Modelling the Transcriptional Regulation of Androgen Receptor in Prostate Cancer

by

Yuqian (Eugene) Hu

B.Sc., University of Waterloo, 2018

Thesis Submitted in Partial Fulfillment of the
Requirements for the Degree of
Master of Science

in the
Department of Physics
Faculty of Science

SIMON FRASER UNIVERSITY
Spring 2021

Declaration of Committee

Name:	**Yuqian (Eugene) Hu**
Degree:	**Master of Science**
Thesis title:	**Modelling the Transcriptional Regulation of Androgen Receptor in Prostate Cancer**
Committee:	**Chair:** Malcolm Kennett Associate Professor, Physics
	Eldon Emberly Supervisor Professor, Physics
	David Sivak Committee Member Associate Professor, Physics
	Nancy Forde Committee Member Professor, Physics

Abstract

Transcription of genes and production of proteins are essential functions of a normal cell. If disturbed, misregulation of crucial genes leads to aberrant cell behaviour and in some cases, leads to the development of diseased states such as cancer. One major transcriptional regulation tool involves the binding of transcription factor onto enhancer sequences that will encourage or repress transcription depending on the role of the transcription factor. In prostate cells, misregulation of the androgen receptor(AR), a key transcriptional regulator, leads to the development and maintenance of prostate cancer. Androgen receptor binds to numerous locations in the genome, but it is still unclear how and which other key transcription factors aid and repress AR-mediated transcription. Here I analyzed the data that contained the transcriptional activity of 4139 putative AR binding sites (ARBS) in the genome with and without the presence of hormone using the STARR-seq assay. Only a small fraction of ARBS showed significant differential expression when treated with hormone. To understand the underlying essential factors behind hormone-dependent behaviour, we developed both machine learning and biophysical models to identify active enhancers in prostate cancer cells. We also identify potentially crucial transcription factors for androgen-dependent behaviour and discuss the benefits and shortcomings of each modelling method.

Keywords: Transcription, Transcription regulation, Prostate Cancer, Machine learning, Mean-field, Ensemble Learning, PCA

Acknowledgements

There are many people who have helped me during my master's degree and who have helped contribute to this thesis. Firstly, I need to give my thanks to my supervisor, Dr. Eldon Emberly who has helped me shape my interests throughout the past two years. His knowledge and curiosity have inspired me to continue to constantly question my surroundings and learn about new and interesting methodologies. More important than anything, he has taught me to slow down and truly conceptualize what I am trying to do which has helped me tremendously throughout my research. Additionally, I would like to thank Dr. Nathan Lack and Tunc Morova for their help in the biology of prostate cancer and androgen regulation. Their work in prostate cancer cell lines and experimental data are the foundation to my mathematical modelling and are the reason why this thesis is possible. Next, I would also like to thank my family and my wonderful girlfriend who has supported me through this journey.

Table of Contents

List of Figures

Chapter 1

Introduction

1.1 Transcriptional Logic and Cancer

The diversity of human tissues and cell functions require accurate gene expression and protein production. Despite all cells sharing the same genome, the proteins required for cellular function will vary from tissue to tissue. To generate such diverse functions, the transcription of genes by RNA Polymerase II (Pol-II) is regulated by a wide range of different signalling proteins, DNA structural changes, epigenetic modifications, and the binding of transcription factors (TFs) [1, 2]. While all of these play a major role in transcriptional regulation, this thesis will be focused on deciphering how the binding of transcription factors changes the resulting transcription and gene expression.

Pol-II initiates transcription by binding to a DNA sequence upstream of a gene. The binding of Pol-II is regulated via its interactions with TFs that bind to DNA sequences close to the gene of interest or at a distal region that becomes spatially close due to formation of DNA loops (Fig. 1.1). The binding of TFs along with secondary co-factors forms a protein complex that can encourage or repress the binding of Pol-II depending on the role of the TFs. Therefore, an understanding of the binding locations of TFs and the function of each TF will give insight into the regulation of Pol-II binding to promoters. Overall, this presents a rich and complex transcriptional regulatory logic that contains tens if not hundreds of various confounding factors with specific functions. The human genome largely contains 3 types of regulatory elements that TFs can bind to: promoters, enhancers, and insulators (Fig. 1.2) [3].

Here we offer a brief introduction to each regulatory sequence to emphasize the differences among the three. Promoters represent regulatory sequences to which the transcription machinery such as Pol-II can bind and initiate transcription. Promoter regions are typically located close to the start of a gene. Enhancer regions are regulatory sequences in which transcription factors can bind to the DNA and promote the binding of Pol-II to promoters. One can imagine them as a switch for the promoter regions which turns the promoter region on or off, once a specific signal or transcription factor is present. These regions can be located

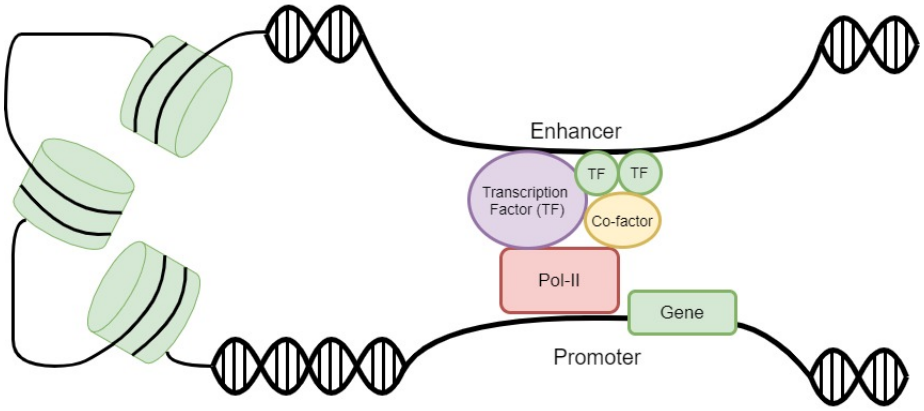

Figure 1.1: Gene transcription is dependent on the binding of RNA Polymerase II (Pol-II) onto the promoter region. Transcription can be regulated by distant binding of transcription factors (TFs) which due to DNA looping are spatially close to the gene's promoter. TFs help regulate Pol-II through the formation of protein complexes that can encourage or repress binding of Pol-II.

from several thousand base pairs (cis) to an entire chromosomal region (trans) away from the start codon of a gene [4]. A given gene may have several enhancers that interact with a single promoter (sometimes over long genomic distances) to regulate the binding of Pol-II. The binding of a transcription factor to its various target enhancers can be influenced by external signals such as DNA methylation and other epigenetic changes. As a result, the binding of multiple TFs and external signals leads to the complex combinatorial logic that influences when and where a gene will be expressed by stabilizing the binding of Pol-II to the gene's promoter [2, 5]. Lastly, insulator regulatory elements function as blockers and prevent enhancers from affecting other nearby genes. Although we will not discuss insulators any further in this thesis, they form an integral part of the transcription regulatory logic by isolating the regulatory effects of an enhancer to the gene of interest. Thus, a cell contains several layers of regulation at its disposal to regulate the appropriate expression of genes in the genome.

The binding of Pol-II is essential for the initiation of transcription, and the regulatory logic becomes important in diseased states such as cancer where mutated enhancer regions or altered signalling pathways lead to detrimental cell behaviour [6–9]. Possessing a thorough understanding of the aberrant regulatory logic may be crucial to treatment and subsequent cure for such disease states. Throughout this thesis, we will be exploring various methods to model a cell's regulatory logic using modern sequencing data, and revealing potential activators and repressors of transcription in prostate cancer. By modelling the binding of different transcription factors and the resulting transcription, we hope to tease out important regulatory factors from a pool of candidates which will provide us with a better understanding of the underlying logic.

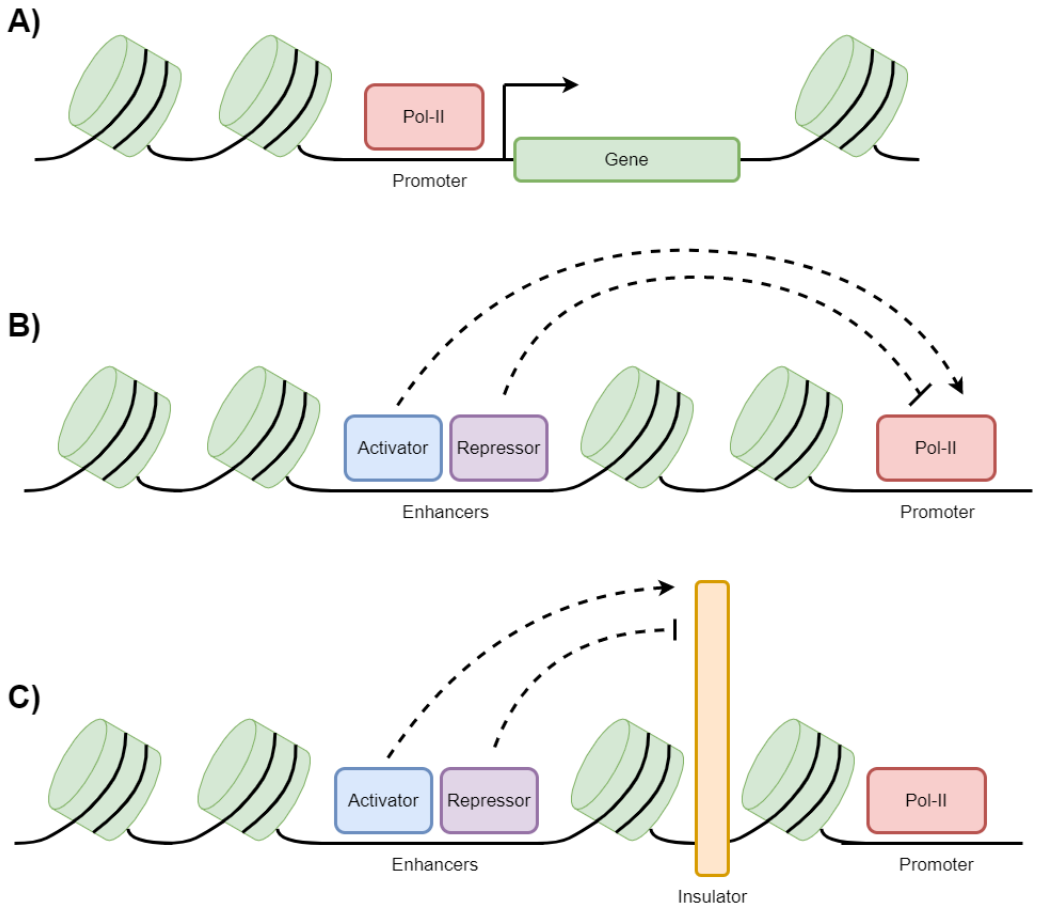

Figure 1.2: Schematic diagram of promoters, enhancers, and insulators in transcriptional regulation. (A) Promoter regions are located at the transcription start site of genes, and the binding of Pol-II initiates transcription. (B) Enhancer regulatory elements can be located thousands of basepairs away from the transcription start site and contain binding motifs for various activator and repressor TFs which will activate and inhibit transcription respectively. (C) Insulator regulatory elements block enhancers by isolating their effect to specific genes of interest.

1.2 Chromatin Immunoprecipitation Sequencing

As stated earlier, the binding of Pol-II to promoter sequences and TFs to enhancers are key steps in regulating gene expression. Therefore, having a genome-wide mapping of TF binding and other epigenetic marks will be vital to understand the underlying transcriptional regulatory network. Chromatin Immuno-Precipitation Sequencing (ChIP-seq) assays make use of protein-specific antibodies and immunoprecipitation techniques to measure the protein-DNA binding *in vivo*[10]. To achieve this, cells are treated with formaldehyde which crosslinks the DNA and everything that is bound to the DNA with larger macromolecules and facilitates detection. The extracted DNA-protein complex is then sonicated to break it into fragments [11]. Fragments of the DNA-protein complex are then collected by immunoprecipitation with antibodies specific to the protein of interest. These fragments are finally treated and sequenced using next-gen sequencing and mapped to the genome. Genomic regions that are enriched for sequencing reads are likely to have been bound by the protein of interest. The same process can be repeated for epigenetic marks such as histone marks with additional treatment and antibodies that target the epigenetic marks of interest. Since the entire library of fragments is generated from the whole genome, a single ChIP-seq experiment can generate a genome-wide mapping of DNA-protein interactions [12]. ChIP-seq can also be performed in disease states such as cancer cells where the regulatory network has been perturbed leading to detrimental cell function, potentially due to the mis-regulation of binding of TFs [13].

With a genome-wide mapping of where transcription factors and epigenetic marks are bound, we can identify potential enhancer regions in the genome. If multiple transcription factors and epigenetic marks have been shown to cluster around a distinct genomic region, then the region can be tested for its enhancer capabilities and may function as the regulatory hub for a gene of interest. This, in addition to transcription output assays, can lead to the identification of activator and repressor TFs. The identification of crucial TFs is critical in disease states where the regulatory network can lead to disrupted TF binding and altered gene expression.

1.3 Massively Parallel Reporter Assays Data

Recent years have seen the continued application of innovative sequencing technologies and techniques leading to an explosion of sequencing data in both wildtype and mutated disease states [10]. For instance, massively parallel reporter assays (MPRA) are capable of measuring the transcriptional output of hundreds of thousands of putative enhancer sequences in parallel, due to the low cost of sequencing [14]. These assays when used in conjunction with the ChIP-seq genome-wide mapping of TF are rich datasets from which computational and biophysical models can decipher the transcriptional regulatory network and mine the data set for key regulators of transcription. Throughout this thesis, our models

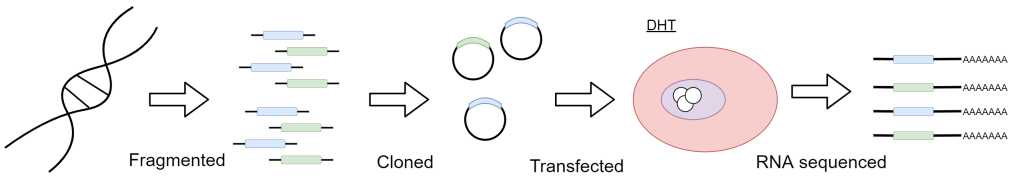

Figure 1.3: Enhancer regions are captured from clinical prostate cancer cells and tested for enhancer activity using STARR-Seq. Enhancer sequences are cloned into plasmids and then transfected into cells and treated with either androgen (DHT) or control (EtOH). Enhancer activity can be measured in the resulting RNA sequencing.

will be developed and applied to a specific type of MPRA known as Self-transcribing Active Regulatory Region Sequencing (STARR-Seq), which we will now detail [15, 16].

1.3.1 Self-transcribing Active Regulatory Region Sequencing

To determine enhancer activity, potential enhancer sequences are cloned into a reporter assay containing a fixed promoter and a reporter gene. Once cloned, enhancer sequences will recruit transcriptional machinery and activate transcription of the reporter gene. Stronger enhancer activity would result in a higher abundance of the reporter and thus evidence of high transcriptional activity. Self-transcribing Active Regulatory Region Sequencing (STARR-Seq) takes in a library of potential enhancer sequences, each of which is then cloned onto a plasmid downstream of the promoter region and transfected into cells [15, 16]. As such, the original enhancer sequence will be included in the resulting RNA transcript (Fig. 1.3) and can be detected through sequencing. This self-transcription allows for the concurrent testing of an entire enhancer library. Enhancer activity is measured by the frequency of the enhancer sequence within the RNA transcript library. For disease states, enhancer regions can be isolated from a diseased cell and be tested for enhancer activity. This can also be repeated in the presence of different extracellular signals (such as hormones) which might induce binding of specific TFs and activate certain enhancer regions over others. Deciphering how enhancer regions and TF binding may change in the presence of specific signalling pathways will reveal more of the underlying regulatory mechanics and potential targets for treatment.

1.4 Models of transcription

To understand the underlying transcriptional logic, models of transcription make use of either sequence data or binding data of enhancer regions to predict the resulting Pol-II binding and transcription expression. These modelling techniques generally fall into two areas: biophysical modelling and machine-learning techniques. This section will introduce each area and emphasize the major benefits and drawbacks of each methodology.

1.4.1 Biophysical Modelling

Biophysical models typically treat the occupancy of Pol-II as a thermodynamic problem where the probability of Pol-II can be calculated using the Boltzmann distribution over the various bound states of the enhancer(s) and promoter [1, 17]. For example, consider a single promoter region with a specific binding site and then N_{NS} non-specific binding sites for Pol-II in the rest of the genome. We will assume that all Pol-II are bound to the DNA and ignore contribution due to 'free' Pol-II that do not bind to the DNA. Given that there is a large number of Pol-II within a cell and assuming that there is a chemical potential for Pol-II to enter the nucleus, we can assume that is constant number of Pol-II that will bind to the DNA. The statistical weight of all the Pol-II binding non-specifically will be a product of N_{NS} choose P combinations $\binom{N_{NS}}{P}$ and Boltzmann weight,

$$Z_{NS}(P) = \frac{N_{NS}!}{P!(N_{NS}-P)!} \exp\left(\frac{-P\epsilon_{pd}^{NS}}{k_B T}\right) \tag{1.1}$$

where ϵ_{pd}^{NS} represents the nonspecific binding energy of Pol-II and P represents the total number of Pol-II in the system.

One can then write the total partition function as a sum of two contributions; one where all Pol-II binds non-specifically (shown above) and one in which a single Pol-II successfully binds specifically to the promoter site.

$$Z_{Total}(P) = Z_{NS}(P) + Z_{NS}(P-1)\exp\left(\frac{-\epsilon_{pd}^{S}}{k_B T}\right) \tag{1.2}$$

where ϵ_{pd}^{S} represents the specific binding energy of Pol-II to the promoter. Then using the Boltzmann distribution, the probability, P_{bound}, of a single Pol-II being specifically bound to the promoter is,

$$P_{bound} = \frac{Z_{NS}(P-1)\exp\left(\frac{-\epsilon_{pd}^{S}}{k_B T}\right)}{Z_{Total}(P)}. \tag{1.3}$$

Substituting Eq 1.2 into Eq. 1.3 and making the assumption that $N_{NS} \gg P$ leads to a simple expression for a single promoter to be bound by Pol-II,

$$P_{bound} = \frac{1}{1 + \frac{N_{NS}}{P}\exp\left(\frac{\Delta\epsilon_{pd}}{k_B T}\right)} = \frac{[P]}{[P] + K_D} \tag{1.4}$$

$$K_D = \frac{N_{NS}}{V}\exp\left(\frac{\Delta\epsilon_{pd}}{k_B T}\right) \qquad [P] = \frac{P}{V}$$

where $\Delta\epsilon_{pd} = \epsilon_{pd}^{S} - \epsilon_{pd}^{NS}$ represents the change in binding energy from specific to non-specific binding and K_D represents the equilibrium dissociation constant. $[P]$ symbolizes the Pol-II concentration and V represents the volume within the nucleus. This effectively

maps the binding of Pol-II to a classical discrete canonical ensemble problem and given the appropriate type of data can determine the corresponding binding energy of Pol-II [18].

Now to include the regulation due to the binding of transcription factors, let us first consider the presence of an activator TF which when bound will increase the probability of Pol-II binding. Assuming that all activator TF binds to the DNA, one needs to modify the partition function to include the binding of an additional activator factor,

$$Z_{NS}(P, A) = \binom{N_{NS}}{P}\binom{N_{NS} - P}{A} \exp\left(\frac{-P\epsilon_{pd}^{NS} - A\epsilon_{tf}^{NS}}{k_B T}\right) \tag{1.5}$$

where ϵ_{pd}^{NS} and ϵ_{tf}^{NS} represent the non-specific binding of Pol-II and TF respectively, and A represents the number of activators binding to the DNA. Following the same notation as before, one can then write the probability of Pol-II binding as

$$P_{bound} = \frac{Z_{NS}(P-1, A)\exp\left(\frac{-\epsilon_{pd}^{S}}{k_B T}\right) + Z_{NS}(P-1, A-1)\exp\left(\frac{-\epsilon_{pd}^{S} - \epsilon_{tf}^{S} - \epsilon^{int}}{k_B T}\right)}{Z_{Total}(P, A)} \tag{1.6}$$

where ϵ^{int} symbolizes the interaction energy between Pol-II and the activator and $Z_{Total}(P, A)$ represents the full partition function which will consider all possible states of specific and non-specific binding of Pol-II and the activator TF.

Additional TFs would all require separate modelling and exponentially increase the number of possible states. As such, the enumeration of Z_{Total} can be difficult especially given the vast number of possible factors. The lack of scalability makes it difficult to apply this technique to modern sequencing data. There are often 10s to 100s of potential factors which makes the number of possible states expensive to enumerate. Additionally, the binding energies and protein-protein interaction energies in the generalization of Eq. 1.6 are typically not known. While there have been efforts to simplify the parameters demanded by theory [17], the addition of each activation/repression transcription factor increases the complexity to model.

Nevertheless, biophysical approaches are typically easier to interpret compared to those used in machine learning. Fitted parameter estimates can be related to a chemical/physical process and interactions between factors (eg. ϵ^{int} shown above) are clearly defined. A simplification of the biophysical approach presented above would be to use a mean-field approach that relates Pol-II binding to the equilibrium occupancies of the various TFs. We will develop this approach in Chapter 3.

1.4.2 Machine-learning Methods

Contrasting the biophysical approaches where the learned parameters have direct physical interpretation, are purely statistical machine learning approaches where fitted parameters

may lack interpretability. Machine-learning approaches generally view the binding data of transcription factors as the feature/input data and the resulting transcription as the response/output data. This view compresses the complexity of binding and interaction with Pol-II into the parameter estimates of each factor and can obscure the physical interpretation of some of the model's parameters. The simplest forms of machine learning are ordinary least squares (OLS) for regression and logistic regression for classification. Both methods have been utilized in a variety of biological systems to identify important transcription factors [19, 20]. Drawbacks to these approaches will be discussed in Section 1.5, but they often require modification to compete with more modern approaches. One such modification which we will discuss in greater detail in Chapter 2 involves the use of ensemble methods to sample separate parts of data space.

Recent years have seen a growing interest in both the development and application of deep neural nets in the field of transcriptional regulation. Deep neural nets (DNN) and convolutional neural nets (CNN) are the most common deep learning approaches to model transcription. Modelled after neurons within the brain, DNNs contain multiple hidden layers that pass and transform information from the input layer before finally mapping it to the output.

$$Y(\boldsymbol{X}, \boldsymbol{W}) = h\left[\sum_{j=1} w_j^n h\left(\sum_{i=1} w_i^{n-1} h\left(...\right) + B_j^{n-1}\right) + B^n\right] \tag{1.7}$$

where Y,$\boldsymbol{X}$ represent the output variable and input matrix, respectively. w^n and B^n corresponds to the weights and biases at layer n. In this formulation, h represents the hidden layer activation function which can be non-linear to improve training and n represents the number of hidden layers within the model.

This feedforward method of information maintains the continuous mapping from layer to layer while containing enough complexity to map high-order non-linear functions onto the input space. As such, given enough depth, feedforward fully connected neural networks are capable of approximating any continuous functions between two Euclidean spaces [21]. This universal approximation theory makes the application of deep neural networks attractive, as given enough data, the model would be capable of deciphering any internal functional dependence between the input and output data. While powerful, I would like to stress that there are still inherent drawbacks to the universal approximation theory and deep neural networks. Firstly, while the approximation works for Euclidean spaces, non-Euclidean transformations (i.e. spherical geometry) might not be as simple, though some research shows the universal theory might work on non-Euclidean geometry [22]. However, I would still be cautious as these results are still contested. Secondly, while the theory states that DNNs are capable of approximating any functional form given enough depth and information, the exact amount of depth and information is still largely debated. Depending on the data provided, some neural network tasks may require hundreds of hidden layers which might make training and computation expensive. As such, one should be careful when making use of

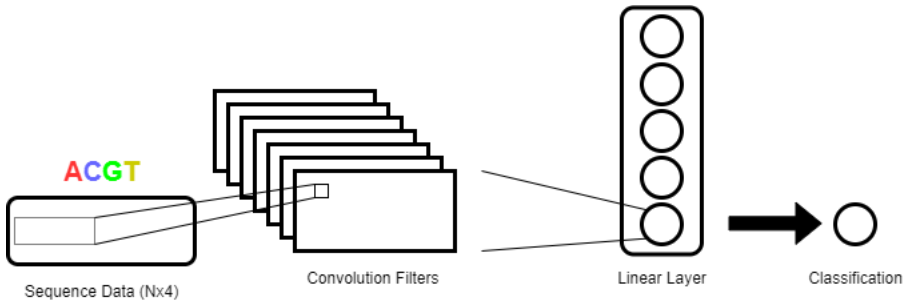

Figure 1.4: Schematic figure of a convolutional neural net (CNN) with multiple convolution filters and a single linear layer. Kernels from each convolution filter scan the input data and identify important spatial patterns that get passed on to the next layer. Deeper convulutional nets will contain multiple convolution layers where the spatial patterns of the input data are processed and passed to a fully connected linear layer before classification.

deep neural networks over other machine-learning approaches. In some situations, applications of simpler models such as logistic regression or support vector machines (SVM) may be sufficient.

Similarly to a DNN, a CNN contains hidden layers between input and output data, but employs additional kernels to scan the data for recurring patterns. As such, compared to DNNs, CNNs are more effective at identifying spatial patterns and are commonly used in machine learning tasks such as object identification and image reconstruction. In regards to transcriptional regulation, CNNs are generally paired with DNA sequence data of enhancer regions (Fig. 1.4) where the kernels will identify important DNA patterns [19]. Given enough data, CNNs can mine for key regulatory motifs and function as a way to verify the binding motifs of transcription factors from other experimental methods. Unfortunately, CNNs also contain a high number of free parameters and hence require a large amount of data to train. Compared to other machine learning methods, CNNs take significantly longer to train and may require dedicated computational hardware.

From biophysical to machine-learning models, there is a diverse array of modelling techniques for transcriptional regulation. All of them have their significant benefits and drawbacks. Therefore deciding which one to use will depend heavily on the problem that one is trying to examine. Throughout this thesis, we will explore how different approaches can be used on the same dataset, but identify different parts of the underlying regulatory picture.

1.5 Statistical Learning

While the methods that we have developed contain modelling techniques from machine-learning to equilibrium statistical physics, all of them will make use of statistical learning techniques to fit the models to experimental data. This section will serve as an introduction to statistical learning.

Assume a response variable, Y and a set of all possible input variables, ϕ such that,

$$Y = G(\phi) + \delta \tag{1.8}$$

where $G(\phi)$ represents the functional mapping from the input space to the response variable and δ represents the natural error in the response variable. Generally, it is not possible to have ϕ, and so we must estimate Y with X where X is a set of p input variables and a subset of ϕ. While $G(X)$ can take any functional form, most statistical learning techniques default to a linear approximation of $G(X)$ due to mathematical ease,

$$\hat{Y} = X\beta + \beta_o \tag{1.9}$$

$$\epsilon = Y - \hat{Y} \tag{1.10}$$

where β is the linear mapping between X and $\hat{Y}$ and $\hat{Y}$ represents the best possible linear prediction of Y given X. As such, ϵ represents the bias error of the model or the inherent error in our model construction. One popular notation is to include bias parameter β_o as an extra dimension of the input matrix X which now represents a $n \times (p+1)$ matrix. n is the number of samples/data and p represents the number of features or independent variables in our input space.

To find the β, one usually minimizes sample mean squared error which in linear regression becomes the ordinary least square (OLS) estimation,

$$\sum_i (y_i - [\beta_o + \sum_j \beta_j X_{i,j}])^2 = (Y - X\beta)^T (Y - X\beta) \tag{1.11}$$

This results in a closed-form expression of β and a projection matrix from response variable to predicted values

$$\beta = (X^T X)^{-1} X^T Y \tag{1.12}$$

$$\hat{Y} = X\beta = X(X^T X)^{-1} X^T Y = P_X Y \tag{1.13}$$

$$P_X = X(X^T X)^{-1} X^T$$

$$\epsilon = Y - \hat{Y} = (I - P_X)Y \tag{1.14}$$

This expression gives rise to an interesting interaction where one can determine the relative error without the need to fit the input matrix onto the response variable. In large datasets or cross-validation across multiple subsets of data, this interaction can mean a significant drop in computation[23].

Ordinary Least Squares (OLS) forms the basis of modern statistical learning. While quite old, successors to OLS (LASSO, Ridge) are still an active research topic in the field of

statistics. The reason being that while extremely useful, OLS has some serious limitations. For instance, OLS cannot remove irrelevant features or perform feature selection once constructed and has difficulties when handling large datasets. Additionally, OLS regression is unable to deal with highly linearly correlated or collinear features. To understand why this is an issue for regression, one can imagine two perfectly collinear features. The addition of the second feature adds no additional information to model while increasing the dimensionality of the model by one. This creates informational imbalance where the model has an additional free parameter to overfit to the data. Mathematically, high multicollinearity among the feature space results in a singular $(X^T X)$ matrix and makes the inverse in Eq.1.12 computationally impossible. To bypass this, Ridge [24], a successor to OLS, suggests the addition of λ along the diagonal of the singular $(X^T X)$ matrix,

$$\beta = (X^T X + \lambda I)^{-1} X^T Y \tag{1.15}$$

where I represents the diagonal identity matrix. The addition of the diagonal "ridge" (hence the name) on the singular matrix reduces the multicollinearity among the data and makes the inverse possible. Generally, Ridge regression is formulated through its dual optimization problem which adds an L_2 norm term to the mean-squared error,

$$\sum_i (y_i - [\beta_o + \sum_j \beta_j X_{i,j}])^2 + \lambda \sum_j \beta_j^2. \tag{1.16}$$

There are many ways to interpret the meaning behind Ridge regression and this can change depending on the field. In the machine learning field, one generally considers Ridge regularization as a modification to reduce overfitting, while a Bayesian views Ridge or L2 norm as a prior Gaussian distribution on β. However regardless of interpretation, by introducing the Ridge, we are artificially introducing bias to our model to lower the overall variance. This introduces a bias-variance tradeoff that we must change depending on the model and the data. Selection of an optimal hyperparameter, λ can be difficult depending on the data and is generally selected as the value that produces the best predictions. Additionally, although Ridge regression makes training possible, the problem of multicollinearity still makes interpretation of parameter estimates difficult. Reducing the effects of collinearity on parameter estimates will require changes to the modelling approach which we will discuss later in this thesis.

While Ridge helped solve the issue of collinear regression in OLS, it did not help with the lack of feature selection. To do this, we must rely on another successor to OLS, Least Absolute Selection and Shrinkage Operator (LASSO) [25]. LASSO follows a similar optimization problem as Ridge, but rather than an L_2 norm, LASSO applies an L_1 norm to the mean square error:

$$\sum_i (y_i - [\beta_o + \sum_j \beta_j X_{i,j}])^2 + \lambda \sum_j |\beta_j|. \tag{1.17}$$

While the change seems small, LASSO changes the shrinkage geometry and resulting parameter estimates. For instance, in the event that β is small ($\beta \ll 1$), there is an ever-decreasing impact on the L_2 regularization term, and therefore, Ridge's parameter estimates tend to keep features despite little to no impact. On the other hand, small β values have the same impact on Lasso's L_1 norm which will shrink small parameter estimates to zero. This results in the identification of crucial factors and zeroed parameter estimates for features that contribute little to none to model performance. The ability to select for features is vital for modern datasets, as they often contain a large number of possible features. Similar to Ridge regression, the introduction of Lasso adds additional bias while reducing variance in the final predictions. Both serve a similar purpose in regularizing the model and are generally lumped together in the field of machine learning. However, I would stress that the two methodologies are fundamentally different and change the outlook of the model. While Ridge regression generally presents better predictions by keeping features non-zero, Lasso enhances the sparsity and interpretability of the model. Deciding which method to use or a combination of the two (Elastic-Net) depends heavily on the model and the dataset.

Throughout this thesis, the problems of feature selection and collinearity will be a major focus as we design models to unveil the underlying transcriptional logic in prostate cancer. Since we want to identify crucial TFs for enhancer activity, interpretable parameter estimates will critical for our models. Enhancer regions contain binding motifs for a multitude of different TFs which often require simultaneous binding and correlated expression to form protein complexes. As such, binding data of various TFs are often correlated and collinear in nature. This correlated features makes the identification of crucial TFs difficult, as LASSO encourages sparsity among parameter estimates for features. Feature selection among highly collinear features will result in the identification of a single crucial factor rather than the array of different TFs which are required for transcription. To solve the issues of collinearity and feature selection, more sophisticated models are required and will be vital to uncover important TFs for enhancer activity.

1.6 Prostate Cancer and Androgen Receptor

Throughout this thesis, we will be focusing on the transcriptional regulation in prostate cancer cells and how the presence of male sex hormone changes the underlying transcriptional logic. Prostate cancer is the leading cause of cancer in men and resulted in almost 350000 deaths worldwide in 2018 [27]. This corresponds to $\sim$3% of all deaths due to cancer in men. It has been shown that high levels of androgens, a general term for the male sex hormones, play a big role in both the growth and maintenance of prostate cancer. Therefore, decreasing the levels of androgen through castration has been shown to be an effective way to combat prostate cancer [26]. In terms of transcriptional regulation, while androgen cannot directly bind to DNA, its activator effects are mediated through the androgen recep-

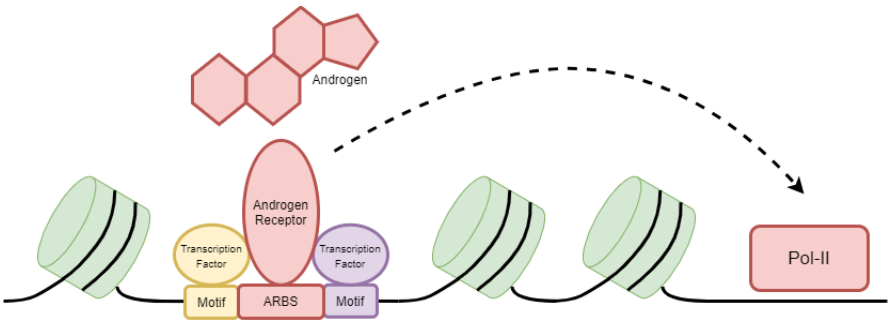

Figure 1.5: Schematic of AR-mediated transcription. Androgen receptor (AR) in the presence of androgen, a general term for the male sex hormones, binds at AR binding sites (ARBS) and recruits the binding of other TFs that help induce transcription. AR-mediated transcription plays a major role in the maintenance and proliferation of prostate cancer cells [26].

tor (AR). Androgen receptor (AR)-mediated transcription is the primary driver of prostate cancer (PCa) growth and proliferation [28]. Activation of this critical signalling pathway occurs when AR binds to androgens such as testosterone or dihydrotestosterone (DHT) (Fig. 1.5). This induces the translocation of the AR into the nucleus, where it interacts with DNA at AR binding sites (ARBS). While there are thousands of ARBS enhancer regions throughout the genome, only a fraction of enhancers are active and help induce gene transcription. The mechanism which dictates if an enhancer region is active or not is still largely unknown. Within this thesis, we will explore different methods to identify TFs that correlate with active enhancer behaviour and try to understand the underlying regulatory logic behind AR-mediated transcription. Further details about AR-mediated regulation and its role in prostate cancer will be given in the forthcoming chapters.

1.7 Outline of thesis

The work done in this thesis was done in collaboration with Nathan Lack's lab at the Vancouver Prostate Centre and Koc University. They performed the experiments and graciously shared the data with us to generate our models. Chapter 2 focuses on the data that they have generated and an ensemble machine-learning approach that I developed to uncover important factors for ARBS enhancer activity. The work presented in this chapter has been submitted for publication and is under revision. In Chapter 3, we map the complex transcription factor interactions onto a mean-field Ising model. As will be shown, this approach leads to a predictive biophysical model for ARBS transcriptional output based on what TFs are bound. One advantage of this model is that factors can be mutated/knocked down to find the relative impact on transcriptional activity. The work in Chapter 3 is currently being prepared to be submitted for publication. Finally, in Chapter 4, I summarize several other approaches that demonstrate potential that I did not have time to finish. With these

various approaches, we have made predictions for important transcription factors that lead to AR-regulated enhancer activity and potential targets for prostate cancer treatment.

Chapter 2

Functional mapping of androgen receptor enhancer activity

2.1 Introduction

As stated in section 1.6, androgen receptor (AR) binds to ARBS enhancer regions and the resulting transcription and potential misregulation plays a key role in the proliferation of prostate cancer cells (PCa) [28]. Almost all of AR-binding cis-regulatory elements (CRE) are located at distal intergenic or intronic regions [9, 29]. AR binding to ARBS regions is influenced by various transcription factors, including FOXA1, HOXB13, and GATA2 [29–31]. The translocation of AR into the nucleus and binding at AR binding sites (ARBS) recruit numerous co-activators (CBP/p300, SRC/p160), chromatin modifiers (SWI/SNF-BRG1) and co-repressors (HDAC, NCoR) in a highly coordinated manner [32]. This protein complex then physically interacts with gene promoters via chromosomal loops, activating basal transcriptional machinery to drive transcription. Yet similar to other receptors that respond in the presence of hormone, most AR-regulated genes interact with multiple ARBS [33]. There are vastly more ARBS (tens of thousands) than AR-regulated genes (hundreds) [26, 32]. We do not know if these ARBS are inactive or active enhancers that interact in an additive, synergistic or dominant mechanism to induce gene transcription. Characterization of ARBS enhancer activity is critical to interpreting the underlying regulatory logic of this critical transcription factor.

Enhancers have traditionally been identified by correlating transcription factor binding sites with chromatin accessibility, RNA polymerase II, nascent RNA (GRO-seq), or enhancer-associated histone modifications such as H3K27ac [12, 34–36]. These features all broadly correlate with active enhancers but they are not causative and are therefore extremely prone to false positives [37]. Demonstrating this, global loss of enhancer mark H3K27ac was shown to have no functional impact on gene transcription, chromatin accessibility or histone modifications [38]. Therefore, reporter assays, which quantify the enhancer-induced transcription of a gene, remain the cornerstone of enhancer validation [39]. These

assays are not significantly influenced by endogenous chromatin compaction or epigenetic modifications and test the potential enhancer capability of each specific cis-regulatory element [40]. While robust, conventional approaches are very low-throughput. To overcome these limitations several massively parallel reporter assays (MPRA) have been developed including Self-Transcribing Active Regulatory Regions sequencing (STARR-seq) [16, 41]. In this approach, enhancer activity is quantified by measuring the rate of self-transcription of the genomic region cloned downstream of a minimal promoter. The enhancer activity of many thousands of genomic regions can be measured simultaneously and provide locus-specific resolution of enhancer activity.

There is increasing clinical evidence that non-coding mutations can act as oncogenic drivers in PCa [42–44]. Recent studies by the Lack lab have shown that ARBS are highly mutated in a tissue-specific manner [45, 46]. Given the critical role of AR in PCa progression, any changes to the transcriptional landscape could dramatically change the growth of the tumour. However, establishing a causal link of these mutations to a phenotype is extremely challenging due to the lack of functional CRE annotation. Therefore the vast majority of these non-coding mutations remain unexplored in PCa. Better characterization of these CRE in PCa is essential to classify potential mutations that drive cancer development.

To provide the first locus-specific AR regulatory map, we functionally quantified the enhancer activity of all commonly observed clinical ARBS with STARR-seq. We demonstrate that only 7% of ARBS have androgen-dependent enhancer activation, while 11% have enhancer activity independent of AR binding. Surprisingly the vast majority of ARBS (81%) do not have significant androgen-dependent or constitutively active enhancer activity. Supporting these results we observed that our in vitro annotation strongly correlated to clinical PCa samples. To characterize the mechanism of AR enhancers we trained a machine learning classifier that can successfully predict active enhancers and identify key features for active enhancers.

2.2 Contribution

The work shown in this chapter was conducted by Chia-Chi Flora Huang, Shreyas Lingadahalli, Tunc Morova, Dogancan Ozturan, Eugene Hu, Ivan Pak Lok Yu, Simon Linder, Marlous Hoogstraat, Suzan Stelloo, Henk van der Poel, Umut Berkay Altintas, Mohammadali Saffarzadeh, Stephane Le Bihan, Brian McConeghy, Funda Sar, Bengul Gokbayrak, Felix Y. Feng, Martin E. Gleave1, Andries M. Bergman, Colin Collins, Faraz Hach, Wilbert Zwart, Eldon Emberly, and Nathan A. Lack. I was responsible for the implementation of the ARBS classifier as well as the resulting analysis of the predictions. I also helped draft parts of the Methods and Results section. The graphs in 2.1, 2.2, 2.3, 2.8, 2.9, 2.10, 2.11, and 2.12 were made by me, while the remaining figures are kept to provide context. Only the parts of the paper related to my work were kept in this thesis.

2.3 Material and Methods

2.3.1 STARR-seq Data Collection and Data preparation

STARR-seq and ChIP-seq data collection was done by the Lack lab and shared with us for modelling. Here we detail a quick summary of the STARR-seq and ChIP-seq data collection methodology. ARBS(n=4139) were defined as sites that were present in all normal prostate (n=3) or independent PCa tumours (n=13) [29]. LNCaP cells, a cell line for prostate cancer, were electroporated and treated with DHT/EtOH. Finally, the ARBS STARR-seq capture library was PCR amplified with DNA-specific PCR primers. The resulting fragments were sequenced by Illumina and mapped onto a reference genome (hg19). As stated before, STARR-seq transcriptional activity is determined by the frequency of the enhancer fragment within the final RNA transcript library(see Appendix for full details of the STARR-seq procedure).

2.3.2 ChIP-seq Data Collection

ChIP-seq data was generated on prostate cancer tissues from patients and immunoprecipitated and sequenced using Illumina HiSeq. This was then aligned onto the human reference genome (hg19). Previously published ChIP-seq and Gro-seq data were downloaded from the GEO database. Regions were normalized for differences in read counts and controlled for quality (see Appendix for full details of normalization). Additionally, all ChIP-seq data used in machine learning was analyzed with the standardized ChIP-Atlas bioinformatic pipeline [47].

2.3.3 ChIP-Seq/Gro-seq Data Preparation

3229 of the 4139 non-overlapping ARBS regions were categorized into Inducible (N=285), Non-Inducible (N=2479) and Constitutively Active (N=465) based on change in transcription activity under different treatments. Regions that did not fit any class description were removed from the training set. Using the data, we trained a classifier to predict the probability of being in one of the three groups using the bound factors' ChIP-seq signals in a given region as input. We have 90 ChIP-seq profile that measure the genome-wide binding of a single DNA binding factor. There are 49 unique DNA binding factors, with 35 measured in both DHT/androgen and EtOH/control conditions. Additionally, our model contains 6 factors measured in fetal bovine serum (FBS) medium which contains a small amount of DHT and other growth hormones. For each ARBS region, we extracted the ChIP-seq binding/signal scores over the 750 bp of a region, yielding 90 different ChIP-seq profiles for the region. The 90 ChIP-seq profiles formed the input feature vector for each ARBS region.

To correct for variations in scores across factors and to normalize their values in a consistent range, we applied signal extraction scaling (SES) normalization [48] to estimate a score cut-off that separates non-specific from specific binding in each ChIP-seq dataset. The

method finds the score where the difference between the cumulative sum of the sorted observed ChIP-seq scores and a uniform linear background is the largest. This score represents the point at which specific binding starts to be present in the observed ChIP-seq data. For each ChIP-seq dataset, the data across ARBS regions are pooled and analyzed for an optimal cutoff against the uniform linear input. This effectively maximizes the signal-to-noise ratio between non-specific and specific binding.

From there, the median value of the specific scores above the binding cutoff (S_{cutoff}) was then used as μ for the sigmoid transformation below. The sigmoid transforms ChIP-seq scores of individual factors into an occupancy score between 0 and 1.

$$S^* = \frac{1}{1 - e^{-\frac{S-\mu}{\sigma}}} \tag{2.1}$$

where S^* and S represent the normalized and un-normalized ChIP-seq signals respectively. μ represents the median value above the cutoff and $\sigma = \frac{(\mu - S_{\text{cutoff}})}{2}$ represents the width of the transition. This meant that any enriched ChIP-seq binding signals will fall between 0.12 ($S^*(S = S_{\text{cutoff}})$) and 1 and all non-specific occupancy scores will receive a normalized score of ≤ 0.12. Heatmaps of the average occupancy score for each bound factor at a 50 bp resolution for Inducible and non-Inducible classes are shown in Fig. 2.1 and 2.2. Finally, we took the maximum occupancy score over the 750 bp region as the feature of the factor's activity in that region.

2.3.4 ARBS classifier: model, training and prediction

The classifier we chose to fit was a bootstrapped multinomial logistic regression model with a sparsity LASSO regularizer. (Several other regularizations were tried (e.g. Ridge and Elastic-Net) but LASSO was found to give the best accuracy and interpretability.) Essentially, bootstrapped or ensemble models divide up the data space and create smaller models, called base estimators (see Fig. 2.3), that are fitted using a few features and a small amount of training data [23, 24]. Given an input array of ChIP-seq occupancy scores, X, the resulting prediction, Y, for class J of the ensemble model will be an average of the predictions across the N base estimators,

$$M_{\text{Ens}}(Y = J, \boldsymbol{X}) = \frac{1}{N}(\sum_{i}^{N} M_i(Y = J, \boldsymbol{X}_i)) \tag{2.2}$$

where M_i represents the predictions by base estimator i and M_{Ens} represents the ensemble model prediction. Since we want to classify enhancer regions into one of 3 possible classes, the base estimator will be a multinomial logistic regression or softmax model,

$$M_i(Y = J, \boldsymbol{X}_i) = \frac{1}{Z_i} e^{\boldsymbol{\beta}_J \boldsymbol{X}_i} \tag{2.3}$$

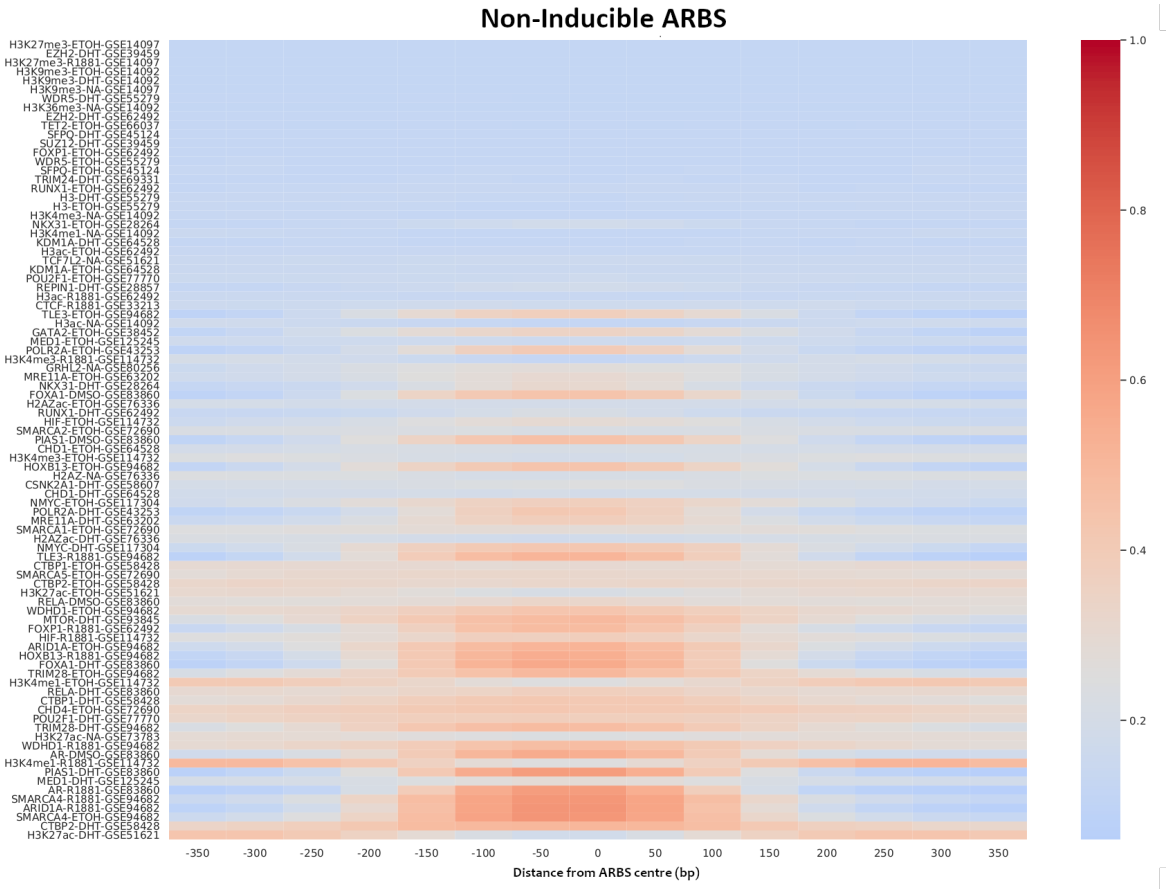

Figure 2.1: Spatial heatmap of average ChIP-seq score (50 bp resolution) after SES and sigmoid normalization (see Eq. 2.1) for all ARBS regions (N=2479) that showed no significant increase in transcription when treated with DHT (Non-Inducible). Blue regions represent low occupancy and red regions represent higher average occupancy.

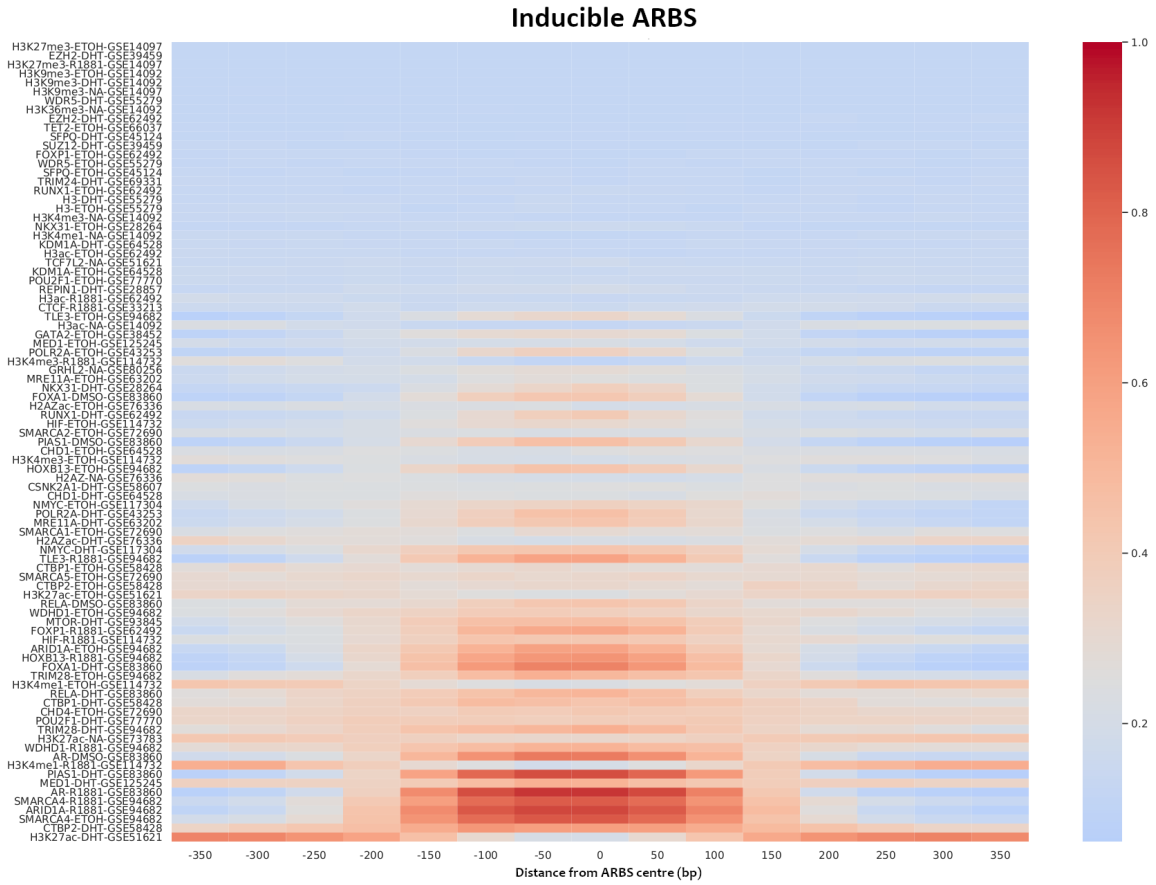

Figure 2.2: Spatial heatmap of average ChIP-seq score (50 bp resolution) after SES and sigmoid normalization (see Eq. 2.1) for all ARBS regions (N=285) that showed significant increased in transcription when treated with DHT (Inducible). Blue regions represent low occupancy and red regions represent higher average occupancy.

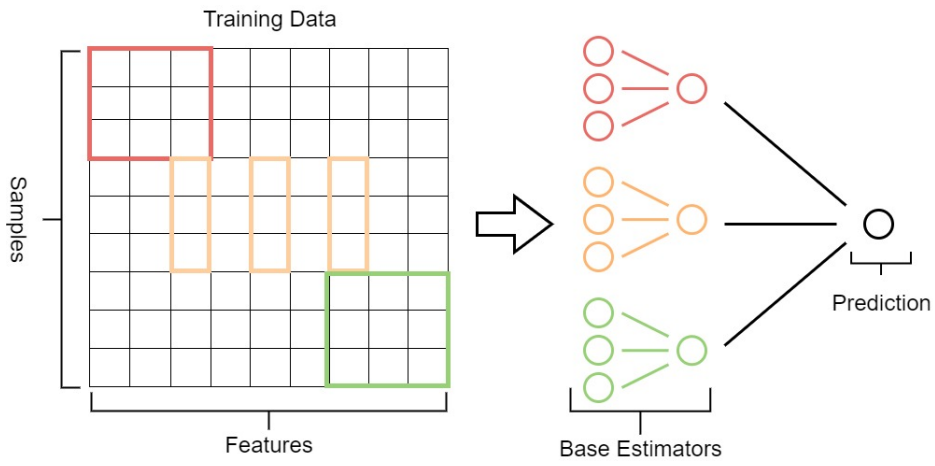

Figure 2.3: Schematic figure of ensemble learning. Training features and samples are randomly assigned to base estimators such that each estimator is trained using a fraction of the training space. Given enough base estimators, parameter estimates will reflect each feature's weight in classifying the data both in the presence and absence of every other feature. The final ensemble model prediction will be an aggregate of the predictions across its base models.

$$Z_i = \sum_{k=1}^{3} e^{\boldsymbol{\beta}_k \boldsymbol{X}_i}$$

where $\boldsymbol{\beta}_J$ represent the parameter estimates for a given class J, and Z_i represents the partition function which ensures that the predicted probabilities sum to one.

In ensemble learning, each base estimator is trained with a fraction of features (X_i) and a subset of the training samples. As such, each base model should be representative of the overall general model but with a perturbation. Therefore given enough base models, the resulting ensemble model will average out any overfitting that had occurred and lower the overall variance of the resulting predictions [24]. Each base estimator will also be subjected to the same training regularizer which sets a prior distribution on the distribution of weights, β. In our classifier, we wanted to identify essential indicators of inducible activity, and therefore LASSO regularization was selected to give us better interpretability of the final weights.

Several variations on the size of base estimators were tested but saw no improvements in prediction power. The final model consisted of base estimators that were trained using 5 of 90 possible features/ChIP-seq profiles ($dim(\boldsymbol{X}) = 5$) and 50 training samples. To ensure proper sampling of each feature, the bootstrapped model consists of 100 thousand ($N = 100,000$) randomly sampled base logistic regression estimators. This meant that each ChIP-seq feature was evaluated in ~ 6000 estimators. The fitted weights in the aggregate model are an average of the fitted weights in the base estimators and reflect the importance of a factor to differentiate the categories in the presence and absence of other factors.

By splitting up the training process and training in smaller batches, this alleviates some problems of collinearity between features and allows for a robust representation of feature importance. Training of the ensemble model was done using the `BaggingClassifier` and `LogisticRegression` function of Python's `sklearn.linear_model` and `sklearn.ensemble` packages.

In an attempt to balance the number of samples between classes, we created dataset samples that consisted of 500 randomly selected samples of the Non-Inducible group alongside all of the samples from the Constitutively Active and Inducible ARBS groups. The data set (N=1250) was further split into 80 percent training and 20 percent testing. This process was repeated three times to have an accurate sampling of the Non-Inducible group, and a separate model was fitted to each sample. The reported average contribution was the average contribution averaged over the three models. To validate the model, we computed the occupancies and constructed the input features for 10000 other putative ARBS regions in the genome that were not previously measured in the STARR-seq assay. The trained classifier was then used to predict the probability of each of the three categories for each of these regions using their occupancy features as input.

To rank the different factors in terms of predictive power for the Inducible class, we compute the contribution of a given factor as the fitted weight times the average occupancy for that factor over all Inducible regions.

$$K_l^{i,\mathrm{Ind}} = [\langle w_l^i \rangle m_l^{\mathrm{Ind}}] \tag{2.4}$$

$$\langle C_l^{\mathrm{Ind}} \rangle = \frac{1}{2}[[K_l^{\mathrm{Ind,Ind}} - K_l^{\mathrm{Non-Ind,Ind}}] + [K_l^{\mathrm{Ind,Ind}} - K_l^{\mathrm{Cons,Ind}}]] \tag{2.5}$$

where $\langle w \rangle$ and m represent the average weight across all models and average occupancy of a class respectively. $\langle C_l^{\mathrm{Ind}} \rangle$ represents the average contribution/binding energy for the Inducible class and a given factor l.

2.4 Results

2.4.1 Functional quantification of AR enhancer activity

To functionally characterize AR CRE, we experimentally tested the enhancer activity of all commonly occurring clinical ARBS sites with STARR-seq (Fig. 2.4A). During optimization, we found that ARBS inserts less than 250bp had marginal activity. As the current synthesis limit of pooled oligos is $\sim$ 200bp, we, therefore, used a capture-based approach to maintain a large insert size in the library. We designed a custom DNA capture assay to enrich the following genomic regions: common ARBS that are found in all normal tissue or primary PCa clinical AR ChIP-seq [29] (clinical ARBS; n=4139), previously identified strong enhancers that are not associated with AR [49] (positive control; n=500) and regions where the AR does not bind in either clinical samples or cell lines yet contains an ARE-motif (negative

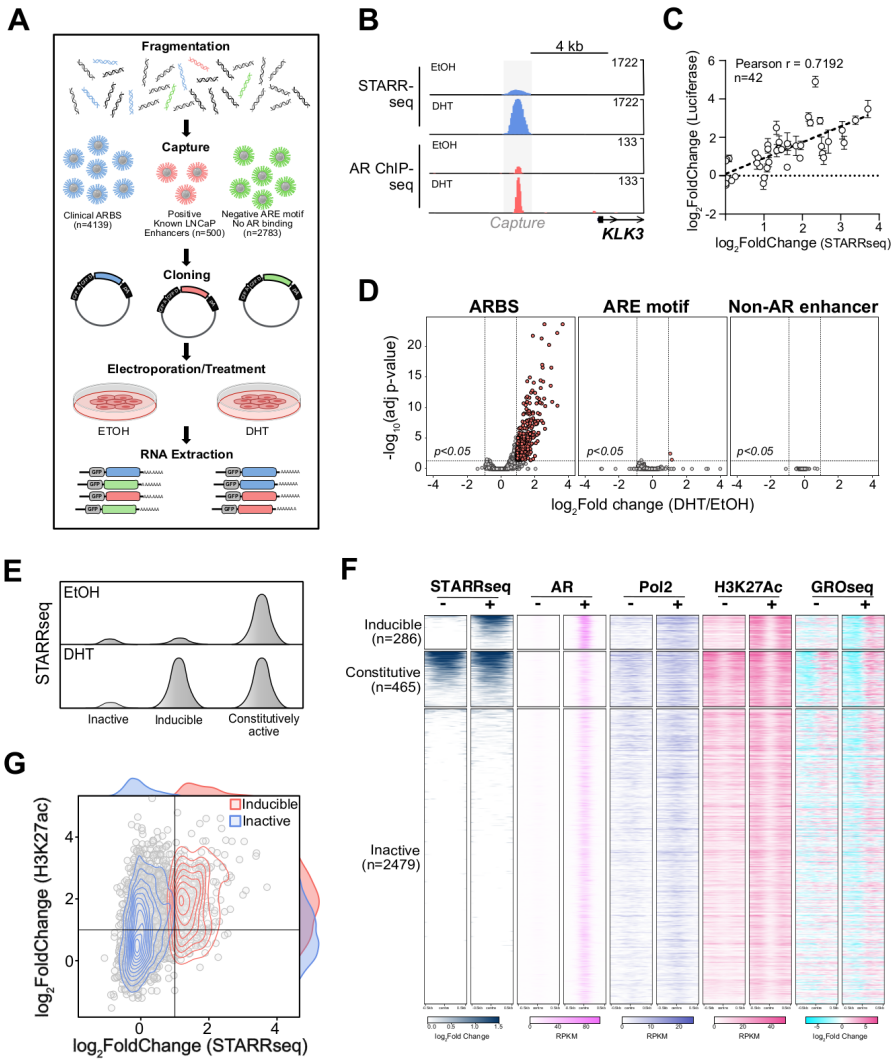

Figure 2.4: (A) Schematic representation of AR STARR-seq. (B) Strong androgen-dependent enhancer activity (red) was observed at known AR binding sites (blue; GSE83860) proximal to KLK3. (C) Enhancer activity of AR CRE with varying levels of STARR-seq signal (n=42) were validated with a luciferase assay (4 biological replicates±SEM). A strong correlation is observed between luciferase and STARR-seq signals. (D) Volcano plot of androgen-dependent changes in STARR-seq enhancer activity for clinical ARBS (n=4139), ARE motif alone (n=2783) and strong non-AR enhancers (n=500). Significantly Inducible enhancers (Log Fold change > 0.7) are highlighted in red. (E) Schematic representation of the different classes of AR enhancers. (F) Heatmap of STARR-seq, publicly available ChIP-seq of AR, Pol2 and H3K27ac as well as Gro-seq in EtOH or DHT treated LNCaP cells. The heatmap is divided based on the functional classes of each enhancer class identified by STARR-seq. (G) Density map of androgen-induced changes to H3K27ac ChIP-seq and STARR-seq at Non-Inducible and Inducible AR enhancers.

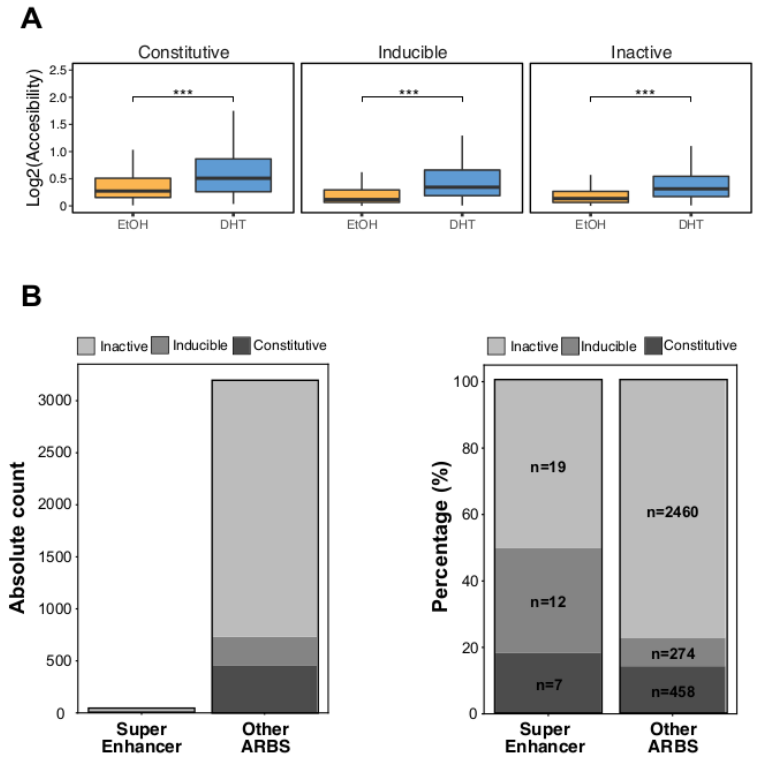

Figure 2.5: (A) Chromatin accessibility of each enhancer class +/- DHT treatment in LNCaP cells. (B) The total number of super enhancers found in each AR enhancer class (Left) and their relative percentage (Right).

control; n=2783). With this, we then captured fragmented normal genomic DNA and cloned it into a second-generation STARR-seq plasmid [16]. A total of 365,265 unique on-target inserts (median 50 inserts/region) were cloned, with a normal distribution of inserts across our capture regions and a median insert size >500bp.

Using this targeted library, we tested for AR enhancer activity in an androgen-dependent PCa cell line (LNCaP). The resulting data demonstrated excellent reproducibility across biological replicas (Pearson correlation 0.84-0.99) and a strong STARR-seq signal at known AR enhancers that regulate KLK3 and FKBP5 (Fig. 2.4B). Similar to previously published work, our results for the STARR-seq enhancer activity was comparable to a conventional luciferase reporter assay [16, 41] (Fig. 2.4C). Importantly, as STARR-seq is a plasmid-based approach the enhancer activity is independent of endogenous chromatin compaction and quantifies the potential activity at each genomic region. This is clearly demonstrated with the non-AR positive controls which had strong enhancer activity regardless of being found in either heterochromatin or euchromatin [49]. When comparing all genomic regions tested, AR-driven enhancer activity was almost exclusively limited to ARBS with only 2/2783 ARE regions showing a significant increase in signal following androgen treatment (Fig. 2.4D).

We observed three distinct classes of enhancer CRE: a "classical" AR enhancer that increases activity when treated with androgen (Inducible), enhancers that were active regardless of androgen treatment (Constitutive) and those ARBS that had no significant enhancer activity (Non-Inducible) (Fig. 2.4E; see Methods). Of these, Non-Inducible CREs were by far the most common (81.8%; 3388/4139) with no enhancer activity either before or after androgen activation. A total of 11.1% (465/4139) and 6.9% (286/4139) were Constitutively Active or Inducible enhancers, respectively. Surprisingly, none of the ARBS that were found only in clinical tissue but not LNCaP cells were Inducible enhancers (n=900). This rate of inactivity is significantly less than expected compared to the clinical ARBS that overlap with LNCaP ($p < 2.1 \times 10^{-16}$). As these regions lacked AR binding they were separated in subsequent analysis (no AR).

To confirm that our AR CRE annotations correlated with enhancer activity in vitro, we compared each group with published enhancer-associated features including H3K27ac, RNA polymerase II (Pol2) and bidirectional eRNA (Gro-seq) (Fig. 2.4F). We observed a strong correlation between each enhancer group and these features. Specifically, Constitutive AR enhancers demonstrated high levels of H3K27ac, Pol2 and eRNA that were comparable between hormone-deprived and androgen containing conditions. In contrast, in Inducible enhancers, these features only increased when cells were treated with androgens. For Non-Inducible CRE, enhancer-associated features were broadly reduced compared to active enhancers though there was some variation observed. Further, we found that there were marked differences in AR-mediated DNase I hypersensitive sites between different AR CREs, with Inducible enhancers dramatically increasing accessibility following androgen treatment (Fig. 2.5A). There was no significant enrichment for any one AR enhancer class

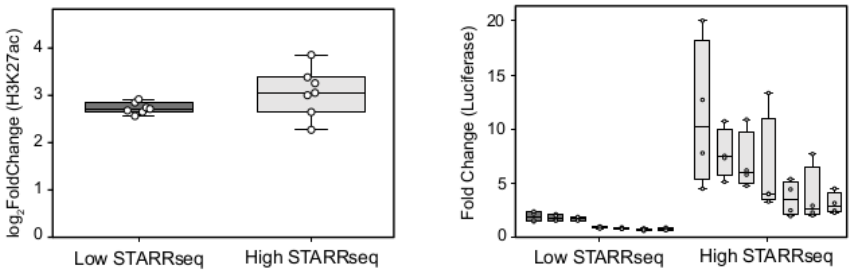

Figure 2.6: The enhancer activity of ARBS with high H3K27ac and either low Inducible (n=7) or high Inducible STARR-seq signal (n=7) were validated by traditional luciferase assay. While both the groups showed high H3K27ac (left) only the high Inducible STARR-seq group had androgen-induced reporter activity with a luciferase assay (right).

at super-enhancers as these elements are relatively rare at ARBS (Fig. 2.5B). Yet while these descriptive features generally correlate with active AR enhancers, they are extremely prone to false positives at individual CRE. For example, while Inducible AR enhancers generally have higher androgen-induced H3K27ac than Non-Inducible ARBS there is significant overlap between these classifications (Fig. 2.4G). Given that Non-Inducible CREs are far more common than Inducible enhancers, this dramatically increases the false positive rate. Specifically, if an active enhancer is called solely on AR and H3K27ac ChIP-seq, there is a >80% false positive rate. Supporting these results, the enhancer activity of high H3K27ac Non-Inducible and Inducible CRE was validated with a luciferase reporter assay (Fig. 2.6). Overall these results demonstrate that functional enhancer testing is needed to provide locus-specific resolution and annotate CRE.

2.4.2 Clinical validation of enhancer annotation

While histone modifications do not accurately identify individual AR enhancer CREs (Fig. 2.4G), these features, particularly H3K27ac, do broadly correlate with active enhancers (Fig. 2.4F). Therefore to determine if our enhancer annotations represented clinical AR activity, we analyzed previously published AR, H3K27ac and H3K27me3 ChIP-seq from primary PCa tissue (n=97) [32]. Supporting our in vitro classifications, we observed significant enrichment of both AR and H3K27ac at Inducible and Constitutive enhancers, as compared to Non-Inducible ARBS (Fig. 2.7A). Further, while not as dramatic, we also found a statistically significant enrichment of the repressive H3K27me3 mark at Non-Inducible ARBS

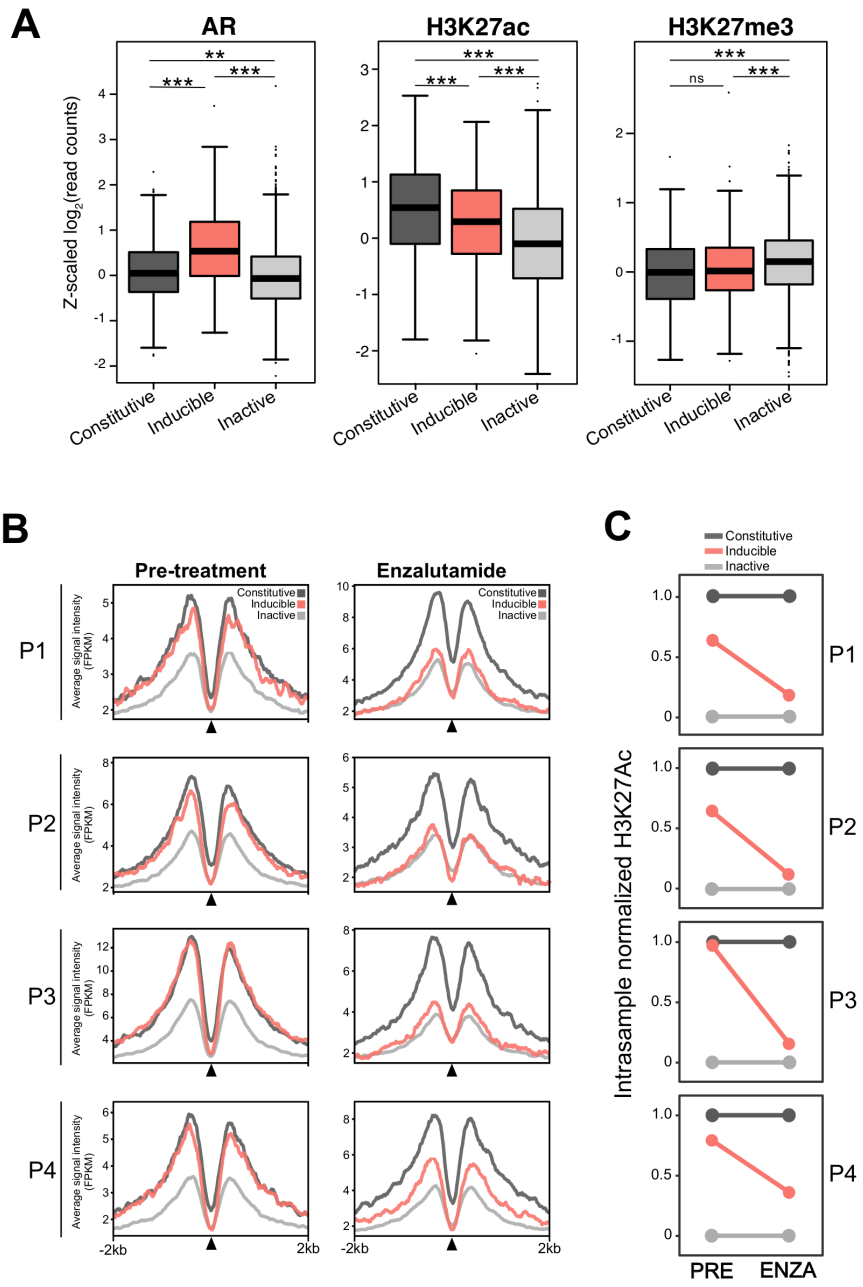

Figure 2.7: (A) Normalized ChIP-seq of AR (n=87), H3K27ac (n=92) and H3Kme3 (n=76) from primary PCa tumours. A significant enrichment of AR and H3K27ac are observed in Inducible and Constitutive compared to Non-Inducible enhancers (ns $p_{adj} > 10^{-1}$, * $p_{adj} > 10^{-1}$, ** $p_{adj} > 10^{-2}$, *** $p_{adj} > 10^{-5}$). H3K4me3 was also significantly enriched in Non-Inducible enhancers compared to Inducible or Constitutive enhancers. (B) H3K27ac ChIP-seq was done in matched patient PCa tissue pre- and post-enzalutamide treatment (n=4). Normalized H3K27ac enrichment ± 2kb around Inducible (red), Constitutive (dark grey), and Non-Inducible (grey) enhancers are shown. (C) H3K27ac enrichment in each class of enhancers was normalized within each tumour, and the normalized scores were compared before and after ENZA treatment. H3K27ac enrichment in Inducible enhancers was markedly reduced after ENZA treatment.

($p_adj < 10^{-1}$). However, as these primary PCa tumours contain physiological levels of androgen, we could not separate Inducible and Constitutive enhancers. Therefore, to further validate our in vitro classifications we conducted H3K27ac ChIP-seq on prostate tumours from patients enrolled in a neoadjuvant antiandrogen ENZA clinical trial (NCT03297385). In this, tumour samples were collected pre-and post-ENZA thereby allowing the impact of AR activity to be quantified in matched clinical samples. As expected, ChIP-seq results from pre-ENZA patients were very similar to the primary PCa samples with an enrichment of H3K27ac in Constitutive and Inducible ARBS as compared to Non-Inducible ARBS (Fig. 2.7B). However, following ENZA treatment H3K27ac was only enriched at Constitutive enhancers, and both Inducible and Non-Inducible CRE had markedly lower histone modifications (Fig. 2.7B). When normalized to Constitutive and Non-Inducible CRE, ENZA treatment markedly reduced H3K27ac at Inducible AR enhancers (Fig. 2.7C). Overall these results demonstrate that our in vitro classifications strongly correlate to clinical AR activity suggesting that this plasmid-based enhancer assay represents AR activity in situ.

2.4.3 Genomic features associated with AR enhancers

Having mapped the AR CRE enhancer activity, we then analyzed the DNA motifs at each ARBS to determine what feature correlated with active enhancers. Unfortunately, this gave very poor results, with almost no difference in DNA motifs at Non-Inducible, Inducible and Constitutively Active ARBS (Fig. 2.8A). This matches our experimental findings, where almost all of the genomic regions that contain an ARE motif but not AR binding had no inducible enhancer activity (Fig. 2.4C). Expanding on our earlier observation that LNCaP AR binding was required for activity, we incorporated all publicly available experimental genome-wide ChIP-seq data from LNCaP cells (n=90) and trained a machine learning classifier to predict enhancer activity at each ARBS (Fig. 2.9A). All transcription factor and histone ChIP-seq was processed and normalized with a standardized bioinformatic pipeline to reduce technical variation. Using this functional genomic information our bootstrapped multinomial logistic regression model with a sparsity LASSO regularizer achieved 65% precision. On test data, our model managed a 65% precision for the Inducible group and a 62% precision for the Non-Inducible group with an overall accuracy of 60%.

While 60% is a significant increase from random guessing, we next investigated regions that were incorrectly classified. Assessing these regions showed both false negatives (Fig. 2.10) and positives (Fig. 2.11). False negatives were likely due to missing values in the ChIP-seq dataset, as this group was much more likely to have a missing value than the Inducible group as a whole. The lack of critical experimental ChIP-seq values may have led the classifier to incorrectly assess these regions as Non-Inducible. While the Non-Inducible regions as a whole followed a normal distribution centred around zero fold change, the distribution of false positives was skewed towards positive fold change and the cutoff point. False positives were still mostly above zero log fold change, indicating potential enhancer

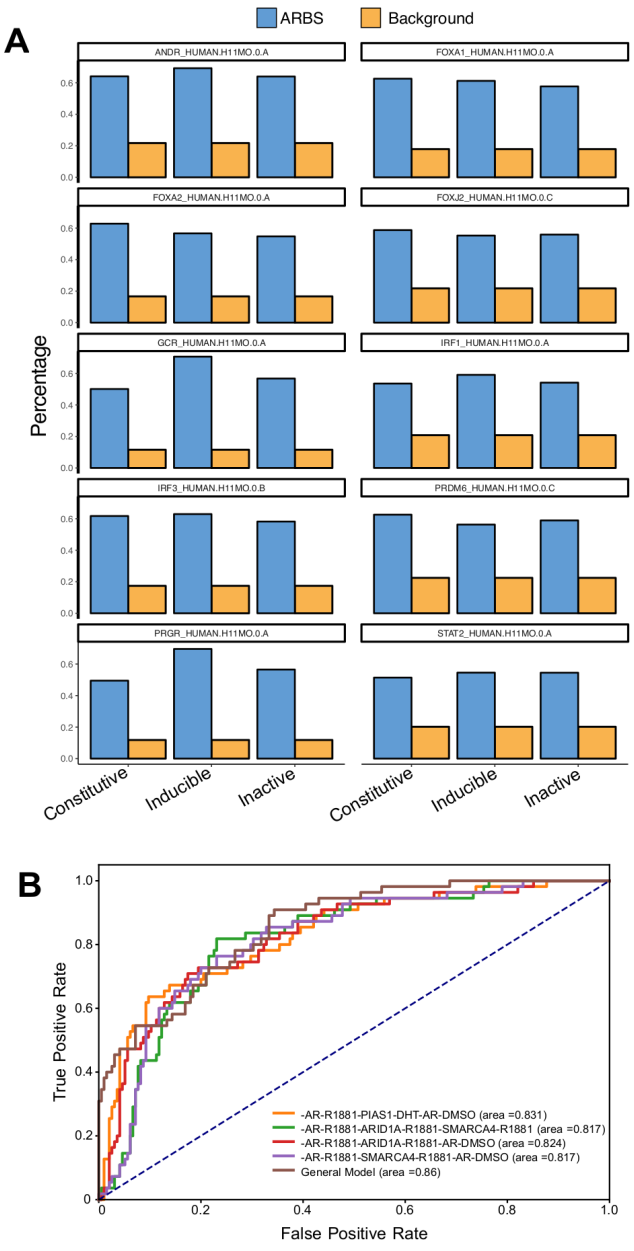

Figure 2.8: (A) Motif analysis was performed to identify the features associated with each enhancer class. The graph shows the percentage of enrichment of the indicated motif in relation to the background in all 3 different classes of enhancers across the most highly enriched motifs (n=10). (B) Receiver operating characteristic curve of the top downsampled three feature combination models (4/117480 permutations) compared to the larger general model.

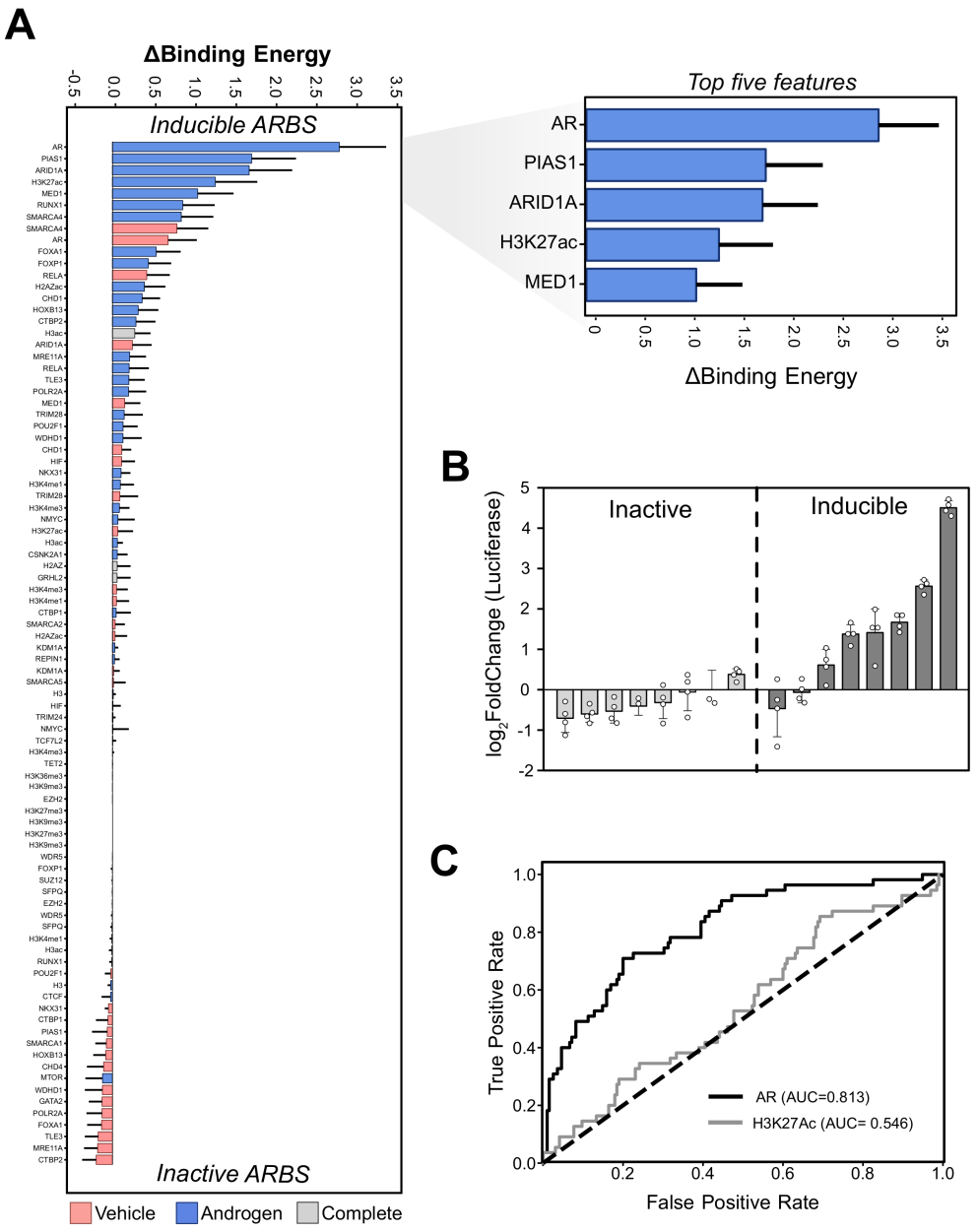

Figure 2.9: (A) Features from all publicly available TFs and histone mark ChIP-seq datasets measured either in DHT(Androgen), EtOH(Vehicle), or FBS(Complete) were ranked based on their calculated average contribution at Inducible enhancers. The figure states binding energy, but this was changed to the average contribution (see Methods) for this thesis to avoid confusion in the next sections. The inset (top right) shows the top 5 features that are predictive of Inducible enhancers. (B) LNCaP ARBS predicted by the classifier as either Non-Inducible or Inducible enhancers were validated by the luciferase assay. (4 biological replicates±SEM). (C) Receiver operating characteristic curve of AR and H3K27ac ChIP-seq to accurately identify Inducible enhancers.

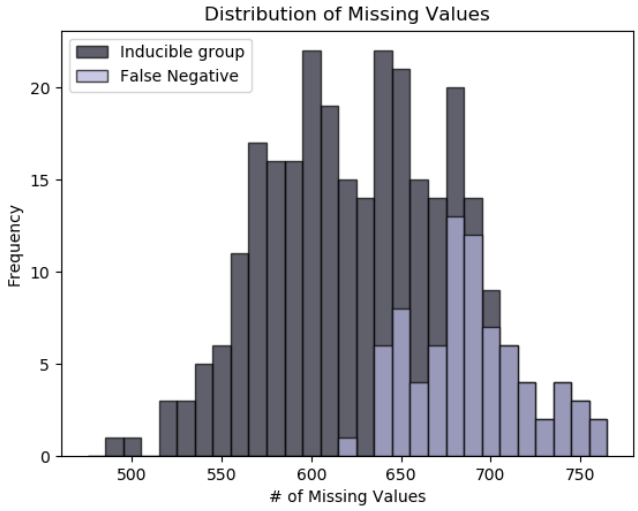

Figure 2.10: Distribution of missing binding data (NaNs) in ChIP-seq data for all Inducible regions and those misclassified Inducible regions. The x-axis represents the number of missing values within a 750bp (at 50 bp resolution) region summed across all 90 ChIP-seq profiles. Misclassified Inducible regions have a larger proportion of missing values, leading to poorly defined inputs to the classifier.

function that fell just below the fold-change cutoff to be considered in the Inducible group (Fig. 2.11).

Finally, to validate this predictive model we experimentally tested those LNCaP-specific ARBS not included in our clinical STARR-seq library (Fig. 2.9B). Over 10000 possible regions were fed into our classifier and 20 (10 Inducible, 10 non-Inducible) regions were experimentally tested. Of the ten newly tested regions, we managed to maintain a 60% and 100% accuracy for the Inducible and Non-Inducible classes respectively. This is a remarkable step-up comparable to the clinical STARR-seq library which only contained a 7% percent active enhancer discovery rate. Given the low frequency of Inducible enhancers in the training data, this is a >10x enrichment compared to random ARBS. More importantly, this demonstrates our model's ability to generalize past its training data and be applied to any LNCaP-specific ARBS regions.

As the base model uses a relatively simple multinomial logistical regression, we can quantify the predictive strength of each DNA-bound factor and identify those features that strongly correlate with Inducible AR enhancers. When calculating the differential average contribution for the Inducible and Non-Inducible groups, we observed that most features associated with Inducible enhancers were found in androgen treated conditions. Specifically, AR, PIAS1, ARID1A, MED1, and RUNX1 binding peaks in androgen treated conditions were strong predictors of Inducible AR enhancers (Fig. 2.9A). In contrast, occupancy of

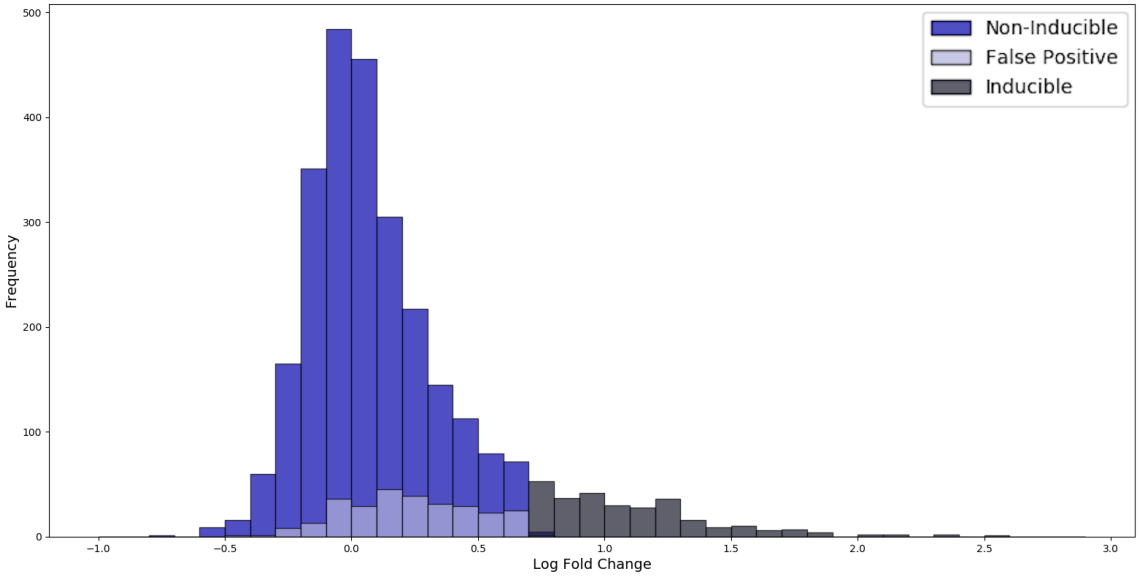

Figure 2.11: Distribution of androgen induced STARR-seq expression (log fold change) for regions that were classified as Non-Inducible, Inducible, and falsely predicted to be Inducible by the classifier. The distribution is both skewed towards positive values and close to the cutoff for the Inducible group, leading to incorrect predictions for some sequences in the Non-Inducible group.

CTBP2, WDHD1, and TLE3 at ARBS in EtOH were generally predictive of Non-Inducible CRE though these did not have comparable predictive strength to inducible features.

To identify which functional genomic features best predicted Inducible AR enhancers we downsampled our model by reducing the number of features and then re-tested each classifier compared to the general model. Since our ensemble classifier simply ranked which feature was most prominent in its base models and did not require all features to be present, downsampled models featuring a single top-ranked feature or a small combination of top-ranked features should achieve some success in classification (Fig. 2.8B). With this, we found that of all individual features, AR+DHT peak height had the best power to identify Inducible enhancers at ARBS and was significantly better than either H3K27Ac or any other transcription factors/histone marks (Fig. 2.9C; 0.81 vs. 0.55 AUC). If expanded to three features, ChIP-seq of AR+/-DHT and PIAS1+DHT gave the best results with a comparable recall to the larger general model (Fig. 2.8B; 0.83 vs. 0.86 AUC).

2.5 Discussion

2.5.1 Features of ARBS enhancers

The AR, like most nuclear receptors, binds to thousands of chromosomal sites but only regulates hundreds of genes [50–52]. The majority of AR-regulated gene promoters therefore

physically interact with multiple ARBS [33]. Yet, how these binding sites work together to induce transcription is poorly understood, as annotation of non-coding CRE has been challenging. While descriptive approaches including histone ChIP-seq or Gro-seq largely correlate with enhancer activity, they cannot provide the locus-specific resolution that is needed to understand these complex interactions. To annotate these individual regions we systematically tested the enhancer activity at all commonly occurring clinical ARBS using an MPRA assay recently optimized for human models [16]. From this we found that only a fraction of ARBS (7%) showed androgen-dependent enhancer activity, while the majority (82%) were Non-Inducible. This is comparable to work in stem cells where only a small percentage of CREs marked with NANOG, OCT4, H3K27ac, and H3K4me1 were found to be active enhancers [14, 53]. While the ARBS enhancer usage will likely change during development, the results from this plasmid-based assay strongly correlate with epigenetic features in both PCa cell lines and clinical tumours suggesting that the enhancer capability of the individual CRE are predictive of activity in situ (Fig. 2.4E and 2.7B). In matched patients pre- and post-ENZA treatment, we observed that H3K27ac at Non-Inducible and Constitutive enhancers was generally not affected by AR inhibition, while Inducible enhancers had a marked decrease when the patient was treated with ENZA. Sorting these results, in a larger patient population of primary PCa samples, H3K27ac was significantly enriched at both Inducible and Constitutively active enhancers as compared to Non-Inducible CRE (Fig. 2.7A). However, these descriptive ChIP-seq results only broadly correlate with AR enhancer activity (Fig. 2.4F). They cannot annotate individual CREs and provide the locus-specific resolution needed to characterize complex protein interactions, identify non-coding mutations or delineate the underlying regulatory logic of AR-mediated transcription. By systematically testing all clinical ARBS for enhancer activity, this provides the first detailed "map" needed to investigate these clinically important problems.

We next identified what features correlated with active enhancers. We initially characterized the different DNA motifs in each AR enhancer class but this provided almost no predictive power, with ARE motifs being equally distributed between Inducible, Non-Inducible and Constitutive enhancers. This is in contrast to work with glucocorticoid receptor that suggested steroid response elements were more likely to be found at active enhancers [52]. Potentially this may be due to the larger size of the ARBS tested in our assay, as many motifs will be frequently found in the large fragments (>500bp) incorporated into our STARR-seq library. Yet while DNA motifs do not stratify enhancer CREs, AR binding was essential for androgen-mediated enhancer activity. Only those clinical ARBS that had an AR peak in LNCaP cells were Inducible enhancers. Based on these results, we expanded our analysis and trained a machine learning classifier to predict enhancer classes from all publicly available ChIP-seq studies in LNCaP including transcription factors and histone marks. Once developed, we experimentally validated this model and could successfully identify those regions likely to be Inducible AR enhancers (Fig. 2.9B). Importantly, as

our classifier was not a black box model we could identify those features that were strongly predictive of AR enhancer activity (Fig. 2.9A). As this classifier is predictive for AR induction, it will not identify those constitutive features found at most ARBS. With this, we found that Inducible enhancers strongly associate with AR, enhancer-associated features including H3K27ac and MED1, and co-activators such as PIAS1 and ARID1A [32, 54]. With this model, we then systematically downsampled each feature and tested the recall compared to the larger model. AR when treated with DHT was more predictive at identifying Inducible enhancers at ARBS than any feature including H3K27ac (Fig. 2.9C). These results help to identify enhancers in those samples that cannot be easily or ethically tested with MPRA but are commonly used for ChIP-seq, including patient-derived xenografts or clinical samples. While further validation is required, our model sorts AR ChIP-seq peaks by height to stratify Inducible AR enhancers in samples where functional testing is not possible.

2.5.2 Benefits of the ARBS classifier and feature selection

Since the introduction of high throughput sequencing and the explosion of sequencing data, there has been an increased interest in applying statistical-learning and other machine-learning methods to these problems. Unfortunately, a cell's transcriptional logic often involves the formation of protein clusters which requires the simultaneous binding of multiple transcriptional factors. As such, the binding data of enhancer regions are often correlated and contain a high number of potential co-factors. This makes fitting such a model difficult using standard statistical approaches such as Ordinary Least Squares (OLS). While newer approaches such as Ridge [55] and Lasso [25] made training possible, the inherent collinearity of the data inflates coefficients and makes feature selection difficult [56, 57]. This is significant as most biological studies emphasize interpretation and the ability to determine function from feature. This has motivated us to employ ensemble learning, a method that is both capable of producing interpretable coefficients and accurate predictions. As stated before, the use of ensemble methods enables us to avoid overfitting in a heavily collinear training space without the need for heavy regularization. This resulted in the predictions on regions that were not in the initial STARR-seq assay, achieving a 60% discovery rate of Inducible regions.

Additionally, ensemble models are able to perform feature selection on the smaller models where the degree of freedom is significantly lower. This is important, as we want our feature coefficients to reflect each factor's significance for the overall model, as well as its ability to separate between classes. For instance, factors such as RUNX1, AR, and ARID1A all share the same increased occupancy from Non-Inducible to Inducible ARBS. Feature selection using Lasso on such a dataset would have resulted in an inflated coefficient for a single feature, rather than an equal distribution of the weights. We demonstrate this in an earlier attempt on regression (Fig. 2.12) using a smaller subset of the data which only resulted in

the recovery of two important features. Randomly re-sampling the training space allowed each feature to be measured both in the presence and absence of every other factor. As a result, the final learned coefficients are more representative of a given factor's ability to differentiate between classes. Despite AR receiving a majority of the average contribution in the models in which it was included, we were still able to identify several other transcription factors that were important for Inducible classification.

Future work could focus on the application of ensemble learning methods to game theory's Shapley values [58]. Shown below, Shapley values come from conditional game theory, and are a way to fairly distribute the contribution for each participant in a group:

$$\phi_i = \frac{1}{|N|!} \sum_{S \subseteq N \setminus \{i\}} |S|!(|N| - |S| - 1)![f(S \cup i) - f(S)] \tag{2.6}$$

where N and S represent the complete set and a single set of features respectively. $f(S)$ is the model output given the set of features, ϕ_i represents the Shapley value for a factor i. Basically, Shapley values measure the corresponding contribution due to the addition of a feature compared to the model output without it. This is repeated through all possible permutations of the data space. Unlike statistical learning methods, Shapley values measure contribution based on the accuracy of the resulting model. As such, Shapley values bypass many pitfalls of fitting (e.g. features selection, collinearity) and provide a mathematically rigorous method to rank contributions. Recently, there has been an increased interest in the application of Shapley values as an interpretation tool for machine learning models [58]. However due to the high complexity of many modern machine learning models, and the current lack of optimization, accurate determination of Shapley values is computationally intensive. One possible solution would be to bridge ensemble learning methods with the calculation of Shapley values. Mathematically, rather than sampling through all possible permutations of the feature space, one can heuristically apply a randomized sampling through ensemble modelling of a certain size. Given enough samples and equal probability for all sets, one can approximate the Shapley value for a given feature by determining the change in prediction power when a model includes and lacks the feature of interest. Given that the current infrastructure is already optimized for the parallelization of ensemble learning, the calculation of the randomly sampled Shapley value would be less computationally intensive.

In summary, we developed an ensemble logistic regression classifier trained on the LNCaP Chip-seq and STARR-seq assays. Cross-validation tests showed that our logistical model was able to achieve 65% precision for the Inducible regions and an overall 60% accuracy on clinical LNCaP test data. Additionally, we demonstrated the effectiveness of ensemble learning methods to perform feature selection even on an extremely large and collinear training space. Lastly, we showed that such an approach was able to identify important co-activators of enhancer activity such as PIAS1 and ARID1A while still main-

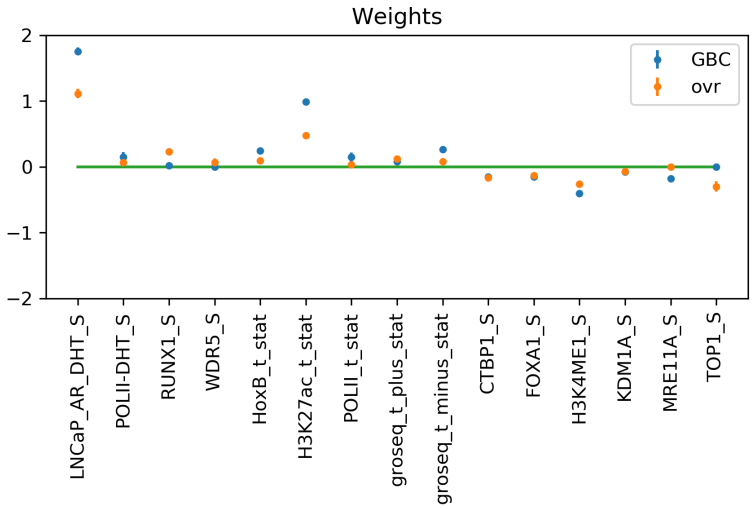

Figure 2.12: Weights of a prior logistic regression model (ovr) and gradient boosted logistic regression model (GBC) trained on a similar set of Chip-Seq and Star-seq data reveals the power of ensemble learning. Despite having a significantly smaller feature space, the prior model was only able to identify AR and H3k27ac as important factors for the Inducible class. Due to the inherent high collinearity of the feature space, other important factors such as RUNX1 and FOX were not identified.

taining the prediction power to generalize past its training data and be applied to any general LNCaP ARBS region.

2.6 Appendix

2.6.1 Cell lines

Cell lines were purchased from ATCC and routinely tested for mycoplasma contamination. LNCaP cells were routinely grown in RPMI 1640 media (Gibco) with 1% Penicillin-Streptomycin and 10% fetal bovine serum (FBS). No activation of the IFNγ pathway by double-stranded DNA was observed in electroporated LNCaP cells which has been shown to systematically alter STARR-seq activity [16].

2.6.2 Generation of ARBS STARR-seq library

Common clinical ARBS were defined as those sites that were present in all normal prostate (n=3) or independent PCa tumours (n=13) [29]. Pooled human male genomic DNA (Promega) was randomly sheared (500-800bp) using a Covaris M220 Focused-ultrasonicator. The fragments were end-repaired and ligated with Illumina compatible adaptors using the NEBNext UltraTM II DNA Library Prep Kit (NEB). The adaptor-ligated DNA was hybridized to custom Agilent biotinylated oligonucleotide probes across a 700bp region (53032 probes; 4.684 Mbp oligo) and then pulled-down by Dynabeads M-270 Streptavidin beads (NEB).

The post-capture DNA library was amplified with STARR in-fusion F and STARR in-fusion R primers, and then cloned into AgeI-HF (NEB) and SalI-HF (NEB) digested hSTARR-ORI plasmid (Addgene plasmid #99296) with NEBuilder HiFi DNA Assembly Master Mix (NEB). The ARBS STARR-seq library was transformed into MegaX DH10B T1R electrocompetent cells (Invitrogen). Plasmid DNA was extracted using the Qiagen Plasmid Maxi Kit.

2.6.3 ARBS STARR-seq

LNCaP cells (>1.3 x 108 cells/replica; 3 biological replicas) were electroporated with 266-300ug (1 million cells:2ug DNA) of the ARBS STARR-seq capture library using the Neon Transfection System (Invitrogen). Electroporated cells were immediately recovered in RPMI 1640 medium supplemented with 10% FBS. After overnight recovery, the media was changed to RPMI 1640 medium supplemented with 5% charcoal-stripped serum (CSS), and 1% Penicillin-Streptomycin. Approximately 72 hours after electroporation, the cells were treated with 10nM DHT or EtOH for 4 hours, washed with PBS and then lyzed using the Precellys CKMix Tissue Homogenizing Kit and the Precellys 24 Tissue/Cell Ruptor (Bertin Technologies). Total RNA was extracted using Qiagen RNeasy Maxi Kit (Qiagen) and the mRNA was isolated using the Oligo (dT) [16] Dynabeads (Thermo Fisher). The isolated mRNA samples were treated with Turbo DNase I (Thermo Fisher), synthesized into cDNA using the gene-specific primer, treated with RNaseA (Thermo Fisher), and PCR-amplified (15 cycles) with the junction PCR primers (RNA-jPCR-f and jPCR-r primers). The ARBS STARR-seq capture library was PCR amplified with DNA specific junction PCR primers (DNA-jPCR-f and jPCR-r primers). After junction PCR and AMPure XP beads clean-up, an Illumina compatible library was generated by PCR amplification with TruSeq dual indexing primers (Illumina) and sequenced on Illumina HiSeq4000 (150bp; PE). The resulting sequencing data is available at GSE151064.

2.6.4 Analysis of STARR-seq data

Reads were mapped to reference genome (hg19) with BWA aligner (v0.7.17) [59] and all mapped reads with a MAPQ score <60 or indels were removed. Captured region coverage was quantified with BamCoverage function (v3.1.3) in DeepTools [60] while discarding all reads on blacklisted regions (ENCODE ENCFF001TDO). Differential enhancer activity was quantified by DESeq2 (v1.26.0) [61]. Inducible enhancers were defined as having a log2 fold-change (LFC)>1 and p.adjusted<0.05 in the DHT/EtOH samples. Constitutively active enhancers had plasmid-normalized reads LFC>1 in both EtOH and DHT but no DHT induction (DHT/EtOH LFC<1). Inactive regions had both minimal DHT inducible activity and low plasmid to RNA ratios (plasmid-normalized LFC<1). The output of the DESeq2 was visualized with ggplot2 [62].

2.6.5 Clinical approval and sample collection

Clinical PCa tissue was collected before and after Enzalutamide (ENZA) therapy from the DARANA study (ClinicalTrials.gov #NCT03297385). The trial was approved by the IRB of the Netherlands Cancer Institute. Informed consent was signed by all participants enrolled in the study and all research was carried out in accordance with relevant guidelines and regulations. Trial participants received three months of neoadjuvant ENZA treatment prior to robotic-assisted laparoscopic prostatectomy. Biopsy-(pre-treatment) and prostatectomy-specimens (post-treatment) were fresh-frozen and sectioned prior to immunoprecipitation. Tissue sections were examined pathologically for tumour cell content and only those samples with a tumour cell percentage of $\geq 50\%$ were used for tissue ChIP-seq analysis.

2.6.6 Tissue ChIP-seq

Chromatin immunoprecipitations (ChIP) on prostate cancer biopsy- and prostatectomy-tissues were performed as previously described [13]. Nuclear lysates of each tissue specimen were incubated with 5 μg of H3K27ac antibody (Active Motif, 39133) pre-bound to 50 µL magnetic protein A dynabeads (Thermo Fisher Scientific, 10008D). Immunoprecipitated DNA was processed for library preparation (Part# 0801-0303, KAPA biosystems kit) and samples were sequenced using an Illumina HiSeq 2500 system (65 bp, single-end). Sequences were aligned to the human reference genome hg19, duplicate reads were removed, and reads were filtered based on MAPQ quality (≥ 20).

2.6.7 Tissue ChIP-seq data processing

Intensity plots were generated using EaSeq [63]. For boxplots, the number of sequence reads per region of interest was calculated using bedtools multicov (v2.25.0) [64]. The data were further processed in R (v3.4.4) (R Core Team, 2018). Region read counts were z-transformed per sample to correct for differences in total read count. Statistical significance in read counts differences was determined using the Mann-Whitney test, based on the median read count over all samples, and adjusted for multiple testing using FDR.

2.6.8 ChIP-seq and Gro-seq analysis

Previously published ChIP-seq and Gro-seq data were downloaded from the GEO database and uniformly processed. The sequencing reads were controlled for quality using FASTQC and the reads were then mapped to the human reference genome (hg19) with bowtie aligner (v0.12.9) [65]. All reads mapped to the blacklisted regions (ENCODE ENCFF001TDO) were discarded. For a direct comparison of the H3K27ac ChIP-seq with the STARR-seq data, average signal values of the Inducible and Non-Inducible regions were calculated using bigWigAverageOverBed software(v377) [66]. For each region, log fold change values were calculated between DHT and EtOH treatments in H3K27ac and STARR-seq experiments.

Scatterplots were generated by R's ggplot2 package (v3.2.0) [62]. ROSE [67] was used to identify super enhancers from DHT H3K27ac ChIP-seq. BigWig signal tracks were generated with BamCoverage function (v3.1.3) of deepTools [60] with RPKM normalization. For Gro-seq data, using log fold enrichment between + strand of DHT over the + strand EtOH a single track was generated. Similarly, a single - track was also generated. Then - strand is subtracted from + strand to the final bigwig file. For each of the 500 positive control enhancer regions, we calculated the average accessibility scores using bigWigAverageOverBed from ENCODE's LNCaP DHT (ENCFF975MZT) and EtOH (ENCFF906QXX) DNase-seq experiments. Then we compared the top and bottom 100 to show differences in accessibility. Later, for the same regions, we calculated the average STARR-seq signal with bigWigAverageOverBed36 from merged DHT and EtOH samples.

Chapter 3

Learning a predictive biophysical model of transcriptional regulation from massively parallel transcription assay data

3.1 Introduction

Transcription of a gene depends on the binding of RNA Polymerase II (Pol-II) to its promoter. This binding is regulated in part through interactions with other proteins, known as transcription factors (TFs) that can either stabilize or inhibit the binding of Pol-II to DNA. Many TFs bind to specific DNA motifs, and in higher organisms, enhancer regions often contain clusters of such motifs. As stated before, misregulation of TF binding can disrupt the Pol-II binding and in extreme situations, lead to the over- or under-expression of crucial genes causing diseased states [68]. As such, insight into the combinatorial logic governing the transcriptional regulation of TFs known to contribute to diseased states can prove to be crucial in our understanding of both healthy and diseased tissue.

The combinatorial logic governing the transcriptional regulation of a set of transcription factors and their target genes can be assessed experimentally using a variety of assays [10, 12]. In local genomic contexts, select enhancer and promoter sequences can be altered to assess the effect on transcription. Such approaches have provided detailed insights into the regulatory logic at specific loci, highlighting the importance of specific biophysical processes such as binding strength, motif spacing and looping [10]. Genome-wide transcriptional assays such as microarrays, later followed by RNA-seq, allowed for bioinformatics approaches to try and mine for regulatory sequences and motifs within them that could explain the observed output [11, 15]. Such methods were always indirect, and specific experiments would have to be done to validate the predictions of the bioinformatic searches. More recently, high-throughput assays that can measure the self-transcriptional output of 100,000's of DNA sequences in parallel have become available, including Massively Parallel Reporter Assay

(MPRA) [69] like Self-Transcribing Active Regulatory Regions sequencing (STARR-seq) [41]. Previous work has shown that increased self-transcription and increased binding at enhancer regulator elements correlates with increased transcription [70]. As such, these methods provide an unprecedented amount of quantitative data on how sequences directly regulate transcription [41, 49, 71].

With the availability of such a large amount of data that directly relate sequence to function (here, the measured transcriptional output), several computational approaches have been applied to identify important sequence motifs [19]. These have predominantly been solely sequence-based machine-learning algorithms [72], aimed at identifying features that either classify the DNA sequences as functional [73] or directly quantifying the amount of measured expression [19, 20, 69, 74, 75]. Many of the learned sequence features correlate with known motifs [20, 69, 75]. With these methods, the transcriptional function of individual bases can be scored, thereby providing a powerful method for predicting the effects of mutations within a particular genomic context [76]. However, there is often a large number of unidentified motifs and sorting through the complex grammars is challenging, though efforts have been made to facilitate interpretability of these models [20, 77].

Complementing the machine learning methods are biophysical models of transcription that aim to reduce the overall complexity inherent in the previous methods [1, 17, 18, 78]. These models are built on the chemistry of the biophysical processes, including the binding of the TFs and Pol-II to the DNA, interactions between the DNA-bound proteins, and potentially the structuring of the DNA itself. Although the cell is a non-equilibrium system, a common starting point for these models is the assumption that transcription is in a steady-state and near equilibrium with the current state of the cell. Given the kinetics of the transcriptional process, compared to other cellular changes that might regulate it, this is usually not unreasonable. Thus, equilibrium statistical physics provides the machinery to approximate the mean steady-state occupancies of the various factors to their cognate DNA binding sites, including that of Pol-II. We assume the cell is behaving in 'quasi-equilibrium' state in which during the time scales of the experiment, the steady-state of the cell can be approximated by equilibrium statistical mechanic techniques. The transcriptional output is then directly related to the likelihood of Pol-II being bound. These models have been validated in a variety of contexts including altering the directions, spacings, and numbers of the types of DNA binding sites [79, 80].

In this chapter, we develop a method to fit a biophysical model of transcription to the quantitative data provided by massively parallel self-transcription assays. Instead of trying to fit a biophysical model that tries to discover DNA motifs within the tested sequences, we use genome-wide binding data for a large number of known transcription factors, co-factors and epigenetic marks as input to the model. Thus the model is constrained to only have a certain number of known bound factors that interact with Pol-II. One advantage of this approach is the ability to use binding data measured across several conditions that include

implicit information about interactions, that may be missed in purely sequence-based models. Sequence-based models such as convolution neural nets (CNNs) are only dependent on the DNA sequence of enhancers region which can be misleading, as the presence of a motif may not lead to binding without secondary chemical signaling such as the presence of a specific hormone. To overcome this, our approach relies on binding and sequence data measured in the presence of hormone and control which will include additional information of the underlying interactions between factors. Also, the model parameters are directly interpretable as the binding site energies of the various factors on the tested sequences and the energies of interaction between the factors, including Pol-II. Once fit, the model can then make predictions for how knocking out factors or altering interactions would affect transcription.

We have applied our model to STARR-seq data generated from a prostate cancer cell line. The specific sequences that were tested were chosen based on their binding to the TF, Androgen Receptor (AR), whose misregulation is associated with prostate cancer [28, 29, 81]. AR binds to androgens including DHT, testosterone and R1881, which confer the ability to enter the nucleus and then bind DNA. Hormone bound AR binds to numerous sites throughout the human genome which persist during prostate cancer development [28]. Over 6000 genomic regions (corresponding to more than 100,0000 sequences assayed by STARR-seq) that bind AR in clinical samples were tested for their self-transcriptional activity under two conditions: with and without androgen. As input to the model, we used the DNA-binding data for AR and 34 (35 in total) other factors measured in both the presence and absence of androgen. After fitting our model, we are able to identify several key factors, besides AR, that are essential in setting both the overall transcription level, and whether a sequence will differentially express under the two test conditions. The fitted model is then used to assess the importance of each factor in altering transcription in hypothetical knock-out experiments. Our work highlights a network of complexes that come together to confer function to AR binding enhancers.

3.2 Contribution

The work shown in this chapter was conducted by Eugene Hu, Tunc Morova, Eldon Emberly, and Nathan A. Lack. I was responsible for the implementation and analysis of the mean-field model. I also drafted sections of methods and results as well as the developed figures in this chapter.

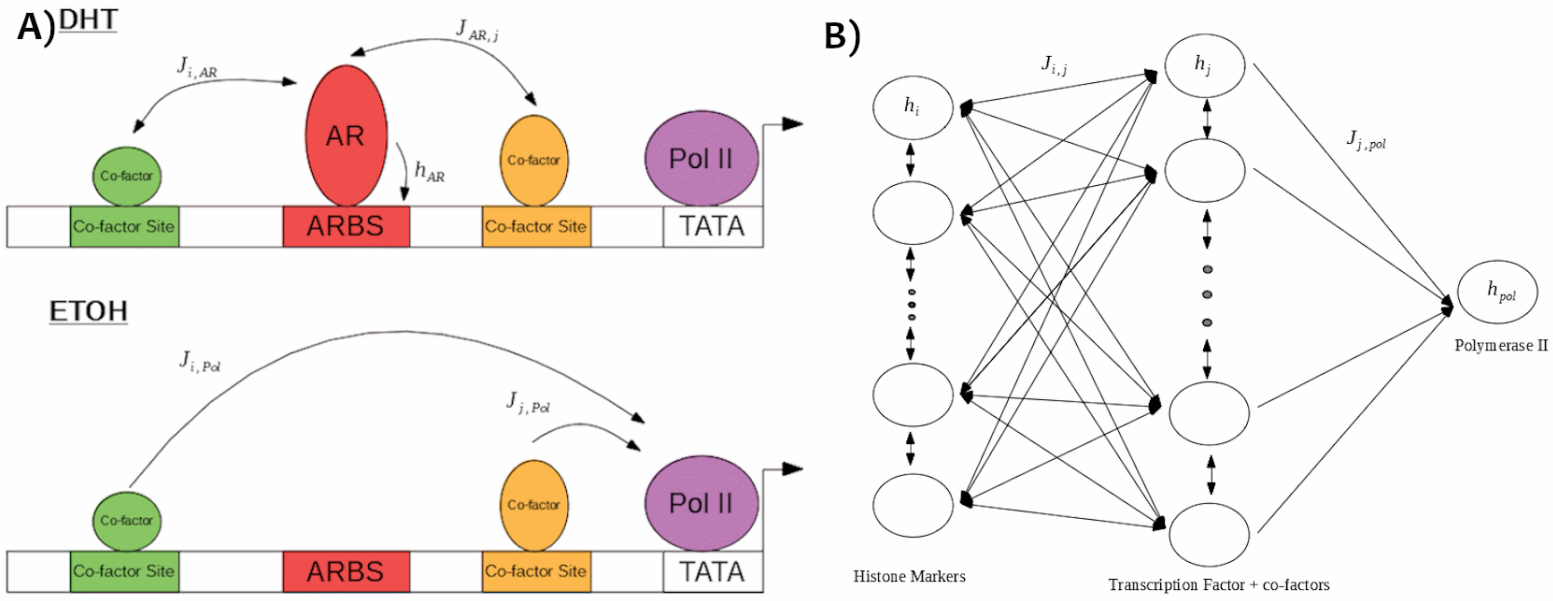

Figure 3.1: Schematic of model for transcriptional regulation. A) Androgen Receptor (AR) activator interacting with the promoter. Binding sites for AR and other transcription factors are shown. Sites can be either bound or unbound, and binding is governed by a site energy at each site, $h^{\alpha}_{k,i}$ and energetic interactions with other factors, $\{J_{i,j}\}$. The equilibrium occupancy of a site will depend on these energies. Modern high-throughput transcriptional sequencing of the enhancer is proportional to the occupancy of the Pol II binding site. Shown are the enhancer in two environmental conditions: with hormone (DHT) and without (EtOH). The environment affects a factor's binding site energy, $h^{\alpha}_{k,i}$, but does not alter the pairwise interaction energies $\{J_{i,j}\}$ between factors. B) A network representation of the model, showing how the various sites interact. Each site is represented by a binary variable, $\sigma^{\alpha}_{k,j} = 0, 1$ representing whether it is unbound or bound. Given measured values of the equilibrium occupancies of each site, the model's energetic parameters can be determined.

3.3 Methods

3.3.1 Biophysical model of transcription

We take each putative enhancer region to potentially be able to bind M different DNA-bound factors (including Pol-II). The i^{th} factor's binding on region k in condition alpha (e.g. cells treated with EtOH would be one condition, and those exposed to DHT would be another) can be described by a binary variable, $\sigma^{\alpha}_{k,i} = 0$ or 1 with zero being unbound and one, bound. When the i^{th} factor is bound it contributes a site energy $h^{\alpha}_{k,i}$ which combines both a specific energy due to binding and the factor's chemical potential (which depends on its concentration which can be condition-dependent). If factors i and j are bound, there can be an additional interaction energy $J_{i,j}$, and we assume that this interaction is treatment/condition-independent, and only depends on the particular proteins (see Fig. 3.1A), which implies that the physical interactions between proteins do not change between conditions. This assumption may not be always true if allosteric effects are condition-dependent but helps to simplify the complexity of the model. Given a particular assignment for the k^{th} region's $\sigma^{\alpha}_{k,i}$, the Hamiltonian (or total energy) of the region is

$$H^{\alpha}_{k} = \sum_{i} h^{\alpha}_{k,i}\sigma^{\alpha}_{k,i} + \sum_{i}\sum_{j\neq i} J_{i,j}\sigma^{\alpha}_{k,i}\sigma^{\alpha}_{k,j} \tag{3.1}$$

Given the above Hamiltonian, one could calculate the equilibrium occupancies if one knew the energetic parameters $\{h^{\alpha}_{k,i}\}$ and $\{J_{ij}\}$. However, usually the site and interaction energies are not known, and only measurements of the occupancies, $m^{\alpha}_{k,i} = \langle\sigma^{\alpha}_{k,i}\rangle$ are given. (Note, given our definition of $\sigma^{\alpha}_{k,i}$, $m^{\alpha}_{k,i}$ is equivalent to the equilibrium probability of factor i being bound on region k in condition α).

To find the $\{h^{\alpha}_{k,i}\}$ and $\{J_{ij}\}$ given only measurements of $m^{\alpha}_{k,i}$, we write the mean-field approximation where

$$\sigma^{\alpha}_{k,i}\sigma^{\alpha}_{k,j} = (m^{\alpha}_{k,i} + \sigma^{\alpha}_{k,i} - m^{\alpha}_{k,i})(m^{\alpha}_{k,j} + \sigma^{\alpha}_{k,j} - m^{\alpha}_{k,j}) \tag{3.2}$$

and assuming mean field, we drop the higher order term of $(\sigma^{\alpha}_{k,i} - m^{\alpha}_{k,i})(\sigma^{\alpha}_{k,j} - m^{\alpha}_{k,j})$ and rewrite the Hamiltonian in the mean-field representation,

$$H^{\alpha}_{MF,k} = \sum_{i} h^{\alpha}_{k,i}\sigma^{\alpha}_{k,i} + \sum_{i}\sum_{j} J_{i,j}m^{\alpha}_{k,i}\sigma^{\alpha}_{k,j} - \sum_{i}\sum_{j\neq i} J_{i,j}m^{\alpha}_{k,i}m^{\alpha}_{k,j} \tag{3.3}$$

From there, one can find the partition function of the mean-field representation by taking the trace of the Boltzmann-weighted Hamiltonian. However, since we assumed that the binding of a factor to a region is binary, we can rewrite the partition function as a product of discrete states,

$$Z_{MF} = \Pi_i^N \sum_{\sigma_i=0}^{1} e^{-\beta H^{\alpha}_{MF,k}} \tag{3.4}$$

where N represents the total number of factors in the system. Expanding out the discrete states and simplifying, one can rewrite the mean-field partition function as,

$$Z_{MF} = e^{\beta \sum_{j\neq i} J_{ij} m^{\alpha}_{k,i} m^{\alpha}_{k,j}} \Pi_{i=1}^{N} [1 + e^{-\beta(h^{\alpha}_{k,i} + \sum_{j\neq i} J_{ij} m^{\alpha}_{k,j})}]$$

To simplify, one can rewrite the discrete states in the form of the hyperbolic cosine,

$$Z_{MF} = e^{\beta \sum_{j\neq i} J_{ij} m^{\alpha}_{k,i} m^{\alpha}_{k,j}} \Pi_{i=1}^{N} e^{-\frac{\beta}{2}(h^{\alpha}_{k,i} + \sum_{j} J_{ij} m^{\alpha}_{k,j})} 2\cosh[\frac{\beta}{2}(h^{\alpha}_{k,i} + \sum_{j\neq i} J_{ij} m^{\alpha}_{k,j})].$$

To make progress, one needs the mean-field representation of free energy and the relationship between the mean occupancy and the system's free energy,

$$F_{MF} = -k_B T \ln Z_{MF} \tag{3.5}$$

$$m^{\alpha}_{k,i} = \partial F / \partial h^{\alpha}_{k,i}$$

One could then solve for the equilibrium occupancy, $m^{\alpha}_{k,i}$ for factor i on region k in condition α,

$$m^{\alpha}_{k,i} = \frac{1}{2} - \frac{1}{2}\tanh[\frac{\beta}{2}(h^{\alpha}_{k,i} + \sum_{j\neq i} J_{i,j} m^{\alpha}_{k,j})] \tag{3.6}$$

where $\beta = 1/k_B T$ which we will set to unity. This effectively maps the regulation problem between Pol-II and the binding of transcription factors onto a fully connected Ising model. Eq. 3.6 allows us to directly relate the mean occupancies of TFs to the binding of Pol-II without the need to enumerate through all possible states. This meant that the lack of scalability which we discussed in Chapter 1 as a major drawback of biophysical models is not an issue for the mean-field approach. Additional TFs will increase the dimensionality of the condition-independent interaction matrix, **J** but can still be related to Pol-II binding using Eq. 3.6.

To simplify regression, Eq. 3.6 can be further linearized around zero into a linear system involving the $\{h^{\alpha}_{k,i}\}$ and $\{J_{i,j}\}$ and a rescaled version of m, that we label Y that takes on values between -1 and 1. The linearized system is,

$$Y^{\alpha}_{k,i} = 2(\frac{1}{2} - m^{\alpha}_{k,i}) \approx [\frac{1}{2}(h^{\alpha}_{k,i} + \sum_{j\neq i} J_{i,j} m^{\alpha}_{k,j})] \tag{3.7}$$

Given a set of measurements of the equilibrium occupancy in a given condition (or conditions), one can in principle perform least-squares fitting to Eq. 3.7 to find the condition-

independent interaction matrix, **J** and the potentially condition-dependent site energies $\{h^{\alpha}_{k,i}\}$. However, due to each region having its own set of site energies, this adds an additional dimension for each sample/region and makes the entire system under-determined.

To make progress, we classify regions based on their transcriptional output (e.g. Inducible and Non-Inducible) and then assign each factor to have condition- and class-specific but not region-dependent site energies, $\{\bar{h}^{\alpha}_{i,c}\}$. This grouping of regions into classes will depend on the nature of the transcriptional data being analyzed. The data we are working with has transcription measured in two conditions, and we define the particular class-based either on the condition (i.e. the class is the condition) or on a region's change in transcription across conditions (i.e. increase in transcription, decrease in transcription or no-change). This allows us to look into the condition/class dependence of each factor and learn the conditional/class average occupancy of each factor and transcriptional expression. One can imagine this average site energy to represent the interactions of unknown factors and other epigenetic changes that were not included in the model which leads to an increase or decrease in factor occupancy when treated with hormone.

The model is implemented using the `sklearn` package in Python and fitted using Ridge regression. LASSO and Elastic-Net regularization were also tested for mean squared error (MSE) in Pol-II regression but they were out-performed by Ridge regression. The final hyperparameter, λ (see Eq.1.16), was chosen based on on minimizing the MSE on the validation set using 5-fold cross-validation. This found $\lambda = 0.1$ to be the best hyperparameter. Finally, the resulting interaction model is an average of the five models generated during cross-validation.

Clustering and sorting of the full interaction matrix were done using the Ward distance method [82]. A correlation matrix was calculated based on each factor's interactions before being sorted by hierarchical clustering. The clustering algorithm is done using the `Scipy` package within python.

3.3.2 Calculation of equilibrium occupancies from sequencing data

Modern high-throughput techniques provide measurements of how often a binding site was occupied. Although cells are by nature out-of-equilibrium systems, we will assume that the measured occupancies represent a steady-state value which we will take to be equilibrium. As such, we should expect that the measured counts of the i^{th} factor on region k in condition α $n^{\alpha}_{k,i}$ should be proportional to the equilibrium occupancy $m^{\alpha}_{k,i}$ defined in the section above.

To transform the observed $n^{\alpha}_{k,i}$ from a ChIP-seq experiment into $m^{\alpha}_{k,i}$, we apply the following transformation (based on a two-state system), that takes an observed count and maps it onto the interval, (0,1). The two-state transformation is

$$m^{\alpha}_{k,i} = \frac{n^{\alpha}_{k,i}}{n^{\alpha}_{0,i} + n^{\alpha}_{k,i}}. \tag{3.8}$$

This equation transforms the observed counts relative to a background count level, $n^{\alpha}_{0,i}$. To determine this background count level we applied SES normalization that is well suited to separate out non-specific from specific binding [48]. In this approach, the background count level is defined to be the count at which the normalized cumulative distribution of the observed and control input counts are maximally different (see section 2.3.3). However, unlike the normalization in chapter 2, to have a stricter background count that can differentiate between strongly and weakly bound TFs for a given factor, we pool together the ChIP-seq data across all 750bp regions and both conditions as the control input. This strict approach allows for conservation of scores across conditions (see Fig. 3.2) and for more separation in scores as a strongly bound TF will be scored highly while a weakly bound TF or not bound TF will be scored little to none. Additionally, a selection of 40 random 10MB regions was chosen as the random background sequences to normalize between DHT and EtOH conditions for a given factor. The SES background count, $n^{\alpha}_{0,i}$ was then calculated for each factor i and Eq. 3.8 was used to convert the ChIP-seq signal to m-values.

The binding data was collected from 35 different ChIP-seq signal profiles across both DHT/R1881 (with hormone) and EtOH (no hormone) conditions on LNCaP cell lines [83], while the enhancer activity data was collected using the STARR-Seq method [84]. For each ARBS region, we extracted the ChIP-seq signal scores over the 750 bp in a 50 bp resolution surrounding a central AR peak. The mean occupancy over the entire region was taken to be $m^{\alpha}_{k,i}$ for factor i on region k.

3.3.3 Predicting occupancies due to mutations

Given a set of interaction energies $J_{i,j}$ and a set of site energies $h^{\alpha}_{c,i}$ for a class of regions in a given condition α, we solve the linearized mean-field equations (Eq. 3.7) for the equilibrium occupancies $m^{\alpha}_{c,i}$. This allows us to use the mean-field model to make predictions when one alters the site energy of a factor, or possibly interaction energies between factors. A factor can be 'knocked out' by simply removing the factor of interest from the model. This effectively mirrors loss-of-function mutations. Each factor is sequentially left out of the model and Eq. 3.7 is solved for the new equilibrium occupancies (including Pol-II) in each mutant condition.

3.4 Results

3.4.1 Fitting a biophysical model of transcription to STARR-seq data reveals activators and repressors of AR-mediated regulation.

Here we present the results of fitting our biophysical model of transcription (see Methods) to recent STARR-seq measurements of sequence regions bound by the transcription factor Androgen Receptor (AR) which is implicated in the growth and maintenance of prostate

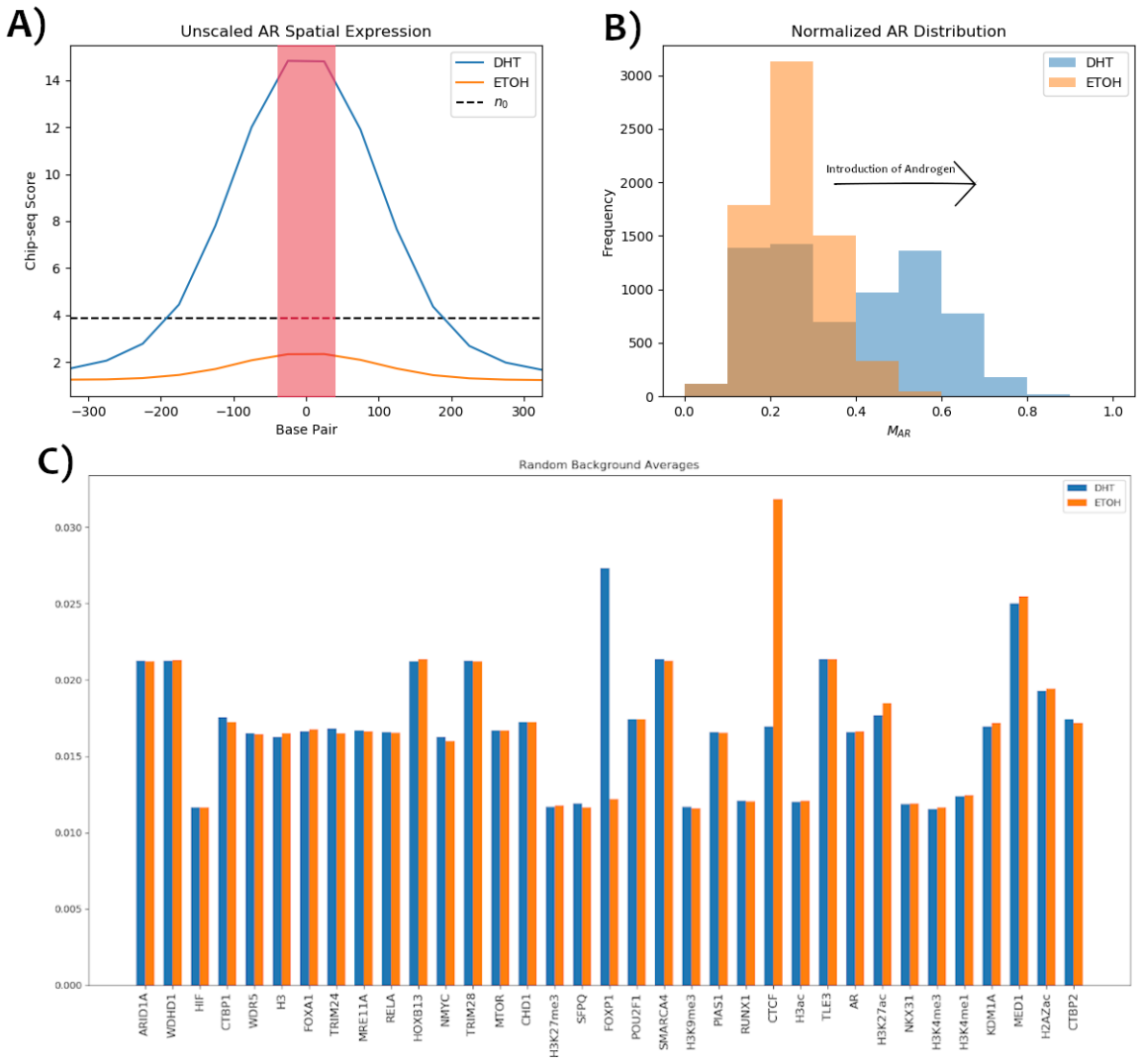

Figure 3.2: (A) Average AR ChIP-seq scores across 6922 ARBS regions as a function of position with respect to the center of a region (x = 0) reveals a central spike (shown in red) when treated with DHT. n_0 (dashed line) determined using the SES method and pooling the surrounding regions across both conditions corresponds to half occupancy or a score of 0.5. (B) Normalized AR m-value Distribution shows a shift towards higher expression with the introduction of Androgen. (C) Random regions of 10 mega-bases were selected from each ChIP-Seq dataset. These scores were averaged to form the background expression which is divided out to normalize the ChIP-seq scores across conditions.

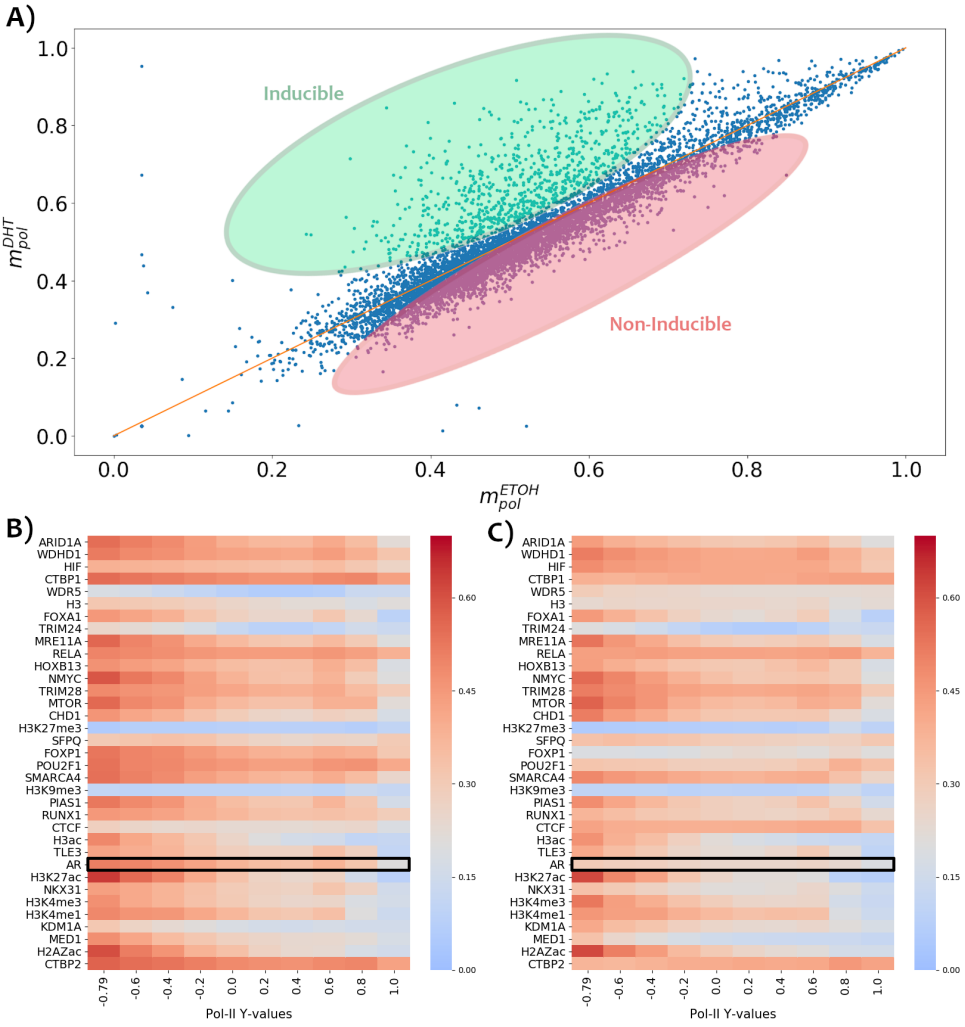

Figure 3.3: (A) Measured Pol-II occupancy of 6922 ARBS regions across the two conditions: hormone (DHT) versus no-hormone (EtOH). Inducible regions (highlighted in green) have increased transcription in the presence of hormone while Non-Inducible regions (highlighted in red) have slightly lower or no change in expression. (B,C) Average regional occupancies (m-values) of various bound factors in (B) DHT (with hormone) and (C) EtOH (no hormone) conditions versus measured transcriptional output (Pol-II Y-value). The 6922 ARBS sequences were sorted by their Pol II Y-value as measured by STARR-seq in both DHT (hormone) and EtOH (no hormone) conditions and were then divided into 10 non-overlapping sequence groups. Groups with more negative Pol-II Y-values have higher Pol-II occupancy and vice versa. The average occupancy over each sequence group, $m_{c,i}^{\alpha}$, was then calculated for each factor, i, in each condition, α =DHT or EtOH. The heat maps clearly show AR (outlined in black), a known key factor in prostate cancer is associated with the most negative Pol-II Y-value groups treated with androgen, but has limited activity when no hormone is present. (B) and (C) also show that some factors show significant changes in occupancy across treatments while others show little to no difference.

cancer. AR is a DNA binding factor and interacts with other TFs and co-factors to regulate the binding of Pol-II (see Fig. 3.1).

A total of 6922 unique AR binding site (ARBS) regions averaging 700 bp in length that showed AR binding in prostate cancer tissue from clinical samples or contained the ARE motif were selected for the testing of their enhancer activity using the STARR-seq assay in two conditions: i) with hormone (DHT condition) or ii) without hormone (EtOH condition). The measured transcriptional output across the two conditions is plotted in Figure 3.3A. As can be seen, many ARBS regions showed little to no differential activity, but some showed significant changes in expression and were defined to be Inducible ARBS regions. Although the binding of AR to these regions is key, which factors also contribute to differential expression is still largely a mystery.

Using a mean-field approach, the biophysical model shown in Fig. 3.1 can be solved resulting in the equilibrium occupancy of each bound factor written in terms of the equilibrium occupancies of all the other factors. We calculate the equilibrium occupancies from experimental measurements of chromatin binding. For Pol-II we take the above transcriptional activity as measured by STARR-seq for the k^{th} ARBS region in condition α=DHT or EtOH, to be proportional to its equilibrium occupancy $m^{\alpha}_{k,\mathrm{Pol-II}}$. For the other DNA-bound factors, we use genome-wide binding data to quantify the equilibrium occupancy, $m^{\alpha}_{k,i}$ of factor i on the k^{th} ARBS region in condition α. Here we have binding data for 35 factors measured in both α =DHT and α =EtOH conditions (see Section 3.3.2 for details). In Fig. 3.3B,C, we show the average occupancy of the various DNA-bound factors in both conditions as a function of the ARBS region's Pol-II occupancy (high occupancy corresponds to negative $Y^{\alpha}_{\mathrm{Pol-II}}$ values, and lower occupancy to positive values). There are clearly several factors besides AR whose occupancy correlate (both positively and negatively) with the Pol-II occupancy across both conditions (e.g. H3k27ac and MTOR). It can also be seen that a factor's overall occupancy often depends on the condition. Clearly, without hormone, AR is not bound and shows considerably less binding to ARBS, but other factors such as RUNX1 and MED1 also show strong condition dependence in their binding to ARBS regions. Given such correlations and dependencies, we expect to be able to discover the relative importance of each factor in mediating AR-regulated transcription.

The parameters of the model are thus the site energies $\{h^{\alpha}_{k,i}\}$ (which depend on the particular ARBS region, k, bound factor, i, and the experimental condition, α=DHT or EtOH) and interaction energies $\{J_{i,j}\}$ between the various factors, which we take to be condition-independent since they represent specific protein-protein interactions.

We first fit the measured Pol-II occupancies to the occupancies of the measured TFs across the two conditions to find the Pol-II site energies and energetic interactions with the other factors. In the STARR-seq assay, the construct is such that every tested genomic region utilizes the same promoter and therefore the same Pol-II binding site. The binding to this site by Pol-II in the absence of any other factors is described in our model by the

site energy parameter, $h^{\alpha}_{\mathrm{Pol-II}}$, which thus has no regional dependence, but in principle could still depend on condition (i.e. with or without hormone). The other fit parameters are the interaction energies, $J_{\mathrm{Pol-II},j}$, between the various TFs and Pol-II. Additionally, STARR-seq is a plasmid-based assay, and as such, the measured enhancer activity should not be affected by the location and orientation of the regions or the endogenous chromatin structure [15]. To model this, the $J_{\mathrm{Pol-II},j}$ between histone marks and Pol-II have been set to zero. However, we do not remove histone marks from the model altogether. As shown in Fig 3.3 and previous research [41], there is a strong correlation between enhancers found using STARR-seq and histone mark binding. This is especially true for H3K27ac whose signal is strongly correlated with Pol-II binding in both DHT and EtOH conditions. Additionally, the inclusion of histone marks allows us to screen the genome for other possible ARBS enhancers where the effects of histone marks would be present.

Using the measured occupancy values for the 6922 ARBS sequences across both conditions, we used 5-fold cross-validation with Ridge regression to fit Eq. 3.7 and determine the above parameters for Pol-II. Using our fitted model, we predicted the Pol-II occupancy of each ARBS in both conditions (see Fig. 3.4A). The Spearman correlation between predictions and the experimentally observed Pol-II occupancy is 0.41, showing that the model was able to relate transcriptional output to TF binding. For comparison, prior work that fit MPRA data to sequence using a CNN resulted in a Spearman correlation of 0.28 [75].

Additionally, the mean-field model was used to predict the transcriptional activity in DHT and EtOH conditions for 20 test regions (10 were predicted to be Inducible and 10 Non-Inducible by our prior machine learning ARBS classifier [84]) that were not in the originally selected 6922 ARBS regions. Shown in Fig. 3.4B are the predicted differential activity of these 20 test regions from the biophysical model versus the measured change across the two conditions. Of the 10 test regions that were predicted to have significant differential expression between the two conditions ($\Delta Y > 0.1$), all show the same behaviour experimentally. The remaining ten test regions that were predicted to have lower/similar transcriptional output when treated with DHT than in EtOH all experimentally demonstrated Non-Inducible behaviour. Our mean-field model also correctly predicts one of the previously misclassified Inducible regions to have similar differential activity as other Non-Inducible regions. In terms of magnitude, the model performs well for the Non-Inducible regions; correctly attributing the lowest change in Y value for the Non-Inducible regions. The model struggles more with the Inducible regions where it was unable to separate the ARBS region with the highest measured differential expression from other Inducible regions.

The resulting fit parameters for Pol-II are shown in Fig. 3.5. The interaction energies, $J_{\mathrm{Pol-II},j}$, between Pol-II and the 35 other factors are shown in blue. Strongly positive interaction energies signify a repressive relationship with Pol-II binding whereas those that are very negative correspond to activating interactions with Pol-II. Since our approach directly models the repressive/activating behaviour of each TF, known transcriptional repressors

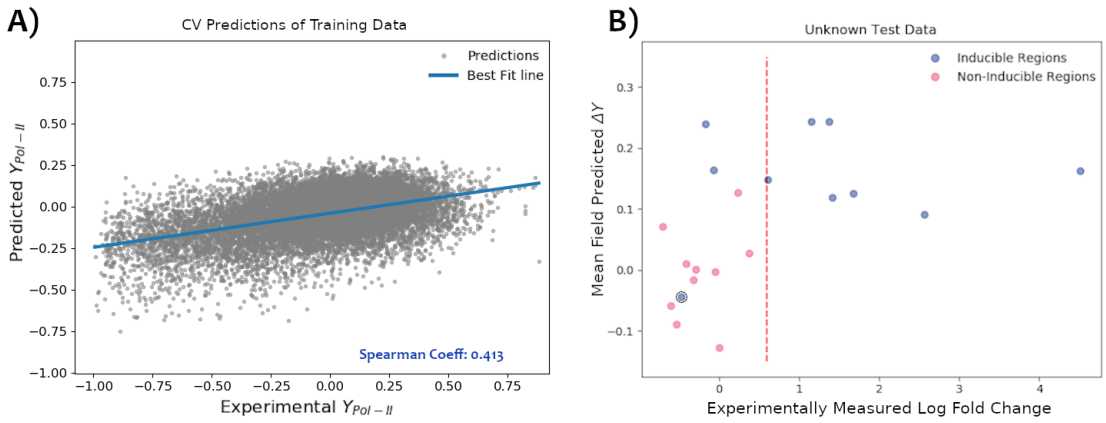

Figure 3.4: Cross-validated predictions of training and test set reveals a strong positive correlation between predicted and measured transcriptional output. (A) 6922 ARBS regions measured for their transcriptional activity by STARR-seq in both DHT (with hormone) and EtOH (no hormone) were fit to our biophysical model of transcription (see Methods). The training data was then re-predicted using a model constructed from the average parameters from the 5-fold fits. Shown are the predicted Y values versus the measured Y values for Pol-II. Predictions of the training data are well correlated with the expected values with a Spearman correlation value of 0.41. (B) Predicted differential expression of 20 test regions (that were not in the originally selected ARBS regions) versus their measured differential expression. A previous classifier of ours classified regions as Inducible (blue) vs Non-Inducible (gray) (the Inducible group was defined to have a log fold-change > 0.7 shown as red dashed line) and managed to achieve an accuracy of 60% on the Inducible group. Our mean-field model correctly identifies a previously misclassified Inducible region (outlined in black).

should correspond to high positive interaction energies. Three of the seven factors with the most repressive interaction energies are either well-known transcriptional repressors or have been implicated with some form of transcriptional inhibition in previous work. For instance, transcriptional co-repressors, TRIM28 and TLE3 [85, 86] have a high positive interaction energy from the model. Similarly, TRIM24, a known activator in prostate cancer, is predicted to have a negative interaction energy with Pol-II [87]. Meanwhile, HOXB13, the factor with the highest Pol-II interaction energy, is not a well known transcriptional inhibitor or promoter. There are mixed reports on HOXB13's role on AR-dependent behaviour and prostate cancer: while some suggest it is a suppressor of tumour growth, others suggest it acts as a critical pioneer factor in AR-mediated transcription [29, 81]. Our model suggests that in terms of overall transcriptional output, HOXB13 has a generally inhibitive relationship. However, HOXB13 can still be a crucial factor for AR-dependent behaviour that recruits the binding of subsequent transcription-dependent factors.

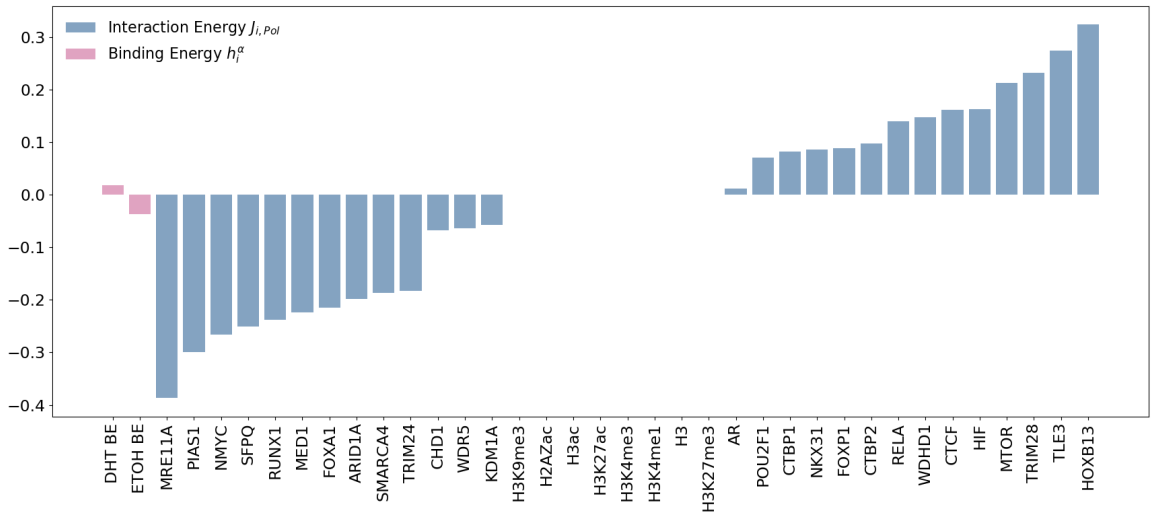

Figure 3.5: Fitted Pol-II interaction energies, $J_{\mathrm{Pol-II},j}$ (Blue), and site energies in the two conditions, $h^{\alpha}_{\mathrm{Pol-II}}$ (Pink). A negative interaction energy signifies an attractive effect on Pol-II binding, whereas positive energy corresponds to an inhibitory effect on binding. The fitted model correctly identifies known transcriptional co-repressors such as TLE3 and TRIM28. Since measurements of transcriptional activity were done in plasmids, histone marks cannot directly affect the binding of Pol-II and are set to have zero interaction energy with Pol-II.

The Pol-II site energy $h^{\alpha}_{\mathrm{Pol-II}}$, which we take to be condition-dependent, serves as a measure both of the Pol-II binding energy to its site in the construct as well as its chemical potential in the particular condition, α. We find that the site energy is slightly more negative in the EtOH condition. This suggests that in the absence of other factors, there would be a tendency for higher expression (i.e. more Pol-II binding) in the EtOH condition. This could

reflect experimental changes during the treatment of DHT or the condition dependence of factors potentially missing from the fitted model.

3.4.2 Determining interactions between all factors and identifying AR-mediated complexes

Next, we sought to quantify all the interaction energies $J_{i,j}$ between the other 35 DNA bound transcription factors in the model. Similar to what was found for Pol-II, the binding of other TFs can either stabilize or inhibit the binding of other TFs and co-factors, creating a network of interactions, as shown in Figure 3.1. In principle, we could use the measured occupancies on the ARBS regions in both conditions and sequentially fit Eq. 3.7 for each factor to find its condition and region-dependent site energies and interaction energies with the other factors. However, unlike Pol-II which has the same binding site in every construct tested in the STARR-seq assay, the same is not true for the other transcription factors. Each ARBS may have its own unique set of binding sites for the 35 measured TFs (including histone marks) that is different from every other region. This makes the model under-determined as there are potentially ~ 20000 possible site energies, $h^{\alpha}_{k,i}$ with only 6922 ARBS regions.

To make progress, we instead sought to fit a model to describe the global average occupancy of ARBS regions in each condition. Thus each factor, i, now just has a single site energy, h^{α}_{i} that sets its overall occupancy in a given condition. The interaction energy matrix, $J_{i,j}$ is as before, independent of the experimental condition and does not depend on regions. Despite the simplification of the site energies, the interaction energy matrix will still capture cooperative or inhibitive interactions that are present in the binding data for the various factors.

We fit Eq. 3.7 using the above simplification, and Fig. 3.6 shows the full fitted interaction matrix, $J_{i,j}$ between all factors. Each row represents the fitted interaction energies from a regression on the factor of interest using the remaining factors. Those with a cooperative binding relationship have a negative interaction energy (blue) in the interaction matrix, while an inhibitory relationship (red) will increase the total energy of the system and thus lower the probability of both being bound. What is also striking in the interaction matrix are groups of factors that share cooperative interactions forming putative complexes. TFs that come together to form protein complexes will have shared interactions with other TFs. To find possible complexes, we calculated the correlation matrix between factors by correlating the interaction energies of one factor to those of another. In order to cluster similar factors, we then performed hierarchical clustering on the correlation matrix. In Fig. 3.7, the resulting dendrogram of the clustering shows three to four major groups of factors that share common interactions among them. In the clustering, we see known prostate cancer transcription factors such as FOXA1, HOXB13, and NKX3 clustered together [30, 88, 89].

Despite being closely clustered within the same group, individual factors can still behave differently in the presence of specific factors. For instance, consider TLE3, ARID1A, and

PIAS1 which are all closely grouped within the same cluster and share strong cooperative interactions with each other (blue). However, they show very different interactions in the presence of the histone mark H3K27ac. Both ARID1A and PIAS1 share a negative interaction energy with H3K27ac, while TLE3 has a positive interaction energy which implies an inhibitive relationship with H3k27ac. This results in a fairly low correlation coefficient of 0.1 across ARID1A, TLE3, and PIAS1 despite strong negative interaction energy between the three factors. In terms of regulatory logic, this cluster highlights how interactions between factors can be dependent on the epigenetic state of the genomic region. The same behaviour can be seen in the interactions with other important factors such as AR and MED1.

Interestingly, the clustering algorithm predicts AR to be closely related to MED1. Both AR and MED1 share a negative interaction energy with each other as well as a high correlation coefficient of 0.6. The high correlation coefficient for interaction energies with other factor and negative pairwise interaction energy between the two factors implies a strong similarity between their interactions with other factors and suggests that MED1 is strongly associated with AR when regulating transcription. Furthermore, in Figure 3.3B,C, MED1 also shows a strong correlation with increased transcription in both DHT and EtOH conditions, suggesting it plays an important role in mediating transcription even in the absence of hormone. This is significant as androgen-deprived growth is the main driver of late-stage lethal castration-resistant prostate cancer.

The fitted model also quantifies the condition-specific site energies for each factor and histone mark. As stated before, the site energy includes the chemical potential for each factor which takes into account any changes in concentration that may change across conditions. As expected, AR shows a lower site energy in the DHT condition compared to EtOH, consistent with the concentration of active AR being higher in the presence of hormone. Other factors such as POUF1 and MED1 share similar behaviour to AR in their condition-dependent site energies. But factors like WDR5 and KDM1A demonstrate the opposite behaviour, having a higher site energy in the DHT condition, suggesting a tendency to be less bound overall in that condition. These changes in site energy across condition prime the differences in factor occupancy that ultimately lead to the differential activity. However, they are not sufficient to predict which factors are most important for enhancer activity as the complex network of interactions ultimately plays a significant role. For example, since AR has a lower site energy when treated with DHT, the increase in AR occupancy can lead to an increased probability of binding for factors that positively interact with AR. This can lead to a cascading effect where more factors are recruited and ultimately change Pol-II binding beyond that set by site energies alone.

3.4.3 Predicting changes in factor occupancies due to mutations

Given a set of site energies $h^{\alpha}_{k,i}$ and an interaction matrix, Eq. 3.7 can (in principle) be solved for the equilibrium occupancies $m^{\alpha}_{k,i}$ of each factor on a given region and condition.

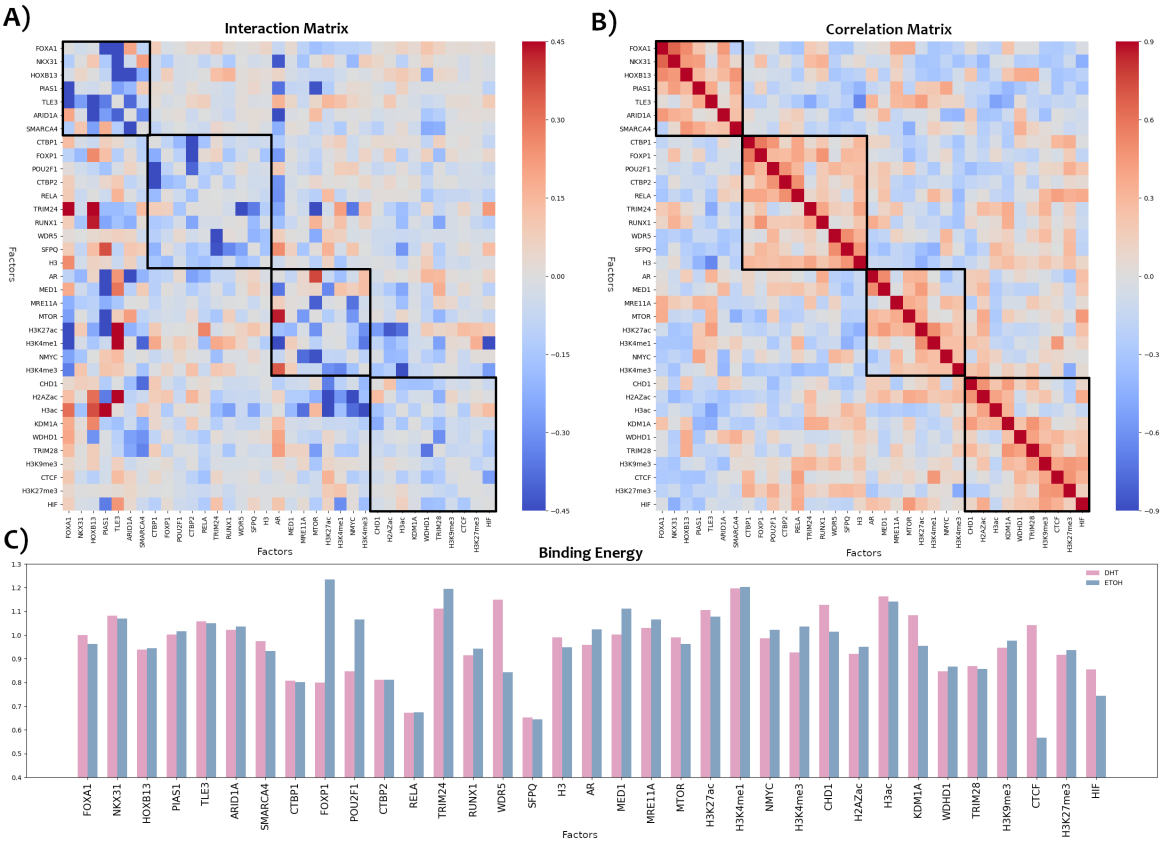

Figure 3.6: (A) The interaction energy matrix, $J_{i,j}$, between every factor as determined by fitting Eq. 3.7 using the experimentally determined equilibrium occupancies, $\{m^{\alpha}_{k,i}\}$ for all 6922 ARBS regions and no regional dependence for the site energies. Energies coloured blue correspond to negative interaction energies and therefore a cooperative/attractive effect on binding between the given factors. Red corresponds to a positive interaction energy that inhibits binding between the given factors. Each factor is fit separately where each row represents the fitted interaction energies for each regression. As such, the interaction matrix is not symmetric. (B) Hierarchical clustering of the correlation(at 50 bp resolution) matrix (see Fig. 3.7) shows four clusters of factors (outlined in black) which represent possible complexes. (C) Fitted condition-dependent site energies h^{α}_{i} in either DHT (pink) or EtOH (blue) conditions. Lower site energy favours an increase in factor occupancy whereas higher site energy corresponds to the opposite.

In section 3.4.2, we made the simplifying assumption of removing the regional dependence of the site energies to examine the broader condition dependence, and more importantly, learn the interaction energies between factors. Nevertheless, we can use the fitted h^{α}_{i} and $J_{i,j}$ to predict the average occupancies of each factor in the two conditions, α=DHT or EtOH, learning the generic changes to binding that may occur. Fig 3.8 shows that the resulting solution of Eq. 3.7 for the average occupancies corresponds exactly to the average occupancy of each factor as determined from the measured binding data in both conditions. The high correspondence to the average state in each condition suggests that our model

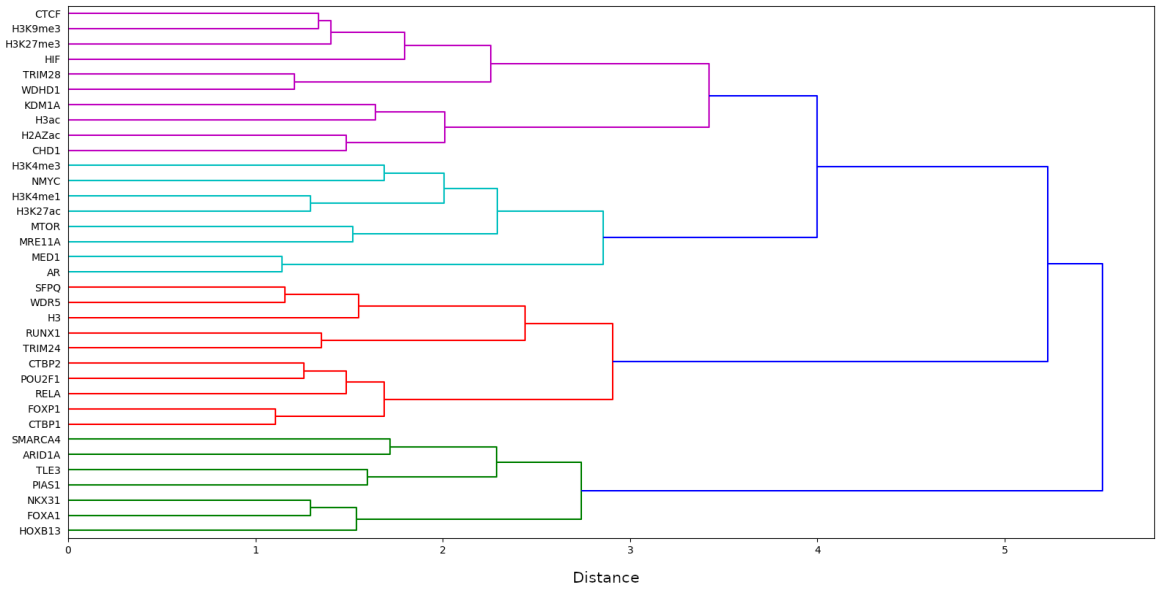

Figure 3.7: Hierarchical clustering of the correlation matrix made from the interaction energy matrix $J_{i,j}$ shows three to four major interacting groups. Clustering was done using the Ward linkage function.

has a self-consistent energy landscape such that the minimum coincides with the actual expected values of each factor.

Given that we can predict factor occupancies, we now try to quantify the importance of each factor to AR-dependent enhancer activity resulting from sequence mutations or knock-downs of factors. To do this, we will systematically remove each factor from the system and calculate the resulting change to the average occupancies and ultimately transcriptional activity (see section 3.3.3). The previous model was fit to all ARBS regions, while most showed little to no differential enhancer activity when treated with hormone (see Fig. 3.3A). Indeed using all 6922 ARBS regions to look at factor occupancy is dominated by these non-functional regions, and the predicted and measured change in transcription across conditions is minimal (see Fig. 3.8). Instead, we now relax our simplification, and group regions based on their fold change in Pol-II occupancy such that we have regional classes that are representative of an i) Inducible, or as a control, ii) Non-Inducible ARBS region. So instead of using all 6922 ARBS regions, we choose to select the top 1000 and bottom 1000 ARBS regions sorted by their measured fold change across conditions. This now introduces condition- and class-specific site energies $\{h^{\alpha}_{c,i}\}$ into the model, where the c represents either i) top 1000 (Inducible) or ii) bottom 1000 (Non-Inducible) ARBS regions.

We fit this less constrained model using the data for the two groups of ARBS regions. Using this fitted model, the predicted equilibrium occupancy state representing the Top 1000 ARBS is predicted to have higher self-transcription output in the presence of DHT

compared EtOH, consistent with the experimental observations of these ARBS (Fig 3.10). Similarly, the predicted state of the Bottom 1000 ARBS behaves exactly the opposite, having slightly decreased expression in the presence of DHT compared to EtOH. With regards to the energetic parameters, the interaction matrix $J_{i,j}$ (see Fig. 3.9) was similar to the one fit using all the ARBS regions with a Spearman correlation of 0.92 when comparing the two interaction matrices (see Fig. 3.6.) In regards to the site energies, there is still a tendency for increased AR occupancy in the presence of DHT even at Non-Inducible regions. This change in expression in the presence of DHT is magnified in the Inducible regions along with an increased expression for crucial factors such as MED1 and RUNX1.

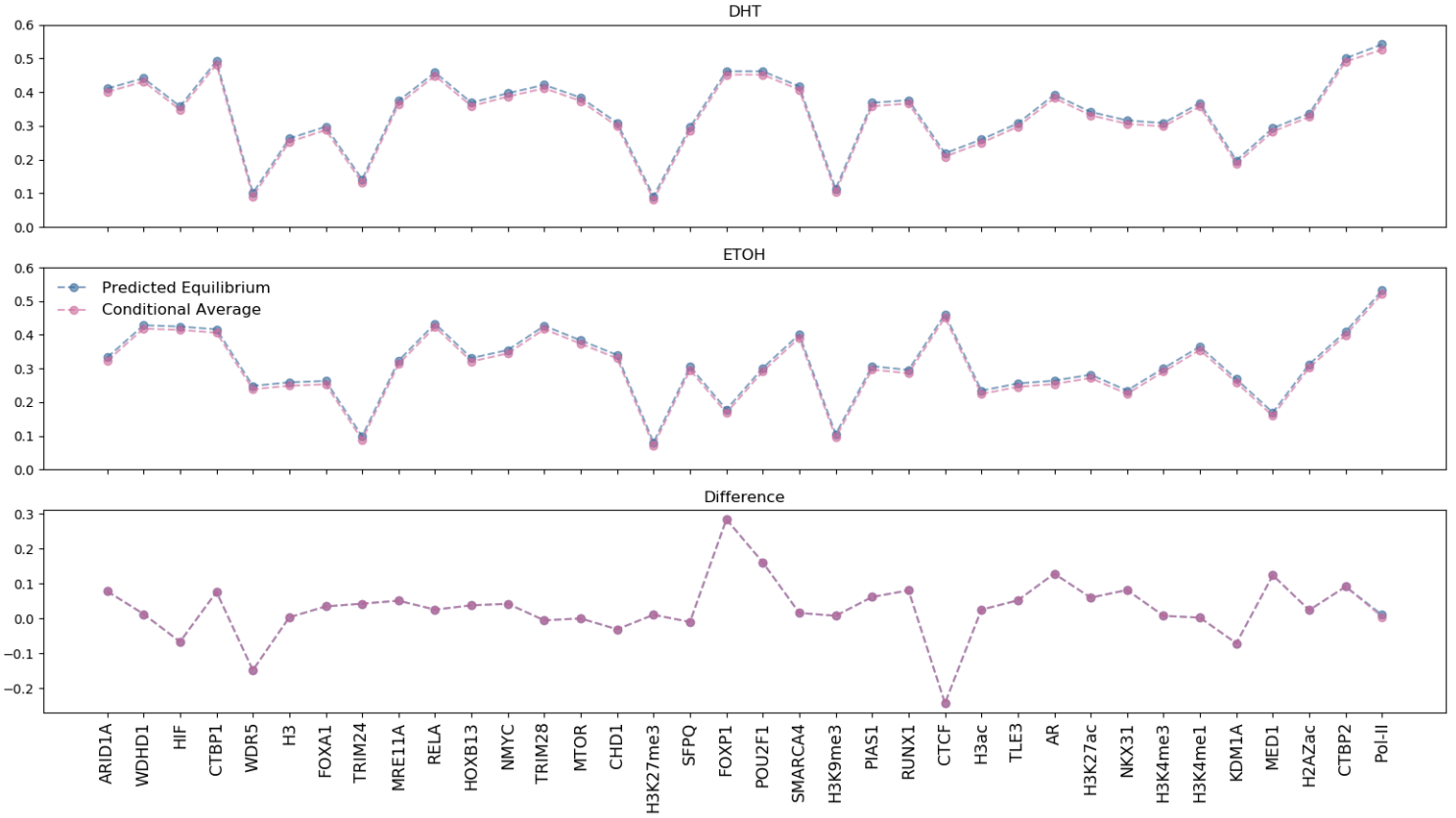

Figure 3.8: Solution for the average occupancies, m_i^α from the model fit to all 6922 ARBS regions corresponds to the measured average occupancies of the bound factors in the two conditions, α = DHT (A), or EtOH (B). (C) The difference in average occupancies of bound factors between the two conditions across the 6922 ARBS sequences for both the measured and predicted from the model. A small offset has been added to the solution for clarity.

Using this model, we can carry out mutations by removing a factor from the system and observe the effects of the mutated system on enhancer activity in the presence and absence of hormone. The loss of key factors to transcription should lead to a significant change in Pol-II binding across conditions. To knock a factor out, we sequentially leave each factor out of the model (i.e. set $m_i^\alpha = 0$), and solve Eq. 3.7 for the new equilibrium occupancies of all the other factors including Pol-II. Due to our self-consistent approach, key factors can potentially alter Pol-II binding directly or indirectly, through their interactions with other important regulators. Loss of factors that do not directly interact with Pol-II may lead to significant changes in the binding of other crucial factors that do interact with Pol-II. Mutations that only result in drops of self-transcription output (i.e. Pol-II occupancy) in the presence of DHT but not in the presence of EtOH are treated as crucial factors for Inducible ARBS behaviour. To quantify changes in Pol-II occupancy across conditions due to the mutations, we use a metric motivated by a t-statistic,

$$t_{c,\mathrm{Pol-II}} = \frac{(m_{c,\mathrm{Pol-II}}^{\mathrm{DHT}} - m_{c,\mathrm{Pol-II}}^{\mathrm{EtOH}})}{\sqrt{\frac{1}{2}(m_{c,\mathrm{Pol-II}}^{\mathrm{DHT}} + m_{c,\mathrm{Pol-II}}^{\mathrm{EtOH}})}} \tag{3.9}$$

where we have assumed the resulting output $\propto m$ would follow a Poisson distribution and $Var(m) = m$

The results of the mutation analysis are shown in Fig. 3.10. For reference, the predicted statistic measuring the change in transcription of the unmutated (wildtype, WT) model for the top 1000 ARBS (red line) class is positive, while for the bottom 1000 ARBS class it is negative (green line). Mutations that lead to changes in Pol-II occupancy below or above the mean value of the other group are considered significant. For the Inducible class (see Fig. 3.10A), mutations to crucial factors for hormone-dependent enhancer activity such as AR and RUNX1 are predicted to significantly lower the statistic of Pol-II expression relative to WT. HOXB13 which in earlier analysis demonstrated a negative relationship with Pol-II binding is predicted to slightly drop the enhancer activity of Inducible regions. This shows an important distinction between potential inhibitors of transcription versus inhibitors of hormone-dependent enhancer activity. Similarly, the prior clustering analysis identified MED1 as a potential transcriptional activator with similar interactions as AR. Loss of MED1 is predicted to lead to a dramatic drop in differential Pol-II binding similar to that of AR. Of the 35 different factors and histone marks, mutations and loss of function to AR and MED1 results in the most significant drop in predicted Pol-II binding when treated with DHT compared to control.

The same mutational analysis was applied to the site energies representing the bottom 1000 ARBS region class (see Fig. 3.10B). Loss of important factors for the top 1000 ARBS class had little to no effect on the predicted Pol-II binding. However, the loss of some factors such as WDR5 and KDM1A is predicted to cause this class of Non-Inducible ARBS to show

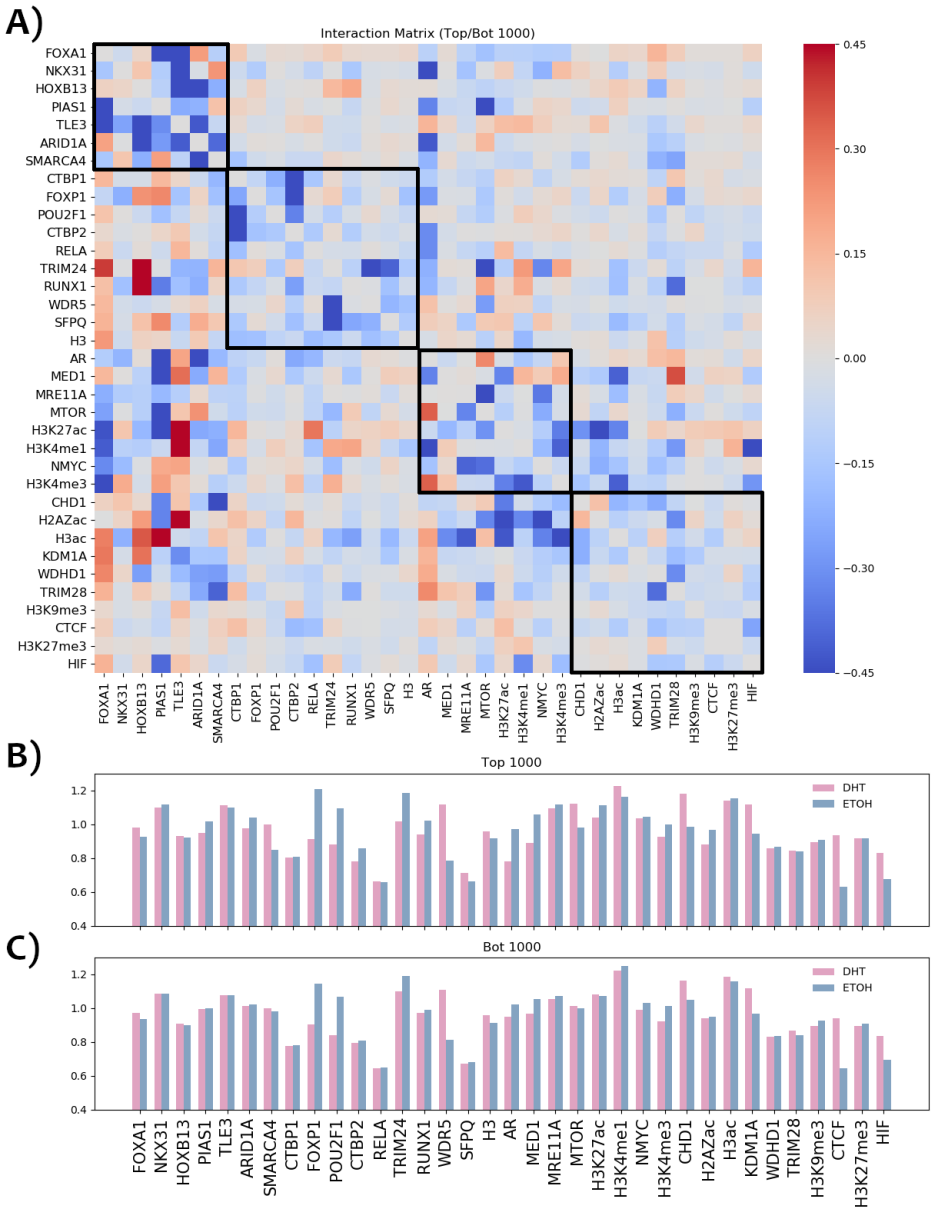

Figure 3.9: (A) Fitted interaction matrix of factors using the top 1000 and bottom 1000 ARBS as sorted by fold change. The interaction matrix is ordered and grouped as in Fig. 3.7. Cooperative pairwise binding relationships are shown in blue while uncooperative/inhibitory relationships are shown in red. Heatmap is sorted using the same distance matrix from hierarchical clustering as before (see Fig. 3.7). The same four dominant clusters are outlined in black. There is good correspondence between the Top/Bot 1000 Interaction and the old interaction matrix with a Spearman correlation of 0.92. (B,C) The binding energies of the Top 1000 and Bot 1000 ARBS regions in both DHT and EtOH conditions. As expected, crucial TFs such as AR and MED1 in the top 1000 ARBS regions all show significantly increased expression in the presence of DHT compared to the bot 1000 ARBS regions.

some DHT-dependent enhancer activity, but would still be classified as an Non-Inducible region.

Lastly, as a check for consistency, instead of fitting Eq. 3.7 directly to the measured occupancies, we instead took the difference of Eq. 3.7 for Pol-II across both conditions, resulting in a model that depends now on the change in occupancy $\Delta m_{c,i} = m_{c,i}^{\mathrm{DHT}} - m_{c,i}^{\mathrm{EtOH}}$. We fit this model using the differential transcriptional output and differential occupancies of the other factors. This fitted model is able to predict the differential expression given the measured difference in occupancies for the various factors (see Fig. 3.11A). We find that the same key factors from the mutagenesis analysis above are also the strongest contributors to differential Pol-II binding. As shown in Fig. 3.11B, PNAS1, AR, ARID1A, and RUNX1 are all critical factors which when mutated are predicted to result in the loss of enhancer activity and whose change in expression contribute the most to the strength of enhancer activity. The two approaches are in agreement, and this provides additional support to the key AR-dependent factors that have been identified in this work.

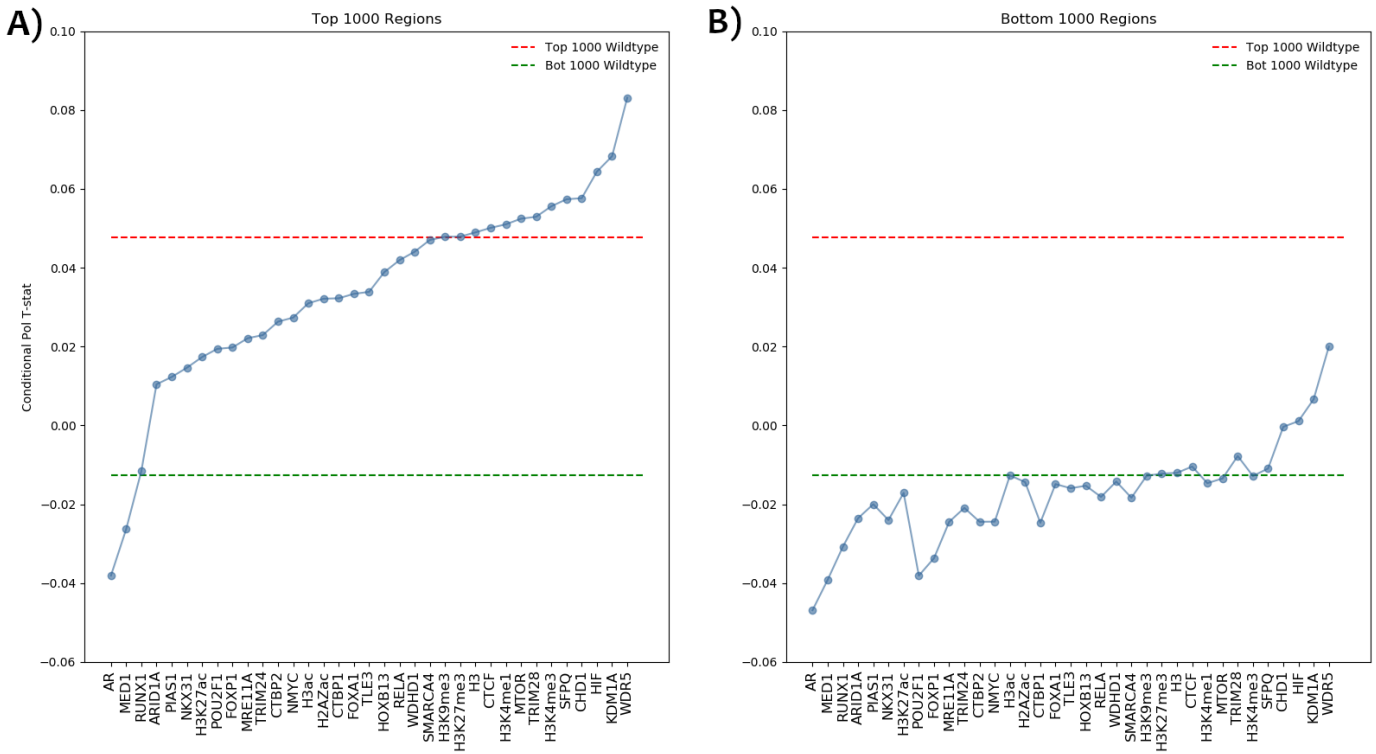

Figure 3.10: Predicted differential Pol-II occupancy between DHT and EtOH conditions when the factor of interest is knocked out. Individual factors were removed from the model in both treated and untreated conditions, and the *mutated* model was then solved for the Pol-II occupancy in both conditions (α = DHT or EtOH), and regional class (c = positive top 1000 fold-change or negative, bottom 1000 fold-change). Also shown are the calculated differential expression of the unmutated/wildtype model for the top 1000 (red dashed line) and bottom 1000 regions (green dashed line). (A) Computed Pol-II statistic, $t_{\text{top,Pol}-\text{II}}$, for the mutated top 1000 ARBS class. Mutations to key factors that are crucial to hormone dependent enhancer activity have a negative statistic. (B) Same as in (A) but now for the bottom 1000 ARBS regions class; factors are sorted in the same order as in (A).

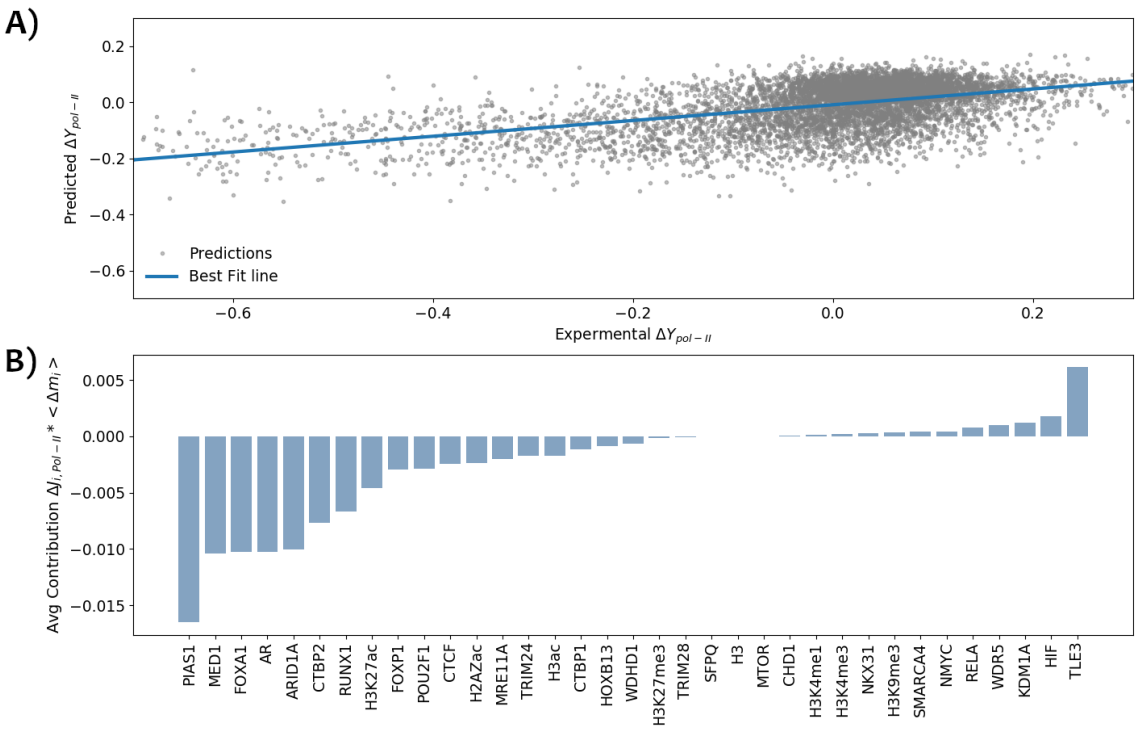

Figure 3.11: (A) Predicted differential expression of ARBS regions between DHT and EtOH conditions versus that measured experimentally. Instead of fitting Eq. 3.7 to the observed occupancy of Pol-II as in Fig. 3.4, the difference in occupancy for Pol-II between DHT and EtOH conditions for the 6922 ARBS regions was fit to the difference in occupancy for the bound factors across conditions. The fitted interaction energies $J_{\mathrm{Pol-II},i}$ now highlight those factors that are essential for differential expression, not overall expression, as for the previous fits. Predicted differential occupancy of Pol-II from the cross-validated model correlates well with the observed differential occupancy with a Spearman coefficient of 0.45. (B) Fitted $J_{\mathrm{Pol-II},i}\langle\Delta m_i\rangle$ for the cross-validated model of differential Pol-II occupancy. Factors such AR, PIAS1, and ARID1A all have large negative changes in interaction energy when treated with DHT and coincide with the crucial factors as predicted by the mutation analysis shown in Fig. 3.10.

3.5 Discussion

3.5.1 Connection to DNNs and CNNs

In recent years, since the introduction of high-throughput assays, there has been an increased emphasis on the development of machine learning models to learn the complex combinatorial transcriptional logic used by a cell. While these models are effective, the high complexity of these models along with the high number of connections (Fig. 4.6) means they often lack the interpretability of biophysical models [19]. This makes teasing out important TFs difficult, as each node within a modern neural net can have hundreds if not thousands of

connections. The lack of interpretability within these models motivated us to develop a hybrid approach that bridges current biophysical modelling while still taking advantage of the high-throughput assays. Our model is still fundamentally a biophysical model as the subsequent transcription output can be related to the equilibrium binding of Pol-II and the total energy of the system. However, the model has enough flexibility that it is able to leverage the high count number of the STARR-seq and ChIP-Seq assays and determine the optimal interaction energies between transcription factors.

Here we draw the connection between our biophysical model and other deep learning methods that have been used on reporter assays. Starting with the sequence data of tested regions, many modern neural nets initiate with convolution layers followed by a section of dense fully connected layers (Section 4.2) before the output [74, 75]. This approach allows the model to determine important DNA motifs from the tested sequences and gives enough flexibility to the model to mine for previously unknown or mutated motifs. Analysis of such filters often involves the use of a sequence alignment algorithm to known motif databases to determine the factor of interest. Rather than starting from the sequence level and discovering important DNA motifs within tested sequences, our biophysical approach starts with the genome-wide binding data for a large number of known DNA-binding factors. This effectively accomplishes the same task as the convolution layers in many neural networks, as the outputs of the convolution filters (assuming that it is representative of a known factor) should correspond to and be correlated with the binding data of the same factor. Subsequently, this greatly decreases the degrees of freedom within our model and thus requires less data to train.

The predicted transcriptional output given the binding data as input had a Spearman correlation of 0.41 with the measured output from STARR-seq. This was an improvement over sequence-based methods used to fit MPRA data that achieved an impressive correlation of 0.28 [75]. The improvement could be due to several sources. Here the measured sequences were for a very specific regulatory program (namely AR) whereas the MPRA modelling tried to fit a much broader spectrum of enhancer activity. The measurement of activity across two conditions along with the incorporation of binding data also provided additional information that is not captured by sequence alone. We believe that further improvement can be achieved by incorporating both sequence and binding information.

Finally, rather than applying a dense linear layer to binding data, we specifically model the interconnected nature of transcription factors. Within our model, all of the TFs are interconnected and will affect each other's binding. By leveraging the large amount of data available, our model is able to tease out the relationships between factors as well as their interactions to Pol-II. Correlated expression between TFs will lead to a negative interaction energy between the two factors. Our approach then allows us to perform theoretical knock-out experiments for each factor which removes a given factor of interest and changes the equilibrium occupancy of the remaining factors. This allows us to differentiate between

non-functioning factors that do not affect enhancer activity and crucial factors that help drive the process (Fig. 3.10). An important distinction here is that while there are quite a few factors that would reduce enhancer behaviour, only three factors (AR, MED1, and RUNX1) would lead to the complete loss of enhancer capabilities. While one can achieve a similar mutational analysis on CNNs and other machine learning approaches, the learned transcriptional grammar is often hidden deep into the network and is limited to the interactions between TFs and Pol-II. Consequently, this means that any mutation to the DNA sequence would not reflect the cooperative binding relationships between TFs. As such, pioneer TFs that help initiate binding for other TFs will not be weighted highly in the this representation. A solution to this problem is to first train CNNs on the binding data of known TFs, which would help identify important binding motifs for both the factor of interest and other pioneer factors [90]. However, as stated before, one can bypass this step by using the binding data as the input for the mean-field approach.

3.5.2 Connection of MED1 and RUNX1 with AR

In prostate cancer, AR-mediated gene dysregulation has been shown to have a significant effect on the growth and maintenance of tumour cells [28]. Here we have analyzed 6000+ AR binding sites (ARBS) found in patient tumours to identify other key AR-dependent regulators, as Inducible enhancers only make up a small fraction of all binding sites [84]. We used the measured transcriptional output and binding data for 35 transcription factors and co-factors for these ARBS regions and fit a biophysical model to determine the energetic interactions that are important for AR-mediated transcription.

Our biophysical model identified a network of interactions between the 35 factors that were necessary not only for transcriptional output but also for differential activity in the presence of hormone. Several crucial factors besides AR that were essential factors for ARBS enhancer activity in the presence of hormone were identified, including RUNX1, ARID1A, and PIAS1. In a previous study, the same factors were shown to be important in classifying ARBS regions with inducible enhancer activity [84].

Additionally, there was a clear separation between androgen-dependent activators and factors that were important for transcriptional activity within our model. For instance, we found known transcriptional co-repressors such as TRIM28 and TLE3 to have a high interaction energy and therefore a negative relationship with Pol-II [85, 86]. Both TRIM28 and TLE3 showcase high occupancy regardless of hormone treatment, signifying a lack of dependence on androgen. Conversely, we found AR, a known androgen-dependent activator, to have effectively zero interaction energy with Pol-II. AR, whose occupancy is heavily dependent on hormone treatment (Fig. 3.2), only shows a strong correlation with transcription activity in the presence of hormone (Fig. 3.3). This suggests that our model is capable of separating between activators/repressors that require androgen to function and constantly active activators/repressors of transcription.

We then quantified the interaction energies between transcription factors as well as each TF's site energy. As expected, important factors that showed correlated occupancy with transcriptional activity such as AR and RUNX1 also showed lower factor-specific site energy in the presence of androgen. To test the newly learned interaction matrix and energy landscape, we reversed the problem and solved for the equilibrium occupancies. The solved equilibrium occupancies perfectly matched the average occupancies of each factor.

Finally, we wanted to test how important each factor was to androgen-dependent behaviour. We fitted a new model using data from two different groups of ARBS regions: regions that showed significant inducible activity and regions that showed no change in activity when treated with hormone. We then mutated each factor by systematically removing it from the model and observing the changes in transcription in both treated and untreated conditions (Fig. 3.10). As expected, we found that the loss of AR generated the greatest loss of androgen-dependent enhancer activity in Inducible regions. However, we found that the loss of several other factors such as MED1 and RUNX1 also led to significant drops in enhancer behaviour for the top Inducible regions despite the presence of AR. This suggests that multiple factors can play a crucial role and are required for androgen-dependent inducible activity. For those regions that showed little to no enhancer activity, knockout mutations were not able to recover the same level of enhancer activity as other top regions. While mutations to WDR5 and KDM1A showed some effect, the resulting change in transcriptional activity from hormone treatment was significantly lower than Inducible regions. We now discuss some of the key transcription factors identified by the model for androgen-dependent regulation.

Interestingly, our model associated HOXB13 binding with lower transcriptional output, based on the strong positive interaction energy with Pol-II. The role of HOXB13 is still largely unclear in prostate cancer. Some argue that HOXB13 functions as a suppressor of prostate tumour growth [89]. However, it is generally believed that HOXB13 helps with the binding of AR by acting as a pioneer factor and influences where AR binds [29, 81]. Looking at HOXB13's interactions with other factors, we saw that it had a strong negative interaction energy and therefore a cooperative relationship with TLE3, a known transcriptional co-repressor, and ARID1A. As such, HOXB13 may have more of an indirect effect on transcription, as it recruits other transcription co-repressors such as TLE3. ARID1A is a member of the SWI/SNF chromatin remodelling complex and has been shown to directly interact with AR [81]. It is possible that ARID1A, as part of the SWI/SNF complex, interacts with HOXB13 in a similar manner and changes the binding relationship of the pioneer factor.

Of the TFs, MED1 showcased the strongest correlation with AR (Fig. 3.6). MED1 is a part of the mediator complex which mediates interactions between enhancers and Pol-II [91]. Our results are consistent with this activity, as we demonstrate that MED1 helps to mediate AR-dependent transcription. In addition to having similar interactions as AR, MED1 also

showcased strong negative interaction energy with Pol-II (Fig. 3.4). This suggests that MED1 is an important mediator factor of binding for a variety of different TFs as well as Pol-II. Despite a drop in the average occupancy when deprived of androgen, the occupancy of MED1 correlates well with increased transcription in both treated and control conditions. As stated before, this is crucial as androgen-deprived growth is the main driver of late-stage lethal castration-resistant prostate cancer [28]. Our results suggest that MED1 may influence the enhancer activity of other transcription factors in castrate conditions. Similar to previous work, we found that the loss of MED1 dramatically reduced the enhancer capabilities of Inducible regions [92, 93]. In castrate-resistant prostate cancer cells, MED1 has been shown to help recruit transcriptional machinery and induce transcription in the absence of AR [94–96]. All of this supports our previous theory of MED1 as a potentially important driver of androgen-dependent inducible activity similar to AR.

RUNX1 was another factor singled out as a critical factor for enhancer activity from the mutational analysis. RUNX1 is a transcription factor, that is also a target gene of androgen-induced and AR regulated transcription [97]. This coincides with what we found; the occupancy of RUNX1 correlated with the self-transcription output in the presence of androgen but not in control experiments. RUNX1 also showed a negative interaction energy with other important factors such as AR, MED1, and PIAS. This suggests a cooperative binding relationship between RUNX1 and other androgen-dependent factors. Additionally, RUNX1 showed a strong positive interaction energy with HOXB13. This is interesting; as generally, it is believed that both RUNX1 and HOXB13 interacts positively with AR [81, 88]. We showed that there might be an antagonistic relationship between HOXB13 and RUNX1 binding. Recent experiments have shown that the knockdown of RUNX1 is related to decreased androgen-dependent cancer cell proliferation [97]. This matches up with our own theoretical knock-out experiments. Apart from MED1 and AR, the loss of RUNX1 in active Inducible regions lead to the greatest drop in enhancer capabilities. As such, it is plausible that a large shift in the transcriptional logic would lead to deteriorated cell function.

In summary, our biophysical approach leveraged the large amount of data from ChIP-Seq and STARR-seq assays to determine interaction energies between transcription factors and identify important TFs for androgen-dependent enhancer activity. Our model generated a self-consistent energy landscape that is representative of experimental values. Finally, knock-out mutations of Inducible ARBS regions ranked each TF in terms of importance for enhancer behaviour in Inducible ARBS regions. Future work will incorporate the spatial information from the binding data to allow for the discovery of a position-dependent grammar between the various factors. Given the recent advancements in deep learning for identifying the important sequence motifs from both binding and transcriptional data, it is only natural to consider merging sequence, binding and transcriptional data to get a basepair

resolution of the regulatory factors. Such work can provide a more in-depth picture behind the complex combinatorial transcriptional logic of androgen and AR in prostate cancer.

Chapter 4

Future Work

During my Master's degree, I have mainly been focused on the two projects shown in the previous two sections. However, there were a few other projects that showed great promise but I simply did not have enough time to finish. Rather than simply leaving them out of the thesis, I think it is important to dedicate a chapter to such works. These projects employ different strategies of tackling the same statistical learning problems as before, and each reveals new information regarding the regulatory transcription logic in prostate cancer.

4.1 Principal Component Regression

In chapter 2 and 3, we developed models that classified and made predictions of enhancer activity based on a single score that represented the occupancy of each factor in that region. However, a factor's binding may depend on specific locations within a region and its position relative to other factors. Thus our approach in the previous chapters is an over-simplification and omits the spatial aspects of binding. To rectify this, we want to include the spatial binding signal of each factor over the entire 750bp of each region. The spatial signal is measured at a 50 bp resolution which means that each factor has 15 values representing its binding over the region. As part of the preliminary analysis, we will be focusing on 30 different ChIP-seq profiles, leaving us with an input space of 450 dimensions. Besides the significant increase in dimensions, the spatial binding of a factor is generally correlated and can lead to high collinearity within the feature space, which can pose problems for regression. To counteract this, we leveraged principal component analysis (PCA) to reduce the dimensionality and collinearity within the input data.

Principal component analysis (PCA) is a powerful statistical learning tool that has become popular due to its ability to reduce the dimensionality of datasets. Essentially, PCA rotates the data space basis in a way that maximizes the variation across the data along only a few dimensions. Mathematically, this is done through an eigendecomposition of the data's covariance matrix. Considering a feature space X with zero mean, and a covariance matrix of Σ, it can be written such that

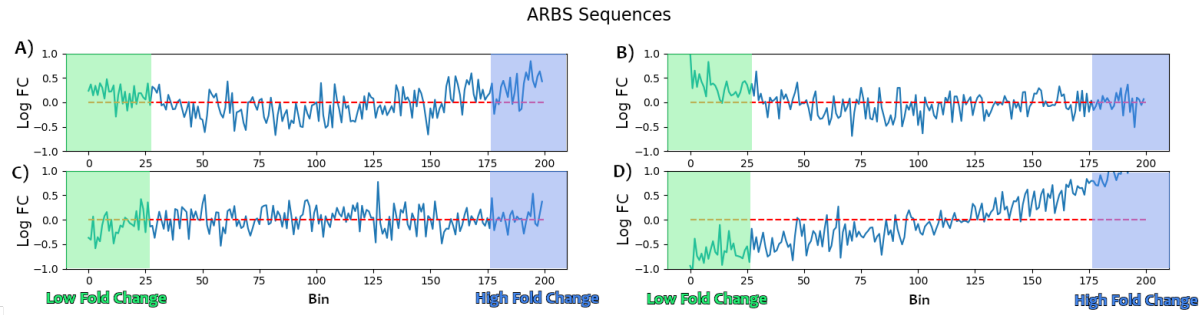

Figure 4.1: (A-D) Averaged PCA representation of the spatial TF binding data on the top 4 principal components. The projections were binned and sorted based on log fold change and can be examined for correlation with enhancer capabilities. Each bin corresponds to the average value of ∼20 ARBS regions.

$$\Sigma = W^T \lambda W \tag{4.1}$$

$$P_L = XW_L \tag{4.2}$$

$$Tr(\lambda) \sim Tr(\lambda_L) \tag{4.3}$$

where W, λ represent the eigenvectors and eigenvalues of the covariance matrix respectively. To reduce the dimensionality of the system, L principal components are selected on which the data will be projected. W_L represents the subset of L eigenvectors from the original W eigenvector matrix.. P_L represents the downsampled projection which is then used for regression or other identification tasks.

The W_L eigenvector matrix forms an orthogonal basis, and thus removes the problem of collinearity across our features. By downsampling data by using only the top eigenvectors, we are effectively removing variances in the original dataset that we believe to be noise. The optimal way to determine L, the number of principal components is still largely debated. Generally, this is done by reducing L as much as possible without sacrificing prediction power.

PCA can be performed in conjunction with regression where the parameter estimates on the new orthogonal basis can be rotated back into the original space using the same eigenvectors,

$$\hat{Y} = P_L \beta_L = X\beta^* \tag{4.4}$$

$$\beta^* \sim W_L \beta_L \tag{4.5}$$

where β^* represents the parameter estimates in the original feature space while β represents the parameter estimates in the reduced principal component space.

In terms of our ARBS work, this approach allowed us to include spatial binding data without fear of overfitting. Highly correlated spatial binding would be represented by a

single principal component. PCA rotates the spatial data onto an orthonormal basis where each principal component can represent the spatial binding of multiple TFs. Variances in the spatial data can be downsampled into principal projections (Fig. 4.1) and measured in terms of predictive power for enhancer activity. From there, parameter estimates in the principal component space can be rotated back into spatial data space where one can assign contributions to the spatial regions of each TF. This approach bypasses the need for more extreme methods such as ensemble learning to reduce the multicollinearity within the data.

4.1.1 Second-Order Polynomial PCA

The reduced dimensionality afforded by PCA also allows us to explore fitting models with higher-order interactions. In many situations where the data is sparse, the addition of higher-order terms rarely adds power to the model and makes it more difficult to perform feature selection. However, in our case, potential second-order interactions can signify the cooperative binding of transcription factors with Pol-II. As a result, this can imply important collocation of binding and thus the formation of protein complexes. To do this for a given region r, we made use of PCA to reduce our dimensionality and then perform second-order regression on the projections,

$$\hat{Y_r} = \sum_{i}^{L} P_{r,i}\beta_i + \sum_{i}^{L}\sum_{j}^{L} P_{r,i}P_{r,j}\hat{\beta}_{i,j} \tag{4.6}$$

where β and $\hat{\beta}$ represent the first-order and second-order projection weights. $P_{r,i}$ represents the projections of the feature space of region r onto eigenvector i.

To find the weights in our original space, we simply need to make use of the same rotation matrix as before and the distributivity and commutativity properties of summation.

Linearly expanding the projections in terms of Eq. 4.2 and letting x_r denote the original $(n \times p)$ input space for a region r,

$$\hat{Y_r} = \sum_{i}^{L}\left(\sum_{k}^{p} w_k^i x_{r,k}\right)\beta_i + \sum_{i}^{L}\sum_{j}^{L}\left(\sum_{k}^{p} w_k^i x_{r,k}\right)\left(\sum_{m}^{p} w_m^j x_{r,m}\right)\hat{\beta}_{i,j}$$

Multiplying out the sums gives,

$$\hat{Y_r} = \sum_{i}^{L}\sum_{k}^{p} w_k^i x_{r,k}\beta_i + \sum_{i}^{L}\sum_{j}^{L}\sum_{k}^{p}\sum_{m}^{p} w_k^i w_m^j x_{r,k}x_{r,m}\hat{\beta}_{i,j}.$$

Rearranging the sum results in,

$$\hat{Y_r} = \sum_{k}^{p}\left(\sum_{i}^{L} w_k^i\beta_i\right)x_{r,k} + \sum_{k}^{p}\sum_{m}^{p}\left(\sum_{i}^{L}\sum_{j}^{L} w_k^i w_m^j\hat{\beta}_{i,j}\right)x_k x_m$$

$$\hat{Y_r} = \sum_{k}^{p} \beta_k^* x_{r,k} + \sum_{k}^{p} \sum_{m}^{p} \hat{\beta}_{k,m}^* x_{r,k} x_{r,m} \tag{4.7}$$

$$\beta_k^* = \sum_{i}^{L} w_k^i \beta_i \qquad\qquad \hat{\beta}_{k,m}^* = \sum_{i}^{L} \sum_{j}^{L} w_k^i w_m^j \hat{\beta}_{i,j} \tag{4.8}$$

where β^* and $\hat{\beta}^*$ represent the PCA recovered first-order and second-order weights in the original space. In event where the full eigenvector matrix W is included ($L = p$), the recovered weights should be identical to the parameter estimates in second-order linear regression.

Additionally, we assume that $\hat{\beta}$, the second-order weights in the projection space, are symmetric:

$$\hat{\beta}_{k,m} = \hat{\beta}_{m,k} \tag{4.9}$$

This further reduces the dimensionality of the feature space and decreases the computation time of the model.

The second-order PCA regression was done using the PCA decomposition and the cross-validation Elastic-Net regression from the `sklearn` package. Hyperparameter selection was done using the 5 fold cross-validation. The resulting model is an average of the models generated during cross-validation.

4.1.2 Data Preparation

Similar to chapters 2 and 3, ChIP-seq and STARR-seq data of the ARBS regions were used to perform PCA regression. As part of the preliminary analysis, the input data will consist of 30 different ChIP-seq profiles measured mainly in DHT (hormone) with a few factors of interest measured in EtOH (no hormone) conditions. Each of the 4139 non-overlapping 750bp ARBS regions consists of 15 bins at a 50 bp resolution for a total of 450 (450=30×15) spatial features. To normalize across ChIP-seq profiles, an arcsinh transformation is applied to the spatial features.

$$S^* = \sinh^{-1}(S) = \ln(S + \sqrt{1 + S}) \tag{4.10}$$

This normalizes the ChIP-seq scores by maintaining the magnitude for small values ($S \leq 1$) while lowering larger ChIP-seq scores into a similar range.

Additionally, we take the STARR-seq log fold-change value of each enhancer region as a measure of its enhancer activity. Although one could expand on this and include the full 90 ChIP-seq profiles from chapter 2, we found that the additional ChIP-seq profiles increased the degrees of freedom significantly and led to trouble recovering the second-order spatial interactions. Future work could focus on the application of this method on larger datasets.

We found that $L = 20$ components captured the majority of the information (55% of the total variance) from the spatial features. While one can include additional PCs, any additional PCs added significantly less information ($\leq$ 0.6% of the total variance) while

Figure 4.2: First-order spatial weights recovered from PCA projection (see Eq.4.8) shows the predicted spatial weight to enhancer behaviour. The model accurately attributed AR expression under the treatment of DHT as the strongest indicator of enhancer behaviour. Features are sorted based on their average spatial weight. Red signifies an interaction that enhances differential activity while blue signifies a repressive interaction. Majority of the ChIP-seq scores were tested in the presence of DHT while a select few factors of interest were tested in both DHT (R1881 = DHT) or control (veh, EtOH, no DHT) conditions.

significantly increasing the degrees of freedom for the model (adding one PC increases the number of parameters by 21). Models with more PCs were fit but showed no significant drop in MSE on a validation data set compared to the model using the first 20 PCs. This meant that we had 20 first-order projections and 190 second-order projections for a total of 210 parameters per ARBS region. This is a significant reduction compared to the 450 1st order spatial features and 202500 possible second-order features.

4.1.3 Results

Here we present our results from fitting our second-order PCA model to enhancer activity from STARR-seq measurements of ARBS regions. 30 profiles of ChIP-seq data of 4139 non-overlapping ARBS regions were used to generate a 750 (15x30) bp spatial binding profile of each region. Principal components are generated using the covariance matrix of the spatial binding data, and the resulting top 20 components (see Section 4.1.2) sorted by eigenvalue are collected to represent the variances in the original space. We then expanded the feature space by generating the 2nd degree polynomial for the 20 principal components.

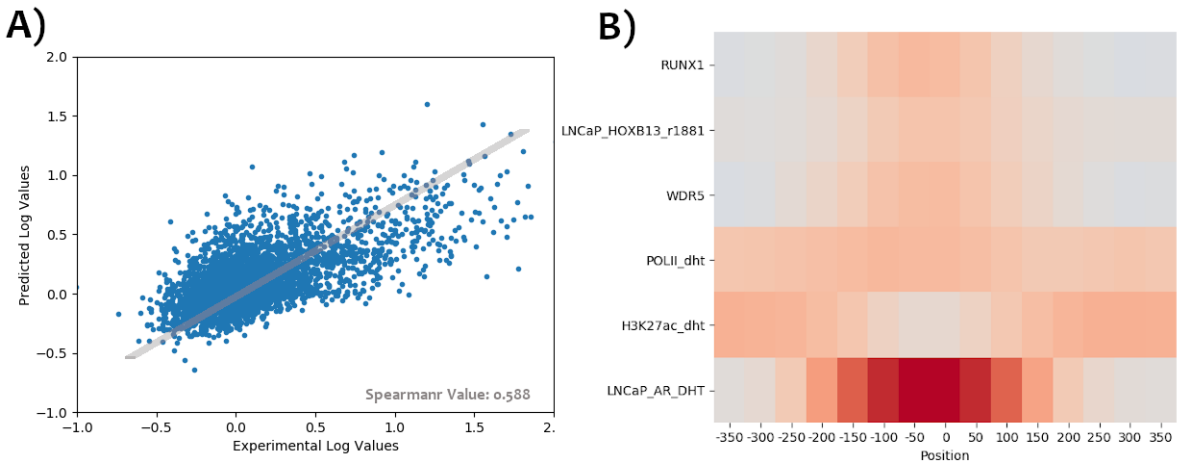

Figure 4.3: (A) Predicted enhancer activity of the training set after cross-validation shows a strong correlation with experimental log-fold change, with a Spearman r correlation of 0.588. The final model was the average model after 5 fold cross-validation training using first and second-order PCA projections. (B) The first-order spatial weights of the top 6 factors sorted by average spatial weight includes many familiar co-factors as previous chapters. AR, HOXB13, and RUNX1 are all factors that were shown to be implicated in AR-regulated transcription in past work. Red signifies an enhancement of differential activity between DHT treatment and control/EtOH conditions.

Using the first and second-order PCA projections from 4139 non-overlapping ARBS regions, we were able to train an Elastic-Net model with 5 fold-cross validation. The averaged model after cross-validation was then used to re-predict the training data and managed to achieve an MSE of 0.074 and a Spearman r correlation of 0.588 with the experimental values (see Figure 4.3A). For comparison, this process was repeated using the spatial binding data without PCA and with first-order PCA regression. Regression using the spatial data without PCA achieved 0.086 MSE and 0.574 Spearman r correlation while first-order PCA regression achieved 0.090 MSE and 0.550 Spearman r correlation. As expected, the loss of dimensionality during PCA lead to a slight drop in predictive power. While the improvements are relatively small, the comparisons between the 3 models become significant when considering the degrees of freedom for each model. Spatial binding regression had 450 parameters while first and second-order PCA regression had 20 and 210 parameters respectively. As such, second-order PCA regression had significantly fewer parameters than the spatial model and managed to outperform in both MSE and spearman r correlation.

As stated previously, PCA eigenvectors form an orthogonal basis that can be used to reduce the number of dimensions by only using a subset of them. We found this to be an effective method to determine the spatial impact of each factor on enhancer behaviour. By first transforming the spatial data onto the eigenvector basis (see Section 4.1.1) and then performing regression on the projections, we were able to recover the first-order spatial weights of each factor (Fig. 4.2). Factors were then sorted based on their average spatial

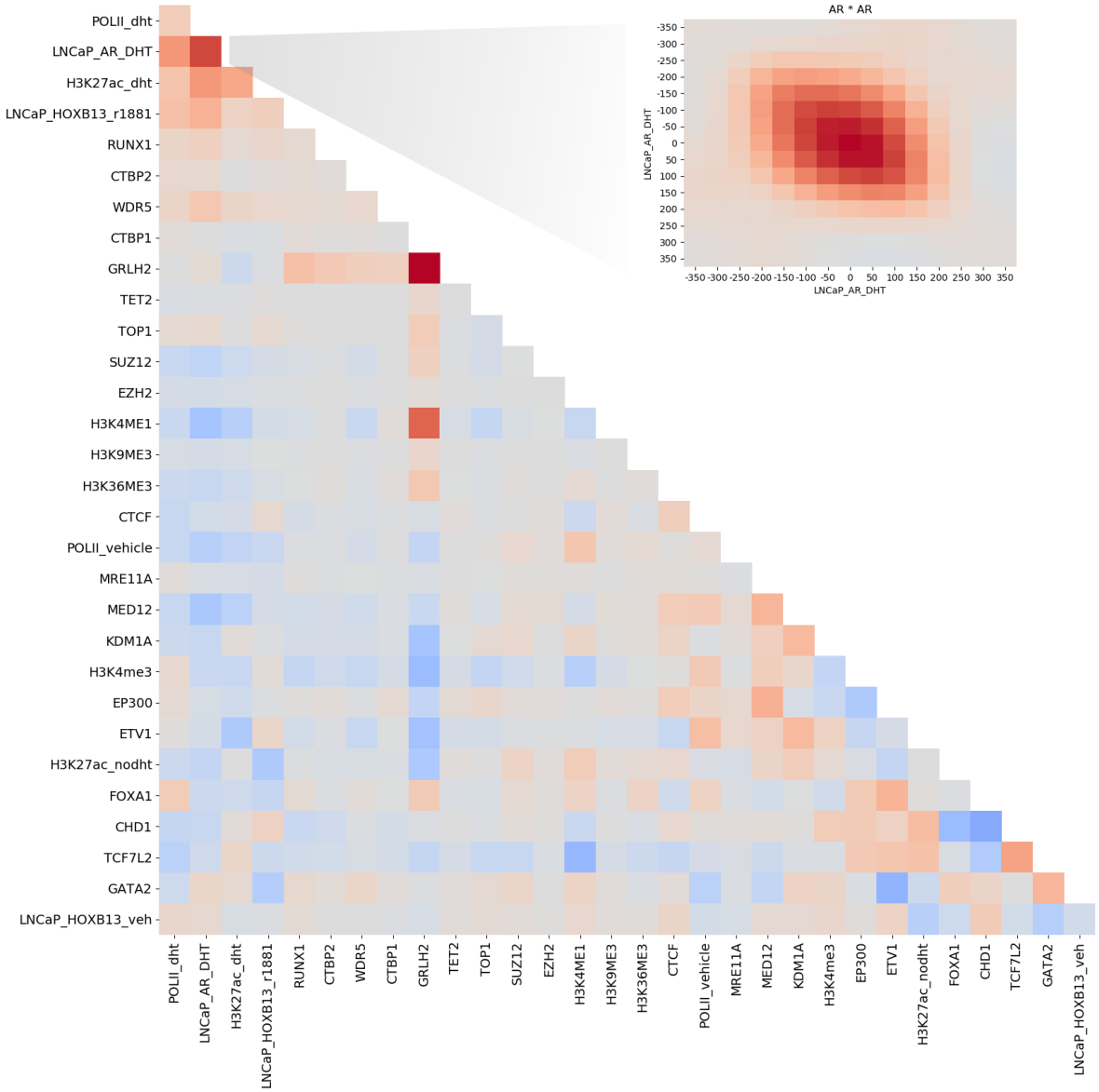

Figure 4.4: Second-order spatial interaction matrix showcases the mean spatial weight across two-factor 15x15 spatial interaction matrices. Red signifies an overall positive interaction whereas blue signifies a negative interaction. The zoomed plot (top right) shows the AR to AR spatial interaction matrix.

weight to enhancer behaviour. As expected, when treated with DHT the strongest indicator of enhancer activity was AR followed by H3k27ac. Interestingly, both H3k27ac and AR were heavily weighted in opposing spatial regions. AR had its weights concentrated on the central part of the region where it is most often bound. On the other hand, H3k27ac showed increased weight on either ends of the region. Factors such as RUNX1 showed similar weighting toward the center, suggesting a cooperative binding relationship with AR.

As shown in Eq. 4.8, the fitted second-order weights in the PCA space can be converted back into the second-order spatial weights by applying the PCA rotational matrix in reverse. The second-order spatial weights (450x450) were then split into individual spatial interaction matrices of size (15x15) for all 30 spatial profiles. To rank contributions and identify similarities, interaction matrices (15x15) between every pair of factors were collapsed into an average interaction value, and the resulting interaction matrix (Fig. 4.4) was sorted based on a Ward distance clustering algorithm [82].

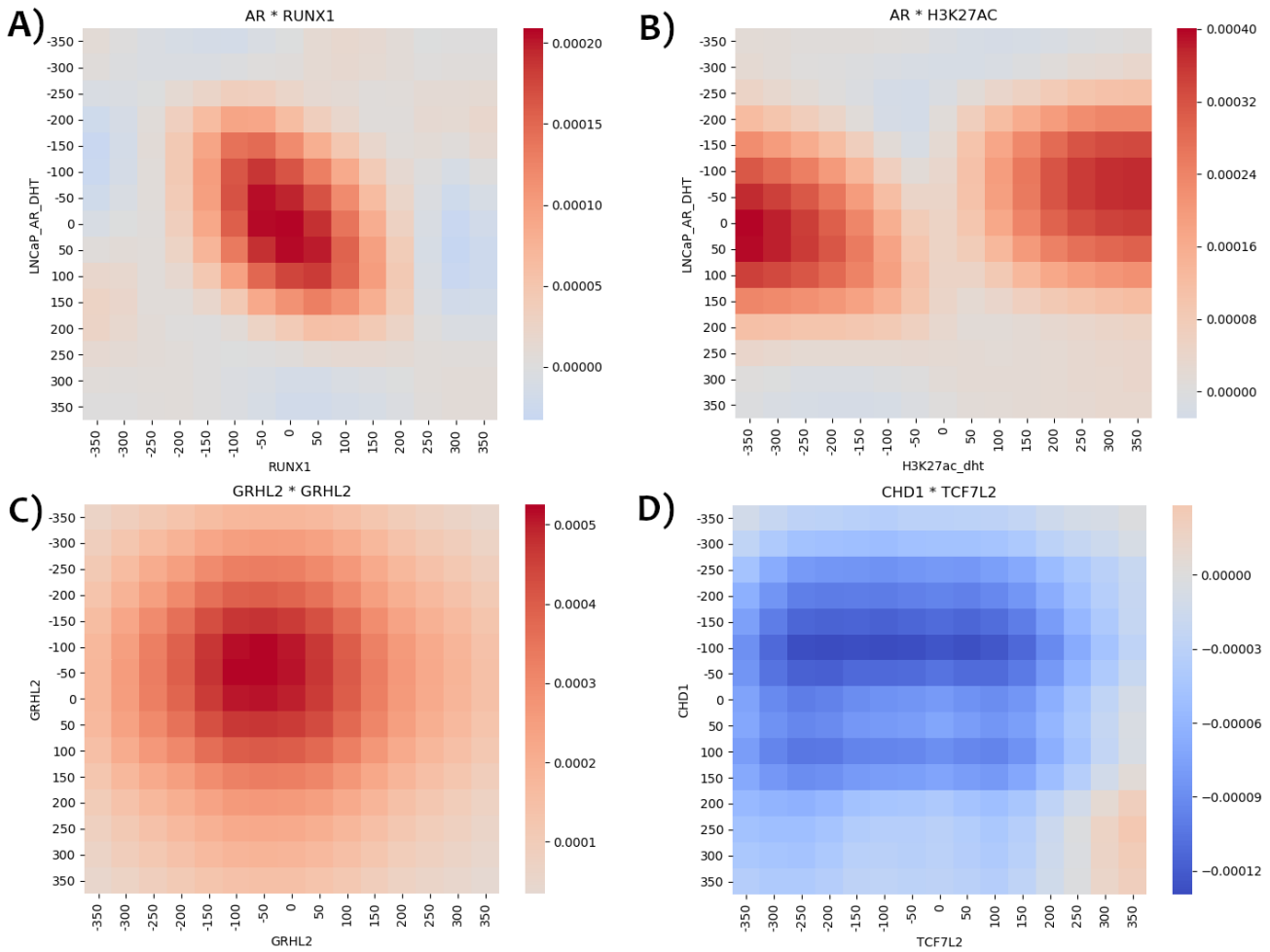

Figure 4.5: (A,B) Spatial interaction matrices between AR, RUNX1 (A) and H3k27ac (B) show positive weight towards DHT-driven enhancer behaviour. (C,D) The highest positively and negatively weighted spatial matrices between two-factor interactions. GRHL spatial interaction with itself shows strong positive weight towards DHT-driven enhancer activity while spatial interaction between CHD1 and TCF7L2 shows strong negative weight against enhancer behaviour.

From there, we could then study the individual interaction matrices between factors of interest. As expected, the interaction matrix of AR with itself was heavily centralized, where the spatial regions that contribute the most are along the diagonal and near center of the region. The same type of behaviour can be seen in the interaction matrix between AR and RUNX1 (Fig. 4.5A). In both interactions, colocalized binding of AR with itself or RUNX1 seemed to positively contribute towards enhancer behaviour. On the other hand, the interaction matrix between AR and H3k27ac showed anti-localized binding of AR and H3k27ac (Fig. 4.5B). Our model predicts the interaction to have a strong positive weight for enhancer activity near the centre of AR binding, but not along the diagonal of the interaction matrix. This suggests that the distal modification of H3k27ac away from AR was more beneficial for enhancer activity. H3k27ac is an epigenetic modification and changes the histone structure in the general area. This makes the predicted spatial interaction noteworthy, as the distal modification would avoid interference with the binding of AR and other co-factors.

Interestingly, of all the possible 2-factor interactions, AR interaction with itself was not the strongest contributor to enhancer activity. GRHL2 interaction with itself was predicted to be the strongest indicator for DHT-driven enhancer activity despite little to no contribution in the first-order spatial weights. GRHL interaction with itself shows a similar onsite interaction with itself, where central binding of GRHL contributed the most towards enhancer behaviour. This might indicate the formation of dimer where two GRHL comes together to form a specific protein complex. Biologically, GRHL2 is a novel co-regulator of androgen receptor discovered in 2017 and is believed to have a role in hormone-dependent cancers [98, 99]. This is significant, as it shows that the inclusion of the second-order principal components is capable of capturing information that is not present in the first-order principal components. On the other hand, CHD1's interaction with TCF7L2 was the most negatively weighted interaction, and the resulting interaction matrix was less clear. While there are distinct regions of the interaction matrix that contributed more to enhancer activity repression, it is hard to ascribe a particular relationship. Again, it is important to stress that these interaction matrices were generated by a transformation from the reduced principal component space. Additional experimental work would be required to show if the relationship plays a major role in DHT-mediated enhancer behaviour.

4.1.4 Discussion

Here we presented a new way of applying PCA in addition to second-order polynomial regression as a way to identify second-order interactions in a high dimension spatial binding dataset. The second-order polynomial PCA regression model performed better than the model constructed with the original spatial data despite fewer degrees of freedom, with a MSE of 0.074 and a spearman r correlation of 0.588 with the experimental values. Comparing this to the differential model in Chapter 3 which had a spearman r correlation of 0.45, this

is a significant jump in accuracy and shows the additional power of the spatial data and secondary order PCA regression. With that being said, the model in Chapter 3 had more samples included and contained regions that showed no AR binding but had the AR motif. As such, one should expect that increase in predictive power was partially due to the removal of regions that showed no AR binding and partially due to the change in methodology.

By transforming the parameter estimates from the principal component space back into the spatial binding, we were able to fully map the first-order spatial weights of each factor. AR and H3k27ac when treated with DHT were the two most impactful factors while displaying opposing binding patterns. AR showed a centralized binding pattern while H3k27ac occupancy away from the center contributed the most towards enhancer behaviour. This shows an important distinction between potential on-site co-factors and distal off-site histone changes for transcription. For instance, RUNX1 was also found to have its weights concentrated around the center of the region, similar to AR, and as such, the two factors likely share a similar co-factor complex and function. On the other hand, H3k27ac functions as a histone modification and as such, would not participate in the formation of a co-factor complex. By mapping out the spatial weights of each factor, we are able to associate a potential role for each factor and identify potential TFs that contribute to the AR complex.

This separation in binding behaviour can also be seen in the second-order interaction matrices where AR's interaction with itself and RUNX1 showed an onsite interaction but with H3k27ac showed an offsite interaction. Interestingly, GRHL's interaction with itself was predicted to be the best two-factor interaction for enhancer activity, while the interaction between CHD1 and TCF7L2 was predicted to be the strongest repressor of log fold change. GRHL is a novel co-factor in the AR signalling pathway discovered in 2017, and its effect in prostate cancer is still a topic of ongoing research [98, 99]. CHD1 is believed to help regulate transcription and nucleosome positioning, while TCF7L2 helps to regulate the Wnt pathway, a type of signalling pathway that passes through the cell surface [100, 101]. Interestingly, mutations to CHD1 and TCF7L2 are commonly found in prostate cancer and thus can signify that both factors potentially play a role in cancer suppression [42, 44, 101]. This shows a clear distinction between the highly positive parameter estimates (AR, GRHL,..) whose overexpression is common in prostate cancer and negatively weighted parameter estimates (CHD1, TCF7L2) which are often lost in prostate cancer and thus can potentially serve as repressors that suppress the aberrant expression of androgen.

To understand the parameter estimates in terms of the biophysics of binding, one can refer to the mean-field modelling in chapter 3. Rather than fitting to the transcriptional output, here we are fitting to the log fold change. As such, the parameter estimates should be viewed as a change in energy when exposed to DHT compared to control. Additionally, rather than binding energies, we are fitting the feature data to the enhancer output and so, one should view the first-order parameter estimates as the interaction energies between Pol-II and the factors of interest. Similarly, the second-order parameter estimates can be

understood as the 3-body interaction energy between Pol-II and the two factors of interest. Rather than finding the full 3-body interaction tensor, the second-order interaction matrix (Fig. 4.4) is a part of the 3-body interaction tensor that is related directly to Pol-II binding. As such, unlike in chapter 3, the weights shown will not dictate how likely a given factor will be occupied, but rather, how much more likely Pol-II is to bind in the presence of androgen compared to control given that the factor of interest is present.

4.2 Convolutional Neural Nets (CNN)

Convolutional Neural Nets (CNNs) are a powerful deep learning tool that has been used in recent years to study transcriptional regulation. Unlike approaches in chapters 2 and 3 which rely on a linear correspondence between input and output, CNNs make use of hidden layers between the input and output layer that contain enough complexity to approximate any functional form [22]. As stated before, CNNs also contain additional convolution layers that scan the input data for similar spatial patterns. This type of pattern recognition makes CNNs the ideal model for image classification and other similar machine learning tasks. In terms of transcriptional regulation, CNNs are given DNA sequence data along with the resulting transcription output to identify potential binding motifs and other important sequence patterns that can affect transcriptional regulation [75]. We attempted a similar sequence approach by designing our own CNN model to classify ARBS enhancer regions into Inducible and Non-Inducible categories. This section will cover our preliminary attempt at using a CNN to uncover binding motifs in ARBS regions that are predictive of differential enhancer activity.

4.2.1 Data preparation

Rather than using the ChIP-seq binding data, we consider a CNN that will only take the DNA sequence of each ARBS region as input. This puts more emphasis on the pattern recognition functionality of CNNs, as it will attempt to identify repeated sequence motifs (that may be binding sites for TFs) that are present within the enhancer regions. As such, training a CNN generally requires a large amount of data. In order to achieve this, we broke down the regions of the STARR-seq assay into smaller tiles. In the STARR-seq assay presented in chapter 2, the ARBS regions were captured and cloned into a library. In this process, a particular ARBS region would get cloned into a set of overlapping tiles of variable length that span the region. Each of these tiles gets assayed for transcriptional activity. Thus in the STARR-seq experiment, the 4139 ARBS regions were actually represented by $300,000$ tiles. As stated in Chapters 2 and 3, this data was pooled into a single regional representation to reduce noise and experimental error. For training the CNN, we will use the sequences for the tiles along with their measured transcriptional output. To meet the

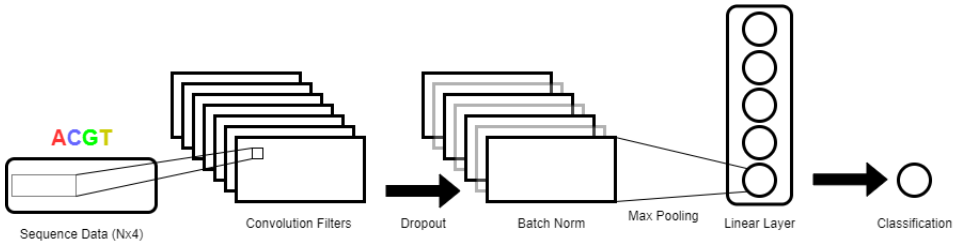

Figure 4.6: Schematic of the CNN. Tile sequence data are scanned for spatial patterns using the learned filters. The convolution results are then passed to a batch normalization layer which helps normalize training variations across mini-batches and a dropout layer which randomly removes filters from the model to reduce the possibility of over-fitting. The spatial results are finally pooled together, and the maximum value are sent to the linear layer for classification.

CNN's requirement of data, separate tile reads were treated as individual data points and predicted for DHT-driven enhancer classification.

To ensure better training, several data cleansing filters were applied to reduce some of the variations in the tile data. Firstly, tiles must be 450-700 bp in length and show reasonable output ($>$ 10 reads) in both DHT and EtOH conditions. Tiles that were too short in length and did not show reasonable output had increased variance presumably due to noise and were thus removed from the training set. Additionally, a random DNA sequence was generated and used to pad all tiles on both sides such that the final tile reached 700 bp in length. Finally, to balance both the forward and reverse directional sequences, the reverse complement was also added to the training data. Tiles that showed significant differential transcription output ($\log(\frac{m_{DHT}}{m_{EtOH}}) \geq 0.5$) were considered inducible while the rest are considered Non-Inducible. This is similar to the Inducible cutoff used in Chapter 2 but slightly lowered to increase the number of possible Inducible tiles. To balance out the tiles between classes and to give an advantage to the classification model, tiles that showed mediocre inducibility ($0.2 < \log(\frac{m_{DHT}}{m_{EtOH}}) < 0.5$) were removed from the training set (Fig. 4.7). This resulted in a final training set of $\sim$ 127000 tiles ($\sim$53000 Inducible, $\sim$73000 Non-Inducible).

4.2.2 CNN development

Rather than focusing on the predictive power of the model, our CNN approach prioritizes the interpretability of the filters and thus is simpler than most CNN models. The model consisted of a layer of 100 convolution kernels 8 bp in length followed by max-pooling of the entire tile, batch normalization, and a dropout linear layer (see Figure 4.6) before classification. As stated in Section 1.4.2, the convolution kernels scan the input data for patterns and will identify important DNA patterns. Batch normalization normalizes the variance between training steps and ensures a more gradual learning curve. The dropout layers randomly remove certain filters from training and ensure that the model does not

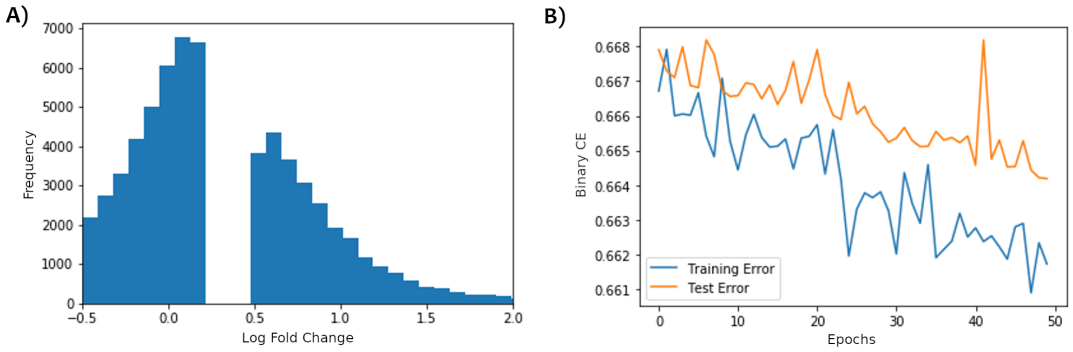

Figure 4.7: (A) Histogram of the log fold change ($\log(m_{DHT}/m_{EtOH})$) across training tiles. Any tiles with significant differential expression ($\log(m_{DHT}/m_{EtOH}) \geq 0.5$) are considered Inducible while all other tile data is considered Non-Inducible. (B) Binary cross-entropy of the CNN model tested using the training and test dataset during training. Gradient-based optimization of the CNN slowly lowers the cross-entropy as training epochs continue.

rely on a single filter. One can imagine the dropout layer as a form of regularization similar to LASSO and Ridge regression which we have discussed before (Section 1.5). This setup contains enough complexity to identify important spatial patterns among the sequence data while still being able to differentiate between crucial filters for classification. Other deeper models were tested but the increase in complexity made the model difficult to train using our current training set. Additionally, complex changes to the CNN structure such as the inclusion of short- and long-term memory units were not tested due to time constraints, but can potentially improve the overall model accuracy.

The training of the CNN was done using the Adam function of PyTorch with a learning rate of 0.001. Adam is an advanced version of gradient-based optimization such as stochastic gradient descent and leverages a moving learning rate to improve the training capabilities of the model. Error measurement for classification was done through the binary cross-entropy,

$$BCE(Y, X) = \frac{-1}{N}\sum_{i=1}((y_i)log(x_i) + (1 - y_i)log(1 - x_i)) \tag{4.11}$$

where Y represents the true classification, and X represents the predicted probability of the class.

4.2.3 Results

After 100 epochs (Fig. 4.7B shows the last 50 epochs) of training, the final results of the CNN model accuracy on the test data are collected and the learned filters are examined. After training, the CNN achieved ~59% accuracy with a low recall of the Inducible tiles. This is lower than the 65% accuracy for Inducible regions in Chapter 2. Additional training time and changes to the CNN structure can increase the prediction accuracy. Furthermore, the drop in accuracy might be due to the nature of the tile dataset which contains more

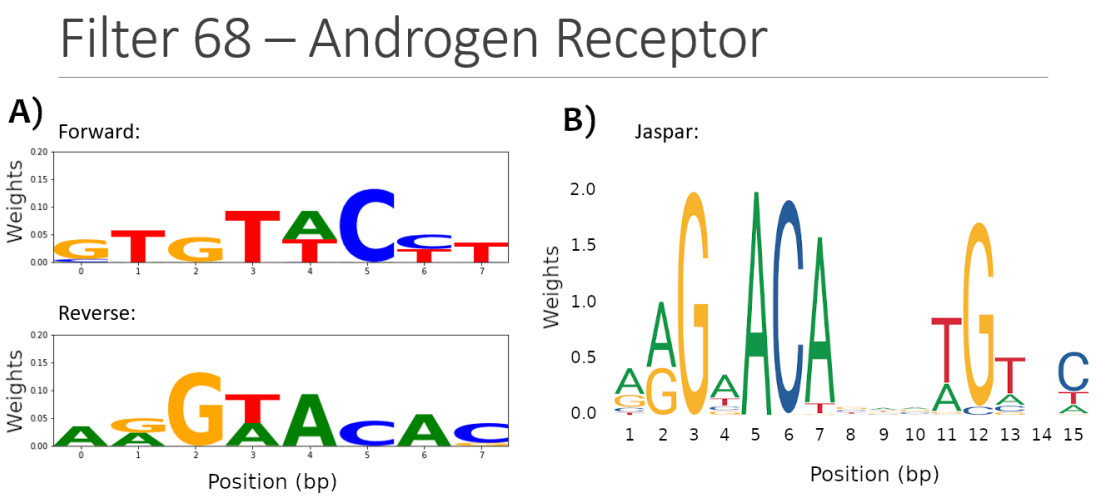

Figure 4.8: Learned filter 68 showed strong similarities to the binding motif of AR in Jaspar. (A) The forward and reverse complement of filter 68 were learned by training the CNN on the tile representation of the STARR-seq data. (B) Jaspar prediction of the AR binding motif. Similar sequences such as the "*AAGAACA*" sequence are strongly weighted in both presentations.

experimental noise than the measurements from ARBS regions. Currently, our approach is to ensure that CNN models used in conjunction with the STARR-seq tile data are capable of identifying important motifs. As such, the level of accuracy is not unexpected as the model was simpler than most CNN models and had less data than other approaches.

Filters 68 and 10 had the strongest weights for the Inducible class, at 0.85 and 0.59 respectively. Converting the weights into a position weight matrix for both filters revealed close matches to binding motifs in Jaspar, a database of transcription factor motifs [102]. Since both forward and reverse complements of the tile data were included, the reverse complements of the learned filters were inspected for similarities to the binding motif. Filter 68 of the CNN model showed similarities to the Androgen Receptor binding motif in Jaspar while filter 10 shared a similar binding pattern to the FOX family of proteins. Filter 68 and AR binding motif both shared the same "*AAGAACA*" sequence (Fig. 4.8) that is believed to be important for AR binding. AR generally binds as a dimer and therefore contains both the forward and reverse sequence as part of its binding motif on Jaspar. The sequence that was learned by the CNN is considered to be the binding motif for a single AR protein. Similarly, filter 10 heavily weights the "*GTAAACA*" sequence (Fig. 4.9) which corresponds with the binding motif for FOXP1 and FOXA1. FOX family proteins have been implicated in facilitating AR binding, and our previous work (see Chapter 2) has shown that FOX family binding is correlated with increased enhancer classification [30]. Other highly weighted filters were analyzed for similarities to other factors' binding motif. Unfortunately, due to the high number of possible factors, none of the other highly weighted

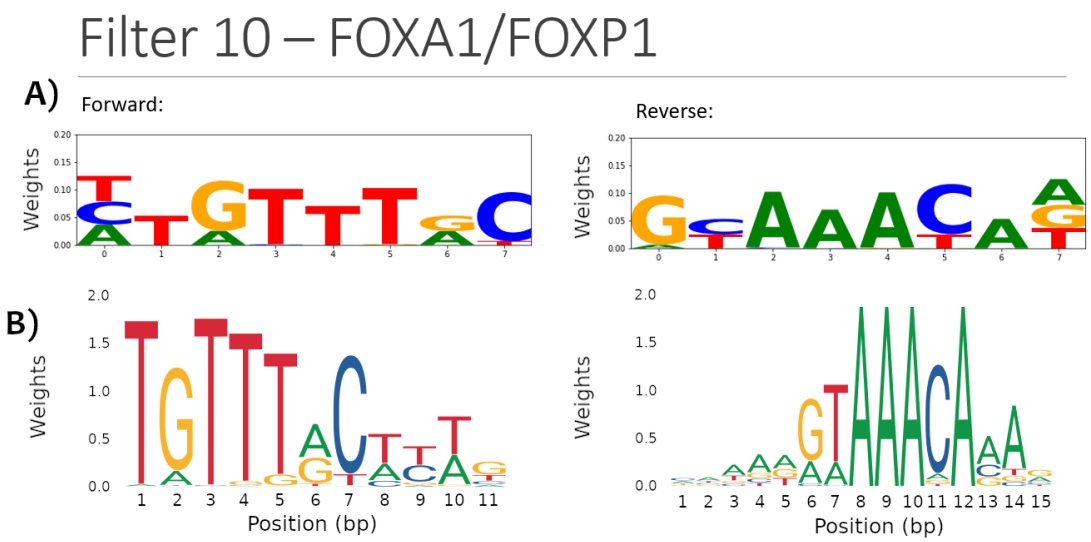

Figure 4.9: Filter 10 of the CNN showed similarities to the binding motif of the FOX family. (A) The forward and reverse complement of filter 10 learned by the CNN. (B) The left figure shows the binding motif of FOXA1 while the right figure shows the binding motif of FOXP1 on Jaspar. Both the learned filters and the binding motifs show a strong preference for the "$GTAAACA$" sequence or its reverse complement.

factors matched known motifs to the degree that we are comfortable with matching the two together. Future work will have to focus on a methodology that is capable of identifying and matching known motifs to predicted filters.

4.2.4 Discussion

By breaking down the STARR-seq data into individual tiles, our preliminary attempt was able to train a CNN that was capable of identifying Inducible tiles with an accuracy of ~59%. While the accuracy was lower than the 65% accuracy in previous approaches, increased training time and samples would increase the accuracy of the resulting model. More importantly, our simple approach was able to recover the binding motifs of two important factors that impacted the enhancer capabilities of ARBS regions. While the two factors were not surprising and had previously been identified to be important for enhancer classification, the similarities with which the filter weights matched the known motif was surprising. Both the AR filter and the FOX filter weights coincided with 7 bp long patterns of the binding motif. The strong similarities between the filter and the motif suggests that the CNN can with limited data help identify important co-factors of transcription. This is promising, as additional sequencing data in the future can then be used to discover novel factors that can impact transcription. Future work could look into a possible metric that could measure the similarities between CNN filters and position weight matrices.

Future work can focus on the combination of binding data and sequencing data. As stated in chapter 3, factors potentially missing from the fitted model will impact the resulting conditional site energies and erroneously attribute additional occupancy in certain conditions. This can lead to inflated site energies for TFs that interact with the missing factors. Having a CNN along with the sequence data will help compensate for some of the effects due to missing factors, as potential binding motifs of the missing factor can be identified through the CNN. One can also regularize the CNN such that it will only try to identify filters for factors that were not included in the binding data, by removing filters that show high correlation with a factor's binding data. This will give a more complete model of transcription and a better picture of the underlying regulatory mechanics behind transcription activity and DHT-driven enhancer capabilities.

Chapter 5

Conclusion

In this thesis, we have discussed both biophysical and machine-learning methods to model transcriptional regulation in prostate cancer and have identified important transcription factors that help drive DHT-dependent enhancer behaviour. The nature of Pol-II's regulation makes regression on the highly collinear TF binding data tricky. As shown in Section 1.6, regression on such a datatset can lead to inflated parameter estimates and high variances in our predictions. As such, developing an approach that is still interpretable despite the high collinearity is vital for our biological analysis. For instance, in chapter 2, we implemented an ensemble logistic regression classifier that leverages random sampling of the data space to overcome the collinear binding data. By random sampling the training data, and training smaller base estimators, we were able to perform feature selection and rank each feature based on its contribution to inducible classification. Without ensemble sampling, feature selection on the dataset would have led to inflated estimates for a few key factors and reduced weights for the remaining factors.

While we did not explicitly discuss this in the thesis, the problem of collinearity is largely minimized in the mean-field model due to our approach to the problem. Unlike in chapter 2 or in chapter 4.1 where the binding data of TFs in both conditions were treated as separate variables, the mean-field approach models the equilibrium binding of Pol-II in both conditions. This meant that correlated expression between TFs must be maintained in both DHT and EtOH conditions. This lowered the possibility of collinear expression in the feature space and reduced the need for drastic methods. Additionally, we explored the entire interaction matrix TF; factors that had significant positive correlations in occupancies had negative interaction energies between them. This could signify a cooperative binding relationship where the occupancy of a factor recruits the binding of a secondary TF. As such, by modelling the interconnected nature of TFs, we also overcame the problem of collinearity in statistical regression.

Another way to reduce the effects of collinearity would be to leverage the rotation matrix of principal component analysis. By transforming the data onto the principal basis, we can project variances in the binding data onto an orthogonal reduced space. This makes

regression and feature selection in projection space easy to perform, and we leveraged this in chapter 4 to explore high orders of regression. Our second-order PCA regression was able to recover parameter estimates for second-order spatial interactions ($\sim$200000 features) that were reflective of previous experimental work with limited experimental data ($\sim$4000 samples). The reduction in both dimensionality and the need for training data makes PCA an attractive approach. However, one must also consider that the learned parameter estimates are heuristic in nature and can change depending on the number of principal components included. Proper selection of principal components is vital for PCA and can change depending on the data and model. One possible solution and an idea for future work would be to combine sampling methods such as the one shown in chapter 2 and the selection of principal components. To achieve this, one can implement a sampling algorithm on the eigenvectors of the covariance matrix and train an ensemble of base estimators using the selected projections. One can then weight each eigenvector based on its contribution to the base estimators or rely on Shapley values from game theory to fairly distribute contributions. This combined approach may lead to a more robust representation of PCA regression that is less dependent on the selection of eigenvectors. Overall, while collinearity remains a delicate problem for transcriptional analysis, we showed several approaches that can reduce its effects and still uncover distinct parts of the regulation picture.

Throughout this thesis, several transcriptional factors have consistently been identified as important drivers of DHT-dependent transcription. As expected, all methods in this thesis found AR when treated with androgen to be the most significant driver of transcription in prostate cancer cells. This is crucial, as it shows that all of our methods are consistent with biological results and are capable of modelling the underlying regulation picture. The consensus among biologists is that AR is the crucial driver for this type of behaviour, and one should view the AR prediction as a control. Models that were incapable of identifying AR as the crucial factor might be unsuitable for this type of analysis. In terms of secondary drivers of enhancer activity, several factors were consistently identified to be important for AR-mediation. For instance, in both the ensemble and mean-field models, PIAS1, MED1, H3k27ac, RUNX1, and ARID1A were identified to be important for DHT-driven enhancer behaviour. This was further confirmed in the PCA spatial model where RUNX1 and H3k27ac were among the top spatial contributors to enhancer activity.

By modelling TF regulation using a variety of approaches, we were also able to tease out the potential roles of TFs. For example, in the mean-field methodology, we were able to show that the loss of MED1 leads to the greatest loss of enhancer activity outside of AR. This in combination with correlated occupancy with transcriptional output in both DHT and EtOH conditions may suggest MED1 has a role in mediating transcription in the absence of androgen and the androgen receptor. Furthermore, in both the mean-field and ensemble learning methodologies, we predicted RUNX1 to be another crucial factor for DHT-dependent enhancer activity. While our method cannot pinpoint the exact role of

RUNX1, second-order PCA regression revealed that both RUNX1 and AR demonstrate a centralized and on-site interaction with each other. This is distinct from the offsite interaction that AR had with H3k27ac and would suggest the formation/sharing of a similar co-factor complex between AR and RUNX1.

While our approaches agreed on a majority of the TFs that were important for enhancer behaviour, there were a few factors that were inconclusive among our methods. For instance, while HOXB13 occupancy when treated with DHT was ranked highly in both the ensemble learning and PCA regression models, theoretical mutation of HOXB13 in the mean-field model did not result in a significant drop in enhancer ability. As stated in chapter 3, this might be due to the antagonistic binding relationship between HOXB13 and RUNX1. This might lead to significantly higher RUNX1 binding in absence of HOX13. Biologically, it has been shown that HOXB13 interacts positively with AR, and a few of our models support that sentiment. However, it is also possible that due to the interaction between RUNX1 and HOX13, the internal relationship between HOXB13 and enhancer behaviour is more complicated than we first expected.

Lastly, while we are able to generate predictions of possible roles for crucial factors, I stress that these are still predictions and would require additional experimental validation. Our approaches leveraged modern-day sequencing data as well as innovation from machine learning to biophysics to provide the best possible estimates of the underlying mechanisms of transcription. Due to the inherent highly dimensional and collinear nature of transcription regulation, individual roles of transcription factors are difficult to tease out. This will hopefully provide the foundation for later experimental work that can explicitly examine how these crucial transcription factors are affecting AR-mediated transcription in prostate cancer.

Bibliography

1. Bintu, L. *et al.* Transcriptional regulation by the numbers: Models. *Current Opinion in Genetics and Development* **15,** 116–124. arXiv: 0412011 [q-bio] (2005).
2. Spitz, F. & Furlong, E. E. Transcription factors: From enhancer binding to developmental control. *Nature Reviews Genetics* **13,** 613–626. ISSN: 14710056 (2012).
3. Riethoven, J. J. Regulatory regions in DNA: promoters, enhancers, silencers, and insulators. *Methods Mol Biol* **674,** 33–42 (2010).
4. Bateman, J. R., Johnson, J. E. & Locke, M. N. Comparing enhancer action in cis and in trans. *Genetics* **191,** 1143–1155. ISSN: 00166731 (2012).
5. De Mendoza, A. *et al.* Transcription factor evolution in eukaryotes and the assembly of the regulatory toolkit in multicellular lineages. *Proceedings of the National Academy of Sciences of the United States of America* **110.** ISSN: 00278424 (2013).
6. He, H. H. *et al.* Nucleosome dynamics define transcriptional enhancers. *Nature Genetics* **42,** 343–347. ISSN: 10614036 (2010).
7. Sur, I. & Taipale, J. The role of enhancers in cancer. *Nature Reviews Cancer* **16,** 483–493. ISSN: 14741768 (2016).
8. Tewari, A. K. *et al.* Chromatin accessibility reveals insights into androgen receptor activation and transcriptional specificity. *Genome Biology* **13.** ISSN: 1474760X (2012).
9. Yu, J. *et al.* An Integrated Network of Androgen Receptor, Polycomb, and TMPRSS2-ERG Gene Fusions in Prostate Cancer Progression. *Cancer Cell* **17,** 443–454. ISSN: 15356108. http://dx.doi.org/10.1016/j.ccr.2010.03.018 (2010).
10. The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489,** 57–74 (2012).
11. Park, P. J. ChIP-seq: Advantages and challenges of a maturing technology. *Nature Reviews Genetics* **10,** 669–680. ISSN: 14710056 (2009).
12. Roadmap Epigenomics Consortium *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518,** 317–329. ISSN: 14764687 (2015).
13. Singh, A. A. *et al.* Optimized ChIP-seq method facilitates transcription factor profiling in human tumors. *Life Science Alliance* **2,** 1–12. ISSN: 25751077 (2019).
14. Ernst, J. *et al.* Genome-scale high-resolution mapping of activating and repressive nucleotides in regulatory regions. *Nature Biotechnology* **34,** 1180–1190. ISSN: 15461696 (2016).
15. Muerdter, F., Boryń, Ł. M. & Arnold, C. D. STARR-seq - Principles and applications. *Genomics* **106,** 145–150. ISSN: 10898646 (2015).

16. Muerdter, F. *et al.* Resolving systematic errors in widely used enhancer activity assays in human cells. *Nature Methods* **15,** 141–149. ISSN: 15487105 (2018).

17. Phillips, R. *et al.* Figure 1 Theory Meets Figure 2 Experiments in the Study of Gene Expression. *Annual Review of Biophysics* **48,** 121–163. ISSN: 19361238. arXiv: 1812.11627 (2019).

18. Barnes, S. L. *et al.* Mapping DNA sequence to transcription factor binding energy in vivo. *PLoS Computational Biology* **15,** 1–29. ISSN: 15537358 (2019).

19. Ching, T. *et al.* Opportunities and obstacles for deep learning in biology and medicine. *Journal of the Royal Society Interface* **15.** ISSN: 17425662 (2018).

20. Liu, Y., Barr, K. & Reinitz, J. Fully interpretable deep learning model of transcriptional control. *Bioinformatics (Oxford, England)* **36,** i499–i507. ISSN: 13674811 (2020).

21. Anastassiou, G. *Intelligent Systems: Approximation by Artificial Neural Networks* ISBN: 9788578110796 (2011).

22. Zhou, D. X. Universality of deep convolutional neural networks. *Applied and Computational Harmonic Analysis* **48,** 787–794. ISSN: 1096603X. arXiv: 1805.10769. https://doi.org/10.1016/j.acha.2019.06.004 (2020).

23. Efron, B. & Tibshirani, R. *An Introduction to the Bootstrap* eng. ISBN: 9780412042317 (CRC Press LLC, Boca Raton, 1994).

24. Hastie, T., Tibshirani, R. & Friedman, J. *Elements of Statistical Learning* eng. ISBN: 0387848576 (Springer, New York, 2009).

25. Tibshirani, R. Regression Shrinkage and Selection Via the Lasso. eng. *Journal of the Royal Statistical Society. Series B, Methodological* **58,** 267–288. ISSN: 0035-9246 (1996).

26. Sharma, N. L. *et al.* The Androgen Receptor Induces a Distinct Transcriptional Program in Castration-Resistant Prostate Cancer in Man. *Cancer Cell* **23,** 35–47. ISSN: 15356108. http://dx.doi.org/10.1016/j.ccr.2012.11.010 (2013).

27. Rawla, P. Epidemiology of Prostate Cancer. *World Journal of Oncology* **32,** 2–4. ISSN: 09281258 (2008).

28. Heinlein, C. A. & Chang, C. Androgen receptor in prostate cancer. *Endocrine Reviews* **25,** 276–308. ISSN: 0163769X (2004).

29. Pomerantz, M. M. *et al.* The androgen receptor cistrome is extensively reprogrammed in human prostate tumorigenesis. *Nature Genetics* **47,** 1346–1351. ISSN: 15461718 (2015).

30. Robinson, J. L. *et al.* Elevated levels of FOXA1 facilitate androgen receptor chromatin binding resulting in a CRPC-like phenotype. *Oncogene* **33,** 5666–5674. ISSN: 14765594 (2014).

31. Rodriguez-Bravo, V. *et al.* The role of GATA2 in lethal prostate cancer aggressiveness. *Nature Reviews Urology* **14,** 38–48 (2017).

32. Stelloo, S. *et al.* Integrative epigenetic taxonomy of primary prostate cancer. *Nature Communications* **9.** ISSN: 20411723. http://dx.doi.org/10.1038/s41467-018-07270-2 (2018).

33. Zhang, Z. *et al.* An AR-ERG transcriptional signature defined by longrange chromatin interactomes in prostate cancer cells. *Genome Research* **29,** 223–235. ISSN: 15495469 (2019).

34. Andersson, R. *et al.* An atlas of active enhancers across human cell types and tissues. *Nature* **507,** 455–461. ISSN: 14764687 (2014).

35. Boyle, A. P. *et al.* High-Resolution Mapping and Characterization of Open Chromatin across the Genome. *Cell* **132,** 311–322. ISSN: 00928674 (2008).

36. Heintzman, N. D. *et al.* Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nature Genetics* **39,** 311–318. ISSN: 10614036 (2007).

37. Inoue, F. *et al.* A systematic comparison reveals substantial differences in chromosomal versus episomal encoding of enhancer activity. *Genome Research* **27,** 38–52. ISSN: 15495469 (2017).

38. Zhang, T. *et al.* Histone H3K27 acetylation is dispensable for enhancer activity in mouse embryonic stem cells. *Genome Biology* **21,** 1–7. ISSN: 1474760X (2020).

39. Banerji, J., Rusconi, S. & Schaffner, W. Expression of a β-globin gene is enhanced by remote SV40 DNA sequences. *Cell* **27,** 299–308. ISSN: 00928674 (1981).

40. Mladenova, V., Mladenov, E. & Russev, G. Organization of plasmid dna into nucleosome-like structures after transfection in eukaryotic cells. *Biotechnology and Biotechnological Equipment* **23,** 1044–1047. ISSN: 13102818 (2009).

41. Arnold, C. D. *et al.* Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science* **339,** 1074–1077. ISSN: 10959203 (2013).

42. Quigley, D. A. *et al.* Genomic Hallmarks and Structural Variation in Metastatic Prostate Cancer. *Cell* **174,** 758–769.e9. ISSN: 10974172 (2018).

43. Takeda, D. Y. *et al.* A Somatically Acquired Enhancer of the Androgen Receptor Is a Noncoding Driver in Advanced Prostate Cancer. *Cell* **174,** 422–432.e13. ISSN: 10974172 (2018).

44. Viswanathan, S. R. *et al.* Structural Alterations Driving Castration-Resistant Prostate Cancer Revealed by Linked-Read Genome Sequencing. *Cell* **174,** 433–447.e19. ISSN: 10974172. https://doi.org/10.1016/j.cell.2018.05.036 (2018).

45. Mazrooei, P. *et al.* Cistrome Partitioning Reveals Convergence of Somatic Mutations and Risk Variants on Master Transcription Regulators in Primary Prostate Tumors. *Cancer Cell* **36,** 674–689.e6. ISSN: 18783686. https://doi.org/10.1016/j.ccell.2019.10.005 (2019).

46. Morova, T. *et al.* Androgen receptor-binding sites are highly mutated in prostate cancer. *Nature Communications* **11.** ISSN: 20411723. http://dx.doi.org/10.1038/s41467-020-14644-y (2020).

47. Oki, S. *et al.* ChIP -Atlas: a data-mining suite powered by full integration of public Ch IP -seq data. *EMBO reports* **19,** 1–10. ISSN: 1469-221X (2018).

48. Diaz, A. *et al.* Normalization, bias correction, and peak calling for ChIP-seq. *Statistical applications in genetics and molecular biology* **11.** ISSN: 15446115 (2012).

49. Liu, Y. *et al.* Functional assessment of human enhancer activities using whole-genome STARR-sequencing. *Genome Biology* **18,** 1–13. ISSN: 1474760X (2017).

50. Nacht, A. S. *et al.* C/EBPα mediates the growth inhibitory effect of progestins on breast cancer cells. *The EMBO Journal* **38,** 1–22. ISSN: 0261-4189 (2019).

51. Palaniappan, M. *et al.* The genomic landscape of estrogen receptor α binding sites in mouse mammary gland. *PLoS ONE* **14,** 1–22. ISSN: 19326203 (2019).

52. Vockley, C. M. *et al.* Direct GR Binding Sites Potentiate Clusters of TF Binding across the Human Genome. *Cell* **166,** 1269–1281.e19. ISSN: 10974172. `http://dx.doi.org/10.1016/j.cell.2016.07.049` (2016).

53. Barakat, T. S. *et al.* Functional Dissection of the Enhancer Repertoire in Human Embryonic Stem Cells. *Cell Stem Cell* **23,** 276–288.e8. ISSN: 18759777 (2018).

54. Toropainen, S. *et al.* SUMO ligase PIAS1 functions as a target gene selective androgen receptor coregulator on prostate cancer cell chromatin. *Nucleic Acids Research* **43,** 848–861. ISSN: 13624962 (2015).

55. Hoerl, A. E. & Kennard, R. W. Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics* **12,** 55–67. eprint: `https://www.tandfonline.com/doi/pdf/10.1080/00401706.1970.10488634`. `https://www.tandfonline.com/doi/abs/10.1080/00401706.1970.10488634` (1970).

56. Craney, T. A. & Surles, J. G. Model-dependent variance inflation factor cutoff values. *Quality Engineering* **14,** 391–403. ISSN: 08982112 (2002).

57. Dormann, C. F. *et al.* Collinearity: A review of methods to deal with it and a simulation study evaluating their performance. *Ecography* **36,** 27–46. ISSN: 16000587 (2013).

58. Rodríguez-Pérez, R. & Bajorath, J. Interpretation of machine learning models using shapley values: application to compound potency and multi-target activity predictions. *Journal of Computer-Aided Molecular Design* **34,** 1013–1026. ISSN: 15734951. `https://doi.org/10.1007/s10822-020-00314-0` (2020).

59. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26,** 589–595. ISSN: 13674803 (2010).

60. Ramírez, F. *et al.* deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic acids research* **44,** W160–W165. ISSN: 13624962 (2016).

61. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology* **15,** 1–21. ISSN: 1474760X (2014).

62. Wickham, H. ggplot2: Elegant Graphics for Data Analysis. *Journal of Statistical Software* **35.** ISSN: 1548-7660 (2010).

63. Lerdrup, M. *et al.* An interactive environment for agile analysis and visualization of ChIP-sequencing data. *Nature Structural and Molecular Biology* **23,** 349–357. ISSN: 15459985 (2016).

64. Quinlan, A. R. *BEDTools: The Swiss-Army tool for genome feature analysis* 11.12.1–11.12.34. ISBN: 0471250953 (2014).

65. Langmead, B. *et al.* Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology* **10.** ISSN: 14747596 (2009).

66. Kent, W. J. *et al.* BigWig and BigBed: Enabling browsing of large distributed datasets. *Bioinformatics* **26,** 2204–2207. ISSN: 13674803 (2010).

67. Whyte, W. A. *et al.* Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell* **153,** 307–319. ISSN: 00928674. http://dx.doi.org/10.1016/j.cell.2013.03.035 (2013).

68. Chen, X. F. *et al.* Transcriptional regulation and its misregulation in Alzheimer's disease. *Molecular Brain* **6,** 1237–1251. ISSN: 17566606 (2013).

69. Melnikov, A. *et al.* Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. *Nature Biotechnology* **30,** 271–277. ISSN: 10870156 (2012).

70. Palstra, R. J. & Grosveld, F. Transcription factor binding at enhancers: Shaping a genomic regulatory landscape in flux. *Frontiers in Genetics* **3,** 1–12. ISSN: 16648021 (2012).

71. Zabidi, M. A. *et al. Enhancer-core-promoter specificity separates developmental and housekeeping gene regulation* 2015.

72. LeCun, Y., Bengio, Y. & Hinton, G. *Deep Learning* 2018.

73. Zhou, J. & Troyanskaya, O. G. Predicting effects of noncoding variants with deep learning-based sequence model. *Nature Methods* **12,** 931–934. ISSN: 15487105 (2015).

74. Kelley, D. R. *et al.* Sequential regulatory activity prediction across chromosomes with convolutional neural networks. *Genome Research* **28,** 739–750. ISSN: 15495469 (2018).

75. Movva, R. *et al.* Deciphering regulatory DNA sequences and noncoding genetic variants using neural network models of massively parallel reporter assays. *PLoS ONE* **14,** 1–20. ISSN: 19326203 (2019).

76. Schaub, M. A. *et al.* Linking disease associations with regulatory information in the human genome. *Genome Research* **22,** 1748–1759. ISSN: 10889051 (2012).

77. Shrikumar, A., Greenside, P. & Kundaje, A. Learning important features through propagating activation differences. *34th International Conference on Machine Learning, ICML 2017* **7,** 4844–4866. arXiv: 1704.02685 (2017).

78. Wong, F. & Gunawardena, J. Gene Regulation in and out of Equilibrium. *Annual Review of Biophysics* **49,** 199–226. ISSN: 19361238 (2020).

79. Jones, D. L., Brewster, R. C. & Phillips, R. Promoter architecture dictates cell-to-cell variability in gene expression. *Science* **346,** 1533–1536. ISSN: 10959203 (2014).

80. Vilar, J. M. & Saiz, L. Reliable prediction of complex phenotypes from a modular design in free energy space: An extensive exploration of the lac operon. *ACS Synthetic Biology* **2,** 576–586. ISSN: 21615063 (2013).

81. Stelloo, S. *et al.* Endogenous androgen receptor proteomic profiling reveals genomic subcomplex involved in prostate tumorigenesis. *Oncogene* **37,** 313–322. www.nature.com/onc (2017).

82. Ward, J. H. Hierarchical Grouping to Optimize an Objective Function. eng. *Journal of the American Statistical Association* **58,** 236–244. ISSN: 0162-1459.

83. Mei, S. *et al.* Cistrome Data Browser: A data portal for ChIP-Seq and chromatin accessibility data in human and mouse. *Nucleic Acids Research* **45,** D658–D662. ISSN: 13624962 (2017).

84. Huang, C.-C. F. *et al.* Functional mapping of androgen receptor enhancer activity. *bioRxiv,* 2020.08.18.255232. https://doi.org/10.1101/2020.08.18.255232 (2020).

85. Czerwińska, P., Mazurek, S. & Wiznerowicz, M. The complexity of TRIM28 contribution to cancer. *Journal of Biomedical Science* **24,** 1–14. ISSN: 14230127 (2017).

86. Jangal, M. *et al.* The transcriptional co-repressor TLE3 suppresses basal signaling on a subset of estrogen receptor α target genes. *Nucleic Acids Research* **42,** 11339–11348. ISSN: 13624962. https://www.oncomine.com (2014).

87. Groner, A. C. *et al.* TRIM24 is an oncogenic transcriptional activator in prostate cancer. *Cancer Cell* **29,** 846–858 (2016).

88. Farina, N. H. *et al.* A microRNA/Runx1/Runx2 network regulates prostate tumor progression from onset to adenocarcinoma in TRAMP mice. *Oncotarget* **7,** 70462–70474. ISSN: 19492553 (2016).

89. Jung, C. *et al.* HOXB13 induces growth suppression of prostate cancer cells as a repressor of hormone-activated androgen receptor signaling. *Cancer Research* **64,** 9185–9192. ISSN: 00085472 (2004).

90. Avsec, Ž. *et al.* Base-resolution models of transcription factor binding reveal soft motif syntax. *bioRxiv* (2019).

91. Russo, J. W., Nouri, M. & Balk, S. P. Androgen receptor interaction with mediator complex is enhanced in castration-resistant prostate cancer by cdk7 phosphorylation of med1. *Cancer Discovery* **9,** 1490–1492. ISSN: 21598290 (2019).

92. Hsieh, C. L. *et al.* Enhancer RNAs participate in androgen receptor-driven looping that selectively enhances gene activation. *Proceedings of the National Academy of Sciences of the United States of America* **111,** 7319–7324. ISSN: 10916490 (2014).

93. Sabari, B. R. *et al.* Coactivator condensation at super-enhancers links phase separation and gene control. *Science* **361,** eaar3958. ISSN: 0036-8075 (2018).

94. Chen, Z. *et al.* Phospho-MED1-enhanced UBE2C locus looping drives castration-resistant prostate cancer growth. *EMBO Journal* **30,** 2405–2419. ISSN: 02614189. http://dx.doi.org/10.1038/emboj.2011.154 (2011).

95. Liu, G. *et al.* MED1 mediates androgen receptor splice variant induced gene expression in the absence of ligand. *Oncotarget* **6,** 288–304. ISSN: 19492553 (2015).

96. Wu, D. *et al.* Three-tiered role of the pioneer factor GATA2 in promoting androgen-dependent gene expression in prostate cancer. *Nucleic Acids Research* **42,** 3607–3622. ISSN: 13624962 (2014).

97. Takayama, K. I. *et al.* RUNX1, an androgen- and EZH2-regulated gene, has differential roles in AR-dependent and -independent prostate cancer. *Oncotarget* **6,** 2263–2276. ISSN: 19492553 (2015).

98. Paltoglou, S. *et al.* Novel androgen receptor coregulator GRHL2 exerts both oncogenic and antimetastatic functions in prostate cancer. *Cancer Research* **77,** 3417–3430. ISSN: 15387445 (2017).

99. Reese, R. M., Harrison, M. M. & Alarid, E. T. Grainyhead-like Protein 2: The Emerging Role in Hormone-Dependent Cancers and Epigenetics. *Endocrinology* **160,** 1275–1288. ISSN: 19457170 (2019).

100. Augello, M. A. *et al.* CHD1 Loss Alters AR Binding at Lineage-Specific Enhancers and Modulates Distinct Transcriptional Programs to Drive Prostate Tumorigenesis. *Cancer Cell* **35,** 603–617.e8. ISSN: 18783686. https://doi.org/10.1016/j.ccell.2019.03.001 (2019).

101. Chen, C. S. *et al.* Genetic interaction analysis of TCF7L2 for biochemical recurrence after radical prostatectomy in localized prostate cancer. *International Journal of Medical Sciences* **12,** 243–247. ISSN: 14491907 (2015).

102. Fornes, O. *et al.* JASPAR 2020: Update of the open-Access database of transcription factor binding profiles. *Nucleic Acids Research* **48,** D87–D92. ISSN: 13624962 (2020).