorithm, as detailed in the "ChIP-seq Peak Intensity—Expectation Maximisation (CPI-EM) algorithm" 1. The algorithm starts with a list of genomic locations where a peak of TF A overlaps a peak of TF B by at least one base pair. (Note that the ChIP-seq data of TF A after TF B knockout is not used as input.) 2. Each overlapping pair of peak intensities is modeled as a sum of two probability functions that represent the likelihood of observing a given intensity pair from either a cooperatively or non-cooperatively bound region. These probabilities are estimated by fitting the model to the input data using the expectation-maximization algorithm whiich is available in the supplementary informatioin of the paper. 3. Bayes' theorem is then applied to these probabilities to determine the likelihood that each peak intensity pair corresponds to a cooperatively bound region. 4. Finally, any binding site with a cooperative probability greater than a chosen threshold is classified as cooperatively bound. The predicted list of cooperatively bound locations is then compared with the list inferred from knockout data to assess the number of correct and incorrect predictions made by the CPI-EM algorithm.

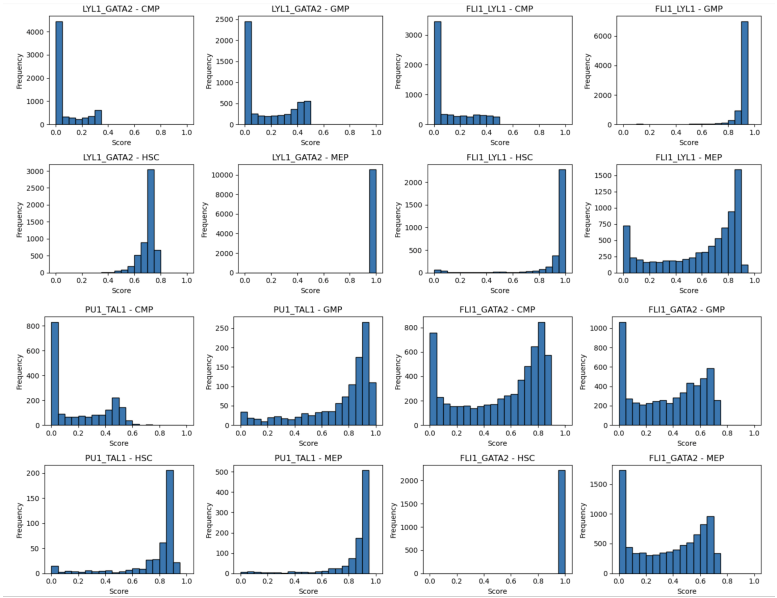

Figure 14: HetCoop 1

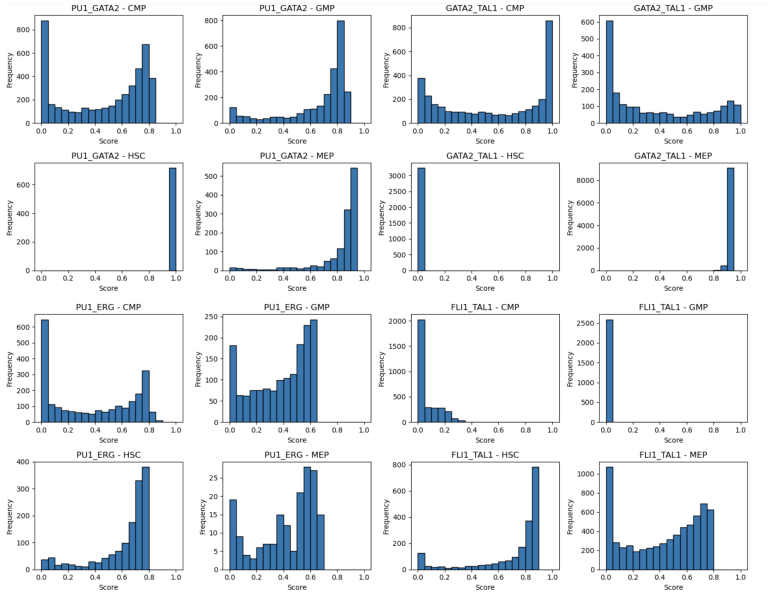

Figure 15: HetCoop 2

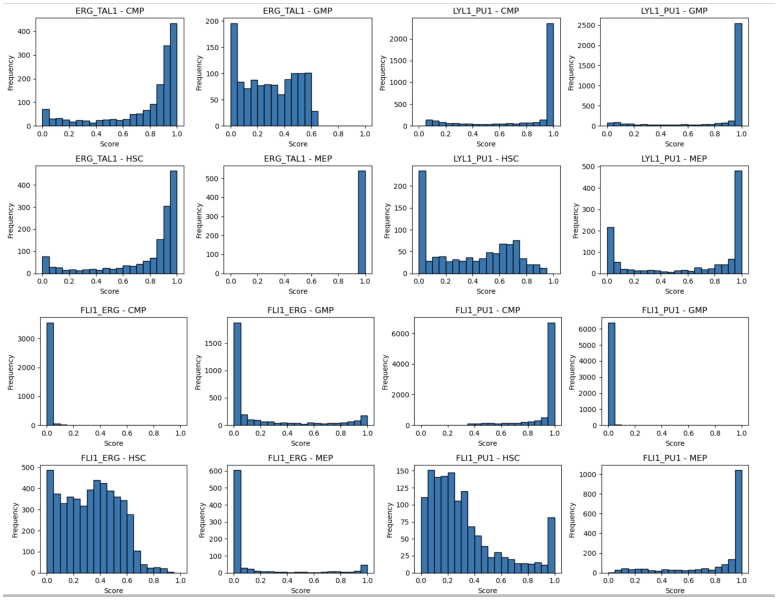

Figure 16: HetCoop 3

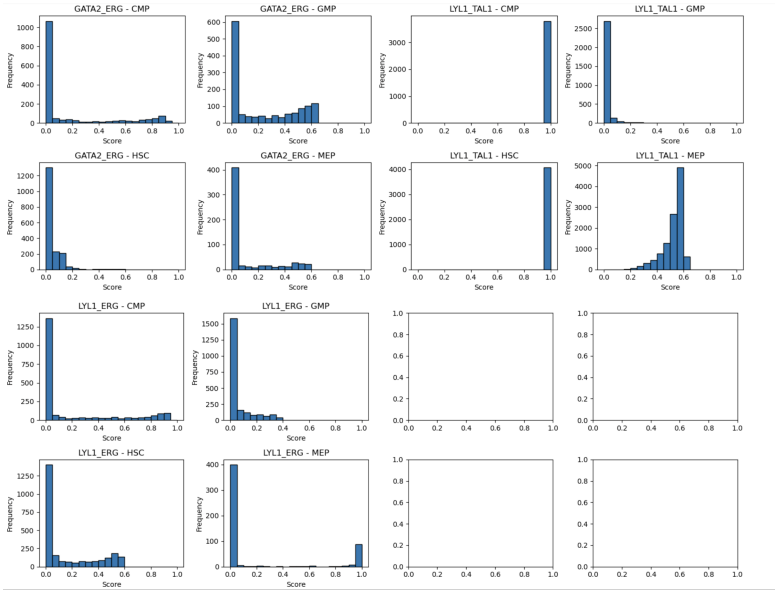

Figure 17: HetCoop 4

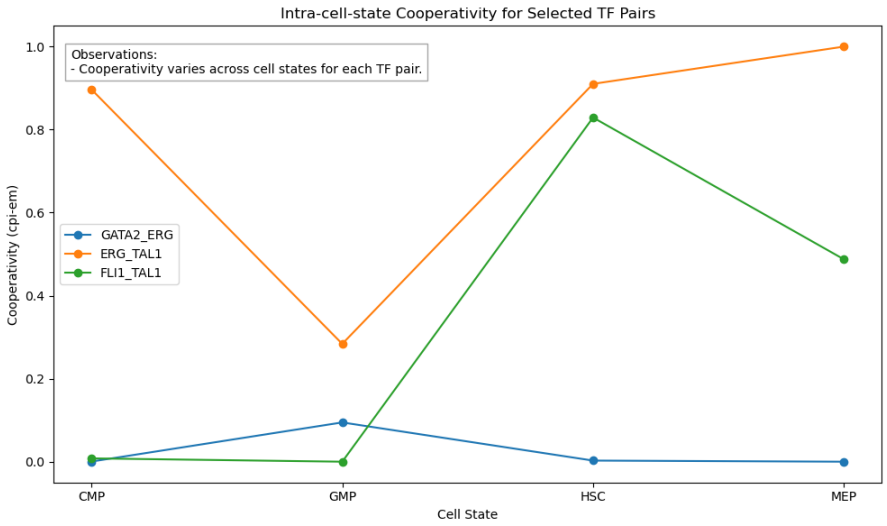

Figure 18: Intra-cell-state Cooperativity: Variation of cooperativity (cpi-em values) for selected TF pairs across cell states (CMP, GMP, HSC, MEP), with the median of all assigned probabilities for all peaks plotted.)

5.3 Deep Learning Models, Training on ChIP-seq Data and Data Augmentation

First GENOSTAN genomic annotations were divided into 250 base pair windows, and the signal was aggregated. Training the CNN model yielded an R^2 of 0.52 in CMP while 0.37 in GMP. The MEP cell state was temporarily discarded due to some erroneous data points. Augmenting the dataset with reverse complement sequences It was clear from the predictions that the model was unable to identify sites where TFs were binding due to a major class imbalance in the dataset; from approximately 800,000 sequences, only about 20,000 contained peaks. Removing 85 percent of unbound sequences resulted in a performance boost of 0.05 R^2, which pointed to the fact that the dataset needed to be augmented. Besides removing a significant portion of the non-binding regions, the reverse complement of all sequencess were added with the same signal/ enrichment value, doublign the dataset.

5.3.1 Architectural Overview

One of the first Architectures to be used was the DEEPSTARR architecture, whcih was trained on a similar task of 249 base-pair sequences with STARR-Seq data, with two heads.

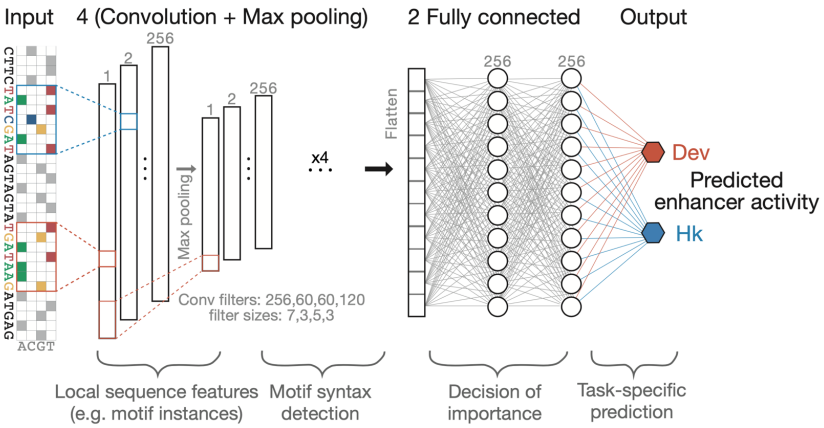

Figure 19: Deepstarr Architecture

DEEPSTARR did not result in good predictions, which could be due to amultitude of reasons, which haven't been pinned fown yet.

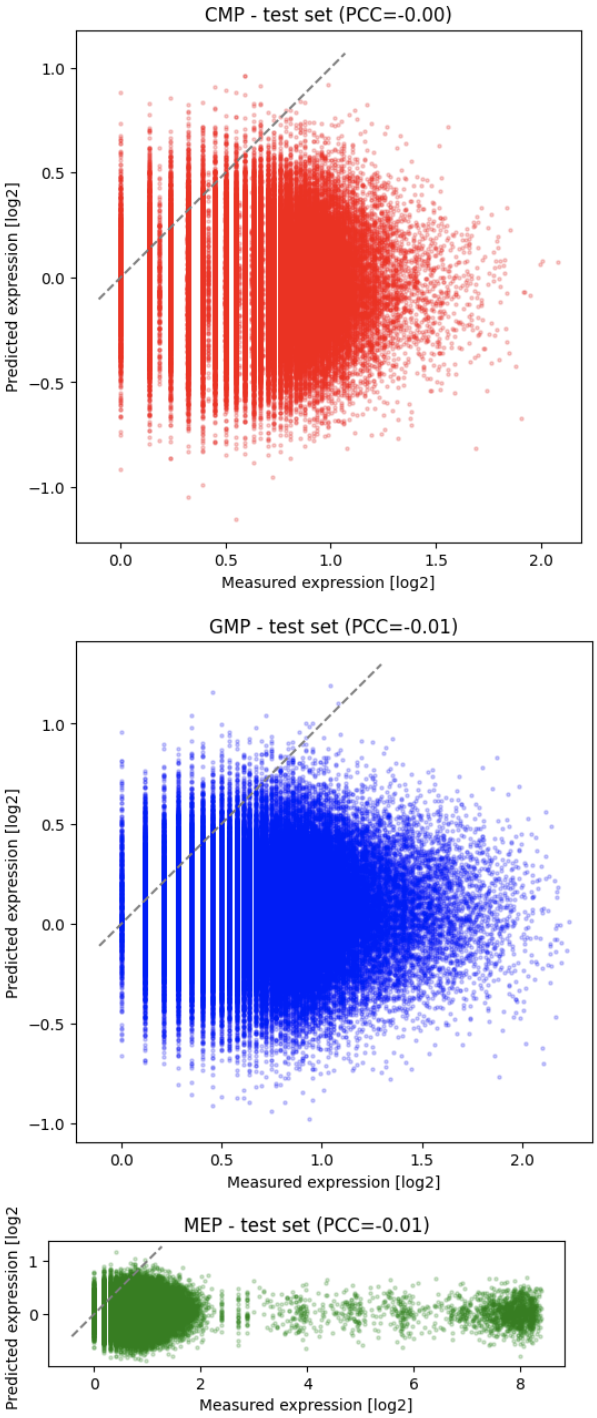

Figure 20: Deepstarr predictions

Taking inspiration from BPNet, we eventually moved from predicting binding in 2 cell states to 2 TFs in the same cell state, hoping to perform post-hoc interpertability analysis post-training.

```
==========================================================================================
Layer (type:depth-idx)                             Output Shape              Param #
==========================================================================================
PrixFixeNet                                        [1, 1]                    --
BHIFirstLayersBlock: 1-1                           --                        --
    ModuleList: 2-1                                --                        --
        ConvBlock: 3-1                             [1, 160, 249]             7,360
        ConvBlock: 3-2                             [1, 160, 249]             12,160
AutosomeCoreBlock: 1-2                             --                        --
```

```
    ModuleDict: 2-2                        --                        --
        Sequential: 3-3                    [1, 320, 249]             420,048
        Sequential: 3-4                    [1, 128, 249]             573,696
        Sequential: 3-5                    [1, 128, 249]             173,856
        Sequential: 3-6                    [1, 128, 249]             229,632
        Sequential: 3-7                    [1, 128, 249]             87,072
        Sequential: 3-8                    [1, 64, 249]              114,816
        Sequential: 3-9                    [1, 64, 249]              45,968
        Sequential: 3-10                   [1, 64, 249]              57,472
        Sequential: 3-11                   [1, 64, 249]              45,968
        Sequential: 3-12                   [1, 64, 249]              57,472
        Sequential: 3-13                   [1, 64, 249]              45,968
        Sequential: 3-14                   [1, 64, 249]              57,472
AutosomeFinalLayersBlock: 1-3              --                        --
    Conv1d: 2-3                            [1, 256, 249]             16,640
    Sequential: 2-4                        [1, 1]                    --
        Linear: 3-15                       [1, 1]                    257
    Conv1d: 2-5                            [1, 256, 249]             16,640
    Sequential: 2-6                        [1, 1]                    --
        Linear: 3-16                       [1, 1]                    257
==========================================================================================
Total params: 1,962,754
Trainable params: 1,962,754
Non-trainable params: 0
Total mult-adds (M): 446.78
==========================================================================================
Input size (MB): 0.01
Forward/backward pass size (MB): 23.09
Params size (MB): 7.85
Estimated Total Size (MB): 30.95
==========================================================================================
```

Finally, a CNN based architecture was used, and details of other architectures which were tried can be found in the appendix.The CNN-based model can be seen below. Performance of the model on the test set and its progression are shown below:

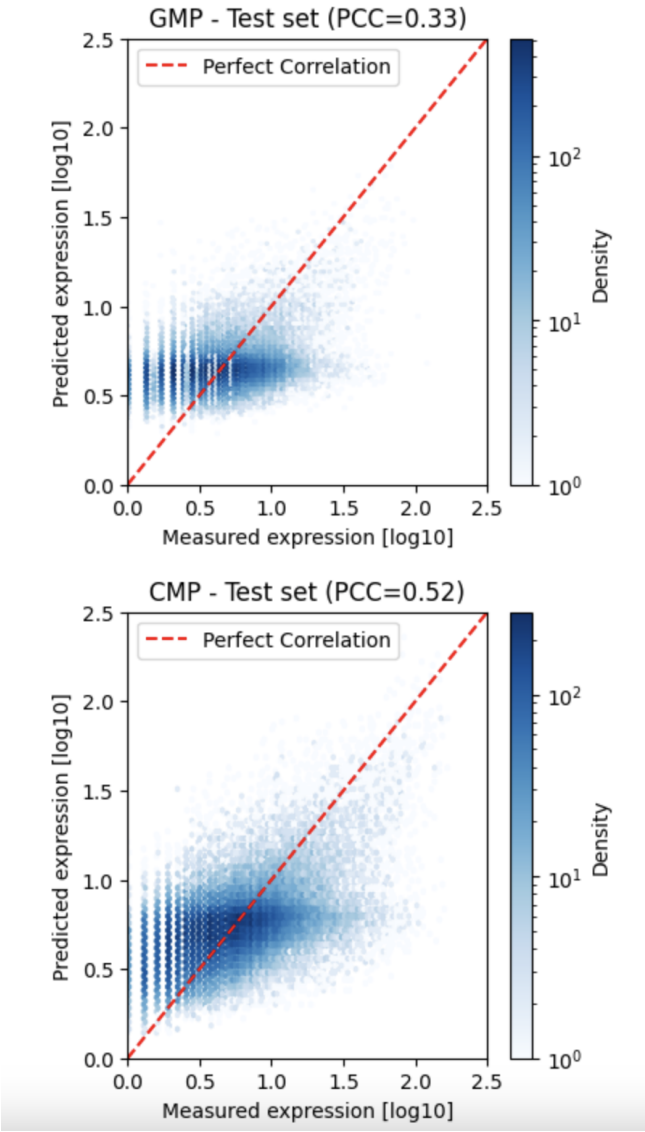

Figure 21: First Predictions on GENOSTAN annotations of enhancer promoter regions

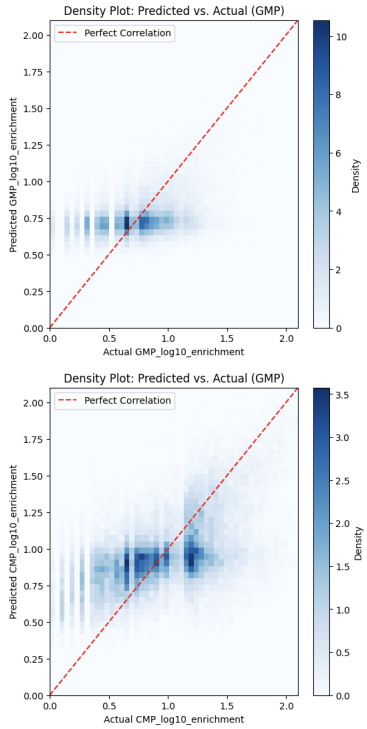

Figure 22: Removing 85 percent of the non-bound regions, using a better data preparation script

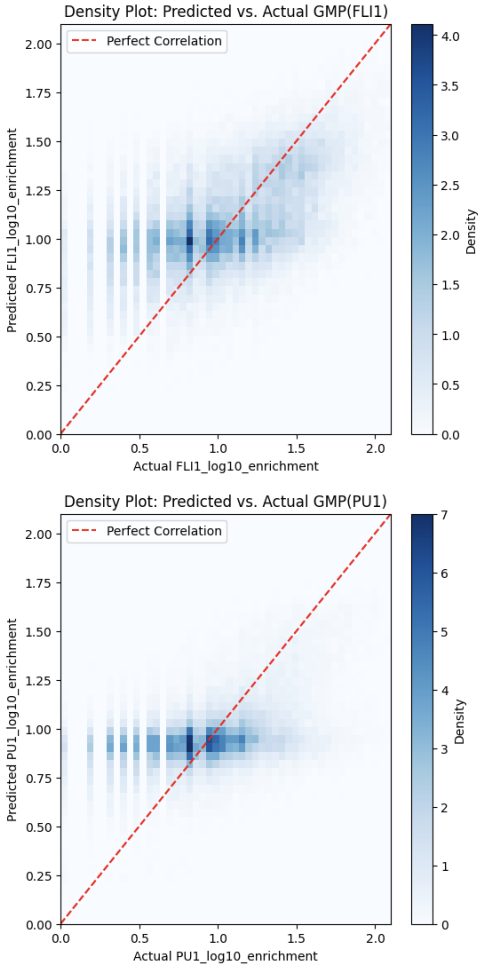

Figure 23: Trying different architectures, hyperparamter tuning

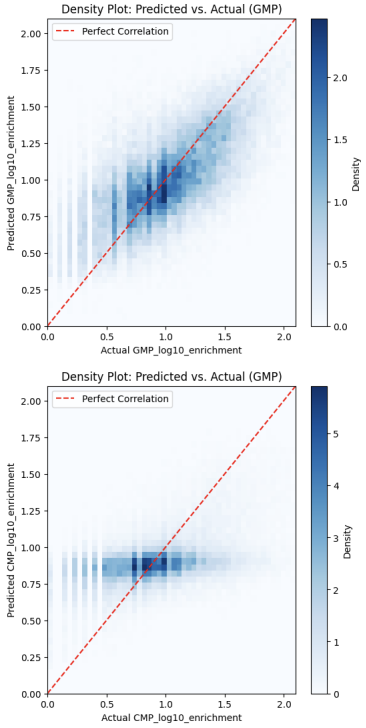

Figure 24: Augmenting the dataset with reverse compliment sequences, early stopping, prdictions for FLI1, GATA2

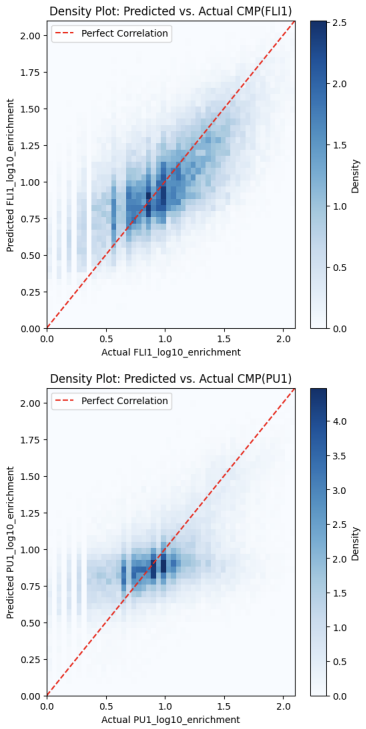

Figure 25: Reaching a PCC on par with DEEPSTARR (0.67) predictions for FLI binding in CMP

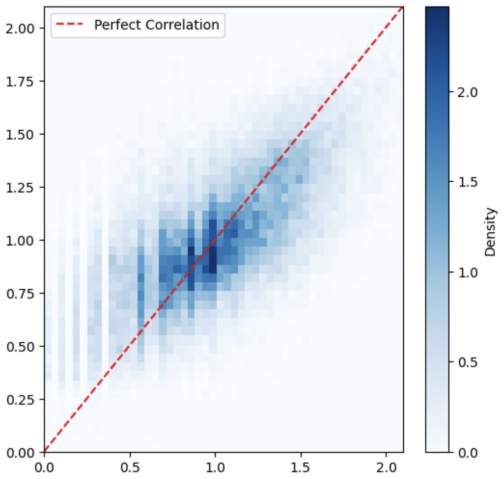

Figure 26: Final Best Predictions in CMP for FLI1

5.3.2 Prediction of Binding on the MPRA Sequences

For predictions on the synthetic enhancer sequences, Trimming of MPRA sequences from 262 bp to 250 bp was done The trained models were used to make predictions of transcription factor binding across all synthetic sequences, instead of the test set.

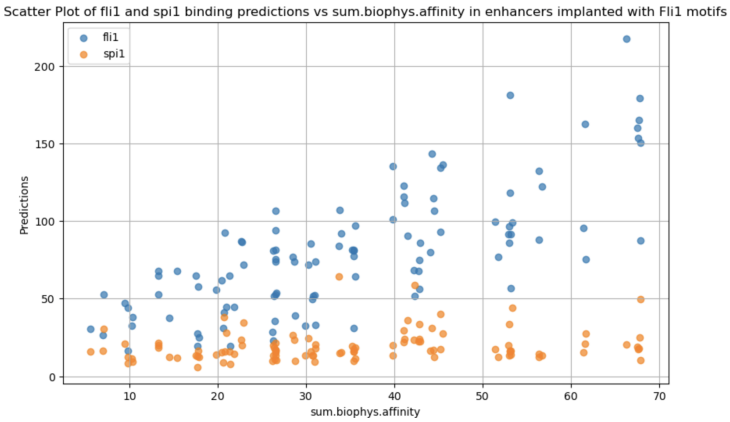

Figure 27: Predictions of the deep learning model, trained on CMP Chip-Seq data on MPRA enhancer sse-quences, implanted with FLI1 Motifs

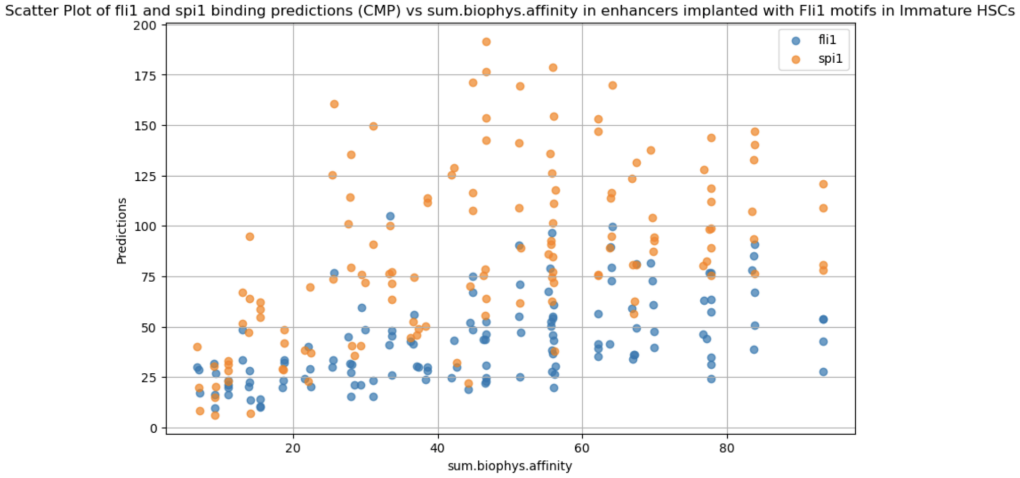

Figure 28: Predictions of the deep learning model, trained on CMP Chip-Seq data on MPRA enhancer sse-quences, implanted with Spi1 Motifs

5.4 Mathematical Modeling

5.4.1 Initial Mathematical Model for binding

The initial models considered are described below.

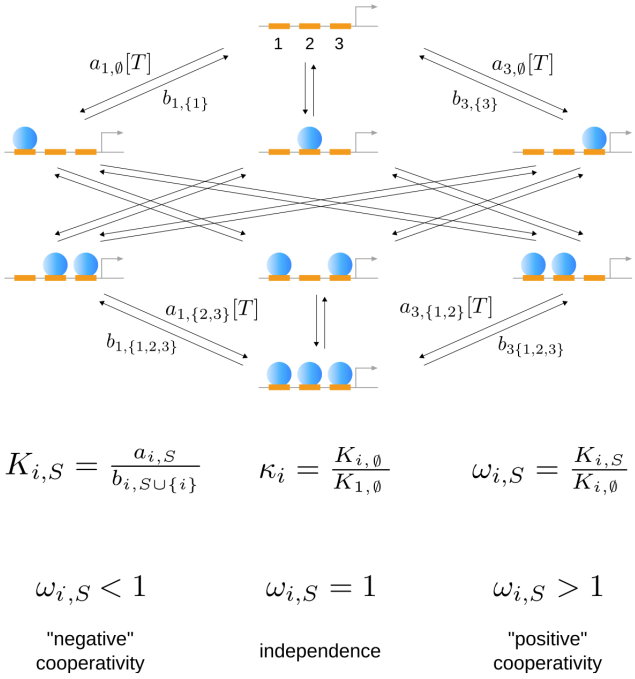

Figure 29: A representative graph displays eight microstates along with several labeled, directed transitions. In the second row, the key thermodynamic equilibrium parameters are shown. The higher-order cooperativity parameter, $w_{i,S}, quantifies how the binding of a factor T to site i is affected by the prior binding of T to a set of sites S; that is, it indicates whe$ $rate a_{i;S}, where i denotes the binding site and S represents the subset of sites (denoted as i_1, \ldots, i_k) already occupied by T. Similar$ $rate b_{i;S'}, where S' is the subset of sites bound by T and i is one of these sites. In the system to which this model is applied, two bindin$

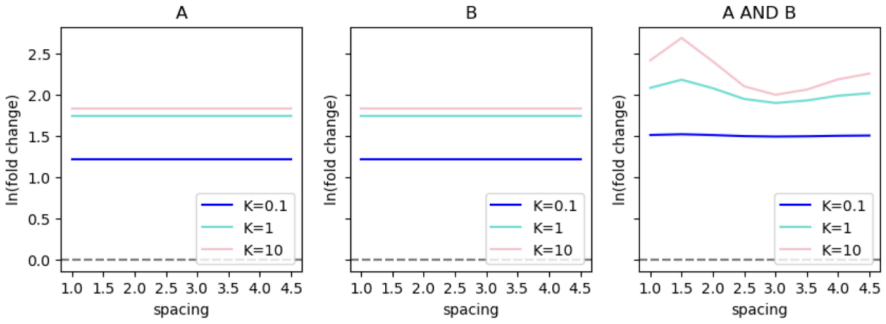

Figure 30: The result of a mathematical model where the y axis is log fold change and the x axis is spacing. Two TFs A and B interact with co-activators(S) and co-repressors(R). S helps Polymerase recruitment while R prevents it. The three sites from figure 5 are occupied by A,B and Polymerase P.

5.4.2 Transition to Deep Mechanistic Modeling

Given the vast complexity of moecular interactions between co-factors, chromatin, dna, TFs etc, we decided to switch to the deep learning model to predict binding. However, by constraining the binding in the model, we could still emplpoy the mathematical framework for biophysical modelling to infer effects of TFs and associated cofactors on other rate limiting steps of the transcriptional cycle, for which data isn't available.

5.5 Development of a Biophysical Model for Transcription Regulation

5.5.1 Model Formulations

Interactions between pairs of TFs can be additive or multiplicative in reducing or increasing the rate of a step leading to transcription. Interactions in the form of other functional interactions are also possible, but weren't considered for analysis.

Multiplicative case

The standard Arrhenius equation for the rate constant is given by:

$$k = A \exp\left(-\frac{E_a}{RT}\right),$$

where

- E_a is the energy of activation to go from an initial to a final state,
- R is the gas constant,
- T is temperature in Kelvin, and
- A is the constant, pre-exponential factor.

The presence of two transcription factors may reduce the activation energy additively:

$$E_a^{TF} = E_a - \Delta g_1 - \Delta g_2.$$

Substituting this modified activation energy into the Arrhenius equation yields:

$$k_{TF} = A \exp\left(-\frac{E_a - \Delta g_1 - \Delta g_2}{RT}\right).$$

This expression can also be written as:

$$k_{TF} = A \exp\left(-\frac{E_a}{RT}\right) \exp\left(\frac{\Delta g_1}{RT}\right) \exp\left(\frac{\Delta g_2}{RT}\right).$$

Defining

$$k_0 = A \exp\left(-\frac{E_a}{RT}\right),$$

the rate constant in the presence of both transcription factors becomes:

$$k_{TF} = k_0 \exp\left(\frac{\Delta g_1}{RT}\right) \exp\left(\frac{\Delta g_2}{RT}\right).$$

Addditive Case

$$f(x; K, n) = \frac{x^n}{K^n + x^n}$$

$$X_A = delay_{A,max,coop} \cdot f(A; K_{A,coop}, n_{A,coop})$$

$$X_B = delay_{B,max,coop} \cdot f(B; K_{B,coop}, n_{B,coop})$$

$$t_{coop} = T_1 - (X_A + X_B)$$

$$k_1 = \frac{1}{t_{coop}} = \frac{1}{T_1 - (X_A + X_B)}$$

$$X'_A = delay_{A',anti} \cdot f(A; K_{A',anti}, n_{A',anti})$$

$$X'_B = delay_{B',anti} \cdot f(B; K_{B',anti}, n_{B',anti})$$

$$t_{anti} = T_2 + (X'_A + X'_B)$$

$$k_2 = \frac{1}{t_{anti}} = \frac{1}{T_2 + (X'_A + X'_B)}$$

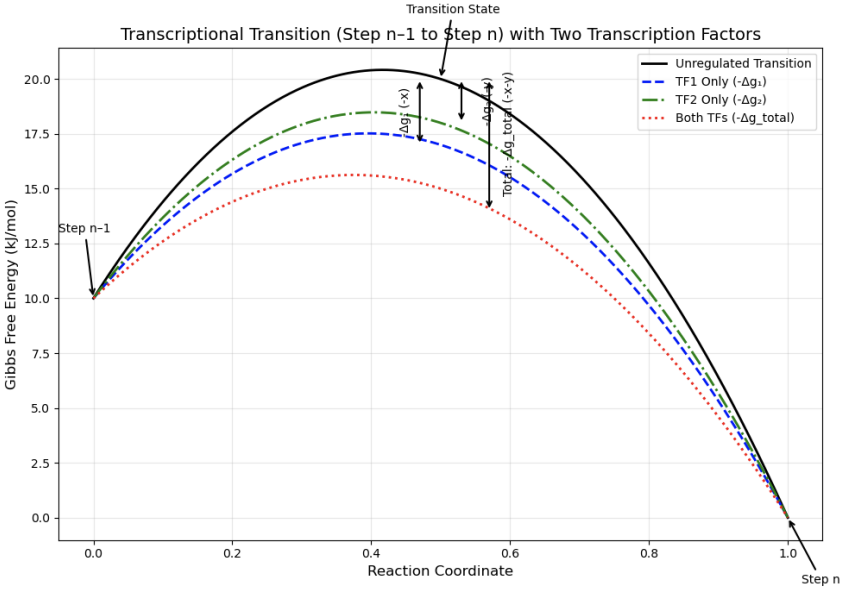

Figure 31: Pairs of TFs lowering the actiavtion energy required to move from one kientic step to another

$$x_{obs} = \frac{k_3\, k_1\, k_2}{k_m\, k_3 + k_2\, k_3 + k_1\, k_3 + k_1\, k_2}$$

This may be derived from the procedure outlined in methods.

$$LFC = \ln\left(\frac{x_{obs}}{x_{basal}}\right)$$

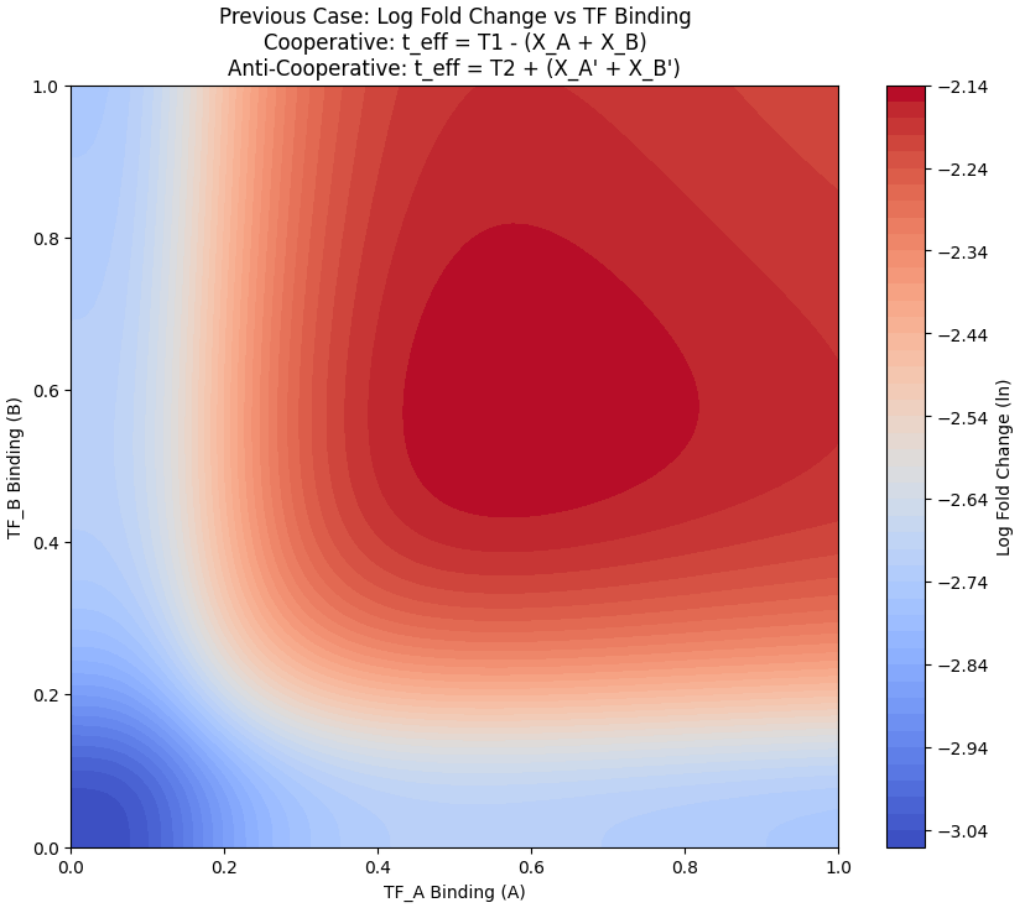

Figure 32: Baseline cooperative time (in seconds) $T_1 = 16.0$; Delays (which subtract from the available time): `delay_A_max_coop` $= 5.9$ (maximum delay for TF_A in cooperative pathway); `delay_B_max_coop` $= 5.9$ (maximum delay for TF_B in cooperative pathway); Hill parameters for cooperative delays: $K_{A,coop} = 0.2$, $n_{A,coop} = 3$, $K_{B,coop} = 0.2$, $n_{B,coop} = 3$; Parameters for the Anti-Cooperative/desaturating Pathway (k): Baseline anti-cooperative time $T_2 = 2.0$; `delay_A_prime_anti` $= 1.0$ (maximum delay for TF_A in anti-cooperative pathway); `delay_B_prime_anti` $= 1.0$ (maximum delay for TF_B in anti-cooperative pathway); Hill parameters for anti-cooperative delays: $K_{A,prime,anti} = 0.8$, $n_{A,prime,anti} = 3$, $K_{B,prime,anti} = 0.8$, $n_{B,prime,anti} = 3$; $k_3 = 10.0$, $k_{m1} = 0.1$, $x_{basal} = 1$

5.6 Parameter Fitting of the Biophysical Model

Finally, we planned to use the `scipy-optimize` library to fit model parameters, using bindig predictions as an inout and fitting the final expression as an ouput, fitted through the mechanistic model. However, results for this section aren't available currently.

6 Discussion and Future Directions

Massively parallel reporter assays in an ex-vivo differentiation system present a powerful framework for decoding cell-state specific cis-regulatory logic, especially in the context of lineage specification. Novel cell-state dependent behaviour between transcription factor pairs and single factors was observed in our analysis for factor pairs like SPi1-FLi1, Cebpa-GATA2 (switches from activation to repression depedning on the cell state and level of binding of either TF) and single factors like Myc and Creb1. Additionally, sinusoidal spacing depedent behaviour was evident for Spi1 and Fli1 pairs. Interestingly, this happend at a 10 base pair interval alluding to the helical twist of the DNA. While it is known that protein-protein interactions aren't the only driving fcators of RNAP recruitment between TF pairs, similar behaviour was also seen in Z. AVsec et al 2021 Nat. Genetics. Moreover, the same paper shows that motifs for composite binding of two TFs might have a pair dependent spacer motif between the two motifs for the 2 TFs as well, which increased spacing might interfere with.

The design for the sequences involved inside the MPRA experiment crucially had random DNA implanted between the different motifs for transcription factors. Through information theoretic and bioinformatics analysis eukaryotic genomes, it was shown that eukaryotic TFs lack sufficient specificity to uniquely specify target genes

(Wunderlich Mirny, 2009). Unlike bacterial TF motifs that can be 15-20 bp long, eukaryotic TF motifs can be as short as 6-7 bp long and a lot of them are degenerate. The effects of random DNA between implanted motifs can be significant, and a deep binding model may capture these features, while a simple, motif scanning convolutional neural network might not. Additionally, our choice in deep learning as opposed to mathematical modelling for prediction of binding relies on the immense predictive capabilites of deep l=models. Deep learning models now seem to generalise predictions across different assays and not just organisms, for example MPRA data trained models could generalise well on ATAC-seq data for yeast (M. Rafi et al 2024 Nature Biotech). Chip-seq data was chosen since it provides in-vivo predictions of binding. Moreover, it was available in 4 cell states, for 7 important transcription factors. Since each "read" in a chip-seq experiment acts as an "event" of binding, different representations of the data can be used to build deep learning model. Most importantly, raw reads should be used to train deep learning models, without normalisation, since the model loses information on normalised sequencing data. Sum of the reads that fall in each base, or binning the results over a window of a genome (typically 50-100 bp), and taking the middle "base" of each read, mapped across the genome are some ways of representing the data.

A pioneering study from Bell et al., 2024 shows that TFs do not have a single way of driving activation. Instead, a unique combination of cofactors, targeting distinct steps in the transcriptional cycle in a TF-specific manner, are necessary. The authors also measure the influence of core promoters in their system, cataloging them as either initiation or pause-release sensitive. This is a highly impactful paper when considering the mathematical modelling that has been carried out over the course of this thesis. Additionally, Our analysis using the algorithm to detect cooperativity between heterotypic pairs of transcription factors reveals that Cooperativity is likely highly cell state dependent. This confirms, or alludes to the fact that co-factors might be very important cell-state specific regulators of not just different effects of the TFs on steps of the transcriptional cycle, but also at the level of binding, especially in the hematopoietic system with the heptad of transcription factors.

Moving on to using the chip-seq data for training the deep learning model, for the purposes of this thesis, binned results were used since chip-seq data was available at lower resolution (50 bp bins) as opposed to 28 bp in Chip-nexus which made it difficult to make base-pair resolution predictions. Moreover, the subsequent treatment of base pair resolution predictions for a mechanistic model would have lead to more complications, which is why it was decided against said approach. Data quality is an important consideration before training deep learning models on Chip-seq data. The ChIP-seq sample should not be too sparse and the regions need to be selected in a way that a model can learn the patterns. Models seem to require require ¿15K-10K peaks to perform well. However, selection of the data to predict binding was also an important,albeit not fully optimized, step. Ideally, one would use the deeptools package multibigwigsummary to create a minimal set of regions for all transcription factors in all 4 cell states using Principal Component analysis. Moreover, a sliding window could also be employed to increase the dataset by 3-4 fold. Additionally, major performance boosts for the deep learning model seemed to come from better data preparation, instead of changing the model architecture in our case. TFs with the highest number of peaks (Fli1 or PU1) could be trained better than all other factors. Longer sequences can also be considered for training the deep learning model, since transcription factors and enhancers can interact with each other in sequence lengths much greater than 250 base pairs.

For training the deep learning model, multiple architectures were tried out over the course of thesis, and it may be concluded that for shorter sequences, multiple model architectures may provide similar performance results. Initially, one dimensional convolution filters followed by an attention based architecture was tried out to capture interactions between identified motifs, following similar attempts in literature (vaishnav et al 2020), even for short sequences (80 base pairs) (M. Thomson et al 2024) . However a CNN based architecture followed by multiple linear units outperformed all other architectures. For the current architecture for binding predictions on Chip-see data, we aim to add a layer in the neural network, which scales the binding term by the concentration of a transcription factor. Moreover, interactions between different bound TFs can also be incorporated, like done for example in (Liu et al 2020), to make the model more mechanistic. Moreover, additional layers that incorporate effects of the bound transcription factors on the transcription cycle need to be included. However, we'd aim to build a model that does not hard-code these interactions, rather learns them while training the data to offer mechanistic insight while simultaneous training on chip-seq and MPRA data through a Multi-Task learning strategy. One major flaw in our analysis is that the Chip-Seq cell states were determined from Bulk RNA-Seq while the data from the MPRA dataset characterised cell states using single cell RNA Seq. Therefore, a proper mapping of the cell states is not possible.

Non-monotonic responses in biology may be very important, for example in order to maintain homeostasis. Recently, in the context of transcriptional regulation, it has been shown theoretically that non-monotonocity is possible when considering models at non-equilibrium. One approach (Mahdavi et al 2024) considers the impacts of energy dissipation while another (Martínez Corral et al 2024) considers the effects of non-monotonicity tuned by affinity, while having distinct functional effects on the rate-limiting steps of the transcriptional cycle.

While non-monotonicity at the level of binding when increasing the number of motifs and their binding affinities is admittedly lacks evidence from our analysis, it is certainly plausible. Through a deep learning

model that is trained on ChIP-seq data, the binding predictions of the transcripting factor SPI1 or PU.1 first increase and then only seem to saturate, as opposed to being non-monotonic. Indeed, non-monotonicity in expression can arise by different molecular contexts or interactions other than at the level of transcription factor binding, which have also been investigated through the use of mathematical and biophysical models in the thesis. However, it has been proposed that macromolecular complexes and crowding may be one of the reasons for this non-monotonic regulation of transcription. (Matsuda et al 2014). Therefore, we hypothesised that non-monotonocity may arise from Transcription factors mediating different rate-limiting steps inchoerently. Whille we were able to see interesting and a wide range of behvaiour governed by the equations described in the methods section for transcription factors, an important unfinished step was to fit the biophysical model with the MPRA expression data, taking as an input the binding predictions from the model trained on Chip-Seq Data. Given known cell-state mappings from single cell experiments, a multi task leanring stratgey with simultaneous training of both datasets would be the ideal approach, where the binding model would be a black box, while the expression model would comprise of a black box as well as a mechanistic part.

What we are attempting is line with one of the most interesting trends in computational biology is the growing importance of inductive bias in AI models. Inductive biases are the assumptions and prior knowledge incorporated into the model by the scientist to help the model make predictions and generalize beyond (often incomplete) training data. Integrating biologically informed inductive biases help fine-tune the model for domain-specific applications. for example, biological systems often display hierarchical organization (e.g., TF binding and pause release). One way to introduce inductive bias is to use a deep learning model with a hierarchical architecture, such as a multi-layer neural network or a graph convolutional network, which would learn the hierarchical structure of transcription, with each layer corresponding to a different level of the hierarchy. On the other hand, theoretical and analytical considerations might help us to constrain the prediction space of the model. Combined, we might be closer than ever to having forward theory and experimental validation with models having immense predictive power of how biological 'systems' behave, not just molecules.

7 Supplementary Information

7.1 Alternative Architectures Used

7.1.1 Attention-based Architecture

The following architecture resulted in slightly lower accuracy with a Pearson Correlation Coefficient of 0.417 and 0.564 for cell states CMP and GMP respectively, as compared to the CNN based architecture on the same dataset.

```
==========================================================================================
Layer (type:depth-idx)                        Output Shape              Param #
==========================================================================================
PrixFixeNet                                   [1, 1]                    --
BHIFirstLayersBlock: 1-1                      --                        --
    ModuleList: 2-1                           --                        --
        ConvBlock: 3-1                        [1, 160, 250]             7,360
        ConvBlock: 3-2                        [1, 160, 250]             12,160
AutosomeCoreBlock: 1-2                        --                        --
    ModuleDict: 2-2                           --                        --
        Sequential: 3-3                       [1, 320, 250]             420,048
        Sequential: 3-4                       [1, 128, 250]             573,696
        Sequential: 3-5                       [1, 128, 250]             173,856
        Sequential: 3-6                       [1, 128, 250]             229,632
        Sequential: 3-7                       [1, 128, 250]             87,072
        Sequential: 3-8                       [1, 64, 250]              114,816
        Sequential: 3-9                       [1, 64, 250]              45,968
        Sequential: 3-10                      [1, 64, 250]              57,472
        Sequential: 3-11                      [1, 64, 250]              45,968
        Sequential: 3-12                      [1, 64, 250]              57,472
        Sequential: 3-13                      [1, 64, 250]              45,968
        Sequential: 3-14                      [1, 64, 250]              57,472
AutosomeFinalLayersBlock: 1-3                 --                        --
    Conv1d: 2-3                               [1, 256, 250]             16,640
    Sequential: 2-4                           [1, 1]                    --
        Linear: 3-15                          [1, 1]                    257
```

```
    Conv1d: 2-5                              [1, 256, 250]             16,640
    Sequential: 2-6                          [1, 1]                    --
        Linear: 3-16                        [1, 1]                    257
==========================================================================================
Total params: 1,962,754
Trainable params: 1,962,754
Non-trainable params: 0
Total mult-adds (M): 448.57
==========================================================================================
Input size (MB): 0.01
Forward/backward pass size (MB): 23.19
Params size (MB): 7.85
Estimated Total Size (MB): 31.04
==========================================================================================
```

7.1.2 RNN-based Architecture

The following architecture resulted in slightly lower accuracy with a Pearson Correlation Coefficient of 0.410 and 0.561 for cell states CMP and GMP respectively, as compared to the CNN based architecture on the same dataset.

```
==========================================================================================
Layer (type:depth-idx)                        Output Shape              Param #
==========================================================================================
PrixFixeNet                                   [1, 1]                    --
BHIFirstLayersBlock: 1-1                      --                        --
    ModuleList: 2-1                           --                        --
        ConvBlock: 3-1                       [1, 160, 250]             7,360
        ConvBlock: 3-2                       [1, 160, 250]             12,160
AutosomeCoreBlock: 1-2                        --                        --
    ModuleDict: 2-2                           --                        --
        Sequential: 3-3                      [1, 320, 250]             420,048
        Sequential: 3-4                      [1, 128, 250]             573,696
        Sequential: 3-5                      [1, 128, 250]             173,856
        Sequential: 3-6                      [1, 128, 250]             229,632
        Sequential: 3-7                      [1, 128, 250]             87,072
        Sequential: 3-8                      [1, 64, 250]              114,816
        Sequential: 3-9                      [1, 64, 250]              45,968
        Sequential: 3-10                     [1, 64, 250]              57,472
        Sequential: 3-11                     [1, 64, 250]              45,968
        Sequential: 3-12                     [1, 64, 250]              57,472
        Sequential: 3-13                     [1, 64, 250]              45,968
        Sequential: 3-14                     [1, 64, 250]              57,472
AutosomeFinalLayersBlock: 1-3                 --                        --
    Conv1d: 2-3                               [1, 256, 250]             16,640
    Sequential: 2-4                           [1, 1]                    --
        Linear: 3-15                         [1, 1]                    257
    Conv1d: 2-5                               [1, 256, 250]             16,640
    Sequential: 2-6                           [1, 1]                    --
        Linear: 3-16                         [1, 1]                    257
==========================================================================================
Total params: 1,962,754
Trainable params: 1,962,754
Non-trainable params: 0
Total mult-adds (M): 448.57
==========================================================================================
Input size (MB): 0.01
Forward/backward pass size (MB): 23.19
Params size (MB): 7.85
Estimated Total Size (MB): 31.04
==========================================================================================
```

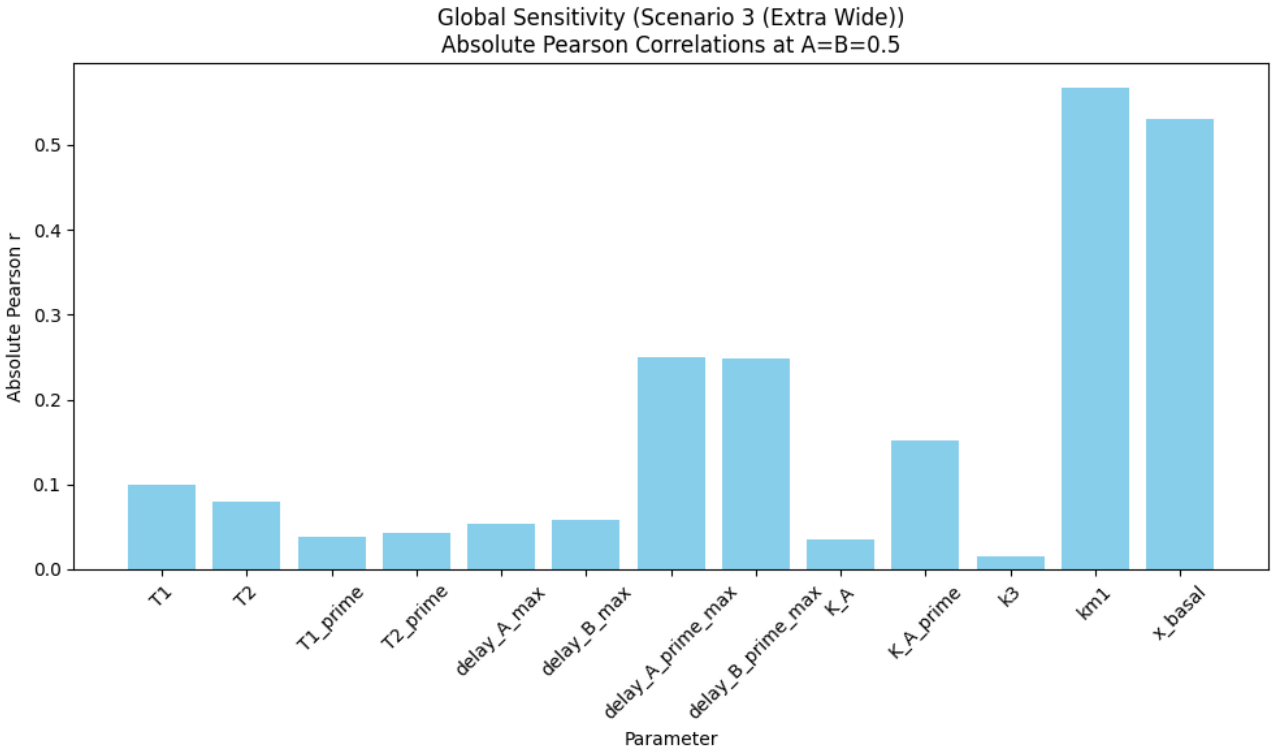

Figure 33: Global sensitivity analysis for the following parameter ranges in the mathematical model: 'T1': (60.0, 80.0), 'T2': (60.0, 80.0), '$T1_prime'$: $(15.0, 25.0)$,'$T2_prime'$: $(15.0, 25.0)$,'$delay_Amax'$: $(2.0, 30.0)$,'$delay_Bmax'$: $(2.0, 30.0)$,'$delay_Aprime_max'$: $(2.0, 1000.0)$,'$delay_Bprime_max'$: $(2.0, 1000.0)$,'K'_A : $(0.4, 0.6)$,'K_Aprime' : $(0.4, 0.6)$,'$k3'$: $(0.01, 100.0)$,'$km1'$: $(0.01, 100.0)$,'x_basal' : $(0.01, 1.0)$

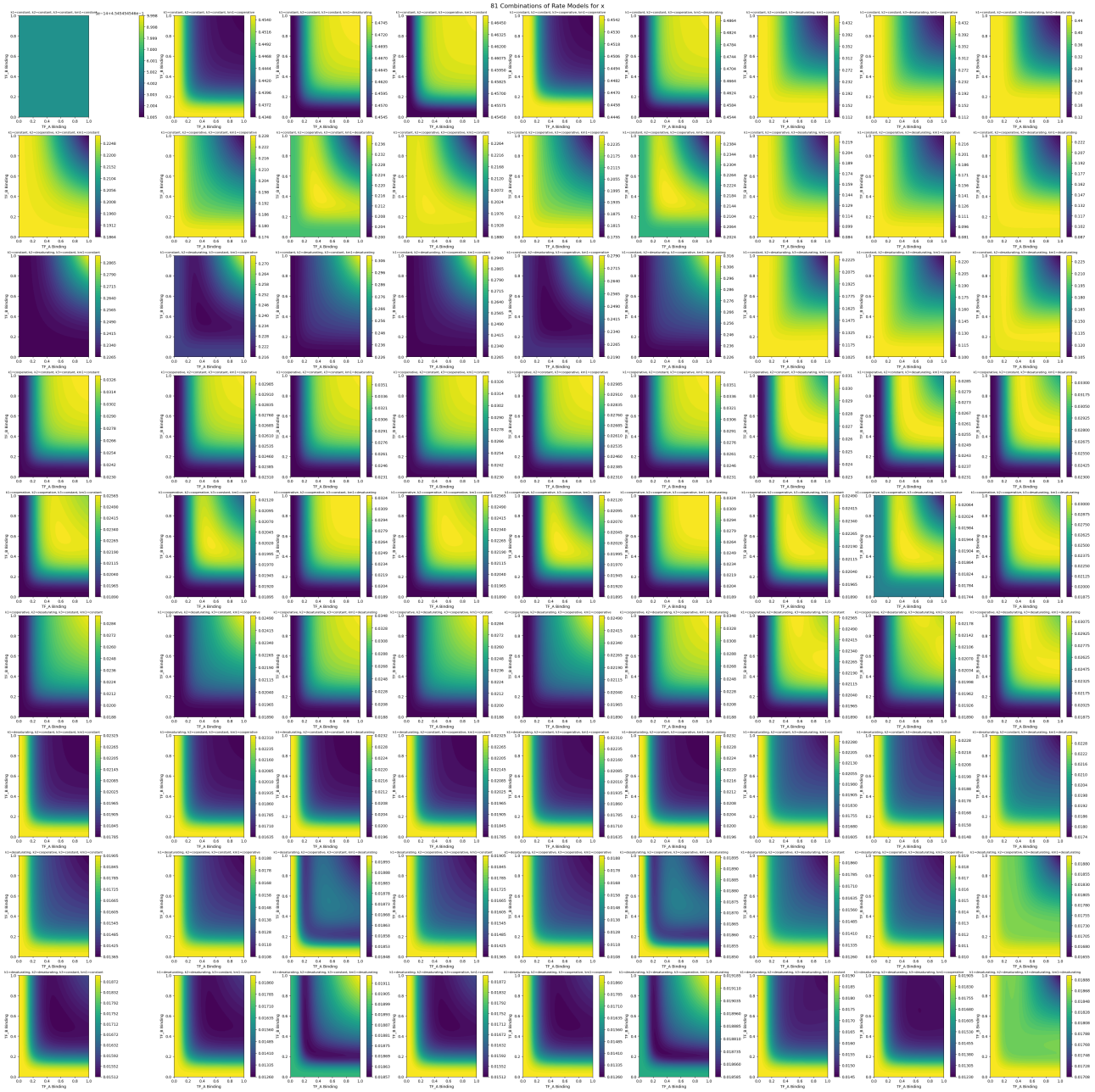

Figure 34: Mathemtical and Parameter space exploration of the mathematical model built. All for 4 kinetic rates were allowed to chage in three ways- remain constant, increase cooperatively, decrease anticooperatively both in the presence of 2 TFs. A wide range of output behvaiour is observed.

7.2 Additional Data Visualisation of the MPRA dataset

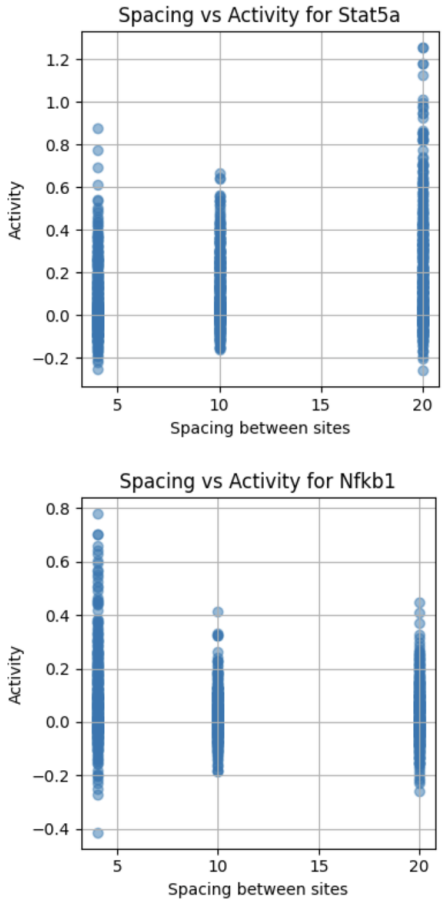

Figure 35:

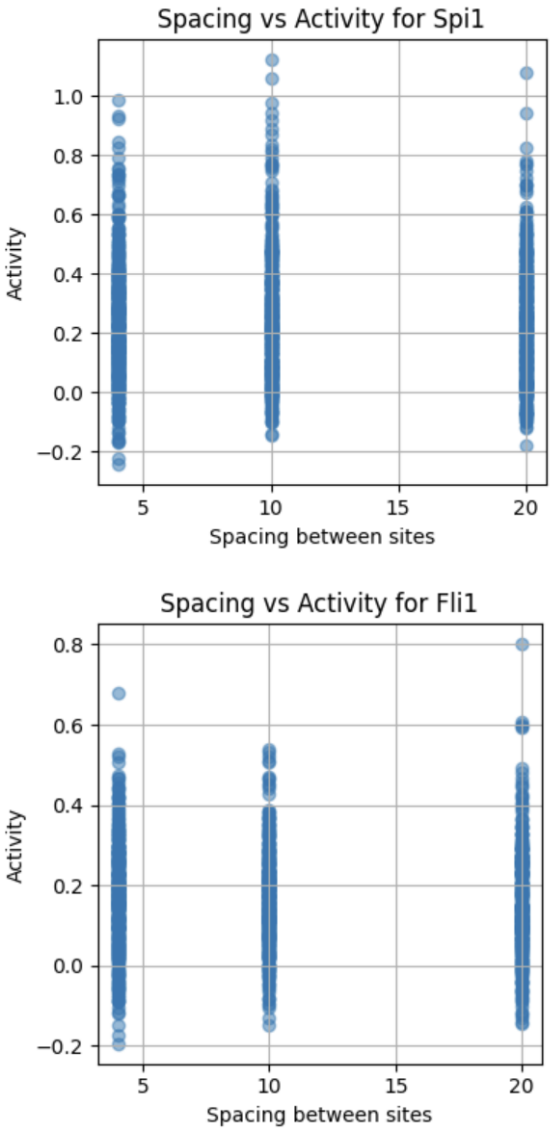

Figure 36:

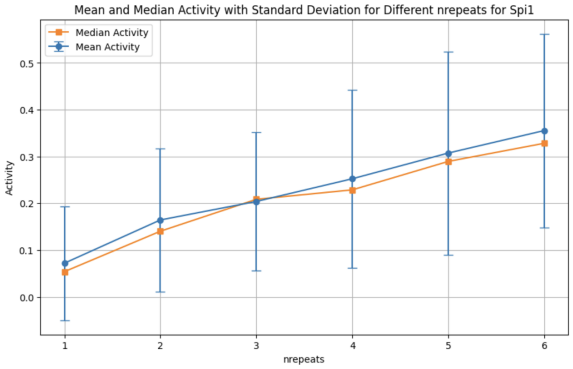

Figure 37:

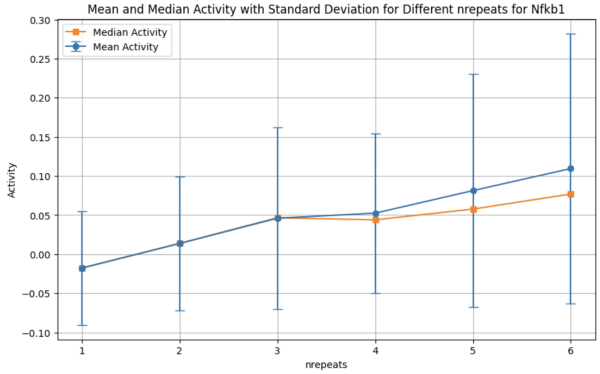

Figure 38:

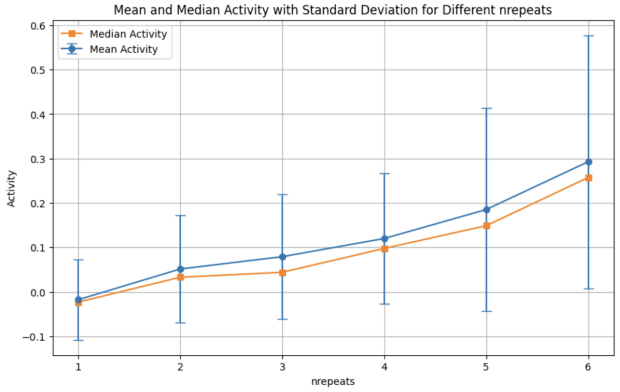

Figure 39:

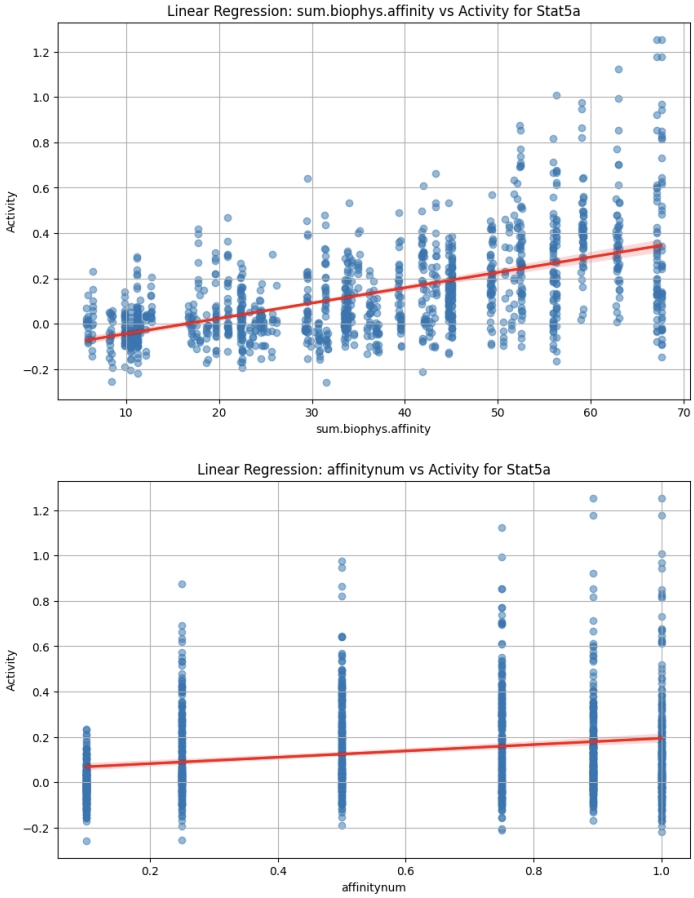

Figure 40:

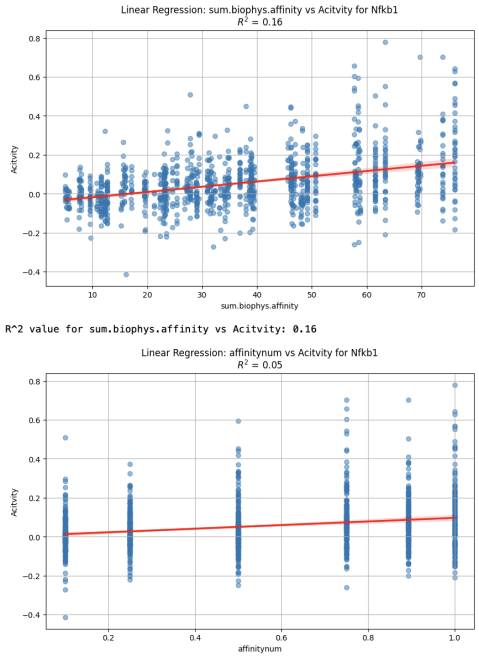

Figure 41:

7.3 Analysis of GENOSTAN annotations with Chip-Seq peak regions

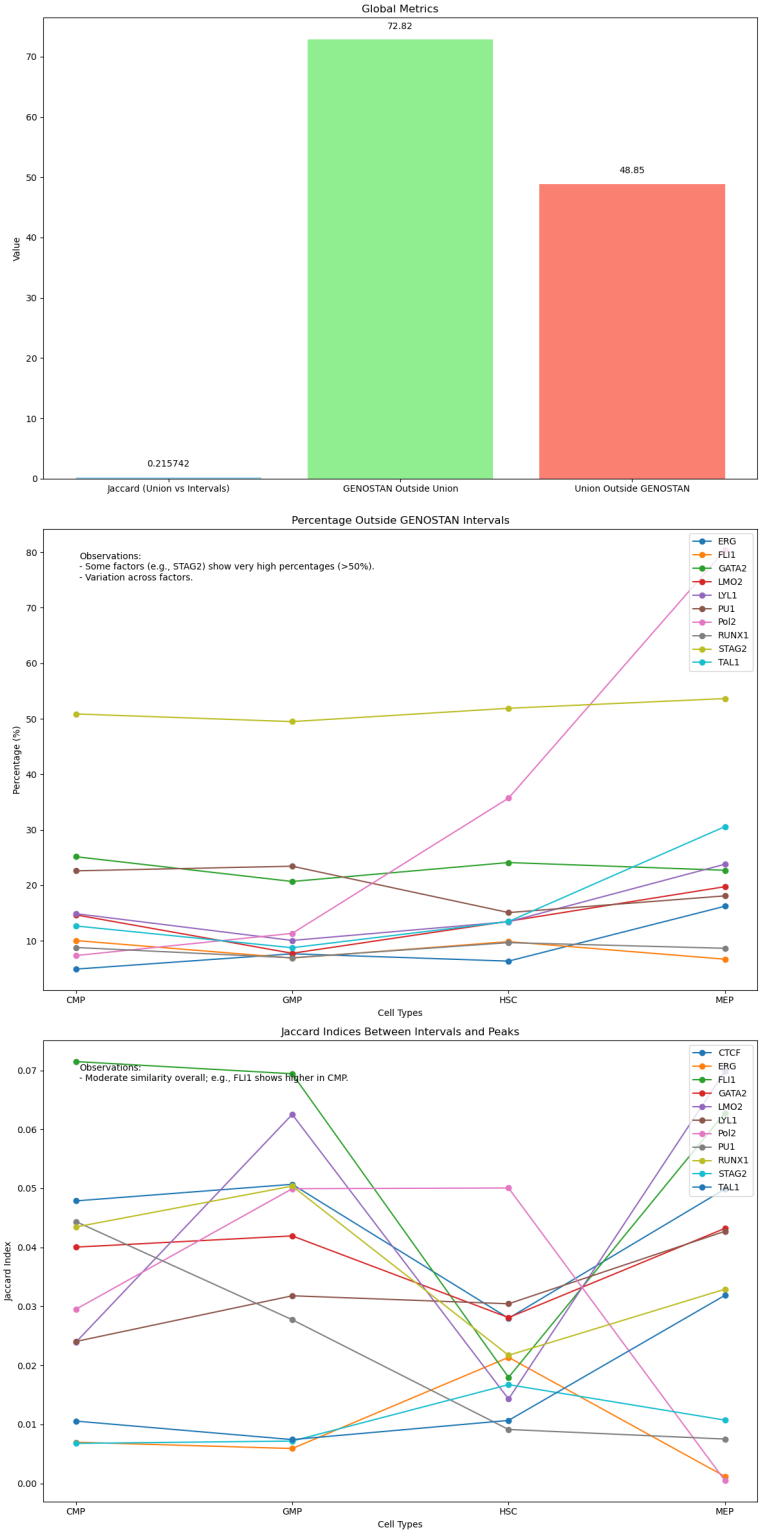

Figure 42: GENOSTAN Comparisons: (Top) Global metrics comparing GENOSTAN intervals with the union of all peaks; (Middle) Percentage of GENOSTAN intervals outside the peak union; (Bottom) Jaccard indices between intervals and peaks.

8 Appendix

A breif description of the unique values of the MPRA dataset taken from R. Froemel et al 2023 is given below for all libraries:

```
Unique values in column 'Unnamed: 0':
[    1     2     3 ... 43671 43672 43673]

Unique values in column 'clusterID':
['State_1M' 'State_2D' 'State_3E' 'State_4M' 'State_5M' 'State_6N'
 'State_7M']

Unique values in column 'CRS':
['LibA.Seq1' 'LibA.Seq10' 'LibA.Seq1000' ... 'LibA.Seq9941' 'LibA.Seq9987'
 'LibA.Seq9988']

Unique values in column 'TF':
['Cebpa' 'Elk1' 'Tcf3' 'Tfap2a' 'Trp53' 'Yy1' 'Zbtb7a' 'Fli1' 'Fos'
 'Gata1' 'Gata2' 'Gfi1' 'Gfi1b' 'Cic' 'Ikzf1' 'Irf7' 'Klf1' 'Klf4' 'Lyl1'
 'Mecom' 'Meis1' 'Meis2' 'Myb' 'Myc' 'Creb1' 'Nfix' 'Nfkb1' 'Nfyc' 'Nr2c2'
 'Pbx1' 'Runx1' 'Rxrg' 'Sp1' 'Spdef' 'Spi1' 'Ddit3' 'Stat5a' 'Tcf12']

Unique values in column 'RNA.1':
[ 217  357  411 ... 8250 5434 3250]

Unique values in column 'DNA.1':
[  77   97   49 ... 1275 1018 1001]

Unique values in column 'RNA.norm.1':
[ 32.21754713  53.00306141  61.0203312  ... 425.14737634 319.012154
 255.3310206 ]

Unique values in column 'DNA.norm.1':
[ 60.39845725  76.08636823  38.43538189 ... 770.76474381 540.17584539
 625.5791411 ]

Unique values in column 'RNA.2':
[ 161  297  363 ... 8245 6420 4469]

Unique values in column 'DNA.2':
[  55   69   43 ...  943 1126  904]

Unique values in column 'RNA.norm.2':
[ 30.43084788  56.13640883  68.61116634 ... 414.21637263 387.96322225
  56.88182582]

Unique values in column 'DNA.norm.2':
[ 47.07602264  59.05901022  36.80489043 ... 509.44101701 588.89512058
 591.23200598]
```

Figure 43: Screenshot from 2024-06-27 at 1:06:38 PM

```
Unique values in column 'Library':
['LibA']

Unique values in column 'Seq':
['aggaccggatcaactgctattgtggaacacataaatttcacacaaggtatctccagacctagtctggcatcccggtgtcctactagacctagaaccatattacaacgacctaa
attccatcccaagtaaccatggcaggaaaccaggagtaacgaaatctggaccattcgaaaaaaatgaagttgacttttattgctacggttaggttcactacgagggacggcgtat
aaaggggttacTTGCGCAAcattgcgtgaaccga'
 'aggaccggatcaactgaagttattcgcgaccgaataccttgatcttaccctgaccagtgtaagtgattgcttctctgtgatcatacctcttggcctccacacatagctaaa
ccggaatcacgttggggtatggtaacgctctcttggggtatgacgaaagagggtgctgttgcccactgttcttaggcaccgcaacttcgtaacccagtcggcttatatgtggaaa
ccggcggtacgTTGCGCAAcattgcgtgaaccga'
 'aggaccggatcaactataaaacaaataagtgtctgctttgggatctacgtccaagcttatttgtctcttagctgcgacaaagatggcacccgggacgttgctatttaatccat
ccccttcgtaactatctcacaaccacttgtctagaccgcgcgtcataccctcgttagagaaatctattctttctacagCTAATTCCGGCtcgtTAAGTTCCGGTgcaaCAACTTC
CAGCacgcCGAGTTCCGGTcattgcgtgaaccga'
 ...
 'aggaccggatcaactttgacaattatctcataccacctagctgtcctgttcgagtggcagaagagagcgagcaaatcaggtagaagaacctcagatagctacacgctctacca
taatttagcgaacgatcgctgtttcatacgaacaaacggccgtcccctcgtagaaccttatgtgccacGCAGGTGCGagatgtagcaAACACCTGTgctcaccagtACAGGTGTG
tgagggttccTACACCTGCcattgcgtgaaccga'
 'aggaccggatcaacttagaacgtccccatttagtactcttcgcagcgttgaaaattggtagcgaagtgtgtctaaacgttctagacgccatcaTCACATGTGatgcttgataa
acgcgcaaaTCCACTTGGtacttgggcaagacctagttTCACATGTAttcccgtaatcgcataccccTACATATGGactaaaatctaaacgacggtCCAGATATTtctgtagttc
agcttaggctAACATCTTCcattgcgtgaaccga'
 'aggaccggatcaactcgttagctttttgtatcacaccacgtcggggaaaagtgtccaagcgtaaaccctgaggctataactgattcctacttaaggggtacccaagagaatga
acgcgacaagacgtagggaaactcagttgtagccgtacgttcctttcgggaggatcgtccCAACGTGGAcccaTAGACCTGCaataGCCGCTGATatatAACATCAGAccgcAAA
CATGTTgctcCAAATCTGGcattgcgtgaaccga']

Unique values in column 'nrepeats':
[1 4 5 6 3 2]

Unique values in column 'affinitynum':
[1.      0.89375 0.25    0.5     0.1     0.75   ]

Unique values in column 'orientation':
['fwd' 'rev' 'tandem']

Unique values in column 'spacer':
[ 4 10 20]

Unique values in column 'mean.norm.raw':
[-0.74969661 -0.29270245  0.76699825 ... -0.26973275 -0.07738041
 -0.53192558]

Unique values in column 'mean.norm.adj':
[-0.24435368  0.21264047  1.27234117 ... -0.04611148  0.14624085
 -0.30830431]
```

Figure 44: Screenshot from 2024-06-27 at 1:06:55 PM

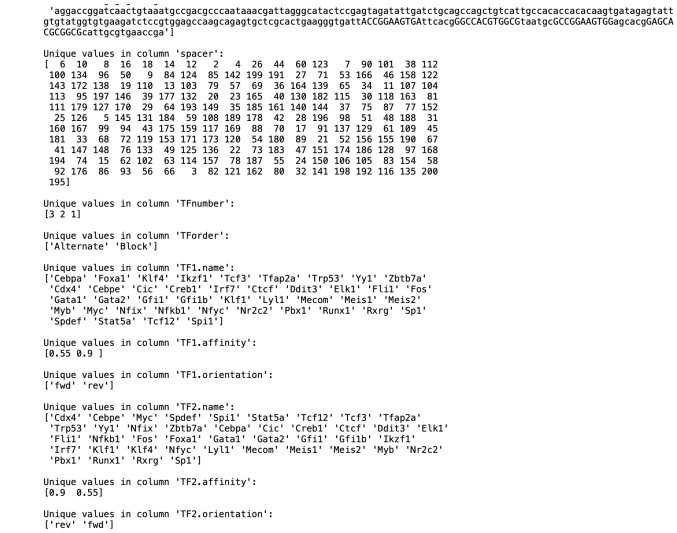

Figure 45: Screenshot from 2024-06-27 at 1:12:55 PM

8.1 Statstical physics based Models of Transcripttional Regulation

Two classes of outcomes are considered: (i) all P RNAP molecules are bound to non-specific sites, and (ii) one RNAP is bound to the promoter while the remaining $P-1$ molecules are distributed among the non-specific sites.

The number of arrangements for placing P number of RNAP II molecules on N_{NS} non-specific sites is given by the combinatorial terms, through simple combinatorics:

$$\frac{N_{NS}!}{P!(N_{NS}-P)!}.$$

Each state is weighted by its Boltzmann factor, $\exp\left(-\frac{\epsilon}{k_BT}\right)$, where ϵ represents the relevant binding energy. Defining the partition function for the promoter-unoccupied states as

$$Z(P) = \frac{N_{NS}!}{P!(N_{NS}-P)!}\,\exp\left(-\frac{P\,\epsilon_{pd}^{NS}}{k_BT}\right),$$

and including the states where the promoter is occupied,

$$Z_{tot}(P) = Z(P) + Z(P-1)\,\exp\left(-\frac{\epsilon_{pd}^{S}}{k_BT}\right),$$

the probability of promoter occupancy becomes

$$p_{bound} = \frac{Z(P-1)\,\exp\left(-\frac{\epsilon_{pd}^{S}}{k_BT}\right)}{Z_{tot}(P)}.$$

Assuming $P \ll N_{NS}$, this expression simplifies to

$$p_{bound} = \frac{1}{1+\frac{N_{NS}}{P}\,\exp\left(\frac{\Delta\epsilon_{pd}}{k_BT}\right)},$$

where

$$\Delta\epsilon_{pd} = \epsilon_{pd}^{S} - \epsilon_{pd}^{NS}.$$

To incorporate regulation by transcription factors (activators and repressors), a regulation factor F_{reg} is introduced such that the effective number of RNAP molecules available for promoter binding is modified. The resulting expression becomes:

$$p_{bound} = \frac{1}{1+\frac{N_{NS}}{P\,F_{reg}}\,\exp\left(\frac{\Delta\epsilon_{pd}}{k_BT}\right)}.$$

This framework is further extended to account for the combined occupancy of RNAP and activator molecules on the promoter and adjacent specific sites by constructing a total partition function:

$$Z_{tot}(P,A) = Z(P,A) + Z(P-1,A)\, e^{-\frac{\epsilon^S_{pd}}{k_B T}} + Z(P,A-1)\, e^{-\frac{\epsilon^S_{ad}}{k_B T}} + Z(P-1,A-1)\, e^{-\frac{\epsilon^S_{pd}+\epsilon^S_{ad}+\epsilon_{pd}}{k_B T}},$$

where A denotes the number of activator molecules and ϵ^S_{ad} is the specific binding energy for activators. This quantitative approach enables a mechanistic understanding of basal transcription and its regulation by TFs, linking microscopic binding events to macroscopic transcriptional outputs.

8.2 Modelling transcriptional regulation considering effects from co-factors

The section was introudced in the introduction, and the model is summarised below in greater detail.

8.2.1 Three-Layer Model of Transcriptional Regulation (Janssens et al., 2006)

This model decomposes gene regulation into three layers:

1. **Fractional occupancy of DNA-binding factors (including quenching)** Activators bind DNA without cooperative interactions, but short-range repressors can sequentially reduce activator occupancy.
2. **Cofactor recruitment** Each activator contributes additively to recruiting cofactors, simplifying complex molecular interactions into a single "recruitment" term.
3. **Arrhenius-type calculation of transcription rate** Cofactor accumulation lowers an effective energy barrier for transcription, often yielding near-exponential increases in output with an optional threshold to reflect saturation effects.

1. Fractional Occupancy of Transcription Factors The fractional occupancy $f_{m_i}(n_i)$ for an activator m_i with concentration (or ligand) n_i often follows a Hill-like or mass-action form:

$$f_{m_i}(n_i) \;=\; \frac{K_i\, n_i^a}{1 \;+\; K_i\, n_i^a \;+\; \ldots},$$

where K_i is an affinity constant, a is a Hill coefficient, and the denominator sums over all relevant factors. In the presence of short-range repressors, activator occupancy is reduced (quenched) by a multiplicative term:

$$F^a_{A,m_i} \;=\; 1 \;-\; q(d_k)\, E_b\, f^a_{m_i},$$

where $q(d_k)$ is a quenching factor (potentially distance-dependent), E_b is an efficiency parameter, and $f^a_{m_i}$ is the unquenched activator occupancy.

2. Cofactor Recruitment Each bound activator can recruit cofactors ("adapters") with a certain potency. The model assumes all cofactors are functionally equivalent, contributing additively to a recruitment term R:

$$R \;=\; \sum_k c_A\, E_b\, f^A_{m_k},$$

where c_A is the activator's recruitment strength, E_b is again an efficiency or scaling factor, and $f^A_{m_k}$ denotes the fractional occupancy (potentially adjusted for quenching) of each activator m_k. A simplified multiplicative term M can also be used to capture cofactor availability or other regulatory influences:

$$M \;=\; f^A\, N\, f^S,$$

where f^A and f^S represent different classes of regulatory factors or states, and N is a normalization constant.

3. Transcription Rate via an Arrhenius-Type Expression The final step translates cofactor recruitment into an effective lowering of the activation energy for transcription initiation. An Arrhenius-like expression is used:

$$\alpha_R \;=\; R_0\, \exp\Big(E_Q - Q\, M k_B\, T\Big) \quad if \quad Q\, M < E_Q,$$

otherwise α_R is set to zero (or a basal level) if $Q\, M \geq E_Q$. Here,

- R_0 is a basal rate constant,
- E_Q is an effective activation energy threshold,
- Q is a scaling factor for cofactor influence,
- M is the multiplicative term from the previous layer,
- k_B is the Boltzmann constant, and T is in K.

Because the exponential term can grow quickly, an upper threshold is often imposed to maintain biologically realistic limits on transcription. Alternatively, a sigmoidal (e.g., hill) function may replace the Arrhenius form to yield a more gradual activation response.

8.3 Hybrid mechanistic and deep model for prediction expression from sequence

Interpretable Deep Learning Model for Gene Regulation

Developed in the Velten lab, based on Liu et al 2020 and Segal et al 2008 the following interpretable moddel was develpoed to predict mRNA expression from a DNA sequence by integrating biophysical insights about transcription factor binding, occupancy, and activation. It consists of several conceptual layers:

- **Input Sequence**
  A one-hot encoded representation of the DNA, capturing the nucleotides at each position.
- **PWM Module – Computing Binding Affinities**
  Uses Position Weight Matrices (PWMs) to evaluate how well each segment of the DNA matches a particular transcription factor motif. Computes affinities for both the forward and reverse strands, and concatenates the results so that each transcription factor's binding affinity is obtained at every position for both strands.
- **Expression Scaling – Incorporating TF Concentrations**
  Each transcription factor's binding affinity is scaled by its expression level (i.e., concentration), ensuring that higher TF concentrations yield stronger overall binding signals.
- **Direct Interaction Module – Modulating Binding by Proximity**
  Considers how binding at one site may enhance or suppress binding at nearby sites. Learns distance-dependent parameters that modulate the binding affinity based on spatial proximity, producing a refined affinity tensor that incorporates interaction effects among transcription factors.
- **Fractional Occupancy Module – Estimating Site Occupancy**
  Optionally computes how much each site is actually occupied, taking into account that nearby binding events may compete or cooperate. If not enabled, the model uses the raw binding affinities directly.
- **Activation Module – Converting Binding to mRNA Expression**
  Translates binding affinities or occupancies into a predicted transcription rate. This includes summing over all positions, applying Hill-like kinetics, and calculating the final steady-state mRNA expression. Additional parameters (e.g., basal rates, kinetic constants) capture gene-specific regulatory behaviors.
- **Output**
  Returns a prediction of mRNA expression (or another relevant metric). Overall, the model progresses from DNA sequence recognition (via PWMs) and transcription factor concentrations to spatial interaction effects and kinetic activation, culminating in a biologically interpretable estimate of gene expression.

The model was attempted to improve upon by constraining the binding part of the model, as explained in the parts of the thesis containing the non-mechanistic deep model training on chip-seq data. The binding predictions would then feed into the mechanistic biophysical model which incorporates effects of bound TFs on the transcriptional cycle.

8.4 Scripts for data preparation

8.4.1 Prepare windows in multiples of 50

```
    #!/bin/bash
# Usage: ./process_windows.sh <window_size>
# Example: ./process_windows.sh 450
#          ./process_windows.sh 950

if [ "$#" -ne 1 ]; then
    echo "Usage: $0 <window_size>"
    exit 1
fi

WINDOW_SIZE=$1
HG38_FASTA="./hg38.fa"  # Reference genome FASTA file (assumed to be indexed)

# Loop over all BED files starting with "modified"
for INPUT_FILE in modified*.bed; do
    echo "Processing $INPUT_FILE with window size ${WINDOW_SIZE}..."
```

```
    # Create output file name with a prefix indicating the window size
    OUTPUT_FILE="${WINDOW_SIZE}bp_${INPUT_FILE}"

    # Define temporary files
    TEMP_BED="temp_bedfile.bed"
    AGGREGATED_BED="aggregated_bedfile.bed"
    FASTA_OUTPUT="temp_fasta_output.fa"
    CLEAN_FASTA="clean_fasta_output.txt"

    # Step 1: Map each coordinate to its WINDOW_SIZE window.
    # This rounds the start down to the nearest multiple of WINDOW_SIZE.
    awk -v ws="$WINDOW_SIZE" '{
        start = int($2 / ws) * ws;
        end = start + ws;
        print $1, start, end, $4, $5;
    }' OFS="\t" "$INPUT_FILE" > "$TEMP_BED"

    # Step 2: Aggregate the signal for both columns (columns 4 and 5) for each window.
    awk '{
 key = $1"\t"$2"\t"$3;
        sig1[key] += $4;
        sig2[key] += $5;
    } END {
 for (k in sig1) {
            print k "\t" sig1[k] "\t" sig2[k];
        }
    }' "$TEMP_BED" | sort -k1,1 -k2,2n > "$AGGREGATED_BED"

    # Step 3: Retrieve FASTA sequences for each window using Bedtools.
    bedtools getfasta -fi "$HG38_FASTA" -bed "$AGGREGATED_BED" -fo "$FASTA_OUTPUT"

    # Step 4: Clean up FASTA headers.
    # Assumes that the FASTA file is in the usual two-line format (header then sequence).
    awk 'NR % 2 == 0 {print}' "$FASTA_OUTPUT" > "$CLEAN_FASTA"

    # Combine the aggregated BED file (columns 1-5) with the corresponding sequence.
    paste <(cut -f1-5 "$AGGREGATED_BED") "$CLEAN_FASTA" > "$OUTPUT_FILE"

    # Cleanup temporary files
    rm -f "$TEMP_BED" "$AGGREGATED_BED" "$FASTA_OUTPUT" "$CLEAN_FASTA"

    echo "Output written to $OUTPUT_FILE"
done
```

8.4.2 Create composite files with pairs of transcription factors in a given cell state

```
     # Define cell states
cell_states=("CMP" "GMP" "MEP")

# Define transcription factors
tf_list=("FLI1" "PU1" "GATA2" "RUNX1" "LYL1" "TAL1")

# Create all TF pairs
tf_pairs=()
for ((i=0; i<${#tf_list[@]}; i++)); do
    for ((j=i+1; j<${#tf_list[@]}; j++)); do
        tf_pairs+=("${tf_list[i]}_${tf_list[j]}")
    done
done
```

```
# Loop through cell states and TF pairs
for cell in "${cell_states[@]}"; do
    for pair in "${tf_pairs[@]}"; do
        # Extract individual TFs
        tf1=$(echo "$pair" | cut -d'_' -f1)
        tf2=$(echo "$pair" | cut -d'_' -f2)

        # Define input file names
        file1="split50size_normchroms_GSE231422_Coverage_${tf1}_${cell}_merge_clip_50peaks.bed"
        file2="split50size_normchroms_GSE231422_Coverage_${tf2}_${cell}_merge_clip_50peaks.bed"

        # Define output file name
        output_file="composite_${tf1}_${tf2}_${cell}.bed"

        # Check if both files exist
        if [[ -f "$file1" && -f "$file2" ]]; then
            echo "Processing intersection for $tf1 and $tf2 in $cell..."

            # Ensure files are sorted correctly before bedtools
            sort -k1,1 -k2,2n "$file1" -o "$file1"
            sort -k1,1 -k2,2n "$file2" -o "$file2"

            # Perform bedtools intersect with TAB-delimited formatting and integer coordinates
            bedtools intersect -a "$file1" -b "$file2" -wa -wb | awk -v OFS="\t" '{
                print $1, int($2), int($3), $4, $5, $9, $10;
            }' > "$output_file"

            echo "Created: $output_file"
        else
            echo "Skipping: Missing file(s) for $tf1 and $tf2 in $cell."
        fi
    done
done

echo "All composite files created!"
```

8.4.3 Aggregate enrichment over window length and retrieve sequence for a given window from the hg.38 reference genome

```
#!/bin/bash

# Input BED file
INPUT_FILE="pu1_mep_genostan.bed"

# Output file
OUTPUT_FILE="2pu1_mep_genostan_250bp_windows_with_sequence.bed"

# Temporary files
TEMP_BED="temp_bedfile.bed"
AGGREGATED_BED="aggregated_bedfile.bed"
FASTA_OUTPUT="temp_fasta_output.fa"
CLEAN_FASTA="clean_fasta_output.txt"

# Path to the reference genome (current directory)
HG38_FASTA="./hg38.fa"

# Check if reference genome file exists
```

```
if [ ! -f "$HG38_FASTA" ]; then
    echo "Error: Reference genome file hg38.fa not found in the current directory."
    exit 1
fi

# Step 1: Extend and merge the BED file into 250 base pair bins, preserving chromosome information
awk '{
    start = int($2 / 250) * 250;
    end = start + 250;
    print $1, start, end, $4, $5;
}' OFS="\t" "$INPUT_FILE" > "$TEMP_BED"

# Step 2: Aggregate signal (column 5) for each 250 base pair window per chromosome
awk '{
    key = $1"\t"$2"\t"$3;
    signal[key] += $5;
} END {
    for (k in signal) {
        print k"\t"signal[k];
    }
}' "$TEMP_BED" | sort -k1,1 -k2,2n > "$AGGREGATED_BED"

# Step 3: Retrieve FASTA sequences for each window
bedtools getfasta -fi "$HG38_FASTA" -bed "$AGGREGATED_BED" -fo "$FASTA_OUTPUT"

# Step 4: Clean up FASTA headers and align sequences with BED entries
awk 'NR % 2 == 0 {print}' "$FASTA_OUTPUT" > "$CLEAN_FASTA"

# Step 5: Combine the BED file and cleaned sequences
paste "$AGGREGATED_BED" "$CLEAN_FASTA" > "$OUTPUT_FILE"

# Cleanup temporary files
rm -f "$TEMP_BED" "$AGGREGATED_BED" "$FASTA_OUTPUT" "$CLEAN_FASTA"

echo "Output written to $OUTPUT_FILE"

\begin{figure}[H]
    \centering
    \includegraphics[width=0.8\textwidth]{Unknown-4.png}
    \caption{Baseline cooperative time (in seconds) $T_1 = 16.0$; Delays (which subtract from the ava
    \label{fig:figure3}
\end{figure}

\subsection{Parameter Fitting of the Biophysical Model}
    Finally, we planned to use the \texttt{scipy-optimize} library to fit model parameters, using bi
```

9 Acknowledgements

I would like to thank Dr. Rosa Martinez-Corral and Dr. Lars Velten for allowing me to conduct my master's thesis research with their respective groups, and the lab members in both groups for their help and insightful discussions. My thesis is financially aided by a stipend, courtesy of Dr. Velten and Dr. Martinez. I would like to thank both of my supervisors for keeping my best interests in mind, and encouraging scientific thought and critical thinking. Lastly, I would like to thank my mom, dad, Pratyush and my friends, Sid, Ryth, Naman, Pritam, Manasvi, and Hrishi.

10 References

1. Synthetic enhancers reveal design principles of cell state specific regulatory elements in hematopoiesis Robert Frömel, Julia Rühle, Aina Bernal Martinez, Chelsea Szu-Tu, Felix Pacheco Pastor, Rosa Martinez Corral, Lars Velten bioRxiv 2024.08.26.609645; doi: https://doi.org/10.1101/2024.08.26.609645

2. Gunawardena, J. (2012). A linear framework for time-scale separation in nonlinear biochemical systems. PLoS One 7, e36321.

3. Ahsendorf, T., Wong, F., Eils, R., and Gunawardena, J. (2014). A frame- work for modelling gene regulation which accommodates non-equilibrium mechanisms. BMC Biol. 12, 102.

4. Nam, K.-M., Martinez-Corral, R., and Gunawardena, J. (2022). The linear framework: using graph theory to reveal the algebra and thermodynamics of biomolecular systems. Interface Focus 12, 20220013.

5. Evaluation and optimization of sequence-based gene regulatory deep learning models Abdul Muntakim Rafi et al. bioRxiv 2023.04.26.538471; doi: https://doi.org/10.1101/2023.04.26.538471

6. Dmitry Penzar, Daria Nogina, Elizaveta Noskova, Arsenii Zinkevich, Georgy Meshcheryakov, Andrey Lando, Abdul Muntakim Rafi, Carl de Boer, Ivan V Kulakovskiy, LegNet: a best-in-class deep learning model for short DNA regulatory regions, Bioinformatics, Volume 39, Issue 8, August 2023, btad457, https://doi.org/10.1093/bioinformatics/btad457

7. John W Biddle, Rosa Martinez-Corral, Felix Wong, Jeremy Gunawardena (2021) Allosteric conformational ensembles have unlimited capacity for integrating information eLife 10:e65498

8. Martinez-Corral R, Park M, Biette KM, Friedrich D, Scholes C, Khalil AS, Gunawardena J, DePace AH. Transcriptional kinetic synergy: A complex landscape revealed by integrating modeling and synthetic biology. Cell Syst. 2023 Apr 19;14(4):324-339.e7. doi: 10.1016/j.cels.2023.02.003. PMID: 37080164; PMCID: PMC10472254.

9. Combinatorial Gene Regulation through Kinetic Control of the Transcription Cycle Scholes, Clarissa et al. Cell Systems, Volume 4, Issue 1, 97 - 108.e9

10. Steven W. Lane, Fatemeh Vafaee, Emily S. Wong, Berthold Göttgens, Hamid Alinejad-Rokny, Jason W. H. Wong, John E. Pimanda; Genome-wide transcription factor–binding maps reveal cell-specific changes in the regulatory architecture of human HSPCs. Blood 2023; 142 (17): 1448–1462

11. Datta V, Siddharthan R, Krishna S (2018) Detection of cooperatively bound transcription factor pairs using ChIP-seq peak intensities and expectation maximization. PLoS ONE 13(7): e0199771.

12. Zacher B, Michel M, Schwalb B, Cramer P, Tresch A, et al. (2017) Accurate Promoter and Enhancer Identification in 127 ENCODE and Roadmap Epigenomics Cell Types and Tissues by GenoSTAN. PLOS ONE 12(1): e0169249. https://doi.org/10.1371/journal.pone.0169249

    doi:10.1371/journal.pone.0199771

13. Subramanian S, Thoms JAI, Huang Y, Cornejo-Páramo P, Koch FC, Jacquelin S, Shen S, Song E, Joshi S, Brownlee C, Woll PS, Chacon-Fajardo D, Beck D, Curtis DJ, Yehson K, Antonenas V, O'Brien T, Trickett A, Powell JA, Lewis ID, Pitson SM, Gandhi MK, Lane SW, Vafaee F, Wong ES, Göttgens B, Alinejad-Rokny H, Wong JWH, Pimanda

14. Jacob F, Monod J. 1961Genetic regulatory mechanisms in the synthesis of proteins. J. Mol. Biol 3:318–56

15. Knezetic JA, Luse DS. The presence of nucleosomes on a DNA template prevents initiation by RNA polymerase II in vitro. *Cell.* 1986 Apr 11;45(1):95–104. doi: 10.1016/0092-8674(86)90541-6. PMID: 3955658.

16. Lorch Y, LaPointe J, Kornberg RD. Nucleosomes inhibit the initiation of transcription but allow chain elongation with the displacement of histones. *Cell.* 1987;50(1):95–104. (Details based on common citation data; please verify with your sources.)

17. Schones DE, Cui K, Cuddapah S, Roh TY, Barski A, Wang Z, et al. Dynamic regulation of nucleosome positioning in the human genome. *Cell.* 2008;132(4):887–898. doi: 10.1016/j.cell.2008.01.026. PMID: 18304389.

18. Vierstra J, Lazar J, Sandstrom R, et al. Global reference mapping of human transcription factor footprints. *Nature.* 2020;584(7821):120–126. doi: 10.1038/s41586-020-2439-8.

19. Soufi A, Garcia MF, Jaroszewicz A, Osman N, Pellegrini M, Zaret KS. Pioneer transcription factors target partial DNA motifs on nucleosomes to initiate reprogramming. *Cell*. 2015;161(3):555–568. doi: 10.1016/j.cell.2015.03.017. PMID: 25892263.

20. Zaret KS, Carroll JS. Pioneer transcription factors: establishing competence for gene expression. *Genes Dev*. 2011;25(21):2227–2241. doi: 10.1101/gad.615811. PMID: 21808970.

21. Haberle V, Stark A. Eukaryotic core promoters and the functional basis of transcription initiation. *Nat Rev Mol Cell Biol*. 2018;19(10):621–637. doi: 10.1038/s41580-018-0053-1. PMID: 30118703.

22. Roeder RG. The role of general initiation factors in transcription by RNA polymerase II. *Trends Biochem Sci*. 1996;21(9):327–335. doi: 10.1016/0968-0004(96)10038-4. PMID: 8821625.

23. Egly JM, Coin F. [Review on transcription initiation and TFIIH]. In: *Cold Spring Harb Symp Quant Biol*. 2011;76:167–174. (Exact title details are not fully available; please verify if a more complete citation is needed.)

24. Kornberg RD. Mediator and the mechanism of transcriptional activation. *Trends Biochem Sci*. 2005;30(5):235–239. doi: 10.1016/j.tibs.2005.03.004. PMID: 15800818.

25. Landick R. The regulatory roles and mechanisms of transcriptional pausing. *Curr Opin Genet Dev*. 2006;16(5):477–483. doi: 10.1016/j.gde.2006.08.007. PMID: 16915562.

26. Cheung AC, Cramer P. Structural basis of RNA polymerase II backtracking, arrest and reactivation. *Nature*. 2011;471(7340):249–253. doi: 10.1038/nature09703. PMID: 21281704.

27. Eick D, Borkamm R. [Citation details not available – please update with full reference information].

28. Weirauch MT, Yang A, Albu M, Cote AG, Montenegro-Montero A, Drewe P, et al. Determination and inference of eukaryotic transcription factor sequence specificity. *Cell*. 2014;158(3):1431–1443. doi: 10.1016/j.cell.2014.06.009. PMID: 25041268.

29. Brodsky AS, et al. [Full details not available – placeholder citation]. (Please check your sources for complete information.)

30. Liu X, et al. [Full details not available – placeholder citation]. (Please verify and update as needed.)

31. Boija A, Klein IA, Sabari BR, et al. Transcription Factors Activate Genes through the Phase-Separation Capacity of Their Activation Domains. *Cell*. 2018;175(7):1842–1855.e16. doi: 10.1016/j.cell.2018.09.037. PMID: 30217466.

32. Hnisz D, Shrinivas K, Young RA, et al. A Phase Separation Model for Transcriptional Control. *Cell*. 2017;169(1):13–23. doi: 10.1016/j.cell.2017.03.035. PMID: 28498233.

33. Banani SF, Lee HO, Hyman AA, Rosen MK. Biomolecular condensates: organizers of cellular biochemistry. *Nat Rev Mol Cell Biol*. 2017;18(5):285–298. doi: 10.1038/nrm.2017.7. PMID: 28300287.

34. Rahl PB, Lin CY, Seila AC, Flynn RA, et al. c-Myc regulates transcriptional pause release. *Cell*. 2010;141(3):432–445. doi: 10.1016/j.cell.2010.03.030. PMID: 20339731.

35. Pomp A, et al. [Citation details not available – placeholder citation]. (This is a recent reference; please update with complete information when available.)

36. Jolma A, Kivioja T, Toivonen J, et al. DNA-binding specificities of human transcription factors. *Cell*. 2013;152(1–2):327–339. doi: 10.1016/j.cell.2012.12.009. PMID: 23299642.

37. Jolma A, Yin Y, Nitta KR, et al. DNA-dependent formation of transcription factor pairs alters their binding specificity. *Nature*. 2015;527(7578):384–388. doi: 10.1038/nature15518. PMID: 26182410.

38. Gertz J, Savic D, Varley KE, et al. Distinct properties of cell-type-specific and shared transcription factor binding sites. *Mol Cell*. 2013;52(1):25–36. doi: 10.1016/j.molcel.2013.08.037. PMID: 24056586.

39. Nakato R, Shirahige K. Recent advances in ChIP-seq analysis: from quality control to genome-wide profiling. *Brief Bioinform*. 2021;22(4):1476–1492. doi: 10.1093/bib/bbaa297. PMID: 33927759.

40. Park PJ. ChIP-seq: advantages and challenges of a maturing technology. *Nat Rev Genet*. 2009;10(10):669–680. doi: 10.1038/nrg2641. PMID: 19725973.

41. Inukai S, et al. [Full citation details not available – placeholder for a DAP-seq or similar method]. (Please update with complete information.)

42. van Steensel B, Furlong EE. The role of transcription in shaping the spatial organization of the genome. *Nat Rev Genet*. 2019;20(8):327–337. doi: 10.1038/s41576-019-0113-6. (Verify volume/pages as needed.)

43. Chong S, Dugast-Darzacq C, Liu Z, et al. Imaging dynamic and selective low-complexity domain interactions that control gene transcription. *Science*. 2018;361(6400):694–699. doi: 10.1126/science.aar2555. PMID: 30321441.

44. Sabari BR, Dall'Agnese A, Boija A, et al. Coactivator condensation at super-enhancers links phase separation and gene control. *Science*. 2018;361(6400):387–392. doi: 10.1126/science.aar3958. PMID: 30446475.

45. Avsec Z, Agarwal V, et al. Effective gene expression prediction from sequence by integrating long-range interactions. *Nat Genet*. 2021;53:134–142. doi: 10.1038/s41588-020-00740-9. PMID: 33216766.

46. Smith RP, Taher L, Patwardhan RP, et al. Massively parallel decoding of mammalian regulatory sequences supports a flexible organizational model. *Nat Genet*. 2013;45(9):1021–1028. doi: 10.1038/ng.2693. PMID: 23906859.

47. Farley EK, Olson KM, Zhang W, et al. Suboptimization of developmental enhancers. *Science*. 2015;350(6258):325–328. doi: 10.1126/science.aac5846. PMID: 25613420.

48. Ernst J, Kellis M. Chromatin-state discovery and genome annotation with ChromHMM. *Nat Protoc*. 2017;12(12):2478–2492. doi: 10.1038/nprot.2017.124. PMID: 28962880.

49. Lara-Astiaso D, Weiner A, Lorenzo-Vivas E, et al. Chromatin state dynamics during blood formation. *Science*. 2014;345(6199):943–949. doi: 10.1126/science.1256271. PMID: 25394262.

50. Graf T, Enver T. Forcing cells to change lineages. *Nature*. 2009;462(7273):587–594. doi: 10.1038/nature08533. PMID: 19693574.

51. Erwin DH, Davidson EH. The evolution of hierarchical gene regulatory networks. *Nat Rev Genet*. 2009;10(2):141–148. doi: 10.1038/nrg2507. PMID: 19154285.

52. Corces MR, et al. Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution. *Nat Genet*. 2016;48(10):1193–1203. doi: 10.1038/ng.3646. PMID: 27580941.

53. Beer MA, Tavazoie S. Predicting gene expression from sequence: a reexamination. *PLoS Biol*. 2004;2(8):E207. doi: 10.1371/journal.pbio.0020207. PMID: 15343076.

54. Avsec Z, Agarwal V, et al. Effective gene expression prediction from sequence by integrating long-range interactions. *Nat Genet*. 2021;53:134–142. doi: 10.1038/s41588-020-00740-9. PMID: 33216766.

55. Avsec Z, John Jumper et al. Enformer: Predicting gene expression from long-range interactions. [Details similar to reference above; if a distinct publication by DeepMind is intended, please update accordingly].

56. Barbadilla-Martínez J, et al. [Citation details not available – placeholder citation]. (Please update with full details when available.)

57. Gosai S, et al. [Citation details not available – placeholder citation]. (Please update with full details when available.)

58. de Boer CG, et al. Deciphering gene regulatory logic with deep learning and synthetic reporter assays. *Nat Biotechnol*. 2020;38(7):948–956. doi: 10.1038/s41587-020-0505-2. PMID: 32730996.

59. Park, P. ChIP–seq: advantages and challenges of a maturing technology. Nat Rev Genet 10, 669–680 (2009). https://doi.org/10.1038/nrg2641

60. Ryuichiro Nakato, Toyonori Sakata, Methods for ChIP-seq analysis: A practical workflow and advanced applications, Methods, Volume 187, 2021, Pages 44-53, ISSN 1046-2023, https://doi.org/10.1016/j.ymeth.2020.03.005

61. Inukai S, Kock KH, Bulyk ML. Transcription factor-DNA binding: beyond binding site motifs. Curr Opin Genet Dev. 2017 Apr;43:110-119. doi: 10.1016/j.gde.2017.02.007. Epub 2017 Mar 27. PMID: 28359978; PMCID: PMC5447501.

62. Segal E, Raveh-Sadka T, Schroeder M, Unnerstall U, Gaul U. Predicting expression patterns from regulatory sequence in Drosophila segmentation. Nature. 2008 Jan 31;451(7178):535-40. doi: 10.1038/nature06496. Epub 2008 Jan 2. PMID: 18172436.

63. Phillips, R., Kondev, J., Theriot, J., Garcia, H. (2012). Physical Biology of the Cell (2nd ed.). Garland Science. https://doi.org/10.1201/9781134111589

64. Bialek, William (2012). Biophysics: Searching for Principles, Princeton University Press.

65. Biophys J. 2014 Apr 15;106(8):1801–1810. doi: 10.1016/j.bpj.2014.02.019