# QUANTIFYING THE FUNCTIONAL AND EVOLUTIONARY RELATIONSHIPS AMONG SEQUENCES, TRANSCRIPTION FACTOR BINDING AND GENE EXPRESSION

BY

PEI-CHEN PENG

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Computer Science
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2018

Urbana, Illinois

Doctoral Committee:

Professor Saurabh Sinha, Chair
Professor ChengXiang Zhai
Professor Olgica Milenkovic
Professor Remo Rohs, University of Southern California

# ABSTRACT

A central challenge in regulatory genomics today is to understand the precise relationship between regulatory sequences, transcription factor (TF) binding and gene expression. Many studies have discussed how TFs recognize their DNA binding sites. However, it is not well understood how the various factors that influence TF-DNA binding alter the cascade of gene expression. Moreover, mutations in regulatory sequences are a key driving force of evolution and diseases. A number of studies have examined the sequence motif turnover and divergence in TF binding across species. However, there is currently a lack of clarity on what these changes mean to enhancer function. In this thesis, we used computational and statistical methods to quantitatively and systematically examine the relationships among regulatory sequences, TF binding, and gene expression, from both functional and evolutionary perspectives.

At the functional level, we extended thermodynamics-based statistical models of the genetic sequence-to-function relationship to accurately predict gene expression. We incorporated chromatin accessibility and structural biological data into the models, described in Chapter 2 and 3. In doing so, we aimed to better identify transcription factor binding sites likely to influence gene expression, and thus, enhance the models' capacity to predict gene expression. We demonstrated these improvements to gene expression modeling in *Drosophila melanogaster* by integrating DNaseI hypersensitivity assays and DNA shape. At the evolutionary level, we focused on regulatory variations between two distant *Drosophila* species to access inherent properties of enhancers, as described in Chapter 4. We used statistical and computational approaches to quantitatively examine the extent to which sequence and accessibility variations can predict TF occupancy divergence and enhancer activity change. We also found combinatorial TF binding can buffer variations at individual TF level to avoid drastic gene expression changes.

*To my family, for their love and support.*

# ACKNOWLEDGMENTS

First and foremost, I am grateful to my advisor, Professor Saurabh Sinha. Since my first day joining the lab, Saurabh believed in me and gave me continuous support. He guided me to identify broad-impact research questions, fostered my research skills, and mentored me to justify research findings under scrutiny. Moreover, he prepared me to pursue an academic career by providing unique opportunities for collaborative projects, teaching me how to clearly motivate and deliver concepts in writing papers, and encouraging me to present my works to the public, no matter how big or small. His openness and support made me feel my success is aligned with his.

Besides my advisor, I would like to thank our collaborator, Dr. Eileen Furlong, and my thesis committee: Professor ChengXiang Zhai, Professor Olgica Milenkovic, and Professor Remo Rohs. They inspired me on rigorous attitude of conducting research, and provided insightful feedbacks and research directions on this thesis, from either biological or computational perspectives.

I would also like to express my gratitude to my fellow labmates. Dr. Hassan Samee and Dr. Charles Blatti kindly shared their experience in regulatory genomics research in my junior years. My sincere thanks also go to Casey Hanson, Laura Sloofman, and Farzaneh Khajoui, for giving me space to speak freely in both constructive and irrelevant discussions. It was my great pleasure to work with other members in Sinha's group, including Dr. Thyago Duque, Wei Yang, Shayan Tabe Bordbar, Bryan Lunt, Xiaoman Xie, Saba Ghaffari, Payam Dibaenia, and Shounak Bhogale. I am also grateful to friends in UIUC, for all the fun we have had in the last five years.

Most of all, I thank my parents, Han-Chang Peng and Yue-E Hung, for their warm love, endless support, and continuously being on my back. I thank my sister, Ju-Yi Peng, for being my long-time best friend whose laugh is funnier than my jokes. I thank my once best friend and now husband, Dr. Po-Liang Wu, for stimulating new research ideas and enriching my life.

# TABLE OF CONTENTS

# CHAPTER 1: INTRODUCTION

## 1.1 BASICS OF GENE REGULATION

Gene regulation refers to the mechanisms that activate or repress the expression of a gene. Different cells in a multicellular organism may express very different sets of genes, despite the fact that almost all cells contain the same DNA. These different patterns of gene expression cause various cell types to have different sets of functional mRNAs and proteins, giving each cell type unique properties. How this orchestration of thousands of genes acts in a precise way has been an open question. Answering this question can lead to advances in our understanding of developmental programs [1], molecular basis of phenotypes [2], evolution of morphological diversity [3, 4], and ability to prioritize variants in diseases [5].

Among the different levels at which genes may be regulated, the one that has received most attention to date is transcriptional regulation [3, 6]. Transcriptional regulation controls the conversion from DNA to mRNA. Several key players work in concert to finely tune the amount of RNA transcripts being produced, as depicted in Figure 1.1. Instructions of generating RNA transcripts are encoded in regulatory DNA sequences, called enhancers, or *cis*-regulatory modules (CRMs) in some contexts. Enhancers are about 1 kbp long sequences that harbor binding sites for one or more transcription factors. Transcription factors (TFs) are proteins that bind to enhancers with sequence specificity. The binding preference of a TF is known as a motif, and is often represented formally by a Position Weight Matrix (PWM). Enhancers harbor transcription factor binding sites (TFBSs) that are strong matches to the motif. A set of TFs bind to their corresponding TFBSs and act together to regulate a gene's expression pattern, by facilitating or inhibiting recruitment and assembly of the basal transcription machinery (BTM), which is an essential complex to initiate transcription. The availability of TFBSs on DNA is determined by the chromatin structure. More open or relaxed chromatin makes TFBSs accessible to TFs, while TFBSs in condensed DNA regions are inaccessible to TFs, and preclude the region from regulating gene transcription.

Recent technological breakthroughs such as genome-wide chromatin state profiling [7, 8] and massively parallel reporter assays [9, 10] are leading the way in rapid and effective discovery of enhancers. The next frontier [11] is to learn to interpret an enhancer's sequence and predict the expression level driven by the enhancer in a given trans-regulatory context, e.g., a particular tissue or cell type [12, 13, 14]. Various studies have attempted to meet this challenge, and a line of attack that has met with considerable initial success is that of thermodynamics-based models [15, 16, 17, 18, 19, 20, 21, 22].

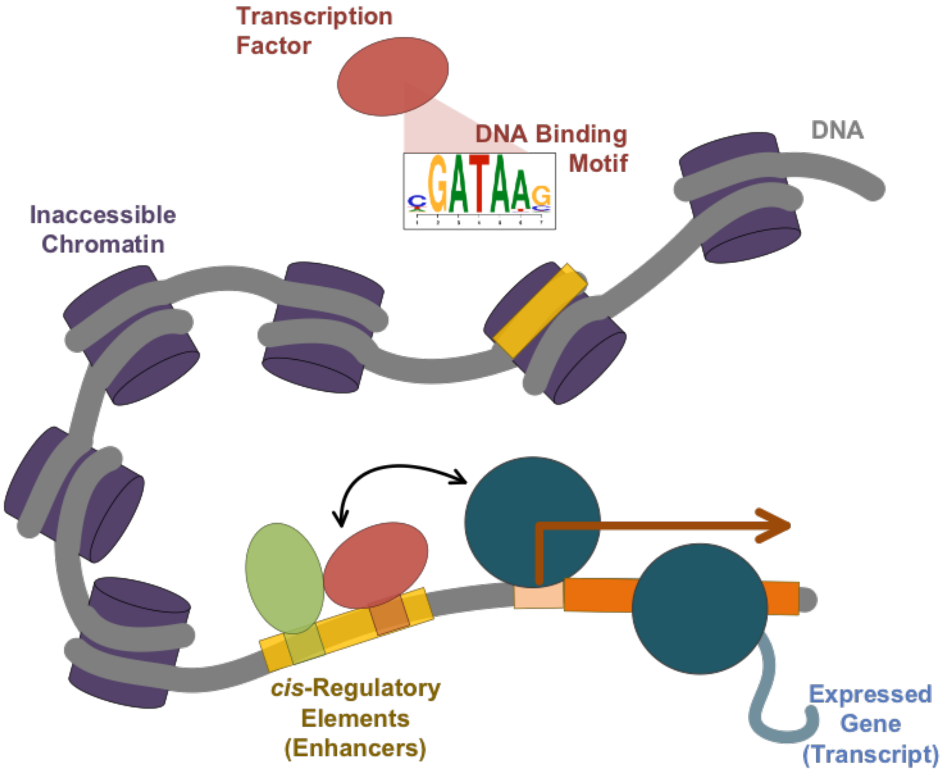

Figure 1.1: **Key players in transcriptional regulation.** Enhancers harbor binding sites for one or more transcription factors to act together to regulate a gene's expression. TFs recognize binding sites by motif specificities. Chromatin structure determines the accessibility of TF binding sites. Basal Transcription Machinery is recruited and assembled by TFs and essential to transcribe a gene.

## 1.2 THERMODYNAMICS-BASED SEQUENCE-TO-EXPRESSION MODELS

Thermodynamics-based sequence-to-expression models have proven capable producing highly accurate fits to complex gene expression patterns. The hallmark of these models is that they are built around molecular interactions involving TF proteins, DNA and the basal transcriptional machinery, and use the language of statistical thermodynamics to map combinations of interactions, both strong and weak, to gene expression levels. Fits of these models to sequence and expression data capture underlying mechanistic details of gene regulation at a convenient level of abstraction. For instance, DNA-binding strengths of TFs and the potency of activation or repression by a DNA-bound TF appear as free parameters of these models, and their optimal values learned from data provide quantitative insights into underlying regulatory mechanisms.

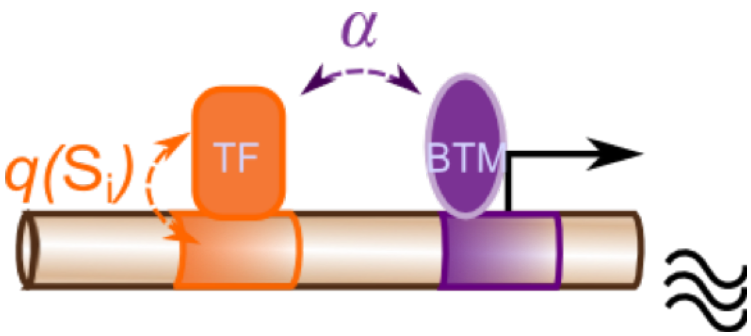

Figure 1.2: **Transcriptional regulation is modelled on major components:** TF (orange), BTM (purple), and DNA (brown tube). The interactions of TF-DNA, BTM-DNA, TF-BTM are assumed to occur in thermodynamic equilibrium. Presumably, gene expression level is proportional to the fractional BTM occupancy at the promoter.

### 1.2.1 GEMSTAT model

Here, we introduce a previously published sequence-to-expression model, GEMSTAT (Gene Expression Modeling Based on Statistical Thermodynamics) [19], which is a statistical thermodynamics-based model of enhancer sequence readout. As shown in Figure 1.2, transcriptional regulation can be modelled as the interaction of three components: DNA sequence, TFs, and the basal transcriptional machinery (BTM). A TF can bind on any site of the DNA sequence with a site-specific probability or affinity. The BTM can bind on the core promoter and initiate transcription. The model assumes that, the interactions of TF-DNA, BTM-DNA, and TF-BTM occur in thermodynamic equilibrium. Following Shea and Ackers [23], GEMSTAT further assumes that the gene expression level is proportional to the fractional BTM occupancy at the promoter.

GEMSTAT computes the fractional occupancy of the BTM by considering an ensemble of molecular configurations, each of which is denoted by $\sigma$ and specifies which sites are bound and which are free. All configurations assume one of two states: one where the BTM is bound or another where the BTM is unbound. The statistical weights of the two states are $W(\sigma)Q(\sigma)$ and $W(\sigma)$ respectively. $W(\sigma)$ represents the contribution of TF-DNA interactions, calculated based on TF concentrations and binding affinities of bound sites. $Q(\sigma)$ represents the contribution of TF-BTM interactions, modelled as $\alpha$, a vector of free parameters with one scalar for each TF, as indicated in Figure 1.2. Given this, the relative probability of bound BTM is the following, where the gene expression level is proportional to $E$:

$$E = \frac{\sum_\sigma W(\sigma)Q(\sigma)}{\sum_\sigma W(\sigma)Q(\sigma) + \sum_\sigma W(\sigma)} \tag{1.1}$$

In this paragraph, we detail the derivation of the statistical weight $W(\sigma)$, and descriptions of parameters used in GEMSTAT are listed in Table 1.1. The sub components of the statistical weight are the contributions of each binding site in a configuration $\sigma$. As shown in Figure 1.2, $q(S)$ represents the contribution of a binding site $S$ to $W(\sigma)$ and is given by the following equation:

$$q(S) = K(S_{\mathrm{opt}})\,\nu\,[\mathrm{TF}\ ]_{\mathrm{rel}}\exp\left[\mathrm{LLR}(S) - \mathrm{LLR}(S_{\mathrm{opt}})\right] \tag{1.2}$$

In this formulation, $[\mathrm{TF}]_{\mathrm{rel}}$ represents the relative TF concentration, so that $\nu\,[\mathrm{TF}]_{\mathrm{rel}}$ is the absolute concentration, for some TF-specific constant of proportionality $\nu$ . $\mathrm{LLR}(s)$ for any site $S$ represents the log likelihood ratio score of the site $S$ for the PWM motif of TF, $S_{opt}$ is the optimal binding site. For simplicity, we define $E_0(S) = \exp\left[\mathrm{LLR}(S) - \mathrm{LLR}(S_{\mathrm{opt}})\right]$ as the TF-DNA binding energy at site S henceforth. $K(S_{\mathrm{opt}})$ represents the association constant of TF-DNA binding for the optimal site. Since both $K(S_{\mathrm{opt}})$ and $\nu$ are unknown constants, GEMSTAT treats the product of the two as a free parameter. The statistical weight $W(\sigma)$ is then given by the following equation:

$$W(\sigma) = \prod_i q(S_i)^{\sigma_i} \tag{1.3}$$

### 1.2.2 Model training

Three different goodness-of-fit functions are used at various stages of optimization, to compare between real and predicted expressions of enhancer sequences: average correlation coefficient (Avg. CC), root mean square error (RMSE), and weighted Pattern Generating Potential (wPGP, taken from [24] and described in the following subsection). To avoid being trapped in local optima, parameter optimizations were done in multiple runs while alternating between Avg. CC and RMSE as the objective functions. The optimization starts with a set of default parameters and Avg. CC as the objective function. Upon convergence, the resulting set of parameters is used to initiate optimization with RMSE as the objective function, which is run to convergence. This procedure of optimizations alternating between Avg. CC and RMSE as objective functions is repeated twice, and the resulting set of parameters initiates the final optimization step that uses wPGP as the objective function. Each optimization is done by alternating between the Nelder-Mead simplex method and the quasi-Newton method, as in He et al. [19].

Table 1.1: **Parameters used in GEMSTAT.**

| **Parameter** | **Description** | **Number** |
|---|---|---|
| bindingWt$_\mathrm{i}$ ( $q\left(S_i\right)$ ) | Represents the dissociation constant of the equilibrium reaction between the i-th TF, TF$_\mathrm{i}$ and its optimal binding site when the concentration of TF$_\mathrm{i}$ is maximum | One per TF |
| q$_\mathrm{BTM}$ | A phenomenological parameter that captures the combined effect of all molecular species that act downstream of the TF recruitment step and initiate transcription (such molecular species are collectively known as the basal transcription machinery or BTM) | One global parameter |
| txpEffect$_\mathrm{i}$ ( $\alpha_i$ ) | Represents the strength of TF$_\mathrm{i}$'s effect on the BTM | One per TF |
| $\omega_{ij}$ | Strength of interaction between molecules of two TFs, TF$_\mathrm{i}$ and TF$_\mathrm{j}$ (i and j may be the same), which are assumed to bind cooperatively to the DNA | One per pair of TFs (TF$_\mathrm{i}$ and TF$_\mathrm{j}$) that are assumed to have cooperativity in DNA binding |

### 1.2.3 Evaluation of model predictions using wPGP (weighted pattern generating potentials)

Given the predicted and real expression profiles, the wPGP score [24] is defined as follows:

$$\mathrm{wPGP} = 0.5 + 0.5 \times \left(\mathrm{reward} - \mathrm{penalty}\right), \tag{1.4}$$

where reward = $\frac{\sum_i r_i \times \min(r_i, p_i)}{\sum_i r_i \times r_i}$ , and penality = $\frac{\sum_i (\max_r - r_i) \times (p_i - r_i) \times \mathrm{I}(p_i > r_i)}{\sum_i (\max_r - r_i) \times \sum_i (\max_r - r_i)}$ . Here, $p_i$ and $r_i$ are the predicted and the real expression in bin $i$, respectively, $\max_r$ is the maximum level of real gene expression, and I(B) is a binary variable indicating the truth of condition "B" . The wPGP score ranges from 0 to 1, with higher scores indicating better matches between the predicted and the endogenous expression. The wPGP score was used as the objective function during parameter training, as well as for assessing if one model fits the data better than another.

## 1.3 CHROMATIN ACCESSIBILITY AND REGULATION OF GENE EXPRESSION

One key aspect missing from the mechanistic view adopted in today's thermodynamics-based models is that of chromatin state. A significant advance in recent years in the field of regulatory genomics has been the realization that chromatin state, e.g., specific histone modifications and general accessibility patterns, of *cis*-regulatory regions strongly correlates with expression and with regulatory events leading to expression [25, 26, 27]. Genome-wide profiling of DNaseI hypersensitive sites (DHS), representing regions of relatively accessible chromatin, or of specific histone modifications such as H3K27ac, has proven to be a powerful strategy to map regulatory DNA and pinpoint active enhancers [28, 29, 30, 31]. For instance, genome-wide, high-resolution, *in vivo* mapping of DHS sites has helped chart the regulatory DNA landscape of *Drosophila* early embryo development [32], showing how chromatin accessibility may influence genome-wide, overlapping patterns of TF binding during embryogenesis [33, 25, 26]. Additionally, we now know that chromatin state (e.g., accessibility) of a genomic segment is an effective predictor of its regulatory activity [34, 29, 35] and an important feature in predicting TF occupancy therein [36]. In particular, incorporation of accessibility data has significantly improved the accuracy of predicting *in vivo* TF occupancy over baseline models that used sequence-specific motifs alone [37, 38]. These findings naturally raise the question: *does chromatin state information also improve our ability to quantitatively predict expression levels driven by an enhancer?* To our knowledge, this question has not been systematically and empirically answered so far, and is the subject of Chapter 2.

Based on our knowledge today, we might expect an affirmative answer to the above question. If chromatin accessibility data improves our ability to predict TF-DNA binding, which it does [33, 25, 38], and since it is generally accepted that better prediction of TF-DNA binding should lead to better expression prediction, it follows that accessibility data ought to improve sequence-to-expression prediction. However, testing this hypothesis requires coupling the two computational aspects mentioned above, i.e., accessibility data → TF-DNA binding prediction and binding prediction → expression prediction, and evaluating the integrated approach on an appropriate data set. This was the methodological challenge facing us in Chapter 2.

Moreover, it was not clear to us going in to this study if the resolution of available data and the expressivity of today's sequence-to-expression models are adequate to demonstrate the advantage of incorporating accessibility data, even if such an advantage exists. Note that our goal was not to use accessibility data to *identify* enhancers and then predict expression from their sequence; rather we wanted to test if variations of accessibility *within known enhancers*, at about 20 – 25 bp resolution [26, 39], can inform sequence-to-expression models in useful ways. This required that the models be sensitive enough to register quantitative variations of DNA-accessibility at individual binding sites, and that the accessibility data pertain to the same cell types for which we do have accurate sequence-to-expression models.

In Chapter 2 we build and evaluate a quantitative model that maps regulatory DNA sequence to the regulated gene's expression while integrating DNA accessibility data. Several studies [17, 18, 19, 40, 22, 41] have proposed quantitative models of the sequence-to-expression relationship. One such quantitative model is "GEMSTAT", a statistical thermodynamics based model that we previously showed to successfully model dozens of enhancers involved in specification of the anterior/posterior (A/P) axis in early *Drosophila* embryos [19] (see Section 1.2 for details). GEMSTAT is the only available general purpose tool that can be trained to model the regulatory activities of a set of enhancers with a common assignment of free parameters. Moreover, its thermodynamics-based formulation lends itself to incorporation of accessibility data in an intuitive and semi-mechanistic manner, to an extent that one may study how accessibility of individual binding sites may impact expression. These considerations, along with our extensive experience with GEMSTAT made it a natural choice as the modeling framework adopted here. The regulatory system we chose comprises the above-mentioned A/P patterning enhancers from *Drosophila*, in part because this system has been the subject of several modeling studies by us [19, 24, 42] and others [20, 43, 22, 44], and also because chromatin accessibility data are available for the developmental stage represented by this data set.

## 1.4 THE ROLE OF DNA SHAPE IN TF-DNA BINDING AND GENE EXPRESSION

A key aspect of transcriptional regulation is the sequence-specific DNA-binding of transcription factors, and in recent years there has been a strong push towards precise characterization of transcription factor (TF)-DNA binding and its underlying mechanisms [45]. The extent to which a TF's binding specificity at a site is dictated by the TF directly interpreting the nucleotide sequence ("base readout") or DNA shape at the site ("shape readout") is a topic of considerable debate [46, 47], as is the role played by secondary TFs [48, 33, 49] that cooperatively or competitively influence *in vivo* DNA-binding.

A number of high throughput assays have been developed to generate data sets on which our understanding of TF-DNA binding can be rigorously tested [50]. To support the study of biochemical mechanisms underlying TF-DNA binding, various computational models have emerged to describe these mechanisms and use them to fit experimental data [51]. The de facto leader of this pack is the "position weight matrix" or PWM model, which prescribes a multinomial distribution over four nucleotides for each position of the binding site, the distributions at different positions being independent of each other [52, 53]. The PWM model has been extensively used in regulatory sequence analysis and numerous algorithms are available for inferring a PWM model from a TF's binding sites [54, 55, 56, 57, 58].

At the same time, several reports have pointed out deficiencies in the model and presented alternative models that are claimed to be in greater agreement with binding data [59, 60, 61]. In short, the high intensity of on-going experimental and computational work in this field has taken us much closer to a quantitative and predictive model of a TF's DNA-binding specificity.

The ultimate goal in modeling TF-DNA binding is to use this ability to understand gene regulation. Achieving this goal will allow us to "read" and interpret non-coding sequences and hence their relationship to organismal form and function [62], and their evolution [63]. It will enable major advances in the genomics of human health, by providing accurate predictions of the effects of single nucleotide polymorphisms at the cellular level. Precise models of *in vitro* and *in vivo* binding take us only part of the way to this grand goal, and must be incorporated into sequence-specific models of gene expression (sequence-to-expression models) for their value to be truly realized. Sequence-to-expression models are steadily gaining popularity, and have been used, among other things, to predict precise levels of gene expression in different regions of the developing embryo [17, 18, 19, 20, 64, 22, 44] or to predict tissue-specific gene expression in humans [65, 27, 66]. However, there is a disconnect today between these models of gene expression and the burgeoning body of work on TF-DNA binding specificity. Sequence-to-expression models exclusively rely on the PWM model of DNA-binding, and it is unknown if alternative, emerging models of DNA-binding can substantially improve prediction of gene expression. This is the gap that we attempt to fill in Chapter 3.

We considered a model of TF-DNA binding that incorporates local DNA shape at the binding site and asked if it performs as well as a PWM model in predicting gene expression. To answer this question, we considered one of the best-studied regulatory systems today – the set of genes and respective enhancers responsible for anterior-posterior (A/P) patterning of the blastoderm-stage *Drosophila* embryo [19, 22, 44]. We used the thermodynamics-based GEMSTAT model [19] to predict gene expression levels from enhancer sequence and TF concentrations, using the DNA-binding model to parse the enhancer sequence in terms of the types, strengths and arrangements of binding sites within. We used rigorous methods of comparing model fits [24], to find that a DNA-binding model based on "shape readout" [67] performs at least as well as, and arguably better than, the PWM model. We performed additional tests to examine if integrating shape readout and PWM into a single model would achieve better predictions than using either binding model independently. To our knowledge, this is the first successful attempt at quantitatively modeling the function of an enhancer sequence using a description of TF-DNA binding specificity other than the PWM. The shape-based model used here was trained on (binding site) data from the Bacterial

1-hybrid system [68]. With the growing availability of data sets describing TF-DNA binding affinities more comprehensively [45, 50], we expect that it will be possible to train such models more accurately and to demonstrate their ability to predict gene expression better than with PWM models alone.

## 1.5 *CIS*-REGULATORY EVOLUTION IN ASPECTS OF SEQUENCE, ACCESSIBILITY, TF BINDING, AND ENHANCER ACTIVITY

*Cis*-regulatory evolution plays an important role in phenotypic diversity, including morphological [69], physiological [70] and behavioral [71, 72] evolution. Given its importance, many studies have examined cross-species changes in various aspects of gene regulation, including expression [73], enhancer activity [74], transcription factor (TF)-DNA binding [75, 76, 77, 78], TF binding motifs [75, 79, 80, 81, 82], DNA accessibility [83] and chromatin states [84]. Changes have been observed to differing extents in these measurable aspects of gene regulation, leading to some emerging principles underlying their conservation and divergence [79]. At the same time, it is challenging to systematically integrate these diverse qualitative observations about *cis*-regulatory evolution at different regulatory levels given the different phyla, biological systems, and technologies.

A number of studies have examined the evolution of DNA *cis*-regulatory sequences [85, 86]. Some have noted surprisingly high levels of sequence change [87, 88, 89], but regulatory function and gene expression are often conserved despite sequence-level changes [90, 91, 74, 92, 93], revealing considerable flexibility in sequence encoding the same function [94, 95]. Further investigations asked if the observed functional buffering against sequence divergence happens at the level of TF-DNA binding, which is the principle molecular event mediating sequence-expression relationships. ChIP-chip or ChIP-seq assays of the same TFs were performed in multiple species [75, 76, 74, 77] and while genome-wide TF binding landscapes were noted to be conserved overall, many large qualitative as well as quantitative differences in binding were also reported [75]. The evolution of binding landscapes thus emerged as an intriguing aspect of molecular evolution, and researchers sought to identify its main determinants.

Loss and gain of TF ChIP peaks are correlated with changes in the presence of the TF's DNA binding motif [75, 79, 80, 81, 82], but this relationship, though significant in its extent, was far from a satisfactory explanation for TF binding differences. For instance, many peaks are lost though the motif is conserved, and conversely, peaks are often conserved despite motif loss. Some studies noted the influence of co-binding TFs at or near the peak [77], suggesting roles for co-operative [91] and 'TF collective' modes of occupancy [74].

On the other hand, Bradley et al. [75] interpreted an observed correlation between evolutionary changes of occupancy among multiple TFs as evidence for TF-independent influences such as differences in local chromatin accessibility. Indeed, DNA accessibility is known to be a major correlate of TF-DNA binding [33, 26], and may therefore underlie evolutionary changes in TF binding. For instance, Paris et al. [73] noted that binding divergence is correlated with changes in binding sites for the pioneer factor Zelda, which indirectly implicates accessibility changes. Genome-wide accessibility landscapes are generally evolutionarily conserved [96, 97], but accessibility changes between orthologous genomic elements are also observed and raise the question: how often do they underlie evolutionary changes in TF binding? Surprisingly, there is no direct analysis of this question. In related work, Connelly et al. [96] reported evidence that much of accessibility divergence (between two yeast species) may be inconsequential for gene expression. Alexandre et al. [83] made similar observations for different ecotypes of A. thaliana, but also noted that loci with high sequence variation and accessibility changes were significantly linked to expression changes. However, the extent to which accessibility changes are predictive of TF binding changes between species remains unknown. Is this relationship comparable in extent to the documented relationship between motif change and TF binding divergence? Do changes in accessibility and motif presence carry complementary information related to observed changes in TF ChIP peaks? How often is accessibility conserved, yet a TF's occupancy diverged due to motif turnover, and how commonly do changes in accessibility result in loss or gain of TF ChIP peaks despite conservation of motif presence? These are not mutually exclusive possibilities and teasing apart their relative contributions and potential causal influence requires a formal, quantitative analysis. Insights emerging from such analyses may also fuel discussions of cause-versus-effect in the relationship between TF binding and accessibility [98, 26]. In addition to advancing our basic understanding of *cis*-regulatory evolution, answering these questions may also allow us to predict changes in TF binding using computational models that incorporate data on sequence and accessibility changes, bypassing the need for expensive ChIP profiling of TFs across species and individuals.

Any investigation of these aspects of *cis*-regulatory evolution must also consider promiscuous occupancy of TFs [99] and that a large number of ChIP peaks may not have a functional impacts on gene expression [100], or be functionally redundant. Evolutionary comparisons have strongly suggested that expression changes are poorly explained by TF binding changes [73], underscoring the need to examine evolutionary questions about TF binding in a functional context. It is difficult to generally predict whether a TF ChIP peak is functional, but there are a few well-characterized regulatory systems where detailed prior knowledge of the regulatory network permits such an exercise. One of these systems is the mesoderm specification network in *D. melanogaster*, where extensive prior work has established the role of a small set of TFs in determining spatio-temporal expression patterns of a large number of genes [101, 102, 103, 104, 74, 105, 106, 107, 108, 109, 110]. This has previously led to the cataloging of thousands of putative enhancers responsible for such patterning [74, 110], with hundreds of them being experimentally validated through reporter assays in transgenic embryos. The mesoderm network with its richness of prior knowledge and *cis*-regulatory data sets thus provides a uniquely suited system to investigate cross-species evolution of TF binding and its determinants [74].

In Chapter 4, we studied the evolution of genome-wide binding landscapes of five essential TFs in the mesoderm specification network, between two drosophilids *D. melanogaster* and *D. virilis*, species separated by 40 million years [73] (1.4 substitutions per neutral site [111]). We collected DNase I hypersensitive sites (DHS) data to measure chromatin accessibility at three different temporal stages during early embryonic development in both *D. melanogaster* and *D. virilis*, and recorded conservation and divergence patterns. We built predictive models that use either motif change or accessibility change to predict stage-specific binding divergence of all five TFs, using our previously reported inter-species ChIP data [74, 110]. Using these models and focusing on a large set of previously characterized mesoderm enhancers [74, 110] to increase functional relevance, we found that accessibility and TF binding motif changes have similar predictive relationship with changes in TF binding. We also noted that they bear complementary information and showed that a model using both accessibility and motif information can predict TF binding divergence with significantly greater accuracy than models using either type of information alone. Finally, in a novel analysis, we used machine learning models to examine changes in TF binding of multiple factors in terms of their combinatorial effects on gene expression. We found that motif and accessibility based predictors of TF binding change can substitute for experimentally measured binding change, for the purpose of predicting divergence in gene expression.

# CHAPTER 2: INCORPORATING CHROMATIN ACCESSIBILITY INTO SEQUENCE-TO-EXPRESSION MODELING

Prediction of gene expression levels from regulatory sequences is one of the major challenges of genomic biology today. A particularly promising approach to this problem is taken by thermodynamics-based models that interpret an enhancer sequence in a given cellular context specified by transcription factor concentration levels and predict precise expression levels driven by that enhancer. Such models have so far not accounted for the effect of chromatin accessibility on transcription factor – DNA interactions and consequently on gene expression levels. This chapter describes a thermodynamics-based model of gene expression, called GEMSTAT-A [64], which incorporates chromatin accessibility data and quantify its effect on accuracy of expression prediction. The results demonstrate how DNA accessibility may be useful for sequence-to-expression models.

## 2.1 A THERMODYNAMICS-BASED MODEL THAT INTEGRATES CHROMATIN ACCESSIBILITY DATA

The new quantitative model for predicting gene expression, called GEMSTAT-A (GEMSTAT with Accessibility), is an extension of GEMSTAT (see Section 1.2). GEMSTAT-A integrates chromatin accessibility data to explore the interplay between accessibility, TF-DNA binding strength and gene expression (Figure 2.1). We first assigned a local accessibility score, $\mathrm{Acc}(S)$, on a scale of 0 to 1 (0 = inaccessible), to each TF binding site $S$. Next, the TF-DNA binding energy at site $S$ is modulated by this accessibility score and redefined to be:

$$E(S) = E_0(S) + k_{\mathrm{acc}}(1 - \mathrm{Acc}(S)) \tag{2.1}$$

where $E_0(S)$ is the TF-DNA binding energy at site $S$ as estimated by GEMSTAT using the TF's motif, and $k_{\mathrm{acc}} > 0$ is a free parameter optimized in course of fitting the data and is a phenomenological parameter reflecting the effect of accessibility. Thus, instead of setting a threshold to define accessible and inaccessible TF binding sites, GEMSTAT-A uses quantitative accessibility scores in calculating the binding energy.

In GEMSTAT (i.e., the original model), every TF binding site is considered to be completely accessible, which is equivalent to setting local accessibility to 1. (Note that setting $\mathrm{Acc}(S) = 1$ implies $E(S) = E_0(S)$ in the above formula.) In reality, if the local accessibility is low ($\mathrm{Acc}(S) < 1$), GEMSTAT may overestimate the contribution of the site by ignoring its accessibility score. GEMSTAT-A increases the binding energy (decreases the strength)

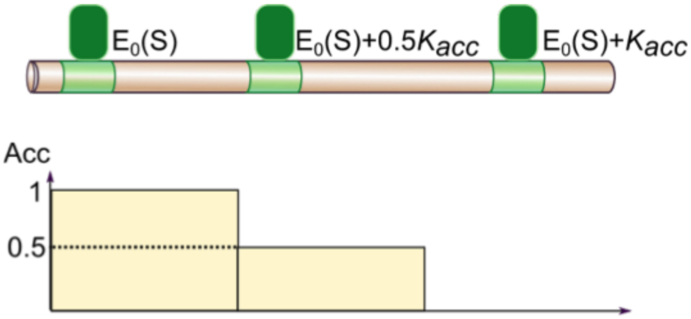

Figure 2.1: **GEMSTAT-A assumes the TF-DNA binding energy at a site $S$ changes according to the accessibility of $S$.** Shown is an example with three identical binding sites where GEMSTAT estimates the same TF-DNA binding energy $E_0(S)$. GEMSTAT-A assigns a local accessibility score Acc($S$) to each site $S$ (bottom, y-axis), and models the TF-DNA binding energy as $E_0(S) + k_{\text{acc}}(1 - \text{Acc}(S))$

of less accessible sites while maintaining the original estimates for sites in highly accessible regions. Other than this modification of how the binding energy is estimated, GEMSTAT-A is identical to GEMSTAT in how enhancer sequence and trans context is mapped to the expression level driven by the enhancer. Note that GEMSTAT-A has one additional free parameter to be optimized, viz., the accessibility effect parameter $k_{\text{acc}}$.

## 2.2 GEMSTAT-A MODELING ON EARLY STAGE *DROSOPHILA* EMBRYOS

We asked if GEMSTAT-A could fit expression profiles of real enhancers better than GEMSTAT, by making use of experimentally measured accessibility variations within the enhancer. To test this, we resorted to a data set that was used in the original GEMSTAT model [19]. The data set comprises: (1) 37 experimentally characterized enhancers involved in the regulation of A/P patterning genes in stage 4-6 *Drosophila* embryos, (2) quantitative profile of the gene expression pattern driven by each enhancer, (3) DNA-binding motifs (expressed as position weight matrices or "PWM"s) of six TFs, namely bicoid (*BCD*), caudal (*CAD*), hunchback (*HB*), giant (*GT*), knirps (*KNI*), and Kruppel (*KR*), and (4) quantitative profile of each TF's concentration (Figure 2.2). He et al. [19] collected the sequences from REDfly database [112], TF concentration profiles from FlyEx database [113, 114], gene expression profiles from Segal et al. [22], the PWM of *BCD* from Bergman et al. [115] and those of the other TFs from Noyes et al. [68]. Following the GEMSTAT study, we chose to model gene expression within 20-80% of the A/P axis.

Additionally, GEMSTAT-A was made to utilize rank-normalized chromatin accessibility data. We gathered chromatin accessibility data from DNase1 hypersensitivity (DHS) assays

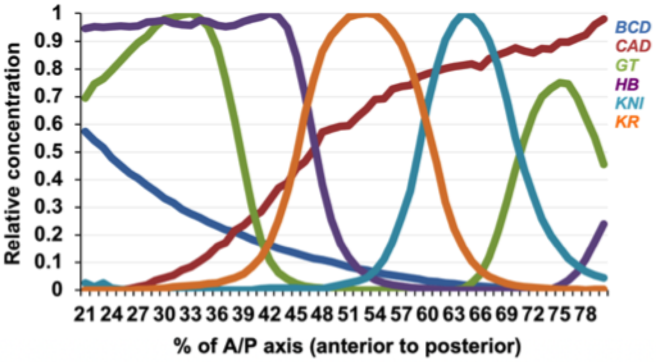

Figure 2.2: TF concentrations (y-axis) for *BCD, CAD, GT, HB, KNI, KR* along the A/P axis (x-axis).

in embryonic stage 5 were gathered from Berkeley Drosophila Transcription Network Project (BDTNP) Release 5 [26, 39]. We ranked the genome-wide DHS scores (at 20 bps resolution), with rank 1 representing the smallest DHS score. The rank-ordered DHS scores were then divided by the total number of windows in the genome. These normalized scores were on the scale of 0 (least accessible) to 1 (most accessible). Rank-based normalized DHS scores within the 37 enhancers were extracted and used to compute the accessibility score Acc($S$) of each annotated binding site S. Acc($S$) was simply the rank normalized score of the 20 bps segment that includes the site S, or the average of multiple segments if the site overlaps with multiple segments.

The GEMSTAT-A model was trained using the same strategy as was used for GEMSTAT [19]. Details of model training are described in Section 1.2. We briefly review the main ideas of GEMSTAT training here. For each TF, all PWM matches in an enhancer with LLR score at least 0.4 times the LLR score of the optimal site were annotated as binding sites. Additionally, in GEMSTAT-A each annotated site is assigned a local accessibility score Acc($S$) as described above, in estimating the TF-DNA binding energy at $S$. Both models considered self-cooperative DNA binding of *BCD* as well as *KNI* and were used in the "DIRECT INTERACTION" mode. The number of free parameters in GEMSTAT was 15 (one transcriptional effect and one binding strength parameters for each TF, one parameter to model basal level of gene expression, and one parameter for each TF that we assumed to have self-cooperative DNA binding. See Table 1.1 for details.), while GEMSTAT-A had one additional free parameter (the "accessibility effect" parameter $k_{\text{acc}}$ ). Model parameters were fit to maximize the average goodness-of-fit score between model predictions and real expression profiles.

State of the art quantitative models of gene expression adopt two common approaches to evaluate their predictions, namely the average correlation coefficient and root mean square error. However, these do not always capture the salient features of a one-dimensional expression pattern, as shown in [40]. To address these issues, a new scoring function, called "weighted pattern generating potentials" (wPGP) was presented by Samee and Sinha [24]. This scoring function was designed to (1) be sensitive to both the shape and magnitude of the predicted expression profiles, and (2) avoid biases towards or against overly broad or overly narrow domains of expression. We therefore used wPGP scores to evaluate the goodness-of-fits (See Section 1.1 for details).

## 2.3 CHROMATIN ACCESSIBILITY DATA IMPROVES EXPRESSION PREDICTIONS

Expression predictions from GEMSTAT and GEMSTAT-A for each enhancer were evaluated using the wPGP score (Figure 2.3 and Table 2.1). Overall, GEMSTAT-A was evaluated at a wPGP score of 0.773 (averaged over 37 enhancers) while GEMSTAT showed an average wPGP of 0.745. Average cross-validation wPGP is 0.741 for GEMSTAT-A and 0.679 for GEMSTAT. Among all 37 enhancers, GEMSTAT-A produced better fits (wPGP score improved by $\geq 0.05$) on 15 enhancers (Figure 2.4A), while its fits were worse than GEMSTAT on 6 enhancers (Figure 2.4B). Within the former group of 15 better-predicted enhancers, the average wPGP score improved by 0.18. These 21 cases included enhancers where one of the models had a wPGP score $\leq 0.5$, which in our experience (also see Figure 2.4C) is a sign that the model failed completely on that enhancer; the differences in fits on these enhancers are likely not due to consideration of accessibility data directly, but due to the different parameter settings the two models utilize. Ignoring these cases, we may identify 11 cases where GEMSTAT-A fits the data better and 4 enhancers where it fits worse than GEMSTAT (last column in Table 2.2). We interpret this as strong evidence that incorporating chromatin accessibility data improves gene expression predictions. To better appreciate the nature of differences between the two models in their fits and to qualitatively assess the improvement due to accessibility information, we plotted the model predictions along with real expression patterns for a selection of enhancers (Figure 2.4 and Figure 2.5).

We noted that on some enhancers GEMSTAT-A fits showed refinements over GEMSTAT predictions resulting in more accurately defined boundaries of expression domains (e.g., btd_head, hb_anterior_actv, eve_37ext_ru in Figure 2.5). On other enhancers there were more qualitative improvements, e.g., GEMSTAT-A correctly models the posterior domain of gt_(-1), correctly removes a spurious anterior domain prediction made by GEMSTAT on the enhancer pdm2_(+1), and dramatically improves upon the boundaries of the predicted expression domain of enhancer nub_(-2). Interestingly, the change in GEMSTAT-A's pre-

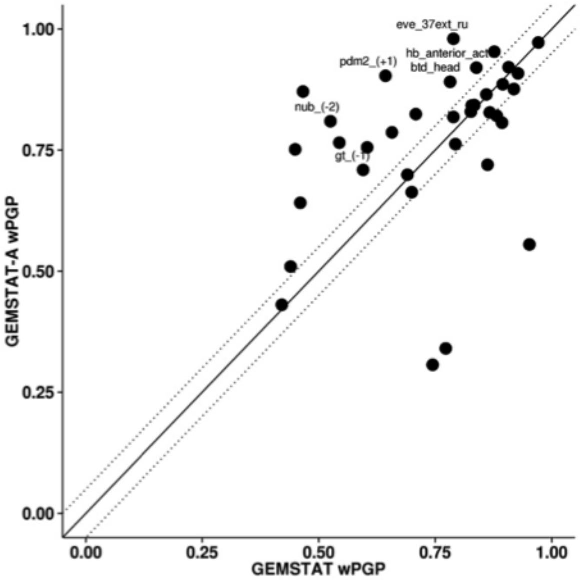

Figure 2.3: **Evaluations of expression predictions from GEMSTAT and GEMSTAT-A.** The goodness of fit between predicted and real expression for each enhancer was assessed by wPGP score, shown here for all 37 enhancers. Dotted lines delineate regions where the difference of w-PGP between the two models is ≥ 0.05. A selection of enhancers where GEMSTAT-A improves fits are labeled and their expression patterns are shown in Figure 2.5.

Table 2.1: **10-fold cross-validation assessment of GEMSTAT and GEMSTAT-A**. Each model was tested with 10-fold cross-validation, repeated five times with different (random) definitions of the ten folds. For each model, shown are the number of free parameters used is shown ("#Pars"), the wPGP score from parameter optimization over all 37 enhancers ("Training wPGP"), and the wPGP score from cross-validation ("CV wPGP"), averaged (with standard deviation, SD, in parentheses) over the five repeats.

| **Model** | **#Pars** | **Training w-PGP** | **CV wPGP (SD)** |
|---|---|---|---|
| GEMSTAT-A | 16 | 0.773 | 0.741 (0.005) |
| GEMSTAT | 15 | 0.745 | 0.679 (0.008) |

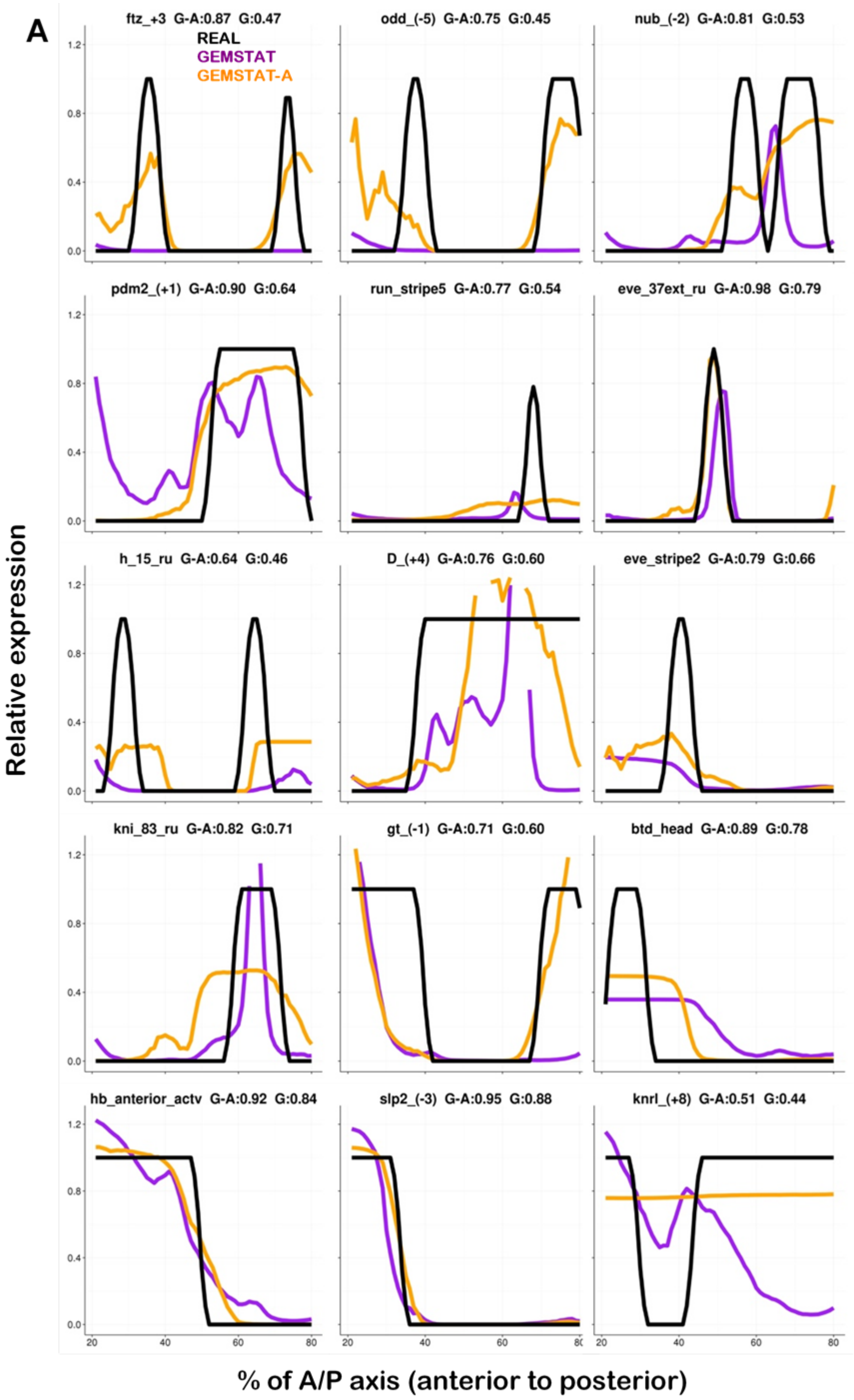

Figure 2.4: **Expression predictions from GEMSTAT and GEMSTAT-A.** The predicted expression profiles of GEMSTAT-A (*orange lines*) and GEMSTAT (*purple lines*) are compared to experimentally determined readouts (*black lines*), for 9 selected CRMs. Each expression profile is on a relative scale of 0 to 1 (y-axis), and shown for the region between 20% egg length and 80% egg length along the A/P axis of the embryo. Title in each panel is in the format of "enhancer, wPGP by GEMSTAT-A (G-A), wPGP by GEMSTAT (G)." (A) 15 enhancers with wPGP score improved by $\geq 0.05$. The order of enhancers is the same as in Table 2.2.

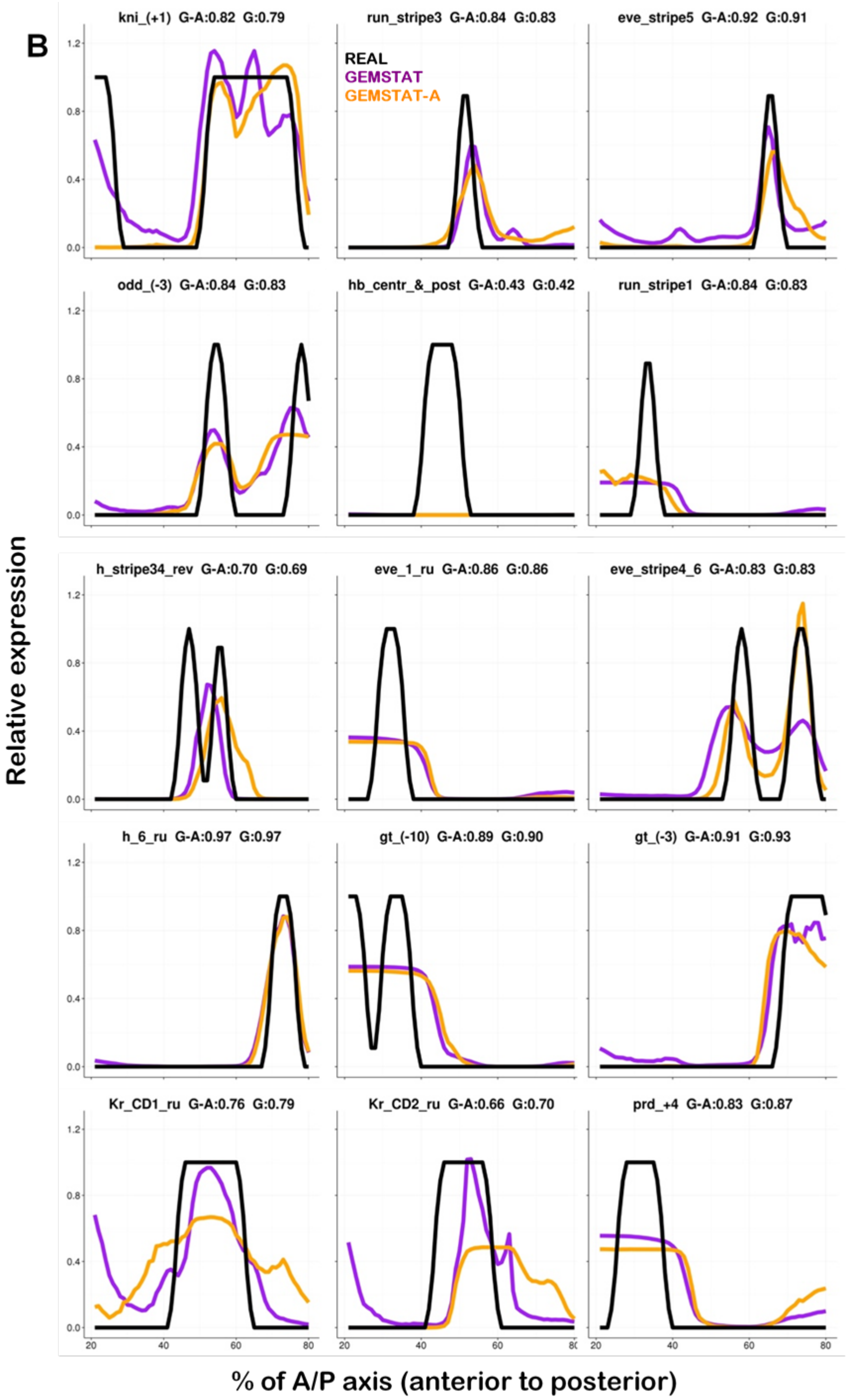

Figure 2.4 (cont'd): **Expression predictions from GEMSTAT and GEMSTAT-A.** (B) 16 enhancers with no substantial change.

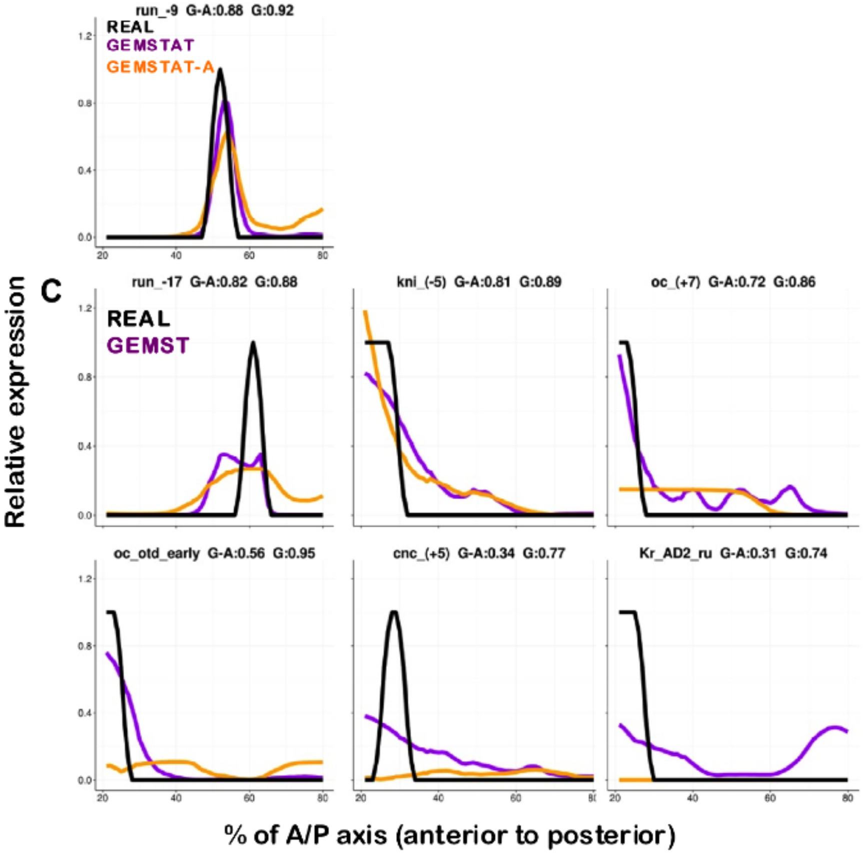

Figure 2.4 (cont'd): **Expression predictions from GEMSTAT and GEMSTAT-A.** (C) 6 enhancers with wPGP scores worsened by ≥ 0.05.

Table 2.2: **Evaluations of expression predictions from GEMSTAT and GEMSTAT-A.** The "goodness of fit" between predicted and real expression for each enhancer was assessed by wPGP score. The wPGP scores from GEMSTAT and GEMSTAT-A over all 37 enhancers are shown, and wPGP scores greater than 0.75 are colored in red.

| **Enhancer** | **GEMSTAT-A** | **GEMSTAT** | **Change $\geq$ 0.05** | **Change $\geq$ 0.05 and both $\geq$0.50** |
|---|---|---|---|---|
| ftz_+3 | 0.87 | 0.47 | + | |
| odd_(-5) | 0.75 | 0.45 | + | |
| nub_(-2) | 0.81 | 0.53 | + | + |
| pdm2_(+1) | 0.90 | 0.64 | + | + |
| run_stripe5 | 0.77 | 0.54 | + | + |
| eve_37ext_ru | 0.98 | 0.79 | + | + |
| h_15_ru | 0.64 | 0.46 | + | |
| D_(+4) | 0.76 | 0.60 | + | + |
| eve_stripe2 | 0.79 | 0.66 | + | + |
| kni_83_ru | 0.82 | 0.71 | + | + |
| gt_(-1) | 0.71 | 0.60 | + | + |
| btd_head | 0.89 | 0.78 | + | + |
| hb_anterior_actv | 0.92 | 0.84 | + | + |
| slp2_(-3) | 0.95 | 0.88 | + | + |
| knrl_(+8) | 0.51 | 0.44 | + | |
| kni_(+1) | 0.82 | 0.79 | | |
| run_stripe3 | 0.84 | 0.83 | | |
| eve_stripe5 | 0.92 | 0.91 | | |
| odd_(-3) | 0.84 | 0.83 | | |
| hb_centr_&_post | 0.43 | 0.42 | | |
| run_stripe1 | 0.84 | 0.83 | | |
| h_stripe34_rev | 0.70 | 0.69 | | |
| eve_1_ru | 0.86 | 0.86 | | |
| eve_stripe4_6 | 0.83 | 0.83 | | |
| h_6_ru | 0.97 | 0.97 | | |
| gt_(-10) | 0.89 | 0.90 | | |
| gt_(-3) | 0.91 | 0.93 | | |
| Kr_CD1_ru | 0.76 | 0.79 | | |
| Kr_CD2_ru | 0.66 | 0.70 | | |
| prd_+4 | 0.83 | 0.87 | | |
| run_-9 | 0.88 | 0.92 | | |
| run_-17 | 0.82 | 0.88 | - | - |
| kni_(-5) | 0.81 | 0.89 | - | - |
| oc_(+7) | 0.72 | 0.86 | - | - |
| oc_otd_early | 0.56 | 0.95 | - | |
| cnc_(+5) | 0.34 | 0.77 | - | |
| Kr_AD2_ru | 0.31 | 0.74 | - | - |

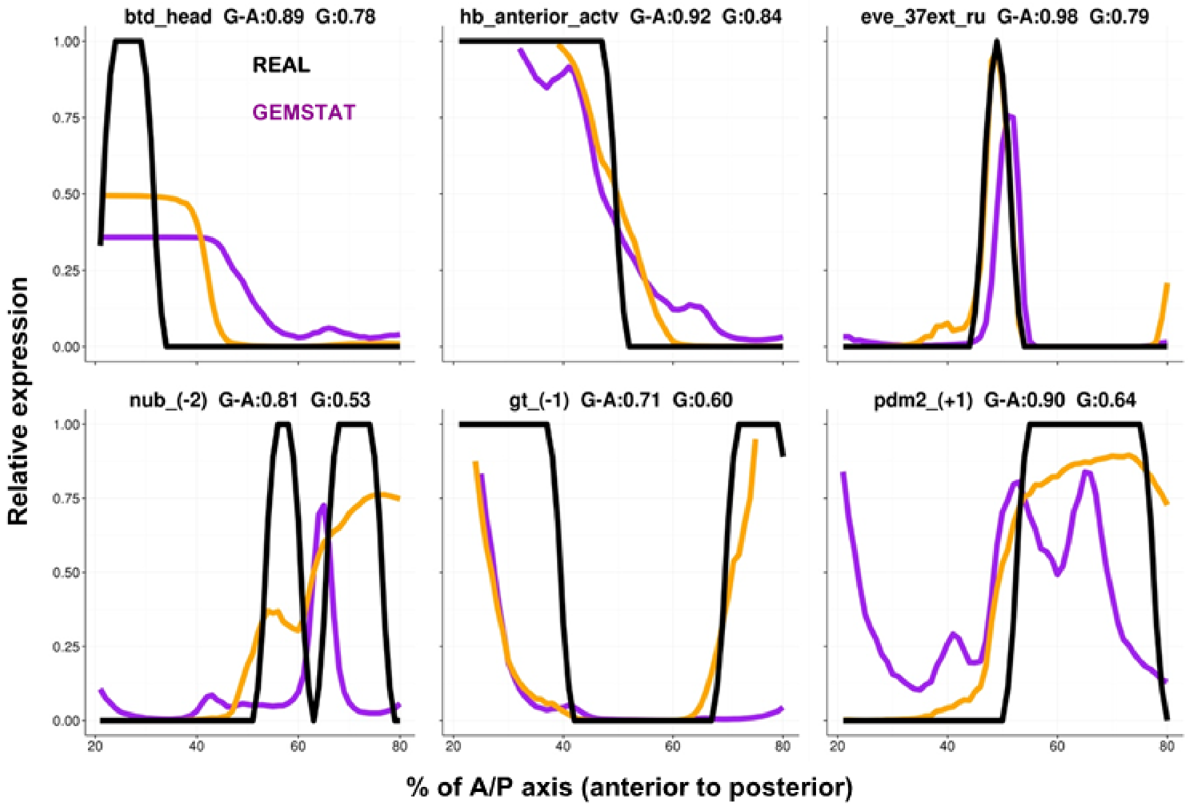

Figure 2.5: **Expression predictions from GEMSTAT and GEMSTAT-A.** The predicted expression profiles of GEMSTAT-A (*orange lines*) and GEMSTAT (*purple lines*) are compared to experimentally determined readouts (*black lines*), for 6 selected CRMs. Each expression profile is on a relative scale of 0 to 1 (y-axis), and shown for the region between 20% and 80% of the A/P axis of the embryo. Title in each panel is in the format of "enhancer name, wPGP by GEMSTAT-A (G-A), wPGP by GEMSTAT (G)."

Table 2.3: **10-fold cross-validation assessment.** GEMSTAT and GEMSTAT-A models were tested with 10-fold cross-validation 5 times. For each 10-fold cross-validation run, the wPGP scores of GEMSTAT and GEMSTAT-A (averaged over 37 enhancers, "Avg. wPGP") are shown.

| **Run #** | **GEMSTAT Avg. wPGP** | **GEMSTAT-A Avg. wPGP** |
|---|---|---|
| 1 | 0.676 | 0.748 |
| 2 | 0.666 | 0.745 |
| 3 | 0.685 | 0.736 |
| 4 | 0.684 | 0.742 |
| 5 | 0.685 | 0.737 |

diction for some cases is more accurate biologically, although the prediction does not match the data. For example, the posterior expression in our predicted readout for eve_37ext_ru is indeed in those locations along the A/P axis where the seventh stripe of the eve gene is formed. Detail comparison of relative successes and failures, as well as examples where one model completely failed to capture the spatial pattern driven by an enhancer while the other model was successful, are shown in Figure 2.4.

Thus, our initial observations on model fits over all enhancers indicated, both quantitatively and qualitatively, a conspicuous improvement due to chromatin accessibility data. Rigorously speaking, GEMSTAT-A fits are expected to be at least as good as GEMSTAT since the former has one extra parameter, the accessibility effect $k_{\text{acc}}$. A common way to compare models of varying complexity is to evaluate their cross-validation accuracy. We therefore performed 10-fold cross-validation with either model, where each "fold" uses 33-34 of the 37 enhancers as training data and the remaining 3-4 enhancers as the testing data. Since partitioning of the 37 enhancers into ten folds is done at random, we repeated the entire 10-fold cross validation exercise five times (with different random partitioning in each repeat) for each model. The average cross-validation wPGP across all 5 runs of 10-fold cross-validation was 0.679 and 0.741 for GEMSTAT and GEMSTAT-A, respectively (Table 2.1). Detailed results from cross-validation are shown in Table 2.3. This analysis clearly shows the improved ability of GEMSTAT-A to predict expression readouts, compared to GEMSTAT, even after accounting for the additional free parameter.

To verify the above effect further, we next repeated the modeling exercise, by (1) using accessibility data from embryonic stage 14, which is a later developmental stage compared to embryonic stage 5 that the expression data corresponds to, (2) using a randomly shuffled version of the normalized accessibility scores across the whole genome and extracting local accessibility profiles in the 37 enhancers, and (3) shuffling the accessibility scores across the 37 enhancers. (The last exercise was motivated by the fact that enhancers are known to have

Table 2.4: **Effect of chromatin accessibility data used in GEMSTAT-A.** Results from GEMSTAT-A trained with different variations on input chromatin accessibility data: data from embryonic developmental stage 5 (stage matching the modeled expression patterns), embryonic stage 14 (mismatched stage), or two different randomly shuffled versions of the stage 5 data (see text). Also shown, in the first row, is the result from GEMSTAT, which does not use accessibility data. For each variation of input accessibility data, shown are the w-PGP score (averaged over 37 enhancers) and optimized value of the accessibility effect parameter ($k_{\mathrm{acc}}$ ). Results in the last two rows (shuffled versions of stage 5 accessibility data) are averaged over three different repeats of the assessment (using different random shuffling).

| **Model** | **DNA accessibility data** | **w-PGP** | $\boldsymbol{k_{\mathrm{acc}}}$ |
|---|---|---|---|
| GEMSTAT | no accessibility data | 0.745 | N/A |
| GEMSTAT-A | Embryonic stage 5 | 0.773 | 17.6 |
| GEMSTAT-A | Embryonic stage 14 | 0.742 | 4.62 |
| GEMSTAT-A | Shuffling across whole genome | 0.734 | 0.91 |
| GEMSTAT-A | Shuffling across all enhancers | 0.735 | 0.97 |

Table 2.5: **Effect of shuffling DNA accessibility data used in GEMSTAT-A.** GEMSTAT-A was applied with two different types of shuffled DNA accessibility data: shuffled across whole genome and shuffled across all 37 enhancers. For each runs of shuffled DNA accessibility data, the average wPGP (“Avg. wPG”) is shown.

| **Run #** | **Shuffling across whole genome** | **Shuffling across all enhancers** |
|---|---|---|
| 1 | 0.739 | 0.735 |
| 2 | 0.732 | 0.731 |
| 3 | 0.733 | 0.739 |

higher accessibility in general, and genome-wide permutation of accessibility scores is likely to assign low accessibility values within enhancers, thus presenting an unrealistic random control.)

GEMSTAT-A was trained on these three different “incorrect” settings of chromatin accessibility data and then evaluated by the wPGP score (Table 2.4 and Table 2.5). In all three cases, the advantage of GEMSTAT-A over GEMSTAT was entirely lost, and the optimal value of the$k_{\mathrm{acc}}$ parameter reported was weak or close to 0, suggesting that the model found no advantage to using the incorrect accessibility data. These negative controls thus confirmed that the improved fits found by GEMSTAT-A are mainly due to use of chromatin accessibility data from the appropriate developmental stage.

## 2.4 GEMSTAT-A LEARNS MUCH STRONGER THERMODYNAMIC PARAMETERS

In the process of training sequence-to-expression models, information about inputs (enhancer sequences and trans-regulatory context) and output (expression pattern driven by enhancers) of a regulatory function is used to automatically learn values for the free parameters of the model. Both GEMSTAT and GEMSTAT-A utilize two free parameters for each TF. One of these TF-specific parameters is called the DNA binding weight parameter ("bindingWt"); it helps estimate the occupancy of the TF at a binding site. The other is called the transcription effect parameter ("txpEffect"), which represents the strength of activation or repression due to a DNA-bound TF molecule. These parameters have intuitive semantics, hence their optimal values reported by a trained model are of interest; for example, these values indicate if TFs bind their respective consensus site strongly or weakly, if one activator is more effective than another, etc. In other words, the trained model parameters paint a quantitative picture of the underlying regulatory mechanisms. It is natural to ask if two models trained on the same data, identical in all respects except that one is aware of accessibility data and one is not, suggest similar quantitative views of the underlying mechanistic reality. We examined the optimal values of the bindingWt and txpEffect parameters for each of the six TFs used in the model, as learned by GEMSTAT and GEMSTAT-A separately. We were surprised to see that the same parameters were often trained to very different values: GEMSTAT-A was found to learn much stronger parameters (in some cases one to two orders of magnitude stronger) than GEMSTAT. The bindingWt parameter of both activators and repressors was assigned a greater value (stronger binding strengths) by GEMSTAT-A compared to GEMSTAT (Figure 2.6 and Table 2.6). The bindingWt parameter of *HB* was around 50-fold greater in GEMSTAT-A, while that of *KNI* was about 13-fold greater. The txpEffect parameter describes the regulatory effect of a TF and takes values greater than 1 for activators and less than 1 for repressors. We observed that GEMSTAT-A assigned about two-fold greater values to the activator TFs *BCD* and *CAD*, compared to values learned by GEMSTAT. Likewise, for three of the four repressor TFs (*GT*, *KNI*, and *KR*), GEMSTAT-A assigned lower txpEffect values reflecting stronger repression ability, especially in the case of *KR*, whose txpEffect was  20-fold stronger in GEMSTAT-A.

## 2.5 GEMSTAT-A IMPROVES EXPRESSION PREDICTION BY REDUCING THE CONTRIBUTION OF INACCESSIBLE BINDING SITES

We showed above that GEMSTAT-A is able to achieve better predictions of enhancer readouts with a simple modification of the estimated binding energy of a TF at its sites.

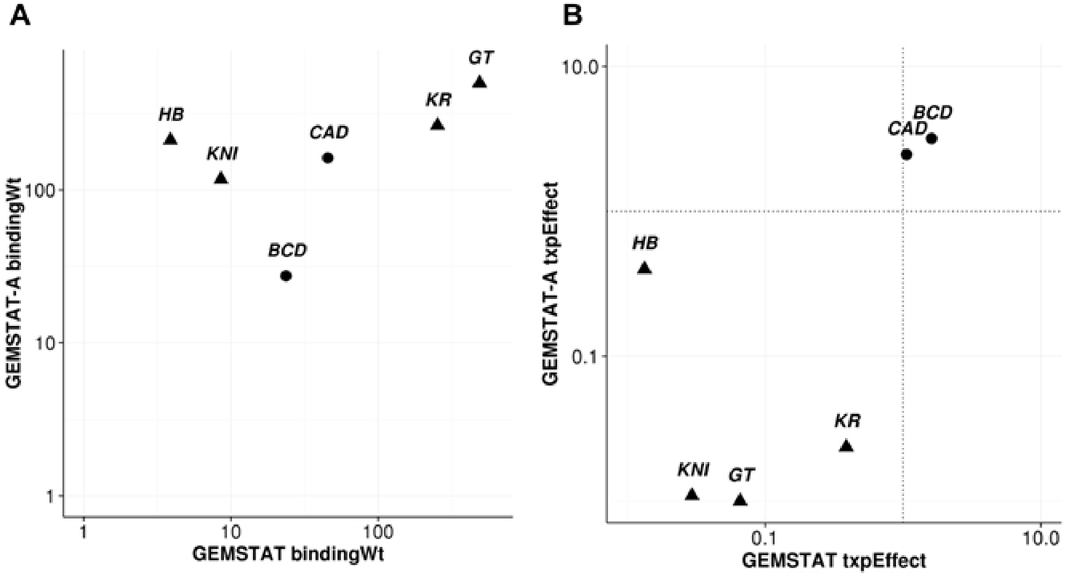

Figure 2.6: **GEMSTAT-A learns stronger parameter values.** The bindingWt (A) and txpEffect (B) parameters of each TF learned from GEMSTAT (x-axis) and GEMSTAT-A (y-axis). Both axes are on logarithmic scale, in either plot. Icons are in triangle for repressors and circle for activators. The txpEffect parameter for an activator is greater than 1 and higher values indicate stronger activation. It is less than 1 for repressors and lower values indicate stronger repression.

Table 2.6: **GEMSTAT-A learns stronger parameters than GEMSTAT on the same data set.** The bindingWt and txpEffect parameters of each TF learned from GEMSTAT-A and GEMSTAT are shown.

| **TF** | **GEMSTAT-A bindingWt** | **GEMSTAT bindingWt** | **GEMSTAT-A txpEffect** | **GEMSTAT txpEffect** |
|---|---|---|---|---|
| *BCD* | 27.38 | 23.70 | 3.18 | 1.61 |
| *CAD* | 161.62 | 45.51 | 2.47 | 1.06 |
| *GT* | 499.98 | 490.17 | 0.01 | 0.07 |
| *HB* | 211.45 | 3.89 | 0.40 | 0.01 |
| *KNI* | 117.55 | 8.58 | 0.01 | 0.03 |
| *KR* | 264.23 | 253.64 | 0.02 | 0.39 |

This suggests the existence of TF binding sites in inaccessible segments within the enhancer, which GEMSTAT was forced to incorporate in its predictions but which GEMSTAT-A could ignore, by exploiting accessibility information. We investigated this potential explanation of why GEMSTAT-A produces better fits. For each annotated binding site within the enhancer (recall that these are identical between the two models), we removed the accessibility information for that site only, designating it as completely accessible (Acc($S$) = 1), and re-computed the expression profile predicted by GEMSTAT-A. The new goodness-of-fit (wPGP) was calculated and compared to the original wPGP score of GEMSTAT-A for that enhancer. The difference in wPGP values, for the same model with or without use of accessibility information on that site, was plotted for each site (Figure 2.7A, ΔwPGP). We also plotted the change in estimated binding energy of each site due to incorporation of local accessibility values (Figure 2.7B, ΔΔE). (Parameters were not re-trained in this analysis.)

Figure 2.7 shows examples of the above-mentioned explanation of how GEMSTAT-A improves fits by weakening the estimated binding energy of sites in less accessible regions. One such example is that of the enhancer gt_(-1), where both GEMSTAT and GEMSTAT-A correctly predict the anterior domain, but the posterior domain ( ~70% to ~80% of A/P axis) is not predicted by GEMSTAT and correctly predicted by GEMSTAT-A (Figure 2.7C, left). A natural explanation for this difference is that binding sites capable of repressing expression in the posterior domain are present in less accessible regions of the enhancer, and while GEMSTAT-A ignores their potential contribution, GEMSTAT includes this contribution leading to the absence of a posterior domain in its prediction. Indeed, Figure 2.7A (left) shows that a binding site of the repressor $GT$ located at position ~250 in the enhancer is one such site: if GEMSTAT-A were to designate this site as accessible its goodness of fit (wPGP) would diminish by ~0.03. Figure 2.7B shows that the estimated binding energy of this $GT$ site was indeed lower due to local accessibility values. The same figure also shows a $KR$ site ( at position ~300 ) that is inaccessible, but whose accessibility score is not relevant to the fits of GEMSTAT-A for this enhancer.

A similar explanation applies to the enhancer pdm2_(+1), for which GEMSTAT incorrectly predicted an anterior domain of expression while GEMSTAT-A correctly predicted lack of expression in the anterior (Figure 2.7C, middle). The natural explanation for this difference is the existence of an activator site capable of driving anterior expression, whose local inaccessibility leads GEMSTAT-A to ignore the site but whose inclusion leads GEMSTAT to predict the spurious anterior expression. Figure 2.7B (middle) shows that there are several $BCD$ sites in the enhancer satisfying this property; $BCD$ is expressed anteriorly (Figure 2.2) and its sites are therefore capable of causing GEMSTAT to predict anterior expression unless their effect is ignored based on local chromatin inaccessibility. Thus, these

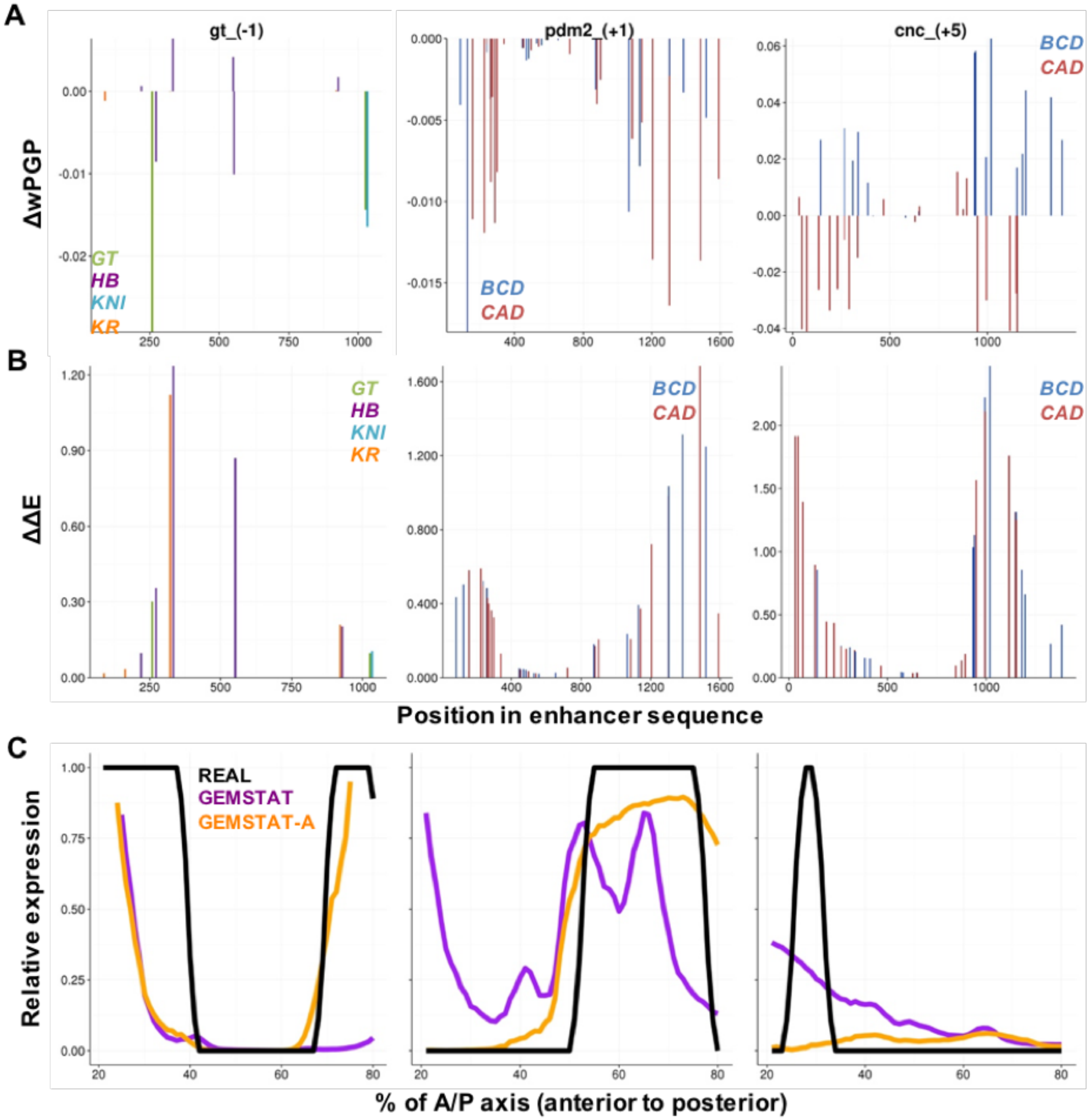

Figure 2.7: **Accessibility of individual sites is utilized by GEMSTAT-A to improve predictions.** Details of GEMSTAT-A modeling on enhancers gt_(-1), pdm2_(+1) and cnc_(+5) are shown in left, middle and right columns respectively. (A) Change in goodness of fit (ΔwPGP) of GEMSTAT-A predictions when a binding site's accessibility score is forced to a value of 1 (maximum accessibility), shown for each site as a function of its location in the enhancer. (B) Reduction in estimated binding energy (ΔΔE) due to local accessibility is shown for each annotated binding site, as a function of the site's location in the enhancer sequence. Only sites for a subset of TFs (all repressors in left panels, both activators in middle and right panels) are shown. (C) Predicted expression profiles of GEMSTAT-A (*orange lines*) compared to GEMSTAT predictions (*purple lines*) and experimentally determined readouts (*black lines*).

two examples provide deeper insights into how GEMSTAT-A can use local accessibility to suppress the activating or repressive effects of binding sites, leading to more accurate predictions of enhancer readout.

The above analysis also explains why GEMSTAT-A performed poorly on a few enhancers. One such example is the enhancer cnc_(+5) where GEMSTAT-A failed to predict the anterior expression domain (Figure 2.7C, right). This enhancer has several *BCD* sites in relatively inaccessible locations (Figure 2.7B, right), and by ignoring or diminishing their potential activating influence GEMSTAT-A loses its ability to predict the anterior domain. Indeed, if it were to ignore the accessibility scores of these sites (i.e., assume that they are accessible), its wPGP value would improve, as revealed by Figure 2.7A (right). Such aberrant cases were rare in our evaluations, and may be attributed to the spatial resolution of accessibility data (see Discussion), among other possibilities.

## 2.6 DISCUSSION

Quantitative models such as GEMSTAT have been shown to have the expressive power to capture the complex relationship between regulatory sequence and precise gene expression patterns, i.e., the so-called *cis*-regulatory "code" [14, 116]. Their appeal lies in achieving this expressiveness within a biophysically motivated framework (so that fit models can be interpreted more easily), while making simplifications that hide mechanistic details on which little data is available. One such simplification heretofore has been to model TF-DNA binding as entirely determined by the binding site and the position weight matrix, by adopting Berg & von Hippel theory [52, 57]. The role of local chromatin structure and epigenetic modifications has been ignored in these models, understandably so since appropriate data for learning this role has been lacking. (Also, the few existing models for predicting nucleosome occupancy profiles [117, 118, 119] have not reached the level of accuracy necessary for coupling them to enhancer models; data not shown.) However, the recent wave of studies profiling the chromatin landscape, especially DNA accessibility, in specific cell types [120] or developmental stages [25] has changed this situation. Our work responds to this exciting new development in regulatory genomics by incorporating DNA accessibility data into sequence-to-expression models and asking if this can at least partly address the limitations introduced by the simplification mentioned above. We find the answer to be in the affirmative, at least in the context of our modeling framework and the data set analyzed here.

We note that the role of chromatin accessibility in sequence-based models of gene expression has not been previously studied. There have been several interesting computational analyses of accessibility data, that have shown the prodigious impact of accessibility on TF-

DNA binding profiles [33, 25, 38], and the correlation between changing accessibility and changing expression [121, 120, 39], but these studies do not quantify the impact of accessibility data on sequence-based prediction of precise spatio-temporal expression patterns. We also note that our answer to the above-mentioned question did not have to be affirmative. Even though accessibility clearly shapes expression [27], its influence might have been simply in making the entire enhancer available for function; in this case a modeling study that already begins the assumption of an "open" enhancer will not gain any significant advantage from accessibility data. Our affirmative answer suggests a more nuanced role where variation of accessibility *within* the enhancer carries information useful for the functional interpretation of the binding sites present in the enhancer.

It is worth noting that GEMSTAT-A is a phenomenological extension that adds accessibility information to GEMSTAT. In reality chromatin accessibility is likely the result of complex processes involving the nucleosome, transcription factors, chromatin remodeling factors and DNA (sequence) [122]. Future sequence-to-expression models may strive to incorporate these processes directly at suitable levels of parameterization, with accessibility being an intermediate *dependent* variable predicted from sequence and the cellular context, rather than an independent variable as is the case in GEMSTAT-A. One example of such future work is to model the influence of pioneer factors [123], which exhibit sequence-specific binding and seem to remodel the accessibility profile locally. The transcription factor *ZELDA* is a strong candidate for this special treatment in the context of our data set, with recent studies recording its widespread and significant regulatory influence [124, 125] on many of the gene expression patterns we have modeled here. Computational [33] and experimental [124] work has strongly suggested that this influence is mediated via accessibility, and the *ZELDA* binding motif has been noted as highly enriched in "hot spots" of multi-TF binding [126]. It is expected that a part of the advantage of using accessibility data will be observed if GEMSTAT was modified to use *ZELDA* as a DNA-binding protein that makes local chromatin more accessible. In this work we chose not to use *ZELDA* as one of the regulatory inputs, so as to get a more accurate view of the role of accessibility variations in shaping expression readouts of enhancers.

In principle the data used as input to GEMSTAT-A should correspond to a cell type – in our case, position along the anterior-posterior axis of the embryo. This is the case for the TF concentration profiles used here, with GEMSTAT-A making separate predictions for each bin along the A/P axis, using relative TF concentration values for that bin. However, this is not the case for the accessibility data used, which correspond to whole embryo measurements. We thus believe that the advantage observed by us is an underestimate of what cell type-specific accessibility data, already available in other contexts [120, 27] can confer upon sequence-to-expression models. For instance, the coarseness of accessibility data might negatively impact the accuracy of GEMSTAT-A on an enhancer that functions for a short period of time (compared to the longer period over which the accessibility data is aggregated), or an enhancer driving expression in relatively few cells of the embryo. This may explain some failures of GEMSTAT-A modeling. For the enhancer oc_(+7), for example, we found that sites for *HB* (a repressor which presumably limits the gene's expression in a narrow anterior domain) are mostly in the inaccessible regions (data not shown). This might have caused GEMSTAT-A to predict a broad ectopic expression pattern for this enhancer (Figure 2.4C). It is also worthwhile to note that we used wPGP score to measure the goodness-of-fit. In some cases, wPGP scores do not reflect our visually perceived quality of fit. The wPGP score has been found to be a superior choice in comparison to the two commonly used goodness of fit scores, namely sum of squared errors and correlation coefficient [40, 24]. Our experiences from these published studies were convincing enough for us to choose wPGP as the goodness of fit scores here. Future work will also have to continue to improve the goodness-of-fit score.

An interesting observation made during our model comparisons (with and without accessibility data) was the stronger parameter values learned in GEMSAT-A fits compared to GEMSTAT fits. Stronger parameter values for a TF imply regarding each binding site of that TF to have greater contribution to the enhancer's function. To see why this might be the case, suppose an enhancer has two TF binding sites for the same TF, with the same binding affinity and concentration, but one of the sites is accessible and the other is not. In GEMSTAT, each TF binding site is supposed to be completely accessible, thus the two sites make equal contributions to the gene expression. However, GEMSTAT-A, is aware that one of the binding sites is inaccessible, and will therefore attribute greater contribution to the accessible site in order to achieve the same level of gene expression. This will result in GEMSTAT-A using stronger parameter values.

In conclusion, we have shown here for the first time how thermodynamic models of enhancer readouts may leverage accessibility information to explain the data with higher accuracy. We have commented above on the limits of the accessibility data used here, and expect that the potential shortcomings of using embryo-wide data may be alleviated by refined, cell type-specific data in the future. The current study also makes it more interesting to assess additional mechanisms of accessibility and the role of histone modifications. Finally, while we demonstrated the utility of our modeling for a model organism, the impact of this modeling framework will be much higher if mammalian data on gene expression levels under a large number of different conditions are available, along with experimentally derived knowledge of the major regulators under those conditions. Extending the current framework to mammalian systems will be a major direction for future research.

# CHAPTER 3: QUANTITATIVE MODELING OF GENE EXPRESSION USING DNA SHAPE FEATURES OF BINDING SITES

An important problem in the study of transcriptional regulation is sequence-to-expression modeling, which interprets the enhancer sequence based on transcription factor concentrations and DNA binding specificities and predicts precise gene expression levels in varying cellular contexts. Such models largely rely on the position weight matrix (PWM) model for DNA binding, and the effect of alternative models based on DNA shape remains unexplored. This chapter introduces a statistical thermodynamics model of gene expression using DNA shape features of binding sites [127]. This work demonstrates that the increasingly popular DNA-binding models based on local DNA shape can be useful in sequence-to-expression modeling. It also provides a framework for future studies to predict gene expression better than with PWM models alone.

## 3.1 TFBS DNA SHAPE SCORE PREDICTED BY RANDOM FOREST CLASSIFIER

The main purpose of this work was to test if a quantitative sequence-to- expression model based on DNA shape at putative binding sites provides better fits to expression data than the PWM-based model that has been tested successfully in multiple prior studies [17, 19, 64, 22]. For this, we first trained a classifier to predict binding scores by using DNA shape information. TF binding sites obtained via bacterial one hybrid (B1H) experiments were downloaded from the Fly Factor Survey database [68]. The DNA shape readouts for all binding sites were obtained by DNAShape [67], which predicts values of minor groove width (MGW) and propeller twist (ProT) at base pair (bp) resolution and values of roll (Roll) and helix twist (HelT) at base pair step resolution; the values are calculated using a window approach around each base pair, which will score all base pairs except for the one or two base pairs at each end of the sequence for which we do not have sufficient flanking residues.

Given a TF, we trained a Random Forest classifier [128], using the R package 'randomForest' [129], to predict the shape scores of its binding sites. As shown in Figure 3.1, a TF binding site is characterized by a set of four 'shape vectors' (MGW, ProT, Roll, and HelT); each vector has $d + 2$ dimensions: d dimensions corresponding to a DNA shape readout at each position except for the terminal one or two base pairs, and two corresponding to the mean and standard deviation of DNA shape readouts over all positions in the binding site. The final feature vector fed into the Random Forest classifier was the concatenation of all four shape vectors, a representation we chose partly based on the work of Zhou et al. [67].

To train each Random Forest, we sampled a set of binding sites for a given TF as the

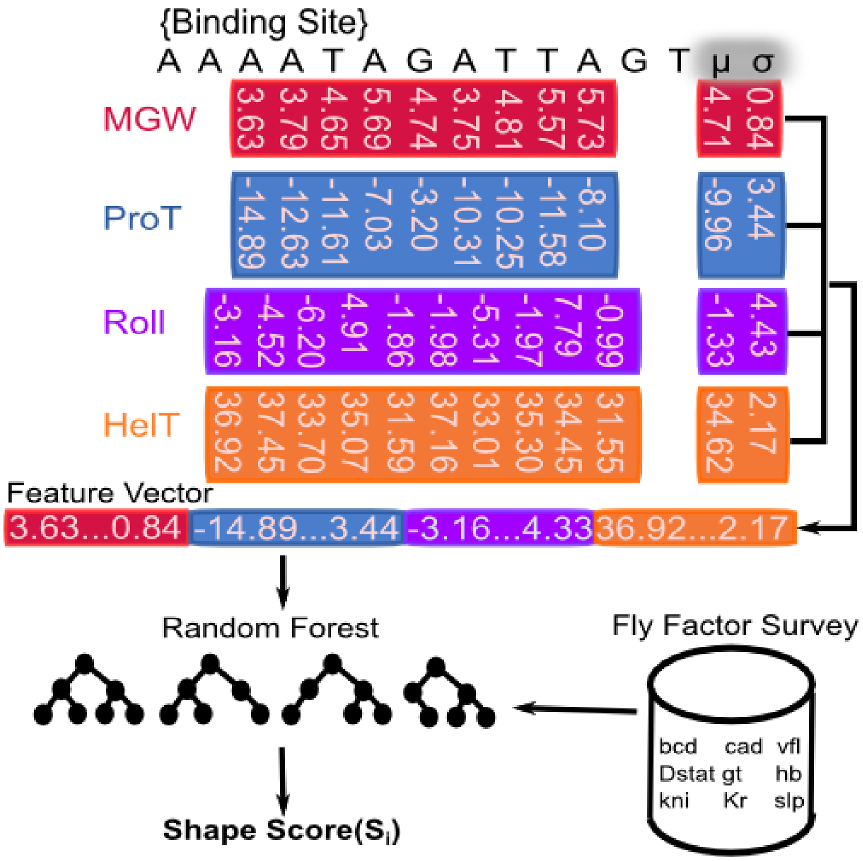

Figure 3.1: **DNA shape-based model of gene expression.** A TF binding site is described by four shape feature vectors: MGW, ProT, Roll, and HelT. Each vector includes the corresponding shape feature at every position of the site, along with the mean and standard deviation over all positions. For a given TF, a random forest classifier is trained on a sample of binding sites from Fly Factor Survey database to predict shape scores for putative binding sites.

positive data and a set of random non-coding genomic regions, each with the same length as the TF's sites, as the negative data. To capture the numerous ways that random sequences can deviate from the TF's preferred binding sequences, we trained each classifier on 10 times as many negative examples as positive examples. We kept the multiplicative factor (10) low as we wanted to prevent the Random Forest from being deluged by negative data to the extent that it suffers from the class imbalance problem [130]. The output of the Random Forest is a probability of the input sequence being 'positive', meaning a TF binding site (TFBS). We denote this probability as the "DNA shape score" in this study.

## 3.2 DNA SHAPE-BASED QUANTITATIVE SEQUENCE TO EXPRESSION MODEL

The DNA shape-based sequence to expression model was adapted from the statistical thermodynamics model GEMSTAT [19]. We review the main ideas of GEMSTAT in Section 1.2 and formulate below the key modification to its architecture that allows it to utilize DNA shape information. The contribution of each binding site to the enhancer's regulatory function is dictated by its 'statistical weight' $q(S)$, given by the following equation:

$$q\left(S\right)=K\left(S_{\text{opt}}\right)\nu\left[\text{TF }\right]_{\text{rel}}\exp\left[\text{LLR}\left(S\right)-\text{LLR}\left(S_{\text{opt}}\right)\right] \tag{3.1}$$

In this formulation, $\left[\text{TF }\right]_{\text{rel}}$represents the relative TF concentration up to some constant $\nu$. $\text{LLR}\left(S\right)-\text{LLR}\left(S_{\text{opt}}\right)$ represents the difference in the log likelihood ratio between the site $S$ and the optimal binding site $S_{\text{opt}}$, and $K\left(S_{\text{opt}}\right)$ represents the association constant of TF-DNA binding. Since both $K\left(S_{\text{opt}}\right))$ and $\nu$ are unknown constants, GEMSTAT treats the product of the two as a free parameter.

In constructing an analogous measure based on DNA shape data and not PWM data, only a single modification needs to be made to the binding site contribution formula, $q(S)$. In particular, the arguments of the exponent are changed to use DNA shape data. In the following formulation, $\text{Shape}(S)$, in the range 0-1, represents the DNA shape score predicted by a Random Forest classifier and $k$ represents a free scaling parameter.

$$q\left(S\right)=K\left(S_{\text{opt}}\right)\nu\left[\text{TF }\right]_{\text{rel}}\exp\left[-k(1-\text{Shape}(S))\right] \tag{3.2}$$

Section 3.3 discusses how to derive this formula and an alternative method for incorporating the shape score into GEMSTAT.

## 3.3 BIOPHYSICAL VIEW OF TF-DNA BINDING

Consider a bimolecular reversible reaction of the TF binding to a short piece of DNA to be represented as

$$\text{TF}+\text{DNA}\xrightleftharpoons{K}\text{TF}\bullet\text{DNA} \tag{3.3}$$

where $K$ is relative binding affinity based on DNA sequence S and can be calculated from the concentration of TF and the concentration of bound complex $\text{TF}bullet\text{DNA}$

$$K=\frac{[\text{TF}\bullet\text{DNA}]}{[\text{TF}][\text{DNA}]} \tag{3.4}$$

Note that the equilibrium probability of a site $S$ being bound is

$$\Pr(S \quad \text{bound}) = \frac{[\text{TF} \bullet \text{DNA}]}{[\text{TF} \bullet \text{DNA}] + [\text{DNA}]} = \frac{K[\text{TF}]}{K[\text{TF}] + 1} \tag{3.5}$$

Let Shape($S$) be the score assigned by the Random Forest classifier to binding site $S$. Assume that the score is normalized to be in the range 0 (minimum) to 1 (maximum). We have tested the following two approaches in combining the shape score into sequence to expression models.

Approach 3.1

Assume that $\frac{\text{Shape}(S)}{2}$ is the probability of site $S$ being bound at conditions where $[\text{TF}] = \frac{1}{K(S_{\text{opt}})}$ , where $S_{\text{opt}}$ is the consensus binding site. The relative binding affinity can be represented $S$ as

$$K = K(S_{\text{opt}})e^{-\Delta E(S)} \tag{3.6}$$

where $\Delta E(S)$ is $E(S) - E(S_{\text{opt}})$, with $E(S) \geq E(S_{\text{opt}})$ and $E(S)$ is the binding energy of the TF to binding site. Therefore, the equilibrium probability of a site $S$ being bound is

$$\Pr(S \quad \text{bound}) = \frac{[\text{TF}]K(S_{\text{opt}})e^{-\Delta E(S)}}{\text{TF}]K(S_{\text{opt}})e^{-\Delta E(S)} + 1} \tag{3.7}$$

Note that since $[\text{TF}] = \frac{1}{K(S_{\text{opt}})}$ at the condition assumed above, we have

$$\Pr(S \quad \text{bound}) = \frac{e^{-\Delta E(S)}}{e^{-\Delta E(S)} + 1} \tag{3.8}$$

and therefore

$$\frac{\text{Shape}(S)}{2} = \frac{1}{1 + e^{\Delta E(S)}} \tag{3.9}$$

Note that for $S = S_{\text{opt}}$ we have $\Delta E(S) = 0$. Therefore, $\frac{\text{Shape}(S_{\text{opt}})}{2} = \dfrac{1}{1+1} = \dfrac{1}{2}$, i.e. $\text{Shape}(S_{\text{opt}}) = 1$, as it should be.

In general,

$$\Delta E(S) = \ln(\frac{2 - \text{Shape}(S)}{\text{Shape}(S)}) \tag{3.10}$$

Use the above formula of $\Delta E(S)$ in calculating the statistical weight of a site as

$$q(S) = K(S_{\text{opt}})[\text{TF}]e^{-\Delta E(S)} = K(S_{\text{opt}})[\text{TF}]\frac{\text{Shape}(S)}{2 - \text{Shape}(S)} \tag{3.11}$$

Approach 2.

Following Pujato et al. [131], the relative binding affinity is defined as

$$K = e^{-\frac{A}{K_B T}(1-\text{Shape}(S))} \tag{3.12}$$

where $A$ is a proportionality constant in units of Kcal/mol, $K_B$ is the Boltzman constant in Kcal/(mol • K) and $T$ is the temperature in Kelvin. In Pujato et al., the best results were observed when $A$=4.74 Kcal/mol at 298 K, we therefore treated $A/(K_B T)$ as one parameter $k$ and set $k = 8.0$ as the default starting value when training the sequence to expression model.

The equilibrium probability of a site $S$ being bound becomes

$$\Pr(S \quad \text{bound}) = \frac{[\text{TF}]e^{-k(1-\text{Shape}(S))}}{[\text{TF}]e^{-k(1-\text{Shape}(S))} + 1} \tag{3.13}$$

and the statistical weight of a site is

$$q(S) = [\text{TF}]e^{-k(1-\text{Shape}(S))} \tag{3.14}$$

## 3.4 MODEL TRAINING AND EVALUATION

For a fair comparison, we focused on the same data set used in one of the original PWM-based modeling studies [19], i.e. GEMSTAT, which includes the following: 37 experimentally characterized enhancers, 37 quantitative profiles of gene expression driven by each enhancer, and quantitative concentration profiles of six TFs - bicoid (*BCD*), caudal (*CAD*), giant (*GT*), hunchback (*HB*), knirps (*KNI*), and Kruppel (*KR*). To supplement this data, we added three additional TF concentration profiles: vielfaltig (*VFL*), Dstat, and sloppy-paired (*SLP*), which were obtained from FlyEx database [132]. Similar to He et al. [19], we limited the gene expression modelling to the 20% - 80% region of the A/P axis, resulting in 60 'bins' of gene expression and TF concentration values. PWMs of all TFs were constructed with MEME [54] applied to binding sites obtained via bacterial one hybrid (B1H) experiments, downloaded from the Fly Factor Survey database [68]. To increase the quality of PWMs, we trimmed MEME-predicted PWMs to have nearly the same length as PWMs in Factor Survey database [68], by removing 0 to 3 degenerate (low information content) positions on either ends (Table 3.1). The DNA shape score predictors were train on the same set of binding sites.

In order to fairly compare DNA shape-based models with PWM-based models, we used the

Table 3.1: **Lengths of trimmed PWMs and positions being trimmed from MEME predicted PWMs**.

| TF | MEME PWM length | Positions trimmed | Trimmed PWM length |
|---|---|---|---|
| BCD | 6 | None | 6 |
| CAD | 8 | Last 1 position | 7 |
| VFL | 11 | First 3 positions | 8 |
| DSTAT | 11 | None | 11 |
| GT | 14 | First 3 positions | 11 |
| HB | 10 | First 3 positions | 7 |
| KNI | 13 | Last 2 positions | 11 |
| KR | 10 | Last 1 positions | 9 |
| SLP | 11 | None | 11 |

same GEMSTAT interaction mode (direct) and only considered self-cooperativity of *BCD* and *CAD*. Following He et al. [19], in the PWM-based model, we annotated a site $S$ with an $\mathrm{LLR}(S) \geq 0.4 * \mathrm{LLR}(S_{\mathrm{opt}})$ as a binding site. To yield a similar number of binding sites for the DNA shape-based model, a site with shape score greater than 0.6 was annotated as a binding site.

To measure the goodness of fit between the real and predicted gene expression, we used the scoring function called "weighted pattern generating potential" (wPGP) [24], which essentially rewards the agreement between endogenous and predicted readouts and penalizes the disagreement. The wPGP score ranges between 0 and 1, with higher values indicating better fits. By choosing wPGP as the measurement, we were able to avoid the following issues common to widely used methods such as correlation or root mean square error: biases from overly narrow or overly board predicted expression and insensitivity to shift and scaling of the expression profiles as previously reported in Kazemian et al. [40] (see Section 1.2 for details).

## 3.5 DNA SHAPE-BASED MODEL PREDICTS GENE EXPRESSION AT LEAST AS WELL AS PWM-BASED MODEL

On the whole, the DNA shape-based model performed as well as and arguably better than the PWM-based model, as shown in Table 3.2 and Figure 3.2. The DNA shape-based model achieved a wPGP score of 0.784, averaged over the 37 enhancers in the training data set while PWM-based model averaged at 0.755. This difference, being taken over averages of scores, is significant based on our prior experience [64] and direct statistical testing (Wilcoxon signed-rank test p-value of 0.003). For 14 out of 37 enhancers we noted better fits using the shape-

Table 3.2: **10-fold cross-validation assessment of various models.** For each model, shown are the number of free parameters used ("#Pars"), the average wPGP scores from parameter optimization over all 37 enhancers ("Avg. wPGP (Training)"), and the wPGP scores from cross-validation ("Avg. wPGP (CV)"), averaged over five repeats of cross validation with different (random) definitions of the ten folds. Standard deviations over the five repeats are also shown.

| **Model** | **#Pars** | **Avg. wPGP (Training)** | **Avg. wPGP (CV)** |
|---|---|---|---|
| Shape-based model | 22 | 0.784 | 0.727 ± 0.020 |
| PWM-based model | 21 | 0.755 | 0.677 ± 0.004 |
| PWM-based model, Perturbed LLR scores | 21 | 0.643 | 0.603 ± 0.021 |

based model (wPGP score improved by greater than 0.05), whereas for 8 out of 37 enhancers the shape-based model produced worse fits (Table 3.3). These results provided clear evidence that DNA shape readout at putative binding sites can lead to accurate quantitative modeling of gene expression, and suggest that it yields arguably better fits than nucleotide readout.

To better appreciate the differences between fits of enhancer readouts from the two models, we plotted the predicted expression profiles of the two models along with real expression patterns for a selection of six enhancers (Figure 3.3). It was evident that the DNA shape-based model improved the expression prediction by predicting more accurately defined boundaries of spatial expression domains. For example, for the enhancers 'eve_stripe5' as well as 'run_stripe1', the shape-based model accurately predicts the posterior and anterior boundary respectively. Qualitative refinements were observed on other enhancers. For instance, the shape-based model reduced a spurious anterior domain predicted by the PWM-based model for the enhancer 'eve_37ext_ru', correctly modeled the anterior peak in 'ftz_+3' which the PWM-based model failed to predict, and correctly suppressed an ectopic posterior domain of 'slp_(-3)' expression predicted by the PWM-based model. More complete comparisons of gene expression profiles where the DNA shape-based model produced better or worse fits than the PWM-based model are shown in Figure 3.4.

The results above have indicated, both quantitatively and qualitatively, that a DNA shape-based characterization of binding sites performed at least as well as the more conventional PWM-based model in sequence-to-expression modeling. It should be noted that while both models used the same set of parameters, the DNA shape-based model had one additional parameter ('$k$', see Section 3.2) to map the site score computed by the Random Forest-based classifier to a pseudo-energy term in GEMSTAT. (The PWM-based model has 21 parameters while the shape-based model has 22 parameters.) A widely accepted method to compare models with different complexities is to assess goodness of fit under cross validation. We

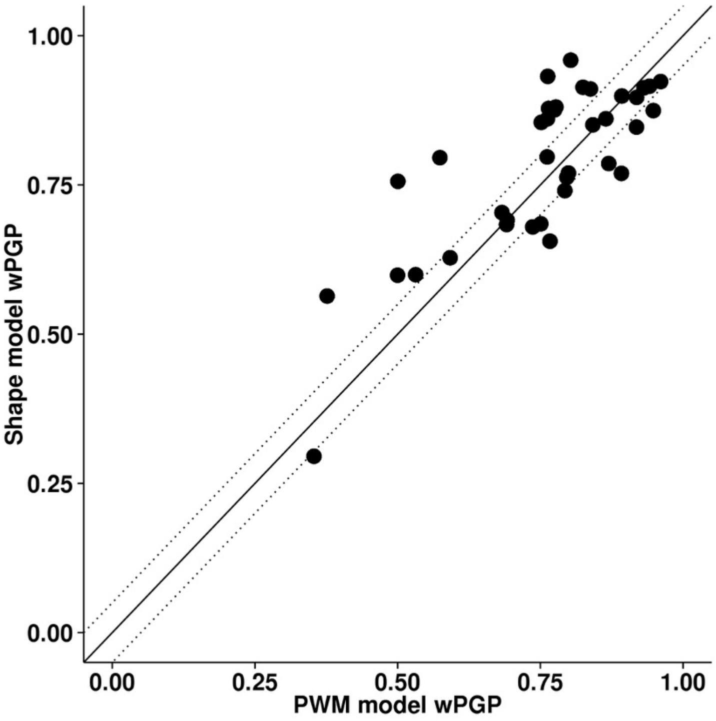

Figure 3.2: **Performance of DNA shape-based model compared to PWM-based model on 37 *Drosophila* enhancers.** The goodness of fit between predicted and real expression for each enhancer was assessed by wPGP scores. Dotted lines delineate regions where the difference in wPGP score between the two models is less than 0.05.

Table 3.3: **Evaluations of expression predictions from DNA shape-based model and PWM-based model.** The "goodness of fit" between predicted and real expression for each enhancer was assessed by wPGP score. The wPGP scores from DNA shape-based model and PWM-based model over all 37 enhancers are shown, and changes of wPGP scores greater than 0.05 are identified.

| **Enhancer** | **DNA shape-based model** | **PWM-based model** | **Change >0.05** |
|---|---|---|---|
| Kr_CD1_ru | 0.76 | 0.50 | + |
| ftz_+3 | 0.80 | 0.57 | + |
| hb_centr_&_post | 0.56 | 0.38 | + |
| run_stripe1 | 0.93 | 0.76 | + |
| eve_37ext_ru | 0.96 | 0.80 | + |
| slp2_(-3) | 0.88 | 0.76 | + |
| kni_83_ru | 0.88 | 0.78 | + |
| eve_1_ru | 0.85 | 0.75 | + |
| nub_(-2) | 0.88 | 0.77 | + |
| eve_stripe2 | 0.60 | 0.50 | + |
| kni_(+1) | 0.86 | 0.76 | + |
| eve_stripe5 | 0.91 | 0.82 | + |
| gt_(-10) | 0.91 | 0.84 | + |
| odd_(-5) | 0.60 | 0.53 | + |
| h_15_ru | 0.63 | 0.59 | |
| hb_anterior_actv | 0.80 | 0.76 | |
| h_stripe34_rev | 0.70 | 0.68 | |
| prd_+4 | 0.85 | 0.84 | |
| run_stripe3 | 0.90 | 0.89 | |
| odd_(-3) | 0.69 | 0.69 | |
| eve_stripe4_6 | 0.86 | 0.86 | |
| knrl_(+8) | 0.68 | 0.69 | |
| btd_head | 0.91 | 0.93 | |
| run_-17 | 0.90 | 0.92 | |
| run_-9 | 0.92 | 0.94 | |
| gt_(-1) | 0.77 | 0.80 | |
| pdm2_(+1) | 0.76 | 0.80 | |
| h_6_ru | 0.92 | 0.96 | |
| run_stripe5 | 0.74 | 0.79 | |
| Kr_CD2_ru | 0.68 | 0.74 | - |
| Kr_AD2_ru | 0.30 | 0.35 | - |
| cnc_(+5) | 0.68 | 0.75 | - |
| oc_otd_early | 0.85 | 0.92 | - |
| oc_(+7) | 0.87 | 0.95 | - |
| kni_(-5) | 0.79 | 0.87 | - |
| D_(+4) | 0.66 | 0.77 | - |
| gt_(-3) | 0.77 | 0.89 | - |

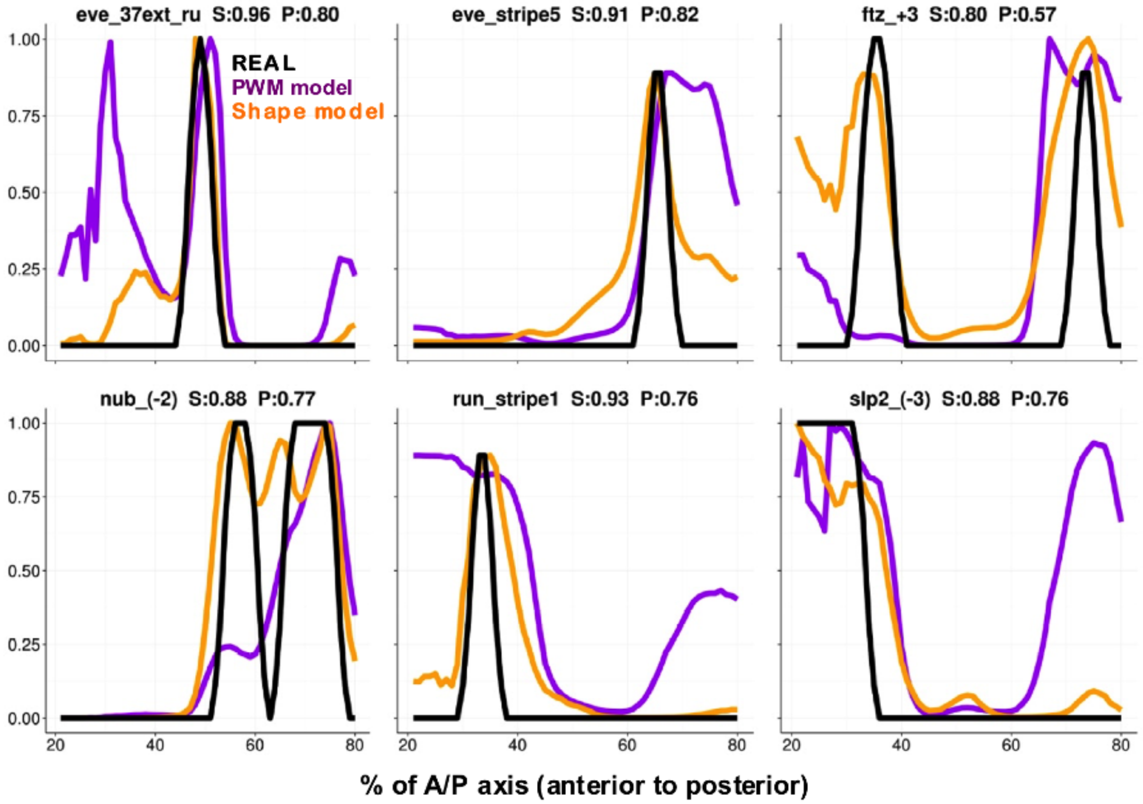

Figure 3.3: **Fits between model and data.** Predicted expression profiles of DNA shape-based model (*orange lines*) and PWM-based model (*purple lines*) are compared to experimentally determined expression profiles (*black lines*), for six selected *Drosophila* enhancers. Each expression profile is on a relative scale of 0 to 1 (y-axis), and shown for the regions between 20% and 80% of the A/P axis of the embryo. Title in each panel is in the format of “enhancer name, wPGP by DNA shape-based model (‘S’), wPGP by PWM-based model (‘P’).” See more enhancers fits in Figure 3.4.

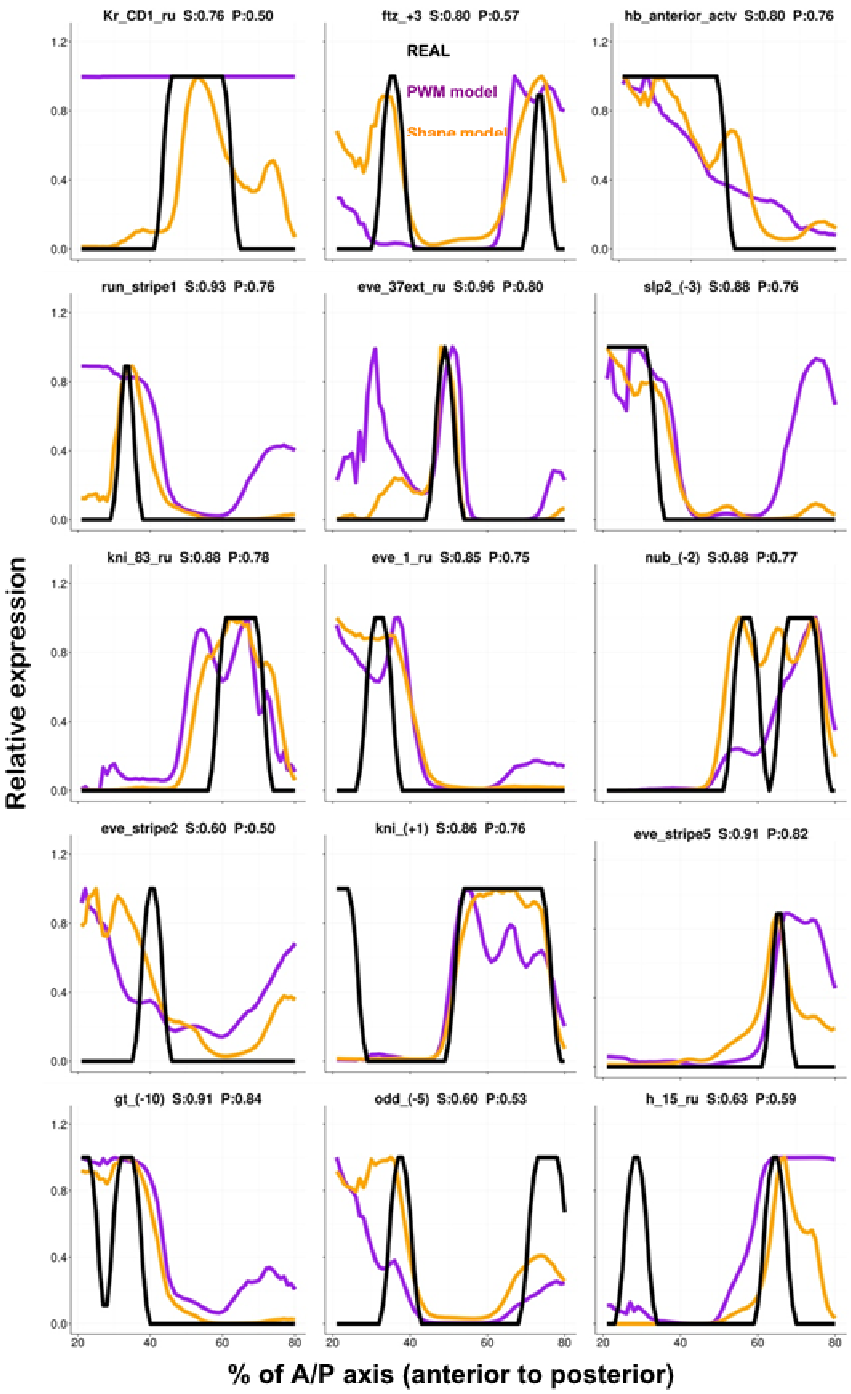

Figure 3.4: **Fits between model and data.** Predicted expression profiles of DNA shape-based model (*orange lines*) and PWM-based model (*purple lines*) are compared to experimentally determined expression profiles (*black lines*), for all 37 *Drosophila* enhancers in this study. Each expression profile is on a relative scale of 0 to 1 (y-axis), and shown for the regions between 20% and 80% of the A/P axis of the embryo. Title in each panel is in the format of "enhancer name, wPGP by DNA shape-based model ('S'), wPGP by PWM-based model ('P')." The order of enhancers is the same as in Table 3.3.

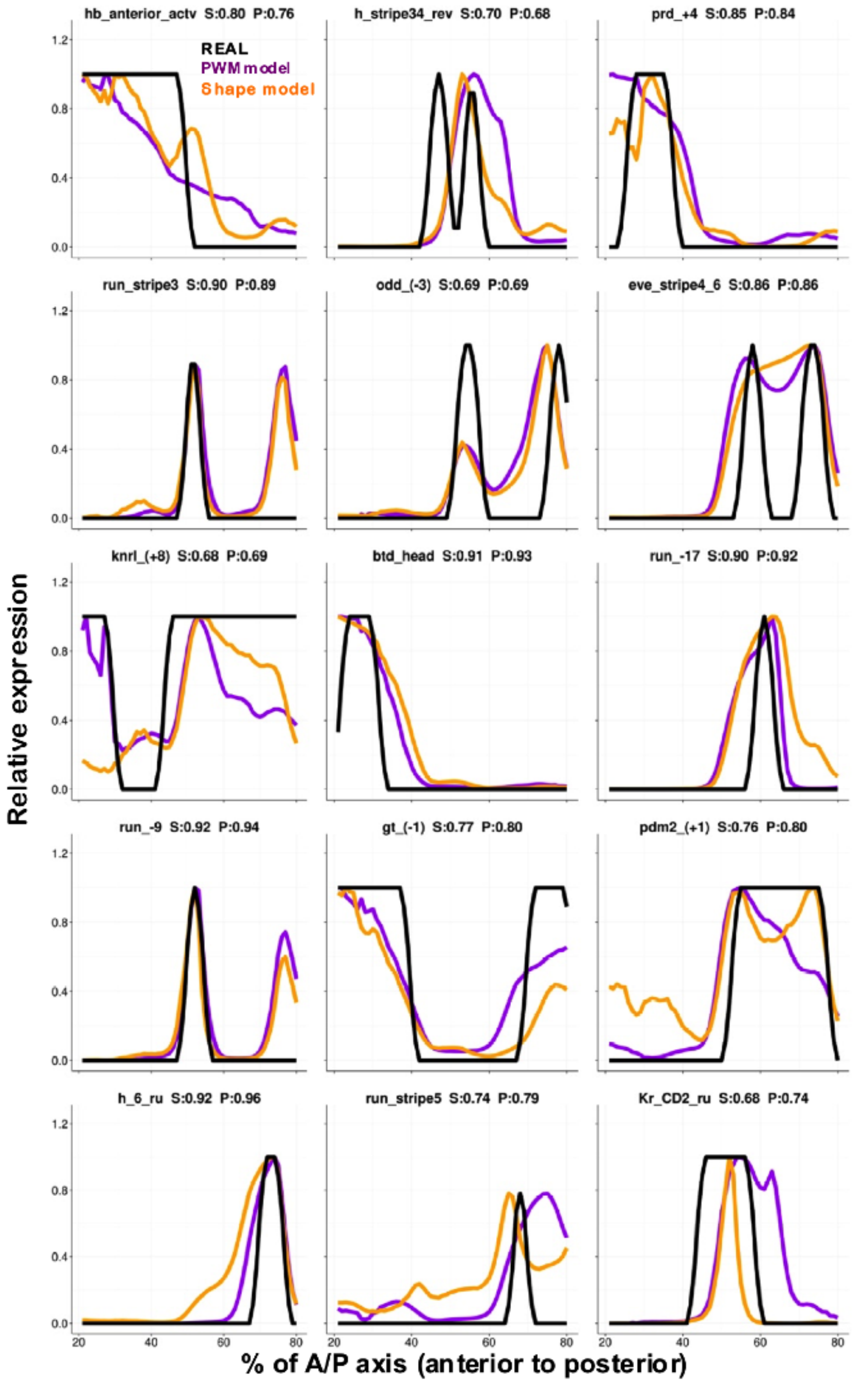

Figure 3.4 (cont'd): **Fits between model and data.**

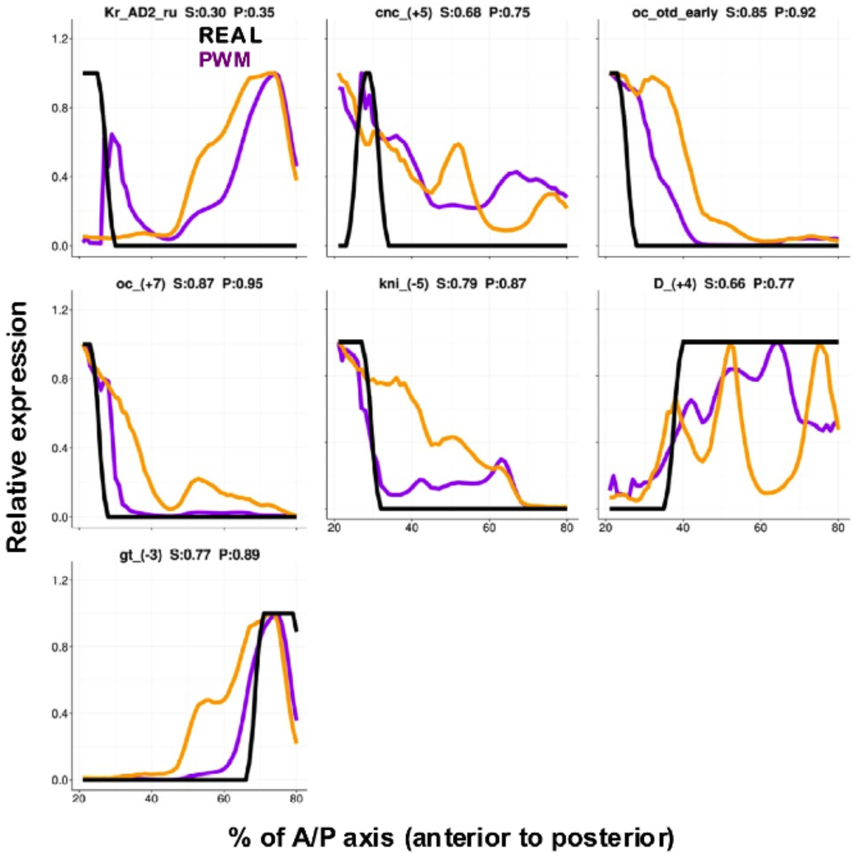

Figure 3.4 (cont'd): **Fits between model and data.**

therefore performed 10-fold cross-validation on all 37 enhancers for each model. Since the partition of the data set into training and test sets was decided randomly, we repeated the exercise ten times with either model. The DNA shape-based model reported a wPGP score (averaged over all 37 enhancers, and over the ten repeats) of 0.727 with standard deviation 0.020 and PWM-based model led to an average wPGP of 0.677 with standard deviation 0.004 (Table 3.4). Thus, we confirmed that the improved fits from the DNA shape-based model are not due to its additional free parameter.

We considered the possibility that the improved fits with the shape-based model are primarily due to a single TF (or a minority of TFs) for which the PWM is not an appropriate model of binding specificity, while for other TFs the PWM model is more suited for use in expression modeling. We tested this possibility and found it to be false. In particular, we repeated the model fitting exercise with the shape-based scoring of binding sites for every TF except one, for which PWM-based scoring was used. We compared the goodness of fit (average wPGP across enhancers) of such hybrid models with that of the purely shape-based model, and noted that for all TFs except *CAD*, the fits deteriorated upon substituting

Table 3.4: **10-fold cross-validation assessment of various models.** For each model, shown are the number of free parameters used ("#Pars"), the average wPGP scores from parameter optimization over all 37 enhancers ("Avg. wPGP (Training)"), and the wPGP scores from cross-validation ("Avg. wPGP (CV)"), averaged over five repeats of cross validation with different (random) definitions of the ten folds. Standard deviations over the five repeats are also shown.

| **Model** | **#Pars** | **Avg. wPGP (Training)** | **Avg. wPGP (CV)** |
|---|---|---|---|
| Shape-based model | 22 | 0.784 | 0.727 ± 0.020 |
| PWM-based model | 21 | 0.755 | 0.677 ± 0.004 |
| PWM-based model, Perturbed LLR scores | 21 | 0.643 | 0.603 ± 0.021 |

shape-based scores with PWM-based LLR scores for that TF's sites. (Figure 3.5A) (The goodness of fit was almost unchanged upon switching from the shape model to the PWM model for *CAD*.) This suggests that for every TF in this analysis the shape-based score is as good or better than the PWM-based score for the purpose of expression modeling.

We wondered if the difference between shape-based and sequence-based models arises from the difference in how the binding site scoring method was trained – as a PWM trained on sample sites versus a Random Forest classifier trained on samples of sites and non-sites. To make the models more similar in this aspect, we trained a Random Forest classifier on 1-mer sequence features (the so-called '1-hot encoding' [133]), using the same training data sets as for shape model. We then incorporated scores predicted by Random Forest into GEMSTAT in the same way as for the DNA shape model. The average wPGP score of this alternative sequence-based model was 0.756 (Table 3.5 and Table 3.6), nearly the same as the PWM-based model. We repeated 10-fold cross-validation ten times, and obtained an average wPGP score of 0.673 with standard deviation 0.014, again very similar to that of the PWM-based model, suggesting that the gap between shape-based and sequence-based models is not merely due to a difference in how underlying binding site scoring methods were trained.

## 3.6 DNA SHAPE-BASED MODEL OUTPERFORM PWM-BASED MODEL UNDER THE SAME SEQUENCE LENGTH

In our direct comparisons between the shape-based and PWM-based models, all other aspects of modeling were identical, including the set of putative sites considered by either model. However, one point of difference was that the site length used to compute shape readouts was in some cases different from the site length used to score for PWM matches.

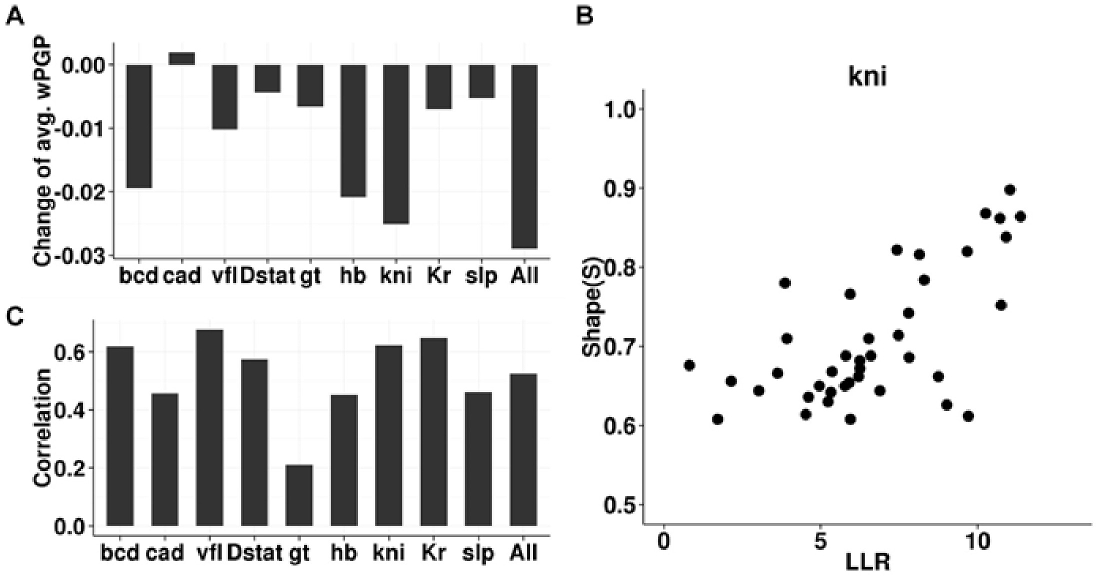

Figure 3.5: **DNA shape is characterized differently from PWM** (A) Change of goodness of fit (avg. wPGP) of DNA shape-based model predictions when binding sites of a specific TF were forced to use LLR rather than shape scores. (B) Visualization of kni binding sites correlation between shape scores and LLR. (C) Pearson correlations of binding sites for each of nine TF in this study and all TFs.

Table 3.5: **Evaluations of various models in this study.** For each model, shown are the number of free parameters used ("#Pars"), the average wPGP scores from parameter optimization over all 37 enhancers ("Avg. wPGP (Training)"), and the wPGP scores from cross-validation ("Avg. wPGP (CV)"), averaged over ten repeats of cross validation with different (random) definitions of the ten folds. Standard deviations over the ten repeats are also shown.

| **Model** | **#Pars** | **Avg. wPGP (Training)** | **Avg. wPGP (CV)** |
|---|---|---|---|
| Sequence Model | | | |
| PWM-based | 21 | 0.755 | 0.677 ± 0.004 |
| RF-1-mer | 22 | 0.756 | 0.673 ± 0.014 |
| RF-1-mer+2-mer | 22 | 0.770 | 0.696 ± 0.012 |
| RF-1-mer+2-mer+3-mer | 22 | 0.765 | 0.705 ± 0.017 |
| Shape Model | | | |
| Shape-based | 22 | 0.784 | 0.727 ± 0.020 |
| Sequence+Shape Model | | | |
| Integrative PWM | 22 | 0.752 | 0.676 ± 0.011 |
| Integrative Shape | 22 | 0.776 | 0.727 ± 0.005 |
| RF-Shape+1-mer | 22 | 0.777 | 0.724 ± 0.013 |
| RF-Shape+1-mer+2-mer | 22 | 0.762 | 0.696 ± 0.012 |
| RF-Shape+1-mer+2-mer+3-mer | 22 | 0.767 | 0.708 ± 0.016 |

Table 3.6: **Comparisons between models in this study.** For each pair of models, shown is the p-value of Wilcoxon signed-rank test over ten pairs of average wPGP scores from ten repeats of 10-fold cross-validation. S: shape; m: mer; Inte: integrative.

| | 1m | 1+2m | 1+2+3m | Shape | Inte PWM | Inte Shape | S+ 1m | S+ 1+2m | S+ 1+2+3m |
|---|---|---|---|---|---|---|---|---|---|
| PWM | 0.421 | 0.003 | 0.003 | 0.003 | 0.288 | 0.005 | 0.003 | 0.003 | 0.003 |
| 1m | | 0.003 | 0.006 | 0.003 | 0.341 | 0.003 | 0.003 | 0.003 | 0.003 |
| 1+2m | | | 0.121 | 0.006 | 0.003 | 0.005 | 0.018 | 0.003 | 0.192 |
| 1+2+3m | | | | 0.008 | 0.003 | 0.003 | 0.006 | 0.018 | 0.192 |
| Shape | | | | | | 0.323 | 0.084 | 0.003 | 0.005 |
| Inte PWM | | | | | | 0.003 | 0.003 | 0.003 | 0.003 |
| Inte Shape | | | | | | | 0.192 | 0.003 | 0.006 |
| S+ 1m | | | | | | | | 0.003 | 0.006 |
| S+ 1+2m | | | | | | | | | 0.003 |

This was motivated by our observation that the PWM-based model yielded better fits when using shorter (‘trimmed’) PWMs than those constructed directly from the available sample of binding sites. Trimming out less informative positions was ideally acceptable because we resisted to deteriorate the performance of PWM-based model so that the DNA shape-based model would look better. However, one may claim that DNA shape obtained more information from positions where its PWM counterpart ignored and therefore fit enhancers more accurately. Here, we applied a thorough analysis on the original untrimmed PWMs whose lengths were identical to the DNA shape-based putative binding sites.

Our intension was first to see how the length of PWMs would affect the modeling. Generally speaking, trimmed PWMs were more suitable for modeling. Figure 3.6A plots the wPGP scores for each enhancer in models using either trimmed or original long PWMs. The average wPGP score was 0.755 for trimmed PWMs model, outperforming the regular PWMs model whose score was 0.734. Detailed fitting of each enhancer can be seen in Table 3.7. At this point, we were confident that trimmed PWMs played a better role in the PWM-based model.

On the other direction, we tried to answer the question: given the same binding site length as DNA shape did, would the PWM-based model be able to gather more information and thus make better predictions? Figure 3.6B and Table 3.7 reports the comparison of the DNA shape-based model and untrimmed PWM-based model over 37 enhancers. In the majority

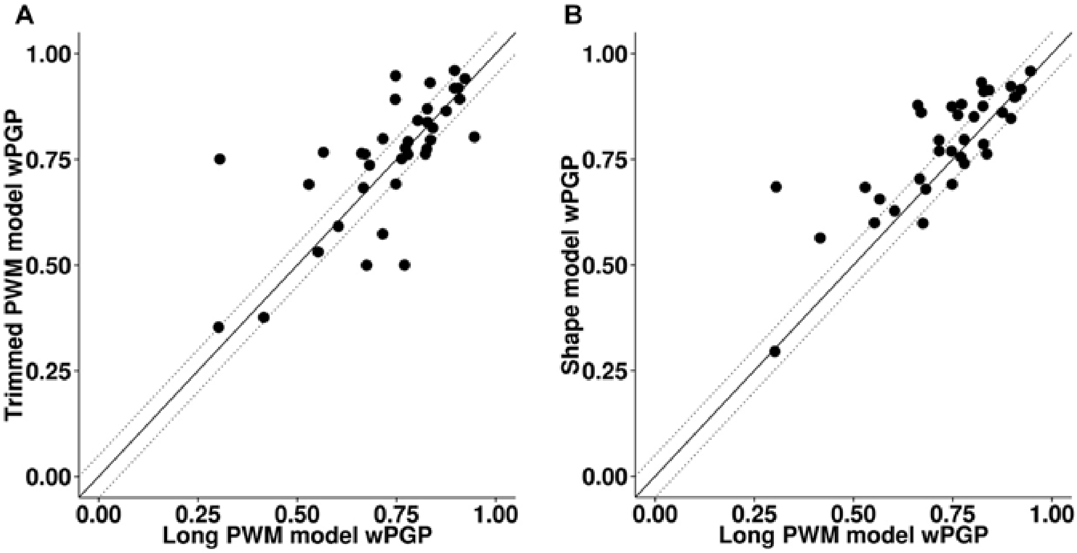

Figure 3.6: **Performance of long PWM models** compared to (A) trimmed PWM model and (B) DNA shape model on 37 *Drosophila* enhancers assessed by wPGP scores.

of cases, DNA shape-based model had considerably better fits than untrimmed PWM-based model. There were 15 out of 37 enhancers having measurable improvements in DNA shape-based model while only three declined. The average wPGP score was 0.784 for the DNA shape-based model compared to 0.734 for untrimmed PWMs model.

We systematically examined the effect of motif length on our claims and confirmed that the comparisons and claims reported above involve a fair treatment of the PWM model. That is, the gap between shape-based and PWM-based models is even greater when the same site lengths are used for both models and PWMs are not 'trimmed'.

## 3.7 DNA SHAPE MODELS CAPTURE INFORMATION DIFFERENT FROM PWM

In light of our aforementioned conclusion that shape-based models perform at least as well as PWM-based models in predicting enhancer readouts, we next asked if the PWM-based score and DNA shape-based score are simply two ways to quantify exactly the same information, differing only procedurally. They are closely related scores, since both are computed from the primary sequence of a binding site. At the same time, each has its own intuitive biophysical explanation: the PWM-based score is related directly to the binding energy of a site [52, 57] assuming positional independence, while the shape-based score reflects how similar a putative site's local DNA shape is to that of the training set of binding sites.

Table 3.7: **Evaluations of expression predictions from long PWM, trimmed PWM, and DNA shape models.** The "goodness of fit" between predicted and real expression for each enhancer was assessed by wPGP score. The wPGP scores from PWM-based models and DNA shape-based model over all 37 enhancers are shown.

| **Enhancer** | **Long PWM model** | **Trimmed PWM model** | **DNA shape model** |
|---|---|---|---|
| btd_head | 0.83 | 0.93 | 0.91 |
| cnc_(+5) | 0.31 | 0.75 | 0.68 |
| D_(+4) | 0.57 | 0.77 | 0.66 |
| eve_1_ru | 0.76 | 0.75 | 0.85 |
| eve_37ext_ru | 0.95 | 0.80 | 0.96 |
| eve_stripe2 | 0.67 | 0.50 | 0.60 |
| eve_stripe4_6 | 0.88 | 0.86 | 0.86 |
| eve_stripe5 | 0.84 | 0.82 | 0.91 |
| ftz_+3 | 0.72 | 0.57 | 0.80 |
| gt_(-10) | 0.83 | 0.84 | 0.91 |
| gt_(-1) | 0.72 | 0.80 | 0.77 |
| gt_(-3) | 0.75 | 0.89 | 0.77 |
| h_15_ru | 0.60 | 0.59 | 0.63 |
| h_6_ru | 0.90 | 0.96 | 0.92 |
| hb_anterior_actv | 0.78 | 0.76 | 0.80 |
| hb_centr_&_post | 0.42 | 0.38 | 0.56 |
| h_stripe34_rev | 0.67 | 0.68 | 0.70 |
| kni_(+1) | 0.67 | 0.76 | 0.86 |
| kni_(-5) | 0.83 | 0.87 | 0.79 |
| kni_83_ru | 0.77 | 0.78 | 0.88 |
| knrl_(+8) | 0.53 | 0.69 | 0.68 |
| Kr_AD2_ru | 0.30 | 0.35 | 0.30 |
| Kr_CD1_ru | 0.77 | 0.50 | 0.76 |
| Kr_CD2_ru | 0.68 | 0.74 | 0.68 |
| nub_(-2) | 0.83 | 0.77 | 0.88 |
| oc_(+7) | 0.75 | 0.95 | 0.87 |
| oc_otd_early | 0.90 | 0.92 | 0.85 |
| odd_(-3) | 0.75 | 0.69 | 0.69 |
| odd_(-5) | 0.55 | 0.53 | 0.60 |
| pdm2_(+1) | 0.84 | 0.80 | 0.76 |
| prd_+4 | 0.80 | 0.84 | 0.85 |
| run_-17 | 0.91 | 0.92 | 0.90 |
| run_-9 | 0.92 | 0.94 | 0.92 |
| run_stripe1 | 0.82 | 0.76 | 0.93 |
| run_stripe3 | 0.91 | 0.89 | 0.90 |
| run_stripe5 | 0.78 | 0.79 | 0.74 |
| slp2_(-3) | 0.66 | 0.76 | 0.88 |

To objectively characterize the relationship between the two scores, we examined their mutual correlation over all putative binding sites for each TF. Figure 3.5B shows the scatter plot of the two scores across all binding sites for the TF *KNI*, where we noted Pearson correlation of 0.623. Figure 3.5C shows Pearson correlation for each of the nine TFs; these correlations are typically around 0.5, ranging between 0.211 (*GT*) to 0.677 (*VFL*), with the correlation over putative binding sites of all TFs being 0.525 (Figure 3.5C, 'All'). We interpreted these observations to mean that the shape-based score, while being closely related to the PWM-based score of a site, is not redundant with the latter and contains additional information not captured by the direct sequence readout. Our tests above (Figure 3.5A) further indicated that the additional information captured by the shape score is useful for predicting gene expression profiles as well as and arguably better than with PWM scores. However, we considered the possibility that this improvement (average wPGP of 0.784 for the shape-based model compared to 0.755 for the PWM-based model) is an artifact of our procedure. Specifically, it was possible that our modeling is fundamentally incapable of discerning an accurate TF-DNA binding model from a noisy version thereof, either due to noise in the data or over-parameterization, or for an unknown reason. To test this possibility, we repeated the PWM-based model-fitting exercise after artificially perturbing the LLR scores of binding sites. For each binding site in each enhancer, an artificial LLR score was assigned at random, sampling from a normal distribution with mean equal to the site's true LLR score and a fixed variance. This added 'noise' was tuned to be such that the Pearson correlation between true and perturbed LLR scores was  0.5, which we noted above to be the overall correlation between shape scores and LLR scores (Figure 3.5C, 'All'). As shown in (Table 3.4), this PWM-based model performed substantially worse than with true LLR scores: the average wPGP score over 37 enhancers dramatically decreased to 0.643 (compared to 0.755) and the 10-fold cross-validation wPGP score (averaged over ten repeats) dropped from 0.677 to 0.603. This exercise strongly suggested to us that the better fits predicted by the DNA shape-based model compared to the PWM-based model cannot be reproduced merely by a good approximation to LLR scores of sites, and that the shape scores carry information that is complementary to LLR scores and useful for sequence-to-expression modeling, ruling out the concern raised above.

Previous work has found sequence models that consider nucleotide inter-dependencies to better fit binding affinity data than the PWM model [59, 61]. We therefore tested if a Random Forest trained to classify sites based on their k-mer profile can lead to improved expression predictions. A '1-mer+2-mer' sequence-based model achieved a wPGP score of 0.770 on average, and a '1-mer+2-mer+3-mer' model yielded an average wPGP of 0.765. (Table 3.5 and Table 3.6; the wPGP score of each enhancer, under either model, can be found in Table 3.8.) Thus, the fits achieved with higher order k-mer models were better than those from a 1-hot encoding or the PWM model, but not better than fits of the shape-based model. This is consistent with the view that DNA shape features provide an alternative and more compact representation of positional interdependencies in binding sites [134].

## 3.8 COMBINING SHAPE AND SEQUENCE READOUT INTO A SINGLE MODEL DOES NOT IMPROVE FITS

The literature suggests that models integrating DNA shape with PWM-based sequence readout can improve prediction of TF-DNA binding over models that use either representation independently [130, 135, 134]. However, sequence-to-expression modeling requires not just the prediction of TF binding strengths, but also quantifying how different configurations of DNA-bound TFs relate to gene expression levels. Given this, it is not entirely clear whether integrating DNA shape and sequence would significantly improve expression modeling. We tested this hypothesis by first comparing a model that integrates DNA shape scores into PWM-based models

(referred to henceforth as 'integrative PWM-based' models) with PWM-based models. To incorporate DNA shape information into the PWM-based model, we replaced the term for binding energy of a site in GEMSTAT to be $\Delta E(S) = \exp[\mathrm{LLR}(S) - \mathrm{LLR}(S_{\mathrm{opt}}) - k(1 - \mathrm{Shape}(S))]$ where $\mathrm{LLR}(S)$ is the log likelihood ratio score of site $S$ under the PWM model, Shape($S$ is the score of site $S$ computed by a Random Forest classifier using the site's shape readout, and $k$ is a free parameter.

As shown in Figure 3.7, for most enhancers the integrative PWM-based model fits expression data nearly as well as the PWM-based model. The wPGP scores are nearly identical with the average over all 37 enhancers being 0.752 and 0.755 respectively. The integrative PWM-based model outperforms the PWM-based model (a wPGP score difference of 0.05 or more) for six of the enhancers, while the PWM-based model outperforms the integrative PWM-based model for five of the enhancers. (The wPGP scores of each enhancer, under either model, can be found in Table 3.9.) Since the integrative model did not perform consistently better than the sequence-based model, we did not explore other formulas for incorporating DNA shape scores into PWM-based models.

As integrating DNA shape information into PWM-based models did not significantly improve average wPGP scores over PWM-based models, we examined the utility of the converse methodology that adds sequence readout to a DNA shape-based model. In order to accomplish this, we added an additional feature to the Random Forest underlying the shape-based model: the LLR score of the binding site according to the TF's PWM. That is,

Table 3.8: **Evaluations of expression predictions from higher order k-mer models.** The “goodness of fit” between predicted and real expression for each enhancer was assessed by wPGP score. The wPGP scores from integrative PWM-based model and integrative DNA shape-based model over all 37 enhancers are shown.

| **Enhancer** | **RF-1-mer** | **RF-1-mer+ 2-mer** | **RF-1-mer+ 2-mer+3-mer** |
|---|---|---|---|
| btd_head | 0.89 | 0.85 | 0.87 |
| cnc_(+5) | 0.37 | 0.66 | 0.64 |
| D_(+4) | 0.74 | 0.66 | 0.71 |
| eve_1_ru | 0.77 | 0.78 | 0.86 |
| eve_37ext_ru | 0.83 | 0.90 | 0.94 |
| eve_stripe2 | 0.64 | 0.67 | 0.73 |
| eve_stripe4_6 | 0.86 | 0.83 | 0.80 |
| eve_stripe5 | 0.85 | 0.91 | 0.82 |
| ftz_+3 | 0.76 | 0.78 | 0.69 |
| gt_(-10) | 0.80 | 0.85 | 0.82 |
| gt_(-1) | 0.75 | 0.74 | 0.66 |
| gt_(-3) | 0.65 | 0.79 | 0.76 |
| h_15_ru | 0.67 | 0.72 | 0.65 |
| h_6_ru | 0.93 | 0.94 | 0.85 |
| hb_anterior_actv | 0.65 | 0.72 | 0.75 |
| hb_centr_&_post | 0.43 | 0.33 | 0.40 |
| h_stripe34_rev | 0.66 | 0.69 | 0.65 |
| kni_(+1) | 0.78 | 0.80 | 0.62 |
| kni_(-5) | 0.84 | 0.85 | 0.95 |
| kni_83_ru | 0.72 | 0.78 | 0.81 |
| knrl_(+8) | 0.65 | 0.57 | 0.59 |
| Kr_AD2_ru | 0.34 | 0.34 | 0.35 |
| Kr_CD1_ru | 0.82 | 0.74 | 0.75 |
| Kr_CD2_ru | 0.88 | 0.75 | 0.77 |
| nub_(-2) | 0.81 | 0.83 | 0.87 |
| oc_(+7) | 0.82 | 0.84 | 0.87 |
| oc_otd_early | 0.91 | 0.85 | 0.90 |
| odd_(-3) | 0.61 | 0.74 | 0.80 |
| odd_(-5) | 0.81 | 0.73 | 0.71 |
| pdm2_(+1) | 0.66 | 0.68 | 0.77 |
| prd_+4 | 0.85 | 0.87 | 0.87 |
| run_-17 | 0.93 | 0.94 | 0.92 |
| run_-9 | 0.91 | 0.94 | 0.88 |
| run_stripe1 | 0.86 | 0.83 | 0.83 |
| run_stripe3 | 0.87 | 0.88 | 0.91 |
| run_stripe5 | 0.82 | 0.89 | 0.73 |
| slp2_(-3) | 0.84 | 0.85 | 0.83 |

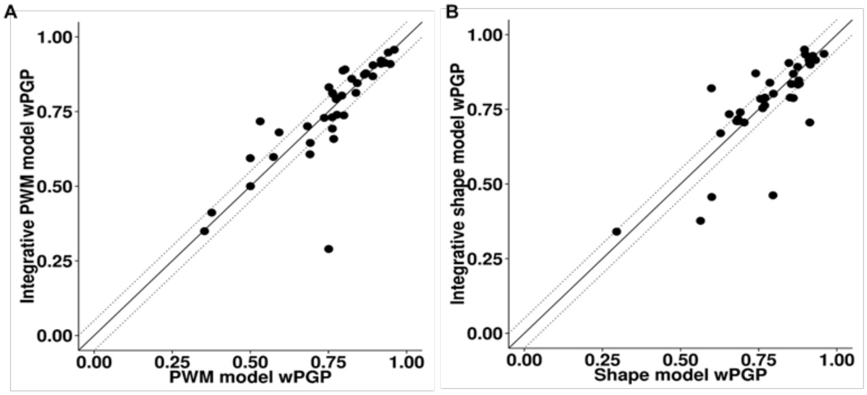

Figure 3.7: **Performance of integrative models** compared to (A) PWM-based model and (B) DNA shape-based model on 37 *Drosophila* enhancers assessed by wPGP scores. Dotted lines delineate regions where the difference in wPGP between the two models is greater than 0.05.

the binding energy term of a site S in GEMSTAT was computed as $\Delta E(S) = \exp[-k(1 - \text{Shape}(S))]$, where Shape($S$) is now computed by a Random Forest classifier trained on predetermined binding sites, using their shape readouts as well as LLR scores. This alternative integrative model (henceforth referred to as a 'integrative shape-based' model) performed as well as the shape-based models (Figure 3.7B), with average wPGP scores over all 37 enhancers being 0.776 and 0.784 respectively. Either model outperformed the other on six of the enhancers (Table 3.9).

In recognition of the fact that there are alternative ways to encode the sequence, we repeated the above test by directly using k-mers of putative sites as features (in addition to shape features) of the Random Forest classifier, and using the resulting score in computing $\Delta E(S)$ as in the previous paragraph. We evaluated three variants of integrative shape+k-mer models, increasing the complexity of models one by one. As listed in Table 3.9, the integrative 'shape+1-mer', 'shape+1-mer+2-mer', and 'shape+1-mer+2-mer+3-mer' models achieved average wPGP scores of 0.777, 0.767, and 0.762 respectively (Table 3.5 and Table 3.6). In short, this section shows that the shape-based model is not improved upon by incorporating sequence-readout into it, nor is the PWM-based model improved upon by including shape readout.

## 3.9 DISCUSSION

Sequence-to-expression models have been effectively used to understand the precise relationship between regulatory sequence and gene expression patterns [18, 19, 20, 22, 44].

Table 3.9: **Evaluations of expression predictions from integrative models.** The "goodness of fit" between predicted and real expression for each enhancer was assessed by wPGP score. The wPGP scores from integrative PWM-based model and integrative DNA shape-based model over all 37 enhancers are shown.

| **Enhancer** | **Integrative PWM-based** | **Integrative shape-based** |
|---|---|---|
| btd_head | 0.91 | 0.90 |
| cnc_(+5) | 0.29 | 0.71 |
| D_(+4) | 0.66 | 0.73 |
| eve_1_ru | 0.83 | 0.83 |
| eve_37ext_ru | 0.89 | 0.94 |
| eve_stripe2 | 0.59 | 0.82 |
| eve_stripe4_6 | 0.87 | 0.87 |
| eve_stripe5 | 0.86 | 0.71 |
| ftz_+3 | 0.60 | 0.46 |
| gt_(-10) | 0.81 | 0.92 |
| gt_(-1) | 0.74 | 0.79 |
| gt_(-3) | 0.87 | 0.76 |
| h_15_ru | 0.68 | 0.67 |
| h_6_ru | 0.96 | 0.93 |
| hb_anterior_actv | 0.73 | 0.80 |
| hb_centr_&_post | 0.41 | 0.38 |
| h_stripe34_rev | 0.70 | 0.71 |
| kni_(+1) | 0.69 | 0.79 |
| kni_(-5) | 0.88 | 0.84 |
| kni_83_ru | 0.74 | 0.84 |
| knrl_(+8) | 0.61 | 0.72 |
| Kr_AD2_ru | 0.35 | 0.34 |
| Kr_CD1_ru | 0.50 | 0.79 |
| Kr_CD2_ru | 0.73 | 0.71 |
| nub_(-2) | 0.79 | 0.83 |
| oc_(+7) | 0.91 | 0.89 |
| oc_otd_early | 0.91 | 0.91 |
| odd_(-3) | 0.65 | 0.74 |
| odd_(-5) | 0.72 | 0.46 |
| pdm2_(+1) | 0.89 | 0.75 |
| prd_+4 | 0.85 | 0.79 |
| run_-17 | 0.92 | 0.95 |
| run_-9 | 0.95 | 0.90 |
| run_stripe1 | 0.81 | 0.91 |
| run_stripe3 | 0.91 | 0.93 |
| run_stripe5 | 0.80 | 0.87 |
| slp2_(-3) | 0.81 | 0.85 |

Table 3.9 (cont'd): **Evaluations of expression predictions from integrative models.**

| **Enhancer** | **Integrative shape+ 1-mer** | **Integrative shape+1-mer+ 2-mer** | **Integrative shape+1-mer+ 2-mer+3-mer** |
|---|---|---|---|
| btd_head | 0.85 | 0.80 | 0.83 |
| cnc_(+5) | 0.57 | 0.64 | 0.67 |
| D_(+4) | 0.60 | 0.46 | 0.74 |
| eve_1_ru | 0.84 | 0.79 | 0.82 |
| eve_37ext_ru | 0.95 | 0.96 | 0.93 |
| eve_stripe2 | 0.58 | 0.67 | 0.72 |
| eve_stripe4_6 | 0.85 | 0.82 | 0.84 |
| eve_stripe5 | 0.72 | 0.82 | 0.89 |
| ftz_+3 | 0.78 | 0.74 | 0.64 |
| gt_(-10) | 0.85 | 0.89 | 0.85 |
| gt_(-1) | 0.72 | 0.71 | 0.65 |
| gt_(-3) | 0.80 | 0.76 | 0.75 |
| h_15_ru | 0.69 | 0.69 | 0.69 |
| h_6_ru | 0.95 | 0.90 | 0.89 |
| hb_anterior_actv | 0.80 | 0.72 | 0.80 |
| hb_centr_&_post | 0.47 | 0.55 | 0.37 |
| h_stripe34_rev | 0.69 | 0.75 | 0.66 |
| kni_(+1) | 0.86 | 0.76 | 0.62 |
| kni_(-5) | 0.83 | 0.82 | 0.89 |
| kni_83_ru | 0.86 | 0.79 | 0.80 |
| knrl_(+8) | 0.78 | 0.53 | 0.58 |
| Kr_AD2_ru | 0.31 | 0.66 | 0.34 |
| Kr_CD1_ru | 0.78 | 0.80 | 0.74 |
| Kr_CD2_ru | 0.73 | 0.77 | 0.72 |
| nub_(-2) | 0.84 | 0.77 | 0.85 |
| oc_(+7) | 0.86 | 0.85 | 0.87 |
| oc_otd_early | 0.92 | 0.91 | 0.89 |
| odd_(-3) | 0.77 | 0.68 | 0.77 |
| odd_(-5) | 0.67 | 0.61 | 0.72 |
| pdm2_(+1) | 0.70 | 0.54 | 0.77 |
| prd_+4 | 0.83 | 0.79 | 0.84 |
| run_-17 | 0.87 | 0.92 | 0.94 |
| run_-9 | 0.92 | 0.90 | 0.90 |
| run_stripe1 | 0.90 | 0.80 | 0.84 |
| run_stripe3 | 0.95 | 0.91 | 0.90 |
| run_stripe5 | 0.81 | 0.86 | 0.88 |
| slp2_(-3) | 0.80 | 0.86 | 0.80 |

TF-DNA binding predictions in these quantitative models typically rely on the PWM representation that assumes every nucleotide in TF binding sites contributes additively and independently to the binding energy at thermodynamic equilibrium, an assumption that does not always hold. A mounting body of work on TF-DNA binding specificity has gone beyond the PWM model by considering the nucleotides dependencies [136, 61], flanking sequences of binding sites [137, 138], and DNA structural features [130, 139, 134] and shown highly promising results in TF-DNA recognition. At the same time, it is not well understood if alternative models of DNA binding can improve the prediction of gene expression. Our work aims at filling this gap by incorporating local DNA shape at the binding site to sequence-to-expression models and asking if it performs as well as a PWM-based model. We found the answer to be affirmative: the DNA shape-based model is arguably better than the PWM-based model in predicting expression. To our knowledge, this sequence-to-expression model based on DNA shape features is the first of its kind.

Previous work has demonstrated that DNA shape-based models compare favorably to sequence-based models for the simpler yet challenging task of modeling TF-DNA binding strength, and that integrative 'shape+sequence' models perform considerably better than sequence-only or shape-only models [130, 135, 134]. However, in this study, we did not see further improvement in our integrative models utilizing both shape-readout and sequence-readout, over models using DNA shape only. This may be in part due to limitations of how our integrative models were constructed, or due to lack of comprehensive data for training our shape models, but it is also a possible indication that better prediction of TF-DNA binding may not always lead to better expression prediction.

Our model succeeds in quantifying the impact of DNA shape on prediction of precise spatiotemporal expression patterns, and also indicates an intuitive and simple approach to deal with DNA shape data. Prior work has suggested several approaches to aggregating shape features as well as learning models, including Random Forest [135, 134] and support vector machine (SVM) [139]. Our approach is in good agreement with the prior use of Random Forest as the learning model, and demonstrates the feasibility of simply using first-order local shape features. We also adopted SVM as an alternative learning model but this appeared to have no further improvement, and we did not pursue deeper investigations thereof.

It is also worth noting that we explored two choices of incorporating shape scores into the original GEMSTAT model. We initially treated the shape score (normalized to a scale of 0 to 1) as being directly related to the probability of a site being bound at a specific TF concentration condition (Section 3.3, Approach 1). This preliminary attempt at incorporating the shape score did not show promising results. The approach used in this study considered the binding affinity of a site, relative to that of the optimal site, as an exponentially decaying function of the shape score (Section 3.3, Approach 2). We expect that future work will continue to improve design of the shape score from the underlying features and integration of shape scores into sequence-to-expression models.

The reader may ask if the thermodynamics-based sequence-to-expression model was necessary for our study. In order to study the effects of particular aspects of data on a higher-level prediction task, one has to make several choices: a modeling or prediction framework, semantic features of the model, and the precise way to quantify those features. In an investigation with so many moving parts, it is natural to first attempt to make reasonable choices about some of those parts, and having fixed them, examine the effect of the one remaining moving part. This is the rationale of our approach. We have extensive experience with the thermodynamic modeling framework and the biological system we utilized here, so we chose to ask questions about shape versus sequence readout in this context.

The DNA shape data used in this study was obtained from computational processing of binding site sequences. This raises the concern that DNA shape scores differ only procedurally from LLR scores (derived from the PWM), but are intrinsically the same information. Our tests suggest that this is not the case and show that DNA shape score captures information different from LLR. It is worth noting that shape features at a single nucleotide position are determined by a pentamer sequence centered at the targeted nucleotide. We have limited information about the flanking sequences of the binding site, so that the shape feature values were unavailable at the terminal positions of some of our TF binding sites. Since it has been reported that DNA shape in the flanking regions of binding sites influences binding specificity [138], we believe that the advantage observed here is an underestimate of how well DNA shape-based models can be used in gene expression predictions. We expect that our modeling approach will be more accurate if and when we have more comprehensive TF binding affinity data sets available.

# CHAPTER 4: EVOLUTIONARY CHANGES IN DNA ACCESSIBILITY AND SEQUENCE PREDICT DIVERGENCE OF TRANSCRIPTION FACTOR BINDING AND ENHANCER ACTIVITY

Transcription factor (TF) binding is determined by sequence as well as chromatin accessibility. While the role of accessibility in shaping TF-binding landscapes is well recorded, its role in evolutionary divergence of TF binding, which in turn can alter *cis*-regulatory activities, is not well understood. In this Chapter, we examine the evolution of genome-wide binding landscapes of five major transcription factors (TFs) in the core network of mesoderm specification, between *D. melanogaster* and *D. virilis*, and study its relationship to accessibility and sequence-level changes. Our collaborators (Dr. Eileen Furlong's laboratory) generated chromatin accessibility data from three important stages of embryogenesis in both *D. melanogaster* and *D. virilis*, and recorded conservation and divergence patterns. We then used multi-variable models to correlate accessibility and sequence changes to TF binding divergence. We found that accessibility changes can in some cases, e.g., for the master regulator Twist and for earlier developmental stages, more accurately predict binding change than is possible using TF binding motif changes between orthologous enhancers. Accessibility changes also explain a significant portion of the co-divergence of TF pairs. We noted that accessibility and motif changes offer complementary views of the evolution of TF binding, and developed a combined model that captures the evolutionary data much more accurately than either view alone. Finally, we trained machine learning models to predict enhancer activity from TF binding, and used these functional models to argue that motif and accessibility-based predictors of TF binding change can substitute for experimentally measured binding change, for the purpose of predicting evolutionary changes in enhancer activity.

## 4.1 INTERSPECIES TF CHIP AND ACCESSIBILITY DATA ACROSS FIVE STAGES OF EMBRYONIC DEVELOPMENT

To understand how evolutionary changes of sequence and accessibility affect TF binding and enhancer activities, we focused our study on an extensively studied regulatory network where prior knowledge of essential regulators and functional enhancers can effectively guide us to functional TF-DNA binding events. We analyzed TF occupancy data for five TFs that form the core of a regulatory network essential for mesoderm development in *Drosophila* [140]: Twist (Twi), Myocyte enhancer factor-2 (Mef2), Tinman (Tin), Bagpipe (Bap) and Biniou (Bin) (Figure 4.1A). We obtained genome-wide TF-DNA binding information on these five TFs across five stages of embryonic development (henceforth, 'time points' or 'TP's), in the form of ChIP-chip and ChIP-seq assays in *D. melanogaster* [110] and *D. virilis* [74] respectively. A total of 14 TF-time point pairs (Figure 4.1B), henceforth called 'TF:TP conditions' or simply 'conditions', were represented in the data, originally reported in Khoueiry et al. [74]. ChIP peaks in close proximity across all TF:TP conditions were clustered to define 8,008 putative ChIP enhancers in *D. melanogaster* [110] and 10,532 putative ChIP enhancers in *D. virilis* [74]. A ChIP score was then assigned to each enhancer, for each TF:TP condition, by extracting the mean ChIP signal (using library size normalized bigwig files) over the enhancer boundaries (performed in Galaxy [141] using the "Compute mean/min/max of intervals" tool version 1.0.0). To make ChIP scores comparable across stages and species, we applied the following normalization on ChIP scores for each TF:TP condition: we set $\mu + 3\sigma$ as the maximum ChIP score, where $\mu$ and $\sigma$ are the mean and standard deviation across all putative enhancers, replaced all ChIP scores greater than this maximum with $\mu + 3\sigma$, and finally applied min-max normalization to set all ChIP scores in a range between 0 to 1.

Orthologous enhancer pairs were defined in Khoueiry et al. [74]. Briefly speaking, we translated *D. virilis* enhancer coordinates into *D. melanogaster* coordinates, overlapped the 10,532 *D. virilis* enhancers with the 8,008 *D. melanogaster* enhancers, and finally obtained a set of 2754 orthologous enhancer pairs. This set of orthologous enhancers served as the targets of our computational analyses in this work. Each enhancer (in either species) was assigned a ChIP score for each TF:TP combination, combining ChIP peaks located within the same *cis*-regulatory element.

To supplement these data, we also collected stage-matched DNase-Seq libraries from both *D. virilis* and *D. melanogaster* in three of the five time points, i.e. TP1, TP3, and TP5 (Figure 4.1B). Accessibility data in *D. virilis* and *D. melanogaster* were obtained using DNase-seq from whole embryos at developmental stages 5-7, 10-11, and 13-15, referred to as TP1, TP3, and TP5. The developmental stages of timed collections were determined exactly as described in [74]. Raw paired-end reads were aligned using BWA [142] on Flybase-R1.2 assembly version for *D. virilis* and on Flybase Assembly 5 (dm3) for *D. melanogaster*. Reads were filtered for optical and PCR replicates using samtools [143]. For peak calling, we used MACS2 (–to-large with -g 1.2E8 for *D. melanogaster* and 1.9E-8 for *D. virilis* and -p 1E-3 as requested for the Irreproducibility Discovery Rate analysis, or IDR [144]). We derived peaks using 1% IDR threshold leading to a unique highly confident and consistent peak sets for biological replicates. For visualization and generation of bigwig score files, reads from BAM files were extended to the average length of the genomic fragments for the corresponding time point, merged and scaled to Read Per Million (RPM) using deeptools [145]. Each enhancer (in either species) was assigned an 'accessibility score' by extracting the mean DNase signal (abovementioned bigwig files) over the enhancer boundaries (performed in Galaxy [141] using the "Compute mean/min/max of intervals" tool version 1.0.0). To make accessibility scores comparable across stages, for each time point, we performed the same normalization as we did for ChIP scores, i.e. replacing accessibility scores greater than $\mu + 3\sigma$ with $\mu + 3\sigma$, where $\mu$ and $\sigma$ are the mean and standard deviation, and then applying min-max normalization. NGS raw sequence data has been deposited in ArrayExpress under accession numbers E-MTAB-3797 (*D. melanogaster* and *D. virilis* DNAse developmental time courses)

## 4.2 EVOLUTIONARY CHANGES IN TF-DNA BINDING AND DNA ACCESSIBILITY IN THE CONTEXT OF A WELL-CHARACTERIZED REGULATORY NETWORK

We calculated evolutionary changes in TF occupancy as the difference of normalized quantitative ChIP scores between orthologous enhancers, for each TF at each TP ('Δ TF:TP'). We noted extensive correlations among different ΔTF:TP measures (Figure 4.1C), i.e., evolutionary changes of TF-DNA binding profiles are correlated. This is especially true of binding profiles of the same TF at different time points, i.e., if a TF loses binding at a location, it tends also to lose binding at the same location at a different developmental stage. For example, Pearson correlation coefficient (PCC) of ΔBin:TP3 (changes in Bin binding at TP3) and ΔBin:TP4 is 0.58 (p-value=2.30E-247), and that between ΔMef2:TP4 and ΔMef2:TP5 is 0.56 (p-value=3.66E-227). The natural explanation for this observation is that loss or gain of the TF's motif plays a significant role in evolutionary changes of TF binding. More interestingly, changes in DNA binding of different TFs at the same time point also show correlations (Figure 4.1D), e.g., ΔTin:TP2 and ΔTwi:TP2 have a Pearson correlation of 0.50 (p-value=3.69E-174). Since the five TFs have different binding preferences (motifs, see Figure 4.2), these correlations most likely arise due to co-binding of specific pairs of TFs – a possibility that we examined in [74], or from changes in accessibility , which is a common contributing factor to DNA binding profiles of different TFs [96, 97].

We also compared the normalized DHS accessibility scores of the same set of ~2,500 orthologous enhancers mentioned above, at each of the three time-points (Figure 4.1E). Most of the accessibility scores are conserved between species, while some enhancer pairs exhibit substantial change. For instance, at TP1, of all the enhancer pairs whose accessibility score is above 0.3 (median score, on a scale of 0 to 1) in at least one of the two species, 6.86% have their orthologous accessibility score below 0.1. We also compared evolutionary changes in accessibility between orthologous enhancers at different developmental stages and found, as expected, that the temporally proximal time points, e.g., TP3 and TP5, or TP1 and TP3,

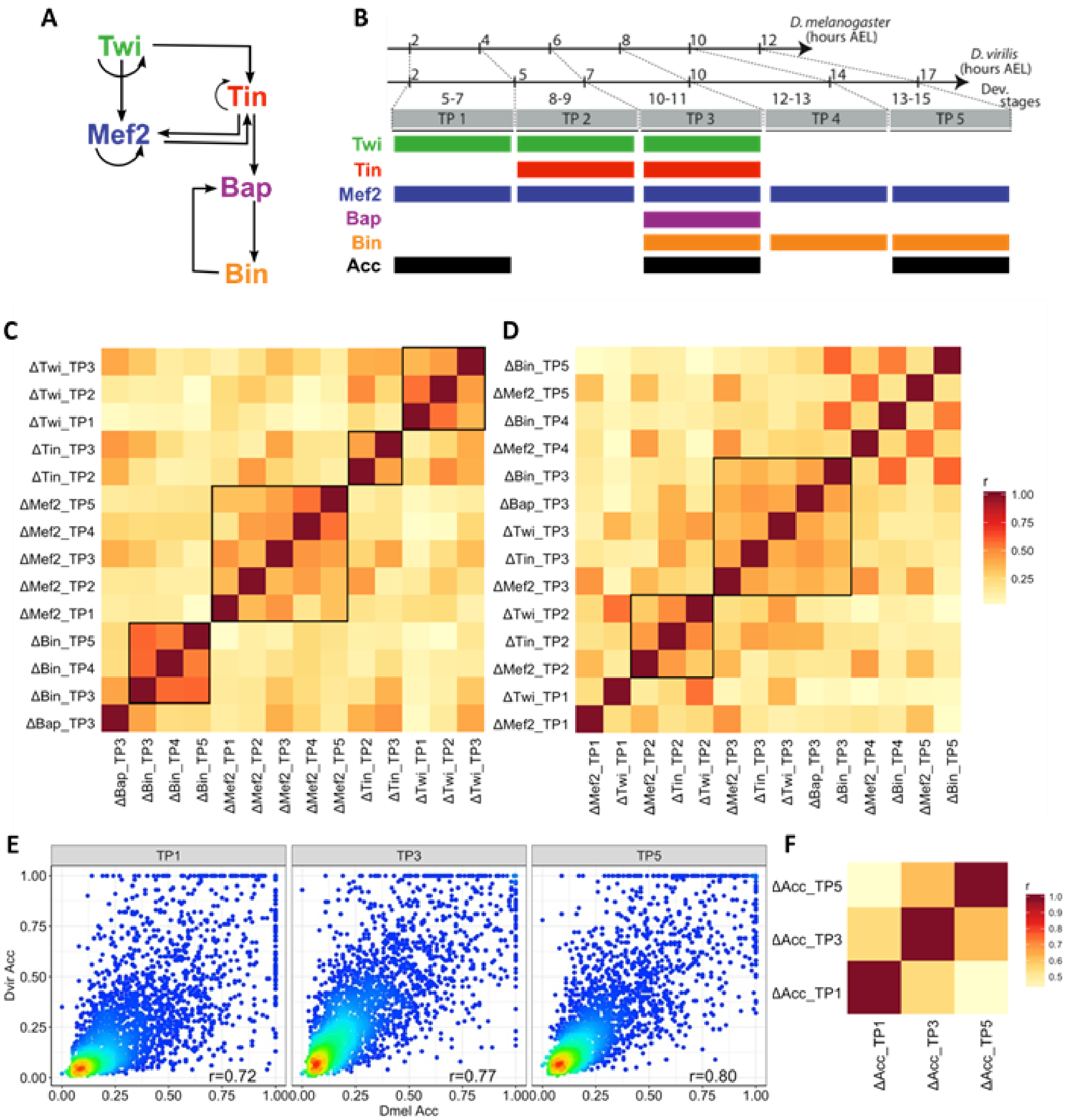

Figure 4.1: **Examining evolutionary changes in TF binding and accessibility across developmental time points.** (A) Regulatory network of five key TFs in mesoderm specification, source:[74]. (B) Data from *D. melanogaster* and *D. virilis* TF ChIP and DNase I hypersensitivity assays were collected. Colored boxes indicate time points (TP1-5) for which each type of genomic profile is available. Orthologous developmental stages between species were mapped according to hours of development in each species, after egg laying (AEL). (C,D) Pairwise Pearson correlations of interspecies ChIP changes, sorted by TF (C) or by time points (D) (E) Normalized accessibility scores of orthologous enhancers for three time points (TP1,3,5). Colors indicate point density, with warmer colors denoting greater density. Pearson correlations between *D. melanogaster* TF ChIP and *D. virilis* TF ChIP are also shown. (F) Pairwise Pearson correlations of interspecies accessibility changes. Data and analysis shown in (C-E) pertain to over 2,500 pairs of putative orthologous enhancers involved in mesoderm specification as defined in text.

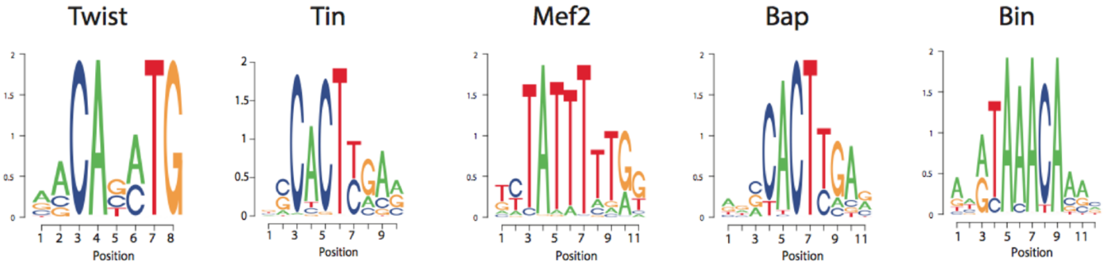

Figure 4.2: **Position weight matrices (PWMs) are shown in the form of sequence logos**, source: [74].

have more correlated evolutionary changes than more separated time points, i.e., TP1 and TP5 (Figure 4.1F). In short, we collated data on TF binding and DNA accessibility at the same stages of embryogenesis in two species, and confirmed previous reports of evolutionary flux in these important measures of the *cis*-regulatory landscape, setting the stage for a closer examination of their mutual relationship.

## 4.3 RELATIONSHIP BETWEEN CHANGES IN CHROMATIN ACCESSIBILITY AND TF BINDING AT ORTHOLOGOUS DEVELOPMENTAL ENHANCERS

We sought to systematically and quantitatively dissect how evolutionary changes in TF binding are related to changes in accessibility. Given the observations above, that the occupancy of different TFs from the same time point tend to change concordantly, it was natural to ask: "how frequently do changes in TF binding between species coincide with changes in DNA accessibility?" We collected orthologous enhancer pairs that are accessible in at least one of the two species (normalized accessibility score $> 0.3$), and examined the relationship between change of accessibility score ('$\Delta$Acc') and change of TF occupancy. As shown in Figure 4.3A, for Twi binding at TP1, enhancers with conserved binding (points closer to diagonal) typically have conserved accessibility (warmer colors), while enhancers with changes of TF binding (off-diagonal points) tend to exhibit changes in their accessibility score (cooler colors) (Pearson correlation between $\Delta$Acc and $\Delta$ChIP is r=0.12, p-value 1.44E-5). Other TF:TP combinations showed the same trend (Figure 4.4).

While the above observations were statistically significant, the strength of relationship between accessibility and TF binding changes revealed by them seemed modest. In part, this may be because the quantified change in TF binding depends not only on the change of accessibility ($\Delta$Acc) but also on the actual accessibility in either species. Thus, to make the above analysis more systematic, we trained a regressor to predict TF binding changes using accessibility scores. For every TF:TP condition, we trained Support Vector Regression (SVR) models, using the R package 'e1071' [146], to predict the interspecies differences in ChIP scores, defined as $\Delta\text{ChIP} = \text{ChIP}_{\text{Dmel}} - \text{ChIP}_{\text{Dvir}}$ for each enhancer. We used the set of 2,754 orthologous putative enhancer pairs to train and evaluate models. For each orthologous pair, two kinds of input features were measured to predict ΔChIP: 1) *D. melanogaster* accessibility score and 2) interspecies accessibility score changes ($\Delta\text{Acc} = \text{Acc}_{\text{Dmel}} - \text{Acc}_{\text{Dvir}}$) for the appropriate time point. Goodness of fit was measured by Pearson correlation coefficient between measured and model-predicted ΔChIP values, using 5-fold cross-validation.

We found that changes in accessibility are modestly predictive of changes in TF binding between species, with correlation coefficients varying substantially across the 9 data sets, averaging about 0.25 (Figure 4.3B). To provide an intuitive calibration of this value, we note that it was computed over 2,754 samples and has a p-value of 1.64E-40. As an alternative evaluation of the predictions, we asked how well the model-predicted ΔChIP values classify the enhancer pairs with the greatest increase in TF binding (measured ΔChIP in top 10 percentile among all 2,754 orthologous pairs) versus those with the greatest decrease in binding (ΔChIP in bottom 10 percentile). We noted an AUROC of 0.78 or greater on such balanced data sets for four of the 9 TF:TP pairs (Figure 4.5). Among the best examples was Twi:TP1, where correlation between measured and predicted ΔChIP on the full set of 2,754 orthologous pairs is 0.44 (p-value 8.94E-131), i.e., about 20% of the variance ($r^2 = 0.19$) of ΔChIP is explained by accessibility changes for this condition (Figure 4.3C). What mechanisms might underlie this relationship? An intriguing but untested possibility is that Twi is the major factor dictating open chromatin i.e., perhaps having a pioneering role at these sties, in keeping with its role as a 'master regulator' being sufficient to convert cells to a mesodermal fate [102]. Alternatively, there may be unmeasured changes in an additional factor required to open chromatin and facilitate Twi binding to these sites. Zelda is a very good candidate, as it is required for Twi binding to some early developmental enhancers [147] and is thought to play a pioneer role in early *Drosophila* embryogenesis [124, 148, 149]. Such mechanistic speculations notwithstanding, the above results – that even in the best example only 20% of variance is explained – emphasize the potential existence of influences other than accessibility, and that operate without major effects on accessibility, on TF binding change.

A natural comparison point for the above correlations is the extent to which accessibility score in a single species can predict TF binding in that species in the same time point, across the same set of enhancers as above. It was not a *priori* clear what the result of this comparison might be. It is possible that accessibility changes are less prominently associated with evolutionary changes of TF binding than the extent to which accessibility is informative of TF occupancy in a single species [96, 97], for instance if most binding changes arise from

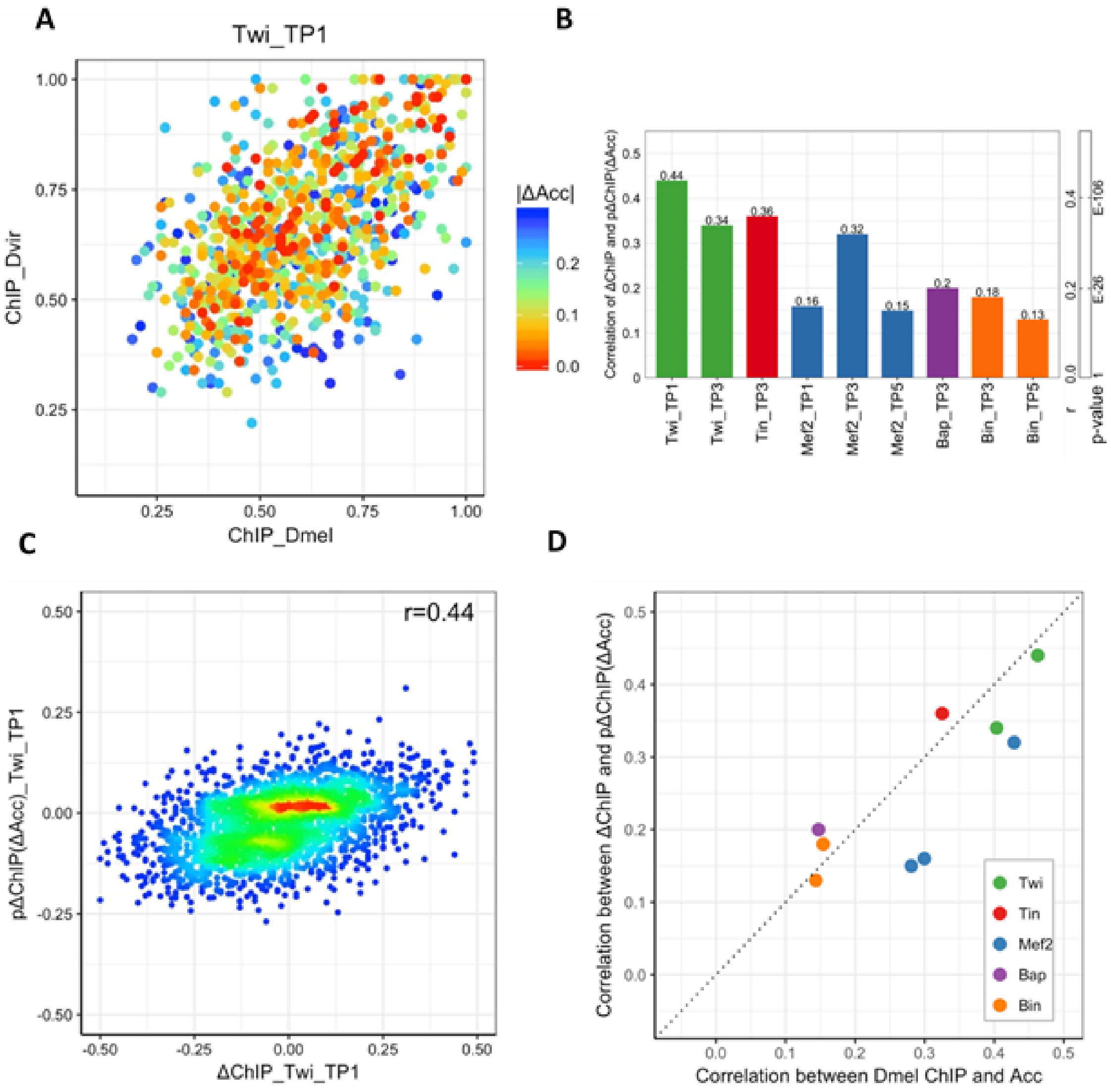

Figure 4.3: **Accessibility changes alone are modest predictors of TF occupancy changes between species.** (A) Scatter plot of *D. melanogaster* ChIP scores versus *D. virilis* ChIP scores for Twi at TP1. Points represent orthologous enhancers that are accessible in at least one species. Colors indicate change of accessibility score. (B) Correlation coefficient between measured ΔChIP and ΔChIP predicted based on ΔAcc, denoted as ‘pΔChIP(ΔAcc)’. P-values of Pearson correlation coefficient (r) with sample size of 2754 are also shown. (C) Scatter plot of ΔChIP versus pΔChIP for Twi at TP1. Warmer colors indicate greater point density. (D) Correlation between ChIP and accessibility in *D. melanogaster* (x-axis) is compared to correlation between interspecies ΔChIP and pΔChIP(ΔAcc).

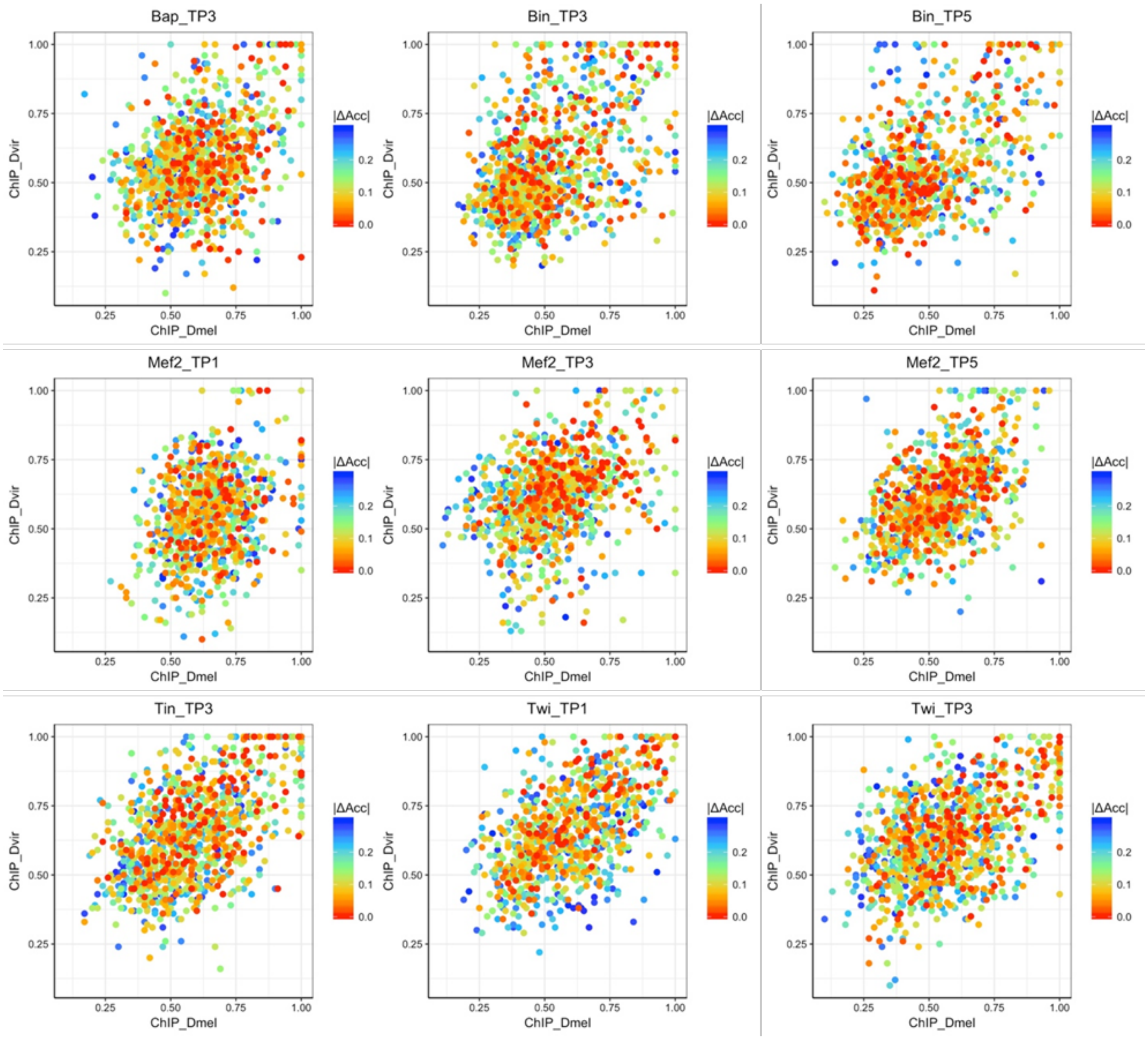

Figure 4.4: **Scatter plots of *D. melanogaster* ChIP scores versus *D. virilis* ChIP scores for each TF:TP combination.** Points represent orthologous enhancers that are accessible in at least one species. Colors indicate change of accessibility score.

turnover of motif hits. On the other hand, the single species correlation values between accessibility and TF binding might not be as high here as reported in some previous studies [33, 26], since our analysis is restricted to putative enhancers, which as a class have high accessibility levels. Our single-species correlation analysis will not reflect the strong genome-wide trend of ChIP peaks coinciding with accessible regions. With these two considerations in mind, it was thus instructive to find that correlations between ChIP score and accessibility score in *D. melanogaster* (Figure 4.3D) were similar to those between evolutionary changes in these scores, i.e., $\Delta$ChIP and predicted $\Delta$ChIP based on accessibility, for every TF:TP condition.

We noted above (Figure 4.1D) that changes in binding for some TF pairs in the same time point, are strongly correlated. We asked if these co-divergence patterns can be explained by changes in accessibility, since accessibility can be simplistically thought of as setting up a 'landscape' for binding, on which different TFs act differently to set up their own binding profiles. Evolutionary changes in accessibility can therefore be expected to impact binding of multiple TFs in similar ways. To test this possibility, we computed a statistic similar to the partial correlation of $\Delta$ChIP between each pair of TFs, given accessibility data. For each pair of TFs, we first computed the residuals of accessibility-based predictors of $\Delta$ChIP for either TF, and then calculated the correlation coefficient between these residuals. This approach removes the effect of accessibility changes in assessing the correlation of $\Delta$ChIP between TF1 and TF2. We found that for the majority of data set pairs (10 out of 16) where $\Delta$ChIP scores of two TFs at the same time point are strongly correlated (PCC $> 0.2$), correlations are lower (a difference of at least 0.04) after excluding the influence of accessibility (Table 4.1), though the (partial) correlations remain strong even after accounting for $\Delta$Acc. For Twi and Tin at TP2, for example, the correlation of $\Delta$ChIP scores drops from 0.5 to 0.45 upon 'removing' accessibility; a similar reduction is observed for the same pair of TFs at TP3, where the correlation drops from 0.39 to 0.32. Another example is that of Mef2 and Tin at TP2, where the correlation reduced from 0.45 to 0.37 upon accounting for accessibility changes. Indeed, previous work reported the potential role of Tin-Twi and Tin-Mef2 co-binding in the evolution of binding sites for these TFs [74]. Our results reveal that changes in accessibility do explain part of the co-divergence of DNA binding exhibited by pairs of TFs, but other causes of co-divergence [150, 74], e.g., cooperative occupancy, functional change of an enhancer and the resulting shared changes of selective pressure, also exist.

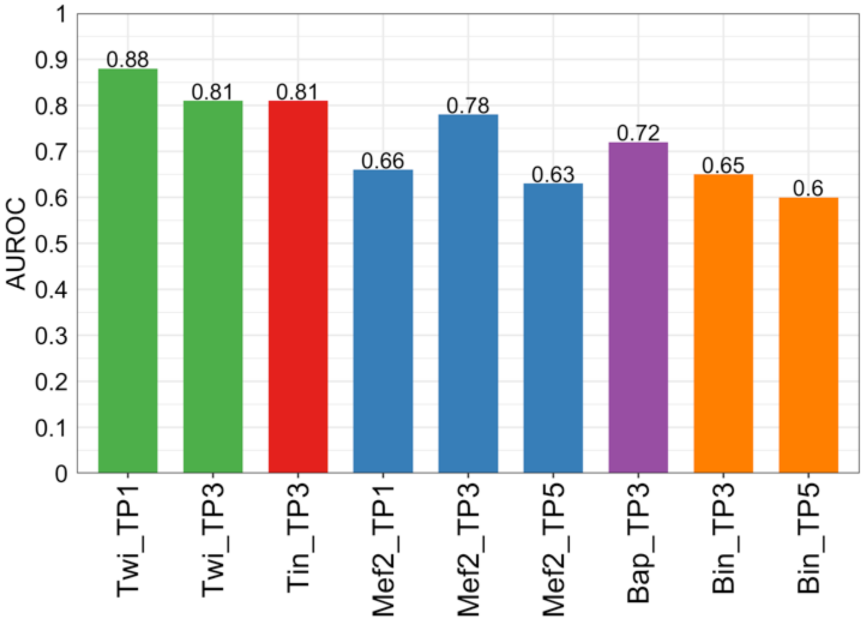

Figure 4.5: **AUROC measure of classification accuracy.** Predicted ΔChIP based on ΔAcc (pΔChIP(ΔAcc)) is used to classify enhancer pairs with the greatest increase in TF binding versus those with the greatest decrease in binding.

Table 4.1: **Effect of accessibility on interspecies ChIP changes.** Pairwise Pearson correlation coefficients (CC) of ΔChIP of TF1 and TF2 at the same time point 'x' are shown, and partial correlation coefficients after excluding the effect of accessibility. Bold fonts indicate cases where the difference between CC and partial CC is at least 0.04.

| **ΔTF1_TPx** | **ΔTF2_TPx** | **CC** | **Partial CC** |
|---|---|---|---|
| ΔMef2_TP2 | ΔTin_TP2 | **0.45** | 0.37 |
| ΔBin_TP3 | ΔMef2_TP3 | **0.32** | 0.25 |
| ΔTin_TP3 | ΔTwi_TP3 | **0.39** | 0.32 |
| ΔBap_TP3 | ΔMef2_TP3 | **0.38** | 0.32 |
| ΔMef2_TP3 | ΔTwi_TP3 | **0.34** | 0.29 |
| ΔTin_TP2 | ΔTwi_TP2 | **0.50** | 0.45 |
| ΔBap_TP3 | ΔTin_TP3 | **0.46** | 0.42 |
| ΔBap_TP3 | ΔBin_TP3 | **0.42** | 0.38 |
| ΔBap_TP3 | ΔTwi_TP3 | **0.41** | 0.37 |
| ΔBin_TP3 | ΔTwi_TP3 | **0.30** | 0.26 |
| ΔMef2_TP3 | ΔTin_TP3 | 0.42 | 0.38 |
| ΔBin_TP3 | ΔTin_TP3 | 0.37 | 0.34 |
| ΔBin_TP4 | ΔMef2_TP4 | 0.24 | 0.23 |
| ΔBin_TP5 | ΔMef2_TP5 | 0.19 | 0.19 |
| ΔMef2_TP2 | ΔTwi_TP2 | 0.30 | 0.31 |
| ΔMef2_TP1 | ΔTwi_TP1 | 0.21 | **0.27** |

## 4.4 CHANGES IN ACCESSIBILITY AND SEQUENCE PREDICT TF BINDING CHANGES TO SIMILAR EXTENTS

The results above quantified the extent to which change of accessibility (ΔAcc) predicts changes in TF-DNA binding (ΔChIP) between orthologous enhancers. We next determined how strongly changes in sequence, in terms of binding motif presence, predict ΔChIP, with the ultimate goal of comparing the relative contributions of changes in accessibility and in sequence to divergence of TF binding. To approach this goal, it is important to have a means of quantifying a TF's motif presence in a given sequence accurately enough to allow quantitative assessment of motif change between orthologous enhancers. We used our previously published STAP (Sequence To Affinity Prediction) model [37] for this purpose. STAP is a thermodynamics-based model that integrates one or more strong as well as weak binding sites, using a given motif, to predict net TF occupancy within a DNA segment. The STAP score is a more realistic estimation of motif presence in a window, compared to using the strength of the best motif match or counting the number of matches above a threshold. Importantly, it is not a confidence score of a single binding site (e.g., CENTIPEDE [38]) and is thus better suited to assess net sequence change in developmental enhancers, which often exhibit homotypic site clustering [151, 152] and suboptimal sites [153, 154].

For each TF:TP condition in each species, we trained a STAP model, following the procedures used in Cheng et al. [33]. We chose the top 1,000 ChIP peaks as the positive training set and 1,000 random windows of the same length as the negative training set, along with their respective normalized ChIP scores. ChIP peaks overlapped with the orthologous enhancers were excluded in training set. The binding motif (position weight matrix, PWM) for each TF was based on the best performing PWMs discovered from *D. melanogaster* and *D. virilis* ChIP data [74] (Figure 4.2). A single free parameter of STAP was learned based on this training set.

To assess the performance of STAP model on each of the 28 ChIP data sets in the given TF, time point, and species combination, we applied four-fold cross-validation on the 2,000 DNA segments training set. Each fold used 1,500 DNA segments to train the single free parameter in STAP, and 500 DNA segments to score. The resulting 2,000 STAP scores, aggregated from each fold, were compared to respective ChIP scores, by Pearson correlation coefficient (PCC). These 28 STAP models fit the ChIP data well, and showed an average PCC of 0.51 (Table 4.2). We also checked the single parameter of STAP learned in each fold, and observed similar values across four folds.

Once the STAP model was trained for every TF, time point, species combination, we used the STAP model to score each enhancer for motif presence. STAP scores were further

Table 4.2: **Evaluation of trained STAP models on 28 TF-ChIP data sets.** Pearson correlation coefficient (PCC) between ChIP scores and STAP scores for each TF- and stage-specific model reported.

| **Data set** | **PCC** | **Data set** | **PCC** |
|---|---|---|---|
| Dmel_Twi_TP1 | 0.48 | Dvir_Twi_TP1 | 0.36 |
| Dmel_Twi_TP2 | 0.53 | Dvir_Twi_TP2 | 0.53 |
| Dmel_Twi_TP3 | 0.56 | Dvir_Twi_TP3 | 0.39 |
| Dmel_Tin_TP2 | 0.58 | Dvir_Tin_TP2 | 0.56 |
| Dmel_Tin_TP3 | 0.56 | Dvir_Tin_TP3 | 0.62 |
| Dmel_Mef2_TP1 | 0.48 | Dvir_Mef2_TP1 | 0.11 |
| Dmel_Mef2_TP2 | 0.42 | Dvir_Mef2_TP2 | 0.32 |
| Dmel_Mef2_TP3 | 0.37 | Dvir_Mef2_TP3 | 0.43 |
| Dmel_Mef2_TP4 | 0.62 | Dvir_Mef2_TP4 | 0.6 |
| Dmel_Mef2_TP5 | 0.59 | Dvir_Mef2_TP5 | 0.6 |
| Dmel_Bap_TP3 | 0.37 | Dvir_Bap_TP3 | 0.53 |
| Dmel_Bin_TP3 | 0.63 | Dvir_Bin_TP3 | 0.59 |
| Dmel_Bin_TP4 | 0.58 | Dvir_Bin_TP4 | 0.59 |
| Dmel_Bin_TP5 | 0.59 | Dvir_Bin_TP5 | 0.56 |

normalized in the same way as ChIP scores, i.e. capping outliers at $\mu + 3\sigma$, where $\mu$ and $\sigma$ are the mean and standard deviation, and then applying min-max normalization.

Once we trained the STAP models to predict TF occupancy based on motif presence, we proceeded to quantify the extent to which change of motif presence predicts changes in TF-DNA binding ($\Delta$ChIP) between orthologous enhancers. For each orthologous enhancer pair, we calculated STAP scores of either ortholog using a TF's motif, and thus obtained a '$\Delta$STAP' score quantifying the evolutionary change in motif presence for that TF. Next, we used the *D. melanogaster* STAP score and the $\Delta$STAP score ($\Delta\text{STAP} = \text{STAP}_{\text{Dmel}} - \text{STAP}_{\text{Dvir}}$) together to predict $\Delta$ChIP for each orthologous enhancer pair, using a Support Vector Regression (SVR) algorithm, similar to what was done for accessibility scores in the previous section. This was repeated for each TF:TP condition. We found that the predicted and measured $\Delta$ChIP are modestly correlated, with average correlation coefficients in the 14 TF:TP conditions being 0.3 (Figure 4.6A, each reported correlation is an average across 5-fold cross validation). It was notable that most conditions exhibited similar correlations, with 9 of the 14 yielding values between 0.27 and 0.33, and the highest correlation (0.38) seen for the Bin-TP3 and Bin-TP5 conditions. By way of calibration, we similarly computed correlations between STAP and ChIP scores in each species separately, across the same set of enhancers as above. We noted that correlation coefficients are 0.68 for *D. melanogaster* and 0.61 for *D. virilis* (Figure 4.7), on average across the 14 conditions. This assured us that STAP provides an accurate estimate of motif content, which is strongly predictive of TF occupancy, and does so in both species. However, it also highlights the poorer predictability of evolutionary changes in binding from change in sequence compared to the ability to predict binding from sequence in a single species.

We next sought to compare the accuracy of $\Delta$STAP-based predictions of $\Delta$ChIP to that of $\Delta$Acc-based predictions, with the intention of assessing the relative contributions of sequence- and accessibility-level changes to TF binding change between species. For this, we modified the accessibility-based predictor introduced above, which used the accessibility scores for the time point matching the ChIP data set, to now use data from all three time points with available data. This allowed us to predict $\Delta$ChIP scores even for the two time points – TP2 and TP4 – for which accessibility data were not available, by basing those predictions on accessibility scores from TP1, TP3 and TP5 (See Figure 4.8 for clarification about a potential methodological concern that this might raise). Correlation coefficients between predicted and measured $\Delta$ChIP scores (Figure 4.6B) had an average value of 0.29 across the 14 TF:TP conditions, which is comparable to the 0.30 average correlation seen above with motif-based predictors (Figure 4.6A), though there is a greater variation across TF:TP conditions when using accessibility-based predictors.

We then made direct comparisons between motif-based and accessibility-based predictors of $\Delta$ChIP scores for every TF:TP condition (Figure 4.6C). In some cases, e.g., Twi at TP1 and Tin at TP2, changes in accessibility shows better predictive power than changes in motifs (PCC values of 0.45 vs. 0.3 for Twi:TP1 and 0.44 vs. 0.33 for Tin:TP2). This is unlikely to be due to inferior strong correlations with ChIP (Figure 4.7). It may be in part because DNA-binding of these two motifs used in the STAP models, as the single species STAP models for both Twi and Tin show TFs is believed to depend not only on their own motif but also on co-binding with each other [74].

In other cases, such as Bin (at all three time points), change in motif presence is a far better predictor of binding change than are changes in accessibility. This is in concordance with our previous studies in a single species – Bin motifs are very predictive of Bin binding [155, 74]. For Mef2, the only TF expressed and with ChIP measurements at all five time points, $\Delta$ChIP values at later time points are predicted better using the motif-based predictor and earlier $\Delta$ChIP values are better predicted using accessibility changes, even though the motif used is the same in all cases. Interestingly, we note that this is part of a general trend for accessibility-based predictions to be better at earlier time points than later ones, such as TP4 and TP5 (Figure 4.8B). This trend may be due increased embryo heterogeneity at later developmental stages having a distortive effect on cell type specific accessibility seen in bulk whole embryo DHS measurements or alternatively due to pioneering roles of early TFs

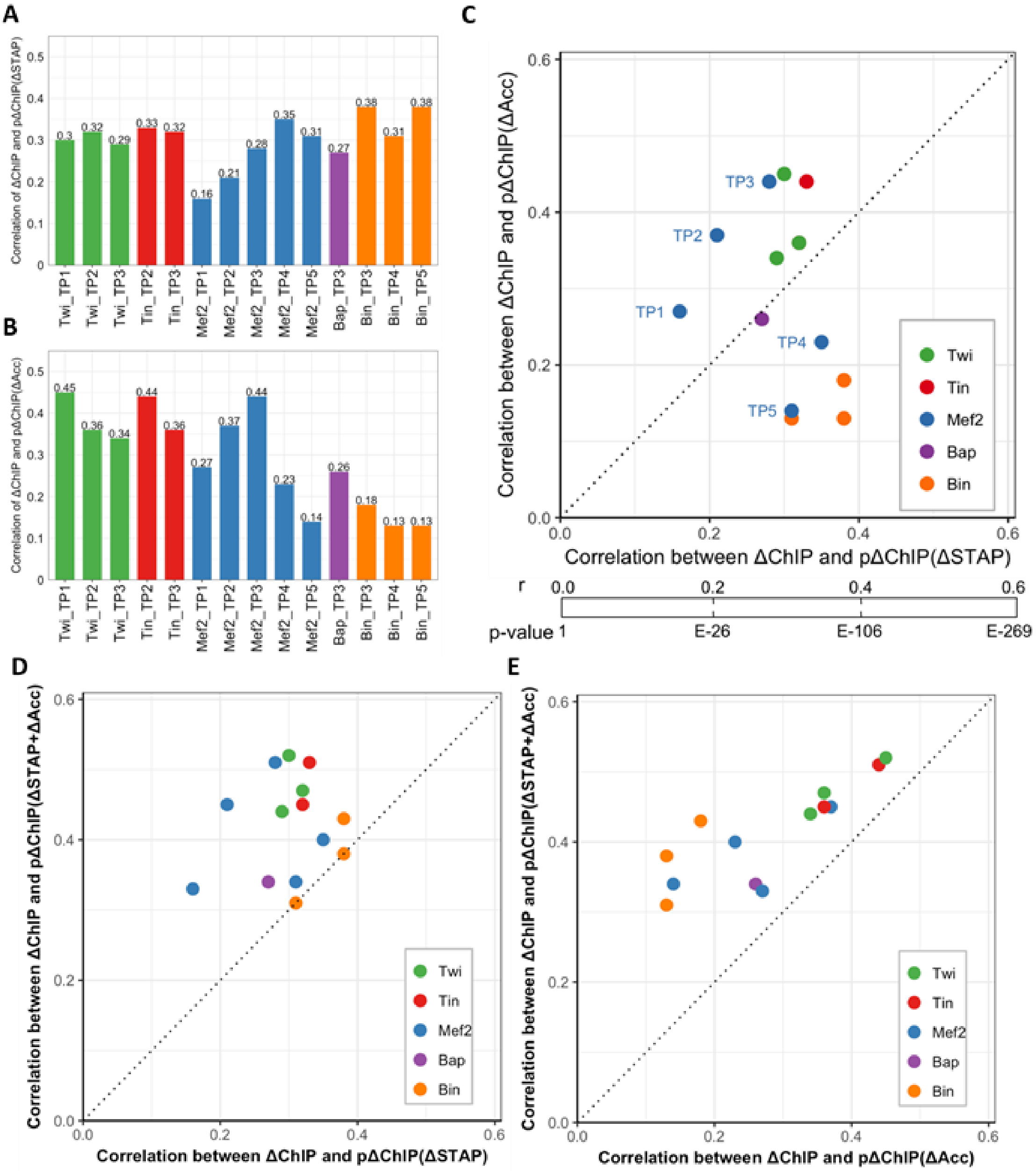

Figure 4.6: **Changes in motif presence and accessibility are both predictive of TF occupancy change.** (A) Correlation between measured ΔChIP and ΔChIP predicted by models based on motif presence changes, denoted by 'pΔChIP(ΔSTAP)'. For each TF-TP condition, average Pearson correlation coefficient from 5-fold cross-validation is reported. (B) Similar to (A), except that the ΔChIP predictions are now based on changes in accessibility, denoted by 'pΔChIP(ΔAcc)'. These values are similar to those reported in Figure 4.3, but with slightly modified models (see text). (C) Comparison of motif-based models (x-axis) and accessibility-based models (y-axis). P-values of Pearson correlation coefficient (r) with sample size of 2754 are also shown. (D,E) Predictions of ΔChIP based on both motif changes and accessibility changes, denoted by pΔChIP(ΔSTAP+ΔAcc)', are better than using only motif changes (D) or only accessibility changes (E).

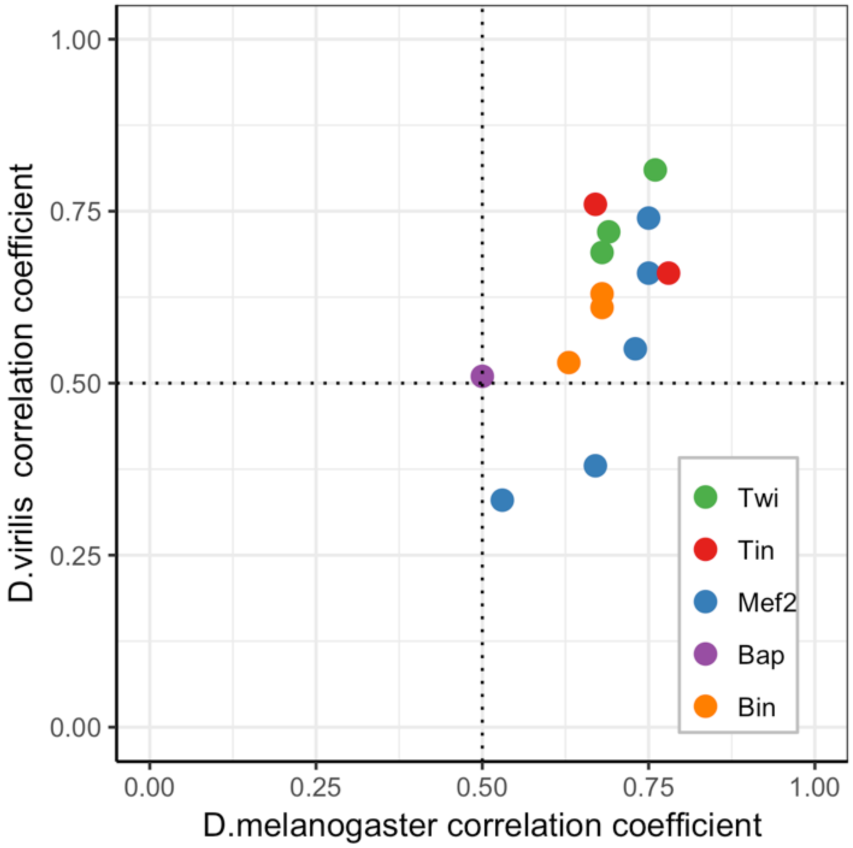

Figure 4.7: **STAP models accurately fit TF occupancy (ChIP) data in single species, either *D. melanogaster* (x-axis) or *D. virilis* (y-axis).** For each TF-time point condition, average Pearson correlation coefficient from 5-fold cross-validation is shown.

priming enhancers for activation at later stages of embryogenesis.

Having found that the contribution of ΔAcc to ΔChIP is similar in extent to the contribution of ΔSTAP (change of motif presence) to ΔChIP, we asked if combining these two pieces of information would further improve our ability to predict binding changes. Generally, the answer is yes, for almost all TF:TP pairs the ΔChIP scores are better predicted with combined models (SVR using *D. melanogaster* STAP, ΔSTAP, Acc and ΔAcc features), achieving correlations in the range of 0.3 to 0.5, with an average of 0.4 across the 14 data sets, compared to ∼0.3 when using accessibility or sequence alone (Figures 4.6D, 4.6E - as before, the reported PCC for each TF:TP condition is an average from 5-fold cross validation). Following the evaluations performed in accessibility-based predictor, we asked how well the combined model-predicted ΔChIP values classify the enhancer pairs with the greatest increase in TF binding (measured ΔChIP in top 10 percentile among all 2,754 orthologous pairs) versus those with the greatest decrease in binding (ΔChIP in bottom 10 percentile). We noted an AUROC of 0.84 or greater on such balanced data sets for four of the 9 TF:TP pairs (Figure 4.9). The performance of this joint predictor is a quantitative summary of how well we understand the determinants of TF-binding changes between orthologous enhancers in a well-studied regulatory system.

Our results suggest that the two types of information (motif and accessibility) are complementary in their contribution to predicting changes in binding (most points are above the diagonal in Figure 4.6D, 4.6E). For instance, the strongest correlation observed with the joint predictor is for TWI-TP1, with a PCC of 0.52, compared to 0.3 when using motif change alone and 0.45 when using accessibility change alone. The only exceptions are data sets for Bin, where predictions of occupancy change based on sequence changes are nearly unaffected after adding accessibility information (Figure 4.6D), which implies that motif change alone is a strong predictor of Bin occupancy divergence. We note that in order to make such direct comparisons between determinants of binding change, we have used an approach that goes beyond testing statistical enrichments of various events, such as motif loss or gain, in regions of binding change.

## 4.5 A STRATEGY TO ASSESS PREDICTIONS OF BINDING CHANGE RELEVANT TO ENHANCER ACTIVITY CHANGE

In the analysis above, we quantified the ability to predict changes in binding by directly correlating experimentally measured ΔChIP of a TF with computationally predicted ΔChIP from accessibility and sequence-level changes between orthologous enhancers. What does this imply for one of the ultimate goals of comparative *cis*-regulatory profiling – to predict

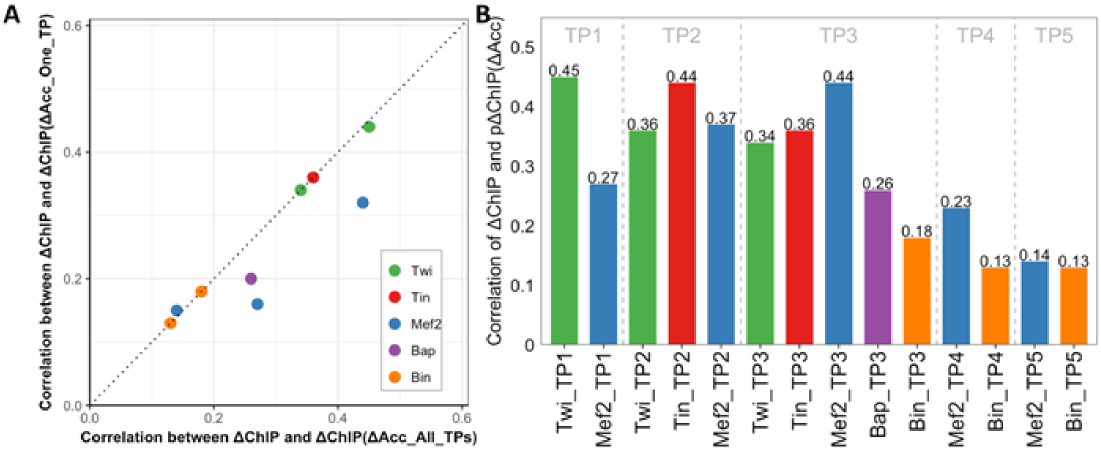

Figure 4.8: **Performance of accessibility-based preditors.** (A) Models using accessibility data from ChIP-matched stage perform nearly as well as models aggregating accessibility data from all available stages. We recognized that using data from multiple time points may inflate the strength of relationship between accessibility and binding changes, leading to unfair comparisons with motif-based predictions of ΔChIP. We therefore compared accessibility-based SVR models that use all three timepoints with models using only the time point matching the ChIP data. We found that for six of nine TF-TP conditions the two models yield equal correlations, while for three conditions the correlation is slightly better when utilizing multiple time points. (B) Performance of accessibility-based predictors is better at early time points. (Adapted from Figure 4.6B, sorted by time points.)

changes in enhancer-driven expression? Prior work has shown that one can predict spatio-temporal activity of mesoderm enhancers based on ChIP data for the set of five TFs studied here [110]. We asked therefore if our ΔChIP predictions agree with the experimentally measured ΔChIP values when examined through the lens of such an activity prediction model, rather than through direct correlations. In other words, if we knew the ChIP values in an enhancer, and the sequence *and* accessibility changes between it and an orthologous enhancer, can we predict ChIP values in the ortholog and use them to determine if the enhancer's spatio-temporal activity is conserved? If so, it would indicate that our understanding of binding changes is accurate enough to be of predictive value. Note that such a comparison must integrate the information from ΔChIP scores for multiple TFs, rather than compare each TF:TP separately as was done above. In this sense, we now aim to assess ΔChIP predictions in a more integrative manner.

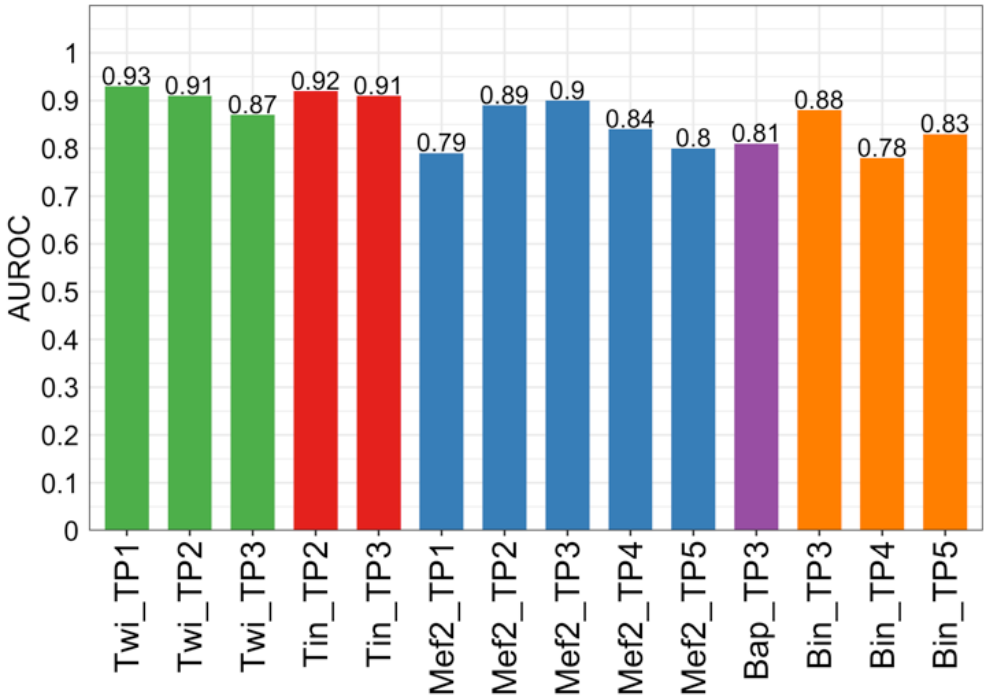

Figure 4.9: **AUROC measure of classification accuracy.** Predicted ΔChIP based on motif changes and accessibility changes (pΔChIP(ΔSTAP+ΔAcc)) is used to classify enhancer pairs with the greatest increase in TF binding versus those with the greatest decrease in binding.

### 4.5.1 Outline of approach

A major obstacle in answering this question is the lack of data on changes in enhancer activity. There is a large collection of *D. melanogaster* enhancers with annotated activities [156, 157, 110], but only a small set of *D. virilis* enhancers whose activities were tested experimentally (in transgenic *D. melanogaster* embryos) [74]. Moreover, this small set of experimentally characterized *D. virilis* enhancers mostly exhibited conserved activity [74], exacerbating the analysis of functional changes. We therefore devised a modeling-based approach to the above question, that can be briefly described as follows (Figure 4.10): (1) Train a model 'A' that predicts an enhancer's activity from its ChIP profile (binding levels for relevant TFs), similar to prior work [110]. (2) Use model 'A' to predict the activities of orthologous *D. melanogaster* and *D. virilis* enhancers, using their respective ChIP profiles, thus characterizing the change in (predicted) activity between the orthologs. (3) Separately, use the ChIP profile of the *D. melanogaster* enhancer and motif and/or accessibility-based predictions of ΔChIP, to *predict* the ChIP profile of the *D. virilis* ortholog. Once again, estimate the activity change between orthologs, but now relying on the predicted ChIP profile of the *D. virilis* ortholog. (4) Compare the activity changes computed in steps (2) and (3), that utilize, respectively, direct ChIP measurements or predicted ChIP profiles in *D. virilis.* The extent to which these changes agree with each other will reveal how well motif and accessibility-based predictions of ΔChIP agree with real ΔChIP when seen through the lens of enhancer function.

### 4.5.2 Predicting enhancer activity from ChIP data

To build a training set for the enhancer activity classifier, we collected known mesoderm enhancers from our previously built CRM Activity Database (CAD) [110], activity information of active tiles from Kvon et al. [157], and a set of new entries from RedFly database [156]. Three activity classes were considered: mesoderm (Meso), visceral musculature (VM), and somatic musculature (SM). Enhancers that drive expression in more than one classes (e.g. Meso and SM or VM and SM) were excluded. We then overlapped the annotated enhancers with our 2,754 orthologous putative ChIP enhancers in *D. melanogaster.* This led to a final training set of 233 enhancers, with 102 expressed in Meso, 65 in VM, and 66 in SM.

We then trained XGBoost [158] models to predict enhancer activity. XGBoost is a supervised machine learning method that uses training data with multiple features to predict a target variable. For each activity class 'C', an XGBoost classifier Ac was trained by using the R package 'xgboost' [158] to discriminate between members and non-member of the class. Thus, for the Meso class, the positive set includes enhancers with Meso annotation, while the negative set includes enhancers with VM or SM annotations. The input features for each enhancer were a 14-dimensional vector of ChIP scores of that enhancer pertaining to the 14 TF:TP conditions. To adjust for the imbalanced distribution of training data set, we used the Synthetic Minority Over-sampling Technique (SMOTE) [159], from R package 'DMwR' [160], to oversample the minority class. We trained the XGBoost classifiers in the mode of 'logistic regression for binary classification (binary:logistic)'. Parameters were set as below: 'eta' = 0.2, 'nrounds' = 50, 'max_depth' = 4, 'subsample' = 0.9, 'colsample_bytree' = 0.8, by following the guidelines from XGBoost documentation. For each class 'C', a separate classifier $A_C$ was trained to predict the enhancer's activity on a scale of 0 to 1, representing the confidence of that classifier, and henceforth called the regulatory potential of that enhancer for the class C. (Below, the numeric prediction of $A_C$ will also be denoted by $A_C$). Leave-one-out cross-validation was applied to measure the classifier performance.

We trained $A_C$ on 223 experimentally characterized enhancers [156, 157, 110] associated with the three expression classes, and noted balanced accuracy values around 0.8 in leave-

Table 4.3: **Classifiers trained from combinatorial transcription factor binding data can accurately predict enhancer activities.** Balanced accuracy from leave-one-out cross-validation is shown for models built for each activity class: mesoderm ('Meso'), visceral muscle ('VM'), and somatic muscle ('SM'). Models were trained (and tested) on 223 experimentally characterized enhancers in *D. melanogaster*; for each activity class, enhancers with that activity were positives while enhancers of the other two classes were negatives. The numbers of correctly and incorrectly classified enhancers for each model are listed. TN: true negative, FN: false negative, TP: true positive, FP: false positive.

| | **Meso** | **VM** | **SM** |
|---|---|---|---|
| **TN** | 100 | 145 | 144 |
| **FN** | 13 | 21 | 16 |
| **TP** | 89 | 44 | 50 |
| **FP** | 31 | 23 | 23 |
| **Balanced Accuracy** | 0.82 | 0.77 | 0.81 |

one-out cross validation for each class (Table 4.3). When estimating accuracy for any class, enhancers of that class were treated as positives, and enhancers of the other two classes were considered as negatives. For each classifier, the specificity is ∼0.9 and sensitivity is ∼0.7. We also assessed the accuracy of the trained functions on held-out transgenic reporter assays of *D. melanogaster* and *D. virilis* enhancers [74, 110]. Among 35 experimentally tested enhancers, the predictions of 23 were correct (drove expression in the predicted domain), 3 were partially correct (one of the active tissues was predicted), whereas 9 predictions failed (did not drive any expression in the predicted domain) (Table 4.4). The experimental assays comprised of enhancers in both species, and the accuracy noted in these evaluations justified our assumption that classifiers trained in *D. melanogaster* can be used to predict the activities of *D. virili*s enhancers as well (though in a *D. melanogaster* context).

We noted that similar enhancer activity predictors had been presented in Zinzen et al. [110], where Support Vector Machines (SVMs) trained from ChIP scores were shown to accurately predict enhancer activities in *D. melanogaster*. We rebuilt the classifiers here mainly because our desired tradeoff between sensitivity and specificity was different; in particular, we sought to achieve high values of balanced accuracy when evaluating classifiers on imbalanced data sets (in our case, there are more negative samples than positive samples); see Table 4.5. In addition, ChIP data for Tin at TP1, an input feature for *D. melanogaster* activity classifiers reported in [110], was not available for *D. virilis* [74], further necessitating rebuilding of classifiers.

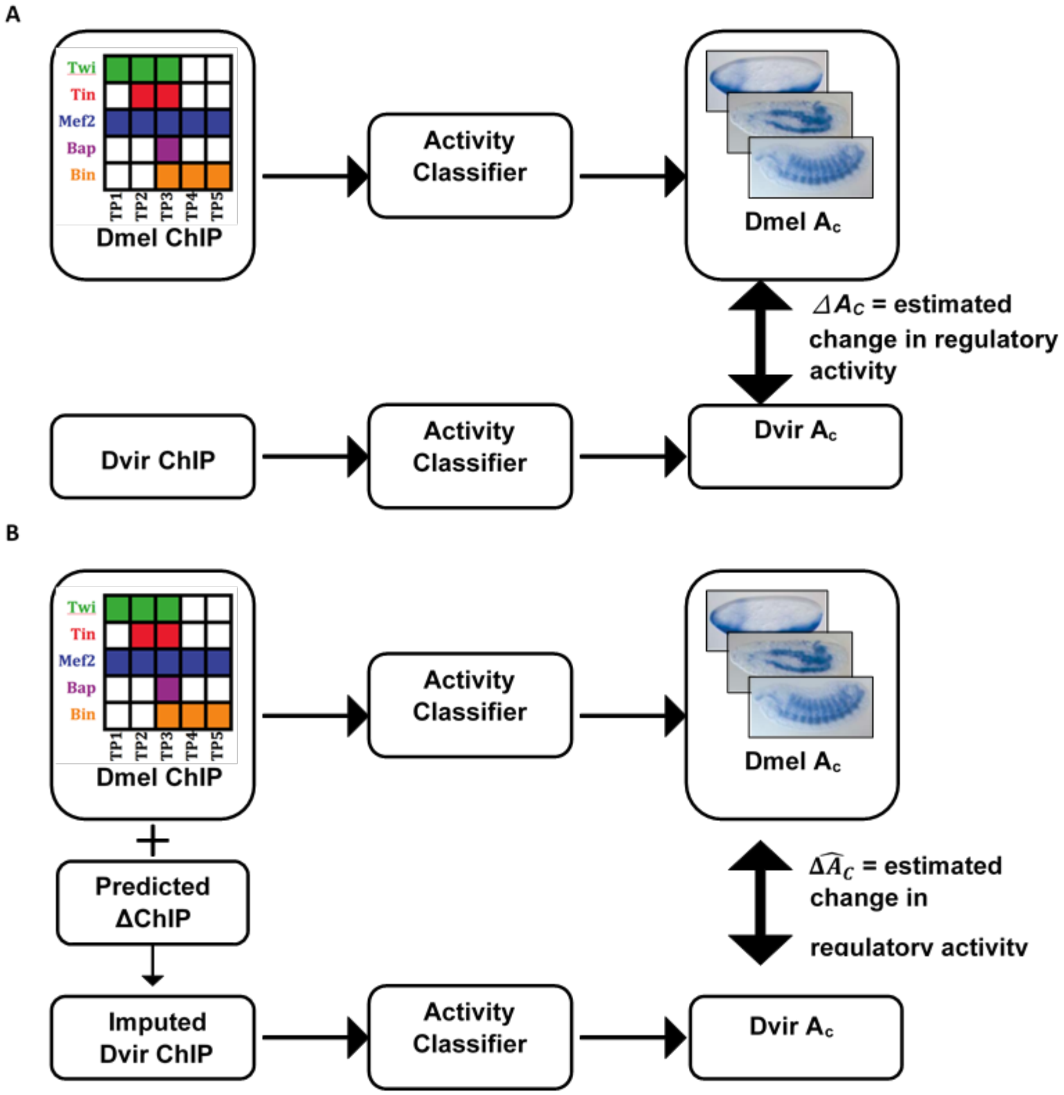

Figure 4.10: **A strategy to assess predictions of binding change through the lens of enhancer activity.** (A) Change in regulatory activity between orthologous enhancers is estimated from difference between output scores of activity classifiers that use *D. melanogaster* and *D. virilis* ChIP profiles respectively as input. (B) An alternative estimate of change in regulatory activity between orthologous enhancers, similar to strategy in (A), except that *D. virilis* activity classifier uses 'imputed' *D. virilis* ChIP profiles as input. Imputation of *D. virilis* ChIP scores is based on *D. melanogaster* ChIP scores and $\Delta$ChIP scores predicted from motif- and/or accessibility-level interspecies changes.

Table 4.4: **Classifier predictions of enhancer activity agree with results of transgenic reporter assays reported in previous studies.** Predicted activities are compared with spatio-temporal expression in three classes: mesoderm (Meso), visceral muscle (VM), and somatic muscle (SM). An enhancer activity class is assigned if the respective classifier prediction value is greater than 0.9. Enhancers in *D. melanogaster* (Dmel) and *D. virilis* (Dvir) were previously tested for *in-vivo* activity in *D. melanogaster* embryos.

| **Species_CRM_ID** | **Experimental essay** | **Predicted activity** | **Prediction correct/failed** |
|---|---|---|---|
| Dmel_CRM_404 | Meso | Meso | correct |
| Dmel_CRM_633 | Meso | - | failed |
| Dmel_CRM_2000 | Meso | Meso | correct |
| Dmel_CRM_2045 | Meso | Meso | correct |
| Dmel_CRM_3407 | Meso | Meso | correct |
| Dmel_CRM_4682 | Meso | - | failed |
| Dmel_CRM_5278 | Meso | Meso | correct |
| Dmel_CRM_6053 | Meso | Meso | correct |
| Dmel_CRM_6176 | Meso | SM | failed |
| Dmel_CRM_388 | Meso weak | - | failed |
| Dmel_CRM_965 | SM | SM | correct |
| Dmel_CRM_1195 | SM | - | failed |
| Dmel_CRM_3215 | SM | SM | correct |
| Dmel_CRM_4575 | SM | SM | correct |
| Dmel_CRM_4725 | SM | SM | correct |
| Dmel_CRM_3027 | SM, VM weak | SM | partially correct |
| Dmel_CRM_160 | VM | VM | correct |
| Dmel_CRM_1560 | VM | - | failed |
| Dmel_CRM_2347 | VM | VM | correct |
| Dmel_CRM_2819 | VM | VM | correct |
| Dmel_CRM_3418 | VM | VM | correct |
| Dmel_CRM_4726 | VM | VM | correct |
| Dmel_CRM_4906 | VM | VM | correct |
| Dmel_CRM_5570 | VM | VM | correct |
| Dmel_CRM_6028 | VM | VM | correct |
| Dmel_CRM_6087 | VM | VM | correct |
| Dvir_CRM_4357 | Meso | - | failed |
| Dvir_CRM_14133 | Meso | - | failed |
| Dvir_CRM_12291 | Meso, SM, VM | SM | partially correct |
| Dvir_CRM_10323 | SM | SM | correct |
| Dvir_CRM_12156 | SM | SM | correct |
| Dvir_CRM_11115 | SM, VM | SM | partially correct |
| Dvir_CRM_459 | VM | VM | correct |
| Dvir_CRM_2469 | VM | SM | failed |
| Dvir_CRM_13100 | VM | VM | correct |

Table 4.5: **Enhancer activity predictions from Support Vector Machine in Zinzen et al. for the 233 experimentally characterized enhancers.** Balanced accuracy is shown for SVM models built for each activity class: mesoderm ('Meso'), visceral muscle ('VM'), and somatic muscle ('SM'). According to Zinzen et al., an enhancer is classified to an activity class if the SVM specificity is greater than 95%. The numbers of correctly and incorrectly classified enhancers for each model are listed. TN: true negative, FN: false negative, TP: true positive, FP: false positive.

| | **Meso** | **VM** | **SM** |
|---|---|---|---|
| **TN** | 130 | 157 | 163 |
| **FN** | 93 | 40 | 48 |
| **TP** | 9 | 25 | 18 |
| **FP** | 1 | 11 | 4 |
| **Balanced Accuracy** | 0.54 | 0.66 | 0.62 |

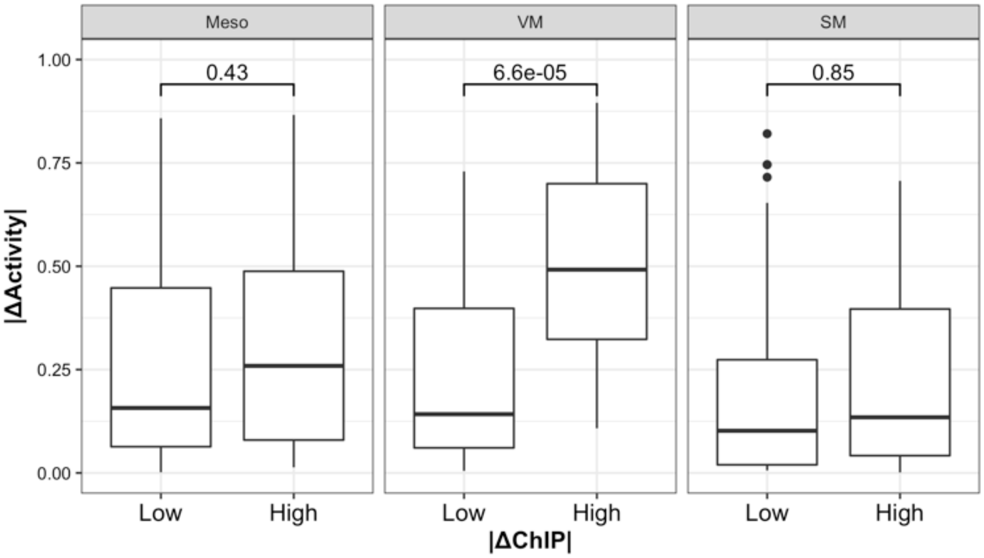

Figure 4.11: **Change in enhancer activities.** Relationship between change of TF binding and model-based change of enhancer activity, examined through 223 experimentally characterized enhancers. For each spatiotemporal expression domain, *D. melanogaster* enhancers with experimentally validated activity in that domain are considered, along with their D. virilis orthologs. Enhancer pairs are divided into "High" and "Low" TF binding change, based on sum of ΔChIP scores for all TFs. Change in predicted enhancer activity ($\Delta A_C$ , see text) is then compared between these two classes.

## 4.6 RELATIONSHIP BETWEEN TOTAL CHANGE IN TF BINDING AND PREDICTED ACTIVITY CHANGES

With an accurate computational model for predicting enhancer activity in hand, we examined TF binding changes between orthologous enhancers in a more contextually informed manner. We began with *D. melanogaster* enhancers that have experimentally confirmed activity in any of the three spatio-temporal expression classes (e.g. 'C') and calculated the regulatory activity $\mathrm{A_C}$ of each enhancer and of its *D. virilis* ortholog, based on their respective ChIP score profiles. We regarded the difference between these two $\mathrm{A_C}$ values ($\Delta\mathrm{A_C} = \mathrm{A_C}(\mathrm{D.mel}) - \mathrm{A_C}(\mathrm{D.vir})$) as an estimate of the change in regulatory activity (specific to class C) between the orthologous enhancers. We then calculated, for each orthologous enhancer pair, the sum of (absolute values of) ΔChIP scores across 14 TF:TP conditions, and used these to categorize the enhancer pairs into two groups of "High" and "Low" change in TF binding (top and bottom 25% respectively), and compared the $\Delta\mathrm{A_C}$ values between these groups (Figure 4.11).

For the 'VM' class, we noted that the enhancer pairs with greater divergence in TF binding ('High' group) tend to exhibit greater change in predicted enhancer activity (P-value = 6.6E-5, Student's t-test). On the other hand, for 'Meso' and 'SM' classes, the two groups exhibit similar distributions of $\Delta\mathrm{A_C}$, suggesting that enhancer activities in these two expression classes are relatively robust to TF occupancy changes. This latter finding is in agreement with our previous study [74], where we assessed the impact of evolutionary changes in TF binding on enhancer activity through *in-vivo* enhancer activity assays, and found five out of seven orthologous enhancer pairs to have conserved activity despite high divergence in TF binding events. The use of activity prediction models developed in the current study allowed us to extend such assessment to 223 experimentally characterized enhancers from *D. melanogaster*, and confirm the finding that observed changes in TF binding at these enhancers may not have a functional impact.

## 4.7 COMPUTATIONALLY IMPUTED CHIP PROFILES AGREE WITH MEASURED CHIP PROFILES IN TERMS OF THEIR PREDICTIONS OF ENHANCER ACTIVITY CHANGES

We next used the strategy outlined in Figure 4.10 to assess if motif and accessibility-based predictions of TF binding change agree with observed binding change when seen through the lens of enhancer function. For each spatio-temporal class 'C' ('Meso', 'VM' and 'SM'), we considered all *D. melanogaster* enhancers with predicted activity $\mathrm{A_C}$ (for that class) in the top 20 percentile, i.e., enhancers with ChIP profiles that are most suggestive of activity in class ‘C’. We further restricted ourselves to the subset of these that exhibited the highest and lowest $\Delta \mathrm{A_C}$ values (in top and bottom 10 percentile for classes ‘Meso’ and ‘VM’, and in the top and bottom 5 percentile for class ‘SM’), i.e., enhancer pairs whose $\Delta$ChIP scores are most strongly indicative of activity change (high $\Delta \mathrm{A_C}$) or conservation (low $\Delta \mathrm{A_C}$). Next, we asked how well these two subsets of orthologous enhancer pairs, with the greatest and least predicted changes in enhancer activity, can be discriminated based on predicted changes of TF binding. To this end, we obtained an ‘imputed’ ChIP profile of the *D. virilis* ortholog, by using the *D. melanogaster* ChIP scores and $\Delta$ChIP scores predicted from interspecies changes in motif presence, accessibility, or both (Figure 4.10B), re-estimated the regulatory activity $\mathrm{A_C}$ of the *D. virilis* ortholog based on this imputed ChIP score profile, and computed its difference from the regulatory activity of the *D. melanogaster* enhancer ($\widehat{\Delta \mathrm{A_C}}$) as an alternative estimate of change in regulatory activity between the orthologs. Finally, we computed the Pearson correlation coefficient between the two estimates $\Delta \mathrm{A_C}$ and $\widehat{\Delta \mathrm{A_C}}$, across all enhancer pairs considered (Table 4.6A), and also noted that AUROC values when $\widehat{\Delta \mathrm{A_C}}$ is used to classify enhancer pairs with high $\Delta \mathrm{A_C}$ versus low $\Delta \mathrm{A_C}$ (Table 4.6B and Figure 4.12).

We noted that when *D. virilis* ChIP score profiles are imputed based on motif and accessibility changes together, the two estimates of activity change have a correlation of 0.47 (P-value 2.21E-7) for the ‘VM’ class, which is substantially greater than the correlation of 0.16 (P-value 0.09) seen in a random control. (In the control setting, *D. virilis* ChIP scores were imputed based on a random permutation of $\Delta$ChIP scores). The high level of agreement between $\Delta \mathrm{A_C}$ and $\widehat{\Delta \mathrm{A_C}}$ is also reflected in the classification AUROC of 0.75, compared to the random control AUROC of 0.59. Similarly, for the ‘SM’ class, a high agreement between the two estimates is borne out by an AUROC of 0.78 (compared to 0.57 in random control), and a correlation coefficient of 0.37 (P-value 0.005), while the random control yields a correlation of 0.09 (P-value 0.51) for this expression class. The correlation and AUROC values are lower for the class ‘Meso’, although clearly statistically significant, e.g., correlation of 0.33 (P-value 0.0004) compared to random correlation of 0.07 (P-value 0.47). Taken together, these results suggest that the accuracy of $\Delta$ChIP predictions demonstrated above (Figures 4.6D, 4.6E), based on modeling interspecies changes in sequence and accessibility, is sufficient for us to make similar predictions of enhancer activity changes as can be made using experimental knowledge of binding changes. It also indicates that much of variation in TF occupancy not predicted by accessibility or sequence may not be critical for fitness related biological output. At the same time, this ability to predict activity changes differs from one expression class to another and there is substantial room for improvement.

We also repeated the above analysis using imputed ChIP score profiles in *D. virilis* from

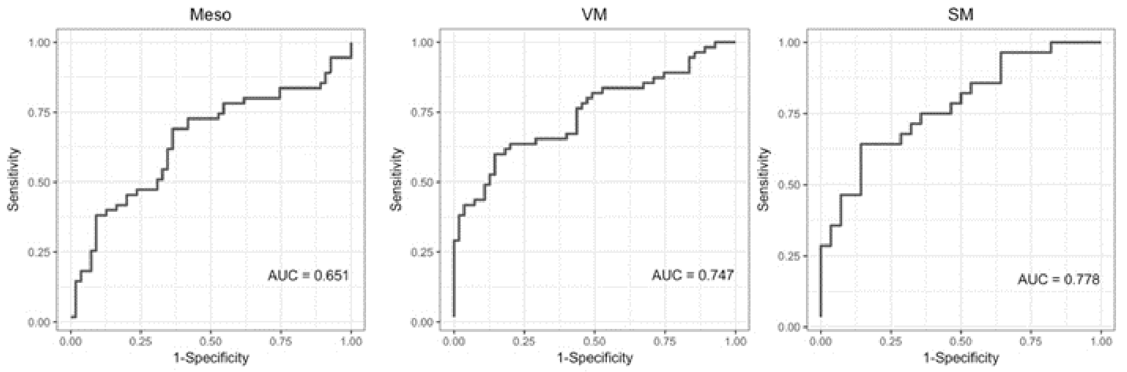

Figure 4.12: **ROC plots for delineating how well $\widehat{\mathbf{\Delta A_C}}$ values can classify high versus low $\mathbf{\Delta A_C}$ enhancer pairs in each activity class.** $\widehat{\Delta \mathrm{A_C}}$ values are based on *D. virilis* ChIP score profiles imputed from *D. melanogaster* scores and predictions of binding change.

$\Delta$ChIP predictions based only on sequence-level changes or only on accessibility changes (Figures 4.6A, 4.6B), rather than both. Our main observation is that sequence-based predictions of binding change are often close to and in some cases even better than the joint predictors that utilize sequence and accessibility changes (Table 4.6, rows '$\Delta$STAP' compared to rows '$\Delta$STAP and $\Delta$Acc'). A noteworthy data point is that for the 'SM' class, sequence-based predictions of $\Delta$ChIP can accurately predict, with an AUROC of 0.82, the enhancer pairs with greatest and least activity change, where activity is defined based on real ChIP profiles in the two species. We also noted that $\Delta$ChIP predictions based on accessibility changes alone are consistently worse in terms of the resulting agreement between $\Delta \mathrm{A_C}$ and $\widehat{\Delta \mathrm{A_C}}$. The is in contrast to the observations in Figure 4.6C, where we did not observe a consistent difference between sequence-based and accessibility-based predictors of binding change for individual TF:TP pairs. This is not surprising: accessibility changes are indeed an important statistical determinant of binding changes, but predicting activity change likely requires correctly predicting binding changes of multiple TFs, and the sequence-based predictors have an advantage in this respect as they use different motifs for each TF, while the accessibility-based predictors utilize the same underlying information in predicting binding change for every TF.

## 4.8 DISCUSSION

We examined the evolution of DNA accessibility in two distant species, and found it to be an important determinant or correlate of inter-species changes in TF binding. It is possible that changes in accessibility are not causal of binding change but rather a consequence; for

Table 4.6: **Changes in motif presence and accessibility can be used to predict enhancer activity change.** (A) Pearson correlation coefficients between two different estimates of activity change: $\Delta A_C$, based on measured *D. virilis* ChIP score profiles and $\widehat{\Delta A_C}$ based on *D. virilis* ChIP score profiles imputed from *D. melanogaster* scores and predictions of binding change ($\Delta$ChIP), which in turn were made from changes in sequence ('$\Delta$STAP'), accessibility ('$\Delta$Acc') or both. As a random control baseline, we used *D. virilis* ChIP scores imputed from *D. melanogaster* scores and a permuted version of the $\Delta$ChIP matrix. These analysis were performed for three expression domains: mesoderm (Meso), visceral muscle (VM), and somatic muscle (SM). (B) AUROC values representing how well $\widehat{\Delta A_C}$ values can classify high versus low $\Delta A_C$ enhancer pairs.

(A)

| ***D. virilis* ChIP imputation based on:** | **Meso** | **VM** | **SM** |
|---|---|---|---|
| **$\Delta$STAP** | 0.33 | 0.42 | 0.36 |
| **$\Delta$Acc** | 0.21 | 0.30 | 0.33 |
| **$\Delta$STAP and $\Delta$Acc** | 0.30 | 0.47 | 0.37 |
| **Random control** | 0.07 | 0.16 | 0.09 |
| **#Samples** | 110 | 110 | 56 |

(B)

| ***D. virilis* ChIP imputation based on:** | **Meso** | **VM** | **SM** |
|---|---|---|---|
| **$\Delta$STAP** | 0.65 | 0.70 | 0.82 |
| **$\Delta$Acc** | 0.57 | 0.68 | 0.70 |
| **$\Delta$STAP and $\Delta$Acc** | 0.65 | 0.75 | 0.78 |
| **Random control** | 0.56 | 0.59 | 0.57 |
| **#Samples** | 110 | 110 | 56 |

instance, the relaxation of selection pressure resulting from a functional loss of TF binding may in turn lead to reduction in local accessibility, which may be the case for Twist. Interestingly, we noted that our ability to predict TF binding changes simply based on accessibility changes rivals our ability to make those predictions based on sequence divergence, i.e., change of TF motif presence. At the same time, there is substantial complementarity between the two, and a model that combines both motif and accessibility changes can predict changes in TF binding more accurately than either alone. A noteworthy feature of our work is that we have approached issues of *cis*-regulatory divergence in a quantitative manner, asking 'to what extent' a relationship (e.g., between accessibility and binding changes) is supported by data, in addition to asking if 'there exists strong evidence' for such a relationship, through hypothesis testing approaches. Such a quantitative approach is also important for comparing how well two different types of information – changes in accessibility and motif presence – correlate with binding change.

It is worth emphasizing that our comparisons of sequence, accessibility and TF binding between species have the advantage of being performed in the context of a system where the examined TFs are all essential regulators that participate in a highly interconnected regulatory network, participating in feed-back and feed-forward regulation of a large number of genes. Thus, by focusing on putative enhancers defined by multiple ChIP peaks in close proximity, we hope to have enriched for evolutionary events with potential consequences for gene expression. Such an advantage is often not possible in other studies of binding evolution, since few regulatory networks have been as well characterized (see [73] for another example).

We also examined how TF binding changes, either experimentally measured or computationally predicted, relate to changes in enhancer spatio-temporal activity within the mesoderm specification network. Enhancers in this network have been previously shown [110] to be amenable to computational models that predict their activity (tissue specificity) from their TF binding profiles within one species (*D. melanogaster*). It was thus natural to ask if evolutionary changes in TF binding can be interpreted in the light of such functional models. However, we were unable to answer this question in the most direct way– whether binding changes for multiple TFs can, via these models, predict changes in enhancer activity – because the available data on regulatory activities of orthologous enhancers are sparse. Instead, we used the ability to model enhancer activity from ChIP data to show that predicted changes in binding (based on accessibility and motif divergence) agree with measured binding changes (ChIP data) in terms of what they imply about activity changes. It is worth clarifying that we defined activity change between orthologous enhancers as the difference in predicted activity in a spatio-temporal class, using a computational model that is meant to predict enhancer activities in *D. melanogaster*. Thus, under this definition, the activity of a *D. virilis* enhancer is in fact the expression pattern we predict it to drive if it was tested through a reporter assay in a *D. melanogaster* embryo. This was necessary since we do not yet have sufficient training data (*D. virilis* enhancers with known expression readouts in *D. virilis*) to learn a classifier for predicting activity in *D. virilis.* It was also a convenient choice since we did not have to make assumptions about conservation of the trans context between *D. melanogaster* and *D. virilis.* The only assumption required was that a "ChIP profile" (14 TF-ChIP values at a enhancer) obtained from *D. virilis* is semantically comparable to a ChIP profile obtained from *D. melanogaster*, which we previously showed is the case [74]. Also, the absolute values in the ChIP profile do not matter (only their relative values), since we worked with normalized ChIP profiles, which have similar distributions in both species (Figure 4.13).

There is precedence in the literature for examining activity changes between orthologous enhancers in a common cellular context [90]. Expression is often conserved despite divergence at sequence level, but that the data is sparse still. We expect that more experimental data on heterologous activity, e.g., of *D. virilis* enhancers, will better address the functional consequences of binding changes and improve our ability to predict functional *cis*-regulatory change from accessibility and sequence data.

In ending, we note that even when using a combined model that integrates sequence and accessibility data, we were able to predict TF binding change with a correlation coefficient of ∼0.5 at best. What is missing in the data and models that might account for the missing predictability? The answer is probably closely tied to the same issue in the context of single-species TF binding prediction, a topic that has received far greater attention [161], and where a number of additional factors, such as co-binding and competitive binding [33, 49], more precise motif characterizations [37, 162], higher resolution mapping of chromatin context [33, 64, 38], etc. have been shown to improve predictive ability. Incorporation of these additional dimensions of data and modeling in the future should further increase our understanding of evolutionary changes in transcription factor binding.

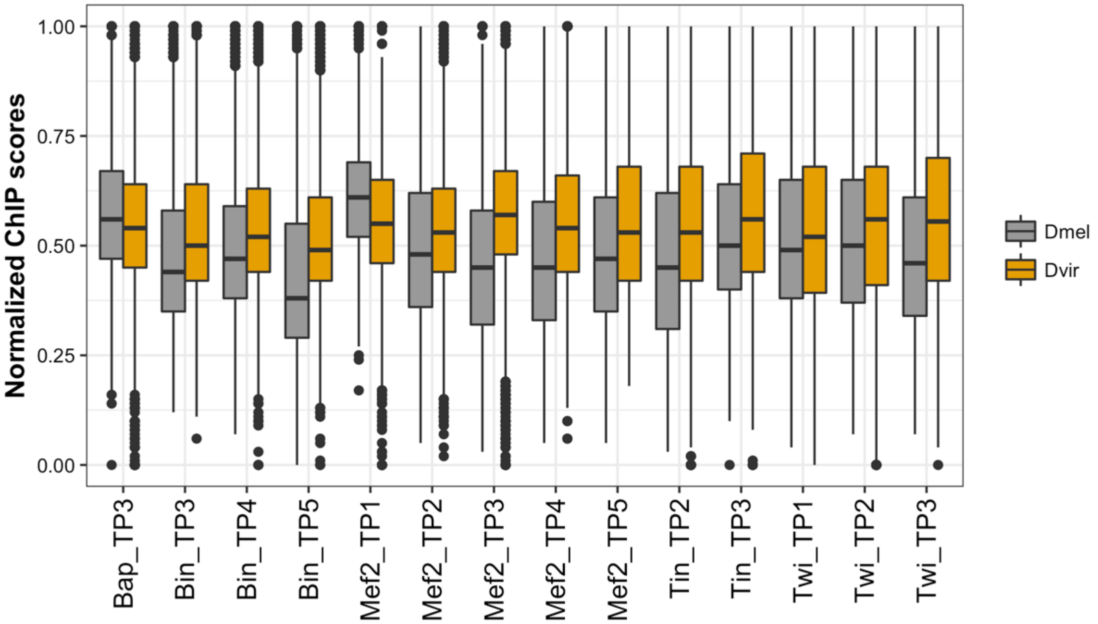

Figure 4.13: **Normalized ChIP scores in *D. melanogaster* and *D. virilis* show similar distributions for all 14 TF-time point conditions.**

# CHAPTER 5: CONCLUSION

Unraveling gene regulation has been a main research focus in quantitative biology. Sequence-to-expression models have been shown to have the expressive power to capture the complex relationship between regulatory sequence, transcription factor binding, and precise gene expression patterns. Their appeal lies in achieving this interpretability within a biophysically motivated framework, while making simplifications that hide mechanistic details on which little data is available. In this thesis, we have shown how thermodynamic models may leverage accessibility and DNA shape information to explain the data with higher accuracy. Moreover, we quantify the evolutionary relationships between enhancer, TF binding, and enhancer activity.

In Chapter 2, we extend a thermodynamics-based model of *cis*-regulatory function to assess the impact of chromatin accessibility on gene expression levels. While recent work in the field has reported a strong relationship between chromatin state (including accessibility) and expression of nearby genes, we focus our attention to the *quantitative* aspect of this relationship: do the quantitative variations in the accessibility levels *within* a *cis*-regulatory sequence such as an enhancer have a quantitative impact on the expression level driven by that sequence? We find the answer to be in the affirmative, and present multiple analyses to support this claim. The literature provides ample evidence that whole-genome accessibility profiles can effectively identify active enhancers, and emerging evidence that higher resolution accessibility data can even identify precise footprints of the DNA-bound transcription factors. We show, for the first time, that accessibility data can significantly improve quantitative predictions of gene expression levels. In doing so, we pose a new challenge for future research in biophysical modeling, which is to incorporate sequence-level determinants of chromatin accessibility, such as pioneer factor binding sites and nucleosome-binding preferences, into quantitative models of gene expression.

We anticipate that this work will nicely complement the intense community-wide activity on measuring chromatin states and correlating the dynamics of these states with cell type-specific gene expression. In particular, it should lead to chromatin states, e.g., accessibility, being utilized in a rigorous mechanistic modeling framework, thereby highlighting their more subtle effects on cellular biology.

In Chapter 3, we assess and demonstrate the impact of DNA shape features on gene expression levels by utilizing a thermodynamics-based model of *cis*-regulatory function. There has been considerable interest over the last few years around the notion that DNA shape features can noticeably improve the prediction of TF-DNA binding. This is part of a broader movement to go beyond the ‘position weight matrix’ (PWM) model of TF-DNA interaction that has ruled for the last 25 years. High throughput data sets, such as from SELEX and PBM technologies, have been generated, and several sophisticated models have been proposed to show that the binding strength of a site can indeed be computationally predicted accurately. DNA shape-based models have been one of the highlights of this on-going movement, whose ultimate goal, of course, is to ‘crack the *cis*-regulatory code’, i.e., to ‘read’ non-coding sequences and predict gene expression. The current literature, despite its emphasis on models of TF-DNA binding, does not provide any evidence at all that these advanced models really help predict gene expression. Our work is the first to bridge this crucial gap.

We focus our attention on the relationship between gene expression and TF-DNA binding: can models using DNA shape perform as well as (or better than) PWM-based models in predicting gene expression directly from sequence? (The ‘directly from sequence’ part is crucial, because the alternative – predicting expression from ChIP (TF binding and epigenomic mark) profiles – is a qualitatively distinct problem, and less useful when interpreting population-level and inter-species differences in regulatory sequences.) We find the answer to be affirmative, and present multiple analyses to support this central claim. The literature provides evidence that models using DNA shape features are highly promising for the goal of TF-DNA recognition, both *in vitro* (e.g., PBM data) and *in vivo* (e.g., ChIP-seq data). We show, for the first time, that a DNA shape-based model is at least as good as and arguably (statistically significantly) better than the PWM-based models in quantitative predictions of gene expression levels. This need not have been the case, since (a) the extent of improvement in prediction of TF-DNA binding may not have been large enough to lead to benefits in predicting expression levels, and (b) the model and system used in testing sequence-to-expression models might not have been sensitive enough to exploit the benefits of a better TF-DNA binding model. Thus, we consider our finding to be an affirmative conclusion to an investigation that could have gone either way.

Our positive conclusion is supported by (a) direct fits of the two models (shaped-based and PWM-based) to the same data sets, (b) cross-validation tests that ensure that the improvement is not due to the one additional parameter present in the shape-based model, and (c) innovative and carefully constructed tests involving perturbations of model inputs, to make sure that the improvement is not accidental. Our model uses a Random Forest classifier to integrate multiple shape features into a single measure of binding site quality and converts this into a pseudo-energy term to be used in established thermodyamics-based models of enhancer function. We anticipate that our work will nicely complement the burgeoning body of work on DNA shape predictions, further energizing that community. Since training of DNA-shape models require high throughput data on TF-DNA binding (e.g., PBM or

SELEX), our work should spur more widespread generation of such data, since it connects those data to the ultimate goal of regulatory genomics, that of predicting gene expression. In the long run, it should lead to DNA shape being utilized in a rigorous mechanistic modeling framework, thereby highlighting their more subtle effects on cellular biology.

In both Chapter 2 and 3, we chose to substantiate our general claim on a specific data set that represents one of the best understood regulatory networks in terms of *cis*-regulatory mechanisms. This is a set of 37 enhancers involved in patterning of the anterior-posterior (A/P) axis of the early *Drosophila* embryo; the data includes the enhancer sequences and their quantitative expression readouts along the A/P axis. Previous work has shown remarkable success in applying equilibrium thermodynamics based models of transcription factor action to this data set, making it a strong baseline against which to examine the effect of using chromatin accessibility and DNA shape features of binding sites. We consider this an important aspect of our work since it avoids the pitfalls of observing a spurious effect in a less-understood regulatory system. In particular, this system is much better suited for studied of *cis*-regulatory code than, for instance, regulatory systems active in human cell lines: the reason being that nearly two decades of rigorous, directed experimental work has established all the relevant regulatory inputs for genes in this system, so that models of gene expression can focus on elucidating detailed mechanisms rather than struggle with identification of the relevant inputs (GRN construction).

We extend our focus from *cis*-regulation to *cis*-evolution in Chapter 4. Mutations in enhancer elements are a key driving force of evolution and disease. A number of studies have examined genome-wide divergence in TF binding across species, providing important information linking sequence motif turnover and DNA accessibility change to TF occupancy. However, there is currently a lack of information on what these changes mean to enhancer function. In Chapter 4, we focus our attention to the quantitative aspect of this relationship: do changes in accessibility and motif presence carry complementary information related to observed changes in TF occupancy? Do quantitative variations in the accessibility and motif presence levels pertain to enhancer-driven expression changes? Our work is the first to examine multiple aspects of *cis*-regulatory divergence in a quantitative and systematic manner.

We chose to substantiate our general claim on a highly connected mesoderm specification network. Such an advantage is often not possible in other studies of binding evolution, since few regulatory networks have been as well characterized. To initiate this study, we generated a unique dataset - interspecies chromatin accessibility across developmental time points in two distally related *Drosophila* species. We also collected interspecies ChIP-seq on five essential tissue-specific TFs in this regulatory system. This actually represents the first interspecies developmental time-course of chromatin accessibility to date and reveals *cis*-regulatory relationships at embryonic enhancers as development proceeds.

We used multi-variable models to integrate accessibility and motif information across thousands of orthologous developmental enhancers and observed three functional properties of TF activity at enhancers. First, using dynamic and combinatorial signatures of co-association in TF occupancy, we demonstrated accessibility changes explain a significant portion of the co-divergence of TF pairs. Second, our models clearly show that accessibility and motif presence bear complementary information that is useful in predicting TF binding divergence. Third, we found that motif and accessibility-based predictors of TF binding change can substitute for experimentally measured binding change, for the purpose of predicting divergence in gene expression.

This thesis demonstrates powerful new claims regarding already popular concepts and techniques. Models and approaches proposed in this thesis pave the way to build a practical tool to add contextual information to non-coding variant for prioritization and interpretation. It will also enable major advances in the genomics of human health, by providing accurate predictions of the effects of single nucleotide polymorphisms at the cellular level.

# REFERENCES

[1] E. Cannavò, P. Khoueiry, D. A. Garfield, P. Geeleher, T. Zichner, E. H. Gustafson, L. Ciglar, J. O. Korbel, and E. E. Furlong, "Shadow enhancers are pervasive features of developmental regulatory networks," *Current Biology*, vol. 26, no. 1, pp. 38–51, 2016.

[2] T. M. Keane, L. Goodstadt, P. Danecek, M. A. White, K. Wong, B. Yalcin, A. Heger, A. Agam, G. Slater, and M. Goodson, "Mouse genomic variation and its effect on phenotypes and gene regulation," *Nature*, vol. 477, no. 7364, p. 289, 2011.

[3] S. B. Carroll, J. K. Grenier, and S. D. Weatherbee, *From DNA to diversity: molecular genetics and the evolution of animal design*. Hoboken, NJ: John Wiley & Sons, 2013.

[4] K. Chen and N. Rajewsky, "The evolution of gene regulation by transcription factors and micrornas," *Nature Reviews Genetics*, vol. 8, no. 2, p. 93, 2007.

[5] M. T. Maurano, R. Humbert, E. Rynes, R. E. Thurman, E. Haugen, H. Wang, A. P. Reynolds, R. Sandstrom, H. Qu, and J. Brody, "Systematic localization of common disease-associated variation in regulatory dna," *Science*, p. 1222794, 2012.

[6] E. H. Davidson, *The regulatory genome: gene regulatory networks in development and evolution*. Cambridge, MA: Academic Press, 2010.

[7] P. V. Kharchenko, A. A. Alekseyenko, Y. B. Schwartz, A. Minoda, N. C. Riddle, J. Ernst, P. J. Sabo, E. Larschan, A. A. Gorchakov, and T. Gu, "Comprehensive analysis of the chromatin landscape in drosophila melanogaster," *Nature*, vol. 471, no. 7339, p. 480, 2011.

[8] R. E. Thurman, E. Rynes, R. Humbert, J. Vierstra, M. T. Maurano, E. Haugen, N. C. Sheffield, A. B. Stergachis, H. Wang, and B. Vernot, "The accessible chromatin landscape of the human genome," *Nature*, vol. 489, no. 7414, p. 75, 2012.

[9] S. S. Gisselbrecht, L. A. Barrera, M. Porsch, A. Aboukhalil, P. W. Estep III, A. Vedenko, A. Palagi, Y. Kim, X. Zhu, and B. W. Busser, "Highly parallel assays of tissue-specific enhancers in whole drosophila embryos," *Nature methods*, vol. 10, no. 8, p. 774, 2013.

[10] A. Melnikov, A. Murugan, X. Zhang, T. Tesileanu, L. Wang, P. Rogov, S. Feizi, A. Gnirke, C. G. Callan Jr, and J. B. Kinney, "Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay," *Nature biotechnology*, vol. 30, no. 3, p. 271, 2012.

[11] D. Shlyueva, G. Stampfel, and A. Stark, "Transcriptional enhancers: from properties to genome-wide predictions," *Nature Reviews Genetics*, vol. 15, no. 4, pp. 272–286, 2014.

[12] M. Levo and E. Segal, "In pursuit of design principles of regulatory sequences," *Nature reviews. Genetics*, vol. 15, no. 7, p. 453, 2014.

[13] E. Segal and J. Widom, "From dna sequence to transcriptional behaviour: a quantitative approach," *Nature Reviews Genetics*, vol. 10, no. 7, pp. 443–456, 2009.

[14] S. Weingarten-Gabbay and E. Segal, "The grammar of transcriptional regulation," *Human genetics*, vol. 133, no. 6, pp. 701–711, 2014.

[15] L. Bintu, N. E. Buchler, H. G. Garcia, U. Gerland, T. Hwa, J. Kondev, and R. Phillips, "Transcriptional regulation by the numbers: models," *Current opinion in genetics & development*, vol. 15, no. 2, pp. 116–124, 2005.

[16] N. E. Buchler, U. Gerland, and T. Hwa, "On schemes of combinatorial transcription logic," *Proceedings of the National Academy of Sciences*, vol. 100, no. 9, pp. 5136–5141, 2003.

[17] W. D. Fakhouri, A. Ay, R. Sayal, J. Dresch, E. Dayringer, and D. N. Arnosti, "Deciphering a transcriptional regulatory code: modeling short-range repression in the drosophila embryo," *Molecular systems biology*, vol. 6, no. 1, pp. 341–354, 2010.

[18] J. Gertz, E. D. Siggia, and B. A. Cohen, "Analysis of combinatorial cis-regulation in synthetic and genomic promoters," *Nature*, vol. 457, no. 7226, pp. 215–218, 2009.

[19] X. He, M. A. H. Samee, C. Blatti, and S. Sinha, "Thermodynamics-based models of transcriptional regulation by enhancers: the roles of synergistic activation, cooperative binding and short-range repression," *PLoS Comput Biol*, vol. 6, no. 9, p. e1000935, 2010.

[20] H. Janssens, S. Hou, J. Jaeger, A.-R. Kim, E. Myasnikova, D. Sharp, and J. Reinitz, "Quantitative and predictive model of transcriptional control of the drosophila melanogaster even skipped gene," *Nature genetics*, vol. 38, no. 10, pp. 1159–1165, 2006.

[21] D. Papatsenko and M. S. Levine, "Dual regulation by the hunchback gradient in the drosophila embryo," *Proceedings of the National Academy of Sciences*, vol. 105, no. 8, pp. 2901–2906, 2008.

[22] E. Segal, T. Raveh-Sadka, M. Schroeder, U. Unnerstall, and U. Gaul, "Predicting expression patterns from regulatory sequence in drosophila segmentation," *Nature*, vol. 451, no. 7178, pp. 535–540, 2008.

[23] M. A. Shea and G. K. Ackers, "The or control system of bacteriophage lambda: A physical-chemical model for gene regulation," *Journal of molecular biology*, vol. 181, no. 2, pp. 211–230, 1985.

[24] M. A. H. Samee and S. Sinha, "Evaluating thermodynamic models of enhancer activity on cellular resolution gene expression data," *Methods*, vol. 62, no. 1, pp. 79–90, 2013.

[25] T. Kaplan, X.-Y. Li, P. J. Sabo, S. Thomas, J. A. Stamatoyannopoulos, M. D. Biggin, and M. B. Eisen, "Quantitative models of the mechanisms that control genome-wide patterns of transcription factor binding during early drosophila development," *PLoS Genet*, vol. 7, no. 2, p. e1001290, 2011.

[26] X.-Y. Li, S. Thomas, P. J. Sabo, M. B. Eisen, J. A. Stamatoyannopoulos, and M. D. Biggin, "The role of chromatin accessibility in directing the widespread, overlapping patterns of drosophila transcription factor binding," *Genome biology*, vol. 12, no. 4, p. R34, 2011.

[27] A. Natarajan, G. G. Yardımcı, N. C. Sheffield, G. E. Crawford, and U. Ohler, "Predicting cell-type–specific gene expression from regions of open chromatin," *Genome research*, vol. 22, no. 9, pp. 1711–1722, 2012.

[28] A. P. Boyle, S. Davis, H. P. Shulha, P. Meltzer, E. H. Margulies, Z. Weng, T. S. Furey, and G. E. Crawford, "High-resolution mapping and characterization of open chromatin across the genome," *Cell*, vol. 132, no. 2, pp. 311–322, 2008.

[29] J. R. Hesselberth, X. Chen, Z. Zhang, P. J. Sabo, R. Sandstrom, A. P. Reynolds, R. E. Thurman, S. Neph, M. S. Kuehn, and W. S. Noble, "Global mapping of protein-dna interactions in vivo by digital genomic footprinting," *Nature methods*, vol. 6, no. 4, pp. 283–289, 2009.

[30] P. J. Sabo, M. Hawrylycz, J. C. Wallace, R. Humbert, M. Yu, A. Shafer, J. Kawamoto, R. Hall, J. Mack, and M. O. Dorschner, "Discovery of functional noncoding elements by digital analysis of chromatin structure," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 48, pp. 16 837–16 842, 2004.

[31] M. Sekimata, M. Pérez-Melgosa, S. A. Miller, A. S. Weinmann, P. J. Sabo, R. Sandstrom, M. O. Dorschner, J. A. Stamatoyannopoulos, and C. B. Wilson, "Ccctc-binding factor and the transcription factor t-bet orchestrate t helper 1 cell-specific structure and function at the interferon-$\gamma$ locus," *Immunity*, vol. 31, no. 4, pp. 551–564, 2009.

[32] S. Thomas, X.-Y. Li, P. J. Sabo, R. Sandstrom, R. E. Thurman, T. K. Canfield, E. Giste, W. Fisher, A. Hammonds, and S. E. Celniker, "Dynamic reprogramming of chromatin accessibility during drosophila embryo development," *Genome biology*, vol. 12, no. 5, p. R43, 2011.

[33] Q. Cheng, M. Kazemian, H. Pham, C. Blatti, S. E. Celniker, S. A. Wolfe, M. H. Brodsky, and S. Sinha, "Computational identification of diverse mechanisms underlying transcription factor-dna occupancy," *PLoS Genet*, vol. 9, no. 8, p. e1003571, 2013.

[34] E. Calo and J. Wysocka, "Modification of enhancer chromatin: what, how, and why?" *Molecular cell*, vol. 49, no. 5, pp. 825–837, 2013.

[35] R. Karlić, H.-R. Chung, J. Lasserre, K. Vlahoviček, and M. Vingron, "Histone modification levels are predictive for gene expression," *Proceedings of the National Academy of Sciences*, vol. 107, no. 7, pp. 2926–2931, 2010.

[36] A. Arvey, P. Agius, W. S. Noble, and C. Leslie, “Sequence and chromatin determinants of cell-type–specific transcription factor binding,” *Genome research*, vol. 22, no. 9, pp. 1723–1734, 2012.

[37] X. He, C.-C. Chen, F. Hong, F. Fang, S. Sinha, H.-H. Ng, and S. Zhong, “A biophysical model for analysis of transcription factor interaction and binding site arrangement from genome-wide binding data,” *PloS one*, vol. 4, no. 12, p. e8155, 2009.

[38] R. Pique-Regi, J. F. Degner, A. A. Pai, D. J. Gaffney, Y. Gilad, and J. K. Pritchard, “Accurate inference of transcription factor binding from dna sequence and chromatin accessibility data,” *Genome research*, vol. 21, no. 3, pp. 447–455, 2011.

[39] S. Thomas, X.-Y. Li, P. J. Sabo, R. Sandstrom, R. E. Thurman, T. K. Canfield, E. Giste, W. Fisher, A. Hammonds, and S. E. Celniker, “Dynamic reprogramming of chromatin accessibility during drosophila embryo development,” *Genome Biol*, vol. 12, no. 5, p. R43, 2011.

[40] M. Kazemian, C. Blatti, A. Richards, M. McCutchan, N. Wakabayashi-Ito, A. Hammonds, S. Celniker, S. Kumar, S. Wolfe, M. Brodsky, and S. Sinha, “Quantitative analysis of the drosophila segmentation regulatory network using pattern generating potentials,” *PLoS biology*, vol. 8, no. 8, p. 1961, 2010.

[41] R. P. Zinzen, K. Senger, M. Levine, and D. Papatsenko, “Computational models for neurogenic gene expression in the drosophila embryo,” *Current Biology*, vol. 16, no. 13, pp. 1358–1365, 2006.

[42] M. A. H. Samee and S. Sinha, “Quantitative modeling of a gene’s expression from its intergenic sequence,” *PLoS computational biology*, vol. 10, no. 3, p. e1003467, 2014.

[43] A.-R. Kim, C. Martinez, J. Ionides, A. F. Ramos, M. Z. Ludwig, N. Ogawa, D. H. Sharp, and J. Reinitz, “Rearrangements of 2.5 kilobases of noncoding dna from the drosophila even-skipped locus define predictive rules of genomic cis-regulatory logic,” *PLoS genetics*, vol. 9, no. 2, p. e1003243, 2013.

[44] R. P. Zinzen and D. Papatsenko, “Enhancer responses to similarly distributed antagonistic gradients in development,” *PLoS Comput Biol*, vol. 3, no. 5, p. e84, 2007.

[45] M. Slattery, T. Zhou, L. Yang, A. C. Dantas Machado, R. Gordân, and R. Rohs, “Absence of a simple code: how transcription factors read the genome,” *Trends Biochem Sci*, vol. 39, no. 9, pp. 381–99, 2014.

[46] R. Rohs, X. Jin, S. M. West, R. Joshi, B. Honig, and R. S. Mann, “Origins of specificity in protein-dna recognition,” *Annual review of biochemistry*, vol. 79, pp. 233–269, 2010.

[47] T. Siggers and R. Gordân, “Protein–dna binding: complexities and multi-protein codes,” *Nucleic acids research*, vol. 42, pp. 2099–2111, 2013.

[48] M. D. Biggin, “Animal transcription networks as highly connected, quantitative continua,” *Developmental cell*, vol. 21, no. 4, pp. 611–626, 2011.

[49] T. Wasson and A. J. Hartemink, "An ensemble model of competitive multi-factor binding of the genome," *Genome research*, vol. 19, no. 11, pp. 2101–2112, 2009.

[50] G. D. Stormo and Y. Zhao, "Determining the specificity of protein–dna interactions," *Nature Reviews Genetics*, vol. 11, no. 11, pp. 751–760, 2010.

[51] M. Weirauch, A. Cote, R. Norel, M. Annala, Y. Zhao, T. Riley, J. Saez-Rodriguez, T. Cokelaer, A. Vedenko, S. Talukder, H. Bussemaker, Q. Morris, M. Bulyk, G. Stolovitzky, and T. Hughes, "Evaluation of methods for modeling transcription factor sequence specificity," *Nature Biotechnology*, vol. 31, no. 2, pp. 126–134, 2013.

[52] O. G. Berg and P. H. von Hippel, "Selection of dna binding sites by regulatory proteins," *Trends in biochemical sciences*, vol. 13, no. 6, pp. 207–211, 1988.

[53] G. D. Stormo, T. D. Schneider, L. Gold, and A. Ehrenfeucht, "Use of the 'perceptron'algorithm to distinguish translational initiation sites in e. coli," *Nucleic Acids Research*, vol. 10, no. 9, pp. 2997–3011, 1982.

[54] T. L. Bailey and C. Elkan, "Fitting a mixture model by expectation maximization to discover motifs in bipolymers," in *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology*, vol. 2, Conference Proceedings, pp. 28–36.

[55] M. Djordjevic, A. M. Sengupta, and B. I. Shraiman, "A biophysical approach to transcription factor binding site discovery," *Genome research*, vol. 13, no. 11, pp. 2381–2390, 2003.

[56] Y. Orenstein and R. Shamir, "A comparative analysis of transcription factor binding models learned from pbm, ht-selex and chip data," *Nucleic acids research*, vol. 42, no. 8, p. e63, 2014.

[57] G. D. Stormo, "Dna binding sites: representation and discovery," *Bioinformatics*, vol. 16, no. 1, pp. 16–23, 2000.

[58] G. D. Stormo, "Modeling the specificity of protein-dna interactions," *Quantitative Biology*, vol. 1, no. 2, pp. 115–130, 2013.

[59] M. Santolini, T. Mora, and V. Hakim, "A general pairwise interaction model provides an accurate description of in vivo transcription factor binding sites," *PLoS ONE*, vol. 9, no. 6, p. e99015, 2014.

[60] J. L. Stringham, A. S. Brown, R. A. Drewell, and J. M. Dresch, "Flanking sequence context-dependent transcription factor binding in early drosophila development," *BMC bioinformatics*, vol. 14, no. 1, pp. 298–310, 2013.

[61] Y. Zhao, S. Ruan, M. Pandey, and G. D. Stormo, "Improved models for transcription factor binding site identification using nonindependent interactions," *Genetics*, vol. 191, no. 3, pp. 781–790, 2012.

[62] P. J. Farnham, “Insights from genomic profiling of transcription factors,” *Nature Reviews Genetics*, vol. 10, no. 9, pp. 605–616, 2009.

[63] T. Duque, M. A. H. Samee, M. Kazemian, H. N. Pham, M. H. Brodsky, and S. Sinha, “Simulations of enhancer evolution provide mechanistic insights into gene regulation,” *Molecular biology and evolution*, vol. 31, no. 1, pp. 184–200, 2014.

[64] P.-C. Peng, M. A. H. Samee, and S. Sinha, “Incorporating chromatin accessibility data into sequence-to-expression modeling,” *Biophysical journal*, vol. 108, no. 5, pp. 1257–1267, 2015.

[65] C. Blatti, M. Kazemian, S. Wolfe, M. Brodsky, and S. Sinha, “Integrating motif, dna accessibility and gene expression data to build regulatory maps in an organism,” *Nucleic acids research*, vol. 43, no. 8, pp. 3998–4012, 2015.

[66] L. A. Pennacchio, G. G. Loots, M. A. Nobrega, and I. Ovcharenko, “Predicting tissue-specific enhancers in the human genome,” *Genome research*, vol. 17, no. 2, pp. 201–211, 2007.

[67] T. Zhou, L. Yang, Y. Lu, I. Dror, A. C. D. Machado, T. Ghane, R. Di Felice, and R. Rohs, “Dnashape: a method for the high-throughput prediction of dna structural features on a genomic scale,” *Nucleic acids research*, vol. 41, pp. W56–W62, 2013.

[68] M. B. Noyes, X. Meng, A. Wakabayashi, S. Sinha, M. H. Brodsky, and S. A. Wolfe, “A systematic characterization of factors that regulate drosophila segmentation via a bacterial one-hybrid system,” *Nucleic acids research*, vol. 36, no. 8, pp. 2547–2560, 2008.

[69] B. Prud’Homme, N. Gompel, A. Rokas, V. A. Kassner, T. M. Williams, S.-D. Yeh, J. R. True, and S. B. Carroll, “Repeated morphological evolution through cis-regulatory changes in a pleiotropic gene,” *Nature*, vol. 440, no. 7087, p. 1050, 2006.

[70] A. Siepel and L. Arbiza, “Cis-regulatory elements and human evolution,” *Current opinion in genetics & development*, vol. 29, pp. 81–89, 2014.

[71] M. C. Saul, C. H. Seward, J. M. Troy, H. Zhang, L. G. Sloofman, X. Lu, P. A. Weisner, D. Caetano-Anolles, H. Sun, and S. D. Zhao, “Transcriptional regulatory dynamics drive coordinated metabolic and neural response to social challenge in mice,” *Genome research*, vol. 27, no. 6, pp. 959–972, 2017.

[72] G. A. Wray, “The evolutionary significance of cis-regulatory mutations,” *Nature Reviews Genetics*, vol. 8, no. 3, p. 206, 2007.

[73] M. Paris, T. Kaplan, X. Y. Li, J. E. Villalta, S. E. Lott, and M. B. Eisen, “Extensive divergence of transcription factor binding in drosophila embryos with highly conserved gene expression,” *PLoS Genet*, vol. 9, no. 9, p. e1003748, 2013.

[74] P. Khoueiry, C. Girardot, L. Ciglar, P.-C. Peng, E. H. Gustafson, S. Sinha, and E. E. Furlong, "Uncoupling evolutionary changes in dna sequence, transcription factor occupancy and enhancer activity," *eLife*, vol. 6, 2017.

[75] R. K. Bradley, X.-Y. Li, C. Trapnell, S. Davidson, L. Pachter, H. C. Chu, L. A. Tonkin, M. D. Biggin, and M. B. Eisen, "Binding site turnover produces pervasive quantitative changes in transcription factor binding between closely related drosophila species," 2010.

[76] A.-R. Carvunis, T. Wang, D. Skola, A. Yu, J. Chen, J. F. Kreisberg, and T. Ideker, "Evidence for a common evolutionary rate in metazoan transcriptional networks," *Elife*, vol. 4, 2015.

[77] K. Stefflova, D. Thybert, M. D. Wilson, I. Streeter, J. Aleksic, P. Karagianni, A. Brazma, D. J. Adams, I. Talianidis, and J. C. Marioni, "Cooperativity and rapid evolution of cobound transcription factors in closely related mammals," *Cell*, vol. 154, no. 3, pp. 530–540, 2013.

[78] E. S. Wong, D. Thybert, B. M. Schmitt, K. Stefflova, D. T. Odom, and P. Flicek, "Decoupling of evolutionary changes in transcription factor binding and gene expression in mammals," *Genome research*, vol. 25, no. 2, pp. 167–178, 2015.

[79] Y. Cheng, Z. Ma, B.-H. Kim, W. Wu, P. Cayting, A. P. Boyle, V. Sundaram, X. Xing, N. Dogan, and J. Li, "Principles of regulatory information conservation between mouse and human," *Nature*, vol. 515, no. 7527, p. 371, 2014.

[80] Q. He, A. F. Bardet, B. Patton, J. Purvis, J. Johnston, A. Paulson, M. Gogol, A. Stark, and J. Zeitlinger, "High conservation of transcription factor binding and evidence for combinatorial regulation across six drosophila species," *Nature genetics*, vol. 43, no. 5, pp. 414–420, 2011.

[81] A. M. Moses, D. A. Pollard, D. A. Nix, V. N. Iyer, X.-Y. Li, M. D. Biggin, and M. B. Eisen, "Large-scale turnover of functional transcription factor binding sites in drosophila," *PLoS Comput Biol*, vol. 2, no. 10, p. e130, 2006.

[82] M. Naval-Sánchez, D. Potier, G. Hulselmans, V. Christiaens, and S. Aerts, "Identification of lineage-specific cis-regulatory modules associated with variation in transcription factor binding and chromatin activity using ornstein–uhlenbeck models," *Molecular biology and evolution*, vol. 32, no. 9, pp. 2441–2455, 2015.

[83] C. M. Alexandre, J. R. Urton, K. Jean-Baptiste, J. Huddleston, M. W. Dorrity, J. T. Cuperus, A. M. Sullivan, F. Bemm, D. Jolic, and A. A. Arsovski, "Complex relationships between chromatin accessibility, sequence divergence, and gene expression in arabidopsis thaliana," *Mol. Biol. Evol*, vol. 35, no. 4, pp. 837–854, 2017.

[84] B. J. Lesch, S. J. Silber, J. R. McCarrey, and D. C. Page, "Parallel evolution of male germline epigenetic poising and somatic development in animals," *Nature genetics*, vol. 48, no. 8, p. 888, 2016.

[85] D. Garfield, R. Haygood, W. J. Nielsen, and G. A. Wray, "Population genetics of cis-regulatory sequences that operate during embryonic development in the sea urchin strongylocentrotus purpuratus," *Evolution & development*, vol. 14, no. 2, pp. 152–167, 2012.

[86] P. J. Wittkopp and G. Kalay, "Cis-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence," *Nature Reviews Genetics*, vol. 13, no. 1, p. 59, 2012.

[87] S. W. Doniger and J. C. Fay, "Frequent gain and loss of functional transcription factor binding sites," *PLoS computational biology*, vol. 3, no. 5, p. e99, 2007.

[88] E. Emberly, N. Rajewsky, and E. D. Siggia, "Conservation of regulatory elements between two species of drosophila," *BMC bioinformatics*, vol. 4, no. 1, p. 57, 2003.

[89] K. D. Yokoyama, Y. Zhang, and J. Ma, "Tracing the evolution of lineage-specific transcription factor binding sites in a birth-death framework," *PLoS computational biology*, vol. 10, no. 8, p. e1003771, 2014.

[90] C. D. Arnold, D. Gerlach, D. Spies, J. A. Matts, Y. A. Sytnikova, M. Pagani, N. C. Lau, and A. Stark, "Quantitative genome-wide enhancer activity maps for five drosophila species show functional enhancer conservation and turnover during cis-regulatory evolution," *Nature genetics*, vol. 46, no. 7, p. 685, 2014.

[91] T. Duque, M. A. H. Samee, M. Kazemian, H. N. Pham, M. H. Brodsky, and S. Sinha, "Simulations of enhancer evolution provide mechanistic insights into gene regulation," *Molecular biology and evolution*, vol. 31, no. 1, pp. 184–200, 2013.

[92] M. Z. Ludwig, C. Bergman, N. H. Patel, and M. Kreitman, "Evidence for stabilizing selection in a eukaryotic enhancer element," *Nature*, vol. 403, no. 6769, p. 564, 2000.

[93] S. Yang, N. Oksenberg, S. Takayama, S.-J. Heo, A. Poliakov, N. Ahituv, I. Dubchak, and D. Boffelli, "Functionally conserved enhancers with divergent sequences in distant vertebrates," *BMC genomics*, vol. 16, no. 1, p. 882, 2015.

[94] E. E. Hare, B. K. Peterson, V. N. Iyer, R. Meier, and M. B. Eisen, "Sepsid even-skipped enhancers are functionally conserved in drosophila despite lack of sequence conservation," 2008.

[95] M. Kazemian, K. Suryamohan, J.-Y. Chen, Y. Zhang, M. A. H. Samee, M. S. Halfon, and S. Sinha, "Evidence for deep regulatory similarities in early developmental programs across highly diverged insects," *Genome biology and evolution*, vol. 6, no. 9, pp. 2301–2320, 2014.

[96] C. F. Connelly, J. Wakefield, and J. M. Akey, "Evolution and genetic architecture of chromatin accessibility and function in yeast," *PLoS genetics*, vol. 10, no. 7, p. e1004427, 2014.

[97] J. Vierstra, E. Rynes, R. Sandstrom, M. Zhang, T. Canfield, R. S. Hansen, S. Stehling-Sun, P. J. Sabo, R. Byron, and R. Humbert, "Mouse regulatory dna landscapes reveal global principles of cis-regulatory evolution," *Science*, vol. 346, no. 6212, pp. 1007–1012, 2014.

[98] M. J. Guertin and J. T. Lis, "Mechanisms by which transcription factors gain access to target sequence elements in chromatin," *Current opinion in genetics & development*, vol. 23, no. 2, pp. 116–123, 2013.

[99] X.-y. Li, S. MacArthur, R. Bourgon, D. Nix, D. A. Pollard, V. N. Iyer, A. Hechmer, L. Simirenko, M. Stapleton, and C. L. L. Hendriks, "Transcription factors bind thousands of active and inactive regions in the drosophila blastoderm," *PLoS biology*, vol. 6, no. 2, p. e27, 2008.

[100] D. A. Cusanovich, B. Pavlovic, J. K. Pritchard, and Y. Gilad, "The functional consequences of variation in transcription factor binding," *PLoS genetics*, vol. 10, no. 3, p. e1004226, 2014.

[101] N. Azpiazu and M. Frasch, "tinman and bagpipe: two homeo box genes that determine cell fates in the dorsal mesoderm of drosophila," *Genes & Development*, vol. 7, no. 7b, pp. 1325–1340, 1993.

[102] M. K. Baylies and M. Bate, "twist: a myogenic switch in drosophila," *Science*, vol. 272, no. 5267, pp. 1481–1484, 1996.

[103] J. S. Jakobsen, M. Braun, J. Astorga, E. H. Gustafson, T. Sandmann, M. Karzynski, P. Carlsson, and E. E. Furlong, "Temporal chip-on-chip reveals biniou as a universal regulator of the visceral muscle transcriptional network," *Genes & development*, vol. 21, no. 19, pp. 2448–2460, 2007.

[104] H. Jin, R. Stojnic, B. Adryan, A. Ozdemir, A. Stathopoulos, and M. Frasch, "Genome-wide screens for in vivo tinman binding sites identify cardiac enhancers with diverse functional architectures," *PLoS genetics*, vol. 9, no. 1, p. e1003195, 2013.

[105] T. Sandmann, C. Girardot, M. Brehme, W. Tongprasit, V. Stolc, and E. E. Furlong, "A core transcriptional network for early mesoderm development in drosophila melanogaster," *Genes & development*, vol. 21, no. 4, pp. 436–449, 2007.

[106] T. Sandmann, L. J. Jensen, J. S. Jakobsen, M. M. Karzynski, M. P. Eichenlaub, P. Bork, and E. E. Furlong, "A temporal map of transcription factor activity: mef2 directly regulates target genes at all stages of muscle development," *Developmental cell*, vol. 10, no. 6, pp. 797–807, 2006.

[107] Z. Yin and M. Frasch, "Regulation and function of tinman during dorsal mesoderm induction and heart specification in drosophila," *genesis*, vol. 22, no. 3, pp. 187–200, 1998.

[108] Z. Yin, X.-L. Xu, and M. Frasch, “Regulation of the twist target gene tinman by modular cis-regulatory elements during early mesoderm development,” *Development*, vol. 124, no. 24, pp. 4971–4982, 1997.

[109] S. Zaffran, A. Küchler, H.-H. Lee, and M. Frasch, “biniou (foxf), a central component in a regulatory network controlling visceral mesoderm development and midgut morphogenesis in drosophila,” *Genes & development*, vol. 15, no. 21, pp. 2900–2915, 2001.

[110] R. P. Zinzen, C. Girardot, J. Gagneur, M. Braun, and E. E. Furlong, “Combinatorial binding predicts spatio-temporal cis-regulatory activity,” *Nature*, vol. 462, no. 7269, pp. 65–70, 2009.

[111] A. Stark, M. F. Lin, P. Kheradpour, J. S. Pedersen, L. Parts, J. W. Carlson, M. A. Crosby, M. D. Rasmussen, S. Roy, and A. N. Deoras, “Discovery of functional elements in 12 drosophila genomes using evolutionary signatures,” *Nature*, vol. 450, no. 7167, p. 219, 2007.

[112] S. M. Gallo, L. Li, Z. Hu, and M. S. Halfon, “Redfly: a regulatory element database for drosophila,” *Bioinformatics*, vol. 22, no. 3, pp. 381–383, 2005.

[113] A. Pisarev, E. Poustelnikova, M. Samsonova, and J. Reinitz, “Flyex, the quantitative atlas on segmentation gene expression at cellular resolution,” *Nucleic acids research*, vol. 37, no. suppl_1, pp. D560–D566, 2008.

[114] E. Poustelnikova, A. Pisarev, M. Blagov, M. Samsonova, and J. Reinitz, “A database for management of gene expression data in situ,” *Bioinformatics*, vol. 20, no. 14, pp. 2212–2221, 2004.

[115] C. M. Bergman, J. W. Carlson, and S. E. Celniker, “Drosophila dnase i footprint database: a systematic genome annotation of transcription factor binding sites in the fruitfly, drosophila melanogaster,” *Bioinformatics*, vol. 21, no. 8, pp. 1747–1749, 2004.

[116] J. O. Yáñez-Cuna, E. Z. Kvon, and A. Stark, “Deciphering the transcriptional cis-regulatory code,” *Trends in Genetics*, vol. 29, no. 1, pp. 11–22, 2013.

[117] N. Kaplan, I. K. Moore, Y. Fondufe-Mittendorf, A. J. Gossett, D. Tillo, Y. Field, E. M. LeProust, T. R. Hughes, J. D. Lieb, and J. Widom, “The dna-encoded nucleosome organization of a eukaryotic genome,” *Nature*, vol. 458, no. 7236, p. 362, 2009.

[118] H. Liu, R. Zhang, W. Xiong, J. Guan, Z. Zhuang, and S. Zhou, “A comparative evaluation on prediction methods of nucleosome positioning,” *Briefings in bioinformatics*, vol. 15, no. 6, pp. 1014–1027, 2013.

[119] T. van der Heijden, J. J. van Vugt, C. Logie, and J. van Noort, “Sequence-based prediction of single nucleosome positioning and genome-wide nucleosome occupancy,” *Proceedings of the National Academy of Sciences*, vol. 109, no. 38, pp. E2514–E2522, 2012.

[120] T. T. Marstrand and J. D. Storey, "Identifying and mapping cell-type-specific chromatin programming of gene expression," *Proceedings of the National Academy of Sciences*, vol. 111, no. 6, pp. E645–E654, 2014.

[121] J. F. Degner, A. A. Pai, R. Pique-Regi, J.-B. Veyrieras, D. J. Gaffney, J. K. Pickrell, S. De Leon, K. Michelini, N. Lewellen, and G. E. Crawford, "Dnase i sensitivity qtls are a major determinant of human expression variation," *Nature*, vol. 482, no. 7385, p. 390, 2012.

[122] C. R. Clapier and B. R. Cairns, "The biology of chromatin remodeling complexes," *Annual review of biochemistry*, vol. 78, pp. 273–304, 2009.

[123] T. Chen and S. Y. Dent, "Chromatin modifiers and remodellers: regulators of cellular differentiation," *Nature Reviews Genetics*, vol. 15, no. 2, p. 93, 2014.

[124] M. M. Harrison, X.-Y. Li, T. Kaplan, M. R. Botchan, and M. B. Eisen, "Zelda binding in the early drosophila melanogaster embryo marks regions subsequently activated at the maternal-to-zygotic transition," *PLoS genetics*, vol. 7, no. 10, p. e1002266, 2011.

[125] C.-Y. Nien, H.-L. Liang, S. Butcher, Y. Sun, S. Fu, T. Gocha, N. Kirov, J. R. Manak, and C. Rushlow, "Temporal coordination of gene networks by zelda in the early drosophila embryo," *PLoS genetics*, vol. 7, no. 10, p. e1002339, 2011.

[126] E. Z. Kvon, G. Stampfel, J. O. Yáñez-Cuna, B. J. Dickson, and A. Stark, "Hot regions function as patterned developmental enhancers and have a distinct cis-regulatory signature," *Genes & development*, 2012.

[127] P.-C. Peng and S. Sinha, "Quantitative modeling of gene expression using dna shape features of binding sites," *Nucleic acids research*, vol. 44, no. 13, pp. e120–e120, 2016.

[128] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[129] A. Liaw and M. Wiener, "Classification and regression by randomforest," *R news*, vol. 2, no. 3, pp. 18–22, 2002.

[130] B. Hooghe, S. Broos, F. Van Roy, and P. De Bleser, "A flexible integrative approach based on random forest improves prediction of transcription factor binding sites," *Nucleic acids research*, vol. 40, no. 14, p. e106, 2012.

[131] M. Pujato, F. Kieken, A. A. Skiles, N. Tapinos, and A. Fiser, "Prediction of dna binding motifs from 3d models of transcription factors; identifying tlx3 regulated genes," *Nucleic acids research*, p. gku1228, 2014.

[132] A. Pisarev, E. Poustelnikova, M. Samsonova, and J. Reinitz, "Flyex, the quantitative atlas on segmentation gene expression at cellular resolution," *Nucleic acids research*, vol. 37, no. suppl 1, pp. D560–D566, 2009.

[133] B. Schmidt, *Bioinformatics: high performance parallel computer architectures.* Boca Raton, FL: CRC Press, 2010.

[134] T. Zhou, N. Shen, L. Yang, N. Abe, J. Horton, R. S. Mann, H. J. Bussemaker, R. Gordan, and R. Rohs, "Quantitative modeling of transcription factor binding specificities using dna shape," *Proc Natl Acad Sci U S A*, vol. 112, no. 15, pp. 4654–9, 2015.

[135] J. Yang and S. A. Ramsey, "A dna shape-based regulatory score improves position-weight matrix-based recognition of transcription factor binding sites," *Bioinformatics*, vol. 31, no. 21, pp. 3445–3450, 2015.

[136] E. Sharon, S. Lubliner, and E. Segal, "A feature-based approach to modeling protein–dna interactions," *PLoS Comput Biol*, vol. 4, no. 8, p. e1000154, 2008.

[137] R. Gordân, N. Shen, I. Dror, T. Zhou, J. Horton, R. Rohs, and M. L. Bulyk, "Genomic regions flanking e-box binding sites influence dna binding specificity of bhlh transcription factors through dna shape," *Cell reports*, vol. 3, no. 4, pp. 1093–1104, 2013.

[138] M. Levo, E. Zalckvar, E. Sharon, A. C. D. Machado, Y. Kalma, M. Lotam-Pompan, A. Weinberger, Z. Yakhini, R. Rohs, and E. Segal, "Unraveling determinants of transcription factor binding outside the core binding site," *Genome research*, vol. 25, pp. 1–12, 2015.

[139] M. Maienschein-Cline, A. R. Dinner, W. S. Hlavacek, and F. Mu, "Improved predictions of transcription factor binding sites using physicochemical features of dna," *Nucleic acids research*, vol. 40, no. 22, p. e175, 2012.

[140] B. Wilczynski and E. E. Furlong, "Challenges for modeling global gene regulatory networks during development: insights from drosophila," *Developmental biology*, vol. 340, no. 2, pp. 161–169, 2010.

[141] E. Afgan, D. Baker, M. Van den Beek, D. Blankenberg, D. Bouvier, M. Čech, J. Chilton, D. Clements, N. Coraor, and C. Eberhard, "The galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update," *Nucleic acids research*, vol. 44, no. W1, pp. W3–W10, 2016.

[142] H. Li and R. Durbin, "Fast and accurate short read alignment with burrows–wheeler transform," *Bioinformatics*, vol. 25, no. 14, pp. 1754–1760, 2009.

[143] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, and R. Durbin, "The sequence alignment/map format and samtools," *Bioinformatics*, vol. 25, no. 16, pp. 2078–2079, 2009.

[144] S. G. Landt, G. K. Marinov, A. Kundaje, P. Kheradpour, F. Pauli, S. Batzoglou, B. E. Bernstein, P. Bickel, J. B. Brown, and P. Cayting, "Chip-seq guidelines and practices of the encode and modencode consortia," *Genome research*, vol. 22, no. 9, pp. 1813–1831, 2012.

[145] F. Ramírez, F. Dündar, S. Diehl, B. A. Grüning, and T. Manke, "deeptools: a flexible platform for exploring deep-sequencing data," *Nucleic acids research*, vol. 42, no. W1, pp. W187–W191, 2014.

[146] D. Meyer, E. Dimitriadou, K. Hornik, A. Weingessel, and F. Leisch, "e1071: Misc functions of the department of statistics, probability theory group (formerly: E1071), tu wien, 2015," *R package version*, pp. 1.6–7, 2015.

[147] J. O. Yáñez-Cuna, H. Q. Dinh, E. Z. Kvon, D. Shlyueva, and A. Stark, "Uncovering cis-regulatory sequence requirements for context-specific transcription factor binding," *Genome research*, vol. 22, no. 10, pp. 2018–2030, 2012.

[148] H.-L. Liang, C.-Y. Nien, H.-Y. Liu, M. M. Metzstein, N. Kirov, and C. Rushlow, "The zinc-finger protein zelda is a key activator of the early zygotic genome in drosophila," *Nature*, vol. 456, no. 7220, p. 400, 2008.

[149] K. N. Schulz, E. R. Bondra, A. Moshe, J. E. Villalta, J. D. Lieb, T. Kaplan, D. J. McKay, and M. M. Harrison, "Zelda is differentially required for chromatin accessibility, transcription factor binding, and gene expression in the early drosophila embryo," *Genome research*, vol. 25, no. 11, pp. 1715–1726, 2015.

[150] T. Duque and S. Sinha, "What does it take to evolve an enhancer? a simulation-based study of factors influencing the emergence of combinatorial regulation," *Genome biology and evolution*, vol. 7, no. 6, pp. 1415–1431, 2015.

[151] D. Ezer, N. R. Zabet, and B. Adryan, "Homotypic clusters of transcription factor binding sites: A model system for understanding the physical mechanics of gene expression," *Computational and structural biotechnology journal*, vol. 10, no. 17, pp. 63–69, 2014.

[152] A. P. Lifanov, V. J. Makeev, A. G. Nazina, and D. A. Papatsenko, "Homotypic regulatory clusters in drosophila," *Genome research*, vol. 13, no. 4, pp. 579–588, 2003.

[153] J. Crocker, N. Abe, L. Rinaldi, A. P. McGregor, N. Frankel, S. Wang, A. Alsawadi, P. Valenti, S. Plaza, and F. Payre, "Low affinity binding site clusters confer hox specificity and regulatory robustness," *Cell*, vol. 160, no. 1-2, pp. 191–203, 2015.

[154] E. K. Farley, K. M. Olson, W. Zhang, A. J. Brandt, D. S. Rokhsar, and M. S. Levine, "Suboptimization of developmental enhancers," *Science*, vol. 350, no. 6258, pp. 325–328, 2015.

[155] G. Junion, M. Spivakov, C. Girardot, M. Braun, E. H. Gustafson, E. Birney, and E. E. Furlong, "A transcription factor collective defines cardiac cell fate and reflects lineage history," *Cell*, vol. 148, no. 3, pp. 473–486, 2012.

[156] S. M. Gallo, D. T. Gerrard, D. Miner, M. Simich, B. Des Soye, C. M. Bergman, and M. S. Halfon, "Redfly v3. 0: toward a comprehensive database of transcriptional regulatory elements in drosophila," *Nucleic acids research*, vol. 39, no. suppl 1, pp. D118–D123, 2011.

[157] E. Z. Kvon, T. Kazmar, G. Stampfel, J. O. Yanez-Cuna, M. Pagani, K. Schernhuber, B. J. Dickson, and A. Stark, "Genome-scale functional characterization of drosophila developmental enhancers in vivo," *Nature*, vol. 512, no. 7512, p. 91, 2014.

[158] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785–794, 2016.

[159] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.

[160] L. Torgo, *Data mining with R: learning with case studies.* CRC press, 2016.

[161] M. Slattery, T. Zhou, L. Yang, A. C. D. Machado, R. Gordân, and R. Rohs, "Absence of a simple code: how transcription factors read the genome," *Trends in biochemical sciences*, vol. 39, no. 9, pp. 381–399, 2014.

[162] M. T. Weirauch, A. Cote, R. Norel, M. Annala, Y. Zhao, T. R. Riley, J. Saez-Rodriguez, T. Cokelaer, A. Vedenko, and S. Talukder, "Evaluation of methods for modeling transcription factor sequence specificity," *Nature biotechnology*, vol. 31, no. 2, pp. 126–134, 2013.