Modeling Spatio-Temporal Enhancer Expression in *Drosophila* Segmentation

Marc Nikolaus Max von Reutern

München 2017

Dissertation zur Erlangung des Doktorgrades
der Fakultät für Chemie und Pharmazie
der Ludwig–Maximilians–Universität München

Modeling Spatio-Temporal Enhancer Expression in *Drosophila* Segmentation

Marc Nikolaus Max von Reutern
aus Freiburg im Breisgau, Deutschland

2017

Erklärung:

Diese Dissertation wurde im Sinne von §7 der Promotionsordnung vom 28. November 2011 von Herrn Prof. Dr. Klaus Förstermann betreut.

Eidesstattliche Versicherung:

Diese Dissertation wurde eigenständig und ohne unerlaubte Hilfe erarbeitet.

München, den . . . .

Marc von Reutern

Dissertation eingereicht am 9. Oktober 2017

1. Gutachter: Prof. Dr. Klaus Förstemann

2. Gutachter: Prof. Dr. Erwin Frey

Mündliche Prüfung am 10. November 2017

Dank

Ich möchte mich bei meiner Doktormutter Prof. Dr. Ulrike Gaul bedanken, die mir die Möglichkeit gab, in diesem spannenden Forschungsgebiet zu arbeiten und die Promotion mit großem Einsatz betreute. Sie hat für mich die Brücke von der reinen Physik zu deren Anwendung in der Biologie gebaut. Außerdem bedanke ich mich bei allen Kollegen und Mitdoktoranden in der Arbeitsgruppe für die großartige Zusammenarbeit und die wertvollen Diskussionen über die letzten vier Jahre. Insbesondere möchte ich Ulrich Unnerstall erwähnen, der die Arbeit engagiert betreute und immer die richtigen Fragen stellte. Prof. Dr. Klaus Förstemann und Prof. Dr. Erwin Frey danke ich für ihre Unterstützung und Betreuung beim Anfertigen dieser Arbeit. Besonderer Dank gebührt auch meinen Eltern, die mich auf meinem Lebensweg stets unterstützt haben.

Am meisten danke ich meiner Frau Dr. Corinna von Reutern-Kulenkamp, die mir bei allen schwierigen Situationen mit gutem Rat und nie endender Geduld beistand. Sie hat mich stets motiviert, wenn die Promotion am schwersten war. Ohne ihre Unterstüzung hätte ich diese Dissertation nicht anfertigen können.

Summary

Thermodynamic models are a key tool to investigate transcription control in the segmentation of *Drosophila*. By modeling the binding of transcription factors to DNA sequences and their effect on transcription initiation, thermodynamic models predict expression patterns directly from the enhancer sequence, given the binding motifs and concentrations of all relevant transcription factors (TFs). However, many parameters of the model are impossible to measure, e.g. the interaction strength between the TFs and the core promoter. Hence, it is necessary to estimate these parameters by training the thermodynamic model on known data, i.e. to fit the model predictions to already measured expression patterns of known enhancers. The quality of the parameter training result, evaluated on independent test data, indicates how well the model recapitulates the biological measurements, which can help us to improve our understanding of the underlaying mechanisms of transcription control. Therefore, proper parameter training is a crucial step for the construction of thermodynamic models.

In this thesis, I develop a thorough parameter training setup that uses the limited amount of available training data efficiently and reduces parameter overfitting significantly. This optimized training setup applies a global parameter training algorithm, a method to artificially increase the amount of training data, called data augmentation, and parameter penalties, which is a technique to limit overfitting. I apply the novel training setup to expand the scope of thermodynamic models of *Drosophila* segmentation considerably by incorporating additional TFs into the model, and to investigate many aspects of transcription control in greater detail than it was possible before. Among these topics are the specificity of TF binding motifs, the nature of TF cooperativity and DNA accessibility. With the help of the here developed impact score, I assess the influence of all relevant TFs *in silico*, delineate the cooperativity range of the key TF bcd, and determine the importance of weak binding sites. Finally, I develop and discuss two alternative models of transcription control that lack the prediction quality of thermodynamic models, but, nevertheless, give valuable insights into the architectural principles of enhancers.

This project is part of a larger effort to advance our current understanding of transcription regulation by reconstructing the segmentation network of *Drosophila in silico*. The results of this thesis facilitate future modeling efforts by optimally leveraging the available data as well as by improving our understanding of thermodynamic models.

Contents

Part I

Introduction

Chapter 1

Overview

The building plan of an organism is encoded in its genome in the form of genes. Yet, the gene sequences are only a part of the genomic information. Not all genes are needed and, therefore, active at the same time. As relevant as the gene sequence is the regulatory information that controls when and where the gene gets activated or repressed according to the requirement of the organism. Controlling gene expression enables organisms to react to external and internal stimuli, and adapt to environmental changes. Furthermore, precise gene regulation is fundamental to metazoan development, during which the organism creates a complex system of different tissues and cell types, although all its cells carry the same genome. In fact, gene sequences are often conserved over remarkably long evolutionary timescales [1] and variations between related species derive from changes in the regulatory code [2, 3, 4].

One crucial form of gene regulation is transcription control, which regulates the initiation of mRNA production and stands therefore at the starting point of gene activation. Transcription control is especially important for development and cellular differentiation. It is reasonable to assume that the organismic complexity is connected to the complexity of transcription control [5, 6]. The information that controls gene expression is encoded in the genome, often in proximity to the coding region of the gene.

In eukaryotes, spatio-temporally controlled genes are not active by default. They get activated by enhancer elements, also called cis-regulatory modules (CRMs) that recruit the polymerase to the genes' core promoter. These enhancers are regulatory regions of the genome, which are typically between 100bp and 3kb long. A single gene can be regulated by multiple enhancers located in its proximity or in its introns. Likewise, an enhancer, taken out of its native context, drives expression of promoters in its vicinity. By fusing it to a reporter construct - a signaling protein with a suitable promoter (hsp70 + LacZ)- one can measure the enhancer's inherent expression pattern.

The information that an enhancer carries is encoded in its sequence and the associated epigenetic modifications. Transcription factors (TFs) read out the information by binding to specific sites (typically 8 bp to 16 bp long) in the enhancer, which they recognize based on a DNA sequence motif. TFs can be activators, which promote gene expression, or repressors, which either prevent polymerase recruitment directly or inhibit activator function. The combined input from activators and repressors - integrated and transmitted by the mediator complex - defines the expression pattern of the enhancer. Some TFs appear to have a dual function as activator and repressor depending on the context in which they bind; however, details are the subject of ongoing research, e.g. [7, 8].

There is some controversy on the degree of TFs interactions in one enhancer [9]. Reportedly, TF bind cooperatively to the DNA, either by nucleosome displacement [10, 11], by protein-protein interaction [12], or, alternatively, DNA bending [13] and allosteric interactions via DNA torsion [14]. Additionally, short-range repression, the local inhibition of activators by a repressor, can be seen as a form of TF interaction although antagonistic in nature [15]. Either way, TF interactions shape the architecture of enhancers. However, little is known about their actual range or whether certain binding site orientations are preferred.

In the language of transcription control, the enhancers function as sentences; they can stand on their own and carry self-contained meaning. The words that assemble the sentences are the TF binding sites. They, too, carry information, but usually do not act individually and are context dependent. One focus of this thesis is to expand our knowledge of the grammatical rules that govern the integration of the TF binding sites and their interactions among each other as well as with the core promoter.

A number of reasons render the language of transcription control difficult to decipher. Unlike human languages or coding regions of genes, transcription control is fuzzily defined on multiple levels. First, TFs recognize their binding sites very unspecifically. Even multiple deviations from the consensus motif do not necessarily inhibit binding. Spelling errors in the words are the norm rather than an exception. Due to their unspecific nature, TF binding sites are distributed ubiquitously over the genome. Second, enhancers do not have a clearly defined start or end. The best way to identify enhancer sequences is to search for clusters of TF binding sites in regions of open chromatin [16, 17]. Furthermore, the rules of site integration are very difficult to grasp and flexible in nature [9, 18]. Little is known about site spacing constraints or the range of site interactions. This fuzzy nature of transcription control together with a high level of redundancy leads to a high functional robustness even under substantial sequence divergence [19, 20].

The paradigm on which this thesis concentrates is the body segmentation of *Drosophila melanogaster* embryos [21, 22]. The network consists of a cascade of TFs, which form increasingly complex patterns and ultimately specify 14 parasegments, see figure 1.1 left. During the first 2.5 h of development, the *Drosophila* embryo undergoes 13 cycles of nuclei division without membrane formation. The missing compartmentalization enables diffusion of the maternally provided TFs, which is crucial for gradient formation and the evolutionary adaption to differing egg-sizes [23]. The maternal TFs are mostly activators, which form broad gradients along the anterior-posterior (AP) and the dorsal-ventral (DV) axes of the egg, breaking the initial body symmetry, and starting the zygotic expression of gap genes. These gap factors are expressed in broad domains and function mainly as short-range repressors. During the blastoderm stage, at the onset of cellularization, maternal and gap genes regulate the expression of pair-rule genes, which are typically expressed in seven thin stripes, and also late gap genes. Typically, multiple enhancers control a single pair-rule or gap gene, each responsible for only one or two stripes. Some expression domains are accounted for by more than one enhancer, further increasing the robustness of the system. Their combined input forms the full expression pattern of the target gene in an additive fashion. Further downstream of the cascade are segmentation polarity and homeotic genes, which provide an even finer layout of the embryo.

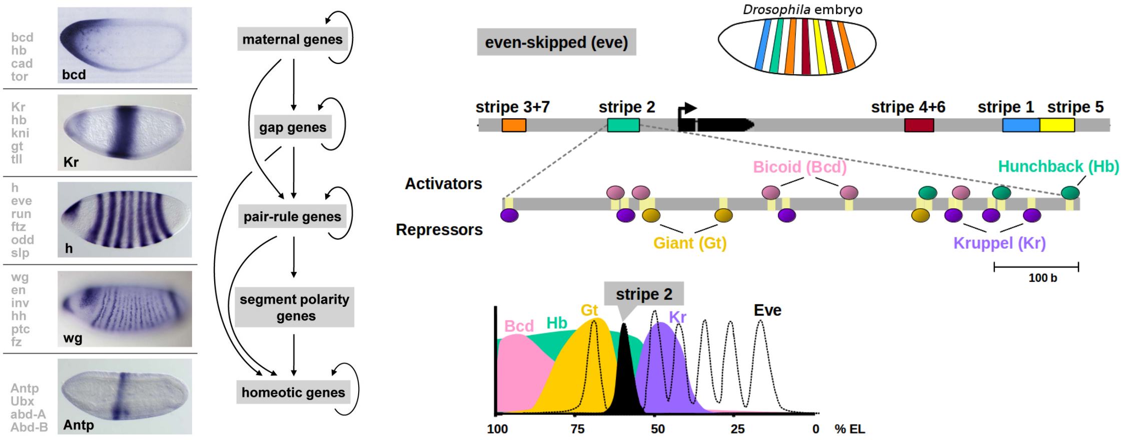


Figure 1.1: (Left) A cascade of TFs defines the map of segmentation for the anterior-posterior axis in *Drosophila* embryos. (Right) Multiple enhancers control expression of the pair-rule gene eve. Depicted as an example: a combination of broad activation and precise repression delineates the expression of stripe 2. (Both illustrations are unpublished figures for presentation purposes.)

Anterior-posterior, AP, pattern formation during cellularization is a widely used paradigm of transcription control for multiple reasons. First, the segmentation network is mostly driven by transcription. Encoded in the enhancers is the full information on the formation of gap and pair-rule patterns. Hence, it is feasible to construct a model of gene regulation solely based on the DNA sequences of the enhancers. Neither post-transcriptional regulation nor diffusion influences the spatial pattern of promoter expression substantially [24]. Second, the events during segmentation are highly precise and robust both in space and time. Although fundamentally stochastic in nature [25], the resulting patterns are reproducible and can be modeled without noise by statistical frameworks. Third, the segmentation cascade has been studied extensively. Countless knock-out experiments have identified all major factors in the system and explored how they affect segmentation. There is an abundance of data about these TFs, their distribution, binding motifs, etc. as well as about the enhancers and their expression. *Drosophila*, in general, has been a model organism for decades, with a fully sequenced genome for *D. melanogaster* and eleven closely related species available. Finally, *Drosophila* embryos are a convenient paradigm. Comprised in one embryo is a multitude of TF concentrations in parallel, making a comparative study of various conditions and levels of input possible in one experiment. Throughout the thesis, the focus will be on relative differences rather than absolute values, thereby circumventing the problem that exact protein numbers are rarely known and more complex to model.

This thesis aims at recreating transcription control *in silico* in order to understand it. The modeling framework of choice is a version of sequence-to-expression models, which are also called thermodynamic models because their approach is rooted in statistical physics, e.g. [15, 26, 27, 28]. Instead of modeling the segmentation in a coarse and simplified manner, thermodynamic models account for single TF binding events and incorporate many mechanistic details like cooperative binding and the quality of binding sites. This makes these models much more complex, and therefore more difficult to construct, but allows for a deeper understanding of the rules governing transcription control. In summary, thermodynamic models predict expression patterns of enhancers from their sequence, with the help of TF concentration profiles and binding specificities. The model used in this thesis is a version of GEMSTAT [28], adapted to greater performance and with additional analytical features.

GEMSTAT models repressors like activators. The transition between a weak repressor and a weak activator is continuous. A parameter, called the activatory potential, determines not only the influence of the TF on the expression level but also its role. An activatory potential above $\beta = 1.0$ indicates an activator, a value below indicates a repressor. While this approach helps to simplify the model, it does not depict the workings of repression. Most repressors in the early segmentation network affect activator binding locally rather than the polymerase directly [29]. Hence, the last chapter of this thesis explores alternative models of repression.

Unfortunately, many parameters needed to model transcription control, like the activatory potential or the absolute binding affinity of the TFs, are difficult or even impossible to measure. These parameters have to be learned by fitting the model to known data, a procedure that is called parameter optimization. A central element of optimization is the quantification of prediction quality in the form of the objective function, which measures how closely the model prediction resembles the measured data. A lower objective function score indicates a better model fit. Especially when the model has many parameters, the task of finding a good parameter optimum is not trivial. There are many different optimization algorithms, which differ in their strategy to navigate the parameter space efficiently.

A good model fit on the training data does not automatically imply a good model in general. One must always consider overfitting. Complex models with many parameters have an advantage during parameter training because they are more flexible and can better adapt to the data. This flexibility makes them prone to assign too much meaning to random noise in the data. For instance, a specific binding site configuration, which is present in only one enhancer, is most likely a random occurrence instead of a functional unit. The difference between a good and an overfit model is that the former predicts unseen data well[1] because it distinguishes between general and random features in the data. In other words, a good model generalizes well. A key to high-quality predictions is to make

[1] As indicated by a good test score, i.e. the objective function score on test data, which was not used during training.

the exploitation of rare data features harder for the model. This, however, reduces the possibility to explore the implications of these rare features and gain new insights into the details of transcription control. For this reason, large parts of this thesis are dedicated to the topic of model training and to find a good balance between model exploration and preventing overfitting.

Thermodynamic models have been studied and applied for research for more than a decade [15, 26, 27, 28]. But most publications emphasize only the application of the model to data and comparably little effort is put into the training aspects of the model. Often, the models are trained on small data sets and without proper Cross-Validation. However, a thorough parameter training set-up, as well as an in-depth analysis of the parameter result, are crucial for any modeling endeavor; not only for reproducibility but also in order to distinguish more clearly between different models and to analyze how the model works and what drives the prediction. To my knowledge, only two publications concentrate on these issues. Suleimenov et al. compared global and local parameter optimization techniques and found that global techniques are superior in some cases [30]. However, their analysis is based on artificial data. Dresch et al. performed a sensitivity analysis to find out which parameters influence the prediction [31]. Both publications are based on a limited dataset of small size. This naturally reduces the number of parameters that can be trained reliably. I know of no publication that tries to measure additional training data or collect the data that is already published.

In this thesis, I build on the result of preceding modeling efforts and improve the parameter optimization setup. I additionally broaden the data foundation by collecting all available enhancers with measured expression pattern from the literature, more than doubling the number of available enhancers. These efforts improve the model predictions substantially.

Improving the model's ability to predict unseen data is not an end in itself. The fundamental goal of modeling is to illuminate the rules governing transcription control as they are encoded in enhancers. A high prediction quality is merely an indication of how strongly the model approximates reality because it learned generalizable rules. This thinking is brought to an extreme when model validation is treated only as an afterthought to prove the soundness of the model. Often, the main modeling effort uses the full dataset, arguing that the focus lies on finding the optimal model and not the optimal prediction, e.g. [15, 32, 33]. This interpretation misses that the parameter result has to be scrutinized just as thoroughly and should always be reviewed on previously unseen data. Therefore, instead of presenting a putative optimal parameter result, I am going to always present a range of results, derived from multiple parsings of training data, and use the spread of the distribution as an indicator for the inherent inaccuracy of the parameter prediction. Furthermore, when determining the influence of a parameter, I have tested its impact not on the training data, but on the test data.

A central aim of this thesis is to describe a modeling and parameter training setup that facilitates the exploration of transcription control at greater depth than was possible before. After presenting the fundamentals of thermodynamic models and the data foundation, I am going to discuss fundamental aspects of training thermodynamic models, which will be applied in the rest of the thesis. The topics will be the choice of the objective function, the parameter optimization algorithm, as well as how data augmentation and hyperparameter training can improve the quality of the model. Using the optimized training setup enables to increase the number of training parameters and incorporate new features into the model. Among them are additional TFs, which have not been considered before, various TF interactions, and DNA accessibility. Applying the impact score, which is a quantitative *in silico* counterpart to an *in vivo* knock-out experiment, I am going to assess the influence of various model elements, like the role of the single TFs, cooperativity, accessibility, but also of model aspects not connected to a parameter, like the role of weak binding sites and steric hindrance. Finally, I am going to explore alternative models, which are a simplified version of GEMSTAT, and apply a new approach to short-range repression.

Chapter 2

Thermodynamic Models

2.1 Models of Protein-DNA Interaction

The central aim of thermodynamic models is to decipher information from regulatory sequence. This means predicting expression based on the sequence of an enhancer. Therefore, the identification and weighting of TF binding sites is a crucial elements of thermodynamic models. However, protein-DNA interaction is a complex process involving the chromatin state, the DNA-sequence, as well as DNA-shape [34]. In order to build a feasible model, I focus solely on the DNA sequence as the main predictor of TF-binding. Further features like chromatin accessibility and DNA shape can be implemented additionally [35, 36]. In the following, I describe a model of a single TF binding site by following the theoretical foundations of Stormo et al. [37] and their implementation by Segal et al. [27] and He et al. [28].

2.1.1 Binding Motifs

Consensus Sequences

Given a set of known TF binding sites, it is possible to identify a sequence pattern that captures the basic mutual features of the sites. This sequence pattern is a tool to identify binding sites, optionally allowing some mismatches. This pattern based binding model is called the consensus sequence. For instance, take the TF krueppel, for which footprinting experiments identified a handful of sites [38]; the first six are:

AAACGGATT
GACCGGGTT
GAAGGGATT
AACTGGGTT
AAACGGGTT
CAAAGAGTT

AAACGGGTT strict consensus

RAMNGGRTT altern. consensus

Here, two possible consensus motifs are depicted. Solely the most frequent base at every position was selected for the strict consensus. The alternative consensus incorporates also degeneracies using IUPAC characters [39]. The limitations of consensus sequences are immediately apparent. First, most identified sites show some mismatches. Only one of the six sites fits the consensus perfectly. However, if ambiguity is allowed, specificity gets lost. To illustrate this point, consider that one expects to find the strict consensus by chance once every 262 kb and the alternative consensus approximately once every 8 kb. Second, consensus sequences are designed for optimal human readability and, therefore, omit information. By reporting only the most important base in every position, information on alternatives is lost.

Position Weight Matrices

An alternative to consensus sequences is the position weight matrix (PWM) [37, 40]. The entries of the PWM are the frequencies, with which the bases A, C, G, and T occur at every position of the binding motif [1].

Based on the six sites mentioned earlier, the krueppel PWM is:

$$\mathrm{PWM}_{Kr} = \begin{array}{cc} \text{pos.} & \begin{array}{cccc} A & C & G & T \end{array} \\ \begin{array}{c} 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \\ 8 \\ 9 \end{array} & \begin{pmatrix} 3 & 1 & 2 & 0 \\ 6 & 0 & 0 & 0 \\ 4 & 2 & 0 & 0 \\ 1 & 4 & 1 & 1 \\ 0 & 0 & 6 & 0 \\ 1 & 0 & 5 & 0 \\ 2 & 0 & 4 & 0 \\ 0 & 0 & 0 & 6 \\ 0 & 0 & 0 & 6 \end{pmatrix} \end{array} \cdot \frac{1}{N_{\text{sites}}}$$

N_{sites} is a normalization constant so that the rows sum to 1. The advantage of the PWM is that one can easily calculate the enrichment, i.e. the likelihood-ratio, of base b at a certain position p in the motif.

$$LR(p, b) = \frac{\mathrm{PWM}(p, b)}{P_{\text{bg}}(b)} \tag{2.1}$$

The PWM in the numerator is simply the probability of finding b at p in the set of known binding sites. P_{bg} is the background probability of finding base b anywhere in the genome. Consider a sequence $S = s_1 \cdots s_n$ of the length of a binding site. The enrichment of S in

[1] Most sources call this definition of the PWM the position count matrix (PCM) or position frequency matrix PFM. The common definition of the PWM is a matrix of log-likelihood-ratios of base occurrence between binding sites and background. I stick to the above given definition because it fits well with the definition of binding weight. Either way, both PWM and PCM carry the same information and can be easily transformed into each other.

the set of binding sites in comparison to the background is:

$$LR(S) = \prod_{s_i} \frac{\mathrm{PWM}(i, s_i)}{P_{\mathrm{bg}}(s_i)} \tag{2.2}$$

The equation above assumes that all positions in the motif are independent of each other. This is not necessarily true, which I am going to discuss later.

Although PWMs are not designed for human readability, there is a clear graphical representation of them in the form of the sequence logo [41]. The sequence logo depicts every position of the PWM as a stack of the base symbols A, C, G, and T. The frequency of the base is depicted by the relative hight of their symbols. The total height of the stack is determined by the relevance of that position. Hence, it is easy to spot the most important base in the relevant positions, e.g. 2.1. The base symbols are often ordered with the most important base at the top, so that reading the consensus is easy.

The relevance of a position is the information content, which is defined as the difference in uncertainty between the background and the motif.

$$R(p) = H(p) - H_{\mathrm{bg}} = \sum_b \mathrm{PWM}(p, b) \log_2(\mathrm{PWM}(p, b)) - \sum_b P_{\mathrm{bg}}(b) \log_2(P_{\mathrm{bg}}(b)) \tag{2.3}$$

A PWM position carries no information ($R = 0$ bit) if its entries just resemble the background base frequencies. For an uniform background model ($P_{\mathrm{bg}}(b) = 0.25$), the maximum information content is $R = 2$ bit.

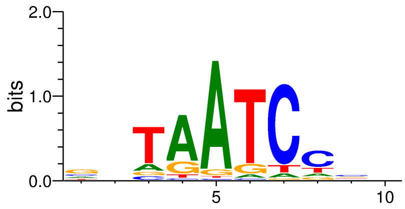


Figure 2.1: The sequence logo of bcd generated by WebLogo [42]

2.1.2 Binding Affinity

Up to now, the equations describe the statistical aspects of TF binding. In the following, I am going to discuss the biophysical aspects of binding. The starting-point is the binding kinetic:

$$TF + S \underset{k_-}{\overset{k_+}{\rightleftarrows}} TF \cdot S$$

S is a binding site on a DNA strand and $TF \cdot S$ the complex formed by the TF bound to the DNA. The $k_\pm$ are the binding and dissociation rates. For thermodynamic models, it is assumed that binding and dissociation are in equilibrium and that both rates are faster than the processes that gets modeled. This is not a trivial assumption and I am going to discuss it later in greater detail. The dissociation constant K_d is defined as the ratio between the concentrations of bound and unbound TFs.

$$K_d = \frac{k_-}{k_+} = \frac{[TF][S]}{[TF \cdot S]} \propto \exp(\Delta G/RT) \tag{2.4}$$

The last term in 2.4 connects the ratio between the bound and unbound state with a change in Gibbs free energy ΔG, which is often just called binding energy. R is the gas constant and T the temperature of the system. From this, it is possible to derive the probability that site S is bound:

$$P(TF \cdot S) = \frac{[TF \cdot S]}{[TF \cdot S] + [S]} \overset{(2.4)}{=} \frac{[TF]}{[TF] + \frac{1}{\alpha}\exp(\Delta G/RT)} \tag{2.5}$$

α is a scaling factor that is called absolute affinity. It accounts for the absolute energy level $E_0 = \log(-\alpha/RT)$. Equation 2.5 resembles a Fermi-Dirac distribution, which can be complicated to handle since α is rarely known. However, if I assume that the binding site is weakly bound (small binding energy, i.e. low concentration and or low affinity), it is possible to transform 2.5 into a Boltzmann distribution.

$$P(TF \cdot S) \approx \alpha \cdot [TF] \cdot \exp(-\Delta G/RT) \tag{2.6}$$

This transformation has multiple advantages. First, in this scenario, the binding probability scales linearly with the factor concentration. And second, comparing different binding sites becomes much easier. Let's take sites S and S' bound by the same TF. Their occupancy ratio becomes

$$\frac{P(TF \cdot S)}{P(TF \cdot S')} = \exp((\Delta G' - \Delta G)/RT) = \frac{w}{w'} \tag{2.7}$$

The absolute terms cancel each other out so that only the site-specific terms $w = \exp(-\Delta G/RT)$ remain. For this reason, I call w the relative binding weight.

Another assumption simplifies the comparison of binding sites even further. If every base in the binding site forms an independent bond with the TF and all these bonds do not interfere with each other, then it is possible to decompose the binding energy in portions.

$$\Delta G = \sum_p \epsilon_p \tag{2.8}$$

Every bond accounts for a share of the total binding energy. This is called the additivity assumption, which I made in a similar form for equation 2.2. The sum becomes a product when one transforms the energy terms into binding weights.

$$\frac{P(TF \cdot S)}{P(TF \cdot S')} = \prod_p \frac{w_p}{w'_p} \tag{2.9}$$

This equation looks similar to 2.2. In fact, it is possible to show — under the assumptions made in this section — that the binding weights relate to the base frequencies of known binding sites [43].

$$w_p = \exp(-\epsilon_p/RT) \propto \frac{PWM(p, s_p)}{P_{\text{bg}}(s_p)} \tag{2.10}$$

At this point, the statistical description of binding sites connects with the energy description of binding kinetics and, therefore, justifies naming the matrix derived from base counts position weight matrix.

2.1.3 Assumptions

Three assumptions are essential for thermodynamic models. Their implications should be taken serious and can not be stressed enough. Here I recapitulate them in more detail.

1. Equilibrium: This is a key assumption to make for the model of protein-DNA binding and is ultimately the motivation why the expression model is called a thermodynamic model [44, 45]. Equilibrium implies that binding and dissociation happens on time scales much faster than changes in the cell environment and therefore transcription factor concentrations correspond directly to DNA occupancy. There is evidence that this assumption is justified. Photo-bleaching experiments indicate that TFs are bound transiently to DNA with residence times of the order of seconds [46, 47], unlike the developmental processes in the fly that happen on timescales of the order of several minutes. Another implication is that all processes are reversible and that therefore the binding energy alone determines TF residence time. Especially, any non-reversible assisting processes are inexistent, e.g. recruiting or remodeling factors. There are indications that this might not always be true [48], however, details are unknown. Nevertheless, most TF display binding *in vitro* as well as in bacterial environments [49, 50], proving that predicted binding energies are a good indicator of TF occupancy and assisting processes are not fundamentally necessary.

2. Weak binding: In equation (2.6) I assumed that the amount of free transcription factors is low and therefore $[TF] << (\alpha w)^{-1}$, resulting in a low binding probability. This enabled us to transform the bulky logistic equation 2.5 into an equation that is linear in the TF concentration and the affinity. Furthermore, this assumption allows us to separate the TF-specific (concentration, affinity) from the site-specific terms (relative weight) of the binding probability.

The TF-specific terms are unknown in many scenarios, e.g. during an analysis of binding site clusters [51]. In this case, the only available information is the relative weight w based on the PWM and the binding sequence. Then, equation 2.7 justifies the use of the relative binding weight as a measure of binding strength although it is only a precursor of binding probability.

Note, that the weak binding assumption is only important in order to connect binding statistics with the concept of binding energies and weights. In the following, it becomes apparent that the thermodynamic model itself assumes a Fermi-Dirac distribution for the occupancy of a binding site and, therefore, is able to model strong as well as weak binding sites (see equation 2.14).

3. Additivity: While calculating the specificity, I assumed that the single positions in the binding motif contribute independently to the binding energy. There are $4^{10} = 1048576$ possible 10-base long binding motifs. Even high throughput techniques struggle to measure all those binding sites thoroughly. Assuming simple additivity reduces the number of measurements to $3 \cdot L + 1 = 31$. (Consensus sequence plus three nucleotide mutations in every position.) Most methods of measuring binding specificities would be unfeasible without this reduction. Similarly, techniques for which binding sites are sampled randomly instead of systematically, e.g. B1H, report rarely enough sites to calculate accurate statistical correlations between the single positions. The additivity assumption is crucial, but controversial [52, 53]. However, a recent comparison of different specificity models shows that for most tested transcription factors, mononucleotide additivity can compete with more complex models [54].

4. Diffusion: There is an additional assumption that is relevant for the calculation of spatial expression patterns but not thermodynamic models in general. TF and RNA diffusion is negligible in the segmentation network of *Drosophila* during the blastoderm stage. This assumption does not affect the model itself. However, it becomes important if one calculates expression rates of neighboring embryo segments. The assumption states that it is possible to calculate the expression level for every position in the embryo independently without considering that expression products diffuse to other positions. Diffusion of TFs is crucial during the early stages of embryogenesis, especially for the formation of the bcd gradient [23]. However, during the blastoderm stage at the onset of cellularization, diffusion does not affect the expression patterns considerably, otherwise the fly would not be able to develop sharp expression patterns like the stripes of the pair rule genes.

2.2 Multiple Binding Sites

Up to know, I modeled a single binding site. The following section describes multiple binding sites and, ultimately, binding configurations of whole enhancers. The weight description, which has been developed in the last sections, will be very useful. To simplify notation, I define the absolute binding weight, notated as capital W, as the relative weight w in combination with the TF-specific parameters:

$$W = \alpha \cdot [TF] \cdot w \tag{2.11}$$

2.2.1 Partition Function

To describe the system of multiple binding sites, one needs micro- and macro-states. A micro-state is a specific configuration in which the binding sites are occupied, e.g. for a system with two sites, the state that the first site is bound and the second is empty. A macro-state is a combination of micro-states, e.g. all states in which only one site is bound. According to Boltzmann's law, the propensity of the system to be in a certain state is dependent on its energy level.

$$P(\text{micro-state } n) \propto \exp(-E_n/RT) =: \omega_n \tag{2.12}$$

Since all micro-states are mutually exclusive, one can easily calculate the probability of a macro-state.

$$P(\text{macro-state } M) = \frac{\sum_{n \in M} \omega_n}{\sum_{\text{all}} \omega_n} = \frac{Z_M}{Z_{\text{tot}}} \tag{2.13}$$

Z is the sum of weights and is called the partition function. By using that all probabilities have to sum to 1.0, it is possible to reduce the calculation of probability to a simple counting of weights. Note that it is not necessary to know the absolute weights ω. Since every term gets divided by Z_{tot}, it is possible to omit the scaling factor. The only crucial term is the relative ratio between the weights. This simplifies many calculations because absolute values are often unknown, e.g. absolute TF concentrations. To see it from another perspective, rescaling of the weights is the same as defining a new zero-point for the energy in equation 2.12.

2.2.2 Calculating Occupancies

Imagine a stretch of DNA with N binding sites s of binding weight W_s. Consider the possibility of multiple TFs although a binding site is always TF-specific. If two TFs bind to the exact same position, the model defines two binding sites, one for each TF. Notice that an unoccupied site has a different weight W_0 from an occupied. It is always possible to scale all weights by a mutual factor and define the zero point of energy, i.e. the reference point, so that $W_0 = 1$.

Consider a single binding site $N = 1$. There are two micro-states: the TF is bound with weight W_1 or not bound $W_0 = 1$. Hence, the occupancy of the site is:

$$p_{\text{bound}} = \frac{W_1}{W_0 + W_1} = \frac{\alpha \cdot [TF] \cdot w}{1 + \alpha \cdot [TF] \cdot w} \tag{2.14}$$

This resembles equation 2.5, which can be derived without the use of the weak binding assumption. The weak binding assumption is not relevant in the following. It has been necessary for the derivation of the relative binding weight w. If there were an alternative way to calculate the binding energies, it would be still possible to use the equations from this section.

Before modeling an arbitrary number of binding sites, consider the scenario with only two sites W_1 and W_2. What is the probability that both sites are bound $(1, 1)$? Since I assume that the sites are independent, it is possible to add their binding energies in order to calculate the full energy of the system. This translates to a multiplication on the weight level:

$$W_{1,2} = \exp(-E/RT) = \exp(-(E_1 + E_2)/RT) = W_1 \cdot W_2 \tag{2.15}$$

The other micro-states can be calculated similarly, with one or both binding weights replaced by W_0. This yields:

$$p(1,1) = \frac{W_1 \cdot W_2}{1 + W_1 + W_2 + W_1 \cdot W_2} \tag{2.16}$$

The macro-state of site one bound $(1,?)$, consists of the states $(1,0)$ and $(1,1)$. Its probability is:

$$p(1,?) = \frac{W_1 + W_1 \cdot W_2}{1 + W_1 + W_2 + W_1 \cdot W_2} \tag{2.17}$$

I call this the occupancy of site 1. Of course, it is possible to simplify it by extracting the factor $1 + W_2$.

$$p(1,?) = \frac{W_1}{1 + W_1} \tag{2.18}$$

As expected, we recover the case of a single site, because both sites are independent. So why bother with multiple sites? To answer this, assume the two sites were not independent. For instance, both binding sites overlap and only one of the sites can be bound at same time. In this case, the micro-state $(1,1)$ would be forbidden and the occupancy of site 1 would be:

$$p(1,?) = \frac{W_1}{1 + W_1 + W_2} \tag{2.19}$$

In this scenario, the TFs compete for binding, which is reflected in the occupancy. If site 2 is weak ($W_2 \approx 0$), the occupancy of site 1 is the same as in equation 2.14. If, on the other hand, site 2 is strongly bound ($W_2 >> W_1$), site 1 is inaccessible $p(1,?) \approx 0$.

It is now easy to generalize to an arbitrary number of sites. Keep in mind that an AND-operation (site 1 and site 2 is bound) requires the multiplication of binding weights, while an OR-operation (site 1 or site 2 is bound) requires the addition of weights as can be seen above.

There is one caveat. The number of possible configurations increases exponentially with the number of sites. It is necessary to calculate all of them for Z_{tot}. However, there is an efficient way to go through all configurations performing a number of steps linear to the number of sites. The main idea is to go through all sites one by one starting at one end of the sequence. Let's define the precursor partition function Z_s as the partition functions of all sites up to position s. Z^*_s is the partition function up to s with the last site explicitly bound. It is possible to calculate both sums by iteration. The starting point is:

$$Z^*_0 = 1 \qquad Z_0 = 1 \tag{2.20}$$

The update rules are as following:

$$Z^*_s = W_s \cdot Z_{o(s-1)} \qquad Z_s = Z_{s-1} + Z^*_s \tag{2.21}$$

Here, I used $o(s-1)$ as an abbreviation for the last site that does not overlap with site s. The update-rule for Z_s resembles the two cases that position s is occupied Z^*_s and that

s is empty Z_{s-1}. The full partition function is simply the last precursor sum $Z_{\text{tot}} = Z_n$. Let's define Y_s and Y_s^* in the same fashion, except that Y counts the sites in the opposite direction starting from the other end of the sequence. Now we have everything we need to calculate the occupancy of any site s.

$$p(s) = \frac{Z_s^* \cdot Y_s^*}{W_s \cdot Z_n} \tag{2.22}$$

Note that the W_s appears in the denominator because the weight of site s has been accounted for twice, once in Z_s^* and the second time coming from the opposite direction with Y_s^*. Nevertheless, $p(s)$ still scales more or less proportional to W_s with some corrections from interacting sites.

2.2.3 Cooperativity

An assumption of the last section was that the binding sites do not interact if they do not overlap. Under this assumption, the binding energy of two sites is simply the sum of their respective energies, leading to a multiplication of their weights. Let's advance beyond independent sites and incorporate interaction terms explicitly.

Interaction of sites is often called cooperativity. The assumption is that neighboring binding events support each other by e.g. opening the chromatin, changing the DNA-shape or forming protein-protein bonds [55, 56, 57]. Although a supportive interaction is mostly assumed, I do not exclude suppressive interactions. In fact, the model described here incorporates a general form of interaction without going into the details of the actual biological mechanism. The key idea is that any interaction changes the energy landscape of the bound protein. I express this change as a correction term ϵ for the total energy E.

$$E = E_1 + E_2 + \epsilon \tag{2.23}$$

If $\epsilon < 0$, then the cooperatively bound TFs are on a deeper energy level than the independently bound TFs and, therefore, are more firmly bound. In contrast, if $\epsilon > 0$, the energy well is not as deep. The extreme case would be $\epsilon >> 0$ in which the state of simultaneously bound TFs would be forbidden energetically. An example of the latter would be overlapping sites that can not be bound at the same time. The weight of the combined state $W_{1,2}$ is therefore

$$W_{12} = e^{-E/RT} = e^{-E_1/RT} e^{-E_2/RT} e^{-\epsilon/RT} = W_1 \gamma W_2 \tag{2.24}$$

γ is the cooperativity parameter. Following the analysis of energy balance, $\gamma > 1$ is a cooperative interaction and $\gamma < 1$ is a repressive interaction. In general, I am going to use γ as a cooperativity function depending on the distance between the binding sites and the type of interacting TFs.

An additional assumption is necessary to efficiently incorporate cooperativity for the calculation of the partition function: two bound TFs can only interact if there is no third TF bound between them, see figure 2.2. This assumption has mainly practical reasons, although, depending on the biological mechanisms of cooperativity, it is definitely plausible. The update rules for the precursor partition functions are now:

$$Z_s^* = W_s \sum_{\sigma=0}^{s-1} \gamma(s,\sigma) \cdot Z_\sigma^* \qquad\qquad Z_s = Z_{s-1} + Z_s^* \tag{2.25}$$

Where $\gamma(s,\sigma) = 0$, if sites s and σ overlap. When implementing the update rule for Z^*, it is not necessary to calculate the full sum from the beginning up to the position $s-1$. If there is a maximum cooperative range and t is the last non-interacting site, i.e. $\gamma(s,\tilde{t}) = 1$ for all $\tilde{t} <= t$, then there is a shortcut:

$$\sum_{\sigma=0}^{s-1} \gamma(s,\sigma) \cdot Z_\sigma^* = Z_t + \sum_{\sigma=t+1}^{s-1} \gamma(s,\sigma) \cdot Z_\sigma^* \tag{2.26}$$

It is easy to prove the latter by induction over t.

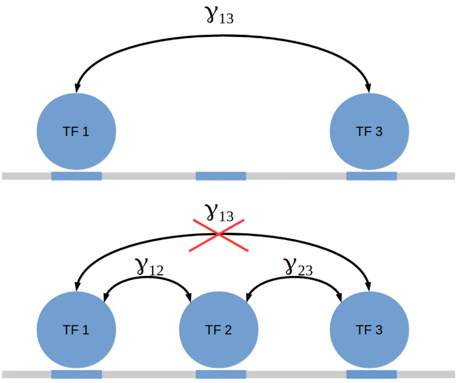


Figure 2.2: Next neighbor assumption: TF 2 interrupts the interaction between site 1 and 3.

2.3 Enhancer-Promoter Interaction

Thermodynamic models simplify gene expression by reducing it to a single step: recruitment of the polymerase to the core promoter. All consecutive steps are assumed to be independent of the enhancer sequence. Therefore, the model does not predict expression levels, which would require a detailed model of the promoter and other factors, but the probability that the core promoter is occupied. By staying in the binding energies framework, one can interpret the core promoter as a binding site with some weight q_{btr}. q_{btr} is called the basal transcription rate because it represents the autonomous binding rate of the polymerase. In this model, the TFs bound to the enhancer help to recruit or repel the polymerase by forming bonds via a mediator complex. These bonds carry binding energy

ϵ_{TF} depending on the type of TF. An activator lowers the energy barrier if $\epsilon_{\mathrm{TF}} < 0$ and, therefore, promotes polymerase binding. Otherwise, if $\epsilon_{\mathrm{TF}} > 0$, the TF is a repressor.

The model distinguishes two macro-states: the promoter is occupied with partition function Z_{ON} and the promoter is not bound with Z_{OFF}. As with TF binding sites, I define q_{btr} so that the empty site has weight 1.0. The probability that the polymerase is bound can be calculates as:

$$p_{\mathrm{bound}} = \frac{q_{\mathrm{btr}} \cdot Z_{\mathrm{ON}}}{Z_{\mathrm{OFF}} + q_{\mathrm{btr}} \cdot Z_{\mathrm{ON}}} \tag{2.27}$$

I already calculated Z_{OFF} in the last section as the unaltered partition function of the full enhancer by applying the update rules 2.21 or 2.25 to incorporate TF-interactions. It is possible to calculate Z_{ON} in the same fashion by incorporating extra energy terms.

$$Z^*_{\mathrm{ON},0} = 1 \qquad Z_{\mathrm{ON},0} = 1 \tag{2.28}$$

The update rules are:

$$Z^*_{\mathrm{ON},s} = \beta_{\mathrm{TF}} \cdot W_s \sum_{\sigma=0}^{s-1} \gamma(s,\sigma) \cdot Z^*_{\mathrm{ON},\sigma} \qquad Z_{\mathrm{ON},s} = Z_{\mathrm{ON},s-1} + Z^*_{\mathrm{ON},s} \tag{2.29}$$

Every time one encounters a bound TF, it is necessary to account for an additional energy term:

$$\beta_{\mathrm{TF}} = \exp(-\epsilon_{\mathrm{TF}}/RT) \tag{2.30}$$

The parameter β_{TF} is called the activatory potential of the TF. I identify activators as TFs with $\beta_{\mathrm{TF}} > 1$ because they increase the $Z_{\mathrm{ON}}/Z_{\mathrm{OFF}}$-ratio and, therefore, promote gene expression. Repressors have $\beta_{\mathrm{TF}} < 1$; they inhibit gene expression. The continuous transition from activators to repressors around $\beta = 1.0$ is one advantage of thermodynamic models because it allows us to treat both types of TF in the same way. Notice that a repressor with β_{R} annihilates the effect of an simultaneously bound activator with activatory potential $\beta_{\mathrm{A}} = 1/\beta_{\mathrm{R}}$. The weights of the TFs are not important for that matter because they only affect the likelihood of these two TFs being bound. In general, the binding weight of a TF matters greatly since the configuration with a firmly bound TF carries a larger share of the total weight than with a loosely bound TF.

2.4 Implementation

The thermodynamic model that I used for all predictions of this thesis is a modified version of GEMSTAT [28]. The core algorithm of GEMSTAT remained relatively unchanged. I implemented modifications that increase the performance of the program, enabled parallelization, and included additional features for improved execution handling and visualization. Additionally, I included data analysis tools into GEMSTAT that calculate the impact score of the most relevant program parameters, see section 4.2.3.

Two additions that I implemented were also implemented by the creators of GEMSTAT in follow-up publications. They included accessibility information to improve binding site predictions, see section 6.3 and [35]. Furthermore, they added a global parameter optimization technique, see section 4.3.2 and [30].

The source code and all changes from the original GEMSTAT version are accessible as a GitHub repository [58].

Chapter 3

Data Foundation

For any kind of model training, it is necessary to have a strong data foundation. The data is the basis to train unknown parameters and assess the model quality. From a modeling perspective, the enhancer elements are the essential data. Their sequences are the model input and their expression profiles are the output. The transcription factors are the second type of data. They build a semantic knowledge base that structures the sequence into binding sites. This chapter presents the composition, source, and quality of the available data.

3.1 Enhancers

Enhancer elements, also called cis-regulatory modules (CRM), are the core data. Thermodynamic models help to understand and computationally reconstruct how enhancers regulate gene expression. As input, they take the enhancer sequences and predict the enhancer expression levels along the anterior–posterior embryo axis as output. Therefore, an enhancer is more than a single data point. Along the embryo axis, it comprises expression rates for multiple TF compositions, which can be compared relatively. For the *Drosophila* segmentation paradigm, it is a standard procedure to model expression with a resolution of 1% egg length, yielding 100 expression levels. Naturally, neighboring sections of the embryo are similar in their TF compositions and show similar expression levels reducing the number of effectively independent data points.

To measure the expression pattern of an enhancer, one fuses its sequence with a reporter construct and integrates it into the genome of *D. melanogaster* embryos, figure 3.1. Reporter constructs contain a basal promoter driving a reporter gene, e.g. *lacZ* or *GAL4*, which gets localized by *in situ* hybridization against its transcript. The enhancer sequence controls expression largely independently of the genome integration site and the gene sequence. The core promoter gets selected to be susceptible to expression control, e.g. by selecting a pair rule gene promoter.

Expression patterns in *Drosophila* embryos are sharp and precisely positioned. The expression rate was measured in a binary fashion, i.e. distinguishing only between expressed

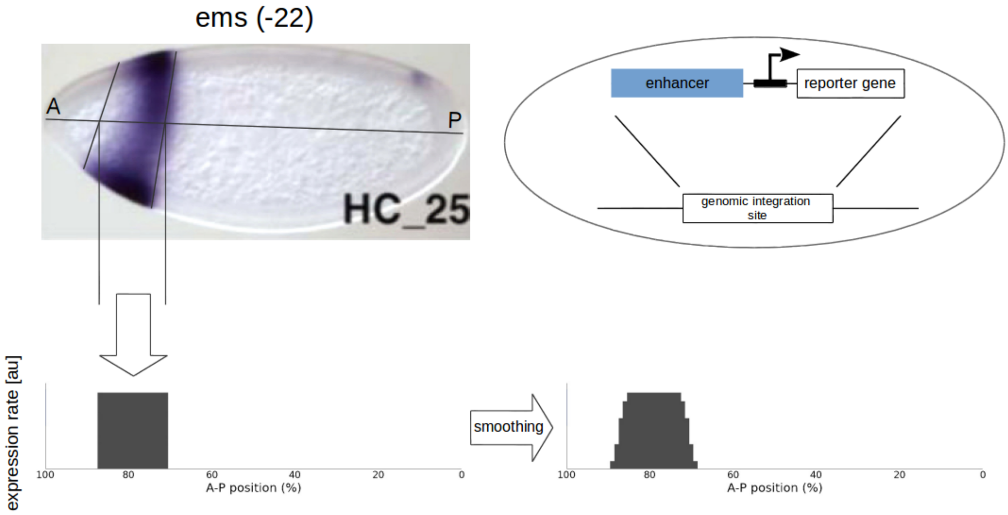


Figure 3.1: An embryo stained with a reporter construct for the enhancer ems (-22). The images were digitalized by hand. Relevant for the localization of the expression domain is the staining at the embryo borders, where the nuclei are located. The first read-out is binary (expressed/not-expressed), which gets smoothed to resemble natural gradients.

and not-expressed, as described in figure 3.1. This is due to the limitations of most reporter construct experiments. Most data in the literature was generated with simple staining methods, although there are methods for high-resolution expression measurements, which measure even absolute expression levels [59].

The expression profiles are artificially smoothed after the measurement. This is done so that the profiles resemble actual patterns, which have a natural gradient at the edges due to diffusion and graded input.

While the measurement of enhancer output is relatively simple, the identification of enhancers and delineating their sequence is not. Enhancers have no clearly defined boundaries and differ in length. The typical size of an enhancer is between 500bp and 1500bp. Although there are tools for enhancer prediction *in silico* [60, 16], only laborious *in vivo* experiments can prove enhancer activity.

I found 98 enhancers in the literature that drive clear expression patterns in the mid-blastoderm stage before cellurarization [16, 22, 61, 62, 17, 63]. I considered only enhancers that are differentially expressed exclusively on the AP-axis. A full list together with expression profiles is in the appendix IV. I chose a uniform naming scheme in which the enhancer is named after the gene, which it presumably controls in combination with a location identifier (approximated distance to the gene promoter in kb, + for downstream and − for upstream).

Only a fraction of the found enhancers is optimally delineated. Hence, the enhancer sequences vary in size between about 360bp and over 2700bp. E.g. Kvon et al. parsed about 13.5% of the *D. melanogaster* genome in 2kb tiles without pre-screening. They measured expression systematically [61]. Thus, the tested regions contain enhancers as well as flanking sequences, which could include additional binding sites that disturb the original enhancer pattern. Nevertheless, I follow the principle to implement exactly what has been measured since, first, any influence of flanking sequence is accounted for by the reporter construct and, second, any *in silico* delineation would be speculative.

The expression domains are unevenly distributed over the embryo axis. Figure 3.2 depicts the number of expressed enhancers for every position along the AP-axis. While there are over 30 enhancers in the future head of the fly, only 7 are expressed in the posterior cap. This imbalance can lead to biases in the model and should be kept in mind during parameter training. As an example, consider a hypothetical TF expressed in the head region that binds equally to all enhancers. Any model that tries to fit the data would be inclined to implement this TF as an activator.

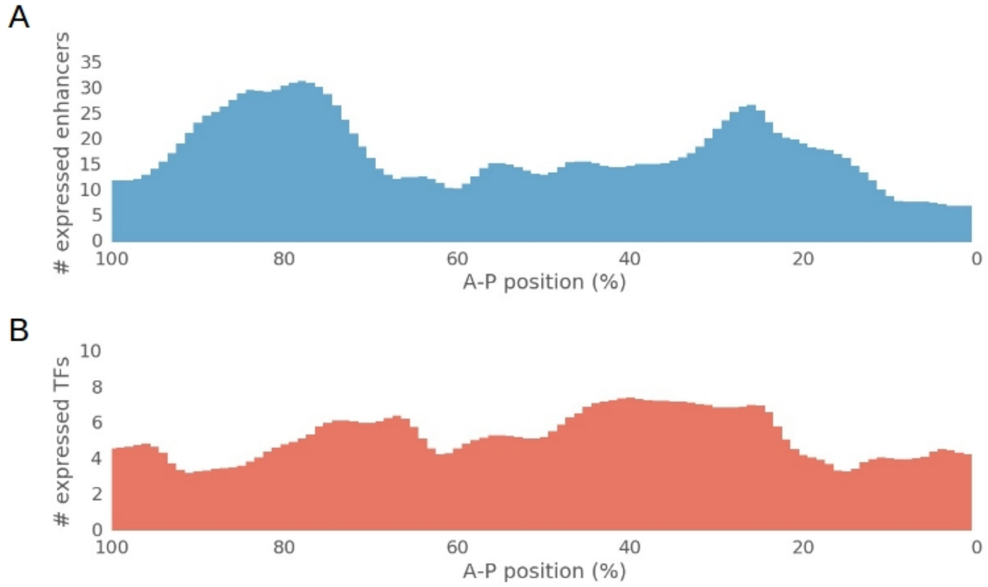


Figure 3.2: Sum of the expression profiles for (A) all 98 enhancers and (B) the 17 TFs, yielding the effective number of expressed enhancers/TFs at every position.

3.2 Transcription Factors

If enhancers are the sentences in the language of transcription control, transcription factor binding sites are the words. As with any language, a dictionary of words is necessary before one can start to identify grammatical rules. In order to compile a full dictionary for AP segmentation during mid-blastoderm, I searched the literature for regulatory TFs. I identified 17 potential factors, see table 3.1. Among them are 6 TFs that – to my knowledge – have never been used for modeling.

Most modeling efforts, e.g. [28, 27], use a set of 8 TFs: the two maternally provided activators bcd (bicoid) in the anterior and cad (caudal) in the posterior, the maternal repressor cic (capicua) as well as the gap gene gt (giant), hb (hunchback), Kr (Kruppel), kni (knirps), and tll (tailless). Together they drive large parts of the mid-blastoderm segmentation expression. I use this set, to which I refer to as the reduced TF set, for a baseline model.

In contrast to the reduced set, I define the 17 TFs as the full set or the expanded set. This set contains the additional gap factors btd (buttonhead), an activator in the head region [64], fkh (forkhead) and hkb (huckebein), two repressor in the embryo termini [65]. Furthermore, the two repressors run (runt) and slp1 (sloppy paired 1) are pair rule genes that show gap-gene-like behavior during early segmentation [66, 67]. The late gap factors D (Dichaete), Nub (Nubbin) as well as pdm2 (POU domain protein 2) come up late in the posterior and affect mainly pair rule genes [68, 69, 70]. Finally, D-Stat is a maternally provided, ubiquitous activator [71, 72].

TFs are able to differentially regulate transcription along the AP-axis because they themselves are expressed in patterns. High-resolution concentration measurements [73] are a source for TF concentration profiles along the AP-axis. For TFs lacking protein distribution data, I used mRNA expression as an alternative [74]. The concentration profiles are depicted in figure 3.4. Unlike the enhancer patterns, the TF concentrations profile are continuous, except for the factors, for which I derived the concentration from mRNA. This difference should be kept in mind for the evaluation of expression predictions because it has implications on the measure of prediction quality. Since predicted enhancer patterns are derived from a continuous input, they can be expected to be much smoother and less sharp than the measured ones. A perfect fit might not always be possible. Overall, the TFs are more evenly distributed over the AP axis than the enhancers, see figure 3.2 B. Hence, there is not a strong bias in the input that favors certain regions in the embryo.

The critical aspect of TFs are the binding specificities depicted by the PWMs. As discussed in section 2.1.1, TF binding sites are defined by a fuzzy logic, which allows for mismatches. Measuring binding specificities is a non-trivial task. Common methods are bacterial one-hybrid B1H [75], Footprinting [76] and Selex [77].

Most methods incorporate a selection step, in which unbound or weakly bound sequences get sorted out. E.g. B1H selects binding sites by making them necessary for the survival of bacteria colonies. Colonies with weak sites are less fit and get sorted-out. This selection step favors the consensus binding site disproportionally and leads to overly specific PWMs [78]. For this reason, it is common to decrease the PWM's specificity artificially by adding uniformly distributed pseudocounts, which serve as a constant offset for every entry in the PWM reducing the relative distance between the minimum and maximum. Especially when the binding site sample size is very low, many entries in the PWM will be zero without pseudocounts rendering many binding options completely impossible independent of the rest of the motif, which is regarded as undesirable [79]. Pseudocounts prevent this and allow the comparison of weak sites, e.g. in equation 2.7.

A method that comes close to the thermodynamic model of protein binding, is High Performance - Fluorescence Anisotropy HIP-FA [78]. HIP-FA was developed to measure the binding affinities of specific sequences directly in a high-throughput fashion. Typically, one measures the binding affinity of all single-base mutations of the consensus sequence to the TF. HIP-FA can measure affinities for weak as well as strong binders with high accuracy. Because the method takes weak binders into account, HIP-FA PWMs have in general a lower information content than alternative PWMs. That means that the binding motifs are less specific than other methods suggest. Nevertheless, the HIP-FA PWMs are superior as input for thermodynamic models in comparison to B1H and Footprinting PWM sets, see [78] and section 5.1. Therefore, HIP-FA PWMs are the default choice as input and I use alternative PWMs only if HIP-FA PWMs are not available.

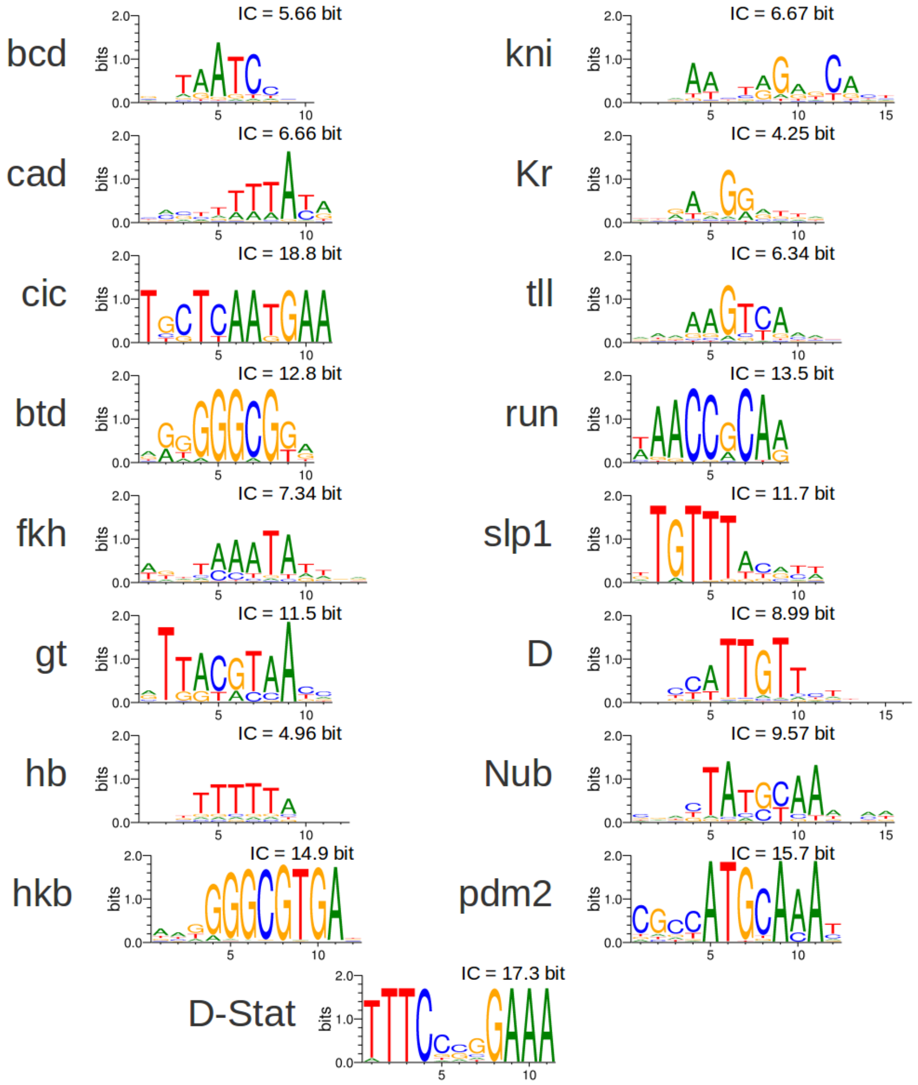


Figure 3.3: The TF's binding motifs as sequence logos. Also depicted is the information content. Alternative sequence logos are in figure 5.2

Table 3.1: All relevant TFs of the AP segmentation system.

Name	Stage	Type	Binding Domain
bcd	Maternal	A	Homeodomain
cad		A	Homeodomain
cic		R	HMG-box
D-Stat		A	STAT-domain
btd	Gap Factor	A	Zinc Finger
fkh		R	Winged Helix
gt		R	B-Zip
hb		R	Zinc Finger
hkb		R	Zinc Finger
kni		R	Zinc Finger
Kr		R	Zinc Finger
tll		R	Zinc Finger
run	Early Pair Rule	R	Runt-domain
slp1		R	Winged Helix
D	Late Gap Factor	A	HMG-box
Nub		R	POU-domain
pdm2		R	POU-domain

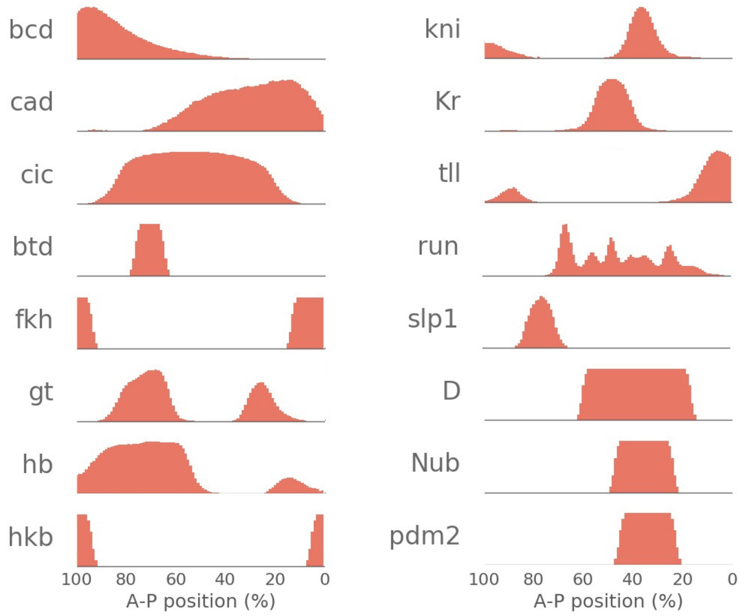


Figure 3.4: Concentration profiles of 16 TFs along the AP embryo axis. Not included is D-Stat, which is ubiquitously expressed. The run profile has already pair-rule characteristics. Substituting the profile for an earlier version does not alter the predictions noticeable.

Part II

Results

Chapter 4

Model Training

Many parameters of thermodynamic models are unknown because they are difficult or impossible to measure. The most important ones are the activatory potentials, the TF concentrations, the absolute binding affinities, the cooperativity parameters and the basal transcription rate. The first four parameters are TF-specific, in contrast to the basal transcription rate, which is a global parameter because it affects all enhancers.

Note that the TF concentration c and the absolute binding affinity α appear together in all equations. Hence, it is possible to split up the concentration into a relative factor, ranging between 0 and 1, and a scaling factor $c = c_{\mathrm{rel}} \cdot c_{\mathrm{scale}}$. I fuse the scaling factor to the affinity resulting in a single parameter $c_{\mathrm{scale}} \cdot \alpha$. This combined parameter, for the sake of simplicity called the absolute affinity α, describes the TF binding strength. The relative concentration, in the following called the concentration c, describes differences in protein abundance. It is comparably simple to measure across the embryo and serves as input data.

If not stated otherwise, homotypic cooperativity is always incorporated. Hence, every TF is associated with 3 unknown parameters. All these parameters plus the basal transcription rate have to be estimated. This is done by searching for the parameter setting that best reproduces the measured expression profiles. The science of training parameters to data is called parameter optimization. The basic principle of parameter optimization is to start with an initial set of parameters and then apply small alterations to them while monitoring the prediction quality of the model.

The two fundamental steps, measuring the prediction quality and updating the parameters, will be the topics of the following two sections. The remainder of the chapter will be about assessing the result of the parameter training and strategies to improve the procedure for thermodynamic models.

4.1 Assessing Prediction Quality - Objective Functions

Objective functions measure quantitatively the quality of the model prediction. They compare the measured data to the prediction of the model and express the deviation as a score[1]. Objective functions enable us to compare and rank different models or sets of model parameters.

The measured data is in this case the expression profile of the enhancer along the anterior-posterior embryo axis. As model resolution, I take 1% egg length, resulting in 100 predictions along the axis. The model identifies occupancy of the core promoter as expression rate. Therefore, the predictions are probabilities. The measurements are binary expressed/repressed readouts from staining experiments. Neither prediction nor measurement represents actual mRNA concentrations. Everything that counts is the relative differences along the AP axis, while the scale of the prediction is irrelevant. For this reason, an objective function that is scale-invariant is preferable.

In the following, p_i is the predicted expression level and m_i the measured expression with i the position along the embryo axis.

4.1.1 Sum of Squared Errors

A commonly used objective functions is the sum of squared errors (SSE) [81].

$$\mathrm{SSE}_0 = \frac{1}{N}\sum_{i=0}^{N}(m_i - p_i)^2 \tag{4.1}$$

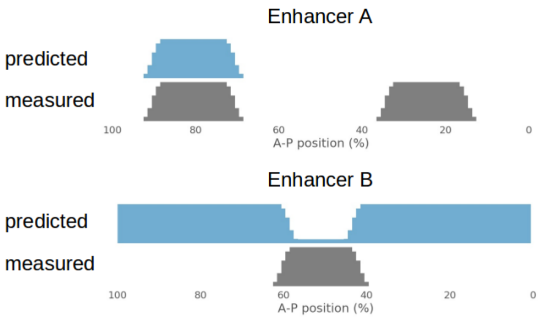
The reason for choosing to square the differences is so that the result is always positive and is easy to derivate. The SSE is on its own not scale-invariant but can be modified to ignore scales. This is possible by scaling the prediction retroactively by a factor β, which improves the fit to the measurement.

$$\mathrm{SSE} = \frac{1}{N}\sum_{i=0}^{N}(m_i - \beta p_i)^2 \tag{4.2}$$

I calculate β by requiring that it should minimize the SSE score. A simple analytical calculation yields:

$$\beta = \frac{\sum_i m_i p_i}{\sum_i p_i^2} \tag{4.3}$$

The main weaknesses of SSE for this type of data is that it scores deviation between patterns rather than differences in shape. This can lead to unintuitive scores rendering a comparison between results difficult. Figure 4.1 gives an example: shown are two enhancers and their flawed prediction. The prediction of enhancer A fits the anterior domain perfectly but misses the expression in the posterior domain completely. This is a realistic scenario, which happens when a posterior activator is missing. The prediction for enhancer B is almost perfectly anti-correlated to the measurement and is, therefore, obviously wrong. Surprisingly, the prediction for B has a better SSE score than A, although, by intuition, A should score better. The reason is that the scaling factor β reduces B's prediction level so



Score	Enhancer A	Enhancer B
SSE	0.19	0.17
Corr	0.6	- 0.95

Figure 4.1: Exemplary comparison of SSE and correlation as objective functions.

that the error in most parts of the embryo is minimal. Hence, the major source of error is B's expression domain, which is similar in size to the missed domain in A.

One can further modify the SSE by weights w_i, if one wants the score to punish deviations in certain positions more than others, e.g. where measurements are more precise or in the expressed domains, as a simple attempt to solve the problem in figure 4.1.

$$\mathrm{SSE}_w = \frac{1}{N}\sum_{i=0}^{N} w_i(m_i - \beta p_i)^2 \qquad \beta = \frac{\sum_i w_i m_i p_i}{\sum_i w_i p_i^2} \tag{4.4}$$

Reliable measurement precisions are not recorded for this data and can not serve as weighting. Based on experience, I advise against the use of manually chosen weights. Applying such weights deteriorates the prediction, judged on qualitative as well as quantitative measures (alternative quality measures like Precision-Recall plots, data not shown). The reason is that hand selecting weights introduces biases, which are, in general, unfavorable for the training process.

4.1.2 Pearson Correlation

An inherent scale free objective score is the Pearson correlation [82].

$$C = \frac{\sum_{i=0}^{N}(m_i - \bar{m})(p_i - \bar{p})}{||m - \bar{m}||_2 ||p - \bar{p}||_2} \tag{4.5}$$

Here, $\bar{x}$ is the mean of x along the AP-axis and $||x||_2 = (\sum_i x_i^2)^{1/2}$ is the norm of x. A perfect fit has a correlation of $C = 1$, while $C = -1$ means that the two patterns are perfectly

[1] For this reason, the objective function is often called loss or cost function, e.g. [80].

anti-correlated. In contrast to the SSE, the correlation scores shape and not deviation. The benefit of this interpretation can be seen in figure 4.1. The Pearson correlation can easily distinguish the decent prediction of enhancer A from the bad prediction of enhancer B. For this reason and because of its simplicity, the Pearson correlation is the objective function of choice in this thesis.

Note that in contrast to the SSE, which has to be minimized to fit the data, the correlation has to be maximized. By convention, most optimization algorithms aim to minimize the objective function. Therefore, I use the negative correlation as the objective function. Nevertheless, results will be reported as positive correlations.

Of course, it is also possible to weight the correlation like the SSE.

$$C = \frac{\sum_{i=0}^{N} w_i(m_i - \bar{m})(p_i - \bar{p})}{||m - \bar{m}||_{w,2}||p - \bar{p}||_{w,2}} \tag{4.6}$$

The same caveat as for weighted SSE applies to the weighted correlation. If there is no objective weighting scheme, e.g. measurement errors, simplicity should be favored.

4.1.3 Precision-Recall Plots

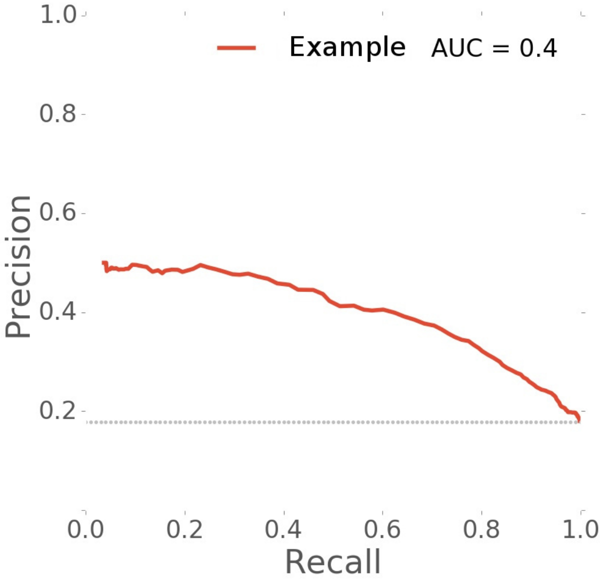


Figure 4.2: Example of a PR-plot. The dotted line indicates the precision of a random guess. By demanding a higher recall, the predictions become less precise. The area under the curve AUC is a summary of the curve's position in the plot.

Not as an alternative objective function but as an independent quality measure, consider the Precision-Recall-plot (PR). The main idea is to define a threshold above which an enhancer counts as active in that position. Much like the ROC-plot, one draws a PR-plot by scanning the full range of thresholds and plotting the respective precision and recall.

The precision is the ratio between the true-positive predictions and the total number of positive predictions. It is the probability that a predicted enhancer activity is correct. The recall is the ratio between true-positive predictions and the total number of positive measurements. It is the fraction of correctly identified expression domains.

The PR-plot carries the same information as the ROC-plot, however, displays it better when there are considerably fewer positive than negative data-points [83]. This is beneficial for this data because most enhancers are expressed in small domains. A good prediction is characterized by having both, high precision and high recall. Figure 4.2 shows an example of an PR-plot. PR-plots help to assess the quality of a prediction by eye and will serve as an alternative quality measure.

4.1.4 Integrating Multiple Enhancers

One enhancer is just a single aspect of the data. Normally, thermodynamic models deal with many enhancers and have to fit all of them at once. The objective function has, therefore, to integrate the scores of multiple enhancers s_i. The immediate solution, which is also bias free, is the average of all enhancer scores. Simple averaging guarantees that all enhancers contribute equally to the total score and hence that the optimization algorithm makes equal effort to fit all enhancers.

$$S = \frac{1}{N} \sum_{i=0}^{N} s_i \tag{4.7}$$

However, there are cases in which a different weighting scheme can be reasonable. In theory, there are dozens of ways in which one can attribute a disproportional share of the total score to single enhancers. One can achieve this by a weighted average in which important enhancers receive a bigger weight. One could put more weight on enhancers from certain 'reliable' sources or enhancers with underrepresented features. An example of the latter would be enhancers expressed at the tips of the embryo. I advise against this integration scheme because a general rule for objective functions is that simplicity is beneficial and preset biases are disadvantageous. Unintuitive and hand-optimized objective functions are difficult to interpret and likely lead to unwanted behavior during parameter optimization.

Capped Score Imagine enhancers that are problematic or have not been measured correctly. They will receive a bad score, e.g. a negative correlation. The algorithm fails to predict these enhancers. However, the difference between a score of -0.3 and -0.7 is more or less pointless. On the other hand, an increase of correlation from 0.3 to 0.7 means usually the difference between a fair and a great prediction. By setting a minimal score of 0, the optimization algorithm is encouraged to improve the enhancers in a regime where changes matter.

$$S = \frac{1}{N} \sum_{i=0}^{N} \max(s_i, 0) \tag{4.8}$$

Alternatively, one can argue that there is little gain above a certain score and everything beyond is certainly over-adaptation, especially since the data quality is limited. Again, the interesting range of scores distinguishes between poor and good predictions. A capped score would stay in this range.

$$S = \frac{1}{N} \sum_{i=0}^{N} \min(s_i, 0.75) \tag{4.9}$$

The examples here are for the correlation as the objective function, but can be adjusted to the SSE.

Median Score The argument for capped scores was that the optimization algorithm should concentrate on the poor but promising predictions. A natural way to do this is to move away from a score average towards the median of scores. The median is the one score that separates the good from the bad half of the predictions. Any change in parameters improves the median only if it results in a net improvement for the majority of enhancers. Concentrating on a handful of high-gain enhancers is, therefore, discouraged.

Vertical Correlation The expression patterns of our enhancer collection are not equally distributed over the AP axis. For instance, many enhancers are expressed in the head of the embryo and very few at the tips of the embryo, recall figure 3.2. All aforementioned modifications do not change the fact that every enhancer is equally important if one does not want to adjust that by hand. Imagine all predictions to be collected in a matrix where every enhancer is one row and every position in the embryo is a column. Instead of calculating the correlation along the horizontal axis, I propose to calculate the correlation between measurement and prediction for every position and average the result for all positions. By using this vertical correlation, all positions contribute equally to the total score.

There is one caveat. The resulting score is not scale-free anymore. It is conceivable that enhancers have large differences in expression strength. By scale-free scoring the enhancers one by one, it is possible ignore the absolute expression rate. The vertical correlation builds on the assumption that the rates are similar.

4.1.5 Comparison

What is the best objective function to use for the parameter training? Comparing objective functions is conceptionally difficult because they are used for two different purposes. First, as a guide for the algorithm to train the parameters to the measured data (the training objective function) and, second, as a quality assessment after the parameter training (the test objective function). One expects that models trained with one objective function excel at the quality measure based on the same objective function. Hence, the test objective function loses its core quality of being objective in this case.

It is not possible to solve this problem. However, comparing the results across different quality measures shows that some training objective functions are clearly better suited for the parameter optimization. I repeated the model training with 5 different training objective functions. The results in table 4.1 are quality measures for the predictions on unseen data (cross validation) using training techniques from the following sections in this chapter. Immediately apparent is that the median correlation as well as the vertical correlation are worse than the other training objective functions regardless of the quality measure. Of the remaining three training objective functions, the capped and the averaged correlation yield almost equal scores. Using SSE results in similar scores only for 3 out of 4 quality measures. This in combination with the conceptional considerations of 4.1.1 suggests to reject SSE as training objective function.

The capped correlation follows the same concept as the average correlation but with some adjustments to improve the prediction. Since the difference is insignificant, I favor the simpler and more intuitive quality measure. In the following, the training as well as test objective function of choice will always be the Pearson correlation averaged over all enhancer results.

Table 4.1: Test scores measured with different objective functions (column) after multiple 10-fold cross validation runs using various objective function for the parameter training (row). The objective functions are CORR (averaged correlation), SSE (averaged sum of squared errors), CCORR (capped correlation), MCORR (median correlation), VCORR (vertical correlation), and AUC-PR (area under the precision recall curve). Marked are the best results for every test function.

test func.	training func.				
	CORR	SSE	CCORR	MCORR	VCORR
CORR	0.411 ± 0.01	0.387 ± 0.014	0.407 ± 0.01	0.355 ± 0.038	0.199 ± 0.014
SSE	0.29 ± 0.002	0.289 ± 0.003	0.289 ± 0.002	0.294 ± 0.003	0.314 ± 0.002
MCORR	0.507 ± 0.036	0.515 ± 0.026	0.5 ± 0.036	0.451 ± 0.046	0.189 ± 0.025
AUC-PR	0.4 ± 0.02	0.38 ± 0.02	0.4 ± 0.02	0.36 ± 0.02	0.26 ± 0.02

4.2 Model Selection

Although the objective function can score the quality of a model fit, it is not possible to compare models simply based on how well they fit the training data. Complex models, i.e. models with many parameters, usually outperform simple ones during parameter training because they have more freedom and can better adapt to variations in the data. In the worst case, a model with enough parameters could simply memorize the presented data and would receive a perfect score. The true indicator of a good model is rather how well it predicts previously unseen data, meaning how well it distinguishes between general features and random noise in the training data. This is expressed in the test score that is defined as the score on an independent data set, called test data. The test data must not be used for the training of the model, otherwise, it would thwart the idea of testing on unseen data.

For instance, take again the overly complex model that simply memorizes the random noise in the training data. Since it does not learn any general features of the data, its predictions on new data are more or less random and would score poorly on the test data, resulting in a low test score. Such a behavior is called overfitting, see for example [80, 81].

4.2.1 Cross-Validation

In order to derive general features from the training data, the model has to be confronted with sufficient training data. Simultaneously, the test data has to be sufficiently diverse and as unbiased as possible, too. In this case, the data, identified enhancers with measured expression, is sparse. Since it is not possible to put an enhancer in the training data as well as in test data, it is necessary to find a balance between both data sets. Cross-Validation (CV) is a common solution, which ultimately enables testing on the whole data set.

For CV, one divides the available data into N disjunct fragments of similar size. The parsing should be random in an unbiased fashion. One then does N parameter optimizations on $N-1$ data fragments and tests the result on the left-out fragment. The final prediction is the concatenation of all test predictions, for which the full test score can be calculated. The obvious advantage is that one does not need to worry about the selection of the test data because every available data point (enhancer) is once in the test data. The disadvantage is that one has to train the model multiple times, which costs computation time and yields multiple parameter results that are not necessarily consistent.

The last aspect is problematic if the parameters are used for a follow-up analysis, where a the single optimal parameter setting is preferable. However, as a positive side effect, this enables to test the sensitivity of the model towards variations of the training data. If the optimization yields similar parameters in every CV-iteration, the model, and its training process are stable and therefore more credible. Figure 4.3 depicts the spread of the activatory potentials after a ten-fold CV training. While the parameter training is very stable for the TFs bcd and hb, the prediction for gt depends strongly on the training data. However, the results are consistent in the sense that the predicted role of the TFs (activator/repressor) is always the same.

Different parsings yield slight variations in the test score even when applying the same training strategy. Especially for small datasets, it is possible that the parsing is biased, e.g. all enhancers of a certain type end up in the same data fragment. Such unwanted clustering could not only impede the parameter training but also favor certain models. For this reason, I repeat all analysis steps 10 times with 10 independent parsings. The final test score is simply the average score for all repeats.

The number of data fragments determines the size of the training data. In the case of leave-one-out LOOCV, every data point is its own fragment. Although LOOCV has the advantage of an almost full dataset allocated to training, it also takes the most computation time. Assuming that the time for a single parameter optimization run is solely proportional to the amount of training data, it is possible estimate the necessary computational resources

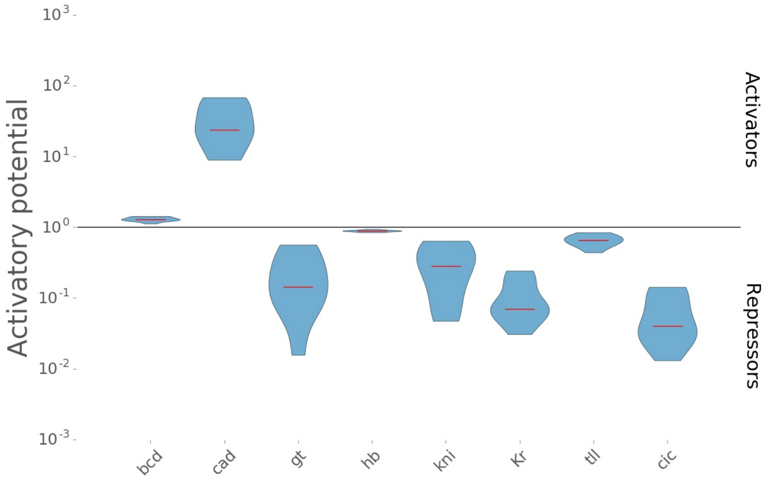


Figure 4.3: Example of the spread of parameter training results. The violin plot represents the distribution of the values. Marked in red is the median. TF predicted to be activators have an activatory potential above 1.0 and can be found in the top half of the graph. Repressors are in the bottom half.

for a N-fold CV.

$$T_N = (N-1)T_{\text{full}} \tag{4.10}$$

I tested a 5 and a 10-fold CV set-up. To do so, I drew 10 parsings for each 5 and 10-fold CV and trained a thermodynamic model with the reduced parameters (8 TFs) and with the extended set of parameters (17 TFs). Not surprisingly, 10-fold CV yields on average noticeably better test scores, however, the difference is not significant, see table 4.2 and in more detail B.3. I nevertheless decided to use 10-fold CV as the standard for all following experiments.

Table 4.2: Test scores (correlation) for 5 and 10-fold CV as well as 10-fold CV with clustering based parsing (c10) each for a small set of parameters (8 TF) and a large parameter space (17 TF).

CV	5	10	c10
8 TF	0.34	0.351	0.353
17 TF	0.36	0.369	0.378

As an alternative to a completely random parsing of the data fragments, I tested a bias-reducing technique. The idea is to prevent an uneven distribution of enhancers of a similar type. To do so, I split the enhancers into six clusters based on their binding site content using Ward's hierarchical clustering method. I distributed the enhancers of the same cluster equally among all data fragments in a random fashion. I did this 10 times and trained the model separately on all 10 parsings. The resulting test scores are noticeable — however not significantly — better than for bias-unaware parsings B.3. For all further experiments, I use the bias-reduced 10-fold CV because they are in all other aspects equal to the completely random parsings.

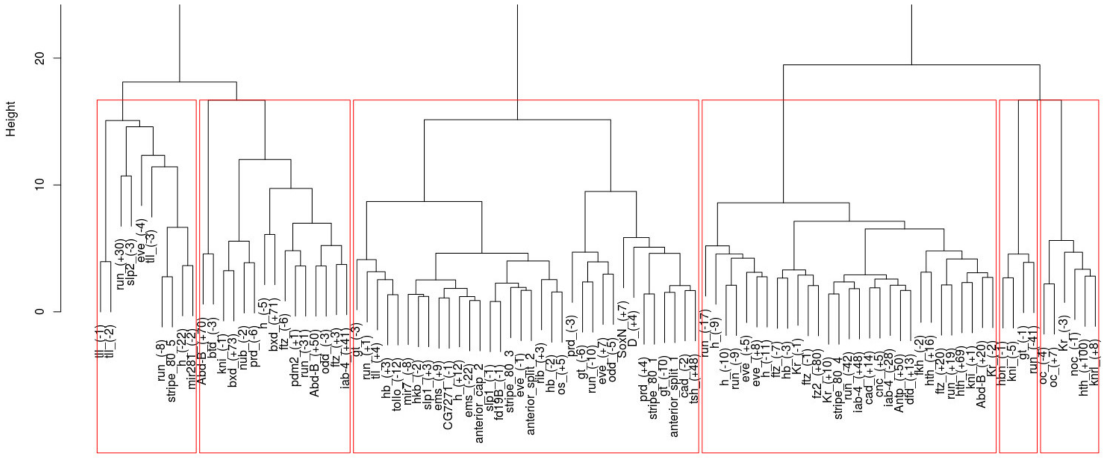


Figure 4.4: Enhancer dendrogram based on TF content. The six clusters are marked in red. The root, where all branches meet, is cropped from the figure.

Having decided on the CV-method, I did not redraw new parsings for every experiment and instead reused the same 10 parsings making the results more comparable. By doing this, I can compare model results using the Wilcoxon signed-rank test, which assumes that results are paired, i.e. performed under the same conditions on the same data.

4.2.2 Parameter Regularization

Relevant for model assessment is the test score rather than the training score. In fact, improving the training score beyond a certain point can lead to a deterioration of the test score. This effect is called overfitting and is one of the main problems of model selection. Given enough freedom, a model can adapt extremely well to the training data. In this case, the model tries to fit the noise in the data (e.g. spurious and weak binding sites, exact enhancer expression etc.) and neglects to learn general features. Extreme parameter values are often an indicator of overfitting because the optimization algorithm is free to exploit any random pattern in the data as long as it improves the training data.

The idea behind parameter regularization is to prevent the algorithm from fitting the noise in the data in order to improve test scores. Here, I describe two simple techniques that reduce the algorithm's freedom during model training by limiting the available parameter range. These approaches are similar to the idea behind Lasso [84] and ridge regression [85], which are regularization methods for linear models.

First, I set a range cut-off for all parameters. This is done by transforming every parameter with a bijective and monotonous function t.

$$t: \qquad [\min, \max] \rightarrow [-\infty, \infty]$$

$$p \mapsto \log\left(\frac{\log(p) - \log(\min)}{\log(\max) - \log(\mathrm{p})}\right)$$

The optimization algorithm works with the transformed parameters $p' = t(p)$, which can take any value. The back-transformed parameters are always positioned between the predefined boundaries. If for example, the optimization increases $p' \rightarrow \infty$, the real model parameter p will slowly converge to its maximum. For this reason, this approach is called a soft boundary, in contrast to a hard boundary, which is reachable.

Second, I apply a parameter penalty by adding an extra term to the objective function.

$$\text{penalty} = \lambda \cdot ||p|| \tag{4.11}$$

λ is a scaling parameter called the absolute penalty. $||p||$ is the parameter norm. It scales with the parameters' deviation from a neutral setting (absolute affinities and cooperativities equal zero and activatory potentials equal 1). A TF with parameters in the neutral setting are neither repressors nor activators and do not bind the DNA. It is as if the TF were not present.

If the optimization algorithm includes a certain TF into the model or increases the TF's importance, the gain in score has to outweigh the penalty, otherwise, the total objective function would deteriorate. Conversely, if the presence of a TF does not improve the prediction, the optimization algorithm has an incentive to push its parameter back to the neutral setting. The gain is twofold; on the one hand, it reduces the parameters to a reasonable range by applying a soft boundary. On the other hand, it makes overfitting harder because adapting to noise becomes unfavorable.

Consider the situation in which a TF has no binding sites in the training data but several in the test data. Without a penalty, its parameters would move together with the other parameters and finish in a random setting because they do not affect the training result. However, the resulting parameters would likely be disadvantageous for the prediction on the test data. The parameter penalty prevents this effect.

There are multiple options for the parameter norm. The L1-norm is just the sum of the absolute parameter values with the exception of the activatory potential for which I add $max(\alpha, 1/\alpha) - 1$ because a repressor with $\alpha_R < 1$ equals in strength an activator with $1/\alpha_R$. The alternative is the L2-norm, which works the same way, but takes the sum of the squared values. A direct comparison between the two options is difficult because the effect depends heavily on the absolute parameter strength λ. In general, the L2-norm penalizes extreme values more than weak ones. The L1-norm is more effective when it comes to pushing parameters back in the neutral setting. For this reason, I choose the L1-norm.

The choice of the absolute parameter penalty λ is crucial. It is a so-called hyperparameter because it affects the final result without being trained by the optimization algorithm. If the penalty is too weak, the technique is inefficient. If the penalty is too strong, it inhibits the algorithm from identifying relevant patterns. There is no obvious choice for λ. One could select λ by trial and error and chose the one value that yields the best test result. This, however, is very controversial because it is a not so obvious case of training parameters on test data. I am going to discuss a technique for hyperparameter training in a later section 4.5. The alternative is to guess a value based on typical parameter norms.

4.2.3 Impact Score

The goal of thermodynamic models is not only to predict expression patterns but to understand the underlying mechanics of the transcription control. The objective function assesses the quality of the prediction but it does not reveal how the model works. One can inspect the thermodynamic parameter after training to see which TFs are most likely relevant for the prediction, however, it is difficult to conclude how much impact a TF has on the basis of the raw parameter values. E.g. different TFs have their unique binding weight distributions, might be blocked by other factors etc.

To render TF predictions comparable, I introduce the impact score. The basic concept follows the idea of an *in silico* knock-out experiment. The impact score $\mathcal{I}$ of a single parameter p is the difference in test score C between the unaltered "wild type" prediction and a modified "mutated" prediction in which the respective parameter is set to its neutral position $\varnothing$ (activatory potential equals 1, absolute affinity and cooperativity equals 0).

$$\mathcal{I}(p) = C(p) - C(p = \varnothing) \tag{4.12}$$

The impact score is positive if changing the parameter deteriorates the prediction, suggesting an overall positive influence of the underling mechanism, and is negative if the test score of the mutated model is better than the full model. The latter case can be an indicator for overfitting.

The impact score of a TF is similarly defined, with all parameters relevant to this TF set to the neutral setting. It quantifies how the prediction would change if the TF were not present. One can calculate the impact score for the full prediction (see figure 4.5) but also per enhancer to see where the TF drives expression (see for example appendix C.6).

In contrast to *in vivo* knock-out experiments, we have an in-depth control of the model and can use the impact score to dissect it on every conceivable level. It is possible to extend the concept of an impact score to hyperparameter (e.g. cooperativity range), groups of binding sites (e.g. strong sites) and basically every mechanistic detail of the model (e.g. steric hindrance). The only caveat is that the impact of the discussed feature has to be strong and consistent enough to be visible above the noise of different predictions. For example, the impact of a single binding site is likely too irrelevant and too variable to be measurable.

An alternative to the impact score is a classical sensitivity analysis [86]. The local parameter sensitivity is defined as:

$$\mathcal{S}(p) = \frac{p}{C}\frac{\partial}{\partial p}C \tag{4.13}$$

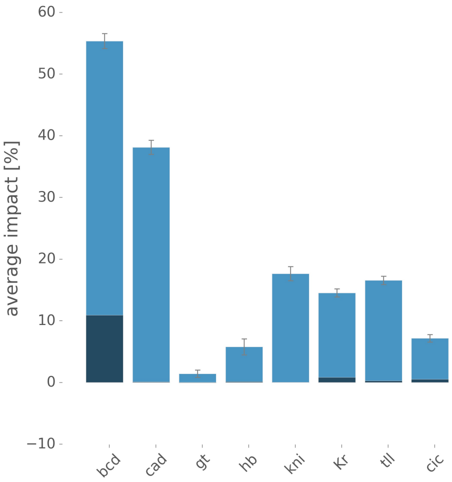


Figure 4.5: The impact score of single TFs as percentage of the total test score (light blue) and the fraction of impact from homotypic cooperativity alone (dark blue). The error bars are based on the results for five different CV-parsings.

$\mathcal{S}$ measures the susceptibility of the model to parameter p at a certain position in the parameter space. Note that both the sensitivity as well as the impact analysis are local in their nature. Both scores depend not only on the value of the single parameter itself but also on the state of all other model parameters, i.e. they analyze a specific parameter result rather than the model in general. The scores neither capture the general susceptibility of the model to a parameter nor do they account for possible compensation or enhancement effects of other parameters. I prefer the impact score because of its biological motivation (mutation/knock-out), which makes it in my opinion easier to interpret.

There are multiple possibilities to extend the local sensitivity $\mathcal{S}$ or impact analysis to a global analysis [87]. Most methods sample from the full parameter space and are therefore computationally very expensive. They additionally extend the concept of a first-order analysis as in equation 4.13 to a higher-order analysis by varying multiple parameters. Dresch et al. performed a global analysis of a thermodynamic model with nine parameters and find modest second-order effects [31]. The second-order effects were especially high for parameters which are naturally connected to other parameters, e.g. cooperativities.

4.3 Parameter Optimization Algorithms

There is a multitude of parameter optimization algorithms. Their basic function is to find a parameter optimum, i.e. a set of parameters that minimizes the objective function. They differ in their strategy how to move through the multi-dimensional parameter space and the stopping criteria, which assesses whether a set of parameters qualify as a satisfactory solution. Figure 4.6 illustrates the trajectories of 16 parameters of an exemplary optimization run.

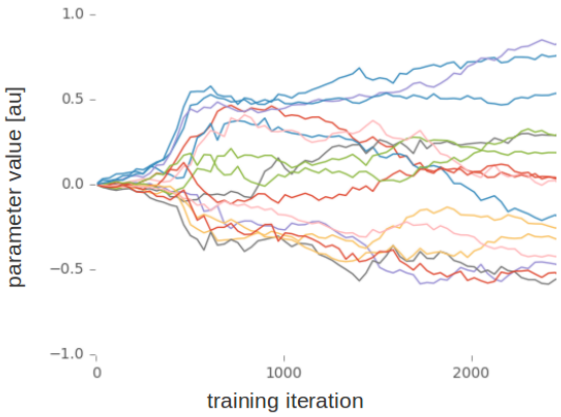


Figure 4.6: Example of parameter trajectories. Depicted are 16 parameters that get trained simultaneously. Three typical stages of model training are visible. First, an exploratory phase, in which the direction of the training is not determined yet, then a fast adjustment close to the final position, followed by a long fine-tuning phase.

Finding the best strategy depends on the structure of the landscape laid out by the objective function. Certain factors complicate finding an optimum; Among them are a high-dimensional parameter space and an ill-conditioned[2] problem setting. Both factors are typical for thermodynamic models, justifying an examination of the topic.

In more detail, I compared two types of algorithms: local and global. Local algorithms search for a local optimum, which is any solution that minimizes the objective function within a local neighborhood. Global optimization algorithms aim to find the best possible solution in the whole parameter space, which is called the global optimum. Global algorithms do not guarantee that their solution is indeed a global optimum, however, they apply strategies that prevent the algorithm from getting stuck in the first encountered optimum and to further explore the parameter space.

4.3.1 Local Optimization

Two local optimization algorithms are commonly used for many optimization applications, among them thermodynamic models [28, 27, 26]: Nelder-Mead Simplex [88] and (quasi-Newton) gradient descent [89, 90].

A simplex in n dimensions is a geometrical object that has $n+1$ vertices (line, triangle, tetrahedron etc.). The idea behind the Nelder-Mead algorithm is to define a simplex in the parameter space, where every vertex is a potential solution. In every iteration, the algorithm moves the simplex by replacing the vertex representing the worst solution.

[2] Ill-conditioned means that small changes in the input (parameter space) lead to large variations in the output (objective function score).

Occasional shrinking narrows the scope of the algorithm until an optimum is found. The strength of the simplex method is that it does not require the derivatives of the objective function.

The basic concept of gradient descent is to start from an initial point and move in the direction of the negative gradient of the objective function. Since the gradient indicates the direction of steepest ascent, moving in the opposite direction reduces the objective function, if the step size is not too large. In few words, quasi-Newton methods improve navigation through the parameter space by additionally approximating the second derivative of the objective function, also called the Hessian matrix, which describes the local curvature. Gradient methods are very efficient if the computation of the gradient is easy and fast. For thermodynamic models, this is impeded by the fact that calculating derivatives is very time-consuming.

As a local optimization strategy, I follow He et al. [28] and try both the Nelder-Mead Simplex and a gradient descent[3], both implemented into C++ via the GNU scientific library [91]. Additionally, I try a strategy that alternates between both algorithms. In order to prevent the algorithms from getting stuck in the first local optimum, I perform three consecutive training iterations with small parameter adjustments between them. In this fashion, the algorithm gets kicked out of the local optimum and is free to further explore the parameter space for better optima, which from my experience is very beneficial (data not shown). Note this training strategy is completely deterministic. Under the same initial conditions, two training runs will always yield the same result.

4.3.2 Global Optimization

Global optimization strategies usually incorporate a stochastic component that allows the algorithm to jump out of a local minimum enabling a more thorough parameter exploration, while still following a global trend to minimize the objective function. The balance between the exploitative and exploratory behavior determines the scope of a global optimization algorithm.

Our global optimization algorithm of choice is the Covariance Matrix Adaptation - Evolutionary Strategy CMA-ES [92]. The biologically motivated evolutionary strategies always consider a population of solutions, which move as a group through the parameter space. Every iteration is a new generation that gets sampled based on the previous generation's fitness, i.e. the objective function score. For CMA-ES, the optimal population size ν depends logarithmically on the number of training parameters n. A rule of thumb is:

$$\nu = 4 + \lfloor 3 \log(n) \rfloor \tag{4.14}$$

The CMA-ES algorithm samples a new generation from a multi-dimensional normal distribution. The covariance matrix defines the spread of the population across the parameter space dimensions. The algorithm constantly adapts the covariance to optimally navigate the parameter landscape much like quasi-Newton methods learn the Hessian matrix. The

[3] In contrast to He et al., I use the quasi-Newton Broyden-Fletcher-Goldfarb-Shanno algorithm [90].

initial population spread σ defines the scope of the algorithm and serves as adjustment parameter for the above-mentioned exploration/exploitation balance.

The CMA-ES algorithm has some advantages. By adapting the covariance, the algorithm is optimally suited for ill-conditioned optimization problems. In theory, the algorithm needs more function evaluations than local methods, however, it is easily parallelizable. Since a whole generation can be evaluated at the same time on a handful (10-16) of nodes, CMA-ES has in my experience a speed advantage (data not shown).

4.3.3 Benchmarking

Suleimenov et al. have benchmarked a local and a global parameter optimization strategy [30]. They constructed synthetic data that looks like Drosophila enhancers, but with predefined thermodynamic parameters. In this fashion, they were able to conclude that a global strategy is superior to a local one for high-resolution input data. For low-resolution expression data, they found that a global strategy provided no benefit justifying the additional computation resources.

In contrast to Suleimenov et al., I tested local and global strategies thoroughly with real enhancer data. I do not use synthetic data and thus do not know the true parameter values. For the model, it is important whether a strategy improves the test scores. I performed 10 independent 10-fold CV training runs with CMA-ES as a global strategy and two local strategies (simplex and simplex mixed with gradient). I repeated the analysis for a model with 17 TFs and with a reduced model with 8 TFs. I found the strategy involving the gradient method to be significantly worse than both the pure simplex method as well as the global strategy, see table 4.3 and appendix B.4. I also found that CMA-ES yields noticeable better test scores than the simplex method but only for 17 TF, i.e. for a model with many parameters is the improvement significant (Wilcoxon signed-rank test $p \leq 0.01$).

Table 4.3: Test results for 8 TF (25 parameters) and 17 TF (52 parameters) trained with a gradient-based method, Simplex, and CMA-ES.

	Gradient	Simplex	CMA-ES
8 TF	0.257	0.363	0.368
17 TF	0.329	0.391	0.41

Figure 4.7 A depicts the results of B.4 as an enwrapped PR plot. Especially for high recall values, the global results are significantly more precise. Does the global strategy overfit less (maybe due to an earlier stopping criteria) or does it sample the parameter space better? Part B shows a scatter-plot of training scores. The x-axis shows the result for the local strategy, the y-axis the global strategy. Every point represents one dataset that was used for training and scoring. It demonstrates that the the global strategy's improvement of test score is preceded by already better training results. In conclusion, CMA-ES searches the parameter space more thoroughly than local strategies and finds better parameter solutions, which in the end also score better on the test data.

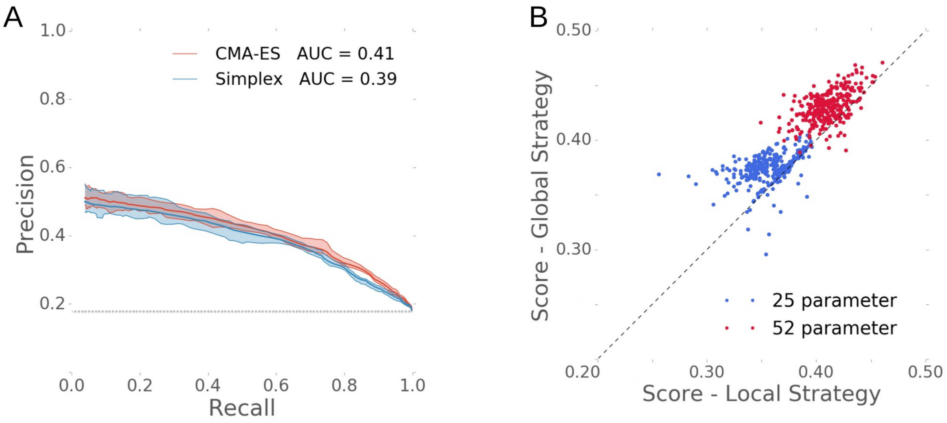


Figure 4.7: (A) enwrapped PR plot of 10 CV test results. (B) Scatter-plot of training results on 100 different training sets for a local strategy (Simplex) and a global strategy (CMA-ES).

CMA-ES is a stochastic algorithm. Thus, training runs are not reproducible and two runs on the same data will yield different solutions. I checked how much variation is possible and repeated the above mentioned training. While the local strategies would yield the exact same solution, we see some differences for the repeated CMA-ES run. Nevertheless, averaged over 10 CV parsings the differences vanish, see table B.4. It is therefore crucial that one test the model on multiple CV parsings to average out outliers.

Although the repeated CMA-ES training run yields on average the same result, there are noticeable differences for single CV parsings. If these differences came from a suboptimal training, it would be possible to improve the overall result by performing a more thorough parameter search. To test this, I started a training run with two full CMA-ES iterations and selected the result with the better training score. This multitrain strategy yielded similar test scores as both single-iteration training runs, see appendix B.4. The difference in average test score is negligible. Since the multitrain strategy takes twice as much computation time, I retained the single-iteration strategy.

4.4 Data Augmentation

The amount of training and test data is crucial for any modeling endeavor. Especially for complex models with many parameters is a rich training set extremely important because these models are prone to overfitting. Novel enhancers, the essential training data, are difficult to identify and measure. Additionally, there is a limited number of naturally occurring enhancers in the segmentation network. Thus, a huge increase in training data from new experiments is unrealistic.

The basic concept of data augmentation is to artificially increase the amount of training data by transforming and distorting existing data. The technique is often used in the context of image analysis, where images get cropped or rotated to generate new data points [93, 94]. Alternative approaches are to add noise [95] or, for sequence data, to simulate evolutionary differences [96].

The goal is always the same: to confront the optimization algorithm with additional noise so that it is forced to learn more general features and can not exploit random patterns in the real data. Much like the parameter penalty, data augmentation raises the cost of overfitting.

I tested three different approaches of generating new data, which I introduce in the following.

4.4.1 Simulated Evolution

In order to augment the training data, I generated three altered versions of the enhancers by introducing single-nucleotide variants into the sequence. This increases the amount of training data 4-fold. I did not change the expression profile. This is justified because evolutionary comparisons between closely related *Drosophila* species reveal strong differences in the enhancer sequences but a high conservation of expression patterns [19, 97]. I tested a 5%, 15% as well as a 25% selection rate. I chose a maximal uninformative approach in which a base that got selected for mutation changes into any base with equal probability. Note that the mutation rate is smaller than the selection rate because 25% of the mutated positions will select the original base (3.75%, 11.25%, 18.5%).

I found no difference in test score between the augmented and the unaltered prediction runs for any mutation rate, see table B.5. If anything the average test score decreased slightly. Although the sequence alternations did not improve the end result, it is remarkable how stable the prediction quality is when trained on strongly mutated sequences.

As an alternative to single-nucleotide variants I tried an approach that crops the sequences on the edges. I removed a fixed percentage of the enhancers total sequence. The proportion, in which both ends get shortened, was chosen randomly. As before I added three altered copies to the training enhancers and did not change their expression profiles. I tested a 10% and 25% cropping rate; however, a significant improvement in test score was not achievable, see table B.5.

4.4.2 Homologous Enhancers

Instead of using artificially mutated sequences to enrich the training data, there is the option to use actual homologous sequences. Besides *D. melanogster*, the genomes of 11 additional *Drosophila* species have been sequenced [98, 99]. These species are in order of their relatedness to *D. melanogaster*: *D. simulans*, *D. sechellia*, *D. yakuba*, *D. erecta*, *D. ananassae*, *D. pseudoobscura*, *D. persimilis*, *D. willistoni*, *D. mojavensis*, *D. virilis*, *D. grimshawi*. Together they span 40 million years of evolution.

I identified homologs based on the two-species alignments with the *D. melanogaster* genome obtained from the UCSC Genome Browser [98]. In order to find a homolog, I required that the boundaries of the enhancer fall in a block of conserved sequence or that

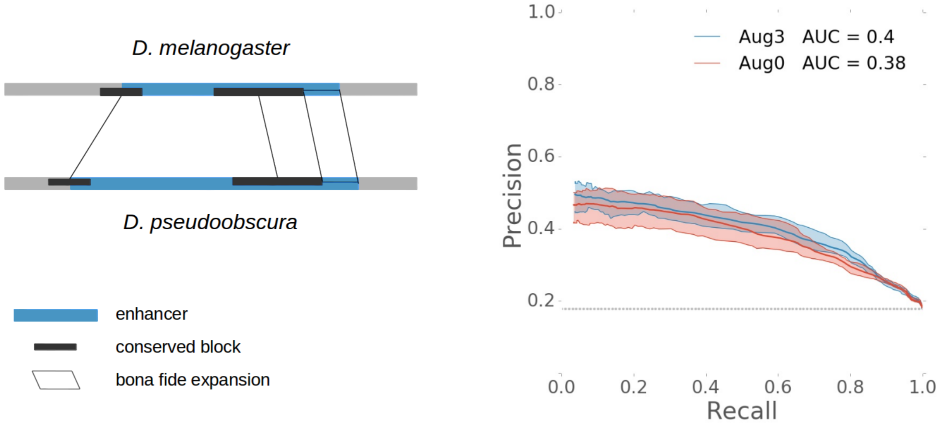


Figure 4.8: (Left) The procedure to identify a homolog of a *D. melanogaster* enhancer in a related species, here *D. pseudoobscura*. If the boundary of the enhancer is not in a conserved block, we expand the enhancer based on the next conserved block. (Right) Performance differences for a 17 TF model with (Aug3) and without (Aug0) augmentation. More details in table B.6.

the enhancer contains a conserved block of at least 100bp length. If one or both enhancer boundaries were not conserved, I estimated the homolog's delineation based on the closest conserved block. I rejected homologs that were shorter than 200bp or longer than 2500bp. 78% of the *D. melanogaster* enhancers are conserved in all of the 11 additional species. For only two enhancers it was not possible to find a homolog in any of the species. For a graphical representation, see the left side of figure 4.8.

Table 4.4: Test scores for different augmentation strategies with two, three and four fold augmentation. Regular augmentation is with closely related species first, alternative augmentation considers weakly related species first. All experiments with 17 TF and without Hyperparameter-training. (** significant at $p \leq 0.01$; significance is depicted for relevant pairs only.)

CV	No Aug	Aug2	Aug2 alt.	Aug3	Aug3 alt.	Aug4
Avg	0.378	0.392	0.386	0.401	0.404	0.396
Std	0.023	0.015	0.014	0.017	0.012	0.012

**

**

Despite substantial sequence divergences, I am confident that these bona fide enhancers drive expression patterns similar to the *D. melanogaster* ones and that it is possible to use them to augment the training data. This comes from the fact that first, the relevant TFs [100, 101] as well as their binding specificities [102] are conserved well beyond the *Drosophila* genus.

Secondly, a small set of homologous enhancers has been identified in other species and the expression driven by these enhancers has been measured in *D. melanogaster* embryos. Their patterns are highly similar to the *D. melanogaster* homologs [103, 104, 19]. Furthermore, earlier modeling efforts demonstrated that these enhancers' expression can be reliably predicted by a model trained to *D. melanogaster* data [27].

I experimented with the number of additional enhancers and with the order of species to select homologs. As a standard approach, I started with closely related species to add enhancer homologs and moved to more distant relatives until a maximum number of additions have been selected. As an alternative approach, I mixed the order by starting with the intermediate species (*D. yakuba*, *D. erecta*, *D. ananassae*) followed by the standard order. I performed training runs with two, three as well as four additions (Aug2, Aug3, Aug4). All forms of augmentation improved the test score, however, only for more than two additions were the results significant, see table 4.4 and in the appendix B.6. Aug4 yielded slightly worse results than Aug3 and requires more computation time. The order of species was irrelevant for the test result. In conclusion, the method of choice is three-fold augmentation with the standard order of species, see 4.8 (Right).

4.5 Hyperparameter Training

The parameter penalty λ is one example of a hyperparameter of the model. In general, hyperparameters are all model parameters that are not subject to optimization. Further examples are the scope of the optimization algorithm and the range of cooperativity. The choice of the parameter penalty has an impact on the predictions, however, it is neither possible to train λ on the training data nor use the test data. Using the training data would thwart the idea of the parameter penalty because reducing the penalty automatically increases the objective function. Using the test data to determine λ is problematic because it would render the test data no longer independent from the parameter training. By tuning λ to improve the test score, one would train it on the test data by hand.

There are two possible solutions to choose the hyperparameters. The first solution is to guess reasonable parameters and not change them retroactively. The second solution is to use a third data set, called the validation data, to chose the best hyperparameters independently of the training and test data.

A common approach is to use nested cross-validation in which the training data is again parsed into multiple fragments to serve as an independent cross-validation set-up, e.g. [105]. After the inner cross-validation run determined the optimal hyperparameter, the model is trained on the full training data and tested on the last data share as with normal cross-validation. This procedure gets repeated for every training run, the outer cross-validation loop. Nested CV greatly increases the number of training runs (one inner cross-validation loop for every hyperparameter and every batch of the training dataset). I do not apply this approach for the simple reason that it is very resource intensive.

To see this, consider that the training in the inner cross-validation instances takes similar computation time as the full training. Even with a small set of hyperparameter settings and a coarse inner cross-validation fragmentation, it would require orders of magnitude more computation time. Since the parameter optimization takes several minutes, sometimes more than one hour, nested cross-validation is simply too computationally expensive.

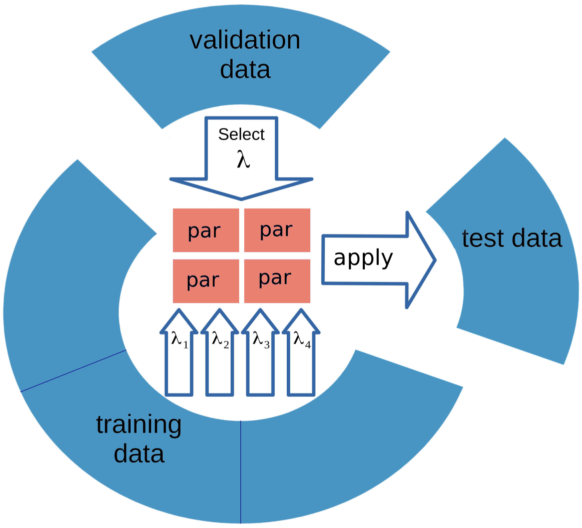


Figure 4.9: The hyperparameter training set-up HTI/II. All training results get tested on the validation data. The results with the best validation score are used for the final prediction.

Alternatively, I tried two hyperparameter training strategies, which I call HTI and HTII. They differ in their choice of validation data. HTI is a simple version of the nested cross-validation set-up. Instead of training on 9 out of 10 data fragments during cross-validation, I hold one data fragment back as validation data and use only 8 shares, see figure 4.9. The advantage of this approach over the nested cross-validation is that for every training instance I need to train the model only once for every hyperparameter setting and can simply use the best set of training parameters to predict the test data. An additional training run with the best hyperparameters is not necessary because the parameters have already been calculated with the full training data. The disadvantage is the reduction of available data. First, the training data is reduced because the validation data is at no point merged back with the training data, and second, the validation data itself is very small because only one data fragment is used to assess the quality of the hyperparameter setting. Since the data fragments contain around 10 enhancers each, the validation data is likely not representative.

HTII solves the problem of HTI by applying ideas similar to data augmentation. Instead of holding data fragments back for validation and therefore reducing the training data, I train on the original training data, 9 out of 10 data shares. The validation data consists of homologous sequences of the training enhancers from two *Drosophila* species (*D. yakuba* and *D. erecta*). If there is no homologous enhancer, I simply used the original sequence.

Table 4.5: Test scores for no penalty (NP: $\lambda = 0$), best-guess penalty (HT0: $\lambda = 10^{-3}$) and the hyperparameter training strategies HTI and HTII ($\lambda \in \{10^{-3}, 5 \cdot 10^{-3}, 10^{-2}\}$), each with and without data augmentation. HTII+ with additionally $\lambda \in \{5 \cdot 10^{-4}, 10^{-1}\}$. (* significant at $p \leq 0.05$)

	No Aug				Aug3				
test score	NP	HT 0	HT I	HT II	NP	HT 0	HT I	HT II	HTII+
Avg	0.362	0.378	0.389	0.383	0.389	0.401	0.417	0.411	0.405
Std	0.029	0.023	0.026	0.015	0.015	0.017	0.032	0.01	0.012
		*					*		

By applying this strategy, the training data does not get reduced and there is a rich validation data set. The caveat is that the validation data has similarities to the training data, especially when data augmentation is used. Hence, it was necessary to test whether this strategy is compatible with data augmentation. Note that test and training data are even with HTII completely independent because no homolog of a test enhancer entered the validation data.

As adaptable hyperparameter, I chose the strength of the parameter penalty λ. Since the available computation resources are limited and single training instances take up to one hour, I limited tests to three values for λ (0.001, 0.005, 0.01). I tested both strategies, HTI and HTII, with and without data augmentation, see table 4.5 and in more detail B.7. Overall, applying any form of parameter penalty improves the final test score noticeably. However, only with hyperparameter training the improvement becomes significant (HTII in the case of augmentation, HTI without augmentation).

The effect of hyperparameter training and augmentation is additive. The best average results are achieved with HTI in combination with 3 fold augmentation. Although the test score for HTI is slightly better than for HTII, I prefer HTII as hyperparameter training scheme because it reduces the variation across different CV parsings significantly (F-test: with augmentation $p \leq 0.01$, without augmentation $p \leq 0.05$). Probably because of the broader data basis in the validation set as well as the training set, the scores for HTII differ less, indicating more consistent results.

The addition of further hyperparameter values is not beneficial as can be seen in table B.7. For HTII+, I tested 5 hyperparameter values λ (0.0005, 0.001, 0.005, 0.01, 0.1) without any improvement of the final test score.

4.6 Focused Discussion: the Optimal Training Strategy

The optimized parameter training set-up is a combination of techniques that I tested on their own as well as in combination with others. It consists of a global parameter optimization algorithm (CMA-ES), L1 parameter penalty, a hyperparameter training set-up based on homologous enhancers (HTII), and three-fold data augmentation (Aug3), also based on homologous enhancers. Table 4.6 summarizes the the improvement in performance for every single technique and their combination, which I measure in the form of percentage gain. There are multiple ways to measure the effect of the training methods. I chose to evaluate the gain by comparing the test scores with and without the technique under otherwise optimal conditions. Of course, this is only an approximation, which helps to visualize the improvements.

Table 4.6: Approximate improvement gained from the training techniques of the last sections as percentage of the base test score. The last column is the gain from all four techniques combined.

	CMA-ES	L1-penalty	HTII	Aug3	Combined
8 TF	1.4%	0.6%	2.8%	1.9%	16.4%
17 TF	4.9%	3.1%	2.6%	7.3%	13.9%

Two things are immediately apparent. First, the improvement in score for every technique on its own is relatively small but their effect can be combined. Taken together, these methods result in a substantial step forward in prediction quality. Particularly successful on its own is data augmentation. This comes mainly from the fact that the number of enhancers, which are the training data, is very low. Since identification and measurement of additional enhancers is very laborious and since there is a limited number of naturally occurring enhancers in *D. melanogaster*, data augmentation is the only way to extended the training data considerably.

It is common in the field of image analysis to augment the data with new images by mirroring them on one axis. The equivalent in this case would be to use the reverse complement of the sequences. However, this is not a valid option for thermodynamic models because they process the enhancers independently of their orientation. Hence, the expression output would be the same and the effective number of training enhancers would remain unchanged.

Second, the improvement is especially noteworthy for the extended model with 17 TFs. The reason is the increased number of parameter for the extended model (52 instead of 25 for the reduced model). This is not surprising since a larger set of parameters increases the risk of over-fitting. All techniques are designed to handle many model parameters: CMA-ES improves the navigation through the parameter space; the penalty (with the penalty strength as trainable hyperparameter) limits parameter values and pushes useless parameters back into the neutral position; augmentation broadens the data foundation substantially. In their combination, they allow us to train a model with more details and evaluate even minor aspects transcription control.

Chapter 5

Model Evaluation

The last chapter depicted all the tools to improve the predictive power of thermodynamic models. However, predicting patterns as accurate as possible is ultimately not the purpose of this thesis. The primary goal is to understand the mechanisms of transcription control and enhancer architecture. We learn these details by probing the trained model and by trying alternative models in order to assess the influence of all system components. Only with high-quality predictions, we can be confident that the model represents mechanisms underlying the real biological system. A high prediction quality, represented by a good test score, enables us to build more complex models and dissect them in much greater detail than a coarse, low-quality model.

5.1 Binding Specificities

5.1.1 Number of Binding Sites

I already introduced the specificities of the binding motifs in the form of a PWM, see chapter 3.2. PWMs are a continuous measure of TF binding site strength. They do not categorize sequences whether they are binding sites or not. In theory, every sequence can be bound by any TF and the identification of binding sites is unnecessary. Since most sequences have a negligible binding weight, they do not affect the prediction; however, they have a strong influence on the computation time of the model. The time required to calculate the partition functions in equation 2.21 depends linearly on the number of sites, quadratically if cooperativity is incorporated. Including all sites regardless of their quality is computationally unfeasible. Hence, I define a threshold T of the binding energy for the inclusion of sites. Equation 2.10 shows that the energy is proportional to the log-likelihood-ratio LLR between the PWM and the background probabilities. To define a uniform threshold for all TFs, I scale the threshold with the binding energy of the consensus site $\max(LLR)$. The thermodynamic model identifies a sequence as a binding site if:

$$LLR \geq (1 - T)\max(LLR) \tag{5.1}$$

The threshold becomes more strict for lower T. At $T = 0$ only the consensus is identified as a binding site. For $T \geq 1$, sites with a negative[1] LLR are included. There is little reason to include sites with a negative LLR, thus, the reasonable range for T is between 0 and 1. Figure 5.1 depicts the distribution of binding weights for all possible sites in the enhancers. Included are sites with binding weight of roughly 3 orders of magnitude. Most sites are much weaker and will be neglected. The right hand side of figure 5.1 shows eight exemplary bcd sites. The first four are the strongest bcd sites in their enhancer, the latter four are the weakest for a threshold of $T = 0.5$. Most differences to the consensus site are in the lesser important flanking region of the motif. The model includes sites with one or two mismatches in the core motif, too, although their weight is one order of magnitude lower.

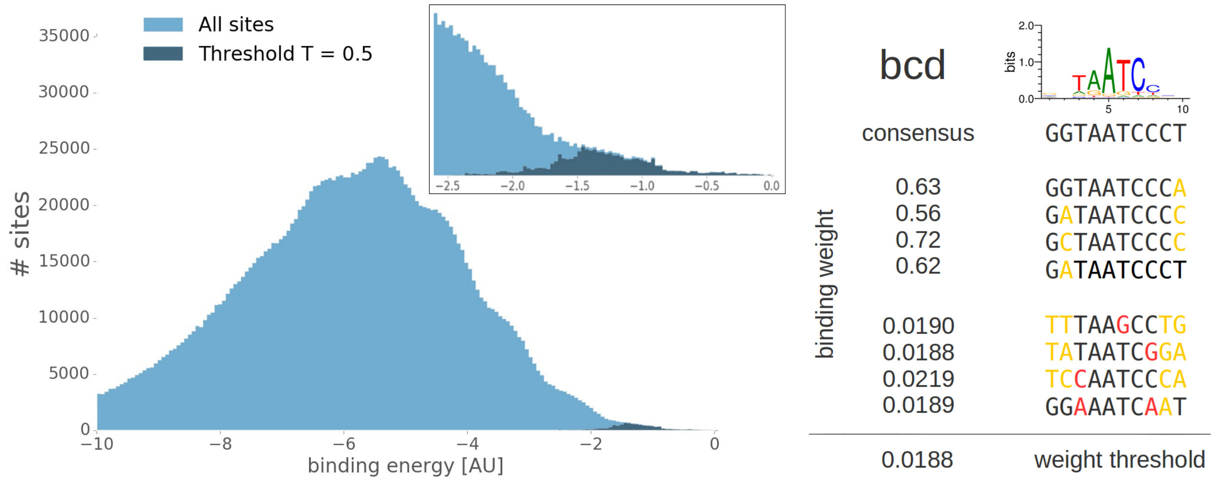


Figure 5.1: (left) Histogram of binding energies ($\log_{10}(w)$) of enhancer binding sites. For the model we use only a subset of approximately 0.4% of all sites, however, they carry 66% of the total binding weight. (Right) exemplary strong and weak binding sites of bcd. Marked are mismatches in the core motif (red) and in the flanking region (yellow).

In the following, I compare predictions with different thresholds T. This analysis is not meant to answer the question to which extent weak sites are relevant for the model because the training algorithm has the possibility to adapt to a scarce site distribution during the training phase. In a later section about enhancer architecture (6.2), I am going to revisit the issue of weak sites and tackle the question with an impact analysis, which is much more suitable for this question because the impact score measures influence without giving the model the chance to adapt. Here, I only determine the influence of site distribution on the model training.

[1] The LLR will be negative if the PWM frequencies of the site are lower than the background frequencies. In other words, it is more likely that this sequence comes from a background sample than from a set of sites.

I compared four thresholds $T \in \{0.4, 0.5, 0.6, 0.7\}$. The steps between the thresholds resemble almost a doubling of incorporated sites and each quadruples the training time of the model, table 5.1. Both, too few and too many sites, can deteriorate the prediction. Certainly, ignoring relevant sites impedes the training because important features of the enhancers are not incorporated.

Table 5.1: Number of sites, computation time and test scores with various site thresholds T for the reduced as well as for the expanded Model. Standard deviation for all test results is $\sigma \approx 0.01$. (* : due to long computation time, the result for $T = 0.7$ with 17 TF is based on only 5 iterations.)

	8 TF				17 TF			
T	0.4	0.5	0.6	0.7	0.4	0.5	0.6	0.7
# sites per enhancer	50	96	181	350	61	123	238	462
# sites per kb	35	67	125	243	42	86	166	321
comp. time [min]	5	15	60	584	15	60	240	2220
test score	0.373	0.37	0.346	0.351	0.384	0.411	0.405	0.375*

In contrast to ignoring sites, it is not immediately clear why incorporating more sites reduces the prediction quality. One reason can be that the PWMs are an imprecise measure for the assessment of weak sites. The HIP-FA PWMs, which are used here, were measured by testing all single base mismatches. Hence, it is reasonable to assume that strong binding sites similar to the consensus are correctly modeled. In the case of multiple mismatches, the PWMs rely on the additivity assumption, which might misinterpret the strength of those sites, see section 2.1.3.

A threshold of $T = 0.5$ seems to be optimal from a model training aspect and to be a good balance between number of sites and computation time. Thus, this threshold will be the default for all following predictions.

5.1.2 Alternative PWM Sets

Next, I compared different sets of PWMs generated by three different methods: HIP-FA, B1H, and Footprinting. Relevant for the interpretation of a PWM is not only the order of the bases but also the overall specificity measured in the form of information content. HIP-FA PWMs are less specific than their alternatives, which makes this comparison also about what level of specificity is optimal for the model. For this reason, I tested the PWM sets with and without pseudocounts PC. As PC for every entry, I choose 0.25 for the count-based PWMs (B1H and Footprinting) summing to one additional unspecific binding site. For the HIP-FA PWMs, which sum to 100, we add 1 in every entry.

To concentrate on the effect of the PWMs, I limit the complexity in the system and use the reduced set of 8 TFs. Notice that there is not a good HIP-FA PWM available for cic and, therefore, I use the B1H PWM instead. Figure 5.2 depicts the differences between the PWM sets.

Table 5.2: Listed are the average number of sites per enhancer and test scores applying various PWM sets, each tested with and without pseudo-counts (PC). Since B1H as well as Footprinting PWMs call fewer binding sites using the same threshold ($T = 0.5$), I tested a prediction with more binding sites (PC and $T = 0.6$). The HIP-FA PWMs score significantly better than the alternative ones in every category (Wilcoxon signed-rank test $p \leq 0.01$).

	No PC		PC		More Sites	
	sites	score	sites	score	sites	score
HIP-FA	96	0.37	101	0.377	181	0.346
Footprinting	37	0.305	41	0.321	79	0.327
B1H	29	0.289	38	0.311	67	0.305

Table 5.2 shows the average test scores for all PWM sets for 10 independent 10-fold cross-validation runs. The HIP-FA PWMs clearly outperform all other PWMs by a large margin (Wilcoxon Signed-rank test against B1H and Footprinting with and without PC $p \leq 0.01$). The second best PWM set is Footprining with PC (against B1H + PC $p \leq 0.05$). It is especially noticeable that PC are very beneficial in the case of B1H and Footprinting but not so much for HIP-FA. This indicates that HIP-FA correctly captures the unspecificity of TF-DNA binding, while B1H and Footprinting are too specific.

An alternative explanation could be that the binding site threshold $T = 0.5$ is too strict for the more specific PWMs. In fact, the number of binding sites identified by B1H or Footprinting is almost one-third of the number of HIP-FA sites. It is possible that predictions with unspecific PWMs would perform better if they were based on a similar number of binding sites as HIP-FA. Hence, I tested the energy threshold $T = 0.6$ also for B1H and Footprinting, table 5.2. Again, the number of binding sites doubles. However, the test scores do not change significantly for B1H and Footprinting. Thus, the better performance of HIP-FA is based on a better site assessment and not necessary on calling more sites.

5.2 Model Expansion

5.2.1 Additional Transcription Factors

Almost all tests that I did up to now, indicate that the expansion of the TF set improves the prediction result. Table 5.3 summarizes the results of a naive and a fully optimized training approach. This proves the importance of parameter regularization. The benefit of the additional TFs is not significant if we apply no form of parameter penalty or data augmentation. Without our advanced optimization set-up, the inclusion of additional TFs leads to over-fitting and an overall suboptimal performance.

Applying the techniques of chapter 4 is especially beneficial for predictions with the expanded set of TFs. Only with an optimized training technique, there is a significant

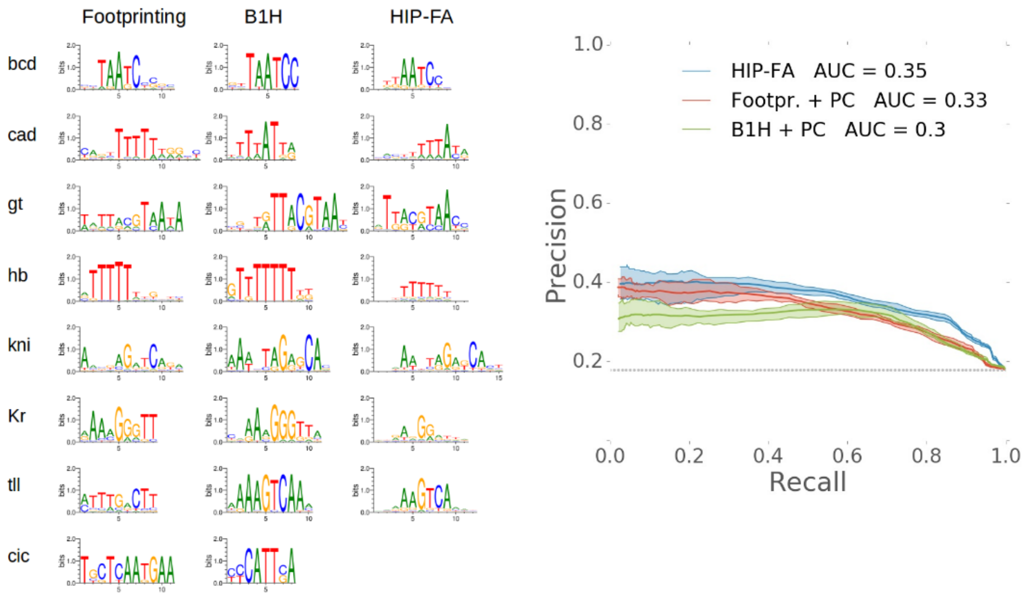


Figure 5.2: (Left) SequenceLogos for all PWMs. The trend that the HIP-FA PWMs are less specific than their B1H and Footprinting counterparts is immediately visible by the height of the logos. (Right) Visualization of the key results from table 5.2. The HIP-FA PWMs clearly outperform alternative PWM sets even without the use of pseudo-counts.

difference between the reduced and the full model. Therefore, it is possible to conclude that the expanded set contains relevant TFs for the pattern formation. Figure 5.3 depicts the improvement in the form of a PR-plot.

Table 5.3: Test results for 8 TF and 17 TF trained without parameter regularization or augmentation (naive training) and all techniques of chapter 4 applied (optimized training).

	Naive Training	Optimized Training
8 TF	0.351	0.37
17 TF	0.362	0.411

Although the test results are promising, three important questions are still open. First, which TFs are relevant for the system, or better, which TFs have the most impact on the model? Second, is the result of the prediction plausible and consistent with experimental data? And third, how do the TFs influence the predictions? Figure 5.4 answers the first two questions. Depicted are the impact of all TFs as a measure of TF influence on the prediction, see section 4.2.3, and the distribution of activatory potentials, which indicates whether a factor is regarded as an activator or a repressor.

All TFs except slp1 are categorized correctly and most very consistently. The classic 8 TFs, bcd, cad, cic, gt, hb, Kr, kni, and tll are all correctly predicted in every training

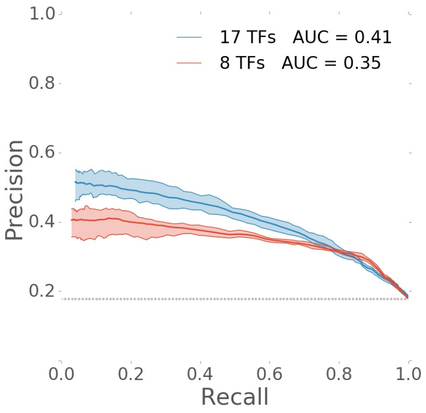


Figure 5.3: The PR-plot sums-up the model predictions for ten independent 10-fold CV runs with our optimized training set-up. The expanded model with 17 TFs clearly outperforms the reduced model with 8TFs.

iteration, except for gt that was falsely predicted to be a weak activator in only one out of 100 training runs. Considering that gt binding sites are in only 61% of the enhancers, a single misclassification is not surprising.

Also not surprising is that these 8 TFs have a huge influence on the patterns. Knocking-out bcd alone accounts for roughly half of the prediction score. This result fits well with the reported role of bcd as a key TF-independent of bcd dependent TFs. The posterior activator cad has less impact, most likely because we have fewer enhancers expressed in the posterior part of the embryo. Also clearly important are cic, hb, kni, Kr, and –although less important– gt. Unexpected is the small impact and the large standard deviation of tll. To understand this, it is necessary to evaluate the impact on the single enhancer level.

Figure C.6 in the appendix shows the impact of the TFs on each individual enhancer. If the impact is positive, the TF is to some degree necessary for the model prediction of this enhancer. Removing the TF from the model deteriorates the prediction, see for example figure 5.5. On the other hand, the impact of a TF on an enhancer is negative if removing it from the model improves the prediction. In this case, the parameters of the TF have been overfit in general or the interaction of the TF with the enhancer can not be explained by the model. All of the important TFs influence the prediction of certain enhancers negatively, although the averaged impact for most of them is positive. Especially noticeable is that tll represses its own enhancers, which is most likely the reason for its low average impact.

5.2.2 Focused Discussion: Role of the Transcription Factors

The additional TFs show a similar behavior, however, their average impact is in most instances much lower. D, Nub and pdm2 are late gap factors that play a secondary role in the segmentation network [68, 69, 70]. If knocked-out or ectopically expressed, they cause slight patterning defects of pair rule genes, but not a complete loss of entire expression domains. This makes it especially challenging to predict their role. However, the model

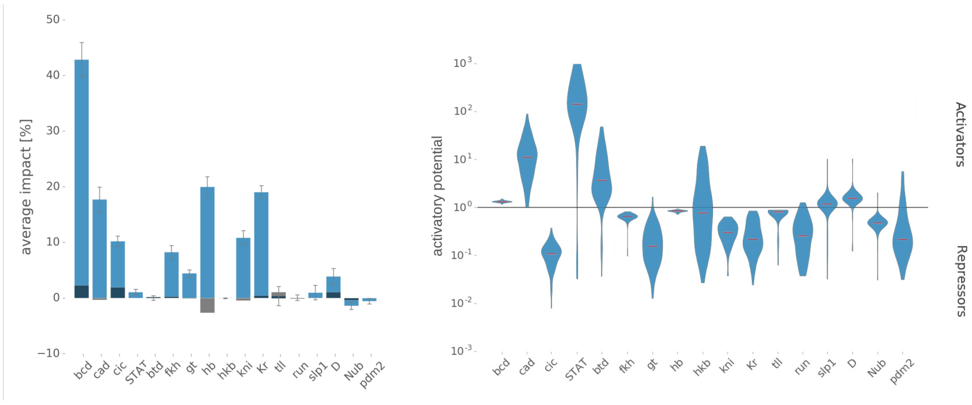


Figure 5.4: (Left) Average impact of all TFs as percentage of total prediction score (blue) and the impact of the cooperativity alone (gray). Negative impact indicates an improvement in the prediction after removing the TFs. (Right) Distribution of the activatory potential for all CV parsings (100 parameter values) in the form of a violin plot. The distribution width depicts the density of parameter predictions, the red lines mark the median prediction. The black line marks the boundary between activators (activatory potential > 1) and repressors (< 1).

correctly predicts D to be an activator (97/100 times) and Nub to be a repressor (99/100 times). D shows an overall positive impact, e.g. by helping to form posterior pair rule expression (run, eve and h), see figure C.6. Figure 5.5 (A) depicts exemplary patterning defects for a ftz and a run enhancer when D gets knocked-out *in silico*. While the stripes are still present, they are less pronounced and weaker expressed. Also shown are the results from *in vivo* deletion experiments from the literature, where a similar effect was observed [68].

Nub is in its impact comparable to tll. The on average negative impact comes mainly from gap-gene enhancers, e.g. an early run enhancer run(+1), D(+4), and kni(+1). Nevertheless, the model predicts that Nub sharpens the gap between pair-rule stripes, e.g. run(-31), run(+30), and ftz(-6), see figure 5.5 (B). An explanation for this result could be inconsistent enhancer staging. For the expression measurement, enhancers were selected from the same time point during embryogenesis. Nonetheless, an imprecision of a couple of minutes is highly likely, especially, since the experimental results come from different sources. It is plausible that the late-expressed gap gene fits pair-rule stripes well while being detrimental to the prediction of early gap-patterns. A solution to this problem is hardly possible with the data situation at the moment and would require remeasuring all enhancers with a more thorough staging or time resolved expression measurements, i.e. from live-reporters [106].

Although pdm2 is correctly classified as a repressor in 83 out of 100 training instances, the model assigns this TF almost no impact in any enhancer. This comes from the fact that pdm2 has binding sites in very few enhancers and receives, therefore, a low weight by the algorithm due to our parameter regularization effort.

The gap factors fkh and hkb act as repressors in the termini [65]. They are expressed

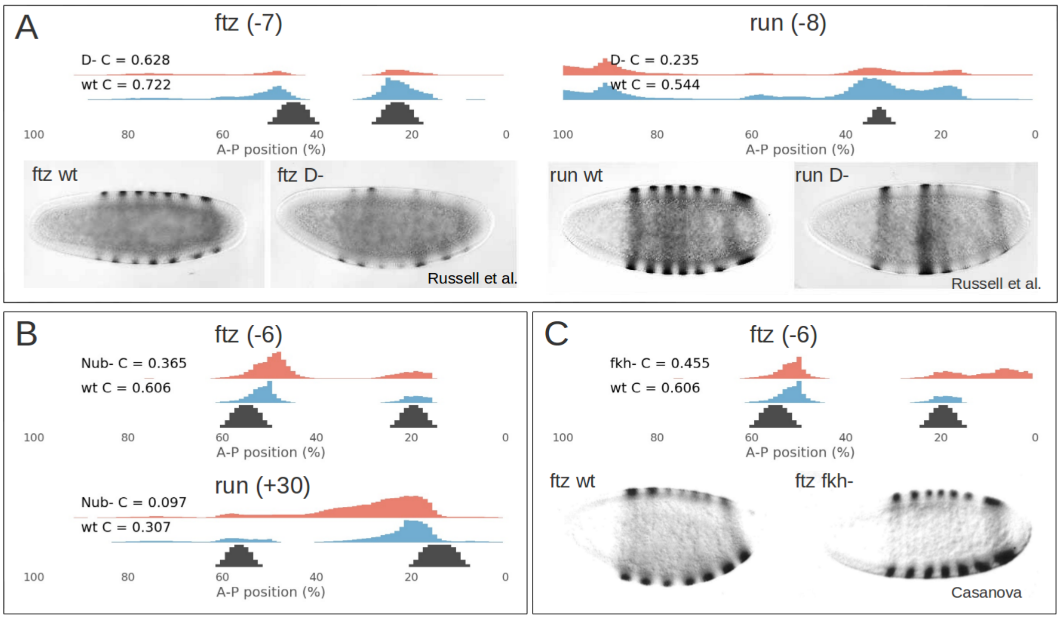


Figure 5.5: *In silico* knock-out experiments for Dichaete (A), Nub (B), and forkhead (C). The wild-type profiles (wt, blue) show the superposed predictions of ten CV runs, the knock-out profiles (red) show the prediction without the deleted TF, scaled to the wild-type track. The measured data is plotted in black. Also included are published *in vivo* knock-out experiments for Dichaete and forkhead [68, 65].

in the anterior and posterior cap of the embryo and constrain expression to the middle segment. The model clearly identifies fkh as a repressor with high impact. Figure 5.5 (C) depicts the patterning defects of a fkh deletion *in vivo* as well as *in silico*. It successfully predicts the posterior expansion of the seventh ftz stripe in the case of lacking fkh input. Among the expanded set of TFs, fkh has the largest impact. hkb, on the other hand, has no measurable impact. The model clearly struggles to assign hkb a role (repressor 54/100 instances). Much like pdm2, hkb binds only a few enhancers. The model identified only 15 binding sites for hkb, mostly because its PWM is very specific, see figure 3.3. Thus, it is not surprising that the training algorithm assigns this TF a very low binding affinity. However, experiments from Casanova suggest that hkb has a similar effect as fkh and only the combined knock-out of both TFs leads to a full patterning defect of ftz stripe 7 [65]. It seems that in the model, fkh compensates for some of the hkb repression, which is missing due to a lack of predicted hkb binding sites. Somehow, the hkb PWM used here and alternative ones from the literature, which are extremely similar [75], do not reflect the *in vivo* binding specificity for which we have no explanation.

The gap factor btd is an activator that helps to form eve stripe 1 [64]. The model consistently predicts btd to be an activator (94/100 training runs), however, with vanishing impact. btd binds to the enhancer of eve stripe 1, eve(+7), and additionally the stripe

1 enhancers of prd and run (prd(-3), run(-10)), but its impact is barely visible above the broad background expression of bcd.

Run and slp1 are early pair rule genes that show gap-gene-like behavior during early segmentation [66, 67]. They are both repressors that constrain the boundaries of anterior expression domains. Indeed, the model predicts run to be repressor in most training instances (96/100) although its impact is very low. Run has a positive impact on some head enhancers (oc and noc) but is clearly a secondary TF.

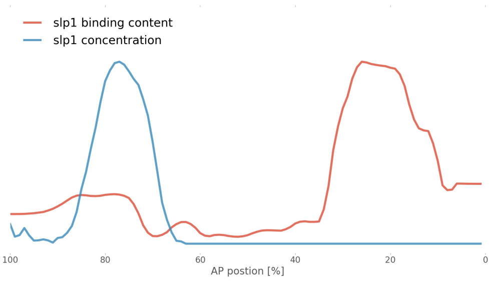


Figure 5.6: Distribution of slp1 position-specific binding content and concentration. The red curve illustrates where the enhancers with slp1 sites are expressed along the embryo axis. We calculate the position-specific binding content by averaging the expression patterns of all enhancers weighted with the summarized slp1 binding weight in their sequence. Most enhancers with strong slp1 binding sites are expressed in the posterior tail of the embryo.

The only TF that the model wrongly classifies is slp1. The model predicts that slp1 is an activator that boosts expression in the embryonic head region (89/100 training instances). Figure 5.6 shows that most of the enhancers that carry slp1 binding sites are expressed posterior to the expression domain of slp1. However, the yield in score from the few head enhancers, if slp1 is an activator, is greater than the mild effect of a repressive slp1. Interestingly, one of the most pronounced targets of slp1 is an enhancer that regulates the slp locus (slp1/slp2) itself. It is unclear whether these binding sites are spurious or whether it is a case of auto-regulation, which the static model can not capture.

Finally, Stat is a ubiquitous activator that targets evidentially multiple pair rule and gap genes [71, 72]. One expects that any ubiquitous TF activity has only a minor impact on the shape of the expression predictions since its concentration profile does not carry positional information. However, it does change the overall expression level much like the basal transcription rate q_{btr} and therefore alters the sensitivity to the input of other TFs. A high sensitivity leads to sharper expression profiles, while a low sensitivity results in broader expression profiles. The algorithm predicts Stat to be a strong activator in most training instances (95/100), thus capturing its reported role.

5.3 Expression Predictions

After establishing the parameter training methods and analysis of the parameter results, I am going to now focus on the actual output of the model: the expression predictions.

Here I present the results for the best model, which incorporates all 17 TF and was trained with the fully optimized training setup. As a reminder, I repeated the parameter training 10 times with different cross validation parsings. Therefore, there are 10 different test results for every enhancer. Figure 5.7 shows a superposition of the ten test results in one profile for each enhancer. The superposed profile depicts the average prediction based on differing models. It is possible to see which expression domains get predicted well on average.

The spread in prediction quality of the differing models is depicted in figure 5.8. It shows the distribution of test scores for every single enhancer in the form of a box plot. The median score is marked in red. The score distribution of most enhancers show a large variance. Hence, the prediction results depend heavily on the parameter results and, therefore, on the composition of the training data. The enhancer that gets predicted is never part of the training data according to the rules of cross validation. However, the test result of a certain enhancer will likely be better, if more enhancers similar to the predicted one are in the training set. Nevertheless, almost all enhancers can be predicted well in at least some instances; 96 out of 98 enhancers have at least once a test score above 0.5. Many have a high test score in all cases, e.g. gt(-10), Kr(-3), and rib(+3). Only few enhancers are consistently difficult to predict, most prominently dfd(-13) and gt(-1).

The large spread of test scores is especially surprising since the roles of the important TF get correctly predicted in almost all training iterations, see section 5.2. It is rather the composition of TF activity that causes the differences between a good and a bad prediction. Take for example gt(-1), which is difficult to predict. The enhancer drives expression in two domains that get separated by Kr repression, see figure 5.7 and C.6. While Kr is necessary for gt(-1), it is deleterious for the enhancer Kr(-2). Depending on the balance of enhancers in the training set, either gt(-1) or Kr(-2) will be suboptimal predicted.

This discrepancy between results from similar models can be seen from two viewpoints. From the modeling viewpoint, there are clearly aspects of the biological system that are not understood and that have to be simplified. The conflicting role of TFs like Kr in the example above means that either some of its binding sites are falsely predicted or that the enhancer architecture is more complex. A reason for the former can be missing accessibility information, i.e. whether the chromatin at the TF binding sites is open for protein binding. In support of the latter, one should mention that the model as stated here ignores the effective range of repressors or heterotypic interactions, both aspects will be topics of later sections in this thesis.

On the other hand, there is the statistical viewpoint. It states that there is much looseness in the system, i.e. the prediction has high variance. In other words, small changes in the composition of the training data can yield large differences in the prediction result. There are typically three approaches to deal with this issue. More training data, stricter parameter regularization, or ensemble approaches. The data augmentation technique showed that a broader data foundation improves the model training greatly. Unfortunately, the data that is available is limited and is laborious to expand. Similarly fruitful was our regularization effort, however, it is hardly feasible to strengthen regularization or reduce the number of parameters without giving up on the goal to build an in-depth model of transcription control. Improving the model should enable the inclusion of more features not the other way around.

Ensembles of models are known to reduce the prediction variance [107]. Instead of training a single model that represents our best understanding of the system and the data, an ensemble relies on multiple diverse models[2]. Their predictions combined constitute the final ensemble prediction. A training technique that generates an ensemble is bagging [108]. The name stands for bootstrap aggregation. For this technique, the model gets trained multiple times with differing fragments of the training data, the so called bootstrap samples. The final result is the superposition of all models. The results in figure 5.7 are generated in a similar spirit. They too are a superposition of multiple models, each trained on a slightly different training data set[3].

Indeed, the superposed prediction performs well in comparison to the single predictions. The average correlation of the single models is $\bar{C} = 0.411 \pm 0.01$, while the correlation of the ensemble model is more than one standard deviation higher $C_e = 0.422$. Although an improvement of the model prediction is desirable, the ensemble model lacks one key feature that is interpretability. By combining multiple models, the resulting predictions depict no longer the biophysical model that was described in chapter 2. For this reason, ensemble models are interesting only on a theoretical level.

[2] The term diverse models includes in this context also the use of the same model framework with differing parameters.

[3] In our case, the model was trained on exactly 90% of the available data, while for bagging each iteration of training data contains on average 63% of the whole data (random sample with replacement).

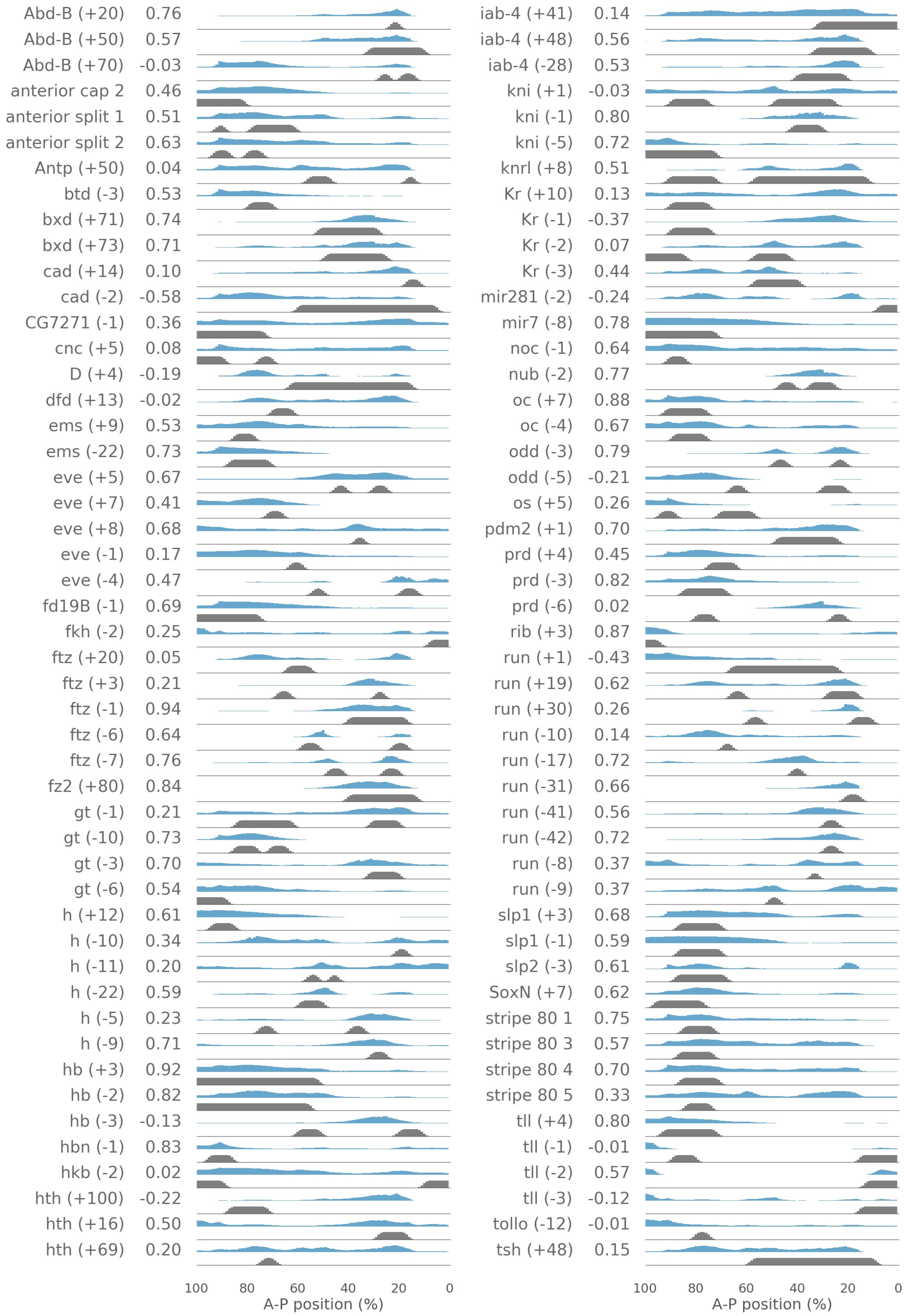


Figure 5.7: Superposed model predictions for ten CV-parsings (blue) and measured profiles (gray). The correlation between prediction and measurement is indicated between the enhancer name and the profiles.

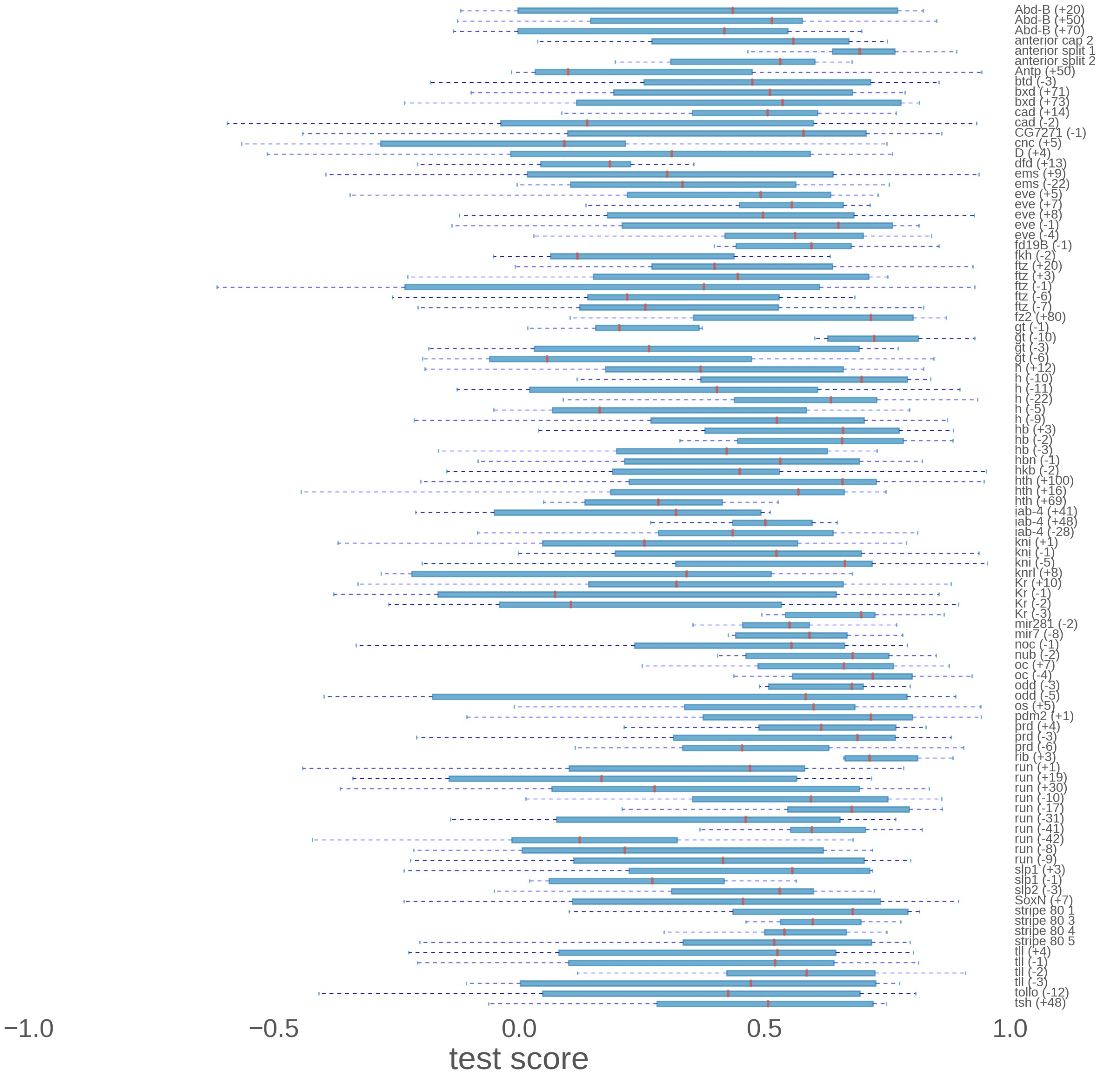


Figure 5.8: Test scores (correlation) of all enhancers as a box plot for 10 data parsings. How well a enhancer gets predicted depends strongly on the composition of the training data.

Chapter 6

Enhancer Architecture

The preceding chapter showed that sequence binding content, i.e. the composition of TF binding weight, does not fully explain the expression patterns. Furthermore, the composition of binding content is often inconsistent, e.g. when enhancers carry binding sites of TF that could distort the correct expression pattern. It is therefore plausible that the information encoded in the sequence is additionally organized on higher levels beyond the mere composition of TFs. This includes features like binding site interaction, steric hindrance, sequence accessibility, and short range repression. I call all those aspects enhancer architecture.

Given our lack of understanding of enhancer expression, architectural features could be part of the reason why the predictions still deviate from the measured expression. The following sections build up on the architecture that is already included in the model, i.e. homotypic interaction, and investigate further aspects in more detail as well as analyze whether including them improves the expression model.

6.1 TF Interactions

There is plenty of evidence that TF binding sites of the same and of different type tend to cluster in enhancers [109, 51]. However, the statistical analysis of site clustering does not necessarily prove actual protein interaction or cooperative binding. If the site clusters are functionally relevant, including their effects in the model should improve the predictions.

Until now, the model gave all TFs the option to interact homotypically based on a fairly simple uniform interaction model. This model assumes a constant level of cooperativity γ regardless of the distance between the binding sites until a maximal range of 50bp and no cooperativity between sites that are farther apart. These assumptions were made to constrain the computation time of the predictions, which increases substantially if heterotypic interactions and longer distances are allowed. Furthermore, additional heterotypic interactions introduce new parameters to the simulation (one per TF pair) and, therefore, increase the risk of overfitting. Regardless of those simplifications, cooperativity of some TFs has a noticeable impact on the predictions, see figure 5.4. In the following, I depict alternative and more complex interaction models.

The cooperativity function defines the level of cooperativity γ for a pair of sites based on their distance and TF types. The simplest model, which was used up to now, is a uniform model with a maximal range. In this model, the range is a hyperparameter of the model and defines a threshold below which TFs interact distance-independently. Alternative interaction functions model a declining cooperativity level, see figure 6.2 left. The linear interaction model assumes a linear decline with the gradient determined so that the function is continuous at the point of the maximal range. The third model assumes a Gaussian decline. For simplicity, the width of the Gaussian is fixed to 1/3 of the maximal range. The training parameter for both, the Gaussian and the linear model, is still the maximal level of cooperativity.

Consider binding sites of TFs i and j and distance d. The maximal range is defined as R and $\gamma_{i,j}$ is the cooperativity parameter, which gets trained during optimization. The interaction function is:

$$\gamma(i,j,d) = 1 + \begin{cases} \gamma_{i,j} & d < R \text{ uniform} \\ \gamma_{i,j}\,(1 - d/R) & d < R \text{ linear} \\ \gamma_{i,j}\exp\left(-\frac{9d^2}{2R^2}\right) & d < R \text{ gaussian} \\ 0 & d \geq R \end{cases} \tag{6.1}$$

6.1.1 Cooperativity Range

Before testing different cooperativity models, let us explore the clustering of binding sites in order to formulate a hypothesis of preferred spacing configurations.

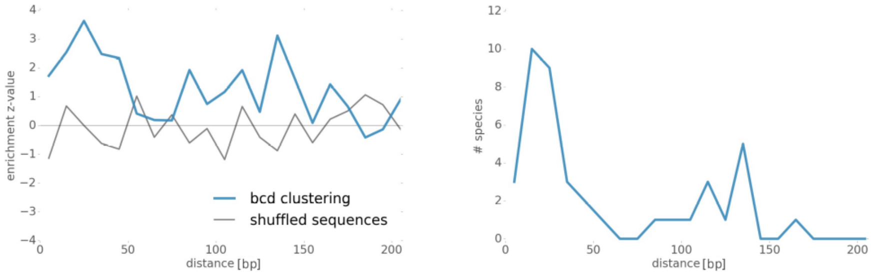


Figure 6.1: (Left) Enrichment analysis of bcd clusters. The z-values were calculated in 10bp windows by comparing clustering of bcd sites in enhancers with randomly placed sites (10,000 repeats). (Right) Conservation of clustering. The clustering analysis on the left was repeated for 12 drosophila species. Depicted are the number of species for which we see clustering for this distance. An enrichment counts as conserved if its empirical p-value is $p \leq 0.01$.

As will be seen later, bcd shows the most pronounced cooperativity. Hence, I concentrate the analysis on bcd homotypic interaction. Figure 6.1 (Left) shows the result of an enrichment analysis for bcd sites. For this analysis, I counted the number of bcd site pairs in the segmentation enhancers weighted with the product of their binding weights. This weighting is rather conservative as it neglects clusters of weak binding sites. An alternative method would be simple counting, which takes clusters of weak sites into account but depends strongly on the binding weight threshold [51]. I counted sites in windows of 10bp site distance and compared the result with randomly placed binding sites (10000 iterations). In addition, it is possible to calculate empirical p-values by counting how often the null model shows similar or higher levels of enrichment. Strong clustering of bcd sites (z-values > 2) for distances below 50bp can be seen, especially pronounced between 20bp and 30bp (empirical p-value 0.0031). As a negative control experiment, I repeated the procedure with random sequences, which resemble real enhancers in length and dinucleotide frequencies. The random sequences exhibit no similar clustering. A second narrow peak at a 140bp distance is most likely an artifact. This gets confirmed by a conservation analysis, figure 6.1 (Right). In this graph, I counted for each distance window the number of species, in which bcd clusters significantly. I based the significance of an enrichment on empirical p-values ($p \leq 0.01$), which I calculated by repeating the enrichment analysis 10000 times with randomly shuffled enhancer sequences. I shuffled dinucleotides so that the sequences are similar to enhancers but without any meaningful binding site architecture. The empirical p-value for a certain distance is the fraction of shuffled data sets that show stronger enrichment at this distance than the real enhancers. In contrast to the probably spurious cluster at a 140bp distance, the site clusters below 50bp distance are conserved in up to 10 *Drosophila* species. The results for other TFs are in the appendix IV.

Although pronounced, the clustering does not necessarily prove actual cooperativity, especially since similar p-values are not unexpected due to the multi-testing problem[1]. Thus, I performed a conservation analysis as further validation. Figure 6.1 (Right) illustrates to which degree enrichment is conserved in other *Drosophila* species. I count an enrichment (or depletion) as conserved in a species if its empirical p-value is below 0.01. The majority of species (more than 75%) exhibit an enrichment of bcd pairs on short interaction distances (10bp – 30bp), indicating a functional importance of bcd clustering.

Judging from the bcd data, an interaction range of 50bp seems to be a good starting point to explore the distance dependence of cooperativity. In theory, a too short range might miss the full extent of the interaction, while a too long range would hamper training the cooperativity parameter due to spurious interactions. In practice, neither the range as a hyperparameter nor the shape of the interaction function has a striking influence on the test results. I tested four interaction ranges for the uniform cooperativity model from 0 (no interaction) to 100bp, as well as a linear and a Gaussian cooperativity model with 50bp range each, see figure 6.2 right. Although some differences are significant, e.g. for both, 8 TFs and 17 TFs, the model with 50bp uniform interaction is an improvement upon a model without interaction, however, the effect is very small. In accordance with the clustering analysis, it seems that a range of 50bp is optimal, but the results are hardly decisive.

[1] The concept of the multi-testing problem is that at any significance level false positives are not rare if enough tests are performed. E.g. one expects one false positive result among 100 tests at the significance level of 0.01.

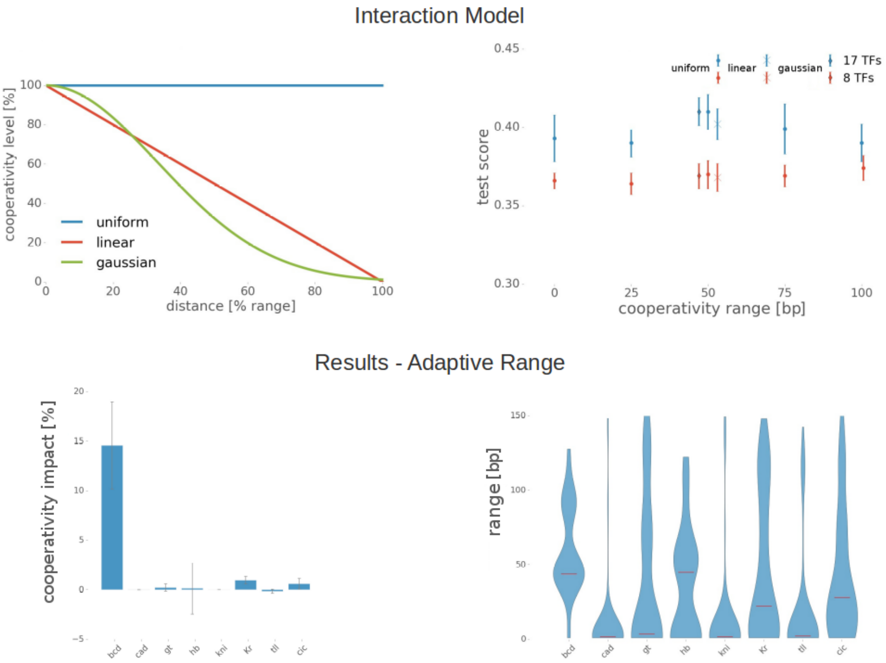


Figure 6.2: (Upper Left) The distance profile for three different models of interaction. Technically, only the linear model is continuous at 100% range, although the gap of the gaussian model is negligible. (Upper Right) test scores for uniform models with various ranges (0 - 100bp) and linear, as well as gaussian interaction (50bp). (Lower Left) Impact of the homotypic cooperativity as percentage of the test score. (Lower Right) Spread of the interaction range: parameter results for 100 predictions.

Defining the maximal range as a fixed hyperparameter has the draw-back that all TF interactions operate with the same range, which is not necessarily the case. Furthermore, the range parameter is only a small element of the whole model. By defining the range at the beginning, the optimization algorithm is able to adapt the training parameters to it, rendering the test results barely distinguishable. Both problems can be avoided by making the range of every interaction pair a training parameter. In this fashion, the optimization algorithm selects the optimal range during the parameter training.

In more detail, I implemented the maximal range as a training parameter for the prediction. In an effort to constrain the number of training parameters I tested the model with the reduced set of TFs and homotypic cooperativity. An additional parameter penalty pushes the range parameter back to zero. Given the earlier results, it is not surprising that training the range does not improve the test score. The final test score is 0.364 ± 0.005, which is worse than the fixed $50b$-range model, possibly due to the additional training parameters. Hence, it is possible to conclude that the small impact of cooperativity does not stem from the fact that we have not found the critical interaction range.

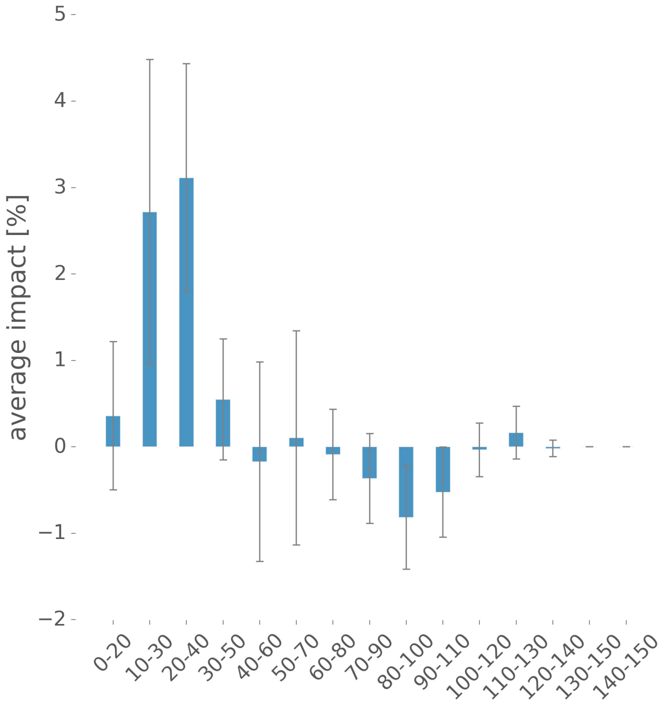


Figure 6.3: Impact of different range segments for the bcd homotypic interaction as percentage of total test score.

More interesting than the score is the parameter training result, figure 6.2 lower right. The TFs cad, kni, tll, and gt show no form of cooperativity except some outliers. Consistently, the optimization algorithm assigns vanishing ranges. The homotypic interactions of Kr and cic have a significant impact but are overall negligible. The data on hb is divided. Some enhancers gain from hb cooperativity while others disagree, resulting in zero average impact. The only clear case is bcd. It is the only TF for which cooperativity has a measurable impact and, therefore, shows the most consistent clustering of the range parameter. The median range is close to 50bp, which is also a cluster for the non-vanishing hb cooperativity.

An impact analysis reveals which interaction-distances drive cooperativity. The idea of this analysis is similar to the impact of TFs or cooperativity parameter but instead of knocking-out all or some TF parameters *in silico*, one deactivates cooperativity for certain distance windows and compare the result with the full model. Figure 6.3 depicts the impact of various overlapping bcd interaction windows. The result is very noisy due to the rather small impact of the cooperativity parameter in general, however, interactions over a distance between 10 bp to 40bp are clearly the main source of impact. Interactions on longer range scales (50b+) show either vanishing or even negative impact. This result agrees well with clustering analyses of bcd binding sites.

Compare the impact to the enrichment and conservation analysis. All indicate accordantly the relevance of clusters of bcd binding sites and homotypic cooperativity of bcd within a range of 50bp.

6.1.2 Orientation

The model is fundamentally strand-invariant on the level of the whole enhancer because it scans both strands for TF binding sites. Using the reverse DNA-strand as input sequence does not alter the result, which reflects the orientation-independent behavior of enhancers in vivo [110]. Nevertheless, the local orientation of binding sites can have an impact on their interaction with their surrounding sites, e.g. in the case of dimer formation [111, 112].

In order to implement an orientation-aware interaction model, let us distinguish whether two TFs bind on the same DNA strand (+ orientation) or on opposite strands (− orientation). The new interaction model is a modification of the Gaussian model with an additional training parameter per interaction, called the skewness s, that controls the orientation bias.

$$\gamma(i,j,d,o) = 1 + \gamma_{i,j}\left(1 + \operatorname{erf}\left(o_{i,j} s \frac{3d}{R}\right)\right)\exp\left(-\frac{9d^2}{2R^2}\right) \tag{6.2}$$

Here, the orientation of sites i and j is $o_{i,j} = \pm 1$. The function erf() is the error function, which is implemented in most programming libraries.

$$\operatorname{erf}(x) = \frac{1}{\sqrt{\pi}} \int_{-x}^{x} \exp\left(-t^2\right) dt \tag{6.3}$$

The error function is point-symmetric in $x = 0$ and maps $[-\infty, \infty] \to [-1, 1]$. An unbiased model has skewness $s = 0$. By altering sign and value of s, the optimization algorithm can choose direction and magnitude of the orientation bias, see figure 6.4.

Introducing the skewness increases the number of training parameters. Hence, additional parameter regularization is necessary to prevent overfitting. This can be done by adding the L1-norm of all skewness parameters to the parameter penalty. In this fashion, the model gets pulled back to an unbiased state with $s = 0$ if the gain does not outweigh the penalty.

I tested the orientation-aware model with the reduced set of 8 TFs and 50bp-range homotypic cooperativity for all TFs although most of them showed marginal signs of cooperativity because some TFs could display orientation biased cooperativity exclusively. However, the resulting orientation bias is vanishing, except outliers and small traces for bcd. Not surprisingly, the impact of the skewness is negative or insignificant. This indicates that - at least at the coarse level of our interaction model - cooperativity is orientation independent.

6.1.3 Heterotypic Interactions

Most thermodynamic models include homotypic cooperativity by default. In the same fashion, as homotypic interaction is implemented, it is possible to implement heterotypic cooperativity by adding a parameter for every TF combination.

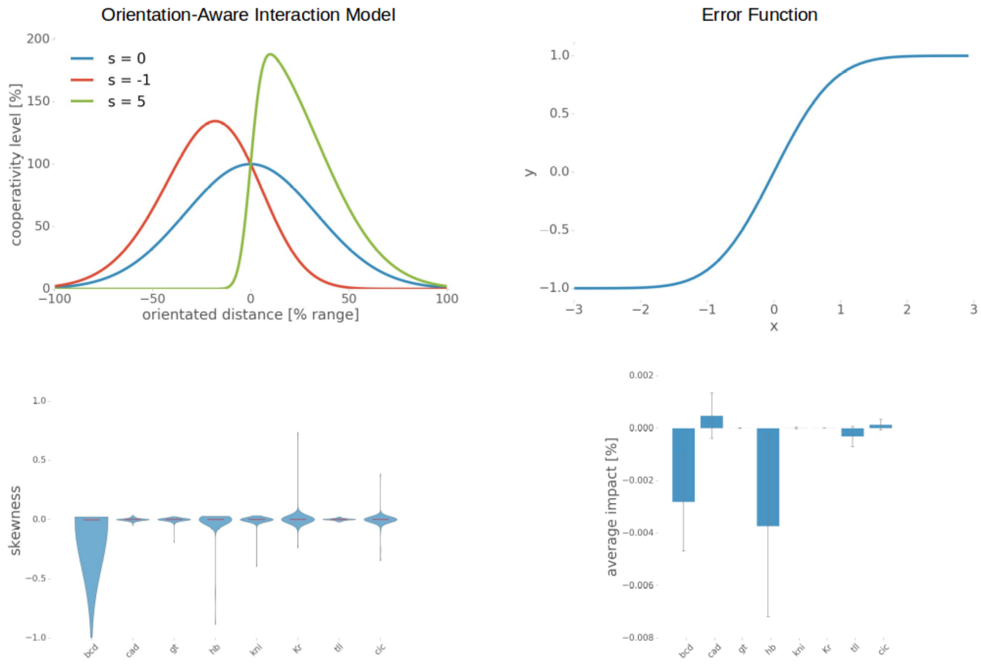


Figure 6.4: (Upper Left) The orientation-aware interaction function with three different values of skewness. (Upper Right) The error function as it is used for the interaction model. (Lower left) Predicted skewness parameters. The parameter penalty holds the skewness at zero for most TFs. Only the high impact bcd homotypic interaction shows preference for opposite strands. (Lower Right) The impact of the skewness parameter and the cooperativity parameters.

Even if only TF pairs that have overlapping expression domains are considered, implementing heterotypic cooperativity increases the number of training parameters substantially. I define TFs as overlapping and consider them for heterotypic interaction if their expression domains have a positive correlation. This increases the number of cooperativity parameters from 8 to 22 in the case of the reduced model and from 17 to 67 for the expanded model. In both cases, the number of training parameters approximately doubles.

Including this amount of parameters increases the computation time of the prediction substantially. Therefore, I tested only 5 different 10-fold CV parsings for the expanded model. The trend is nevertheless clear. Both, the reduced as well as the expanded model, deteriorate significantly when heterotypic cooperativity gets included. The reason is overfitting due to a large number of parameters. In addition, most heterotypic interactions show a negative or vanishing impact, see appendix C.8.

Adding all heterotypic interactions at the same time is therefore not an option. For this reason, I concentrate on bcd and hb. Experiments suggest that hb could change its role from being repressor to an activator in the presence of bcd sites [7, 8], but details of the exact mechanism are unknown. For the model, I included interaction between bcd

Table 6.1: Test results for 8 TF and 17 TF trained with only homotypic and with heterotypic cooperativity. The heterotypic results are significantly worse (Wilcoxon signed-rank test $p_{8\mathrm{TF}} \leq 0.01$, $p_{17\mathrm{TF}} \leq 0.05$)

Cooperativity	homotypic	heterotypic
8 TF	0.37	0.35
17 TF	0.41	0.375

and hb in addition to homotypic cooperativity. I further gave the bcd-hb interaction the possibility to display repressive behavior. By allowing the interaction function to take on parameters between 0 and 1, it can disfavor the simultaneous binding of both factors. To stay in the picture of equation 2.23, binding of one TF builds an energy barrier, hampering simultaneous binding of the second TF. To distinguish both variations of cooperativity from each other, I call them positive and negative cooperativity.

But what does the combined binding of an activator and a repressor mean for the model prediction? The answer to this question depends not only on the cooperativity parameter but also on the activatory potentials (β_a,β_r) of the involved TFs. The combined state of both TFs is effectively activating if the activator is stronger, i.e. the combined potential $\beta_a \cdot \beta_r > 1$, and repressive if the repressor is stronger, $\beta_a \cdot \beta_r < 1$. Thus, if a weak repressor interacts with a strong activator, positive cooperativity promotes expression, while negative cooperativity hampers expression. Training the bcd-hb interaction yields a

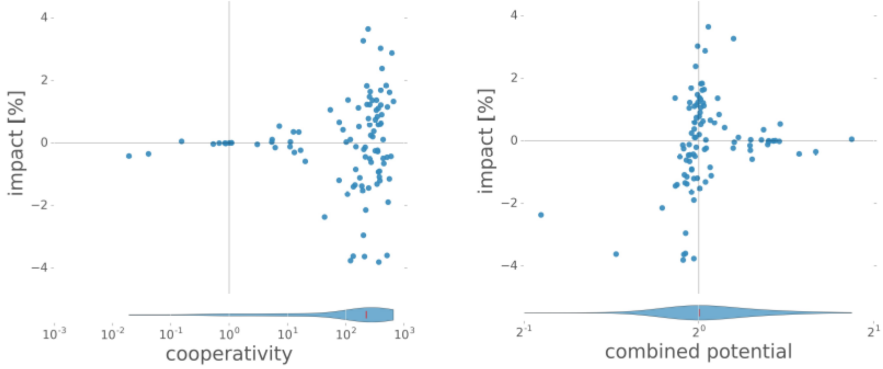


Figure 6.5: (Left) Scatter plot of predicted cooperativity parameters for the heterotypic bcd-hb interaction and their impact for 100 different CV-parsings. The impact is always normalized by the average prediction score. (Right) A similar scatter plot with the impact and the combined potential of bcd and hb $\beta_a \cdot \beta_r$.

positive cooperativity in most cases, however, its impact is inconsistent. The cases with positive and with negative impact balance each other out, leading to a vanishing average impact. In an effort to explain this behavior, I created two scatter plots. The left side of figure 6.5 depicts the spread of the cooperativity parameter and its impact for every CV-parsing. The right side is a scatter plot of the combined potential $\beta_a \cdot \beta_r$ and the impact.

Unfortunately, the data does not show a consistent trend. The impact's magnitude depends on the cooperativity strength, as expected, but no distinctive cooperativity parameter is noteworthy. Interestingly, the activatory potential of bcd and hb cancel each other out in most cases. Neither the predictions with an activating nor with a repressive combined potential show a consistent impact.

In conclusion, although the optimization algorithm clearly favors a strong positive bcd-hb cooperativity, there is not an actual sign of interaction between those TFs at this level. One reason could be that cooperativity modifies the likelihood of combined binding, but does not simulate a switch-like behavior of the TFs' role as it is reported for hb and bcd. A form of interaction that could do that is synergy, which will be discussed in the following.

6.1.4 Synergy

Cooperative binding is one form of TF interaction. As it is implemented in the model, it boosts the combined weight of interacting sites, thereby rendering the configuration with both sites being bound more likely. Once bound, the TFs interact independently with the core promoter according to the model. This holds true for negative cooperativity, which works similarly, but hampers simultaneous binding. From a modeling point of view, cooperativity affects the binding affinity, but not the activatory potential. In order to distinguish it from cooperativity, I call an interaction that affects the activatory potential of simultaneous binding TFs synergy, see figure 6.6.

How is synergy implemented and how is it different from cooperativity? As always, it is best to approach synergy from an energy point of view. Remember that the activatory potential can be seen as a form of binding energy between a TF and the core promoter. A positive energy promotes binding of PolII, negative energy hampers it. The combined potential of two TFs can be seen by adding their activatory energies, which is the same as multiplying their activatory potentials, see equation 2.30. Synergy introduces a correction term for the sum of activatory energies.

$$\epsilon = \epsilon_1 + \epsilon_2 + \epsilon_{\text{synergy}} \tag{6.4}$$

Translated to the level of the training parameters, the synergy parameter is a multiplicative factor for the activatory potentials of two simultaneous bound TFs.

$$\beta = \beta_1 \cdot \beta_2 \cdot \gamma_{\text{synergy}} \tag{6.5}$$

In a way, synergy works similar to cooperativity. When calculating the partition sums, the model inserts γ_{synergy} to the combined potential $\beta_1 \cdot \beta_2$. Unlike cooperativity, this effects Z_{ON} only. Furthermore, with $\gamma_{\text{synergy}} \in [0, \infty)$, the synergy parameter can change the role of the involved TFs by pushing the combined potential above or respectively below 1.0. In this fashion, two repressors could, in theory, become activators, if bound synchronously.

Let's try to replace cooperativity with synergy in the predictions. For the interaction function, I choose simple 50bp-range, uniform, homotypic synergy, similar to the interaction function of the cooperativity. The resulting test scores for the synergy prediction are

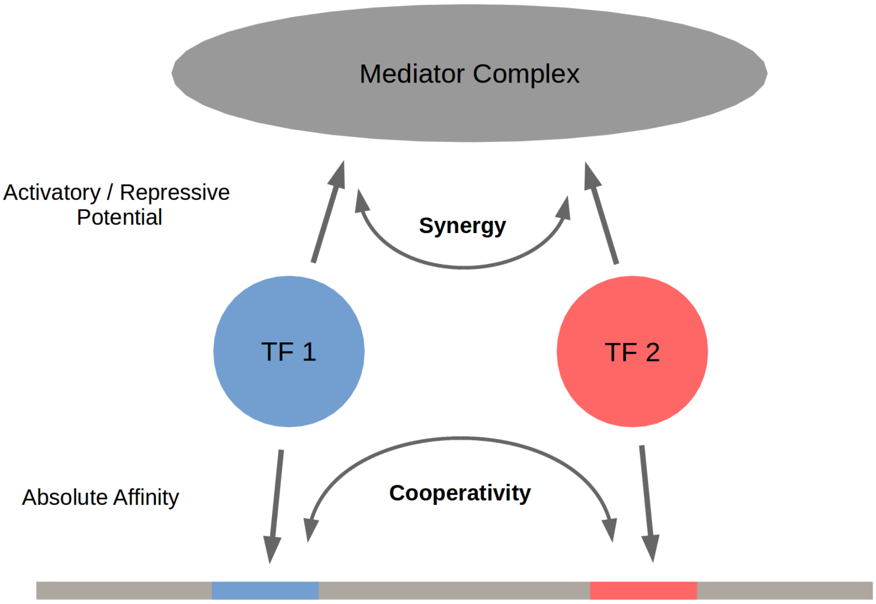


Figure 6.6: (Upper Left) The orientation-aware interaction function with three different values of skewness. (Upper Right) The error function as it is used for the interaction model. (Lower left) Predicted skewness parameters. The parameter penalty holds the skewness at zero for most TFs. Only the high impact bcd homotypic interaction shows preference for opposite strands. (Lower Right) The impact of the skewness parameter and the cooperativity parameters.

significantly worse than with cooperativity (score 0.355 ± 0.01, Wilcoxon signed-rank test $p \leq 0.01$). The predicted synergy parameters show no clear trend and, therefore not surprisingly, the average impact of each synergy parameter is smaller than the standard deviation (data not shown). Hence, synergy is not a compelling alternative model to replace homotypic cooperativity. To explain this result, consider that synergy works similarly to cooperativity. The one thing in which synergy is different from cooperativity is that synergy can change the role of simultaneously bound TFs. However, there is no clear evidence that a homotypic pair of repressors becomes an activator or vice versa.

To find an application for synergy, consider the bcd-hb interaction. I use the standard model with 8 TFs and homotypic cooperativity. In addition, I allow synergistic interaction between bcd and hb within a range of 50bp. The resulting synergy shows a clear trend, but it does not agree with the hypothesis that hb changes its role to an activator if close to bcd. The predicted synergy parameter is almost exclusively below 1.0, indicating a repressive behavior of hb. This coincides well with the reported role of hb as a short-range repressor, although the impact of this interaction is too low to confirm it, see figure 6.7.

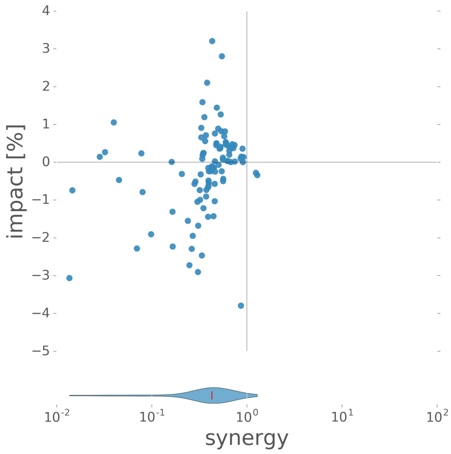


Figure 6.7: Scatter plot of the bcd-hb synergy parameter and its impact for 100 training instances. The model clearly prefers a repressive synergy, but the average impact is close to zero.

6.1.5 Focused Discussion: the Impact of Interaction Terms

TF interactions, synergy or cooperativity, are second order correction terms to the first order single binding site model. The preceding sections showed that the impact of interaction terms is very low in comparison to the impact of the TFs as first-order elements. There are two reasons for this.

First, configurations with two simultaneously bound TFs are usually less likely than single binding events. Consider two TFs that are independently bound 10% of the time (absolute binding probability $p = 0.1$). Thus, both TFs are bound simultaneously only 1% of the time. Even though this calculation is a huge simplification, it exemplifies the weight ratio between single and simultaneous binding events. Cooperativity changes this ratio but only for specific configurations.

This leads us to the second reason. Binding sites are highly abundant, see 5.1.1. Most TFs have at least weak sites in most enhancers. Therefore, the optimization algorithm has a rich data basis to fine-tune the first order parameters (binding affinity and activatory potential). By contrast, two-site configurations, especially when only specific distances or orientations are considered, are comparatively infrequent. Therefore, training the interaction parameters is more difficult and the results are noisy, as we have seen.

The first argument holds true regardless of the data used in the study and argues that interaction terms might inherently have a low impact. The second reason is entirely a shortcoming of the data set. The homotypic bcd cooperativity, which is clearly visible in the results, demonstrates that given enough data, it is possible to identify relevant interactions. Even if enhancer identification and measurement were not a problem, there would still be a limited number of natural enhancers in the anterior-posterior patterning paradigm of *Drosophila*, which is already expanded by considering multiple species.

What can be done to reduce the lack of data? An answer could be synthetically designed enhancers. These enhancers are tailored to a specific research question such as TF interaction. Although still cumbersome to design and to measure they could expand the data basis in many relevant directions. E.g. a set of synthetic sequences could be used to systematically probe the range of a certain TF interaction. Without synthetic enhancers we are limited to analyze the sequences that evolution provided us with.

6.2 Enhancer Structure

6.2.1 Weak Binding Sites

I already analyzed the role of binding site numbers and energy thresholds, see section 5.1.1. The analysis concentrated on the influence of site numbers on the training process. In section 5.1.1, I defined the threshold for the identification of binding sites before the parameter optimization and found the surprising result that the test score does not level for an increasing amount of sites. Instead of assessing the role of weak binding sites, I allowed the algorithm to adapt to the given set of binding sites and measured its response.

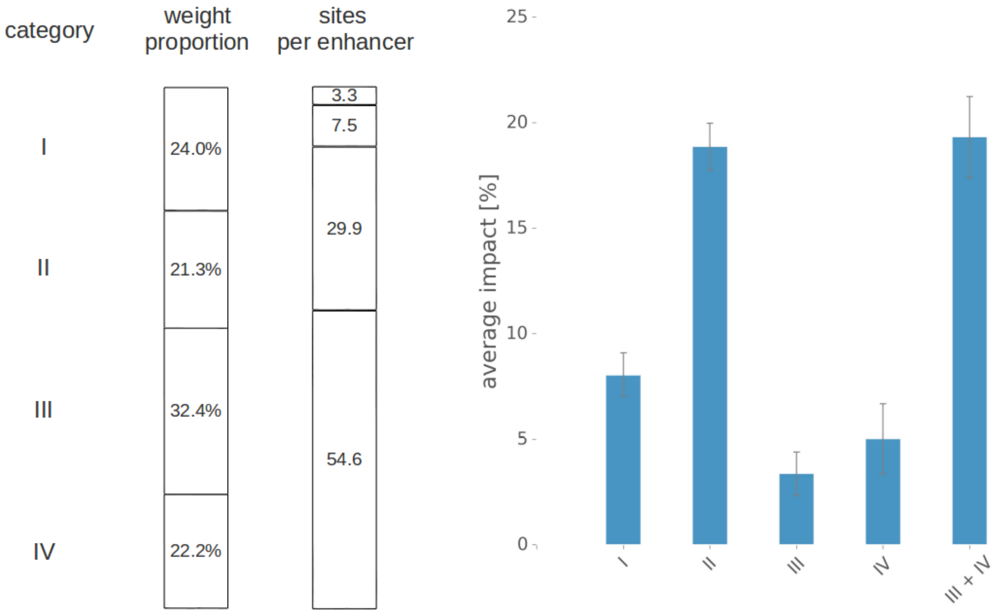


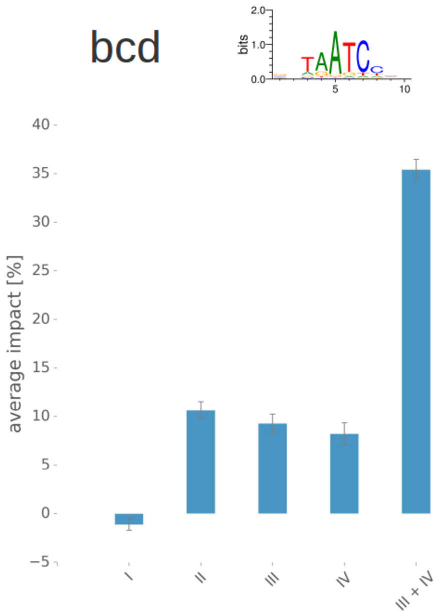
Figure 6.8: (Left) The four categories of sites, their relative share of total binding weight, as well as the number of sites per enhancer. (Right) The average impact of each category on the prediction of the 8 TF model.

In contrast to the above mentioned alteration before the training, which I call in this context prior analysis, I now perform a posterior analysis, in which I alter the model after the parameter training and measure its impact. While a prior analysis measures the adaptability of the model optimization to hyperparameters, a posterior analysis investigates the importance of model elements. A posterior analysis is always specific to the parameter training outcome. I always average over 10 independently trained models (different CV-parsings) to generalize the result of the posterior analysis.

To determine the influence of weak binding sites, it is necessary to measure their impact after parameter training. In theory, it is possible to compute the impact of every single site, however, their impact is too small to be meaningful. Therefore, I partitioned all sites into four categories (I, II, III, IV) based on their relative binding weight ($0 \leq T_I < 0.125$, $0.125 \leq T_{II} < 0.25$, $0.25 \leq T_{III} < 0.375$, $0.375 \leq T_{IV} < 0.5$; for the definition of T, see equation 5.1). Category I comprises the strongest sites, category IV the weakest sites. Although the site counts differ, the sums of relative weight are similar for all categories, see figure 6.8.

I calculate the impact of each category by deleting all sites of this category and comparing the test results. The calculation was performed for the 8 TF model with homotypic cooperativity, figure 6.8. The results are similar for other models (data not shown). The first observation is that the overall impact of most categories is relatively low. One can delete a substantial amount of sites per enhancer without a dramatic effect.

Although categories III and IV do not carry substantial importance on their own, deleting both categories of sites at once has a huge impact. This speaks to the relevance of weak sites. Especially noteworthy is that the strongest sites do not carry the most impact. In fact, the by far most important sites are in category II. These sites, approximately 7.5 per enhancer, are still strong sites but contain some deviation from the consensus sequence.



category	exemplary site	binding weight
I	GCTAATCCGT 44 sites	0.537
II	GGTAATCAGT 136 sites	0.138
III	AGAAATCCCG 289 sites	0.081
IV	GGAAATCACA 677 sites	0.027

Figure 6.9: (Left) The average impact of bcd sites partitioned in four categories of site strength. (Right) Exemplary sites of each category. The number of sites in each category are based on 98 enhancers.

There is one caveat. Every PWM generates a unique weight distribution, which is difficult to compare with distributions of other TFs. Could it be that TFs cluster in certain categories because of their different nature and that the categories represent groups of TFs and not so much binding strength? In other words, is category II so important because it contains, for example, most of the relevant bcd sites?

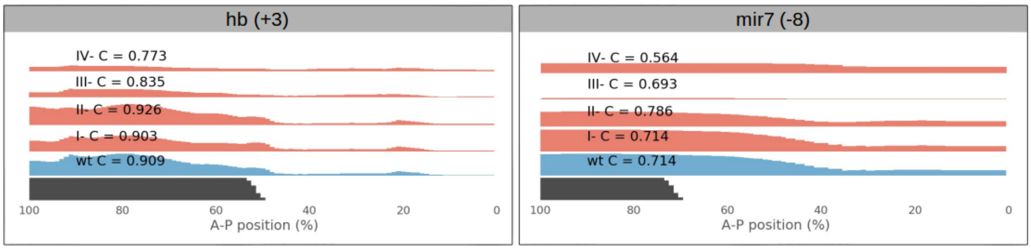


Figure 6.10: Two example enhancers, their expression prediction (wt, blue), and knock-out predictions of bcd categories (I-IV, red). Also depicted, the test correlation C. Relatively weak bcd binding sites (categories III and IV) have the biggest impact in both enhancers.

To address this concern, I repeated the analysis with a single TF. Because of its importance, I choose bcd. Figure 6.9 depicts an exemplary site and its binding weight for each category. In sum, the binding weight of bcd sites is distributed evenly over the four categories. The impact of each category is shown in the same manner as before, but now only the bcd were removed. Interestingly, category I, which represents approximately one bcd site in every other enhancer, has on average a negative impact. These sites, although almost perfect matches, largely do not determine the result of the prediction. Categories II, III, IV carry similar impact, but no category on its own is solely responsible for the majority of bcd activity. This would suggest that single sites do not affect enhancer expression at large. The expression is rather controlled by large groups of sites, even if these sites are comparably weak. Figure 6.10 shows two exemplary enhancers that illustrate this effect.

This fuzzy logic renders the language of transcription control difficult to understand and to model but might be advantageous from an evolutionary perspective. If the correct expression of a gene would rely on single sites, a single base mutation could be lethal for the embryo. Encoding the patterning information into clusters of fuzzily defined sites improves the robustness of the system.

6.2.2 Steric Hindrance

Most thermodynamic models assume steric hindrance, i.e. two factors can not bind overlapping binding sites. Obviously, TFs occupy some space on the DNA when bound and prevent other factors from binding. Bacteria use this as a simple mechanism for repression [113]. Since it is not known how much space a TF occupies, the model excludes all configurations in which overlapping sites are simultaneously bound. Here, the size of a binding site is determined entirely by the PWM. A rough size estimate shows the plausibility of this model. The TFs have a diameter of approximately $4nm - 5nm$ [114], while their motifs are around $10bp - 16bp$ long, equaling $3.4nm - 5.4nm$. Furthermore, one could argue that every relevant position in the motif is certainly in contact with the TF. Whether steric hindrance has an effect on the model or not depends additionally on the site occupancy. If the sites are occupied for only a small fraction of the time, they rarely interact and steric hindrance is irrelevant. On the other hand, if the sites are occupied most of the time, steric hindrance becomes relevant.

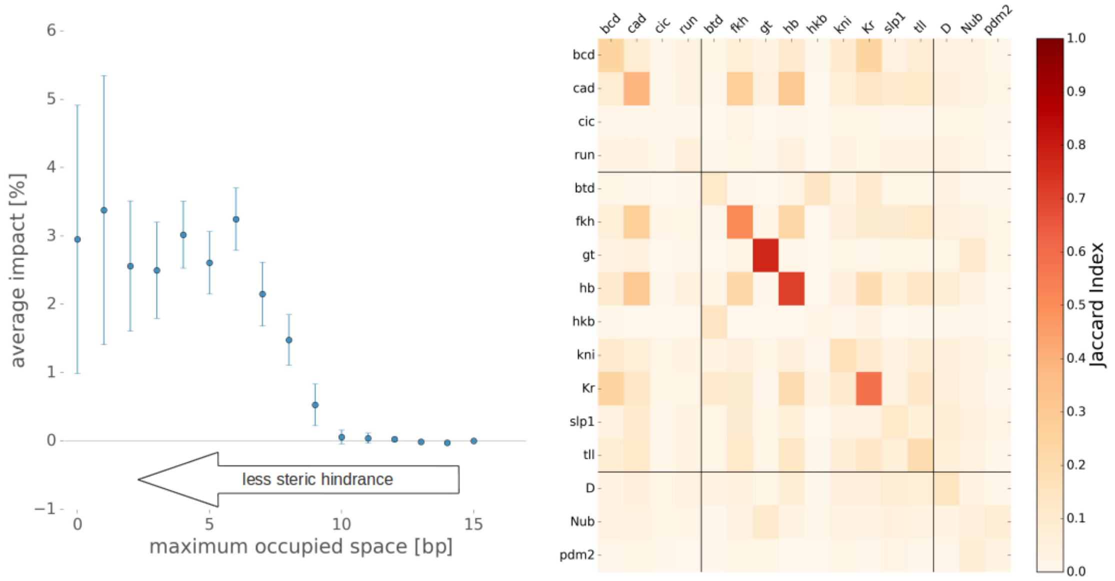


Figure 6.11: (Left) Impact of steric hindrance. By constraining the size of a site, one takes steric hindrance stepwise out of the model. (Right) The level of site overlap. The Jaccard index measures how often sites of two TFs overlap in comparison to the total number of their sites. A Jaccard index of 1 means that all sites overlap at least once with the other TF.

Figure 6.11 (Left) depicts the impact of steric hindrance in a stepwise fashion. I used the 8 TF model, which was trained with steric hindrance and calculated its impact by constraining the maximal occupied space of a TF base by base. At maximal site size of $0bp$, binding sites stop overlapping because they occupy no space. The impact of steric hindrance is small but measurable (approximately 3% of the total test score). The right side of figure 6.11 shows which factors tend to overlap the most and are, therefore, most likely the cause for the impact of steric hindrance. Especially noticeable are hb, Kr, and gt, which overlap homotypically, because they have a repetitive (hb and Kr) or symmetric (gt) motif, see figure 3.3.

6.3 Accessibility

6.3.1 Modeling Open Chromatin

The models discussed until this point assume that all sites are equally accessible to proteins and that binding depends solely on sequence recognition. Nevertheless, epigenetic factors, especially chromatin accessibility, also play an important role. Chromatin accessibility measured by e.g. DNase I correlates well with TF occupancy [115, 116] and incorporating accessibility information improves predictions of TF binding [117, 118]. Since binding sites are ubiquitous throughout the genome due to their unspecificities, open chromatin is an essential feature that distinguishes enhancers from non-regulatory sequences [119].

How can one utilize accessibility data in the model? Certainly, if the enhancer is not properly confined, including accessibility can help us to depreciate spurious binding sites. But are enhancers just open in general or is there an interior fine-structure? Peng et al. investigated this question and came to the conclusion that incorporating accessibility improves the expression predictions [35]. Their heuristic model identifies accessibility with local DNase I cleavage site density in 20bp windows, ranked and normalized on a genome-wide scale (DNase-seq). An additional parameter θ translates the rank-normalized values r_{acc} into a prior for TF binding.

$$W \to e^{-\theta(1-r_{\text{acc}})}\, W \tag{6.6}$$

A value of $r_{\text{acc}} = 1$ indicates the highest genome-wide cut-site density; $r_{\text{acc}} = 0$ indicates a closed site. The data that Peng et al. used for their investigation was fairly limited (39 enhancers modeled with 6 TFs) and of comparably low-quality (B1H PWMs, approximately 13 million mapped DHase I cleavage events). Hence, I aimed at reproducing the result with the improved modeling setup and deep-sequenced data. I compared accessibility data from two different enzymes, DNase I as in the original publication (86 million mapped events) and transposase Tn5 (from the ATAC-seq protocol [120]; 28 million mapped cleavage events). For the concern of this analysis, ATAC-seq works similar to DNase-seq, with the exception that the cutting bias differs.

Furthermore, I changed the accessibility response curve slightly:

$$W \to e^{-\theta(0.5-r_{\text{acc}})}\, W \tag{6.7}$$

This minor change rescales the prior so that open sites ($r_{\text{acc}} > 0.5$) actively gain weight. This change is necessary, otherwise, the absolute weight of all sites would be substantially decreased, leading to higher affinity parameters as compensation, which will be punished by the parameter penalty. Thus, the weight penalty would disfavor strong accessibility scales θ, which this version prevents. In both cases, a parameter value of $\theta \to 0$ reverts all weights back to the plain thermodynamic model. I use this fact to regularize the accessibility parameter by an L1-penalty like the other parameters.

Note that accessibility data is not available for all *Drosophila* species, hence, an accessibility model does not benefit from data augmentation and can only apply hyperparameter

Table 6.2: Average test scores (HTI, no augmentation) for predictions with and without accessibility. Also depicted is the average impact of the accessibility parameter (percentage of total score).

	No Acc	DNase-seq	ATAC-seq	Negative Control
8 TF	0.349	0.359	0.374	0.348
Impact Acc	-	$12.6 \pm 2.9\%$	$34.0 \pm 2.2\%$	$2.1 \pm 0.8\%$
17 TF	0.348	0.362	0.368	0.342
Impact Acc	-	$20.6 \pm 7.0\%$	$39.4 \pm 6.0\%$	$3.2 \pm 2.6\%$

training scheme HTI. Nevertheless, a significant improvement in test score for both the 8 TF and the 17 TF model can be seen (Wilcoxon signed-rank test $p \leq 0.05$), but only for the ATAC-seq data, see table 6.2. DNase-seq based accessibility improves the model, but not significantly. Not surprisingly, the impact of the accessibility parameter is substantial, proving that the improvement is really driven by the new accessibility information. As a negative control experiment, I trained the model with accessibility data derived from gDNA. In this experiment, DNA is first cleared of all proteins before it gets treated according to the ATAC-seq protocol. Hence, the gDNA data resembles solely the cutting bias of the enzyme, but not DNA accessibility in any cellular state. As expected, gDNA data does not improve the model prediction and the accessibility parameter has a vanishing impact.

6.3.2 Enhancer Cores

The improvements introduced by modeling the accessibility are similar to those that can be seen from data augmentation. Unfortunately, it is not possible to combine those two techniques, because there is no accessibility data for other *Drosophila* species. But there is an alternative possibility to use the information about open chromatin. When evaluating the accessibility profiles of enhancers, it is often possible to see a clear peak, which is most likely the core of the enhancer were the chromatin is open for TF binding. Since most of the enhancers are not properly delineated, cropping all flanking sequences and using the enhancer core could reduce the number of spurious sites substantially. I tested whether these cropped sequences can substitute for the full enhancer.

I tested three data sets with different values for the maximum length of the enhancer core: 300bp, 500bp, and 800bp. Any enhancer that is shorter remains unaltered. Otherwise, I identified the core as the stretch of sequence within the enhancer that has the highest cleavage-site density. Furthermore, I repeated the identification of homologous enhancers with those cropped sequences and used them for data augmentation. Everything else was left as before. This applies especially to the parsing of CV fragments.

There are two approaches: first, one retrains all parameters with the new input sequences, and second, one reuses the already trained parameters of the full enhancers and

Table 6.3: Average test scores for the enhancer cores. Maximum enhancer length was defined as 300bp, 500bp, or 800bp. Retrain means that the parameters were trained on the cropped enhancers; reuse means that we used parameters trained on the full sequences. Additionally, I performed a negative control experiment with 500bp long flanking sequences (regions of minimal accessibility).

max length	300bp	500bp	800bp	full	negative control
retrain	0.361	0.382	0.369	-	0.135
reuse	0.335	0.372	0.376	0.37	0.178

apply them onto the cropped test enhancers. Thus, a prior and a posterior analysis approach. Table 6.3 shows the test scores for both analysis steps. The results for 17 TFs are similar (data not shown). It is clear that the 300bp long enhancer cores are probably too short because the test score is clearly lower. For this dataset, the loss of prediction quality is less severe when one retrains the parameters, as it is expected because in doing so the training algorithm can adapt to the smaller number of binding sites. Datasets with longer enhancer cores are even able to improve on the full sequences. Especially pronounced is the improvement for 500bp long enhancers with retrained parameters, which is significant at a confidence level of $p \leq 0.05$.

But one can not use any stretch of sequence from the enhancers. Sequence accessibility is crucial for the selection of the enhancer core. As a negative control experiment, I selected 500bp long flanking sequences, i.e. the least accessible regions of the enhancers. In contrast to the most accessible regions, these sequences can not substitute the full enhancers as can be seen in table 6.3.

However, those are only average scores. In reality, every enhancer is a special case. Figure 6.12 (Left) shows a scatter plot of test scores. In this plot, I compare the predictions for the full and the core sequences. For more comparability, both predictions use the same parameters; only the sequences are different. It is clear that the score difference between core 500bp and full-length enhancer are more or less symmetrically distributed around the diagonal so that the average scores are similar. The right side of figure 6.12 presents two exemplary cases: tll(-3) and kni(-1). The accessibility landscapes within the enhancers have some similarity. Both enhancers are fairly long with an open region at the boundary, but only the prediction for tll(-3) benefits from cropping the sequence. The prediction of kni(-1) deteriorates strongly.

In conclusion, most enhancers are not properly delineated. Accessibility information can help to find the functional core of the enhancer but it is difficult to crop the sequence of all enhancers to a suitable length. A reason could be that the accessibility landscape does not fully capture the chromatins susceptibility to TF binding. The whole process of nucleosome and TF binding might be much more dynamic and dependent on the local concentration ratios of TFs.

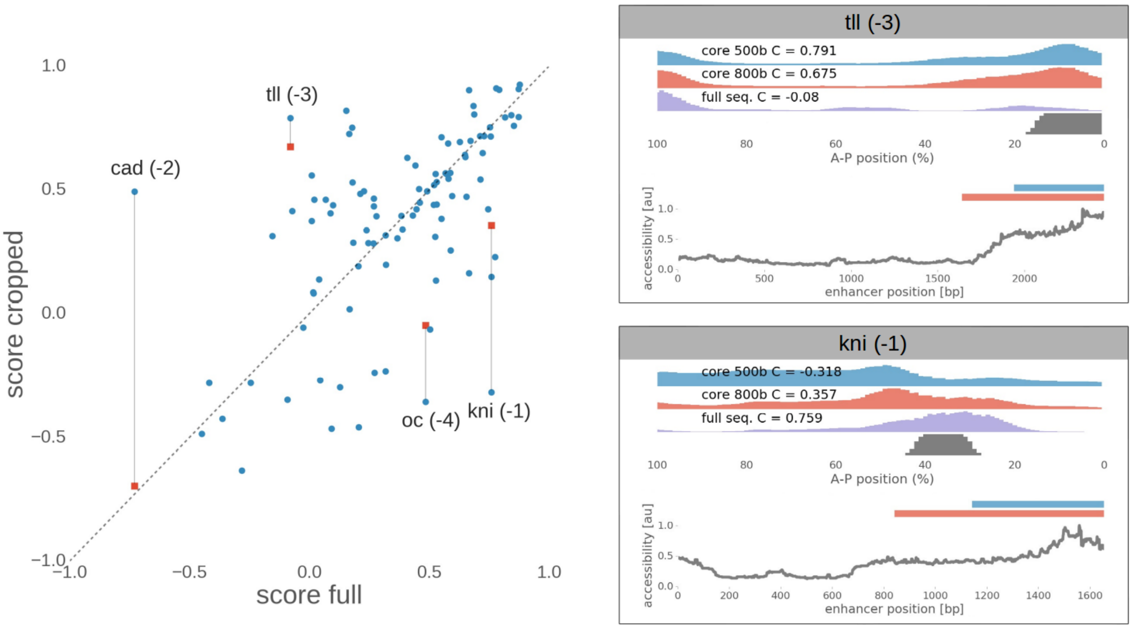


Figure 6.12: (Left) Scatter plot comparing prediction results for the full enhancers with cropped enhancers (blue circles: max 500bp; red squares: max 800bp). The predictions are superpositions of ten independent CV parsings. (Right) Predictions for two exemplary enhancers and their accessibility landscape marked in the scatter plot on the left.

Chapter 7

Alternative Models

Up to this point, only modified versions of the standard thermodynamic model were considered. Characteristic of the standard model is that all binding events are accounted for individually and that activators and repressors are fundamentally similar. In the following, I am going to discuss models that ignore these premises.

7.1 Simplified Model

The standard thermodynamic model is fairly complex and computationally intensive because it needs many training parameters and it accounts for every binding site. Especially the repeated calculation of the partition function requires substantial computation resources. Furthermore, two training parameters determine the impact of a TF in the standard model: the absolute affinity and the activatory potential. The first measures how strong the TF binds to the enhancer, the second measures how effectively it recruits the polymerase to the core promoter. Although these are two different features of a protein, both control ultimately the response of the model to a binding site. Is it possible to combine these parameters into a single one that mediates between sequence input and expression output?

Decreasing the number of parameters reduces the complexity of the model. This could be beneficial, since the output of the standard thermodynamic model has a high variance, which is hallmark of overly complex models, see section 5.3. Therefore, an additional motivation for a simplified model is to reduce the variance of the predictions and in this fashion improve the prediction.

Here, I propose a simplified model that initially ignores much of the enhancer architecture but enables the stepwise inclusion of higher-order features. The central element of the simplified model is the convoluted variable v_j^i, which summarizes the presence of the architectural feature i for the whole enhancer j. An example is the binding content of a certain factor, represented by the summarized binding weight of all its sites.

$$v_j^i = \sum_{i \text{ site in } j} w_i \tag{7.1}$$

This is an example of a first-order convoluted variable because v_j^i represents single binding events. A second-order variable is the sum of all cooperating weights of two factors v_j^{ik}.

$$v_j^{ik} = \sum_{i,k \text{ site in } j} w_i w_k \delta(i \text{ and } k \text{ interact}) \tag{7.2}$$

The delta function δ filters all relevant interaction pairs according to predefined rules, e.g. all pairs within a certain range. The parameter p_i, which I call the expression potential, links the convoluted variables directly to the logistic expression function:

$$p_{\text{expr}} = \frac{\max(E, 0)}{1 + \max(E, 0)} \qquad E_j = q_{\text{btr}} + \sum_i c_i p_i v_j^i + \sum_{i,k} c_i c_k p_{ik} v_j^{ik} \tag{7.3}$$

Depicted here is a second-order model. Higher orders are conceivable but are likely not relevant. The basal transcription rate q_{btr} is the zero-order global scaling parameter. As with the standard model, the TF concentrations c carry the position information because they define the distribution of the input factors. The potential p is a combination of the binding affinity and the activatory potential as it fully describes the effect of the TF on the transcription rate. Activators have a positive potential, while repressors have a negative potential. Notice, that the role of the second-order terms is independent of the first-order. Hence, the simplified model is stratified and allows for a clear separation of the input features, in contrast to the standard thermodynamic model, in which binding affinity and activatory potential affects both layers. The simplified model is less constrained for the same reason. E.g. a TF could be an activator but the combined binding of two TF proteins resembles a repressor. It is this context-dependent role that the simplified model enables.

The prediction quality of the simplified model is significantly lower than the standard model. Figure 7.1 and table 7.1 depict the large margin in prediction quality between the standard and the simplified model. Why does the simplified model fail at predicting expression and what can we learn from this?

One reason could be that the optimization algorithm fails at predicting the correct roles for the TFs. However, the TFs are correctly predicted as activators (bcd and cad) and repressors (hb, gt, Kr, kni, tll, and cic; data not shown). Their individual impact is similar to their impact in the standard model. Hence, the simplified model correctly captures the coarse features of the segmentation network but fails to adapt to the measured data. Hence, there is an inherent characteristic of expression control that is not considered in the simplified model. In order to understand this aspect, let us compare the simplified with the standard model.

The simplified model has a certain linear component. The model itself is not linear because the output is constrained by a response curve, which reflects a saturation behavior, but the convoluted variables summarize the binding content in a linear fashion. In this kind of model, two weak sites can weigh as much as one strong site. The convoluted variables do not account for the number of sites nor do they consider saturation of binding occupancy. In reality, a binding site can be occupied at most 100% of the time. The

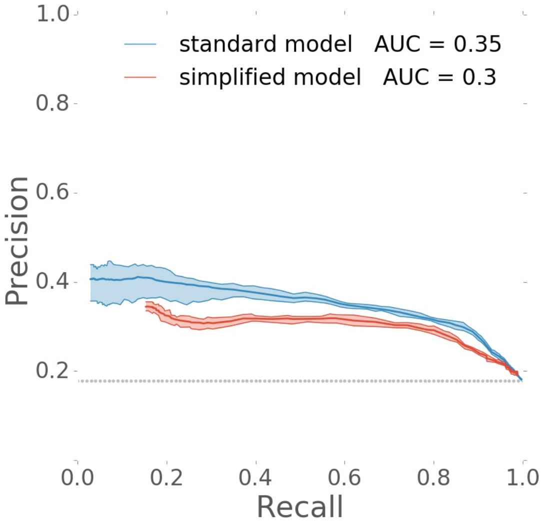


Figure 7.1: Comparison of the standard model with a first-order simplified model. Both predictions are with 8TFs.

standard model acknowledges this and distinguishes between a TF that is firmly bound even to suboptimal binding sites and a TF that is loosely bound. In the case of the latter, the occupancy depends almost linearly on the binding weight and the factor concentration, while the former has a damped response to changes in concentration or mutations in the binding motif since most relevant sites are already bound at saturation. The parameter that controls this behavior is the absolute binding affinity. E.g. bcd and cad play an important role as key activators in the standard model. However, cad has continuously two orders of magnitude lower binding affinities than bcd, but a larger activatory potential, resulting in an overall similar impact score.

But the simplified model is also linear in a different aspect: its training parameters. The potential is a linear scaling parameter of the effect of the TF. In the standard model, there is a non-linear connection between the activatory potential and the expression output. Depending on the number of binding sites in the enhancer, the partition functions resemble high-degree polynomials. The simplified model does not distinguish between the state in which a strong site is bound and a state, in which the activatory potentials of a cluster of weak sites combine in a multiplicative fashion. It is this combination of non-linear effects and the parameters to model protein binding that is missing in the simplified model and most likely the reason for the lack of prediction quality in comparison to the standard model.

7.2 Short-Range Repression

An advantage of thermodynamic models is that they model activators and repressors the same. The transition from an activator to a repressor is continuous in the standard model.

Table 7.1: Average test scores for the simplified model with only first- as well as with additional second-order terms. The standard model refers to the model and training set-up of the preceding chapters.

model	first-order	second-order	standard
8 TF	0.296	0.304	0.372
17 TF	0.293	0.321	0.41

During parameter training, the optimization algorithm classifies the TFs automatically with a remarkable precision in most cases, see section 5.2, which reduces the need for prior knowledge about these TFs. The underlying assumption is that repressors work as anti-activators hindering the recruitment of PolII to the core promoter directly. Although convenient from a modeling point of view, this assumption is certainly false for most TFs. In contrast to pair-rule gene repressors, the repressors in the group of gap factors are rather short-range repressors (SRR), which affect the activators that bind close to them, but not the core promoter directly [29].

In summary, there are two possibilities to suppress activators: first, by hampering their ability to initiate expression and, second, by preventing activator binding, which is also called quenching. In the first scenario, the activators still bind to the enhancer but their interaction with the core promoter is interrupted. For a model that incorporates SRR on this level, see [15] or section 6.1.4. In their work, Fakhouri et al. analyzed the range of the repressor gt with the help of synthetic enhancers [15]. They conclude that the repressive effect of gt has a range of approximately 80bp and does not follow a strictly monotonous decline.

Multiple publications come to the conclusion that the second scenario, in which repressors quench activators, is closer to the reality. E.g. the gap factor kni changes — with the help of the co-repressor Groucho — the chromatin state locally (increased histone density and deacetylation) and reduces activator occupancy [121, 122]. He et al. proposed a model for SRR on the basis of activator quenching [28]. They argue that their quenching model is sufficient to capture the effect of repressors, although the prediction quality deteriorates slightly in comparison to the standard model. In contrast to Fakhouri et al., He et al. based their analysis on native *Drosophila* enhancers, however, they did not explore the influence of the repression range.

Here, I propose an algorithm that models repressors as mediators of chromatin accessibility. This model is similar to the quenching model of He et al. but simplified with less free training parameters. The special nature of the algorithm in combination with the improved parameter training set-up enables us to explore the influence of repression range as well as the ramifications of SRR in much greater depth.

In the SRR model, activators are treated as in the standard model. They interact directly with the core promoter and their impact is controlled by two parameters, the absolute binding affinity, and the activatory potential. Repressors, on the other hand, act by shaping the repressive landscape R in the enhancer, see figure 7.2. Every activator

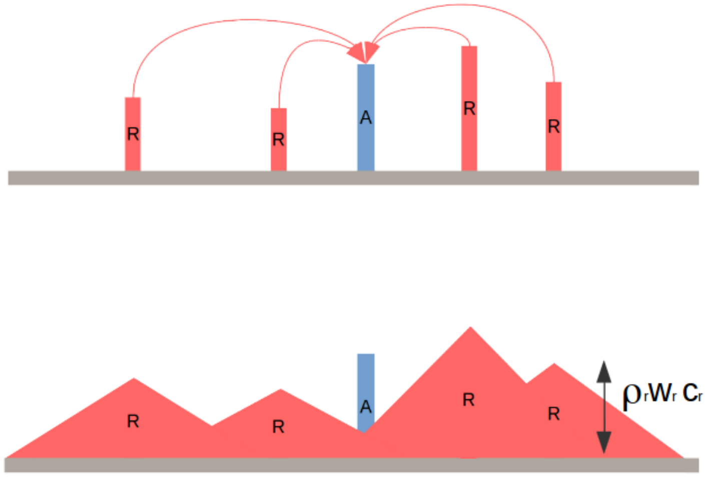


Figure 7.2: An activator site is surrounded by repressors that shape the repressive landscape. In this depiction, the repressive effect declines linearly.

binding site in the enhancer is more or less receptive to binding based on its position in this landscape, which works similar to the accessibility in section 6.3 in the sense that it acts as a prior for binding.

$$W_a \rightarrow \frac{1}{1+R} W_a \tag{7.4}$$

A high R value indicates strong repression as it lowers the binding weight of the activators preventing indirectly the activators effect. The repressive landscape has to be calculated before computing the activator function by considering the distance and strength of all repressor sites.

$$R(a) = \sum_r \rho_r c_r w_r \delta(a, r) \tag{7.5}$$

In the equation above, a is an activator site and r are repressor sites with relative binding weight w_r and concentration c_r; $\delta(a, r)$ is the distance function. An example for the distance function is:

$$\delta(a, r) = \begin{cases} 1 & \text{if distance } a \text{ to } r \ \leq 50bp \\ 0 & \text{otherwise} \end{cases} \tag{7.6}$$

The multiplicative factor ρ_r is called the repressive parameter. It is a TF-specific training parameter that controls both the repressive effect as well as the binding affinity of the TF. ρ_r sums up how strongly the TF shapes the repressive landscape.

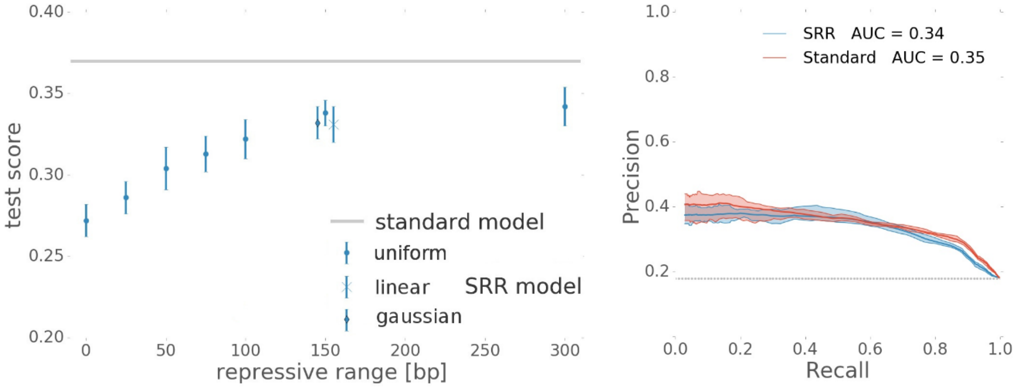


Figure 7.3: (Left) Average test scores for SRR-models with various repression ranges (0bp - 300bp) and different repression functions (uniform, linear, and gaussian). (Right) PR-plot of the 150bp-range SRR model in comparison to the standard 8TF GEMSTAT model.

The primary objectives are to test whether the repressive landscape is a viable model and over which ranges repression most likely works. To simplify the analysis, I concentrated on the reduced set of TFs, although the results for the expanded set are qualitatively similar (data not shown). Only two activators are explicitly modeled like in the standard model: bcd, and cad. Six repressors act over a short range: gt, hb, Kr, kni, tll, and cic. I applied the uniform repression function as delineated in 7.6. Figure 7.3 (Left) depicts the test scores for various repression ranges as well as a control experiment with only activators. The test score monotonously improves with increasing repressive range up to a range of 150bp, beyond this point, the test score saturates. The actual shape of the repression function is of lesser importance. Two predictions with 150bp range, one with a linear declining repressive function, and one with a gaussian decline, yield almost the same test result. The test scores are overall significantly lower than the test scores of the standard model. Additionally, the PR-plot in figure 7.3 (Right) shows that the standard model outperforms the SRR-model.

The analysis above tries to narrow down the repression distance by setting the range as a hyperparameter. In this fashion, the optimization algorithm is able to adapt the training parameters specifically to the predefined range. Hence, a posterior impact analysis can be more informative, similar to the analysis of the cooperativity distance, see section 6.1.1. The left site of figure 7.4 depicts the impact of all TFs, activators and repressors, for the uniform 150bp-range model. The overall trend is similar to the standard model, see figure 4.5; the order of importance is more or less conserved. However, the difference in impact between activators and repressors is much larger, which is not surprising because the effect of the short-range repressors is delivered via the activators. In other words, without the activators, the repressors' function is futile. The right site of figure 7.4 shows the impact of single distance segments, which can be calculated by ignoring this segment

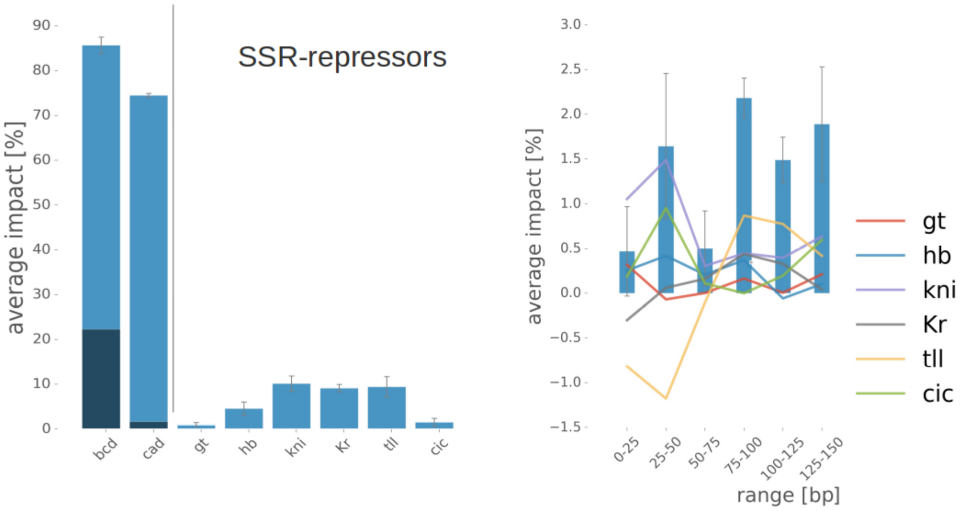


Figure 7.4: (Left) Impact of the single TFs in the 150bp-range SRR model as percentage of the test score. The TFs bcd and cad are activators and are modeled as in the standard model; gt, Kr, kni, tll, and cic are short-range repressors and rely on the activators to deliver their impact. (Right) The impact of 25bp distance segments of the repression function. The blue bars depict the impact of all repressors for this segment. The continuous lines show the impact of the single TFs.

for the computation of the repressive landscape. Overall, the impact is well distributed over the full range of 150bp as can be seen by the low impact of any single segment. The TFs kni, hb, and cic seem to operate mainly within a range below 100bp. Puzzling are Kr and tll, which seem to target activators beyond the distance of 100bp.

Part III

Discussion

Discussion

The topic of this thesis the application of thermodynamic models of transcription control and their parameter training on known enhancer expression patterns in *Drosophila* segmentation. The function of thermodynamic models is to predict expression from enhancer sequences given the binding motifs and protein concentrations of all relevant transcription factors (TF). The intention behind designing these models is to analyze body segmentation in *Drosophila* and to decipher the language of transcription control in general.

After roughly a decade of research with and on thermodynamic models, there are many aspects of transcription control that are still elusive. Key topics of ongoing research concern the interaction of TF binding sites with each other, the details of repressor mechanics, and the interaction of TFs with nucleosomes. All of these topics concern enhancer architecture and require a high degree of complexity to model. However, the small amount of training data, i.e. the measured enhancers, limits our ability to train complex models with many parameters. Most studies use up to 44 enhancers to train thermodynamic models of *Drosophila* segmentation at the blastoderm stage. A thorough literature search has revealed that there are 98 enhancers available, but although this is a substantial addition, thermodynamic models are still limited considering the large number of parameters that need to be trained.

In addition to the large number of training parameters, thermodynamic models are non-linear, non-separable, i.e. the solution can not be reduced to a combination of solutions on independent subspaces, and ill-conditioned, i.e. small variations in the input data yield large differences in the output. Therefore, before dealing with enhancer architecture, I developed an efficient model training setup, which prevents overfitting and optimally utilizes the available data. There is large body of literature on optimization problems with similar difficulties e.g. [84, 85, 93, 94, 92, 95]. I adapted some of the techniques to the setting of thermodynamic models. The resulting training setup comprises a global parameter optimization algorithm (CMA-ES), a L1-parameter penalty with hyperparameter training, and three-fold data augmentation with homologous sequences. Each of these methods independently helps to decrease overfitting and improves the prediction quality that I measured in multiple independent cross-validation instances. I was able to show that when combined, these techniques significantly enhance the prediction quality.

Model Evaluation

The first application of the improved model training setup was to compare different TF binding motifs. I found that binding motifs measured with the HIP-FA method improve the predictions substantially. HIP-FA measures binding affinities in thermodynamic equilibrium, in contrast to alternative methods like B1H and Footprinting, which utilize binding site selection. It is therefore not surprising that the HIP-FA motifs are best suited for thermodynamic models, even though their information content is typically lower, i.e. the motifs are less specific. This suggests that TF binding is indeed relatively unspecific and that transcription information is encoded in clusters of sites rather than in single binding sites.

I continued the analysis of the segmentation network by extending the model. I included additional TFs that have not been incorporated up to this point, among them the late gap factor D (Dicheate), and the early pair rule genes run (runt) and slp1 (sloppy paired 1). Only with the optimized model training setup is it possible to efficiently train the parameters necessary to model the additional TFs and prove that their addition enhances the model quality significantly. In this way, I came closer to the goal of modeling the whole segmentation system during the blastoderm stage and showed that advances in model training enable us to study transcription control in more depth.

An impact analysis revealed which TFs are most relevant for the model prediction. The impact is a quantitative score that compares the test score of the model with and without a certain TF. Note that the impact score does not consider that the remaining parameters could adapt to the missing feature, e.g. by a second parameter training for the remaining parameters (posterior analysis). The impact of a TF depends on the parameter training result. Therefore, I averaged the impact of every factor over ten independent parameter training results. The advantage of the impact score is that it measures the influence of TFs independently of indirect effects from TFs further downstream in the segmentation network. This allows to distinguish direct and indirect impact, which is not possible in *in vivo* experiments. My results reveal that the maternal activators bcd (bicoid) and cad (caudal) are by a large margin the most important TFs, followed by the gap factors cic (capicua), fkh (forkhead), gt (giant), hb (hunchback), Kr (Kruppel), kni (knirps). Especially bcd does not only initiate the segmentation cascade but is also a key TF for pair rule pattern formation during the blastoderm stage.

The impact score is also useful to investigate the importance of weak binding sites for the expression prediction. Since the impact of single sites is too small to be informative, I divided the sites into four categories according to their binding weight. The strongest sites are in category I; the weakest sites are in category IV. The number of sites is different between the categories but the total binding content, i.e. the sum of the binding weights, is similar. Since there are fewer strong than weak sites, the impact of all categories is similar. For all TFs combined as well as for bcd alone, it is not only the strongest sites close to the consensus that determine the predicted expression. Rather, the expression prediction is also based on many weak sites, which often have multiple deviations from the consensus motif. Although enhancers consisting of consensus sites only could also form reliable expression patterns, they might be prone to patterning defects caused by small mutations. The predictions of the thermodynamic model are remarkably stable even when multiple sites are deleted *in silico*. This stability can also be observed in the evolutionary difference between enhancers from diverse species. Many enhancers are conserved over comparatively long evolutionary time scales despite strong sequence divergence. The unspecific nature of TF binding together with a high degree of redundancy probably aids this stability.

Transcription Factor Interactions

It is well established that TF binding sites interact, even though the exact mechanisms are not fully known. I tested multiple alternative forms of interaction. However, introducing interaction terms increases the number of training parameters and renders the model optimization more difficult and prone to overfitting. The depth in which it is possible to explore TF interaction is therefore limited by the amount of available training data. Consider that in order to train a certain TF interaction, many instances of neighboring binding sites of those TFs are needed. Otherwise, the algorithm does not have enough training instances to draw a conclusion about the effect of the interaction. The parameter training setup will, in this case, push the interaction parameter to a neutral position in order to prevent overfitting. Hence, the limitation is especially severe for TFs with few binding sites. This limitation is a prime example of the necessary trade-off between preventing overfitting and including even marginal features into the model.

To model homotypic interactions, i.e. interaction of factors of the same type, I tested two different models of interaction, which I called cooperativity and synergy. Cooperativity, which is the model that thermodynamic models usually utilize, assumes that interacting binding sites gain extra binding energy when bound at the same time. This additional energy, originating from protein interactions, nucleosome depletion or DNA bending, boosts the binding weight and therefore the occupancy of the simultaneously bound sites. Synergy does not affect protein binding but assumes that the interaction with the core promoter is altered. Instead of affecting the binding weight, synergy introduces an extra term that affects the activatory potential of simultaneously bound TFs. The interpretation for synergy is that the activity of bound TFs can be changed by co-factors. Synergy can amplify the effect of the TF but also counteract it. For instance, there are experiments suggesting that hb changes its role from repressor to activator in the presence of bcd binding sites.

Although similar in their implementation, synergy was not able to substitute cooperativity in the model predictions, which indicates that cooperativity is better suited to describe TF interactions. I additionally tested whether synergy could model the unexplained bcd-hb interaction. Simpson-Brose et al. reported that hb shows activatory behavior in the presence of bcd sites. Yet, synergy failed to reproduce this switch-like effect. The data hinted rather towards a repressive effect of hb upon bcd sites, which fits the short-range repressor role of hb. Consider that this analysis might also lack the necessary data foundation because the switch-like effect of bcd-hb synergy as it is described in the literature might apply only to few enhancers.

For the homotypic cooperativity model, I further analyzed whether incorporating binding site orientation, the shape of the interaction function, and the interaction range have an effect on the predictions. I obtained decisive results only for the interaction range. I concentrated the analysis on bcd because it showed the strongest effect. By testing multiple interaction ranges and performing an impact analysis, I was able to narrow the interaction range down to around 50bp. This result is consistent with the strong clustering of bcd sites within the same range, which can be seen in the majority of the analyzed *Drosophila* species.

Alternative Models

The last chapter of this thesis explores two alternative models of transcription control. They are in their nature similar to thermodynamic models but differ in some key features. The simplified model is supposed to serve as an alternative to the full thermodynamic model in order to decrease the computation time and simplify the analysis. The key difference of the simplified model is that protein binding is not modeled explicitly but rather combined with the activatory function. Thus, the simplified model does not distinguish a firmly bound TF with a weak effect on the expression from a weakly bound with strong effect. This lack of detail is probably the reason why the simplified model scores consistently worse than the full thermodynamic model and proves that total binding content is not the only defining feature of enhancers. Rather, the distribution of the binding content among the sites of a TF as it is modeled by the thermodynamic model is important for the proper enhancer expression.

A second alternative model takes a similar approach in order to model short-range repression (SRR). This model depicts SRR by building a repressive landscape. The model is motivated by the idea that the repressors shape the protein binding landscape around their binding site. In this model, repressors do not affect expression directly but through preventing activator binding. For this reason, the activators carry a large impact in the SRR model. The model suggests that short-range repressors act over a fairly long range (beyond 100bp). However, experiments indicate that repressors have in fact a slightly shorter range between 75bp to approximately 100bp [15, 123]. In contrast to expectations, the test score does not deteriorate for models above a certain repressive range, in fact, the success of the standard model suggests that the effect of repression gets integrated over long ranges. Overall, the prediction quality of the SRR model is only slightly worse than the predictions of the full model. This indicates that designing a SRR model, in which activator blocking is the only mode of repression, is possible, although many details are still unclear, which leaves scope for further research.

Residual Error

Regardless of the advancements of the improved training setup, the residual error in the model predictions is still substantial. Some enhancers are consistently difficult to predict and the variance of the prediction result is high, i.e. slight variations of the training data affect the training result disproportionately. This raises the question: what is the limiting factor for further improvement of the prediction? There are three aspects to the issue of prediction quality: the parameter training method, model design, and the available data. This thesis explored parameter training in depth. Based on the currently available literature, it is unlikely that there is much more potential for improvement from this side.

The quality of the model is highly specific to the biological system. It describes how well the model is suited to capture the underlying mechanisms represented by the given data. This thesis explored some alternatives to the standard thermodynamic model. The simplified model was deliberately designed to require fewer parameters than the standard model in order to test whether the prediction quality could be improved by reducing the complexity of the model. The idea was among others to reduce the variance of the prediction, which is a major source of error, see section 5.3. Other models are designed to test alternative mechanisms of transcription control, e.g. the alternative interaction model, which I called synergy, and the SRR model. The intention was to design a model that better captures the biological reality and, thus, describes transcription control more accurately. However, none of these alternative models improve the prediction quality.

What is not considered in the modeling framework is the possibility that the TF binding is not an equilibrium process, which would require a completely different type of model, e.g. [124]. Note that I assume that the TF binding landscape on the DNA reflects the concentration of free TF at the same moment. This assumption implies that the dwell time of TFs on the DNA is much shorter than the biological development processes that the model tries to capture, e.g. the formation of a TF gradient, which takes a couple of minutes. A non-equilibrium model has to consider previous states of the system, too. However, there is evidence that the equilibrium assumption is reasonable [46, 47] and that TF binding follows closely the current TF concentration, see 2.1.3.

At least until now, it has been impossible to design a model that outperforms the standard model, e.g. in the form of GEMSTAT, to my knowledge. Therefore, the remaining aspect that affects prediction quality is the training data. This topic comprises both the amount and the quality of the available data. In order to better understand how the data can help to improve the model, an analysis of the flaws in the predictions will be constructive. The best tool for this analysis is the impact score.

With the impact analysis, it is possible to assess the influence of the TFs on every single enhancer separately. This analysis shows that every TF with non-vanishing impact influences at least some enhancer negatively. Negative impact means that the prediction deteriorates when the TF gets incorporated into the model. For instance, bcd is undoubtedly an activator that initiates transcription in the anterior part of the embryo; but bcd binding sites are also present in purely posteriorly expressed enhancers. On those enhancers, bcd has a negative impact, even though its average impact over all enhancers is positive. For instance, the enhancer D(+4) is expressed entirely outside of the bcd gradient, however, there are multiple strong bcd sites present, which resemble almost the consensus. There are multiple possible explanations for this behavior.

If these sites are spurious binding sites, it means that the predictions utilize a flawed binding motif. A reason for this could be that protein binding differs *in vitro* from binding *in vivo*. Alternatively, the binding model could be too simplistic, e.g. multiple mismatches from the consensus are considered to be independent. It would be easy to adapt the model if better binding motifs were available. The HIP-FA binding motifs, which were used in this thesis, still rely on the additivity assumption. This assumption states that the binding weight of any site can be calculated by independently evaluating deviations from the consensus at every position. Such a binding model is called zero-order, in contrast to first-order models that also incorporate dinucleotide deviations. Incorporating higher-order motifs into the thermodynamic model is simple because the motif affects only the prediction of binding weights. On the experimental side, HIP-FA can measure higher-order motifs easily and accurately too. It is not clear yet how strongly the first-order binding predictions deviate from the zero-order. Nevertheless, given the sensitivity of the thermodynamic model to the quality of the binding motifs, higher-order binding models have the potential to improve the expression prediction further. The success of the HIP-FA motifs demonstrated that this can be beneficial.

It is very likely that most of the sites that apparently do not fit into the model are correctly annotated because they are very similar to the consensus site. In this case, the problem could originate from the fact that thermodynamic models do not predict absolute expression levels but rather relative patterns. Note that the expression patterns are measured in a simplified on-off assay and get smoothed artificially in order to prevent sharp expression boundaries. This method of expression measurements ignores any fine structure and differential expression levels in the patterns. From the perspective of those measurements, it is not possible to distinguish whether the activity of a TF is spurious or whether the TF is necessary to fine-tune the correct expression level or to constrain protein expression. For instance, the TF tll represses its own enhancers. One can only speculate whether this indicates that *tll* is auto-regulated. The alternative is that these sites are inactive because they are repressed locally by other factors and by nucleosomes, or they lack necessary cofactors.

Similar to Peng et al. [35], I found that incorporating accessibility information, which reflects nucleosome binding patterns, indeed improves the model predictions. Even if the resolution is low, accessibility data can help to delineate the enhancer boundaries. Since most enhancer boundaries are not properly defined, many spurious sites could stem from inaccessible regions flanking the actual enhancer sequence. With higher resolution, the model could distinguish continuous levels of accessibility, however, it is not clear whether this directly reflects changes in binding weight. Unfortunately, incorporating accessibility and using data augmentation are mutually exclusive because the accessibility profiles of *Drosophila* species other than *D. melanogaster* are not known. For this reason, it is not possible to check the accessibility model with the fully optimized training setup.

Data as a Limiting Factor

Although heterotypic clustering is common, i.e. grouping of sites of different TFs, it is not possible to efficiently train a thermodynamic model with heterotypic TF interactions. It is implausible that different TFs do not interact. There are two explanations that are more likely. First, the interaction model is flawed. Second, not enough training data is available; or better, there is not enough data tailored to this specific question. Having enough training data is especially crucial because every heterotypic interaction term adds another training parameter to the model and, therefore, increases the model complexity. A handful of enhancers with which one could clearly discriminate correct from incorrect interaction models would improve the data situation substantially.

Unfortunately, all of the available data consists of naturally occurring enhancers that originate from millions of years of evolution. Characteristic for these enhancers is the stochastic distribution of binding sites and that multiple connected mechanisms ensure expression stability and mutation tolerance. Hence, identification of novel enhancer modules in the genome of *Drosophila* will not solve the data shortage. The number of undetected enhancers is likely small since the number of patterned genes is limited. The addition of enhancers from other *Drosophila* species was beneficial for the parameter training although the correct expression profile of these enhancers are unknown. It is possible to expand the set of homologous enhancers further, because the genomes of 12 *Drosophila* species have been sequenced and a large amount of genetic variations in *D. melanogaster* has been identified [125]. However, the positive effect of data augmentation saturated after the addition of more than three homologs. This is not surprising because the augmented enhancers follow the same design principles as the already known enhancers.

Synthetic enhancers could be a source for more training data. There are two variations of synthetic enhancers. De novo enhancers designed from scratch are simple constructs which contain only the necessary TF binding sites mostly implemented as consensus sites. The advantage of the de novo approach is its simplicity and that only a limited number of parameters are necessary to model its expression. The advantages are at the same time disadvantages. An overly simplistic enhancer does not capture the complex interactions that govern real enhancers. For this reason, de novo enhancers are a good tool to probe specific mechanisms, e.g. to test certain TF interactions, but might fall short to fully depict enhancer architecture. An alternative approach is to use derived synthetic enhancers, which are designed by starting with real enhancers and modifying them in specific details. In this fashion, the synthetic enhancers still encompass the architectural principles of real enhancers. However, due to the interwoven structure of real enhancers, it is difficult to design specific mutations targeting a single architectural aspect.

I have already mentioned that the small amount of training data limits the depth to which transcription control can be probed with thermodynamic models. I have demonstrated how smart training techniques like parameter penalization and data augmentation can help with the data shortage but do not solve it. Since the number of naturally occurring enhancers is limited, only synthetic enhancers can increase the amount of training data substantially.

There is an additional aspect of the training data that I have not discussed yet. With increasing amounts of data, data quality becomes the limiting aspect of model training. What does data quality mean? The input sequences are correct although the enhancer boundaries are not always properly delineated. However, the available enhancer data lacks the optimal resolution of the expression patterns in space, time and quantity. The experimental read-out of the enhancers that I have used is binary. Either one detects expression or the reporter is repressed. The reporter constructs neither detect expression gradients — it is common to apply an artificial smoothing to the binary expression data that accounts for diffusion and the graded input — nor fine structure in the patterns, e.g. different peak height. Since the expression prediction is continuous there will always be a discrepancy between the binary measurement and the predictions of the model. The parameter optimization algorithm tries to correct this deviation, which does not necessarily lead to better predictions. Furthermore, I have scored the differences between measurement and prediction always with a scale-free objective function because the absolute expression level is not known. For this reason, all enhancers appear to be of equal expression strength regardless of the number of activator binding sites, which is probably not true. Novel reporter constructs that can measure at least relative differences in expression level would improve the predictions substantially by providing a realistic measure of the expression patterns and gradient levels.

In this thesis, the model predicts expression for one time-point in the middle of the blastoderm stage. However, correctly staging the enhancer expression is difficult. Furthermore, it is impossible to address the dynamics of segmentation because only snapshot information is available. Live-reporters are able to measure enhancer driven expression over several minutes, which would aid embryo staging and could also illustrate the dynamics of enhancer control.

The example of HIP-FA, which provided better and more precise binding motifs, demonstrates that better data quality has an effect on the prediction quality. Similar improvements — if not greater — can also be expected from a better quality of the enhancer data. Remeasuring all enhancers is laborious and time intensive. However, measurements of new enhancers should incorporate improved reporter constructs in the future. The design of novel reporter constructs and their application to the measurement of synthetic enhancers are topics of current research at our lab.

In conclusion, thermodynamic models are a key tool to study transcription control and to decipher the language of enhancers. By improving the prediction quality of the models, it is possible to gain insights into the workings of gene regulation. There are three aspects that can limit the ability to predict expression patterns. Limited training data, model quality, and a naive parameter optimization. This thesis discussed especially the latter aspect: parameter optimization. I have shown how to train a complex, parameter rich thermodynamic model even with limited data, making the most out of the given data quality. The most severe problem that remains is the lack of data and the low data quality. The analyses that I have performed for this thesis demonstrate how beneficial an improvement of data is and that designing new and more complex models is difficult without the necessary data resolution to test and compare them.

Part IV

Appendix

Enhancer

Table A.2: The enhancers used as data basis.

Name	Alias	Length [bp]	Coordinates R5	Source
Antp (+50)	ChIP AHD-1	1500	3R:2774273..2775773	Fisher et al.
Abd-B (+20)	VT42863	2166	3R:12776398..12778564	Kvon et al.
Abd-B (+50)	VT42848	2209	3R:12745522..12747731	Kvon et al.
Abd-B (+70)	VT42837	2212	3R:12726430..12728642	Kvon et al.
btd (-3)	btd head	1799	X:9584106..9585905	Schroeder et al.
cad (-2)	ChIP AHD-2	1500	2L:20768045..20769545	Fisher et al.
cad (+14)	cad (+14)	1636	2L:20784670..20786306	Schroeder et al.
cnc (+5)	cnc (+5)	1420	3R:19020990..19022410	Schroeder et al.
D (+4)	D (+4)	1724	3L:14166136..14167860	Schroeder et al.
dfd (+13)	ChIP-miRNA9	1500	3R:2630027..2631527	Fisher et al.
ems (-22)	HC 25	1094	3R:9705258..9706352	Chen et al.
ems (+9)	HC 18	1041	3R:9736403..9737444	Chen et al.
eve (-4)	eve 37ext ru	2096	2R:5861773..5863869	Schroeder et al.
eve (-1)	eve stripe2	662	2R:5865217..5865879	Schroeder et al.
eve (+5)	eve stripe4 6	601	2R:5871404..5872005	Schroeder et al.
eve (+7)	eve 1 ru	807	2R:5873440..5874247	Schroeder et al.
eve (+8)	eve stripe5	799	2R:5874147..5874946	Schroeder et al.
fkh (-2)	fkh (-2)	1707	3R:24411719..24413426	Schroeder et al.

Name	Alias	Length [bp]	Coordinates R5	Source
ftz (-7)	ftz (-7)	1617	3R:2681761..2683378	Schroeder et al.
ftz (-6)	ftz (-6)	1239	3R:2683373..2684612	Schroeder et al.
ftz (-1)	ftz (-1)	1074	3R:2688614..2689688	Schroeder et al.
ftz (+3)	ftz +3	1744	3R:2692616..2694360	Schroeder et al.
ftz (+20)	VT37580	2091	3R:2709408..2711499	Kvon et al.
fz2 (+80)	Cluster-8458	1234	3L:19145853..19147087	Fisher et al.
gt (-10)	gt (-10)	1744	X:2331789..2333533	Schroeder et al.
gt (-6)	gt (-6)	2181	X:2327322..2329503	Schroeder et al.
gt (-3)	gt (-3)	1208	X:2324294..2325502	Schroeder et al.
gt (-1)	gt (-1)	1238	X:2323048..2324286	Schroeder et al.
h (-22)	VT27671	2139	3L:8645778..8647917	Kvon et al.
h (-11)	h stripe34 rev	911	3L:8657463..8658374	Schroeder et al.
h (-10)	h stripe7 rev	931	3L:8658177..8659108	Schroeder et al.
h (-9)	h 6 ru	867	3L:8659676..8660543	Schroeder et al.
h (-5)	h 15 ru	2677	3L:8662700..8665377	Schroeder et al.
h (+12)	h stripe0	468	3L:8680591..8681059	Ochoa-Espinosa et al.
hb (-3)	hb centr & post	1022	3R:4526520..4527542	Schroeder et al.
hb (-2)	HC 01	859	3R:4524620..4525479	Chen et al.
hb (+3)	hb anterior actv	720	3R:4520323..4521043	Schroeder et al.
hbn (-1)	HC 14	1008	2R:16849286..16850294	Chen et al.
hkb (-2)	hkb ventral elem	589	3R:173891..174480	Schroeder et al.
hth (+16)	Cluster-8277	1645	3R:6448530..6450175	Fisher et al.
hth (+69)	VT39530	2256	3R:6395394..6397650	Kvon et al.
hth (+100)	Cluster-8531	1503	3R:6363647..6365150	Fisher et al.
kni (-5)	kni (-5)	1402	3L:20692603..20694005	Schroeder et al.
kni (-1)	kni 83 ru	1654	3L:20689008..20690662	Schroeder et al.
kni (+1)	kni (+1)	1478	3L:20687055..20688533	Schroeder et al.
knrl (+8)	knrl (+8)	1297	3L:20604991..20606288	Schroeder et al.
Kr (-3)	Kr CD1 ru	1413	2R:21110136..21111549	Schroeder et al.
Kr (-2)	Kr CD2 ru	1811	2R:21111530..21113341	Schroeder et al.
Kr (-1)	Kr AD2 ru	1195	2R:21113325..21114520	Schroeder et al.
Kr (+10)	Cluster-8297	1045	2R:21124126..21125171	Fisher et al.
noc (-1)	HC 34	489	2L:14489159..14489648	Chen et al.
nub (-2)	nub (-2)	1984	2L:12615792..12617776	Schroeder et al.
oc (-4)	oc otd early	1837	X:8547931..8549768	Schroeder et al.
oc (+7)	oc (+7)	1832	X:8537082..8538914	Schroeder et al.
odd (-5)	odd (-5)	1383	2L:3610420..3611803	Schroeder et al.
odd (-3)	odd (-3)	1649	2L:3608812..3610461	Schroeder et al.
os (+5)	HC 06	974	X:18198042..18199016	Chen et al.
pdm2 (+1)	pdm2 (+1)	1622	2L:12678898..12680520	Schroeder et al.

Name	Alias	Length [bp]	Coordinates R5	Source
prd (-6)	Cluster-8520	1077	2L:12091700..12092777	Fisher et al.
prd (-3)	HC 03	1101	2L:12088746..12089847	Chen et al.
prd (+4)	prd +4	1311	2L:12080376..12081687	Schroeder et al.
rib (+3)	HC 11	663	2R:15160750..15161413	Chen et al.
run (-42)	run (-42)	1222	X:20522461..20523683	Schroeder et al.
run (-41)	run (-41)	1282	X:20523501..20524783	Schroeder et al.
run (-31)	run (-31)	2523	X:20533075..20535598	Schroeder et al.
run (-17)	run -17	996	X:20548261..20549257	Schroeder et al.
run (-10)	run stripe1	1616	X:20551039..20552655	Schroeder et al.
run (-8)	run stripe5	1335	X:20552655..20553990	Schroeder et al.
run (-9)	run -9	861	X:20555735..20556596	Schroeder et al.
run (+1)	HC 36	1298	X:20561726..20563024	Chen et al.
run (+19)	run (+19)	2260	X:20583284..20585544	Schroeder et al.
run (+30)	run (+30)	2708	X:20594595..20597303	Schroeder et al.
slp1 (-1)	slpA	370	2L:3824597..3824967	Ochoa-Espinosa et al.
slp1 (+3)	slpB	791	2L:3828818..3829609	Ochoa-Espinosa et al.
slp2 (-3)	slp2 (-3)	2639	2L:3832698..3835337	Schroeder et al.
SoxN (+7)	ChIP-50	1500	2L:8831698..8833198	Fisher et al.
tll (-3)	tll K2	2459	3R:26673280..26675739	Schroeder et al.
tll (-1)	tll P2	2759	3R:26675739..26678498	Schroeder et al.
tll (-2)	HC 07	1036	3R:26675091..26676127	Chen et al.
tll (+4)	tll head	985	3R:26681312..26682297	Ochoa-Espinosa et al.
tollo (-12)	HC 23	908	3L:15216686..15217594	Chen et al.
tsh (+48)	ChIP-2	1500	2L:21876210..21877710	Fisher et al.
fd19B (-1)	CG9571 head	755	X:19986472..19987227	Ochoa-Espinosa et al.
CG7271 (-1)	ChIP AHD-10	1500	3L:18581141..18582641	Fisher et al.
mir7 (-8)	mir7 head	361	2R:16485383..16485744	Ochoa-Espinosa et al.
mir281 (-2)	ChIP-miRNA5	1500	2R:8059538..8061038	Fisher et al.
iab-4 (-28)	VT42796	2111	3R:12646484..12648595	Kvon et al.
iab-4 (+41)	VT42831	2134	3R:12716036..12718170	Kvon et al.
iab-4 (+48)	VT42832	2128	3R:12717726..12719854	Kvon et al.
bxd (+71)	VT42747	2091	3R:12526519..12528610	Kvon et al.
bxd (+73)	VT42746	2155	3R:12524782..12526937	Kvon et al.
anterior split 1	KrRank1	1500	2R:9448948..9450448	Fisher et al.
stripe 80 1	HC 12	1292	2R:7914182..7915474	Chen et al.
anterior split 2	HC 35	1053	3R:20997138..20998191	Chen et al.
stripe 80 2	HC 46	1148	3R:22081560..22082708	Chen et al.
anterior cap	HC 52	1188	2R:19233370..19234558	Chen et al.
stripe 80 3	HC 57	1006	2L:15973654..15974660	Chen et al.
stripe 80 4	HC 58	1202	X:7159751..7160953	Chen et al.

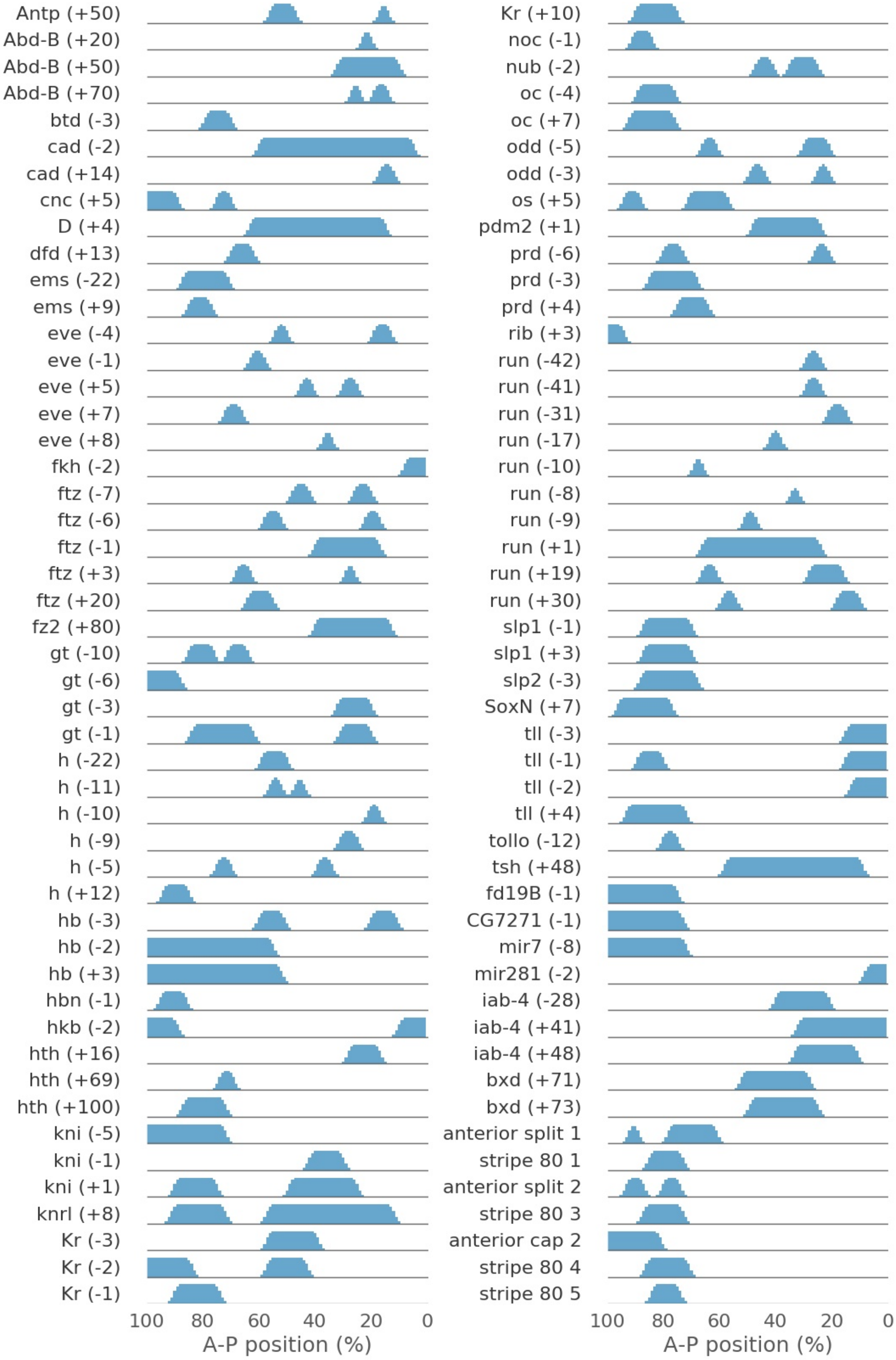


Figure A.5: Expression domains of all 98 enhancers.

Results – Model Training

We use the Wilcoxon Signed-rank test to measure significance in the case of paired results (experiments based on the same ten CV parsings, table A.2 onwards). Levels of significance for important comparisons are marked by: * $p \leq 0.05$; ** $p \leq 0.01$.

Table B.3: Test scores (correlation) for 5 and 10-fold CV as well as 10-fold CV with clustering based parsing (c10) each for a small set of parameters (8 TF) and a large parameter space (17 TF).

	8 TF			17 TF		
CV	5	10	c10	5	10	c10
1	0.367	0.347	0.348	0.386	0.367	0.375
2	0.343	0.339	0.35	0.337	0.367	0.373
3	0.338	0.326	0.361	0.359	0.355	0.345
4	0.309	0.33	0.354	0.354	0.366	0.405
5	0.324	0.353	0.328	0.39	0.382	0.399
6	0.343	0.352	0.347	0.361	0.411	0.389
7	0.354	0.359	0.377	0.36	0.378	0.335
8	0.352	0.366	0.365	0.363	0.343	0.403
9	0.349	0.348	0.334	0.372	0.348	0.379
10	0.317	0.385	0.364	0.32	0.368	0.374
Avg	0.34	0.351	0.353	0.36	0.369	0.378
Std	0.018	0.017	0.015	0.021	0.019	0.023

Table B.4: Comparison of local and global parameter optimization algorithms with CMA-ES repeated. All results are under three-fold augmentation and Hyperparameter training. Repeat: same as CMA-ES with 17 TFs. Multitrain: two training iterations of CMA-ES.

	8 TF			17 TF				
CV	Gradient	Simplex	CMA-ES	Gradient	Simplex	CMA-ES	Repeat	Multitrain
1	0.196	0.373	0.373	0.321	0.402	0.413	0.403	0.402
2	0.237	0.372	0.357	0.309	0.386	0.399	0.394	0.396
3	0.264	0.366	0.373	0.341	0.39	0.427	0.409	0.412
4	0.222	0.365	0.378	0.33	0.398	0.417	0.42	0.425
5	0.302	0.358	0.37	0.319	0.392	0.411	0.415	0.418
6	0.294	0.357	0.372	0.362	0.397	0.413	0.424	0.413
7	0.278	0.376	0.376	0.306	0.395	0.41	0.404	0.415
8	0.224	0.361	0.37	0.359	0.369	0.389	0.423	0.413
9	0.252	0.344	0.354	0.324	0.388	0.401	0.399	0.4
10	0.3	0.360	0.36	0.32	0.393	0.417	0.414	0.420
Avg	0.257	0.363	0.368	0.329	0.391	0.41	0.411	0.411
Std	0.037	0.009	0.008	0.019	0.009	0.011	0.01	0.009

**

Table B.5: Two different types of augmentation. Three-fold augmentation by 5%, 15% and 25% alteration rate as well as 10% and 25% cropping. All experiments with 17 TF and without Hyperparameter-training.

CV	No Aug	Alt5	Alt15	Alt25	Crop10	Crop25
1	0.375	0.398	0.367	0.387	0.393	0.394
2	0.373	0.341	0.364	0.38	0.341	0.36
3	0.345	0.355	0.362	0.374	0.383	0.39
4	0.405	0.378	0.357	0.367	0.374	0.386
5	0.399	0.375	0.379	0.381	0.352	0.37
6	0.389	0.375	0.443	0.376	0.427	0.386
7	0.335	0.382	0.398	0.373	0.388	0.387
8	0.403	0.367	0.353	0.354	0.395	0.391
9	0.379	0.386	0.354	0.324	0.39	0.369
10	0.374	0.387	0.368	0.388	0.381	0.369
Avg	0.378	0.374	0.375	0.37	0.382	0.38
Std	0.023	0.017	0.028	0.019	0.024	0.012

Table B.6: Test scores for different augmentation strategies with two, three and four fold augmentation. Regular augmentation is with closely related species first, alternative augmentation considers weakly related species first. All experiments with 17 TF and without Hyperparameter-training.

CV	No Aug	Aug2	Aug2 alt.	Aug3	Aug3 alt.	Aug4
1	0.375	0.389	0.388	0.4	0.396	0.401
2	0.373	0.381	0.382	0.412	0.406	0.392
3	0.345	0.393	0.382	0.376	0.409	0.384
4	0.405	0.379	0.392	0.439	0.390	0.388
5	0.399	0.409	0.387	0.398	0.423	0.42
6	0.389	0.39	0.41	0.405	0.412	0.411
7	0.335	0.383	0.395	0.39	0.406	0.393
8	0.403	0.424	0.356	0.39	0.410	0.383
9	0.379	0.374	0.372	0.397	0.407	0.399
10	0.374	0.397	0.394	0.403	0.383	0.384
Avg	0.378	0.392	0.386	0.401	0.404	0.396
Std	0.023	0.015	0.014	0.017	0.012	0.012
			**			

Table B.7: Test scores for no penalty (NP: $\lambda = 0$), best-guess penalty (HT0: $\lambda = 10^{-3}$) and the hyperparameter training strategies HTI and HTII ($\lambda \in \{10^{-3}, 5 \cdot 10^{-3}, 10^{-2}\}$), each with and without data augmentation. HTII+ with additionally $\lambda \in \{5 \cdot 10^{-4}, 10^{-1}\}$

	No Aug				Aug3				
CV	NP	HT 0	HT I	HT II	NP	HT 0	HT I	HT II	HTII+
1	0.412	0.375	0.439	0.388	0.36	0.4	0.47	0.403	0.401
2	0.365	0.373	0.373	0.382	0.392	0.412	0.402	0.394	0.385
3	0.337	0.345	0.415	0.387	0.415	0.376	0.385	0.409	0.405
4	0.391	0.405	0.366	0.394	0.389	0.439	0.441	0.42	0.418
5	0.336	0.399	0.39	0.387	0.395	0.398	0.389	0.415	0.411
6	0.343	0.389	0.371	0.384	0.402	0.405	0.369	0.424	0.415
7	0.367	0.335	0.409	0.343	0.378	0.39	0.413	0.404	0.406
8	0.395	0.403	0.384	0.387	0.389	0.39	0.419	0.423	0.386
9	0.354	0.379	0.39	0.396	0.396	0.397	0.451	0.399	0.398
10	0.324	0.374	0.349	0.386	0.376	0.403	0.427	0.414	0.42
Avg	0.362	0.378	0.389	0.383	0.389	0.401	0.417	0.411	0.405
Std	0.029	0.023	0.026	0.015	0.015	0.017	0.032	0.01	0.012
		*					*		

Model Evaluation

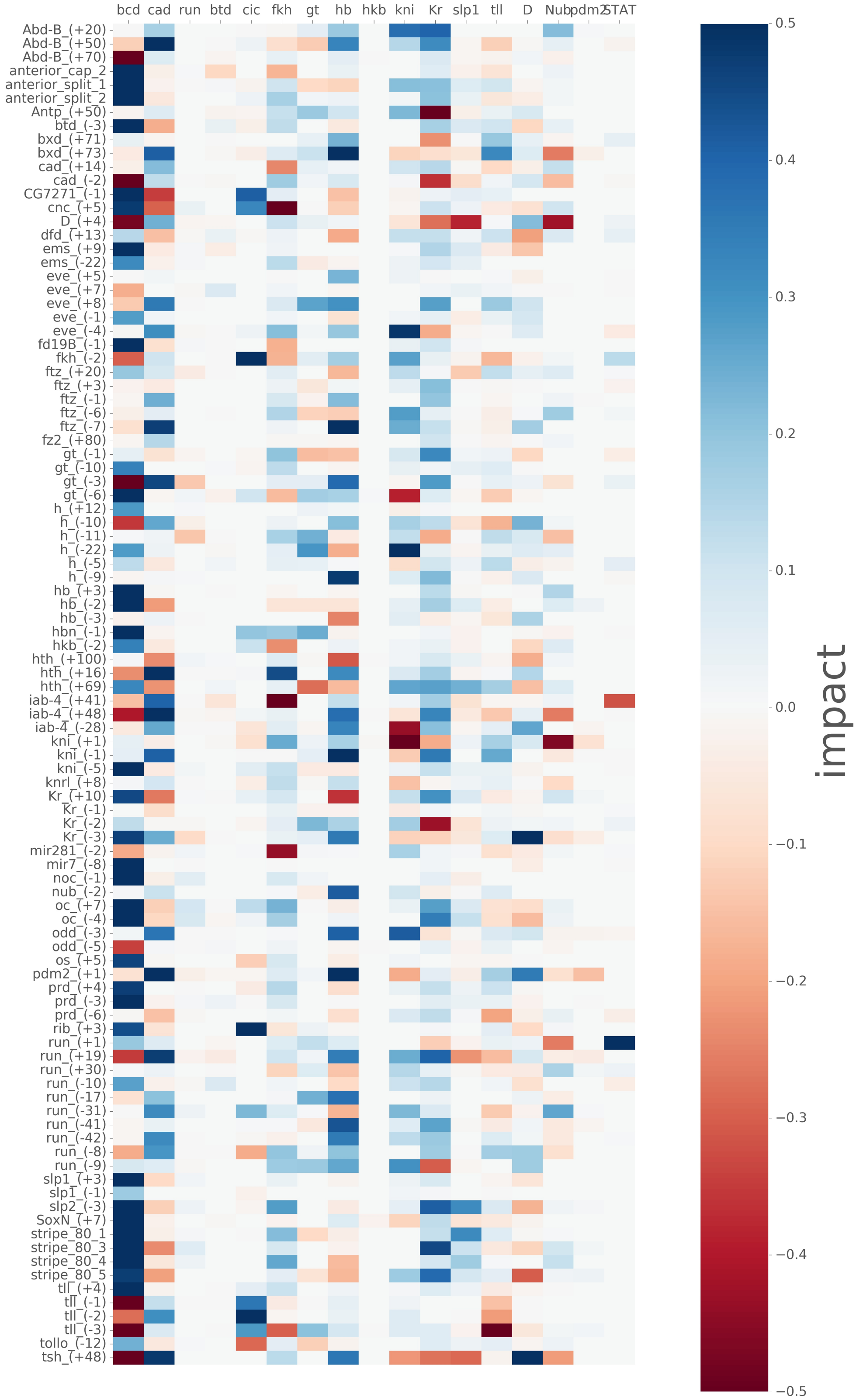


Figure C.6: Impact of every TF on single enhancers. The impact is measured always on the test data. Blue indicates positive impact (deleting the TF deteriorated the prediction), while red indicates that the TF interferes with the correct prediction.

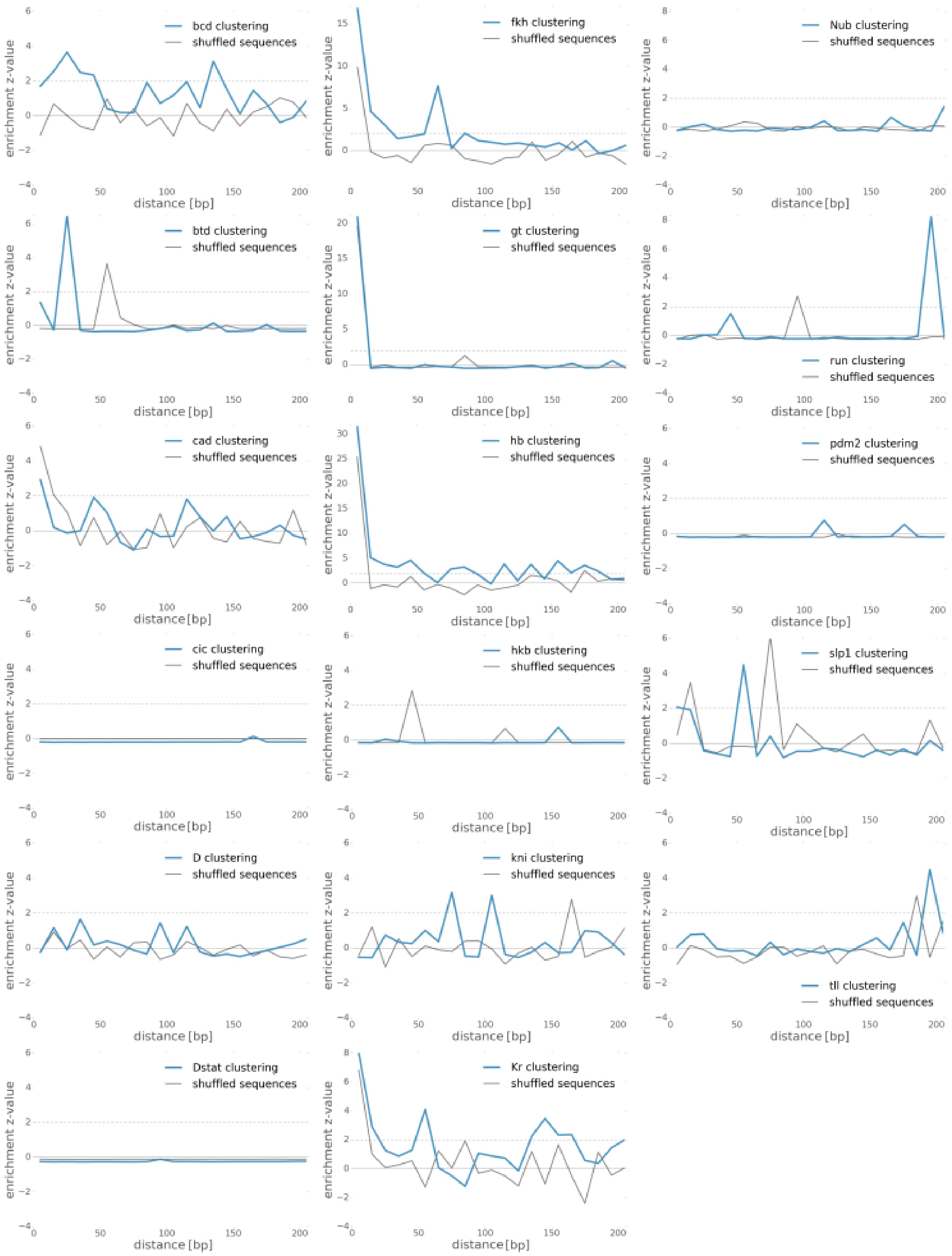


Figure C.7: Enrichment of homotypic clustering for all TFs in 10bp segments compared to shuffled binding sites (10000 iterations). In gray: negative control with 98 pseudo-enhancer (shuffled enhancer sequences).

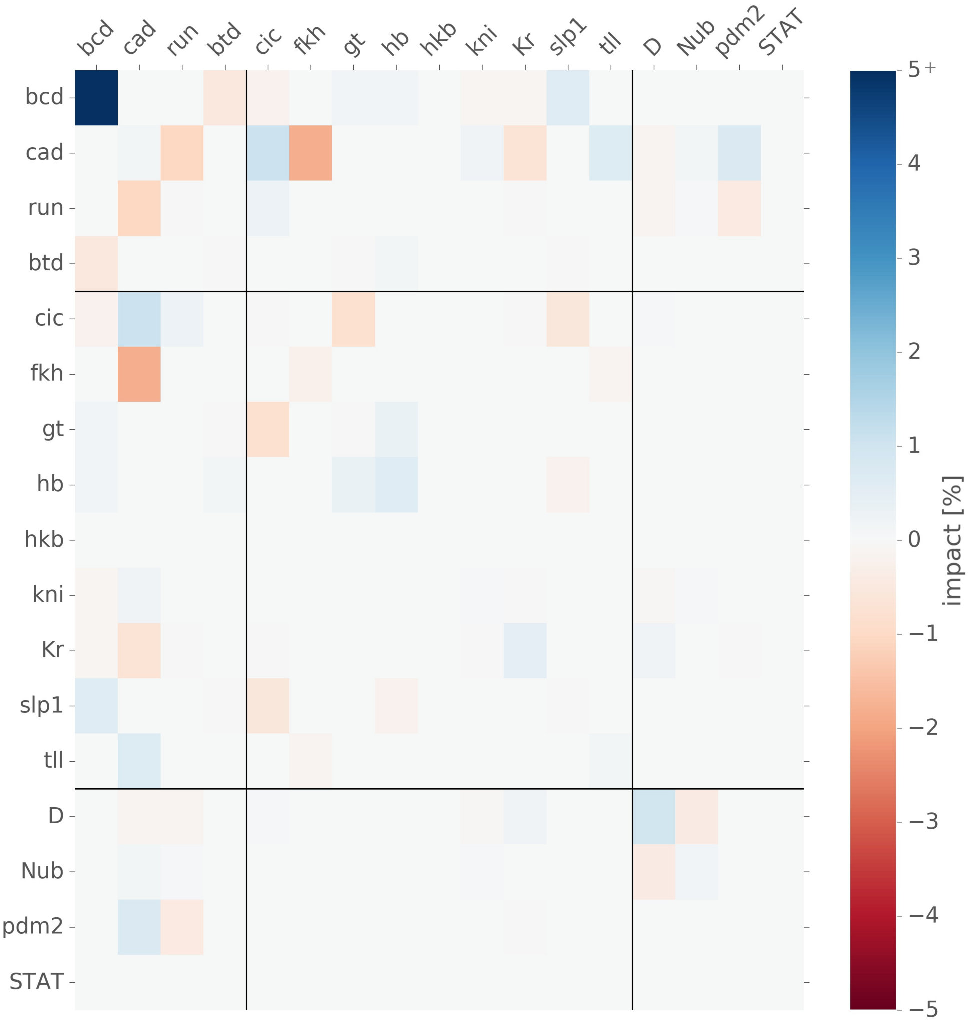


Figure C.8: Impact of heterotypic cooperativity. Notice the symmetric nature of the heatmap due to the symmetry of cooperativity. The scale is cropped at 5% impact although bcd homotypic interaction has more impact. Otherwise most other impact values were not visible.

Bibliography

[1] Pichaud F and Desplan C. Pax genes and eye organogenesis. *Current Opinion in Genetics & Development*, 12(4):430 – 434, 2002.

[2] Wray GA. The evolutionary significance of cis-regulatory mutations. *Nature Reviews Genetics*, 8:206–216, 2007.

[3] Jeong S, Rebeiz M, Andolfatto P, Werner T, True J, and Carroll SB. The evolution of gene regulation underlies a morphological difference between two Drosophila sister species. *Cell*, 132(5):783–793, 2008.

[4] Ciglar L and Furlong EEM. Conservation and divergence in developmental networks: a view from Drosophila myogenesis. *Current Opinion in Cell Biology*, 21(6):754–760, 2009.

[5] Levine M and Tjian R. Transcription regulation and animal diversity. *nature*, 424:147 – 151, 2003.

[6] Yruela I, Oldfield CJ, Niklas KJ, and Dunker AK. Evidence for a Strong Correlation Between Transcription Factor Protein Disorder and Organismic Complexity. *Genome Bio. Evol.*, 9(5):1248–1265, 2017.

[7] Simpson-Brose M, Treisman J, and Desplan C. Synergy between the Hunchback and Bicoid Morphogens is required for Anterior Patterning in Drosophila. *Cell*, 78(5):855–865, 1994.

[8] Staller MV, Vincent BJ, Bragdon MDJ, Lydiard-Martin T, Wunderlich Z, Estrada J, and DePace AH. Shadow enhancers enable Hunchback bifunctionality in the Drosophila embryo. *PNAS*, 112(3):785–790, 2015.

[9] Arnosti DN and Kulkarni MM. Transcriptional enhancers: Intelligent enhanceosomes or flexible billboards? *J. Cell Biochem.*, 94(5):890–898, 2005.

[10] Wasson T and Hartemink AJ. An ensemble model of competitive multi-factor binding of the genome. *Genome Res.*, 19(11):2101–2112, 2009.

[11] Mirnay LA. Nucleosome-mediated cooperativity between transcription factors. *PNAS*, 107(52):22534–22539, 2010.

[12] Ma X, Yuan D, Diepold K, Scarborough T, and Ma J. The Drosophila morphogenetic protein Bicoid binds DNA cooperatively. *Development*, 122:1195–1206, 1996.

[13] Diebold RJ, Rajaram N, Leonard DA, and Kerppola TK. Molecular basis of cooperative DNA bending and oriented heterodimer binding in the NFAT1—Fos–Jun—ARRE2 complex. *PNAS*, 95(14):7915–7920, 1996.

[14] Kim S1, Broströmer E, Xing D, Jin J, Chong S, Ge H, Wang S, Gu C, Yang L, Gao YQ, Su XD, Sun Y, and Xie XS. Probing Allostery Through DNA. *Science*, 339(6121):816–819, 2013.

[15] Fakhouri WD, Ay A, Sayal R, Dresch J, Dayringer E, and Arnosti DN. Deciphering a transcriptional regulatory code: modeling short-range repression in the Drosophila embryo. *Mol. Systems Biol.*, 6(1), 2010.

[16] Schroeder MD, Pearce M, Fak J, Fan H, Unnerstall U, Emberly E, Rajewsky N, Siggia ED, and Gaul U. Transcriptional Control in the Segmentation Gene Network of Drosophila. *PLoS biology*, 2(9):1396–1410, 2004.

[17] Fisher WW, Li JJ, Hammonds AS, Brown JB, Pfeiffer BD, Weiszmann R, MacArthur S, Thomas S, Stamatoyannopoulos JA, Eisen MB, Bickel PJ, Biggin MD, and Celniker SE. DNA regions bound at low occupancy by transcription factors do not drive patterned reporter gene expression in Drosophila. *PNAS*, 109(52):21330–21335, 2012.

[18] Stampfel G, Kazmar T, Frank O, Wienerroither S, Reiter F, and Stark A. Transcriptional regulators form diverse groups with context-dependent regulatory functions. *Nature*, 528:147–151, 2015.

[19] Sinha S, Pearce M, Unnerstall U, Fak J, Dandapani M, Schroeder MD, Siggia ED, and Gaul U. Evolution of transcription control in the segmentation gene network of *Drosophila. In Preperation.*

[20] Hare EE, Peterson BK, Iyer VN, Meier R, and Eisen MB. Sepsid even-skipped enhancers are functionally conserved in Drosophila despite lack of sequence conservation. *PLoS Genetics*, 4(6), 2008.

[21] Scott MP and Carroll SB. The segmentation and homeotic gene network in early Drosophila development. *Cell*, 51(5):689–698, 1987.

[22] Schroeder MD, Greer C, and Ulrike U. How to make stripes: deciphering the transition from non-periodic to periodic patterns in Drosophila segmentation. *Development*, 138(14):3067–3078, 2011.

[23] Gregor T, Bialek W, de Ruyter van Steveninck RR, Tank DW, and Wieschaus EF. Diffusion and scaling during early embryonic pattern formation. *PNAS*, 102(51):18403–18407, 2005.

[24] Becker K, Balsa-Canto E, Cicin-Sain, D Hoermann A, Janssens H, Banga JR, and Jaeger J. Reverse-Engineering Post-Transcriptional Regulation of Gap Genes in Drosophila melanogaster. *PLoS Comp. Bio.*, 9(10), 2013.

[25] Little SC, Tikhonov M, and Gregor T. Precise developmental gene expression arises from globally stochastic transcriptional activity. *Cell*, 154(4):789–800, 2013.

[26] Zinzen RP, Senger K, and Papatsenko D Levine M and. Computational models for neurogenic gene expression in the Drosophila embryo. *Curr. Biol.*, 16(13):1358–1365, 2006.

[27] Segal E, Raveh-Sadka T, Schroeder M, Unnerstall U, and Gaul U. Predicting expression patterns from regulatory sequence in Drosophila segmentation. *Nature*, 451:535–540, 2008.

[28] He X, Samee MAH, Blatti C, and Sinha S. Thermodynamics-Based Models of Transcriptional Regulation by Enhancers: The Roles of Synergistic Activation, Cooperative Binding and Short-Range Repression. *PLoS Comput. Biol.*, 6(9), 2010.

[29] Kulkarni MM and Arnosti DN. cis-Regulatory Logic of Short-Range Transcriptional Repression in Drosophila melanogaster. *Mol. Cell Biol.*, 25(9):3411–3420, 2005.

[30] Suleimenova Y, Ay A, Samee MAH, Dreschd JM, Sinha S, and Arnosti DN. Global parameter estimation for thermodynamic models of transcriptional regulation. *Methods*, 62(1):99–108, 2013.

[31] Dresch JM, Liu X, Arnosti DN, and Ay A. Thermodynamic modeling of transcription: sensitivity analysis differentiates biological mechanism from mathematical model-induced effects. *BMC System Bio.*, 4(1):142–152, 2010.

[32] Bieler J, Pozzorini C, and Naef F. Whole-Embryo Modeling of Early Segmentation in Drosophila Identifies Robust and Fragile Expression Domains. *Biophys. J.*, 101(2):287–296, 2011.

[33] Sayal R, Dresch JM, Pushel I, Taylor BR, and Arnosti DN. Quantitative perturbation-based analysis of gene expression predicts enhancer activity in early Drosophila embryo. *eLife*, 5, 2015.

[34] Rohs R, Jin X, West SM, Joshi R, Honig B, and Mann RS. Origins of Specificity in Protein-DNA Recognition. *Annu. Rev. Biochem.*, 79:233–269, 2010.

[35] Peng P, Samee MAH, and Sinha S. Incorporating Chromatin Accessibility Data into Sequence-to-Expression Modeling. *Biophys. J.*, 108(5):1257–1267, 2015.

[36] Peng PC and Sinha S. Quantitative modeling of gene expression using dna shape features of binding sites. *Nucleic Acids Res.*, 44(13), 2016.

[37] Stormo GD, Schneider TD, Gold L, and Ehrenfeucht A. Use of the 'Perceptron' algorithm to distinguish translational initiation sites in E. coli. *Nucleic Acids Res.*, 10(9):2997–3011, 1982.

[38] Rajewsky N, Vergassola M, Gaul U, and Siggia ED. Computational detection of genomic cis-regulatory modules applied to body patterning in the early drosophila embryo. *BMC Bioinformatics*, 3(30), 2002.

[39] Cornish-Bowden A. Nomenclature for incompletely specified bases in nucleic acid sequences: recommendations 1984. *Nucleic Acids Res.*, 13(9):3021–3030, 1985.

[40] Stormo GD. DNA binding sites: representation and discovery. *Bioinformatics*, 16(1):16–23, 2000.

[41] Schneider TD and Stephens RM. Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res.*, 18(20):6097–6100, 1990.

[42] Crooks GE, Hon G, Chandonia JM, and Brenner SE. Weblogo: a sequence logo generator. *Genome Res.*, 14(6):1188–1190, 2004.

[43] Djordjevic M, Sengupta AM, and Shraiman BI. A biophysical approach to transcription factor binding site discovery. *Genome Res.*, 13(11):2381–2390, 2003.

[44] Shea MA and Ackers GK. The OR control system of bacteriophage lambda. A physical-chemical model for gene regulation. *J. Mol. Biol.*, 181(2):211–230, 1985.

[45] Buchler NE, Gerland U, and Hwa T. Predicting expression patterns from regulatory sequence in Drosophila segmentation. *PNAS*, 100(9):5136–5141, 2003.

[46] Hager GL, McNally JG, and Misteli T. Transcription dynamics. *Molecular Cell*, 35(6):741–753, 2009.

[47] Sprague BL, Pego RL, Stavreva DA, and McNally JG. Analysis of binding reactions by fluorescence recovery after photobleaching. *Biophys. J.*, 86(6):3473–3495, 2004.

[48] Lickwar CR, Mueller F, Hanlon SE, McNally JG, and Lieb JD. Genome-wide protein–DNA binding dynamics suggest a molecular clutch for transcription factor function. *Nature*, 484:251–255, 2012.

[49] Nitta KR, Jolma A, Yin Y, Morgunova K, Kivioja T, Akhtar J, Hens K, Toivonen J, Deplancke B, Furlong EEM, and Taipale J. Conservation of transcription factor binding specificities across 600 million years of bilateria evolution. *eLife*, 4, 2015.

[50] Noyes MB, Christensen RG, Wakabayashi A, Stormo GD, Brodsky MH, and Wolfe SA. Analysis of homeodomain specificities allows the family-wide prediction of preferred recognition sites. *Cell*, 133(7):1277–1289, 2008.

[51] Dörner K. Computational dissection of enhancer architecture in the Drosophila segmentation gene network. B.sc. thesis, LMU München, 2015.

[52] Benos PV, Bulyk ML, and Stormo GD. Additivity in protein-DNA interactions: how good an approximation is it? *Nucleic Acids Res.*, 30(20):4442–4451, 2002.

[53] O'Flanagan RA, Paillard G, Lavery R, and Sengupta AM. Non-additivity in protein–DNA binding. *Bioinformatics*, 21(10):2254–2263, 2005.

[54] Weirauch MT et. al. Evaluation of methods for modeling transcription factor sequence specificity. *Nature Biotechnology*, 31(2):126–134, 2013.

[55] Mirny LA. Nucleosome-mediated cooperativity between transcription factors. *PNAS*, 107(52):22534–22539, 2010.

[56] Kim S, Broströmer E, Xing D, Jin J, Chong S Ge H, Wang S, Gu C, Yang L, Gao YQ, Su X, Sun Y, and Xie XS. Probing allostery through dna. *Science*, 339(6121):816–819, 2013.

[57] Shirokawa JM and Courey AJ. A direct contact between the dorsal rel homology domain and twist may mediate transcriptional synergy. *Mol. Cell. Bio.*, 17(6):3345–3355, 1997.

[58] von Reutern M. Thermodynamic models. `https://github.com/Reutern/thermodynamic_models`, 2017.

[59] Morrison AH, Scheeler M, Dubuis JO, and Gregor T. Quantifying the Bicoid morphogen gradient in living fly embryos. *Cold Spring Harb Protoc.*, 4:398–406, 2012.

[60] Sinha S, van Nimwegen E, and Siggia ED. A probabilistic method to detect regulatory modules. *Bioinformatics*, 19:292–301, 2003.

[61] Kvon EZ, Kazmar T, Stampfel G, Yanez-Cuna JO, Pagani M, Schernhuber K, Dickson BJ, and Stark A. Genome-scale functional characterization of Drosophila developmental enhancers in vivo. *Nature*, 512(7512):91–95, 2014.

[62] Ochoa-Espinosa A, Yucel G, Kaplan L, Pare A, Pura N, Oberstein A, Papatsenko D, and Small S. The role of binding site cluster strength in Bicoid-dependent patterning in Drosophila. *PNAS*, 102(14):4960–4695, 2005.

[63] Chen H, Xu Z, Mei C, Yu D, and Small S. A System of Repressor Gradients Spatially Organizes the Boundaries of "Morphogen-dependent" Target Genes. *Cell*, 149(3):618–629, 2012.

[64] Vincent A, Blankenship JT, and Wieschaus E. Integration of the head and trunk segmentation systems controls cephalic furrow formation in Drosophila. *Development*, 124(19):3747–3754, 1997.

[65] Casanova J. Pattern formation under the control of the terminal system in the Drosophila embryo. *Development*, 110(2):621–628, 1990.

[66] Chen H, Xu Z, Mei C, Yu D, and Small S. A system of repressor gradients spatially organizes the boundaries of Bicoid-dependent target genes. *Cell*, 149(3):618–629, 2012.

[67] Tsai C-C, Kramer SG, and Gergen JP. Pair-rule gene runt restricts orthodenticle expression to the presumptive head of the Drosophila embryo. *Dev. Genet.*, 23(1):35–44, 1998.

[68] Russell SR, Sanchez-Soriano N, Wright CR, and Ashburner M. The Dichaete gene of Drosophila melanogaster encodes a SOX-domain protein required for embryonic segmentation. *Development*, 122(11):3669–3676, 1996.

[69] Nambu PA and Nambu JR. The Drosophila fish-hook gene encodes a HMG domain protein essential for segmentation and CNS development. *Development*, 122(11):3467–3475, 1996.

[70] Cockerill KA, Billin AN, and Poole SJ. Regulation of expression domains and effects of ectopic expression reveal gap gene-like properties of the linked pdm genes of Drosophila. *Mech. of Dev.*, 41(2–3):139–153, 1993.

[71] Yan R, Small S, Desplan C, Dearolf CR, and Darnell JR. Identification of a Stat gene that functions in Drosophila development. *Cell*, 84(3):421–430, 1996.

[72] Tsurumi A, Xia F, Li J, Larson K, LaFrance R, and Li WX. STAT Is an Essential Activator of the Zygotic Genome in the Early Drosophila Embryo. *PLOS Genetics*, 7(5), 2011.

[73] Surkova S, Kosman S, Kozlov K, Manu, Myasnikova E, Samsonova A, Spirov A, Vanario-Alonso CE, Samsonova M, and Reinitz J. Characterization of the Drosophila segment determination morphome. *Developmental Biol.*, 313(2):844–862, 2008.

[74] Hammonds AS, Bristow CA, Fisher WW, and Weiszmann R et al. Spatial expression of transcription factors in Drosophila embryonic organ development. *Genome Biol.*, 14(12), 2013.

[75] Meng X, Brodsky MH, and Wolfe SA. A bacterial one-hybrid system for determining the DNA-binding specificity of transcription factors. *Nat. Biotech.*, 23(8):988–994, 2005.

[76] Galas DJ and Schmitz A. DNAse footprinting: a simple method for the detection of protein-DNA binding specificity. *Nucleic Acids Res.*, 5(9):3157–3170, 1978.

[77] Oliphant AR, Brandl CJ, and Struhl K. Defining the sequence specificity of DNA-binding proteins by selecting binding sites from random-sequence oligonucleotides: analysis of yeast GCN4 protein. *Mol. Cell Biol.*, 9(7):2944–2949, 1989.

[78] Jung C, Bandilla P, von Reutern M, Lange S, Unnerstall U, and Gaul U. High sensitivity measurement of transcription factor-DNA binding energies by automated fluorescence microscopy. *In preparation*, 2017.

[79] Durbin R, Eddy S, Krogh A, and Mitchison G. *Biological sequence analysis: probabilistic models of proteins and nucleic acids.* Cambridge University Press, Cambridge, UK, 1998.

[80] Bishop CM. *Pattern Recognition and Machine Learning.* Springer, Cambridge, England, 2006.

[81] Hastie T and Tibshirani R and Friedman J. *The Elements of Statistical Learning.* Springer Series in Statistics. Springer, Stanford, California, 2008.

[82] Pearson K. Note on Regression and Inheritance in the Case of Two Parents. *PRSL*, 58:240–242, 1895.

[83] Davis J and Goadrich M. The relationship between Precision-Recall and ROC curves. *Proc. ICML '06*, pages 233–240, 2006.

[84] Tibshirani R. Regression Shrinkage and Selection via the Lasso. *JRSS*, 58(1):267–288, 1996.

[85] Hoerl AE and Kennard RW. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

[86] Ingalls B. Sensitivity analysis: from model parameters to system behaviour. *Essays Biochem.*, 45:177–193, 2008.

[87] A Saltelli and Ratto M and Andres T and Campolongo F and Cariboni J and Gatelli D and Saisana M and Tarantola S. *Global Sensitivity Analysis: The Primer.* John Wiley & Sons, Chichester, England, 2008.

[88] Nelder JA and Mead R. A simplex method for function minimization. *Comp. J.*, 7(4):308–313, 1965.

[89] Fletcher R and Reeves CM. Function minimization by conjugate gradients. *Comp. J.*, 7:149–154, 1964.

[90] Fletcher R. *Practical Methods of Optimization, 2nd Edition.* John Wiley & Sons, New York NY, USA, 1987.

[91] Gough B. *GNU Scientific Library Reference Manual - Third Edition.* Network Theory Ltd., 2009.

[92] Hansen N and Ostermeier A. Completely Derandomized Self-Adaptation in Evolution Strategies. *Evolutionary Computation*, 9(2):159–195, 2001.

[93] Simard PY, Steinkraus D, and Platt JC. Best Practices for Convolutional Neural Networks Applied to Visual Document Analysis. In *International Conference on Document Analysis and Recognition*, page 958–962, 2003.

[94] Krizhevsky A, Sutskever I, and Hinton GE. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, 2012.

[95] Sáiz-Abajoa MJ, Mevikb BH, Segtnanb VH, and Næs T. Ensemble methods and data augmentation by noise addition applied to the analysis of spectroscopic data. *Analytica Chimica Acta*, 533(2):147–159, 2005.

[96] Kumar A and Cowen L. Augmented training of hidden Markov models to recognize remote homologs via simulated evolution. *Bioinformatics*, 25(13):1602–1608, 2009.

[97] Kazemian M, Suryamohan K, Chen J, Zhang Y, Samee MAH, Halfon MS, and Sinha S. Evidence for Deep Regulatory Similarities in Early Developmental Programs across Highly Diverged Insects. *Genome Bio. Evol.*, 6(9):2301–2320, 2014.

[98] Rosenbloom KR, Armstrong J, Barber GP, Casper J, Clawson H, and Diekhans M et al. The UCSC Genome Browser database: 2015 update. *Nucleic Acids Res.*, 43:D670–D681, 2015.

[99] Drosophila 12 Genomes Consortium. Evolution of genes and genomes on the Drosophila phylogeny. *Nature*, 450(7167):203–218, 2015.

[100] Bonneton F, Shaw PJ, Fazakerley C, Shi M, and Dover GA. Comparison of bicoid-dependent regulation of hunchback between Musca domestica and Drosophila melanogaster. *Mech. of Dev.*, 66:143–156, 1997.

[101] Lynch JA, Olesnicky EC, and Desplan C. Regulation and function of tailless in the long germ wasp Nasonia vitripennis. *Dev. Genes. Evol.*, 216:493–498, 2006.

[102] Nitta KR, Jolma A, Yin Y, Morgunova E, Kivioja T, Akhtar J, Hens K, Toivonen J, Deplancke B, Furlong EEM, and Taipale J. Conservation of transcription factor binding specificities across 600 million years of bilateria evolution. *eLife*, 4:e04837, 2015.

[103] Ludwig MZ, Patel NH, and Kreitman M. Functional analysis of eve stripe 2 enhancer evolution in Drosophila: rules governing conservation and change. *Development*, 125(5):949–958, 1998.

[104] Wittkopp PJ. Evolution of cis-regulatory sequence and function in Diptera. *Heredity*, 97:139–147, 2006.

[105] Alipanahi B, Delong A, Weirauch MT, and Frey BJ. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nature Biotechnology*, 33(8):831–839, 2015.

[106] Bothma JP, Garcia HC, Esposito E, Schlissel G, Gregor T, and Levine MS. Dynamic Regulation of Eve Stripe 2 Expression Reveals Transcriptional Bursts in Living Drosophila Embryos. *PNAS*, 111(29):10598–10603, 2014.

[107] Meir R. Bias, variance and the combination of estimators: The case of linear least squares. *Department of Electrical Engineering, Technion Israel Institute of Technology*, 1994.

[108] Breiman L. Bagging Predictors. *Machine Learning*, 24, 1996.

[109] Kazemian M, Pham H, Wolfe SA, Brodsky MH, and Sinha S. Widespread evidence of cooperative DNA binding by transcription factors in Drosophila development . *Nucleic Acids Res.*, 41(17):8237–8252, 2013.

[110] Lydiard-Martin T, Bragdon M, Eckenrode KB, Wunderlich Z, and DePace AH. Locus architecture affects mRNA expression levels in Drosophila embryos. *bioRxiv*, 2014.

[111] Kurokawa R, Yu VC, Näär A, Kyakumoto S, Han Z, Silverman S, Rosenfeld MG, and Glass CK. Differential orientations of the DNA-binding domain and carboxy-terminal dimerization interface regulate binding site selection by nuclear receptor heterodimers. *Genes & Dev.*, 7:1423–1435, 1993.

[112] Jolma A, Yan J, Whitington T, Toivonen J, Nitta KR, Rastas P, Morgunova E, Enge M, Taipale M, Wei G, Palin K, Vaquerizas JM, Vincentelli R, Luscombe NM, Hughes TR, Lemaire P, Ukkonen E, Kivioja T, and Taipale J. DNA-Binding Specificities of Human Transcription Factors. *Cell*, 152(1–2):327–339, 2013.

[113] Browning DF and Busby SJW. The regulation of bacterial transcription initiation. *Nature Rev. Microbio.*, 2:785–790, 2004.

[114] Erickson HP. Size and Shape of Protein Molecules at the Nanometer Level Determined by Sedimentation, Gel Filtration, and Electron Microscopy. *Bio. Proc. Online*, 11:32–51, 2009.

[115] Pique-Regi R, Degner JF, Pai AA, Gaffney DJ, Gilad Y, and Pritchard JK. Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. *Genome Res.*, 21(3):447–455, 2011.

[116] Li X-Y, Thomas S, Sabo PJ, Eisen MB, Stamatoyannopoulos JA, and Biggin MD. The role of chromatin accessibility in directing the widespread, overlapping patterns of Drosophila transcription factor binding. *Genome Biol.*, 12(4), 2011.

[117] Kaplan T, Li X-Y, Sabo PJ, Thomas S, Stamatoyannopoulos JA, Biggin MD, and Eisen MB. Quantitative Models of the Mechanisms That Control Genome-Wide Patterns of Transcription Factor Binding during Early Drosophila Development. *PLoS Genetics*, 7(2), 2011.

[118] Arvey A, Agius P, Noble WS, and Leslie C. Sequence and chromatin determinants of cell-type–specific transcription factor binding. *Genome Res.*, 22(9):1723–1734, 2012.

[119] Sabo PJ, Hawrylycz M, and Wallace JC et al. Discovery of functional noncoding elements by digital analysis of chromatin structure. *PNAS*, 101(48):16837–16842, 2004.

[120] Buenrostro JD, Giresi PG, Zaba LC, Chang HY, and Greenleaf WJ. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nature Methods*, 10(12):1213–1218, 2013.

[121] Winkler CJ, Ponce A, and Courey AJ. Groucho-Mediated Repression May Result from a Histone Deacetylase-Dependent Increase in Nucleosome Density. *PLoS One*, 5(4), 2010.

[122] Li LM and Arnosti DN. Long- and Short-Range Transcriptional Repressors Induce Distinct Chromatin States on Repressed Genes. *Curr. Biol.*, 21(5):406–412, 2011.

[123] Arnosti DN, Gray S, Barolo S, Zhou J, and Levine M. The gap protein knirps mediates both quenching and direct repression in the Drosophila embryo. *Curr. Biol.*, 15(14):3659–3666, 1996.

[124] Ahsendorf T, Wong F, Eils R, and Gunawardena J. A framework for modelling gene regulation which accommodates non-equilibrium mechanisms. *BMC Biology*, 12(1), 2014.

[125] Mackay TFC, Richards S, and Stone EA et al. The Drosophila melanogaster Genetic Reference Panel. *Nature*, 482:173–178, 2012.