NOVEL COMPUTATIONAL METHODS FOR STUDYING THE ROLE AND INTERACTIONS OF TRANSCRIPTION FACTORS IN GENE REGULATION

TUOMO HARTONEN, MSC

Integrative Life Sciences Doctoral Program
Research Programs Unit
Faculty of Medicine
University of Helsinki
and
Department of Biochemistry
University of Cambridge

Academic dissertation
To be publicly discussed with permission of
the Faculty of Medicine of the University of Helsinki,
in Lecture Hall 2, Biomedicum 1, Haartmaninkatu 8, Helsinki
on 2nd May 2022, at 17.00.

Helsinki 2022

SUPERVISORS:
Jussi Taipale, PhD, Professor, University of Cambridge, University of Helsinki, Karolinska Institute
Teemu Kivioja, PhD, Docent, University of Helsinki

REVIEWERS:
Veli Mäkinen, PhD, Professor, University of Helsinki
Markus Heinonen, PhD, Academy Research Fellow, Aalto University and Finnish Center of AI

OFFICIAL OPPONENT:
Julia Zeitlinger, PhD, Faculty, The Graduate School of the Stowers Institute for Medical Research. Associate Professor, Department of Pathology and Laboratory Medicine, Division of Cancer and Developmental Biology, University of Kansas School of Medicine

The Faculty of Medicine uses the Urkund system (plagiarism recognition) to examine all doctoral dissertations.

ISBN 978-951-51-8007-0 (paperback)
ISBN 978-951-51-8008-7 (PDF)
http://ethesis.helsinki.fi
Unigrafia Oy
Helsinki 2022

I agree with you that the subject is most interesting.
But to express myself in regard to it would
necessitate a concentration of thought which,
in the midst of my present labors, is impossible for me.

— Nikola Tesla, 1908

CONTENTS

ACRONYMS

ADM	Adjacent Dinucleotide Model
AI	Artificial Intelligence
ASB	Allele Specific Binding
ATAC	Assay for Transposase-Accessible Chromatin
ATI	Active TF Identification
AUprc	Area Under precision-recall curve
BaMM	Bayesian Markov Model
bp	base pair
BWA	Burrows-Wheeler Aligner
CACWM	CNN Activity Contribution Weight Matrix
CAGE	Capped Analysis of Gene Expression
ChIA-PET	Chromatin Interaction Analysis by Paired-End Tag Sequencing
ChIP	Chromatin Immunoprecipitation
CNN	Convolutional Neural Network
CRISPR	Clustered Regularly-Interspaced Short Palindromic Repeats
CUT&RUN	Cleavage Under Targets and Release Using Nuclease
DBD	DNA Binding Domain
DL	Deep Learning
DNA	Deoxyribonucleic Acid
ENCODE	ENCyclopedia Of DNA Elements
EPD	Eukaryotic Promoter Database
gAR	genomic Allelic Ratio
GCN	Graph Convolutional neural Network
GEO	Gene Expression Omnibus
GERP	Genomic Evolutionary Rate Profiling
GPU	Grpahics Processing Unit
GWAS	Genome Wide Association Study
HARS	High-Affinity Recognition Sequence
HT-SELEX	High-Throughput Systematic Evolution of Ligands by EXponential enrichment
IDR	Intrinsically Disorcered Region
ISM	In Silico Mutagenesis
KZFP	KRAB-Zinc Finger Protein
Lasso	Least Absolute Shrinkage and Selection Operator
LFC	Locarithmic Fold Change
LR	Linear Regression

MACE	Model-based Analysis of ChIP-Exo
MACS	Model-based Analysis of ChIP-Seq
MAPQ	Mapping Quality
MI	Mutual Information
ML	Machine Learning
MNase	Micrococcal Nuclease
MPRA	Massively Parallel Reporter-gene Assay
MSA	Multiple Sequence Alignment
NMR	Nuclear Magnetic Resonance
ORF	Open Reading Frame
PBM	Protein Binding Microarray
PCR	Polymerase Chain Reaction
PDB	Protein Data Bank
PFM	Position Frequency Matrix
PWM	Position Weight Matrix
RBP	RNA Binding Protein
ReLu	Rectified Linear Unit
RNA	Ribonucleic Acid
RNN	Recurrent Neural Network
SNP	Single Nucleotide Polymorphism
SNV	Single Nucleotidde Variant
SSD	Signal Sensing Domain
STARR	Self-Transcribed Active Regulatory Region
SVM	Support Vector Machine
TAD	Transactivation Domain
TAD	Topologically Associated Domain
TF	Transcription Factor
TPM	Transcripts Per Million
TSS	Transcription Start Site
UMI	Unique Molecular Identifier
WGS	Whole Genome Sequencing

PUBLICATIONS AND AUTHOR'S CONTRIBUTIONS

PUBLICATIONS

PUBLICATION I: Tuomo Hartonen, Biswajyoti Sahu, Kashyap Dave, Teemu Kivioja, and Jussi Taipale. "PeakXus: comprehensive transcription factor binding site discovery from ChIP-Nexus and ChIP-Exo experiments." In: Bioinformatics 32.17 (2016), p. i629-i638, special issue, 15th European Conference on Computational Biology (ECCB 2016).

PUBLICATION II: Biswajyoti Sahu, Tuomo Hartonen, Päivi Pihlajamaa, Bei Wei, Kashyap Dave, Fangjie Zhu, Eevi Kaasinen, Katja Lidschreiber, Michael Lidschreiber, Carsten O Daub, Patrick Cramer, Teemu Kivioja, and Jussi Taipale . "Sequence determinants of human gene regulatory elements." In: Nature Genetics 54 (2022), p. 283-294.

PUBLICATION III: Tuomo Hartonen, Teemu Kivioja, and Jussi Taipale. "PlotMI: Interpretation of pairwise interactions and positional preferences learned by a deep learning model from sequence data." In: Manuscript (2022).

Author's contributions

PUBLICATION I: The Author designed the peak calling and the allele specific binding algorithms together with TK and JT. The Author designed the computational experiments together with JT and TK. The Author implemented all algorithms and performed all analyses. The Author participated into discussions regarding all analyses, results and interpretation of results. The Author wrote the manuscript, and the other authors participated in planning, commenting and editing of the manuscript.

PUBLICATION II: The Author designed the machine learning models and analyses together with JT, BS and TK. The Author performed all machine learning specific data preprocessing, model training and machine learning related analyses. The Author performed all mutual information and genomic enhancer conservation analyses as well as the analysis of the genomic feature overlap in GP5d cells. The Author performed all preprocessing and analyses of the CAGE data. The Author contributed to discussion related to all analyses and results. The manuscript was mainly written by BS and JT, but the Author participated in writing, commenting and editing the text together with other authors. The Author wrote the descriptions of all analyses he performed, as listed above.

PUBLICATION III: The Author designed the machine learning model interpretation method together with JT. The author designed the computational experiments together with JT and TK. The Author implemented all algorithms and performed all analyses. The Author participated into discussions regarding all analyses, results and interpretation of results. The Author wrote the manuscript, and the other authors participated in planning, commenting and editing of the manuscript.

RELATED PUBLICATION: Kimmo Palin, Esa Pitkänen, Mikko Turunen, Biswajyoti Sahu, Päivi Pihlajamaa, Teemu Kivioja, Eevi Kaasinen, Niko Välimäki, Ulrika A Hänninen, Tatiana Cajuso, Mervi Aavikko, Sari Tuupanen, Outi Kilpivaara, Linda van den Berg, Johanna Kondelin, Tomas Tanskanen, Riku Katainen, Marta Grau, Heli Rauanheimo, Roosa-Maria Plaketti, Aurora Taira, Päivi Sulo, Tuomo Hartonen, Kashyap Dave, Bernhard Schmierer, Sandeep Botla, Maria Sokolova, Anna Vähärautio, Kornelia Gladysz, Halit Ongen, Emmanouil Dermitzakis, Jesper Bertram Bramsen, Torben Falck Orntoft, Claus Lindbjerg Andersen, Ari Ristimäki, Anna Lepistö, Laura Renkonen-Sinisalo, Jukka-Pekka Mecklin, Jussi Taipale, Lauri A Aaltonen. "Contribution of allelic imbalance to colorectal cancer." In: Nature communications 9.1 (2018), p. 1-9.

Authors contribution to the related publication

The Author analyzed the ChIP-nexus data and contributed to writing of the manuscript.

ABSTRACT

Regulation of which genes are expressed and when enables the existence of different cell types sharing the same genetic code in their DNA. Erroneously functioning gene regulation can lead to diseases such as cancer. Gene regulatory programs can malfunction in several ways. Often if a disease is caused by a defective protein, the cause is a mutation in the gene coding for the protein rendering the protein unable to perform its functions properly. However, protein-coding genes make up only about 1.5% of the human genome, and majority of all disease-associated mutations discovered reside outside protein-coding genes. The mechanisms of action of these non-coding disease-associated mutations are far more incompletely understood.

Binding of transcription factors (TFs) to DNA controls the rate of transcribing genetic information from the coding DNA sequence to RNA. Binding affinities of TFs to DNA have been extensively measured *in vitro*, and the genome-wide binding locations and patterns of TFs have been mapped in dozens of cell types. Despite this, our understanding of how TF binding to regulatory regions of the genome, promoters and enhancers, leads to gene expression is not at the level where gene expression could be reliably predicted based on DNA sequence only.

In this work, we develop and apply computational tools to analyze and model the effects of TF-DNA binding. We also develop new methods for interpreting and understanding deep learning based models trained on biological sequence data. In biological applications, the ability to understand how machine learning models make predictions is as, or even more important as raw predictive performance. This has created a demand for approaches helping researchers extract biologically meaningful information from deep learning model predictions.

We develop a novel computational method for determining TF binding sites genome-wide from recently developed high-resolution ChIP-exo and ChIP-nexus experiments. We demonstrate that our method performs similarly or better than previously published methods while making less assumptions about the data. We also describe an improved algorithm for calling allele-specific TF-DNA binding.

We utilize deep learning methods to learn features predicting transcriptional activity of human promoters and enhancers. The deep learning models are trained on massively parallel reporter gene assay (MPRA) data from human genomic regulatory elements, designed regulatory elements and promoters and enhancers selected from totally random pool of synthetic input DNA. This unprecedentedly large set of measurements of human gene regulatory element activities, in total more than 100 times the size of the human genome, allowed us to train models that were able to predict genomic transcription start site positions more accurately than models trained on genomic promoters, and to correctly predict effects of disease-associated promoter variants. We also found that interactions between promoters and local classical enhancers are non-specific in nature. The MPRA data integrated with extensive epigenetic measurements supports existence of three different classes of enhancers: classical enhancers, closed chromatin enhancers and chromatin-dependent enhancers. We also show that TFs can be divided into four different, non-exclusive classes based on their activities: chromatin opening, enhancing, promoting and TSS determining TFs.

Interpreting the deep learning models of human gene regulatory elements required application of several existing model interpretation tools as well as developing new approaches. Here, we describe two new methods for visualizing features and interactions learned by deep learning models. Firstly, we describe an algorithm for testing if a deep learning model has learned an existing binding motif of a TF. Secondly, we visualize mutual information between pairwise k-mer distributions in sample inputs selected according to predictions by a machine learning model. This method highlights pairwise and positional dependencies learned by a machine learning model. We demonstrate the use of this model-agnostic approach with classification and regression models trained on DNA, RNA and amino acid sequences.

TIIVISTELMÄ

Monet eliöt koostuvat useista erilaisista solutyypeistä, vaikka kaikissa näiden eliöiden soluissa onkin sama DNA-koodi. Geenien ilmentymisen säätely mahdollistaa erilaiset solutyypit. Virheellisesti toimiva säätely voi johtaa sairauksiin, esimerkiksi syövän puhkeamiseen. Jos sairauden aiheuttaa viallinen proteiini, on syynä usein mutaatio tätä proteiinia koodaavassa geenissä, joka muuttaa proteiinia siten ettei se enää pysty toimittamaan tehtäväänsä riittävän hyvin. Kuitenkin vain 1,5 % ihmisen genomista on proteiineja koodaavia geenejä. Suurin osa kaikista löydetyistä sairauksiin liitetyistä mutaatioista sijaitsee näiden ns. koodaavien alueiden ulkopuolella. Ei-koodaavien sairauksiin liitetyiden mutaatioiden vaikutusmekanismit ovat yleisesti paljon huonommin tunnettuja, kuin koodaavien alueiden mutaatioiden.

Transkriptiotekijöiden sitoutuminen DNA:han säätelee transkriptiota, eli geeneissä olevan geneettisen informaation lukemista ja muuntamista RNA:ksi. Transkriptiotekijöiden sitoutumista DNA:han on mitattu kattavasti *in vitro*-olosuhteissa, ja monien transkriptiotekijöiden sitoutumiskohdat on mitattu genominlaajuisesti useissa eri solutyypeissä. Tästä huolimatta ymmärryksemme siitä miten transkriptioitekijöiden sitoutuminen genomin säätelyelementteihin, eli promoottoreihin ja vahvistajiin, johtaa geenien ilmentymiseen ei ole sellaisella tasolla, että voisimme luotettavasti ennustaa geenien ilmentymistä pelkästään DNA-sekvenssin perusteella.

Tässä työssä kehitämme ja sovellamme laskennallisia työkaluja transkriptiotekijöiden sitoutumisesta johtuvan geenien ilmentymisen analysointiin ja mallintamiseen. Kehitämme myös uusia menetelmiä biologisella sekvenssidatalla opetettujen syväoppimismallien tulkitsemiseksi. Koneoppimismallin tekemimen ennusteiden ymmärrettävyys on biologisissa sovelluksissa yleensä yhtä tärkeää, ellei jopa tärkeämpää kuin pelkkä raaka ennustetarkkuus. Tämä on synnyttänyt tarpeen uusille menetelmille jotka auttavat tutkijoita louhimaan biologisesti merkityksellistä tietoa syväoppimismallien ennusteista.

Kehitimme tässä työssä uuden laskennallisen työkalun, jolla voidaan määrittää transkriptiotekijöiden sitoutumiskohdat genominlaajuisesti käyttäen mittausdataa hiljattain kehitetyistä korkearesoluutioisista ChIP-exo ja ChIP-nexus kokeista. Näytämme, että kehittämämme menetelmä suoriutuu paremmin, tai vähintään yhtä hyvin kuin aiemmin julkaistut menetelmät tehden näitä vähemmä oletuksia signaalin muodosta. Esittelemme myös parannellun algoritmin transkriptiotekijöiden alleelispesifin sitoutumisen määrittämiseksi.

Käytämme syväoppimismenetelmiä oppimaan mitkä ominaisuudet ennustavat ihmisen promoottori- ja voimistajaelementtien aktiivisuutta. Nämä syväoppimismallit on opetettu valtavien rinnakkaisten reportterigeenikokeiden datalla ihmisen genomisista säätelyelementeistä, sekä aktiivisista promoottoreista ja voimistajista, jotka ovat valikoituneet satunnaisesta joukosta synteettisiä DNA-sekvenssejä. Tämä ennennäkemättömän laaja joukko mittauksia ihmisen säätelyelementtien aktiivisuudesta - yli satakertainen määrä DNA-sekvenssiä ihmisen genomiin verrattuna - mahdollisti transkription aloituskohtien sijainnin ennustamisen ihmisen genomissa tarkemmin kuin ihmisen genomilla opetetut mallit. Nämä mallit myös ennustivat oikein sairauksiin liitettyjen mutaatioiden vaikutukset ihmisen promoottoreilla.

Tuloksemme näyttivät, että vuorovaikutukset ihmisen promottorien ja klassisten paikallisten voimistajien välillä ovat epäspesifejä. MPRA-data, integroituna kattavien epigeneettisten mittausten kanssa mahdollisti voimistajaelementtien jaon kolmeen luokkaan: klassiset,

suljetun kromatiinin, ja kromatiinista riippuvat voimistajat. Tutkimuksemme osoitti, että transkriptiotekijät voidaan jakaa neljään, osittain päällekkäiseen luokkaan niiden aktiivisuuksien perusteella: kromatiinia avaaviin, voimistaviin, promotoiviin ja trankription aloituskohdan määrittäviin transkriptiotekijöihin.

Ihmisen genomin säätelyelementtejä kuvaavien syväoppimismallien tulkitseminen vaati sekä olemassaolevien menetelmien soveltamista, että uusien kehittämistä. Kehitimme tässä työssä kaksi uutta menetelmää syväoppimismallien oppimien muuttujien ja niiden välisten vuorovaikutusten visualisoimiseksi. Ensin esittelemme algoritmin, jonka avulla voidaan testata onko syväoppimismalli oppinut jonkin jo tunnetun transkriptiotekijän sitoutumishahmon. Toiseksi, visualisoimme positiokohtaisten k-meerijakaumien keskeisinformaatiota sekvensseissä, jotka on valittu syväoppimismallin ennusteiden perusteella. Tämä menetelmä paljastaa syväoppimismallin oppimat parivuorovaikutukset ja positiokohtaiset riippuvuudet. Näytämme, että kehittämämme menetelmä on mallin arkkitehtuurista riippumaton soveltamalla sitä sekä luokittelijoihin, että regressiomalleihin jotka on opetettu joko DNA-, RNA-, tai aminohapposekvenssidatalla.

1

INTRODUCTION

The human body contains numerous different cell types [1] and they all share the same genetic code that is guiding their organization and function. The vast spectrum of different functions of human cells, from oocytes to neurons and from skeletal muscle cells to photosensitive retinal ganglion cells of the eye, is achieved through regulation of what part of the "genetic blueprint" encoded into the DNA is read and how it is interpreted. Regulation of gene expression is a complex, multi-step process involving virtually all steps required in producing functional proteins starting from the genetic code. Transcription factors (TFs), the focus of this study, are proteins that recognize specific DNA-sequences through their DNA-binding domains (DBDs), a domain unique for TFs [2]. How TFs read the regulatory DNA sequences and control gene expression, is understood on a very conceptual level only.

The words of the regulatory code, the sequences that bind given TFs have been measured *in vitro* (e.g. [3]), but the mechanisms by which the regulatory sentences - combinations of regulatory DNA elements controlling the expression of a given gene - are constructed at a systems-level, are not well understood. This is in spite of major efforts in determining the binding locations of TFs *in vivo* in multiple cell types (e.g. [4]). Detailed understanding of how TFs regulate gene expression is of utmost importance when trying to understand the mechanisms of action of non-coding disease-associated variants. Majority of the disease-associated variants are located outside of the coding sequence, but their mechanisms of action are far more incompletely known than those variants that hit the protein-coding genes [5].

The computational methods developed and applied in this work concentrate on studying one of the early steps in the chain of processes involved in regulation of gene expression; initiation of transcription. Transcription is the process where specific proteins, known as RNA polymerases, synthesize RNA from the DNA templates located at the coding regions of the genome. DNA in the cell nucleus is wound around nucleosomes and packed into a tight structure where the coding genomic sequences are inaccessible to the large protein complexes required for initiation of transcription, such as Mediator and the RNA polymerases. Making this tightly-packed, chromatin-bound DNA accessible for the transcriptional machinery is achieved by a combination of different processes such as post-translational modification of the histone proteins forming the nucleosomes, demethylation of the DNA and binding of TFs to DNA.

In addition to DBDs, TFs can also contain trans-activation domains (TADs) harboring binding sites for other proteins, or signal-sensing domains (SSDs) that bind to external (non-protein) ligands. TFs regulate transcription by binding to regulatory regions of the genome, to distal enhancers, and to gene proximal promoters. Binding affinities of TFs have been extensively measured *in vitro* (e.g. [3, 6–8]) and their binding locations in different human cell types cataloged *in vivo* (e.g. [4, 9, 10]). In Publication I we develop a novel computational tool for separating true TF-DNA binding event signals from noise and for accurately detecting the binding sites of TFs genome-wide in high-resolution ChIP-nexus [11] and ChIP-exo experiments [12].

Recently developed massively parallel reporter gene assays (MPRAs) such as STARR-seq [13] have allowed researchers to start measuring the effects of TF binding to gene expression in humans more directly [14–19]. However, the limited diversity of sequences in the human genome and evolutionary background from processes not directly related to regulation of transcription can lead to models of transcriptional regulation trained based on human genomic DNA to over-fit [20]. Recently, completely random synthetic DNA was used to probe the promoter activities of TFs in yeast [21]. In Publication II, we build on this idea, and use an array of MPRA experiments to test enhancer and promoter activities of human genomic sequences, designed regulatory elements and elements enriched from completely random synthetic input DNA.

Studies of the three dimensional organization of the genome using methods such as Hi-C [22] and ChIA-PET [23] have increased our understanding of the longer-range interactions in the genome, for example by revealing the existence of topologically associated domains (TADs) [24] in mammalian genomes. TADs are regions of the genome with more frequent interactions withing the TAD than between the TAD and outside regions. The functions of TADs are still debated due to some conflicting results [25], but disruption of some TAD boundaries has been shown to be associated with wide range of diseases (see e.g. [26]).

Despite all these advances, we do not currently understand the activities and interactions of TFs driving gene expression well enough to be able to predict gene expression from the DNA sequence alone. Examples of strict transcriptional regulatory logic, such as the Interferon enhanceosome [27], exist, but also enhancers and promoters with looser regulatory logic have been described (see for example [28, 29]). Also the mechanisms of how contacts between promoters and enhancers are established are incompletely understood. Recently, liquid-liquid phase separation, where so called intrinsically disordered regions (IDRs) of trans-activation domains of TFs mediate formation of phase-separated compartments containing higher concentrations of TFs and other proteins, has been studied as at least one of the possible mechanisms [30–33].

Some examples exist in the literature, where researchers have been able to show disease-associated mutations creating or destroying binding sites for TFs (e.g. [34–37]). In Publication I we present an improved pipeline for studying allele-specific binding (ASB) of TFs from ChIP-exo/nexus experiments and in Publication II we train deep learning models that are able to predict and explain the effects of known disease-associated variants. Being able to predict and explain the effects of non-coding variants is one of the main goals of modern deep learning-based models of gene regulation.

The applications of deep learning and artificial intelligence (AI) have sky-rocketed in the last years in all areas of society. According to the AI Index Report for 2021 [38], for example both corporate investment in AI and the number of peer-reviewed scientific AI-related publications beat the previous annual records by large margins. The beginning of the so called "deep learning revolution" is usually dated to around 2011-2012, when fast implementations of convolutional neural network (CNN) algorithms on graphics processing units (GPUs) allowed breakthroughs in benchmark problems such as first time superhuman performance of a machine learning model in a visual recognition contest [39]. Several inventions and improvements on the deep learning model architectures have since been described, such as max pooling [40] or batch normalization [41], that allow faster training of deeper networks.

The breakthrough application of deep learning, and more specifically CNNs, to model biological sequence data was DeepBind [42], that pioneered the use of CNNs in predicting binding sites of TFs and RNA binding proteins (RBPs). Since then, deep learning has been

successfully applied to for example learning the regulatory code of the accessible regions of the human genome [43] and famously, to state-of-the-art *in silico* protein folding starting from multiple sequence alignments (MSAs) of protein families [44, 45]. In Publication II, we train CNNs on unbiased MPRA data probing in aggregate more than hundred times larger sequence space than what is available for models trained on human genomic data. This allows us to for example train a model of human promoters that can correctly identify and predict the effects of known disease-associated variants and to predict the positions of active transcription start sites in the human genome more accurately than similar models trained on the genome itself.

Deep learning has been traditionally viewed as a "black box" method, that offers superior performance at the cost of model interpretability compared to simpler methods like linear regression, where the contributions of features can be read directly from the regression coefficients. Deep learning model interpretation has been a hot research topic of late, especially in context of biological research where model interpretation is often times as important as model performance in prediction. The early interpretation of CNNs trained on biological data relied on designing the network architecture so that visualization of the first layer filters reveals the types of sequence features learned by the model or testing all possible single position variants of an input sample and scoring them with the model [42, 43]. In Publication II, we develop novel approaches for analyzing features learned by CNNs trained on DNA sequence data and for comparing the features used by simpler models to features learned by the CNNs. In Publication III, we develop a general tool for interpreting and visualizing pairwise dependencies and positional preferences learned by virtually any type of machine learning models trained on sequence data.

In the following sections, methods and concepts central to this study are introduced in more detail.

1.1 High-throughput methods for studying transcription factor binding to DNA

On a high level, high-throughput methods for studying binding of TFs to DNA can be divided into *in vitro* and *in vivo* methods. The *in vitro* methods, such as protein binding microarrays [46] (PBMs) or HT-SELEX [47] measure binding of purified TFs or DNA-binding domains (DBDs) of TFs to designed or random DNA sequence probes, whereas *in vivo* experiments such as ChIP-seq [48] or CUT&RUN [49] identify the positions bound by a TF of interest genome-wide. The aim of the *in vitro* experiments is to measure the binding affinities of the TFs towards DNA sequences and to build biochemical models of TF-DNA binding that explain how TFs find their correct binding sites in the genome. The *in vivo* experiments aim at pinpointing the exact binding locations of the TF of interest in the cell type of interest, but can also be used to study and model the DNA sequences bound by the TF of interest in the cellular conditions.

The *in vitro* TF-DNA binding experiments are based on the idea of introducing a designed or random set of DNA sequences to the TF of interest and observing which sequences the TF binds. This idea was introduced already in the 1980s for studying the common features in sequences bound by a single TF [50]. The development of DNA microarrays during the 1990s [51, 52] paved way for development of PBMs [46, 53] allowing high-throughput characterization of the sequence specificities of TF-DNA interactions. Shortly, in PBMs, a robot is used to print a library of hundreds of thousands of double-stranded DNA oligonucleotides on a glass slide. Then, an epitope-tagged TF of interest is introduced, and

the TF then binds to those DNA oligonucleotides that contain a sequence recognized by the DBD of the TF. The microarray slide is then washed to get rid of non-specific binding, and the TF-bound sequences are retrieved using the epitope-tag.

Development of HT-SELEX (high-throughput systematic evolution of ligands by exponential enrichment) [47] allowed probing of larger sets of longer DNA oligonucleotides, and with smaller amounts of purified protein needed than PBMs. SELEX was originally described already in 1990 [54, 55]. The SELEX protocol starts with a synthesis of a large library of DNA oligonucleotides with a randomized middle region flanked by sequencing primers. The library is then exposed to a TF (or other ligand) of interest. After washing and elution, the resulting population of sequences more specific to the ligand of interest is amplified by polymerase chain reaction (PCR). The resulting amplified population is then sequenced, and as the name suggests, in HT-SELEX the this is done using the modern high-throughput ("next generation") sequencers. These elution and washing cycles are usually repeated multiple times starting from the enriched sequence population from the previous cycle. The versatile HT-SELEX method has subsequently been successfully applied to e.g. measure binding specificities of hundreds of human monomer and dimer TFs [3, 56], the effect of cytosine methylation on TF-DNA binding [8] and the binding affinities of TFs to nucleosome-bound DNA [57].

Many of the current high-throughput, genome-wide *in vivo* experimental methods for studying TF binding rely on chromatin immunoprecipitation (ChIP), first described in the 1980s [58]. In ChIP assays, DNA-binding proteins are cross-linked to DNA in living cells usually using formaldehyde (as pioneered in [59]). The DNA cross-linked to the proteins is then sheared into fragments and the DNA bound by protein of interest is selected using an antibody specific to the protein. Similarly to PBMs, microarray techniques were used to develop a high-throughput version of ChIP, known as ChIP-on-ChIP [60]. In ChIP-on-ChIP, the sequences selected using the antibody are purified to single stranded DNA, and introduced to a DNA microarray surface, where single-stranded fragments of the genome of interest are fixed. By observing which fixed genomic DNA fragments (with known genomic coordinates) are bound, the binding sites of the TFs can be mapped to the genome. While ChIP-on-ChIP can cover only part of the genome, as the microarray can only accommodate a limited number of genomic fragments, the ChIP-based experiments coupled with next generation sequencing, ChIP-seq [48], ChIP-exo [12] and ChIP-nexus [11], can measure TF binding to all of the mappable genome. These three techniques are described in more detail in the Methods chapter.

The recently developed CUT&RUN (cleavage under targets and release using nuclease) technique [49] is a promising alternative for replacing ChIP-seq, as it does not employ cross-linking of proteins to DNA, the step in ChIP protocols that is known to easily produce false-positive binding events [61, 62]. In CUT&RUN, a protein capable of recognizing antibodies from a certain organism (such as protein-A for human proteins), is genetically fused with micrococcal nuclease (MNase) that cleaves double-stranded DNA. When this complex binds to the antibody-tagged TF of interest, the MNase cleaves the bound DNA and the sequence corresponding to the site bound by the TF can be retrieved and later identified with high-throughput sequencing.

1.2 MODELING THE BINDING AFFINITIES OF TFS TO DNA SEQUENCES

In principle, the binding affinities of a given TF towards any DNA could be completely described by listing relative binding affinities of the TF to each possible DNA sequence. In

practice this kind of a model is not very easily interpretable by humans when the length of the DNA sequences is longer than a few base pairs and is thus not very useful. Models describing the affinity of a TF to DNA sequences have two main purposes: 1) to summarize, and present the binding affinities in a format easily interpretable by humans and 2) to be able to predict which DNA sequences the TF binds and which not. Usually some models fulfill the purpose 1 better and some others purpose 2.

The current baseline model for TF-DNA binding affinities is the position weight matrix (PWM) [63]. The PWM model has stood the test of time even though it is known to predict the *in vivo* binding sites of TFs rather poorly (see for example the comparison in [42]). The main reason of this is likely its simplicity as it assumes independence between positions of the model. Thus, the sequence logo representation of the PWM model very clearly and concisely describes the DNA sequences favored by the TF (see example from Figure 1.1). Moreover, the PWM is still a good model of TF-DNA binding in the absence of for example nucleosomes and interactions with other proteins.

The main assumption in the PWM model is that the positions of the model are independent from each other. The model is constructed by counting the occurrence of each nucleotide at each position either from simple alignment or using more involved algorithms designed for finding PWM models from sets of DNA sequences. Review of these algorithms is outside the scope of this work. A matrix that reports the frequencies of nucleotides along the positions of the model is called a position frequency matrix (PFM). The position weight matrix is created from the PFM by computing the position-specific log-likelihoods relative to expected frequencies of nucleotides given by a background distribution. If the position-specific nucleotide frequencies from the PFM are denoted as $f_{b,x}$, where b marks the nucleotide and x marks the position along the model, PWM is then calculated as

$$F_{b,x} = \log_2(f_{b,x}/B_b), \tag{1.1}$$

where B_b is the frequency of nucleotide b from the background model. Thus the PWM model can be used to express the probability of any sequence with same length than the PWM to bind the TF modeled by the PWM assuming the given background model and independence of positions.

Numerous more complex models of TF-DNA binding have been proposed over the years, and some of them are briefly outlined in the following to give a general idea of the different approaches. Adjacent dinucleotide model (ADM) [64] is an inhomogeneous Markov chain of order 1, where dependencies are modeled only between adjacent positions of the model. An example visualization of the ADM model is shown in Figure 1.1. Bayesian Markov Models (BaMMs) [65] are higher order Markov models where lower order models function as priors to higher order models. Recently, also deep learning models have been used to model TF-DNA binding achieving state-of-the-art predictive performance [42], but the interpretation of these models starts to become more challenging (see example from Figure 1.1). Deep learning models in genomics and their interpretation are discussed in more detail below.

1.3 DEEP LEARNING IN GENOMICS

Due to the developments in computing hardware and deep learning algorithms during the last decade, deep learning methods have become a mainstay also in biological research (see e.g. [67] for a comprehensive review). One of the first applications of modern deep learning methods in genomics was to predict TF-DNA binding using DNA sequence as input. These

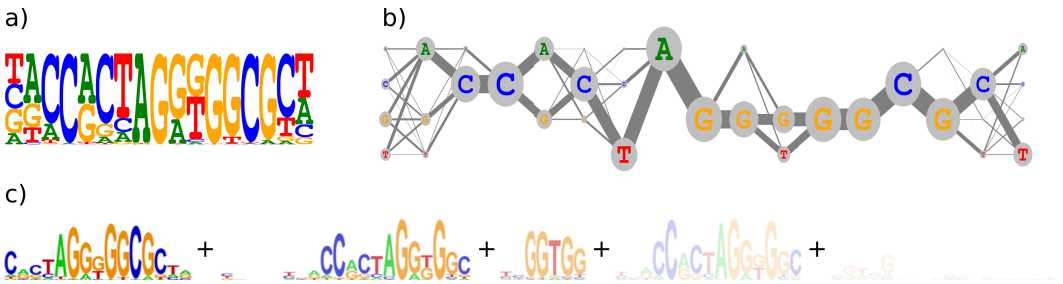

FIGURE 1.1: Examples of models describing the affinity of the CTCF protein to DNA sequences. a) PWM-model (position weight matrix), where the height of each letter is proportional to the affinity for the given nucleotide at each position. Model positions are independent of each other. b) ADM (adjacent dinucleotide model), where the size of each letter at each position is proportional to the affinity for the given nucleotide at that position and the thickness of the lines connecting adjacent positions is proportional to the transition probability from position $x-1$ to x given the nucleotides connected by the line. Adjacent positions are dependent. c) DeepBind [42] CNN model (convolutional neural network), where each logo corresponds to one convolutional filter (motif detector) learned by the neural network (top 5 filters shown for simplicity). Opacity of the logo corresponds to the weight of the given filter in the model. The more complex the models become, the harder it is usually to intuitively visualize the types of sequences favored by the model. The CNN model can learn interactions between any positions of the model and between the different filters, and thus the visualization shown here does not fully describe the model. All models have been generated from the CTCF HT-SELEX experiment published in [3]. The PWM and ADM were generated using Moder2 program [66], and the DeepBind model was downloaded from the website provided by the authors `http://tools.genes.toronto.edu/deepbind/`.

convolutional neural network (CNN) models (e.g. [42, 68]) outperformed earlier machine learning methods and thus paved the way for larger scale application of deep learning methods to study DNA sequence elements controlling gene expression.

The innovation in applying the CNN methods, that had previously achieved massive successes in computer vision (see for example [69, 70]), was to consider the DNA sequence as an "image" of a sort. In computer vision applications, the input for the CNN is a two dimensional image with usually three color channels (red, green, blue). Analogously, DNA sequence is a one dimensional image with four channels (A, C, G and T). This means that if one replaces the two dimensional convolutions used in computer vision with one dimensional convolutions, the power of the CNN methods can be harnessed to analyze the DNA sequence. Large datasets of TF-DNA binding were readily available both *in vitro* (e.g. [3]) and *in vivo* (e.g. [4]) allowing training of these new, training data intensive models. In a recent systematic comparison of neural network architectures for predicting DNA and RNA binding specificities [71], mixed CNN/RNN (Recurrent Neural Network) architectures were found to perform better than pure CNN or RNN architectures. Shortly, in RNNs the hidden layers of the network are connected both to the input and to the internal state of the model containing information about the previous inputs. This serves as a sequential memory allowing the RNN to learn interactions between distant elements in the input sequence. RNNs have been especially successfully used in natural language processing [72].

A logical next step from models predicting binding of a single TF is predicting activities of entire gene regulatory elements, promoters and enhancers. The ultimate aim of these models is to be able reliably explain and predict the effect of non-coding variants to gene expression. The earliest of these types of methods, DeepSEA [73], was trained to predict 919 chromatin features including DNAse, TF and histone features, using DNA sequence from the genome as input. This work demonstrated that a CNN based model trained on genomic DNA sequence was able to predict the effects of individual SNPs on TF binding and to prioritize functional SNPs (single nucleotide polymorphisms) from non-functional based on the model predictions. The authors of DeepSEA later presented an even more comprehensive CNN based model [74] trained on over 2,000 different histone mark, TF and DNA accessibility profiles in over 200 different cell types. With this updated method called ExPecto, the authors also introduced a spatial feature transformation module that allowed integrating signal from a 40 kb window.

CNN models have been shown to perform well in classifying between promoter and non-promoter sequences [75]. Enhancers are more difficult to model mostly because no single biochemical assay can reliably identify all enhancers. Because of this, most enhancer models are actually predicting DNA accessibility, as genome-wide measurements of chromatin accessibility are available in many tissues (see e.g. [4]). Basset [43] used CNNs to predict chromatin accessibility in over 100 cell types and showed that a CNN-based method learned known TF binding motifs associated with open chromatin and predicted greater chromatin accessibility changes for likely causal GWAS (Genome-Wide Association Study) SNPs. In contrast to predicting chromatin accessibility profiles like Basset, DeepEnhancer [76] trained CNN classifiers to separate genomic enhancers from non-regulatory sequences using an enhancer set defined by bidirectional transcription signal in CAGE experiments from the FANTOM5 project [77]. With Basenji [78], the receptive field of Basset model was expanded to cover a much larger 131 kb region of the input DNA sequence to be able to model distal regulatory interactions, and to predict high-resolution quantitative genomic measurement profiles instead of binary chromatin accessibility. The increase in the

receptive field was made possible by use of dilated convolutional layers [79]. Also another recently published model, BPNet [80] was trained to predict quantitative TF-DNA binding profiles, but on base-resolution ChIP-nexus data of genome-wide binding of selected TFs. This led to novel insights into soft motif syntax for mouse TFs Oct4, Sox2, Nanog and Klf4.

Recently, Enformer [81] introduced the use of so called transformer models, that have been previously successfully applied to for example natural language processing problems (see e.g. [82]), into modeling gene regulatory elements. In practice, the main difference between transformers and CNNs is that in transformer models, each position directly sees (or *attends to*), all other positions in the model and this can allow a more efficient flow of information between elements separated by long distances in the input sequences, such as promoters and enhancers. Using CNNs, distal elements can only be seen by adding more and more layers to the network. Using this feature of transformers, Enformer seems to be able to better integrate distal enhancers with promoters to predict gene expression than earlier CNN-based models [81].

1.4 Interpretation of deep learning models in genomics

Increasing popularity and successes of deep learning models in genomics have created a demand for tools that help translate the rules of gene regulation learned by these complex models into human understandable format. Even if the ability of deep learning models to predict the effects of variants is already a great achievement in itself, understanding how these predictions are made is key to new mechanistic insights of the gene regulatory processes. Thus already the very first applications of deep learning to prediction of TF-DNA binding and chromatin accessibility also developed ways to visualize the features learned by the models.

Directly visualizing the filters of the first convolutional layer, also called the "motif detector" layer [42], of a CNN is one of the earliest model interpretation strategies. As the motif detector layer is directly reading in the one-hot encoded DNA sequence, the weights of the filters in this layer can be visualized as sequence logos similar to PWMs, where the weight of each letter at each position corresponds to the importance of that given nucleotide. Subsequent research has shown that care must be taken when designing the CNN model architecture if the motif detector logos are used to draw conclusions about biologically meaningful features learned by the model, as the features learned by the first layer filters heavily depend on the model architecture [83]. An alternative to direct visualization of the motif detector logos is to align the sequences corresponding to highest activation of each first layer filter and present this alignment, weighted by the model prediction, as a sequence logo [43].

Another deep learning model interpretation strategy used already by the early deep learning applications to genomics [43, 73] is so called *in silico* saturation mutagenesis (ISM). In ISM, systematic mutations are introduced to a certain position of input sequences and the change in the deep learning model predictions is recorded for each mutant. ISM is motivated by so called saturation mutagenesis experiments, and it is the standard way of predicting the effects of variants to DNA sequences. The main drawback of ISM is that it is computationally expensive, as the whole model needs to be evaluated for each variant scored. Recently developed fastISM [84] alleviates this problem for certain types of CNN architectures by restricting the calculation of variant effects to those parts of intermediate convolutional layers that are affected by the variant.

Similarly to ISM, so called feature attribution methods highlight the positions most important for the deep learning model predictions in a specific input sample scored by the model using backpropagation (e.g. [85–87]). Because of the efficiency of backpropagation, the feature attribution methods can run orders of magnitudes faster than ISM. However, methods such as DeepLIFT [86] compute the feature attribution scores against certain reference sequences and the choice of the reference sequence can affect the feature attributions. Strength of both ISM and feature attribution methods in deep learning model interpretation is that the feature importances and predicted effects of variants can be intuitively visualized for each input sample as a sequence logo. A drawback is that they operate on the level of an individual input sample which can sometimes complicate making general conclusions about features learned by the model. To overcome this caveat, approaches like motif discovery guided by the predicted attribution scores of the deep learning model [88], sampling the maximum entropy distribution around sample inputs [89] and visualization of feature maps learned by the deep learning model [90] have been developed.

In addition to scoring single nucleotide variants, a pre-trained deep learning model can also be used to score synthetic sequence inputs with embedded known features. This approach has been described as using the deep learning model as an "oracle" [80] as the researcher is using the deep learning model directly to test hypotheses. For example in [80] the BPnet model trained on ChIP-nexus binding profiles was used to score sequences with different spacings between TF binding sites to discover preferred spacings.

1.5 OUTLINE

Despite major developments in both "wet-lab" and "dry-lab" methods in the recent years, as discussed above, sequence determinants of human gene regulatory elements remain incompletely understood. Rapid development of high-throughput genomics measurements has created an increasing demand for novel computational method development and application of state-of-the art computational modeling tools from other fields to genomics data.

In this study, we have utilized both approaches to 1) gain novel insights into the sequence determinants of human promoters and enhancers and interactions between them and 2) to publish computational methods for the research community to use in analysis of high-throughput genomics data. In Publication I, we describe software tools to determine TF-DNA binding sites genome-wide from ChIP-exo/nexus experiments and to analyze allele-specificity of TF-DNA binding. In Publication II we use massively parallel reporter gene assays to directly assess transcriptional activities of genomic, designed and completely random DNA-sequences. Testing transcriptional activity of human gene regulatory elements selected from synthetic sequences with uniform nucleotide frequencies allows for the first time de-coupling the transcriptionally active features from biases present in the human genomic sequence such as the GC-content bias in the human promoters. In Publications II and III we apply and develop new methods for interpreting deep learning models trained on DNA sequence data that can be applied also to other types of models trained for example with RNA or protein sequence.

In this thesis, I will first introduce the specific aims of the study in chapter 2. In chapter 3 I will describe the data analyzed and used by the computational methods developed and applied in this work. I will also go into details of the central computational methods utilized here. Chapter 4 will present an overview of the results, including both descriptions

of the novel algorithms and tools developed, and biological results. The results described in this thesis are discussed in context of the previous literature in chapter 5.

2

AIMS OF THE STUDY

The primary aim of this study is to develop novel computational approaches to model and interpret results from large-scale regulatory genomics experiments. The goal is to better understand the mechanisms of how transcription factors regulate gene expression in humans. The specific aims can be summarized as follows:

A. Develop a tool to call transcription factor binding sites from novel ChIP-exo and ChIP-nexus experiments to allow accurate and unbiased determination of transcription factor binding *in vivo*.

B. Apply and develop state-of-the-art machine learning methods to model and discover the DNA sequence elements and their interactions regulating transcription in humans.

C. Apply and develop novel tools that help to interpret machine learning methods in genomics and especially dependencies and interactions learned by deep learning models.

3

MATERIALS AND METHODS

In this chapter, I will describe in detail the main computational methods and resources used in this study. I will focus on the methods applied and developed by the author, as listed in the *Author contributions* section. The experimental data generated in this study by co-authors of the publications are listed, and details of generation of these data are in the corresponding publications. Generation of experimental data used directly as input in analyses and methods developed by the Author are described in more detail in the following. Computational analyses performed by co-authors are described in more detail when they are directly needed for description of methods and analyses performed by the Author.

3.1 DESCRIPTION OF THE EXPERIMENTAL DATA USED AS INPUT FOR THE COMPUTATIONAL METHODS DISCUSSED IN THIS STUDY

3.1.1 *ChIP-seq, ChIP-exo and ChIP-nexus*

ChIP-seq [48] has become the standard experiment for measuring TF-DNA binding patterns genome-wide (see e.g. [4]). In ChIP-seq, the aim is to use an antibody specific to a given TF or other protein associated with DNA to select fragments of the DNA of a living cell that were bound by the protein of interest. These DNA fragments can later be mapped back to the genome, giving researchers a genome-wide map of binding of the protein of interest, as most of sufficiently long fractions of DNA are unique in the genome.

The first step of ChIP-seq experiments is to crosslink proteins with DNA using formaldehyde. Then, the DNA of the cells is sheared into random fragments of desired size (usually 100 bp - 300 bp). Longer fragments map more likely to a unique position in the genome, but the longer the fragment, the worse is the resolution of the experiment as due to the random nature of DNA shearing, the true binding site of the TF can be anywhere within the borders of the DNA fragment (see Figure 3.1a). After shearing of the DNA, the antibody specific to the protein of interest is used to select those fragments of DNA that are bound to the protein of interest. It is of crucial importance, that the antibody used is truly specific to the protein of interest, else the experiment can produce significant amounts of false binding sites due to the antibody recognizing unintended proteins. Once the DNA fragments of interest have been selected with the antibody, the cross-linking between the DNA and the proteins is broken, the DNA purified, and finally the purified DNA is read using DNA sequencers.

The data from the sequencer consists of strings of nucleotides, called sequencing reads, describing the observed DNA sequences, and quality scores of the base calls that can be used to filter the data. These reads are then aligned to a reference genome of the organism in question by an aligner software such as the Burrows-Wheeler Aligner [91]. The final step in ChIP-seq data analysis is peak calling, where the reads aligned to the reference genome are used as an input for a specific software that aims at fitting a model to the aligned reads so that reads resulting from true binding events of the protein of interest and DNA are separated from noise generated for example by biases stemming from cross-linking,

antibody non-specificity, uneven PCR amplification of the sequencing library (e.g. [92]), base calling errors made by the sequencing machine (e.g. [93]) or alignment uncertainty (e.g. [94]). Numerous peak calling software have been developed for analysis of ChIP-seq data over the years, and the performance of the most widely used peak callers have been benchmarked for example in [95]. The term peak calling comes from the fact that when the signal from a ChIP-seq experiment is visualized by counting read or fragment coverages, i.e. the number of mapped reads overlapping each other at each genomic position, TF-DNA binding sites can be recognized as peaks (see Figure 3.2) in the coverage signal.

In [12], the authors describe a modification to ChIP-seq, called ChIP-exo, adding a step where the DNA fragments bound to protein of interest and selected using an antibody specific to the protein are digested using λ-exonuclease. This step greatly improves the resolution of the experiment, as the λ-exonuclease digests double-stranded DNA in 5′ to 3′ direction until stopped by a physical barrier, in this case the protein cross-linked to DNA [12]. This means that the 5′ ends of the fragments in ChIP-exo experiment are in theory always located exactly at the border of the bound protein (Figure 3.1b). ChIP-nexus [11] is essentially a more efficient ChIP-exo protocol, where re-ligation of the adapter sequences to the 5′ ends of the reads, removed initially by the λ-exonuclease digestion, is done using circular ligation instead of more inefficient intermolecular ligation used in the original ChIP-exo protocol.

The main steps of ChIP-exo/nexus data analysis after the reads have been obtained by sequencing are similar to ChIP-seq. Reads are aligned to a reference genome and can be filtered based on base call and alignment quality. Due to the λ-exonuclease digestion, the signal from ChIP-exo/nexus binding events is however different from ChIP-seq signal, and thus specialized algorithms are needed to fully leverage the increased resolution of ChIP-exo/nexus. In ChIP-seq, the reads will pile up into peaks with relatively shallow slopes due to the random shearing of DNA which means that reads overlapping with a given binding site have a distribution of start points of alignment. In ChIP-exo/nexus, the 5′ ends of the reads in theory always map to the same positions flanking DNA-bound proteins, forming "boundaries" around the bound protein. In reality, the reads at TF-DNA binding site boundaries never pile up to exactly the same position in ChIP-exo/nexus, but nevertheless the signal is much more accurately localized, as illustrated with an example in Figure 3.2.

This feature of the ChIP-exo/nexus signal, however, introduces one problem for data preprocessing: traditionally in ChIP-seq data preprocessing, duplicated reads caused by uneven PCR-amplification of the sequencing library have been removed by discarding all but one of sets of reads that map to identical position. In ChIP-exo/nexus, this cannot be done since the expected signal from the experiment is reads mapping exactly to the same position. To tackle this problem, Unique Molecular Identifiers (UMIs [96]) have been used in ChIP-exo/nexus [11] experiments. In short, the idea of UMIs is to tag each molecule in the initial sequencing library randomly with a constant-length barcode. Given that the number of UMIs is sufficient relative to the library size, it is extremely unlikely that two different molecules will have the same UMI label *and* will map to the same position in the genome. Thus the final input to a ChIP-exo/nexus peak caller are essentially the 5′ end positions of the reads describing unique molecules mapped to the genome, and from this input, the aim of the peak caller is to report positions where the protein of interest was binding in the genome, and give estimate of the rank of the binding sites by strength as well as estimate the uncertainty of the binding site calls. This data is used as input for the PeakXus software described in Publication I.

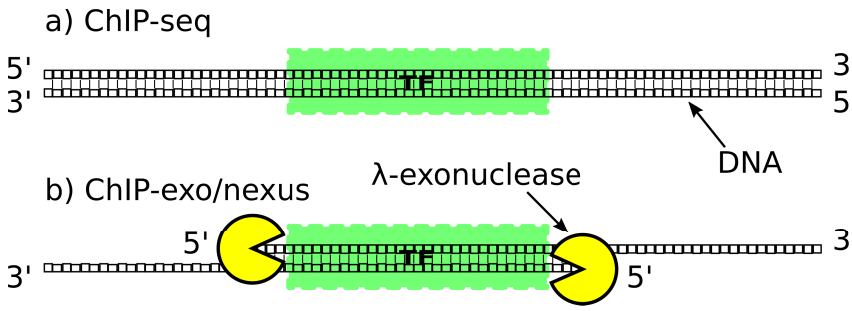

FIGURE 3.1: Schematic presentation of the difference between the TF positions within ChIP-seq (a) and ChIP-exo/nexus (b) reads. In ChIP-seq, random shearing of DNA means the real TF binding site can be anywhere within the read boundaries. The λ-exonuclease treatment in ChIP-exo/nexus digests the 5′ ends of the fragments until the 5′ end corresponds with the boundary of the bound TF.

3.1.2 *ATAC-seq*

ATAC-seq [99] (Assay for Transposase-Accessible Chromatin using sequencing) is a high-throughput, genome-wide method for detecting regions of accessible or open chromatin, meaning those parts of the genome where DNA is not packed around nucleosomes. In ATAC-seq, a hyperactive mutated Tn5 transposase cleaves accessible chromatin and tags it with sequencing adapters. The fragments of accessible DNA are then read using high-throughput sequencing and mapped to the corresponding reference genome using an aligner software similarly to ChIP-seq/exo/nexus. The open chromatin, or accessible regions of the reference genome are then determined using a peak calling software that separates the reads at the real open chromatin regions from background noise. In Publication II, MACS2 [100] peak caller was used to determine the open chromatin regions from the ATAC-seq experiments, as per the current default analysis [101]. These open chromatin regions defined using MACS2 were used as the basis for creating the classification data set for machine learning.

In Publication II, the GP5d ATAC-seq data was used to train a classifier and to compare the classification performances of models trained on STARR-seq and ATAC-seq data in classifying between open and closed chromatin regions in the GP5d colon cancer cells. To this end, 170 bp long sequences were fetched from the ATAC-seq peaks to match with the 170 bp long sequences from the random enhancer STARR-seq experiment. The standardization of the input sequence lengths between the models was done to allow comparison of models using exactly the same test sets. For the class 1 signal set (corresponding to open chromatin), fragments overlapping with any ATAC-seq peak were selected and the 170 bp sequence closest to the overlapping peak summit was retrieved. Exact duplicate sequences were discarded. A balanced negative set (class 0, corresponding to closed chromatin) was created by sampling random 170 bp long sequences from the human genome requiring that they do not overlap with ATAC-seq peaks.

In addition, all regions covered by a so called "extended blacklist" created for the machine learning analyses were discarded from the genomic machine learning datasets (both ATAC-seq and STARR-seq). The extended blacklist was created to remove genomic regions problematic for read mapping that could cause the machine learning models to learn biases. The use of a blacklist is motivated by the blacklisting of problematic genomic regions in the

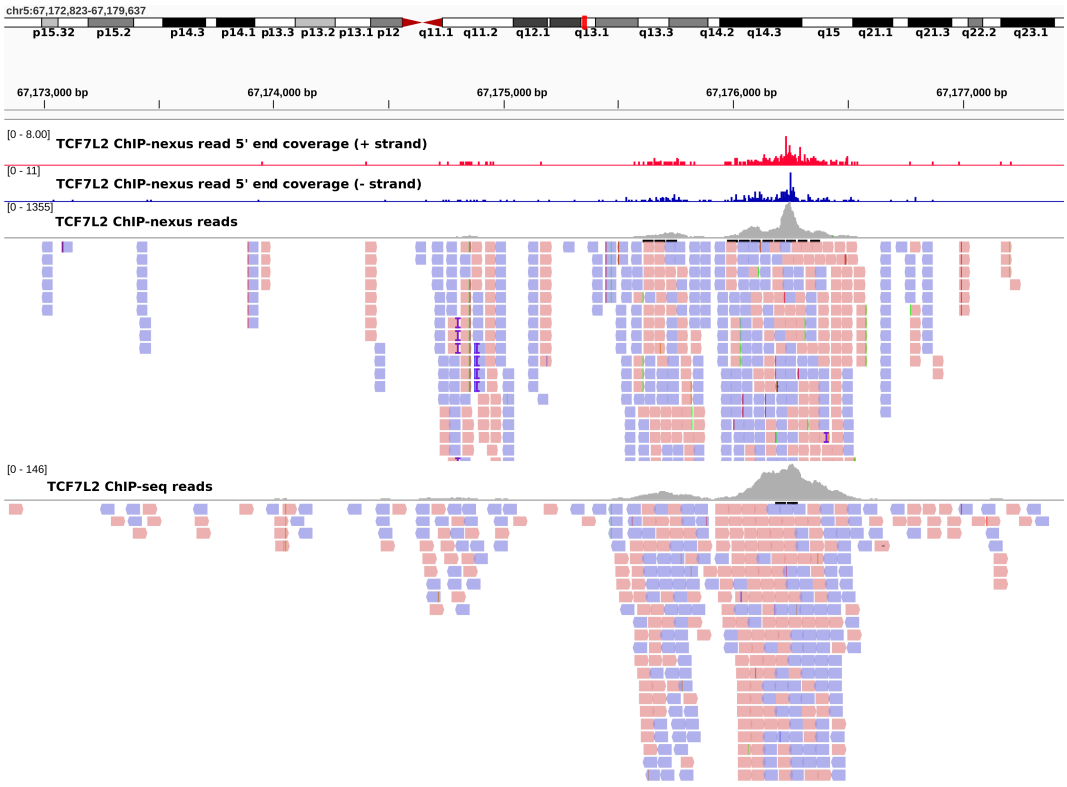

FIGURE 3.2: IGV genome browser [97] tracks showing the reads mapped to the same reference genome region from TCF7L2 ChIP-nexus (middle, data from [98]) and TCF7L2 ChIP-seq (bottom, unpublished data) experiments. The blue and red arrows correspond to individual reads mapping to the sense (red) and antisense (blue) strands, respectively. Note that not all individual reads are shown in this visualization for positions with many mapping reads. The gray coverage signals corresponds to the number of overlapping reads at each position, regardless of strand. The topmost tracks show the UMI counts of read 5′ end positions from the ChIP-nexus experiment (red = sense strand, blue = antisense strand). Signal range for each track is shown on the left. Both experiments have detected TF binding at the same region, but the ChIP-nexus signal is concentrated as a much more narrow peak. Note that the ChIP-nexus peak summit is flanked from left by the position with locally highest number of unique read 5′ ends mapping to the sense strand, and from right by the the position with locally highest number of unique read 5′ ends mapping to the antisense strand. This feature of the ChIP-nexus/exo signal allows mapping the TF binding positions with much better resolution than what can be achieved using ChIP-seq.

ENCODE analyses [102]. The standard ENCODE blacklist is carefully curated by examining the vast collection of ENCODE high-throughput sequencing datasets for problematically behaving genomic regions regardless of the cell type or the exact experimental technique. To accommodate the blacklist for the machine learning analyses, we added the following regions to the extended blacklist: all positions ± 1 Mb from centromeres, all positions with Ns in the hg19 reference genome (machine learning models were trained only on A, C, G and T), and non-uniquely mapping regions defined as follows: all unique 55-mers present in the hg19 reference genome were fetched and aligned back to hg19 reference genome with bwa aln aligner [91]. Each position not covered by reads mapping with sufficiently high quality (MAPQ>20), was added to the extended blacklist. Note that some of the added regions already overlapped with the standard ENCODE blacklist. This extended blacklist covers around 12% of the hg19 reference genome and should remove possible biases stemming from mappability issues fairly conservatively.

After preprocessing, the final input used by the machine learning classifiers trained on the ATAC-seq data in Publication II are sets of 170 bp long sequences which are divided to two classes: class 1 (open chromatin), and class 0 (closed chromatin).

3.1.3 *STARR-seq*

In Publication II, we measure activities of gene regulatory elements in human cells using STARR-seq [13] MPRA (Massively Parallel Reporter gene Assay) experiments. The STARR-seq reporter libraries used in this study are designed so that the DNA sequence whose enhancer activity is investigated is included in the RNA transcript and thus the elements driving gene expression can be directly identified by sequencing the transcribed RNA. Figure 3.3 shows the design of the different STARR-seq libraries utilized in this study. Design i, which we call the motif library, measures enhancer activities of designed 49 bp long sequences that contain either single or multiple copies of known TF binding motifs in different orientations and spacings. Design ii, the genomic library, measures enhancer activities of approximately 500 bp long fragments of the human genome. Design iii measures enhancer activities of 170 bp long completely random synthetic DNA sequences sampled from uniform nucleotide background. In designs i-iii, transcription is initiated at the position defined by a weak minimal promoter included in the constructs. In design iv, both the promoter and the enhancer sequence are synthetic random 150 bp long sequences. In design iv, only the enhancer sequence is captured by RNA-seq, but the corresponding promoter sequence can be retrieved by mapping the transcribed enhancer to the input DNA and taking the corresponding promoter sequence. The different STARR-seq reporter libraries are transfected into human cells and the RNA produced by the cellular transcriptional machinery is sequenced after 24 hours (Figure 3.3) to observe the transcriptionally active sequences. Preprocessing of the STARR-seq datasets used in the machine learning analyses described in this work is discussed in detail below.

3.1.3.1 *Genomic STARR-seq data preprocessing (design i)*

In the genomic STARR-seq design the fragments of the human genome capable of driving gene expression as enhancers were mapped back to the genome using Bowtie2 [103] aligner. Similarly to ChIP-seq or ATAC-seq, the STARR-seq reads aggregate to peaks when mapped back to the genome when activities of overlapping genomic fragments are tested. An example genome browser view of a STARR-seq peak is shown in Figure 3.4. In Publication

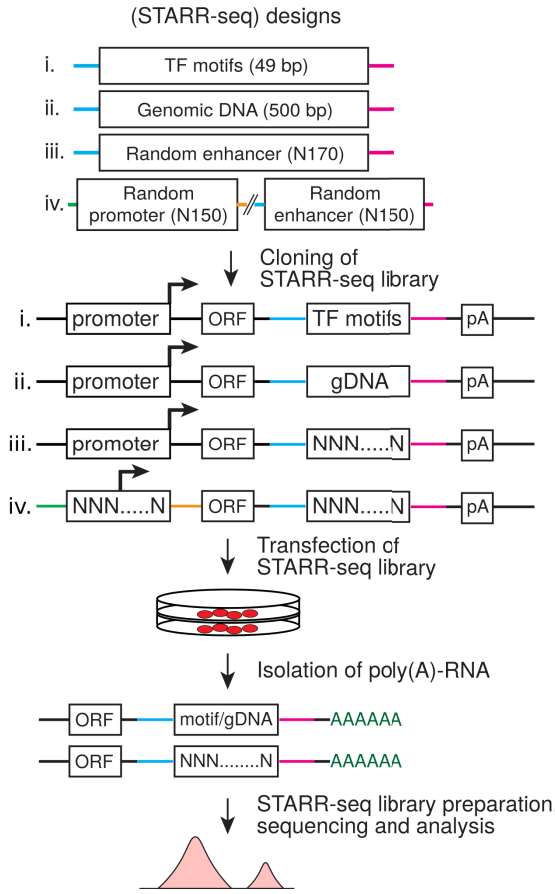

FIGURE 3.3: Schematic presentation of the STARR-seq designs used in this study: i) Motif library of designed sequences containing known TF binding motifs and their combinations cloned to the enhancer position. ii) Genomic library containing fragmented human genomic sequences cloned to the enhancer position. iii) Random enhancer library containing synthetic random 170 bp long sequences cloned to the enhancer position. iv) Random binary STARR-seq library containing synthetic random 150 bp sequences cloned both to the promoter and to the enhancer positions. The STARR-seq libraries were transfected into human cell lines and total RNA was isolated 24 hours after transfection followed by sequencing and data analysis. In designs i-iii, the promoter is a minimal promoter (see Publication II), while in design iv, only those random sequences that by chance have elements required for a functional promoter, act as promoter. ORF = open reading frame, gDNA = genomic DNA. Figure adapted and modified from Publication II.

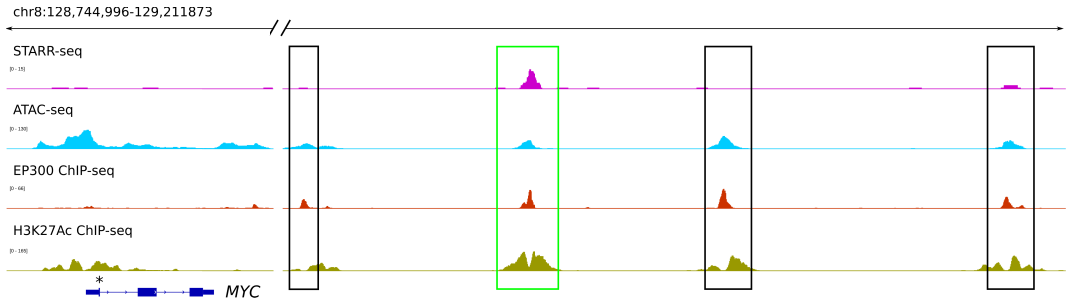

FIGURE 3.4: IGV genome browser [97] tracks showing the *MYC* gene locus, its promoter region (transcription start site, TSS, is marked with asterisk) and four enhancers located downstream of the *MYC* gene. The tracks are from top to bottom: genomic STARR-seq fragment coverage (magenta), ATAC-seq fragment coverage (cyan), EP300 ChIP-seq read coverage (red), and histone H3 lysine 27 acetylation (H3K27Ac) ChIP-seq read coverage (yellow). H3K27Ac marks acetylation of the lysine residue at the N-terminal position 27 of the histone H3 protein and is a mark of active enhancers and promoters [104]. EP300, or histone acetyltransferase 300 is a chromatin remodeling enzyme that marks active enhancers and promoters [105–107]. For STARR-seq and ATAC-seq tracks, a fragment means positions between each mapped paired-end read pair. Fragment coverage means total number of such fragments overlapping a genomic position. The EP300 and H3K27Ac ChIP-seq coverage signals were downloaded from the ENCODE data portal as described in Publication II, and they describe fold change over control computed from single-end sequencing reads. The *MYC* promoter is marked with open chromatin according to ATAC-seq and associated with higher transcriptional activity according to the H3K27Ac signal, and weak EP300 signal. The active enhancers, highlighted with boxes, are also marked with ATAC-seq, H3K27Ac and EP300 signals. One of the enhancers is marked with STARR-seq signal (green box), others only by the three other types of signal (black boxes). These are examples of *classical* and *chromatin dependent* enhancers, respectively, defined and discussed in more detail in the Results section. Notice that the genome browser tracks have been cut just before the first of the four enhancers for clarity of visualization, as indicated by a gap in the signal tracks. The *MYC* TSS and the first enhancer are separated by approximately 425 kb. The datasets used are described in Publication II. Figure modified and adapted from Publication II.

II, the active enhancers were determined with peak calling using MACS2 [100] software. The STARR-seq input library DNA, meaning a sequenced sample of the reporter library comprising of the genomic DNA fragments before transfection, was used as a control in peak calling. Similarly to as described for ATAC-seq above, no sequences mapping to the genomic regions covered by the extended blacklist were included in the machine learning analyses.

Similarly to ATAC-seq, the GP5d genomic enhancer STARR-seq fragment lengths were standardized to 170 bp to allow comparison with the models trained with random enhancer STARR-seq data. The signal set (class 1, active enhancers) sequences were created by taking the 170 bp closest to the peak summit from each genomic STARR-seq fragment that overlaps with a GP5d genomic enhancer STARR-seq peak. Sequences overlapping with the extended blacklist were discarded. The class 0 sequences (no enhancer activity) were drawn at

random from the pool of input library sequences, but discarding sequences overlapping with GP5d genomic STARR-seq peaks or the extended blacklist regions.

In principle it is possible that a CNN classifier could learn some feature correlating with different input library coverage of the class 1 genomic enhancer STARR-seq sequences relative to randomly sampled sequences from the input library, if such feature would exists. Here, input library coverage means the number of copies of a specific sequence present in the sequenced input library. To control for this possible source of bias, the class 0 sequences used in the machine learning analyses were sampled so that the input library coverage histogram of the class 0 sequences matched the input library coverage histogram of the class 1 sequences.

3.1.3.2 *Random enhancer STARR-seq preprocessing (design iii)*

Preprocessing of the random STARR-seq data is different from the genomic library, as the random STARR-seq data is not aligned to a reference genome at any point. The 170 bp long active random enhancer STARR-seq sequences were first filtered for duplicates. Sequencing errors of PCR duplicates of the same initial molecule can lead to multiple sequencing products that have few mismatches but originate from the same initial DNA molecule. To control for this, the random enhancer STARR-seq 170 bp long sequences were sorted four times based on 40 bases long non-overlapping subsequences from base 6 to 165 and only one sequence per identical subsequence at each sort step was kept. This ensured that only one sequence out of a set of sequences separated by Hamming distance less than 4 was kept. The final class 1 (corresponding to active enhancers) sequences for machine learning analyses were these unique 170 bp long sequences. The class 0 (inactive enhancers) sequences were sampled at random from the random enhancer STARR-seq input library (sequenced sample of the initial reporter library before transfection) so that their number matched the number of class 1 sequences for each set (training, test, validation) separately. Note that each STARR-seq design has its own input library.

3.1.3.3 *Random binary STARR-seq preprocessing (design iv)*

In the binary STARR-seq experiment, only the enhancer part of the reporter construct was read using RNA sequencing. Thus the correct promoter responsible for the transcriptional activity was identified by mapping the 150 bp long active enhancer sequences back to the original binary STARR-seq input library promoter-enhancer pairs by requiring an exact match of the first 40 bases to the sequence at the enhancer position. In addition, duplicate filtering was performed similarly to the random enhancer STARR-seq. The pairs of active 150 bp long promoters and enhancers were used as the class 1 sequences (active promoter-enhancer pairs) in the machine learning analyses. The class 0 (inactive) pairs were sampled at random from the input library pairs matching their number to the number of class 1 pairs for each set (training, test, validation).

3.1.3.4 *Determining the TSS position from the binary STARR-seq active elements*

A template switch experiment was conducted to determine precise transcription start site positions in the binary STARR-seq experiment. Importantly, in the binary STARR-seq experiment the promoters are completely random synthetic sequences meaning when such a sequence by chance contains elements required for a functional promoter, the TSS position is determined by these randomly enriched features and cannot be known without

a measurement. The template switch experiment was designed to capture the 5′ end of the transcript, which allowed mapping the transcripts back to the input library promoters using the sequence after the TSS in the sequenced transcripts. The details of the template switch chemistry and preprocessing are given in Publication II. Two GP5d template switch libraries were processed separately and later merged so that only one transcript was kept for each unique input DNA promoter sequence to prevent including duplicate promoter sequences in the subsequent analyses. The TSS positions from the template switch experiments were used to create class 1 (active promoter) sequences for training the STARR-seq promoter models such that a 120 bp long sequence, where the TSS was at position 100, was fetched around each TSS from the experiment. Those TSSs where the position of the TSS within the 150 bp promoter sequence did not allow fetching a 120 bp long sequence described above, were discarded. A balanced number of sequences for each set (training, test, validation) were sampled at random from the binary STARR-seq input library promoter sequences and used as class 0 (inactive promoters) in the machine learning analyses.

3.2 NOTE ON TF BINDING MOTIF NAMING

In Publication II, the STARR-seq experiments measure the ability of TF binding motifs to drive gene expression either at enhancer or at promoter position. We refer to this as the (transcriptional) activity of the TF binding motifs. The STARR-seq measurements only detect the RNA with transcriptional activity, not the TF proteins bound to the DNA responsible for the transcriptional activity. Many TF binding motifs are recognized and bound by different proteins, for example the p53 family motif is bound by p53, p63 and p73 proteins. Thus, in cases where binding motifs for several TFs are highly similar, the motifs have been named in figures according to TF class or subclass based on previous literature. Same principle has been applied also when specificities of closely related TFs have not been measured but can reasonably be expected to be similar. Dimeric TF binding motifs are named so that the orientations of their core consensus sequences (GGAA for ETS, ACAA for SOX, AACCGG for GRHL and GAAA for IRF) with respect to each other are listed: HH head to head, HT head to tail, TT tail to tail, followed by gap length between the core sequences. Asterisk indicates an A rich sequence 5′ of the IRF HT2 dimer. Exact PWMs corresponding to the named motifs are listed in Publication II.

3.3 MACHINE LEARNING METHODS

In Publication II, several different machine learning models were applied to model data from STARR-seq experiments. I will describe these models in detail in the following. The classification performance of each CNN and logistic regression model, on each tested hyperparameter combination, on their respective (unseen) test data is shown in Figure 3.7.

3.3.1 *Modeling the STARR-seq experiments with machine learning classifiers*

Essentially, in STARR-seq experiments the cell is used to classify the transfected enhancer and promoter sequences from the input library into those that activate transcription and to those that do not. The classification is performed by the cellular transcriptional machinery that requires certain sequence elements to be present in order to activate transcription. These required sequence elements can be learned from the sequences using

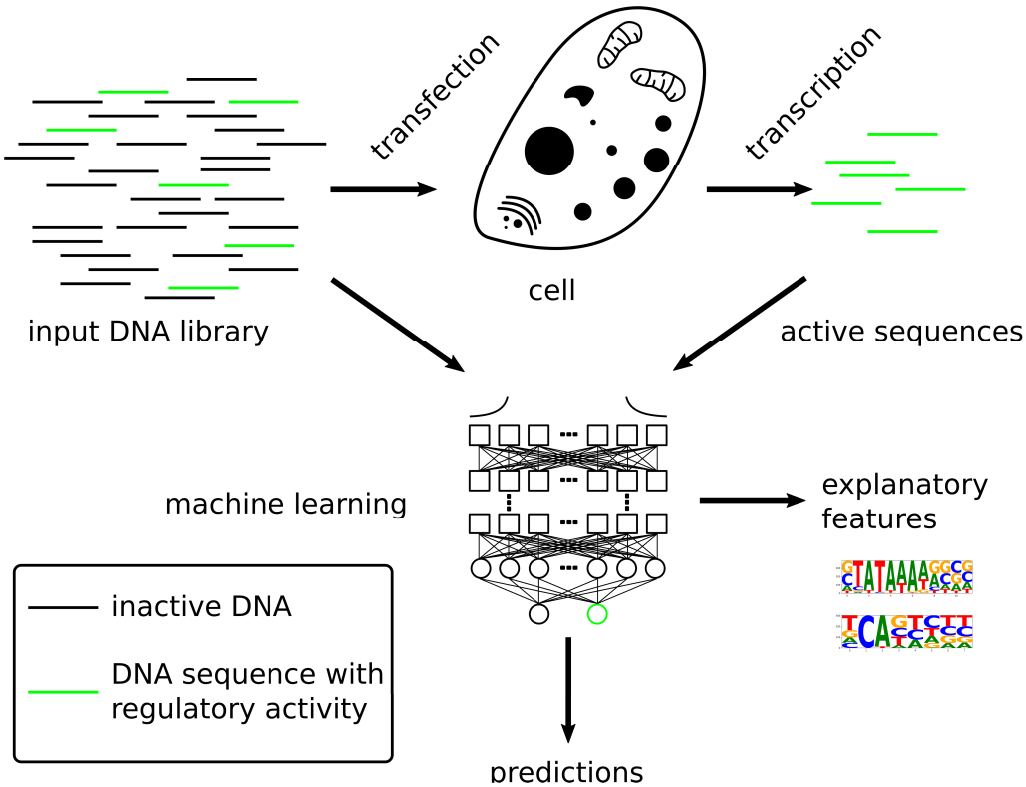

FIGURE 3.5: Schematic presentation of modeling the STARR-seq experiment as a classification problem using machine learning.

machine learning models that try to replicate the classification performed by the cell. Figure 3.5 shows a schematic describing how the essential sequence elements required for transcriptional activity can be extracted from a machine learning model that is trained to act as a similar binary classifier of STARR-seq input library sequences as the cell. The STARR-seq experiments performed in Publication II are extremely well suited for modern machine learning methods as they test millions of sequences that can be used in the model training. In this modeling, the input for machine learning methods are DNA sequences labeled based on the STARR-seq experiment either as class 0 (inactive) or class 1 (active). The convolutional neural network models take the raw sequences as input as such, but for the logistic regression models feature engineering is needed, which is described next.

3.3.2 *Logistic regression classification random of STARR-seq data*

To test whether activities of gene regulatory elements could be determined by linear combinations of effects of known TF binding motifs, we used logistic regression to fit models that use as features either single TF binding motifs only, or both single TF binding motifs and terms accounting for pairwise cooperative binding of selected TF pairs. We used the L1 norm, also known as lasso, to regularize the logistic regression models for easier interpretability. As the set of 880 known TF binding motifs used in this work contains some

motifs that can be highly similar, a non-regularized model, or a model regularized with L2 norm (ridge), could converge into a solution where the learned effects are split among multiple correlating features corresponding to similar TF binding motifs. The L1 norm penalizes solutions with a higher number of non-zero coefficients and thus enables finding the best performing model with the lowest number of individual motifs contributing to the model. Regularization with L1 norm is known to alleviate problems with correlated features, given that the regularization strength is tuned properly [108]. To model the activities of random enhancer STARR-seq sequences with logistic regression, we need to determine the binding sites of each of the 880 known TF binding motifs used in this study in each of the sequences and estimate the binding probabilities of the corresponding TFs to these sites. In the following, I will describe how this was implemented.

3.3.2.1 *Finding the TF binding sites using MOODS*

We used the MOODS [109] software to calculate "PWM match scores" for each of the 880 TFs/DBDs in the set of motifs used in this study against each position in each of the GP5d random enhancer STARR-seq sequences. The PWM match scores were calculated separately for both strands using the strand-specific mononucleotide distributions of the random enhancer STARR-seq input library (see Table 3.3.2.1) as background for the PWM models.

strand	*A*	*C*	*G*	*T*
+	2.288e-01	1.915e-01	2.390e-01	3.408e-01
-	3.408e-01	2.390e-01	1.915e-01	2.288e-01

Table 3.1: The strand-specific mononucleotide biases in the random enhancer STARR-seq input library.

3.3.2.2 *TF-specific occupancy probabilities of DNA sequences*

To create regression features that account for either single TF binding, or cooperative binding of TF-TF pairs to DNA, we estimated occupancy probabilities of STARR-seq sequences by TFs based on PWM models of their binding affinities. A detailed derivation of these probabilities, which follows the derivation in [110] with some modifications, is presented in Appendix A. In the following, only the outcome of this derivation is shown. Given a "PWM match score" $S_{X,i}$ of a PWM corresponding to TF X at position i of a sequence s, computed using MOODS, the probability of s to be occupied by X is

$$P_{X,s} = 1 - \prod_{i=1}^{N_{sites}} \left(\frac{1}{1 + \exp\left(S_{X,i} - S_{X,10000}\right)} \right). \tag{3.1}$$

The product runs over all binding sites of X in the sequence and $S_{X,10000}$ corresponds to the activity of 10,000th strongest binding site of X in the genome, used to estimate the concentration of X in the input library (see Appendix A for details). Similarly, the probability of s to be occupied by a pair of TFs X and Y is

$$P_{XY,s} = 1 - \prod_{i=1}^{N_{sites}} \left(\prod_{j=1}^{M_{sites}} \left(\frac{1}{1 + \exp\left(S_{X,i} + S_{Y,j} - S_{X,10000}\right)} \right) \right). \tag{3.2}$$

3.3.2.3 *Positional logistic regression classifiers*

In contrast to the simple logistic regression described above, in positional logistic regression the PWM match scores were weighted using the position-specific enrichment of the corresponding PWM over the whole set of sequences. Thus, instead of occupancy probabilities, scores

$$A_{X,seq} = \sum_i a_{X,i} \times S_{X,i}, \tag{3.3}$$

were calculated for each sequence and PWM feature X, where $a_{X,i}$ is the positional activity score of PWM X at position i, and i runs over all matches of PWM X in a sequence. Regression coefficients were learned separately for both strands for each PWM.

The positional activity scores required for this analysis were computed by matching motifs to the TSS-aligned promoter sequences from the binary STARR-seq experiment. The number of motif matches for each motif was counted separately at each position and strand so that only the highest affinity motif match per sequence was considered for each motif. The positional activity scores used are $\log_2$ fold changes of the motif match counts between the TSS-aligned promoter sequences and a control set of sequences from the input library estimated with the lfc R-package [111].

3.3.2.4 *Training of the logistic regression classifiers*

The logistic regression models were implemented with the LogisticRegression function in scikit-learn [112] Python library using L1-regularization. The STARR-seq random enhancer, and TSS-aligned binary STARR-seq promoter sequences were divided into training (70%), validation (15%) and test (15%) sets for logistic regression classification. Logistic regression models using single and pairwise TF features were trained on the random enhancer STARR-seq sequences and positional logistic regression models on the binary STARR-seq TSS-aligned promoter sequences. Regularization strength was the only optimized hyperparameter. The optimal regularization strength was chosen based on area under precision-recall curve (AUprc) on the validation data. Otherwise regression was run on default parameters. First, a logistic regression classifier was trained using only features that count matches of individual PWMs (880 features). After this, a more complex classifier was fit with additional features counting all self-pairs

$$A_i + A_i, \tag{3.4}$$

where i runs over all the 880 features, and all pairs of the top 20 strongest individual features (20 features with largest absolute value of the regression coefficient) from the simpler model (see Publication II for the exact features), with all other PWM features

$$S_j + A_i, \tag{3.5}$$

where i runs over all the 880 PWMs, and j runs over the top 20 PWMs from the simple model of 880 single TF features.

Positional logistic regression models were trained using the positional enrichment patterns of PWM matches in the binary STARR-seq TSS-aligned promoters. The same 880 PWM features included in the simple logistic regression were used, plus additional 7 core promoter PWMs from the literature (see Publication II for details). Figure 3.7 shows the performance of each trained logistic regression classifier on unseen test data with the

specific hyperparameter combination resulting to best-performing model on the validation data highlighted for each model type.

3.3.3 *Convolutional neural network classification of STARR-seq data*

Convolutional neural networks (CNNs) learn the features and the parameter weights simultaneously, so no complicated feature design similar to the logistic regression classification is needed. In the following I will describe the CNN architectures and model training strategies used in this study.

3.3.3.1 *Convolutional neural network architectures*

The CNN architectures used in this study are motivated by several recent successful applications of CNNs in learning and modeling information from DNA sequence data (see e.g. [42, 43, 80]). The main building blocks of the networks in this study are convolutional filters and fully connected neurons. The important difference between a convolutional filter and a fully connected neuron is that a convolutional filter only sees part of the input from the preceding layer simultaneously, and it is slid through the input from the preceding layer calculating a convolution between the filter weights and the input, whereas a fully connected neuron integrates the whole input from the preceding layer simultaneously. In this study, layers of convolutional filters form so called convolutional modules, where a 1D convolutional layer is followed by batch normalization, ReLu activation and a dropout layer. Each convolutional module contains $N_{filters}$ filters, which is a hyperparameter optimized during training.

The ReLu (Rectified Linear Unit) is an activation function applied to the input of a convolutional filter or a neuron that is defined as

$$f(x) = x^{+} = \max(0, x), \tag{3.6}$$

where x is the input from the preceding layer. ReLu activation allows the neural networks to learn non-linear functions (without proper activation function the neural network would only be able to compute linear matrix products) and has been shown to enable better training of deeper neural networks [113]. Dropout [114] is a simple technique where a randomly chosen fraction f of the filters/neurons of a layer are omitted during each mini-batch training step. Using dropout greatly reduces overfitting and forces the network to learn generalizeable rules. Dropout is usually (and also in this work) applied only during training, and the whole model is evaluated when making predictions. Batch normalization [41] is used to re-center and re-scale the input of a layer and it has been empirically observed to speed up and stabilize neural network training, even though the reason for this is still debated.

All CNNs in this study use so called dilated convolutions [115, 116], introduced in the context of CNNs in [79]. Dilated convolution is a way to expand the receptive field of a convolutional filter exponentially when the number of parameters (the number of layers in the network) grows linearly without losing resolution on the input of the network. Receptive field of a filter means the area of input sample covered by the filter when computing the filter output at any given time. Dilated convolution is essentially a convolution where the receptive field of the convolutional filter is expanded by making "holes" into the filter. In a dilated convolution, $l - 1$ spaces or holes are inserted between convolutional kernel elements, where l is a parameter called dilation rate. Notice that the conventional

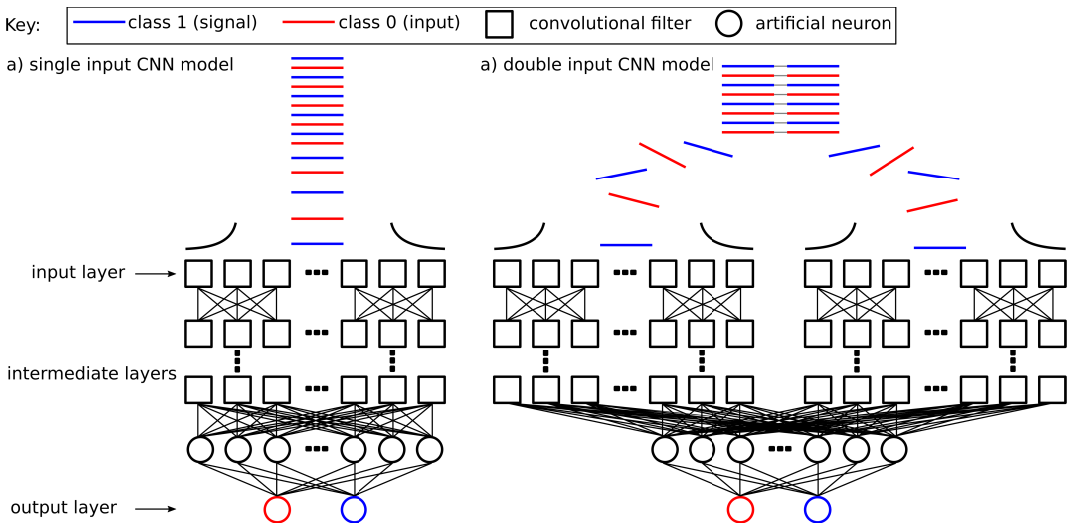

FIGURE 3.6: Schematics of the convolutional neural network (CNN) architectures used in this study. a) CNN models with single input head were used to model the genomic and random enhancer STARR-seq, ATAC-seq, genomic promoter and STARR-seq TSS-aligned promoter data. b) CNN models with two input heads were used to model the binary STARR-seq data, where one input head was used to read in the promoter part and the other input head the enhancer part of the sequence. Notice that the signals from the two input heads are integrated by a fully connected layer.

convolution is a special case of dilated convolution when $l = 1$. In this study, we used exponentially increasing dilation rates, so that for first layer $l_1 = 1$, for second layer $l_2 = 2$, for third layer $l_3 = 4$ and so on. When the dilation rate is increased like this, the receptive field of the filters grows exponentially while the number of parameters grows linearly, making the network training easier and faster than training of a traditional CNN with the same receptive field for the final layer and the same resolution on the input [79].

The two CNN designs used in Publication II are outlined in Figure 3.6. The design with one input head in Figure 3.6a was used for all other CNN classifiers except the binary STARR-seq classifiers, which require two input heads (one for the promoter sequence and one for the enhancer sequence, Figure 3.6b). The other difference between these two architectures is that the double input model includes a fully connected layer of neurons that integrates the inputs from the two input heads while the single input model only contains convolutional modules up until the final output layer. The first layer, or the input layer of the models is a so called motif detector layer [42] that consists of convolutional filters of fixed size x (bp). The first layer contains $N_{filters}$ convolutional filters. The motif detector layer is the only layer of the network that sees the DNA sequences used in the training directly. The filters of this first layer are somewhat similar to PWMs in the sense that each of them has a weight for each possible nucleotide at each filter position. The subsequent layers, whose number is a hyperparameter optimized during training, will gradually combine and abstract the outputs of the motif detector. The final layer is a dense layer of two nodes with sigmoid activation that outputs probabilities of belonging to either of the classes for a given input sequence.

3.3.3.2 *Convolutional neural network training*

For training the CNN classifiers, the STARR-seq random enhancer, the binary STARR-seq promoter-enhancer pairs, and the TSS-aligned binary STARR-seq promoter sequences were divided into training (70%), validation (15%) and test (15%) sets. The genomic sequences from STARR-seq and ATAC-seq were divided into training, validation and test sets in similar proportions based on chromosomes so that no sequences from the same chromosome are in two or more sets (see Table 3.3.3.2). The sequences were fed to the CNN models as such after one-hot encoding (A=[1 0 0 0], C=[0 1 0 0], G=[0 0 1 0] and T=[0 0 0 1]).

The CNN classifiers trained on the random enhancer CNN and the genomic enhancer CNN data were fed also the reverse complement sequences of the training data as it was observed to slightly boost the performance of the models. This is an example of data augmentation that aims at boosting the performance of a deep learning model by generating additional artificial training data by applying some realistic distortion or permutation operation on the original training data samples. When training CNNs on images, possible augmentation strategies could be to for example rotating the images or adjusting their contrast. For enhancers, reverse complementing is a reasonable way to expand the number of training samples as according to the original functional definition of enhancers, they should affect gene expression regardless of orientation [117]. For the models including promoter sequences reverse-complementing was not done to preserve the orientations and positions of sequence features relative to the TSS in the training data.

We also experimented with weighting the training data based on position-specific mononucleotide biases present in the STARR-seq input library sequences, but this did not help the model training indicating that the model was able to easily learn to discard these biases from the data (as the bias is present in both input library sequences and the sequenced transcripts). The CNN models were implemented using Keras [118] with TensorFlow [119] back-end and training was conducted using the Adam optimizer with default parameter values.

training set	*validation set*	*test set*
chr1	chr4	chr2
chr3	chr6	chr10
chr5	chr8	chr11
chr7	-	-
chr9	-	-
chr11	-	-
chr13	-	-
chr14	-	-
chr15	-	-
chr16	-	-
chr17	-	-
chr18	-	-
chr19	-	-
chr20	-	-
chr21	-	-
chr22	-	-
chrX	-	-

TABLE 3.2: Genomic data splits to training, validation and test sets for machine learning analyses.

Early stopping was used for the CNN training so that training was stopped if binary accuracy on validation data did not improve within 200 epochs or when the total training time on a single Nvidia Volta V100 GPU exceeded 72 hours (an exception being the "double input" CNN models trained to classify the binary STARR-seq data where training was continued up until 144 hours if needed due to the larger size of the networks). Model parameters were initialized using the He uniform variance scaling initializer [120]. Optimal hyperparameter combinations (see Publication II) were selected by maximizing binary accuracy on validation data. Figure 3.7 shows the performance of each trained CNN classifier on unseen test data, with the specific hyperparameter combination resulting to best-performing model on the validation data highlighted for each model type.

3.3.4 *Gapped k-mer support vector machine classification of random STARR-seq data*

In Publication II, we used previously published gapped k-mer SVM (support vector machine) framework [121, 122] to train an additional random enhancer STARR-seq model for comparison with the logistic regression and CNN models. The gapped k-mer SVM models were trained on balanced sets of high-confidence random enhancer STARR-seq sequences, defined as active enhancers observed in both GP5d random enhancer STARR-seq replicates (70% of sequences in training, 15% in validation and 15% in test sets). This is because the full random enhancer STARR-seq dataset was too big for the gapped k-mer SVM run to finish in approximately one month. The full GP5d random enhancer STARR-seq dataset consists of approximately 11.5 million sequences, while the high-confidence set has

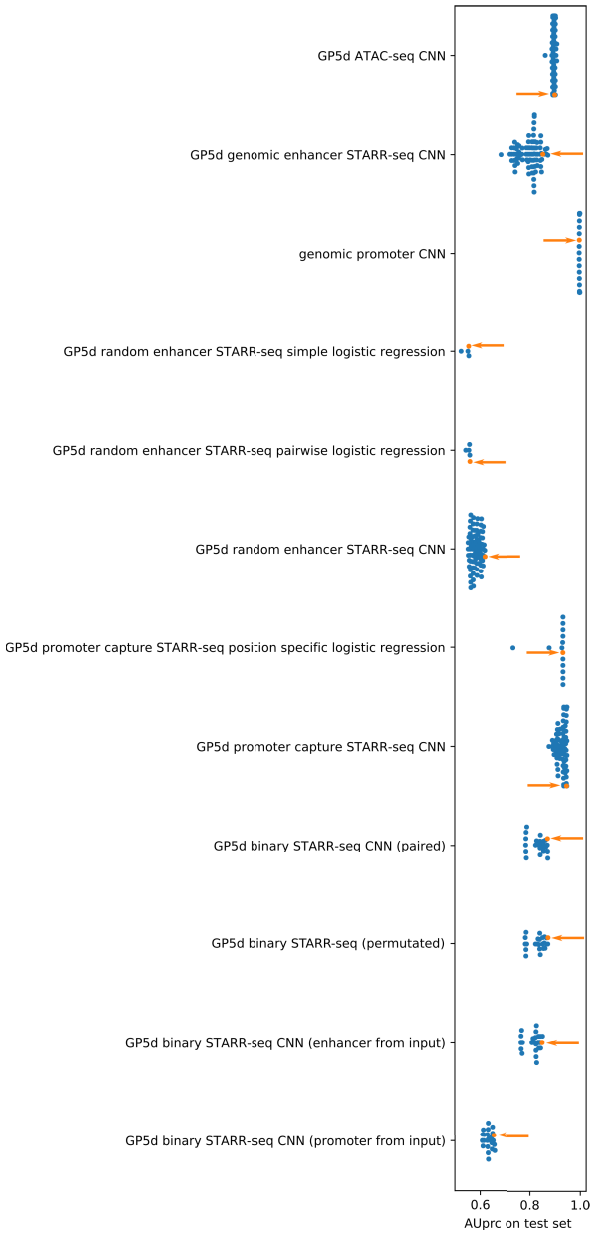

FIGURE 3.7: Test set classification performance (area under precision-recall curve, AUprc, balanced test set) for the main CNN and logistic regression models trained in this study. Each tested hyperparameter combination is marked with a dot and the representative "best" model, based on classification performance on separate validation data, is highlighted with orange color and arrow. Figure modified and adapted from Publication II.

less than three percent of that number of sequences. Optimal hyperparameter combination was selected based on area under precision-recall curve on the validation data.

3.3.5 *Prediction of differential gene expression using lasso regression*

In Publication II we used lasso regression [123] to predict differential gene expression between the HepG2 and GP5d cell lines and especially to study which types of enhancers and promoters are important in these predictions. Lasso regression was used because L1 regularization should not split the effects of correlating features but to select only one of the features and drive the regression coefficient of the other one to zero [108]. Non-zero regression coefficients can be interpreted as evidence towards such features having real predictive effect on differential gene expression. Logarithmic fold change between GP5d and HepG2 expression values (transcripts per million, tpm) was used as the target variable for regression and the STARR-seq and ATAC-seq peaks were divided into 12 features based on the cell line in which the peaks were present and the information about whether the ATAC-seq peaks were promoter proximal (less than 1kb distance to any gene in the target gene set) or distal. The features were named in such a way that for example "Common.STARR.noATAC" means STARR-seq peaks present in both cell lines that do not overlap with an ATAC-seq peak. All features used in the regression model are shown in Figure 4.7.

Using these features we built a somewhat heuristic model, where the effect of each feature was expected to decay following an exponential function. For each feature, the logarithmic fold change at peak summit reported by MACS2 (LFC) was used as the "strength" of the feature - the higher the peak, the more the feature affects gene expression. Peak summit position was used as the position of the feature. ATAC-seq peak summits and fold changes were used for all features except the ones that had no overlap with ATAC-seq. For STARR-seq-only features, STARR-seq peak summit positions and fold changes were used. The effect of a feature to a TSS that is d bp away from the peak summit corresponding to that feature was calculated as

$$S(peak) = LFC \times \exp(-c \times d / d_{max}), \tag{3.7}$$

where LFC is the logarithmic fold change at peak summit, c is a scaling parameter and d_{max} is the maximum distance of a peak from the TSS. The peak score S was used to quantify the effect of each feature in the regression model. No intercept term was included in the model.

3.3.5.1 *Training the differential expression predictors*

The differential gene expression predictors were implemented using scikit-learn [112]. The target genes were split into training (8815 genes), validation (1449 genes) and test (2321 genes) sets according to chromosomes they are in as listed in Table 3.3.3.2. The genes in this analysis were filtered so that genes with smallest overall tpm (transcript per million) values (mean tpm < 2 over all experiments) were discarded to avoid them dominating the fit of the model, in the spirit of "independent filtering" [124].

Optimal regularization strength of the lasso regression model was determined using 5-fold cross-validation during training The other model hyperparameters c and d_{max} were optimized using the validation data set with a grid search. Coefficient of determination was used to select the optimal hyperparameters.

3.3.6 *Pre-trained machine learning models used*

In Publication III, the following pre-trained machine learning models were used to demonstrate the ability of the proposed machine learning model interpretation method to highlight dependencies learned by different machine learning models: DeepBind [42] models D00198.001 (RBMS1) and D00123.001 (MSI1) predicting RNA-binding protein binding to DNA sequences; N-score model [125] predicting nucleosome favoring DNA sequences; The sequence convolutional neural network, graph convolutional neural network and linear regression models predicting GB1 protein domain fitness [126].

3.3.7 *Convolutional neural network classifier interpretation strategies used*

In addition to the novel PlotMI and Nsweep approaches described in detail in Results, several other strategies were used to interpret the features learned by the CNN models trained in this study. These approaches mostly relied on generating synthetic DNA sequences *in silico*, scoring the sequences with the CNN models and either observing the scores for sequences with known features embedded, or studying features enriched in high-scoring random samples. Also the previously published deep learning model interpretation tool TF-MoDisCo [80, 88] was used in Publication II.

To test which TF binding motifs the CNN trained on the random enhancer STARR-seq data had learned, the CNN was used to score random sequences with a known embedded binding motif in them. 100 sequences drawn randomly from each 880 PWMs used also as features in the logistic regression classifiers were embedded to random enhancer STARR-seq input library sequences in such a way that each sequence was embedded at a random position to one of 100 different randomly chosen input sequences (same input sequences used for each PWM) and the average enhancer probability over the 100 sequences was calculated for each PWM. When embedding a single PWM per input sequence, first, the position for the embedding was drawn from uniform distribution. Next, the embedded sequence was drawn at random from the corresponding PWM. When embedding a motif pair, first the positions of the embedded sequences were drawn at random, but not allowing overlap between the embedded sequences. Then, both of the embedded sequences were drawn independently from the corresponding PWMs. This means that both the positions of the embedded sequences and the distance between the embedded sequences are random. The expected enhancer probability for a sequence with two embedded PWMs given that there are no interactions between them is

$$p_2 = 1 - (1 - p_1)^2, \tag{3.8}$$

where p_i is the enhancer probability of a sequence with i PWMs embedded. Thus p_2 is the cumulative probability for geometric distribution with two trials.

As an orthogonal and supporting approach to scoring sequences with known features in them, we also used the CNN to score completely random synthetic DNA sequences, selected a set of highest-scoring sequences and performed *de novo* motif mining on these sequences to discover the most enriched motif patterns in them. This approach will reveal the motif patterns in enhancers that the trained CNN model predicts to be the most likely active enhancers. For this analysis, we created a set of 10 million 170 bp long *in silico* DNA sequences that were sampled from uniform nucleotide distribution. These sequences were scored with the CNN model trained on the random enhancer STARR-seq data, and the top

0.5% of the sequences (50,000) obtaining the highest predicted enhancer probabilities were selected for motif mining analysis using the STREME [127] program.

The sequence features found from the TERT promoter and its variants by the CNN model trained on the binary STARR-seq TSS-aligned promoters were visualized with DeepLIFT [86] software. The activation values calculated with DeepLIFT using the CNN model from the wild type and mutated promoter sequences were compared against 15 randomly chosen sequences from the random promoter STARR-seq input and their average activation signals were visualized as sequence logos.

3.3.8 *Validation of the predicted variant effects with saturation mutagenesis data of the TERT promoter*

Saturation mutagenesis study of the TERT promoter [128] was used to test if the variant effects predicted by the CNN model trained on the binary STARR-seq TSS-aligned promoters correlate with measured activity changes. Note that the CNN model has not seen any saturation mutagenesis data during training. The statistical significance of the mutation effects predicted by the CNN was estimated by first scaling the predicted promoter probabilities ($P_{promoter}$) between $-\inf$ and $\inf$ by transforming them into log odds scores:

$$logit(p) = \log(p/(1-p)). \tag{3.9}$$

Next, the predicted mutation effect (ME) for each mutation was calculated as the logarithm of odds ratio between the predicted promoter probability of the mutated and the wild type sequence:

$$ME = logit(p_{mutated} - p_{wt}). \tag{3.10}$$

To asses the significance of the predicted effects, an empirical p-value was calculated for each predicted TERT promoter ME by comparing if the predicted effect in the TERT promoter is more extreme than the predicted effect in shuffled TERT promoter sequences at the same position and for the same type of mutation.

For this, 10,000 shuffled versions of the wild type TERT promoter were generated where dinucleotide frequencies were preserved. All possible SNPs were introduced into each of the shuffled sequences. Then, the predicted ME was calculated for each of the SNPs as the logarithm of odds ratio against the $P_{promoter}$ of the corresponding shuffled wild type promoter sequence. For each position and mutation type, the empirical p-value was calculated as the fraction of the predicted mutation effects on the shuffled sequences that were more extreme than the predicted ME on the wild type TERT promoter. The non-significant mutations with $p - value > 0.05$, either according to the empirical p-value described above or the p-value from the saturation mutagenesis study, were filtered out from the correlation analyses.

3.4 COMPUTING MUTUAL INFORMATION BETWEEN POSITIONAL K-MER DISTRIBUTIONS IN SETS OF SEQUENCES

In Publication III we use mutual information (MI) to visualize pairwise dependencies learned by machine learning models trained on sequence data. In Publication II, MI is also used to discover interactions from STARR-seq sequences similarly to a previous study where MI was used to study interactions between TFs and the nucleosome [57]. In the

following, I will describe how MI is computed given a set of sequences S of equal length l in any defined alphabet (for example DNA, RNA or protein code).

Let us denote with $P_i(a)$ the observed frequency of k-mer a at position i in S, and with $P_{ij}(a,b)$ the observed joint frequency of k-mer a at position i and b at position j. A k-mer means a continuous subsequence of length k. The frequencies $P_i(a)$ are

$$P_i(a) = \frac{1}{N_S + \gamma}\left(\gamma/\alpha^k + \sum_{n=1}^{N_S} \delta(\kappa_i^n = a)\right). \tag{3.11}$$

Here the summation runs over the N_S sequences in set S and $\delta(\kappa_i^n = a)$ is Kronecker delta that equals to 1 only if k-mer κ at position i of sequence n is a.

To account for unobserved k-mers when dealing with finite samples, we add a "pseudocount mass" γ to the total count of k-mers. The normalization comes from the fact that there are α^k k-mers (where α is the alphabet length, for example 4 for DNA), and pseudocount mass γ is divided between all k-mers, while N_s k-mers are observed from the input sequences. Estimation of MI from finite samples is a non-trivial problem, and different approaches for estimating it have been thoroughly discussed in [129]. The assumptions of our pseudocount-based approach are 1) when the observed sample count of a k-mer approaches zero, the estimated probability of observing this k-mer approaches a non-zero constant and 2) $P_i(a) = \sum_{j,b} P_{i,j}(a,b)$.

Similarly, the observed joint frequency of k-mers a and b is

$$P_{ij}(a,b) = \frac{1}{N_S + \gamma}\left(\gamma/\alpha^{2k} + \sum_{n=1}^{N_S} \delta(\kappa_i^n = a, \eta_j^n = b)\right), \tag{3.12}$$

where the number of 2k-mers is α^{2k}. Note that this is equivalent to counting gapped 2k-mers where gap length is equal to the distance between the k-mers. Pseudocount mass γ is added also to the 2k-mers, as each 2k-mer is a combination of two k-mers, and thus each k-mer pair was added a pseudocount of γ/α^{2k}. With these observed frequencies, one can compute MI (originally described in [130]) between pairs of positions in the set S as

$$MI_{ij} = \sum_{a\in K}\sum_{b\in K} P_{ij}(a,b)\log_2\left(\frac{P_{ij}(a,b)}{P_i(a)P_j(b)}\right), \tag{3.13}$$

where K is the set of all k-mers. Only position pairs where the k-mer distributions do not overlap with each other are considered in the visualization.

3.5 EXPERIMENTAL DATA GENERATED IN THIS STUDY

In Publication I, Dr. Biswajyoti Sahu performed the CTCF ChIP-nexus experiment in human LoVo colon cancer cells.

In Publication II, Dr. Biswajyoti Sahu performed the following experiments with help from Dr. Päivi Pihlajamaa: motif STARR-seq in GP5d human colon cancer cells; random enhancer STARR-seq in GP5d and HepG2 human liver cancer cells; binary STARR-seq in GP5d; HepG2 and RPE human retinal pigmented epithelium cells; genomic STARR-seq in GP5d and HepG2 cells; generation of TP53-null GP5d cell line using CRISPR-Cas9 genome editing; template switch for TSS position determination from binary STARR-seq experiment; H3K27 acetylation, TP53 and IRF3 ChIP-seq experiments in HepG2 cells; H3K27 acetylation, TP53, CTCF, SMC1, H3K9 trimethylation, FOXA1, HNF4A, MYC, TCF7L2 and H3K27

trimethylation ChIP-seq experiments in GP5d cells; ATAC-seq experiments in GP5d and HepG2 cells; RNA-seq in HepG2 and GP5d cells.

In Publication II, Drs. Kashyap Dave and Carsten O. Daub generated the GP5d CAGE experimental data and Dr. Bei Wei the GP5d ATI experimental data.

3.6 DATABASES AND PUBLISHED DATASETS USED

The MAX and TWIST *Drosophila melanogaster* ChIP-nexus experiments used in Publication I were published in [11]. The human CTCF ChIP-exo experiments used in Publication I were published in [37] and [12]. Variant calls used in Publication I were obtained from the 1000 Genomes database [131].

Most of the PWM models of TF binding motifs used in Publications I, II & III were originally published in [3, 8, 56], with the exception of *Drosophila melanogaster* PWMs (MAX: MA0058.3, TWIST: MA0249.1) used in Publication I which were obtained from the JASPAR-database [132] and the core promoter motifs used in Publication II which were curated from the following publications: TATA box, Initiator, CCAAT-box, GC-box from [133] and BRE, MTE, DPE from [134].

The structure of GB1 protein domain (Protein Data Bank ID: 2QMT [135]) was downloaded from the Protein Data Bank [136].

The HepG2 TF ChIP-seq, ATAC-seq and histone modification ChIP-seq data were downloaded from the ENCODE data portal [4]. The human transcription start sites used in Publications II and III were obtained from the Eukaryotic Promoter Database [137]. The pre-computed hg19 reference genome GERP conservation scores used in Publication II were published in [138].

For gene set enrichment analysis in Publication II, the gene lists of p53 and interferon signaling pathways were obtained from the Molecular Signatures Database [139].

3.7 PUBLISHED SOFTWARE USED

The Burrows-Wheeler aligner [91] was used in Publications I and II to align next generation sequencing data to reference, in Publication II, also Bowtie2 [103] was used. In Publication II, RNA-seq transcript-level counts were estimated using Kallisto [140] and differential expression analysis was performed with Sleuth [141]. Samtools [142] and Bedtools [143] were used to process bam- and bed-files in Publications I and II. MACS2 [100] was used for peak calling in Publication II.

FastQC `https://www.bioinformatics.babraham.ac.uk/projects/fastqc/` was used for sequencing data quality assessment in Publications I and II and CutAdapt [144] and TrimGalore `https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/` were used for adapter trimming in Publication II.

Matching of known TF binding motifs to DNA sequences was conducted using MOODS [109] in Publications I and II. MEME suite [145] tools were used for *de novo* motif discovery and motif identification in Publications I, II and III. In Publication II, also Autoseed [146], and HOMER [147] were used for *de novo* motif discovery.

In Publication II, also the following software tools were used: ROSE pipeline [148] for super enhancer calling, Picard `http://broadinstitute.github.io/picard/` for ATAC-seq preprocessing, lfc [111] for estimating fold changes between RNA and input DNA, FLASH [149] for STARR-seq data preprocessing, Preseq [150] and Starcode [151] for com-

plexity estimation of the STARR-seq libraries, IDR [152] for merging replicate experiments, and paraclu [153] for calling CAGE read clusters.

4

RESULTS

4.1 PEAKXUS: A COMPUTATIONAL TOOL FOR ACCURATE TRANSCRIPTION FACTOR BINDING SITE DISCOVERY FROM CHIP-EXO AND CHIP-NEXUS EXPERIMENTAL DATA

Modifications of ChIP-seq, ChIP-exo [12] and ChIP-nexus [11] introduce an additional λ-exonuclease digestion step that brings the resolution of *in vivo* TF-DNA binding studies to one base pair regime. However, this modification leads to different binding signal meaning software developed for ChIP-seq data analysis are not optimal for ChIP-exo/nexus. In Publication I, we develop a peak calling algorithm, PeakXus, tailored for ChIP-exo and ChIP-nexus data and show that it has several desired properties over earlier methods in the literature: 1) PeakXus reports more TF-DNA binding sites that overlap with TF-specific binding motif. 2) PeakXus makes less assumptions about the shape of the signal than its competitors. 3) By using Unique Molecular Identifiers (UMIs) [96] to filter out duplicated reads, PeakXus is better able to separate true binding events from experimental artefacts such as PCR bias.

4.1.1 *The peak calling algorithm*

Figure 4.1a lists the main steps of PeakXus algorithm. The algorithm takes as input sequencing reads from a ChIP-exo/nexus experiment and the UMI corresponding to each read (UMIs can be omitted if necessary) and outputs a list of TF binding sites, or peaks, in the genome. The first step of the algorithm is to filter out reads with identical UMIs that map to an identical position. Each molecule in the initial library is assigned its own UMI at random. Given that the UMIs have been designed properly, it is highly unlikely that two different molecules with same UMI map exactly to the same position in the genome. Thus UMIs offer a robust and simple way to avoid including for example PCR duplicates into the downstream analyses.

Next, read 5′ end counts at each position of the genome are saved for use in downstream analysis. Read 5′ ends are used as these are a proxy of the λ-exonuclease stop positions and contain the most accurate information of the 5′ end location of the bound TF. A list of candidate binding sites is then produced following an algorithm outlined in Figure 4.1b: we iterate through the genome looking for positions where the total read 5′ end count $c_+(i) - c_-(i) < 0$ and $c_+(i-1) - c_-(i-1) \geq 0$. Here, $c_+(i)$ is the + strand read 5′-end count at position i, and $c_-(i)$ the same for the - strand. For each such transition point where total read 5′ end count changes sign, index of the left border of the candidate peak is $k = \{c_+(k) - c_-(k)\}$ such that $i - w < k < i$ and $c_+(k) - c_-(k) > 0$. Similarly for each such position, index of the right border of the candidate peak is $j = \{c_+(j) - c_-(j)\}$ such that $i \leq j < i + w$ and $c_+(j) - c_-(j) < 0$.

The list of candidate peaks computed as described above can contain overlapping peak candidates. In the final step of the algorithm, we perform significance testing, peak ranking and removal of overlapping candidate peaks. The significance testing method employed here is motivated by two main features of ChIP-exo/ChIP-nexus TF-DNA binding events.

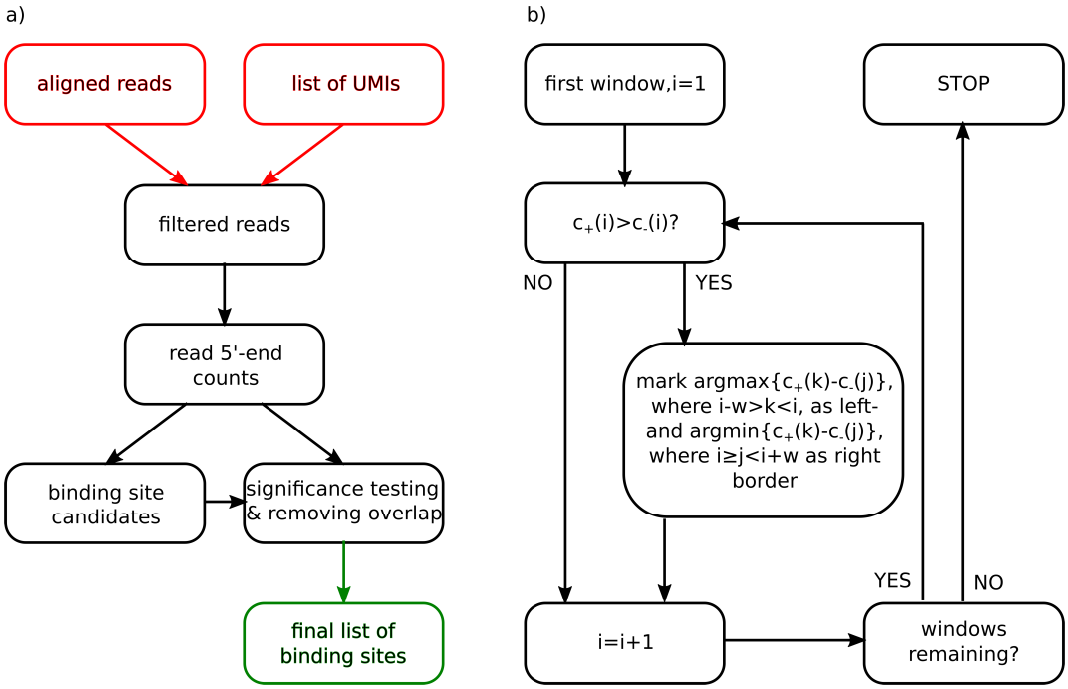

FIGURE 4.1: a) Schematic presentation of the main steps of PeakXus algorithm. Input for the algorithm is marked with red and output with green. b) Schematic presentation of candidate binding site discovery. Whenever the total read 5′ end count is negative, the position within w bp to the left from i with highest sense strand read count $c_+(i)$ is marked as the left border of a candidate peak/binding site. Similarly, the position within w bp to the right from i with the highest antisense strand read count $c_-(i)$ is marked as the right border of a candidate binding site. Here w denotes the widest allowed peak and is an adjustable parameter.

First, true binding events have a high total read 5′ end count around the binding site, and secondly, stopping of the λ-exonuclease when it encounters a bound protein creates borders, or narrow peaks of mapped read 5′ ends, to opposite strands flanking the binding site. See Figure 4.2a for a schematic example of how a ChIP-exo/nexus binding event looks like on the read/UMI level data.

Because the λ-exonuclease digests DNA from 5′ to 3′ direction, theoretically only reads pointing towards the center of a candidate peak come from true binding events. Thus the positions of a mixture of reads resulting from the binding events together with background noise point towards the binding site center. On the other hand, the reads pointing away from the binding site center contain only the background noise. Thus, we compute the distribution of distances between read 5′ ends and the candidate binding site center for the background reads only (reads pointing away from the candidate binding site center, blue background in Figure 4.2) and compare it with the distribution of distances between read 5′ ends and the candidate binding site center for the reads containing the signal and the background reads (reads pointing towards the candidate peak center, red background in Figure 4.2). If these two distributions are significantly different, the candidate binding site is accepted as a real TF-DNA binding event.

The significance testing is conducted using the G-test (see e.g. [154]) and multiple hypothesis correction is done using the the Benjamini-Hochberg procedure [155]. Finally, we compute a peak score SC^{kj} for each candidate peak that is used for final ranking of the peaks. Peak score is also used to filter overlapping peaks so that only the peak with highest peak score among overlapping peaks is kept. The peak score for a peak with borders k and j is defined as

$$SC^{kj} = G_p^{kj} \left(\sum_{i=k_{kj}-d}^{\lfloor m \rfloor} c_+(i) - c_-(i) + \sum_{i=\lceil m \rceil+1}^{j_{jk}+d} c_-(i) - c_+(i) \right), \tag{4.1}$$

where the middle position of the candidate peak is $m = (k_{kj} + j_{kj})/2$, d is a parameter allowing some variation for the λ-exonuclease stop position (by default $d = 5$) and G_p^{kj} is the G-test test statistic value. Default value of pseudocount used in computing the distance histograms is $p = 1$. To summarize, the following steps are performed for all candidate peaks:

1. Calculate the p-value using G-test. If p-value is higher than a predefined threshold value, discard the candidate.
2. Calculate the peak score for all remaining peaks.
3. Find all sets of overlapping peaks, and discard all others but the peak with the highest peak score from each set.
4. Calculate false discovery rates using the Benjamini-Hochberg procedure, using the initial number of candidate peaks as the total number of tested null hypotheses.

4.1.2 *Comparison to other peak callers*

In Publication I, we show that PeakXus finds more peaks overlapping with high-affinity recognition sites (HARSs), defined as genomic sites with a high-affinity match to a known binding motif of a specific TF, than the earlier published methods (Peakzilla [156], MACE

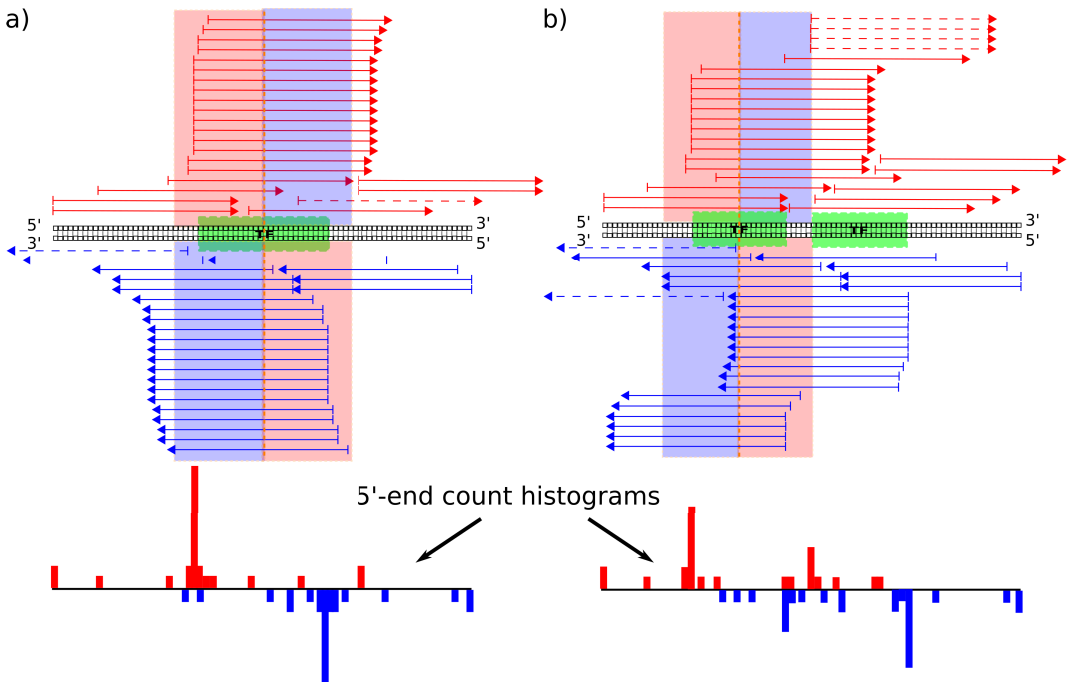

FIGURE 4.2: Schematic presentation of determining the significance of peak candidates with PeakXus. a) Determination of a candidate peak in the presence of one true TF-DNA binding event (green area marks the positions occupied by the TF). b) Determination of a candidate peak in the presence of two true binding events. Red arrows correspond to reads mapped to the sense strand and blue to the antisense strand. Reads point from 5′ to 3′ direction. The red and blue bar charts below the reads correspond to counts of 5′ ends of reads (or UMIs) on the sense and antisense strands, respectively. Because the λ-exonuclease stops at the border of the bound TF, reads pointing towards the candidate peak center (the middle position between the borders on the sense- and antisense strands) are assumed to be true signal, while reads pointing away from the candidate peak center are assumed to originate completely from other sources such as noise. Distance distributions of read 5′ ends and the candidate peak center are compared between the regions on red (signal) and blue (background/noise) backgrounds. If the distributions are significantly different, candidate binding site is marked as a true TF-DNA binding site. Figure modified and adapted from Publication I.

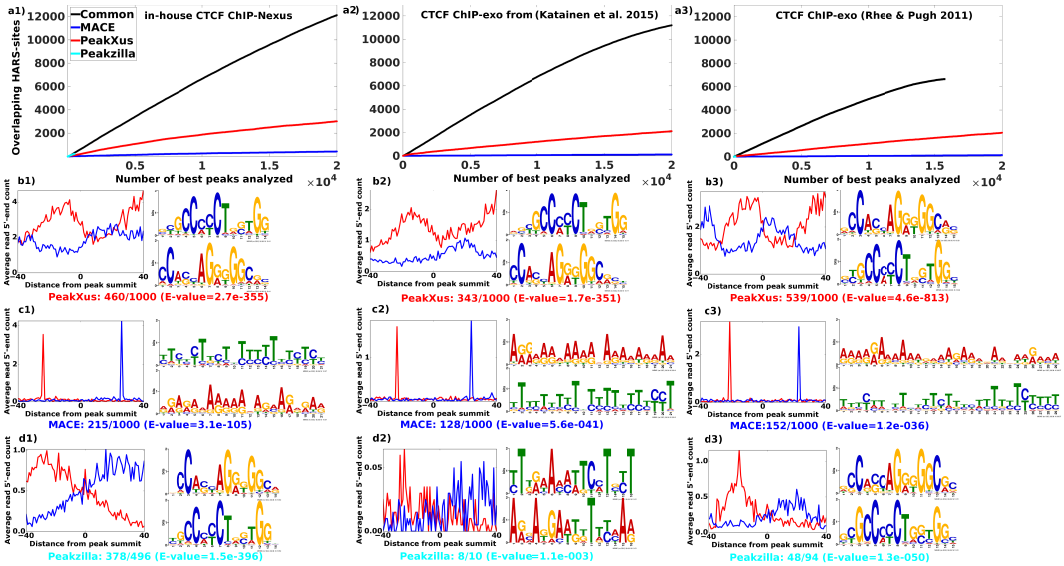

FIGURE 4.3: Analysis of peak caller-specific peaks in three experiments measuring binding of CTCF TF. The peak callers are color coded as: cyan = Peakzilla, blue = MACE, red = PeakXus, and the peaks reported by all methods are plotted in black. Left column (from a1 to d1) shows results from the in-house ChIP-nexus experiment, middle column (from a2 to d2) from the ChIP-exo experiment from [37] and right column (from a3 to d3) from the ChIP-exo experiment from [12]. a1-a3) The number of high-affinity binding sites (HARSs) overlapping with x (value shown on x-axis) top peaks (ranked by method-specific peak score) is shown on y-axis. For clarity, 20,000 highest-scoring peaks are shown. A peak was considered to match with an HARS if distance between the HARS and the peak centers was $\leq$ 20 bp. Only the peaks not overlapping with any peaks found by the other two methods were included in the sets of peak caller-specific peaks. Peakzilla-specific peaks are rare (in-house ChIP-Nexus: 496, Katainen et al. ChIP-exo: 11, Rhee & Pugh ChIP-exo: 94), rendering the corresponding curves hardly visible. b1-b3) More detailed analysis of the top 1000 peaks specific to PeakXus. Each column shows on the left the average read 5' end count profile around the peak center for sense (red) antisense (blue) strands. On the right are shown both orientations of the highest-confidence *de novo* motif from the top 1000 method-specific peaks reported by MEME. The number of occurrences of the best scoring MEME-motif is shown below each of the motif pairs along with the corresponding MEME E-value. c1-c3) More detailed analysis of the top 1000 (or less, if 1000 were not found) peaks specific to Peakzilla and d1-d3) MACE. Figure modified and adapted from Publication I.

[157] or GeneTrack [158]). This does not directly measure "goodness" of a peak caller, as it is well established that TFs also bind locations without HARS to their corresponding binding motif *in vivo* (possibly due to for example non-specific binding to open chromatin), but is a good indicator of how many of the sites where the binding mechanism can be explained by the binding motif the peak caller finds. Moreover, we also studied the peaks commonly reported by the three peak callers PeakXus, Peakzilla and MACE, and the peaks reported specifically by each of the methods and not by the others.

Unsurprisingly, the peaks found by all three methods overlapped more likely with HARSs than peaks specific to any of the individual methods (Figure 4.3a1, a2 & a3). Importantly, PeakXus-specific peaks overlap in total more likely with an HARS than peaks specific to MACE or Peakzilla. Peakzilla reported only less than 100 peaks the other two methods did not from the three tested experiments which is why the Peakzilla curves are not visible in Figure 4.3. Curiously enough, *de novo* motif mining with MEME [159] does not find the CTCF binding motif from MACE-specific peaks, contrary to PeakXus (3/3 experiments) and Peakzilla (2/3 experiments). Moreover, when reads with identical strand and 5′ end positions were removed and the analysis re-run, MACE reproduced only 2,005 of the 99,564 peaks from ChIP-exo experiment by Katainen et. al. [37], and 10,549 of the 54,510 peaks from CTCF ChIP-exo experiment by Rhee & Pugh [12]. MACE was unable to find any peaks from our in-house ChIP-nexus data after duplicate read removal. This observation highlights the difficulty of separating true ChIP-exo/nexus binding events from artefacts created by PCR-duplicates if UMIs [96] are not used in the experiment and supported by the peak caller.

4.1.3 *Allele specific binding analysis algorithm for ChIP-exo/nexus data*

As an application example of PeakXus and the use of UMIs to filter out duplicated reads, we developed an improved algorithm for studying allele specificity of TF binding with ChIP-exo/nexus experiments in Publication I. The main improvements in our algorithm compared to earlier work in studying allele specific binding (ASB) with ChIP-seq experiments are: 1) ChIP-exo and ChIP-nexus are better suited for ASB analysis than ChIP-seq as the λ-exonuclease treatment causes the reads originating from a single binding event to cluster more tightly on top of the binding site giving higher coverage of reads on top of polymorphism sites overlapping binding motifs of TFs. 2) Use of UMIs to filter out duplicated reads makes controlling PCR bias easier compared to earlier methods used in ChIP-seq ASB studies [160]. 3) To control for uncertainty of genomic allelic ratios (gARs) [161, 162] determined using whole genome sequencing (WGS), we use the Audic-Claverie test [163] to compute significance of allele specific binding instead of the binomial test which assumes the genomic allelic ratios are unbiased.

Figure 4.4 summarizes the main steps of the ASB analysis algorithm. The algorithm takes as input the locations of called peaks from PeakXus, locations of known single nucleotide polymorphism (SNP) sites, reads from a WGS experiment performed in the same cell type as the ChIP-exo/nexus experiment, and the aligned reads and UMIs from the ChIP-exo/nexus experiment. ASB analysis is carried conducted for peaks that overlap with a SNP or SNPs. For the analysis, we compute the number of unique reads mapping to each of the alleles both in the WGS and in the ChIP-exo/nexus experiment. Other steps of the algorithm, except significance testing, are rather straightforward. In the following I will describe how significance of ASB is assessed.

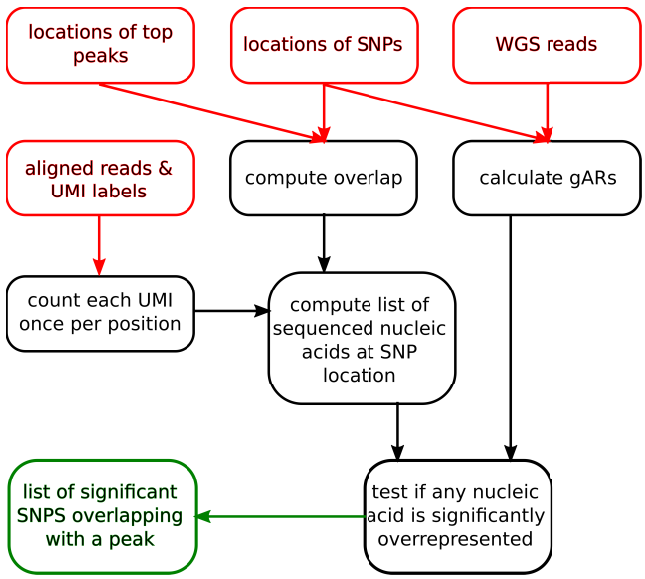

FIGURE 4.4: Schematic presentation of the main steps of the allele specific binding analysis algorithm. UMIs are used to filter reads so that each unique UMI label is counted only once per position and strand. Input of the algorithm is marked with red and output with green.

The goal is to compute if the fraction of reads mapping to the reference allele in the ChIP-exo/nexus experiment is significantly different from the fraction of reads mapping to the reference allele in the WGS experiment for a given SNP. The Audic-Claverie test [163] requires that the observed events tested are rare and part of a large population of possible outcomes. This assumption is satisfied as each read/UMI has approximately a probability of one over the size of the genome for mapping to a specific position. The number of possible mapping locations is proportional to the size of the genome. The Audic-Claverie distribution is a general probability distribution governing occurrence of the same rare event in duplicate experiments. The WGS and the ChIP-exo/nexus experiments are viewed as replicate experiments with k reads that overlap with the SNP i mapping to reference allele in the ChIP-exo/nexus experiment and n reads mapping to the reference allele in the WGS experiment. Following [163], the probability of observing k given n by chance is

$$P_i(k|n) = \left(\frac{N_2}{N_1}\right)^n \frac{(k+n)!}{k!n!(1+N_2/N_1)^{k+n+1}}, \tag{4.2}$$

where N_1 is the total number of reads overlapping with the SNP i in ChIP-exo/nexus and N_2 in the WGS experiment. $P_i(k|n)$ is calculated for each SNP. If $P_i(k|n) < 0.01$, we conclude that there is significant difference in binding the TF measured in the ChIP-exo/nexus experiment between the two alleles.

We demonstrated the ASB analysis algorithm described above with an in-house CTCF ChIP-nexus experiment. In Publication I, we show that using UMIs to filter out duplicated reads, a higher fraction of SNPs overlapping a CTCF HARS show significant allele specific binding compared to SNPs not overlapping a CTCF HARS, than if the analysis is conducted using raw read counts. It is reasonable to expect that SNPs disrupting the binding motif of CTCF affect the binding of the TF more than more distal SNPs. This result suggests that

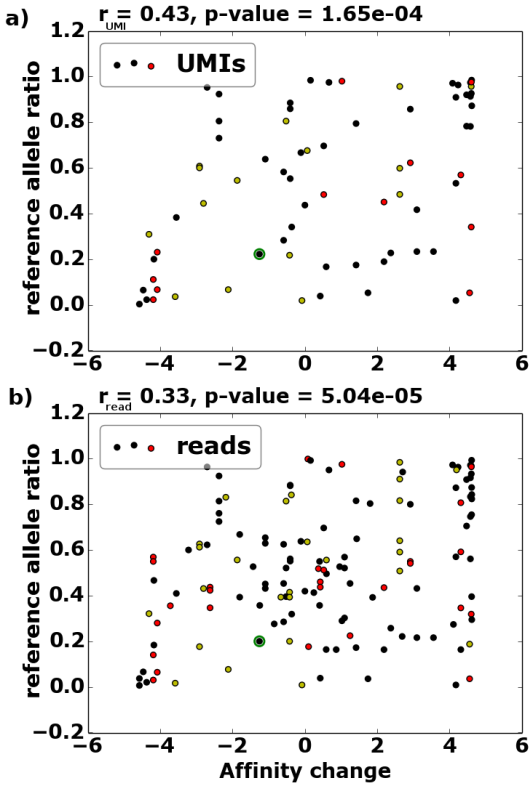

FIGURE 4.5: Reference allele ratio as a function of binding affinity change in the in-house CTCF ChIP-nexus experiment. Reference allele ratio is shown on the y-axis, while x-axis is the affinity change (reference minus alternate sequence affinity to the CTCF binding motif). SNPs with p-value > 0.01 are not shown. Red dots represent SNPs where alternate allele creates a CG-site to the sequence but reference does not. Yellow dots represent SNPs where sequence with reference allele has an extra CG dinucleotide. Other SNPs are black. The green circle marks the only SNP overlapping with an imprinting control region from [164] that overlapped with the ChIP-nexus peaks. Values of Pearson correlation coefficients (r) along with the corresponding p-values are shown above the panels. a) p-values were computed using UMIs (73 significant SNPs). b) p-values were computed using raw read counts (142 significant SNPs). Figure modified and adapted from Publication I.

using UMIs in ASB analysis leads to fewer false positive ASB events than conducting the analysis without UMIs.

In Figure 4.5, we show how the CTCF binding motif affinity change caused by a SNP correlates with observed ASB in the ChIP-nexus experiment using UMIs (upper panel) and without using UMIs (lower panel). The correlation between the reference allele ratio in the ChIP-nexus experiment and the affinity change of the binding motif induced by the SNP is stronger when using UMIs indicating that UMIs help to discard false positive ASB events. Moreover, when the analysis was repeated otherwise similarly, but calculating the significance using the two-sided binomial test according to the previous literature, the correlation coefficients were: $r_{UMI,binom} = 0.38$ (p-value=$2.20e-05$, 115 significant SNPs vs 73 significant SNPs using the Audic-Claverie test) for analysis with UMIs and $r_{read,binom} = 0.31$ (p-value=$5.23e-05$, 168 significant SNPs vs 142 significant SNPs using the Audic-Claverie test). Testing the significance of ASB with the Audic-Claverie test results to fewer significant SNPs but also to stronger correlation between the reference allele ratio and the affinity change. This suggests that using a significance test that accounts for the uncertainty of the gARs helps to capture those ASB events that correlate with the affinity change of the binding motif.

To summarize, in Publication I we describe a novel peak calling algorithm PeakXus, designed to accurately call TF-DNA binding events from ChIP-exo and ChIP-nexus experimental data while making as few assumptions about the binding events as possible to allow discovery of possible new binding modes and patterns. We show that PeakXus reports more peaks that overlap with the TF-specific HARSs than the methods published earlier. In addition, we describe an improved algorithm for studying allele specificity of TF-DNA binding in ChIP-exo/nexus experiments. Both algorithms were made publicly available for other researchers to use via GitHub.

4.2 Sequence determinants of human gene regulatory elements

In Publication II, we employ massively parallel reporter assays (MPRAs), and modern machine learning methods to directly study transcriptional activities of human promoters and enhancers. MPRAs have been previously utilized in genome-wide studies of gene regulatory element activities in yeast [21, 165], fruit fly [13, 166] and humans [14–19]. The MPRA designs we utilized in Publication II (see Figure 3.3) have two major advantages over the previous studies in human cells: First, in addition to a traditional design using human genomic sequences, we also measure enhancer and promoter activities of designed sequences and completely random synthetic sequences. These approaches help to avoid the problems related to discovery of sequence determinants of gene regulatory activity from the human genome which is repetitive and evolved to perform also other functions in addition to transcription. Second, we measure gene regulatory activities of sequences that in total cover 100 times the sequence space of the human genome. In the following, the focus is on the machine learning results from Publication II as those were the major contribution of the Author.

4.2.1 *Sequence determinants of enhancers needed for transcriptional activity*

In Publication II, we performed several STARR-seq experiments to characterize the set of DNA sequence features of human enhancers needed for transcriptional activity. We trained machine learning models both on the genomic fragments that functioned as active

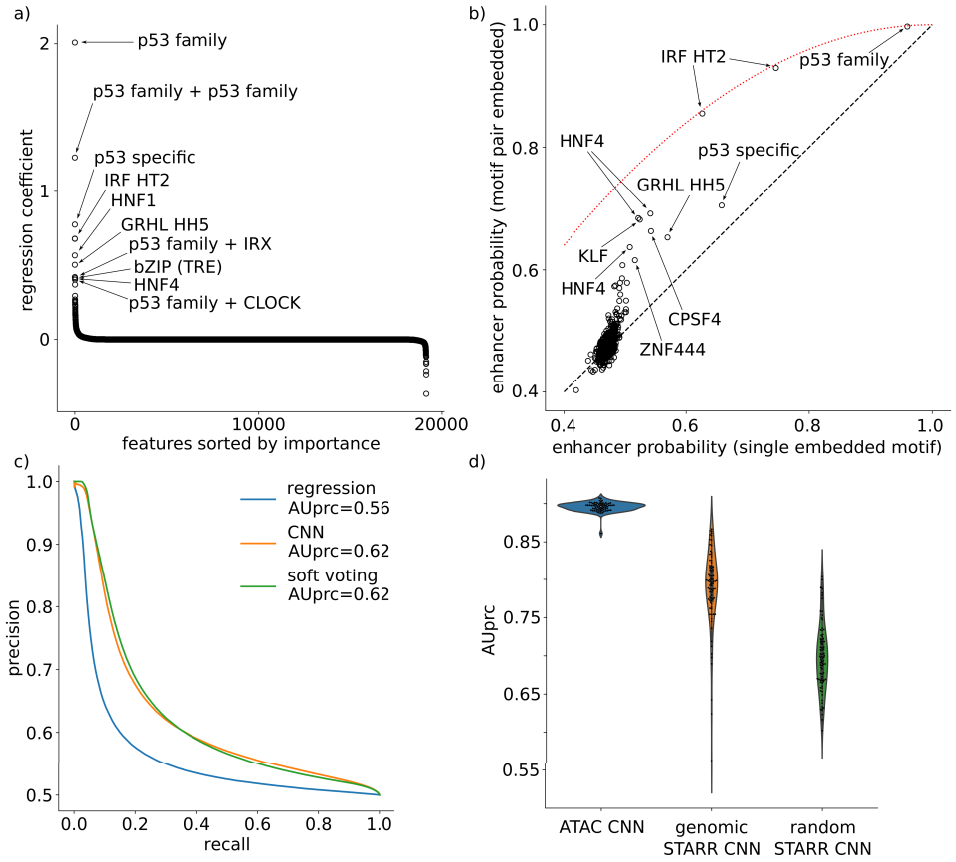

FIGURE 4.6: a) Regression coefficients for TFs and TF pairs from logistic regression (LR) analysis of GP5d random enhancer STARR-seq, with the most important features labeled. b) Inserting two instances of the same motif to randomly selected input library sequences increased the predicted enhancer probability by the random enhancer STARR-seq CNN above that expected from a single motif (dashed black line), but not above expectation from a model assuming independent binding to two motifs (red dotted line). c) Classification (balanced test set) performance of models trained to separate inactive (input) and active GP5d random enhancer STARR-seq sequences. CNN (orange) outperforms the LR model (blue). Soft voting (green) combining the predictions of the CNN and the LR does not improve over the CNN (area under precision-recall curve, AUprc), indicating that the predictive features of the LR are also learned by the CNN. d) Violin plots showing AUprc in binary classification (balanced test set) between GP5d ATAC-seq peaks and control genomic sequences for CNNs trained on: ATAC-seq (blue), genomic enhancer (orange), and random enhancer (green) data. Dots are unique hyperparameter combinations. Figure modified and adapted from Publication II.

enhancers (design ii, genomic enhancer STARR-seq), and on the enhancers enriched from completely random synthetic sequences (design iii, random enhancer STARR-seq). We modeled the random enhancer STARR-seq experiment with machine learning classifiers as described in Methods. The most predictive features in a logistic regression classifier using single TF binding motifs, and a selected subset of pairs of binding motifs as features (in total 19,360 features) were highly similar to the motifs observed from motif matching analyses from the motif, and the random enhancer libraries (see Figure 4.6a, and Publication II). Very few features describing pairs of motifs had strong predictive power, with the "p53 family-p53 family" being the most prominent of the pairwise features, consistent with the strongest observed spacing in an analysis searching for enriched spacings between motifs observed between p53 motifs (see Publication II).

We then trained convolutional neural network (CNN)-based classifiers on the STARR-seq data. These methods do not rely on the assumption of known TF binding motifs, but can in principle learn any type of sequence features responsible for the observed activities. The CNNs classified unseen test data better than the logistic regression classifier (Figure 4.6c). Moreover, a soft voting classifier combining the predictions of the CNN and the regression models did not improve over the classification obtained by the CNN suggesting that the logistic regression model did not learn information that would be missing from the CNN.

Comprehensive analysis of the features learned by the CNN classifier presented in Publication II indicated that the features learned by the CNN were largely similar to the known TF binding motifs, but in many cases the weaker bases of the motifs learned by the CNN were different than in the known binding motifs, suggesting that this was largely the reason for better model performance. Also the activities of the known binding motifs learned by the CNN were similar to activities observed using the motif matching analysis from the motif and random enhancer libraries with highest learned activities for motifs such as p53 family, p53 specific, IRF HT2 and GRHL HH5 (Figure 4.6b, Publication II). Consistent with the other analyses, the CNN also did not learn evidence for beyond additive interactions between instances of the same binding motif.

We next used the CNN model trained on the random enhancer STARR-seq data to predict which sequences in the GP5d cells reside within open chromatin and which not, according to an ATAC-seq experiment in the same cell line. CNN models trained on the random and genomic enhancer STARR-seq experiments were able to classify the ATAC-seq data relatively well, but were still clearly outperformed by a CNN trained on the ATAC-seq data (Figure 4.6d). This indicates that only part of the ATAC-seq peaks in the genome are explained by classical enhancer activity.

4.2.2 *Differential gene expression predictor supports the observation of different enhancer classes*

To characterize the human genomic enhancers in detail, we combined the genomic STARR-seq experiments performed in the HepG2 liver cancer cell line with comprehensive characterization of the regulatory regions in HepG2 cells with approximately 100 TF ChIP-seq experiments, histone modification ChIP-seqs and ATAC-seq, partly downloaded from the ENCODE data portal [4]. This comprehensive integrative analysis, discussed in more detail in Publication II, revealed six classes of gene regulatory elements including three classes of active enhancers (see Figure 4.7a): i) closed chromatin enhancers (STARR-seq +, ATAC-seq -), ii) cryptic enhancers (silenced STARR-seq + regions), iii) promoters (ATAC-seq + with or without STARR-seq), iv) chromatin-dependent enhancers (STARR-seq -/low, ATAC-seq + with active histone mark H3K27ac), v) structural chromatin elements (STARR-seq -, ATAC-

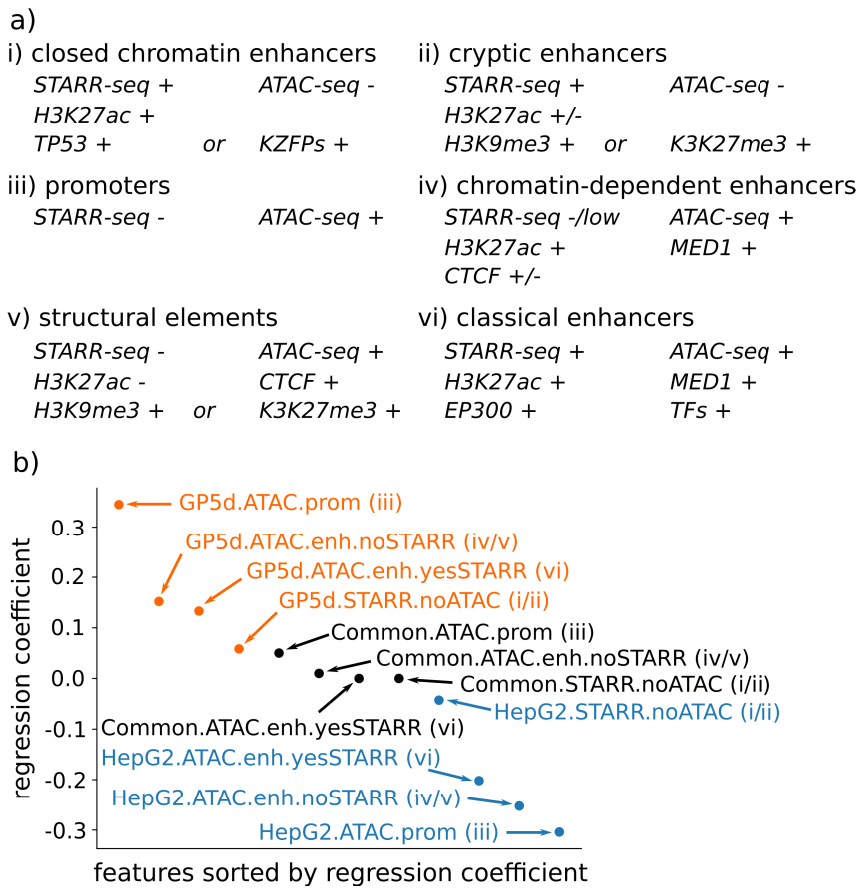

FIGURE 4.7: Integrative genomic analysis reveals three types of transcriptionally active enhancers. a) Six types of regulatory elements classified on the basis of STARR-seq signal and chromatin features such as accessibility (ATAC-seq), TF binding, and epigenetic modifications (KZFPs = KRAB-Zinc Finger Proteins). Derivation of this classification is discussed in detail in Publication II. b) Regression coefficients of a lasso regression model trained to predict differential gene expression between GP5d and HepG2 cell lines based on the features shown (see Methods for details). The model was able to explain approximately 12% of the observed variance in target gene differential expression. Positive values mean overexpression in GP5d and negative overexpression in HepG2. Features were constructed based on distance from TSSs (distal/proximal) and overlap with ATAC-seq and/or STARR-seq peaks in corresponding cell lines. Features are colored so that GP5d-specific features are orange, HepG2 specific features blue, and common features black. Numbers after the feature names indicate to which regulatory element classes from panel a the features correspond to. Figure modified and adapted from publication II.

seq +, CTCF +) and vi) classical enhancers (STARR-seq +, ATAC-seq +). All three types of active enhancers detected had independent predictive power in predicting differential gene expression between the GP5d and the HepG2 cell lines (Figure 4.7b). Cryptic enhancers appeared silenced by their co-occurrence with repressive chromatin marks H3K27me3 (see also poised enhancers [167]) and/or H3K9me3 and HP1 (see [168]). Detailed analysis of the genomic enhancers in the GP5d cells presented in Publication II supported this picture observed from HepG2 cells.

4.2.3 *Additive and non-specific local promoter-enhancer interactions*

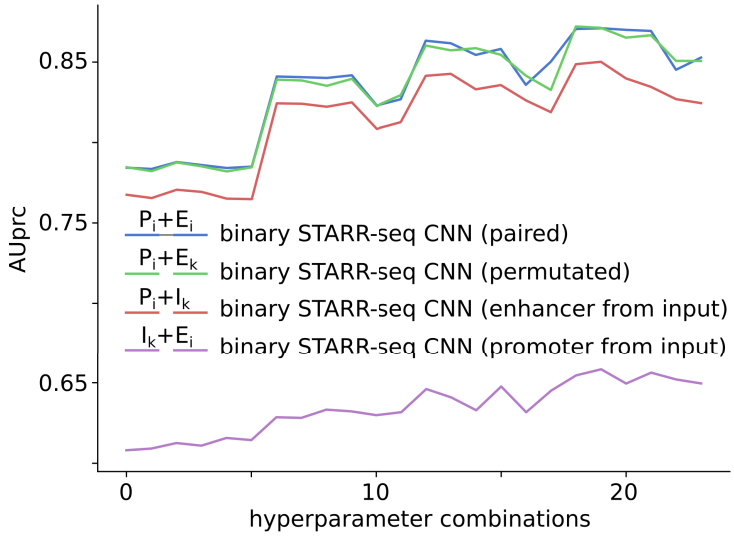

FIGURE 4.8: Machine learning identified no sequence features determining specificity of promoter-enhancer interactions. Four CNN classifiers with identical architecture (see Figure 3.6b) were trained on different data sets to classify between active and inactive promoter-enhancer pairs. In the "paired" training data the promoter-enhancer pairing was kept intact, whereas in the "permutated" data, the pairs were shuffled disrupting any specific interactions between the promoters and the enhancers. In the "enhancer from input" and "promoter from input" data, the promoters and enhancers, respectively, were paired with randomly sampled inactive sequences from the input library. Separate models were trained for 24 different hyperparameter combinations (x-axis). The area under precision-recall curve (AUprc) values show that the CNNs trained on paired data (blue) outperform CNNs trained on enhancer (violet) or promoter (red) data, but not those trained on permutated data (green, paired Student's t-test p-value $\approx 1.34 \times 10^{-1}$). Figure modified and adapted from publication II.

In addition to studying differences between sequence features enriched at promoters and enhancers (see Publication II), the binary STARR-seq experiment also allows studying interactions between promoters and relatively local enhancers separated by few hundreds of base pairs. To establish if there are any specific interactions between the promoters and the enhancers in the binary STARR-seq experiment, we trained a series of CNN-based classifiers with two input heads for separately learning the sequence features from the enhancer

and the promoter positions, which were then combined with a fully connected layer that can learn interactions between the two inputs (see Methods for details). We trained four different types of models with identical architectures but with differently arranged training data (Figure 4.8). The rationale for this *in silico* experiment was that if there are any specific interactions between the promoters and the enhancers, these interactions can be removed by disrupting the pairing of the active promoter and enhancer sequences - by either shuffling the pairs or by pairing the promoters or the enhancers with inactive sequences from the input library. Figure 4.8 shows that promoters have more predictive power than enhancers, when paired with inactive sequences from input. The predictive performance of the model further improves when pairing the active promoters with active enhancers, but with a different enhancer than in the experiment thus disrupting any specific interactions. Keeping the information about the original pairing intact does not improve on this (paired Student's t-test p-value $\approx 1.34 \times 10^{-1}$), indicating that the enhancers interact with the promoters in a non-specific manner. Analysis of motif match counts between the promoters and the enhancers presented in Publication II further supported this result.

4.2.4 *Prediction of genomic transcriptional activity using sequence features from machine learning models*

To study how well a promoter model trained on the promoters enriched from random sequences can predict gene expression in the human genome, we trained a CNN model to classify between the active TSS-aligned promoters from the GP5d promoter capture template switch experiment and corresponding inactive promoter sequences from the input library. Figure 4.9a shows that the CNN trained on the TSS-aligned sequences from the promoter capture STARR-seq is able to predict the active GP5d TSS positions in the genome (promoter activity measured using CAGE, see Methods) even more accurately than a similar CNN model trained on the genomic promoters themselves (promoters and TSS positions downloaded from the EPD database [137]). The CNN trained on the STARR-seq data outperforms the CNN trained on the EPD data even when not restricting to active TSSs in the GP5d cells as defined by CAGE, but considering all unseen test set TSSs from the EPD database (Figure 4.9b). Also a position-specific logistic regression model trained on the STARR-seq data outperforms the CNN trained on the EPD data suggesting that the training data plays a major role in observed better performance. Moreover, both the STARR-seq and the EPD-based CNN models were trained for 72 different hyperparameter combinations, and the model trained on the STARR-seq data outperformed the model trained on the EPD data in predicting the TSS position at active GP5d promoters in all but one of these combinations (Figure 4.10, paired Student's t-test p-value $\approx 9.68 \times 10^{-23}$ for rejecting a null hypothesis of similar performance across the hyperparameter combinations).

To investigate why the CNN trained on STARR-seq data is able to outperform the CNN trained on the genomic promoters, we used the mutual information-based approach described in Publication III to visualize the pairwise dependencies learned by these two models (see Figure 4.11). This analysis revealed that the CNNs trained on STARR-seq data had learned a stronger position-specific signal of TF enrichment than the CNNs trained on the genomic promoters, which relied more on information at a relatively short region around the TSS. This likely makes it easier for the STARR-seq-based CNN to more accurately predict the TSS position.

In Publication II, we use external validation data not seen by the model during training to test how well the CNN model trained on the promoters enriched from random

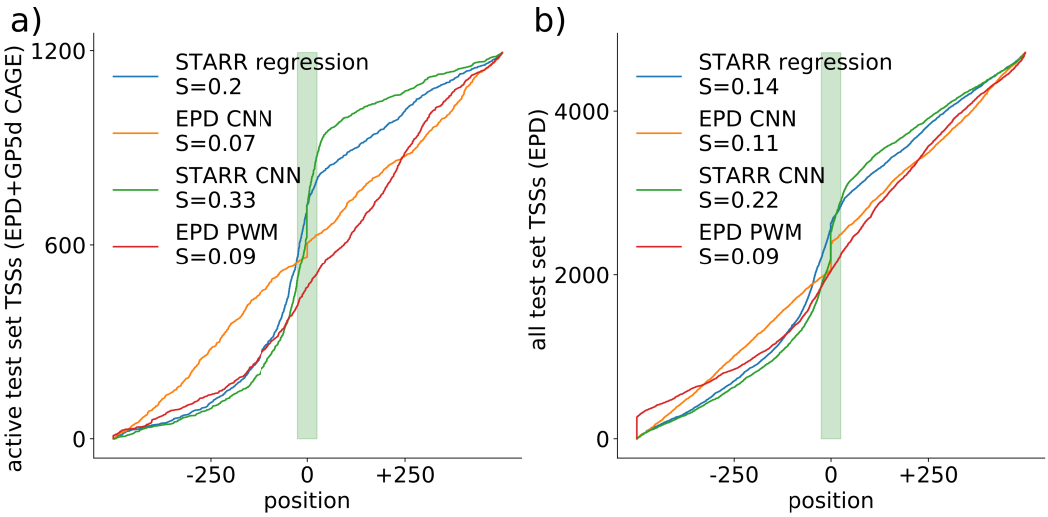

FIGURE 4.9: Prediction of genomic transcription start site (TSS) positions. Cumulative distance of the predicted TSS positions from annotated genomic TSSs is shown for a CNN trained on human genomic TSS data (orange) and for a PWM model of human genomic TSSs (red). Same is shown also for a regression model using positional enrichment of TFs as features (blue) and for a CNN (green) trained on *de novo* promoters enriched from random sequences and aligned based on the TSS positions measured using the template switch experiment described in Methods. The test set genomic TSS positions are aligned at 0; the curves mark predicted TSS positions for each model, sorted by distance from the annotated TSS position. The TSS probability of each position within $\pm$500 bp from the known TSS position was evaluated for each model, and the most likely TSS position, according to each model, was used as the prediction of the model. The score S in the figure legend indicates the fraction of predicted TSS positions within $\pm$25 bp (the area shaded with green) from the annotated TSS positions. a) Prediction of active GP5d TSSs defined as intersection of Eukaryotic promoter database (EPD) TSSs and active promoters in GP5d cells defined using CAGE. b) Prediction of all human TSSs from EPD. Figure modified and adapted from publication II.

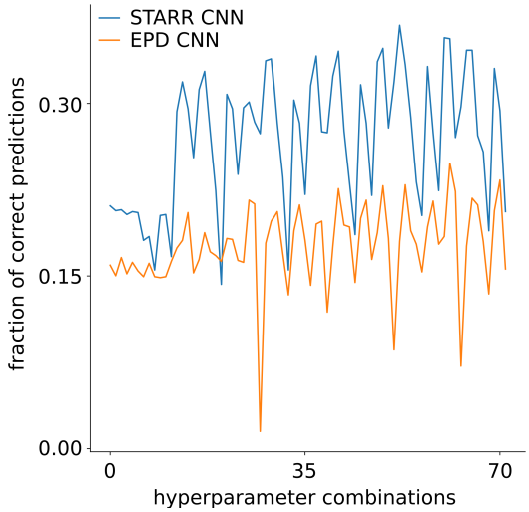

FIGURE 4.10: The CNN classifiers trained on the promoters evolved from random sequences consistently outperform the CNNs trained on genomic promoters in predicting the TSS position on unseen test data of active GP5d genomic TSSs across the tested hyperparameter space. Both models were trained on the same set of 72 different hyperparameter combinations. Fraction of predicted test set TSS positions that are within ± 25 bp from the annotated TSS position is shown on y-axis. The CNNs trained on STARR-seq data (blue) outperform the CNNs trained on the genomic promoters from the Eukaryotic Promoter Database (EPD, orange) data on all but one hyperparameter combination tested (paired Student's t-test, p-value$= 9.68 \times 10^{-23}$) Figure modified and adapted from publication II. .

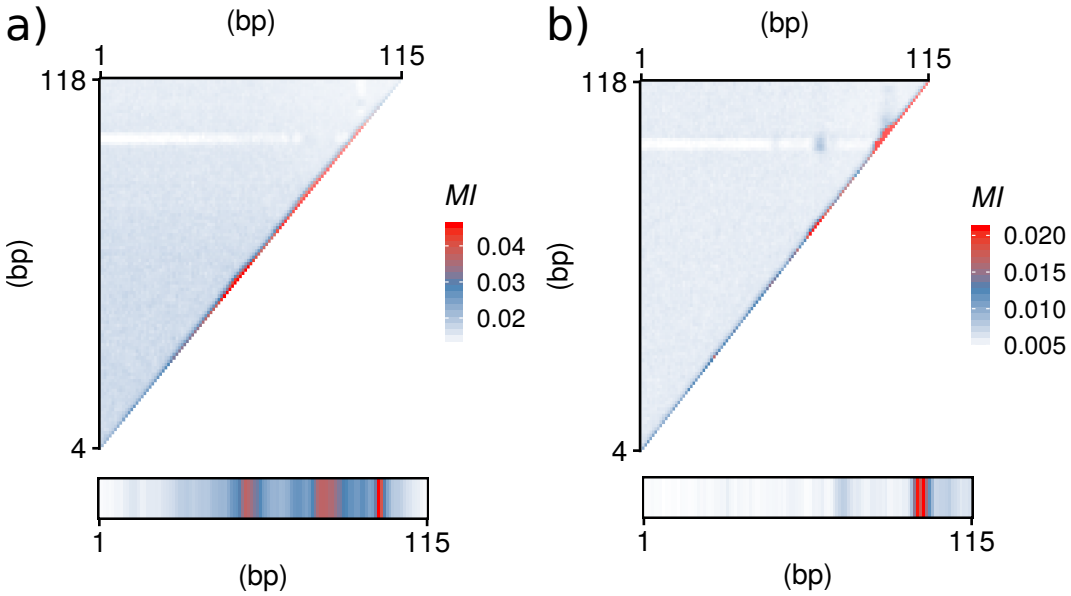

FIGURE 4.11: Mutual information (MI) based comparison of pairwise interactions learned by the CNN models trained on the STARR-seq active promoters (a) and the human genomic promoters from EPD (b). The triangle-shaped upper panels show the values of MI between 3-mer distributions at each position of the models. Below is shown a zoomed-in view of the diagonal of the MI matrix showing the positional enrichment of TF binding sites. The EPD-trained CNNs seem to rely more strongly on the presence of the Initiator-motif at the TSS (position 100, see also Publication II), while the STARR-seq-trained CNNs have learned a stronger pattern of positional enrichment of TF binding sites downstream of the TSS that is likely to make exact positioning of the TSS easier. For both models, random sequences with predicted promoter probability over 0.9 according to 10 best individual CNNs (with different hyperparameter combinations) were used for the MI analysis. Figure modified and adapted from publication II.

sequences has learned to predict effects of mutations on human promoters. The predicted mutation effects on the TERT promoter activity based on the CNN model correlate well with previously published saturation mutagenesis experiments [128] in HEK293T cells achieving Spearman correlation 0.74 (Figure 4.12a; Spearman correlation 0.60 in SF7996 primary glioblastoma cells). The cross-correlation of the HEK293T and SF7996 saturation mutagenesis experiments is only slightly better at 0.79 suggesting the CNN model has learned to predict effects of single nucleotide variants on promoter activity. Figure 4.12b shows the predicted promoter probabilities for all possible TERT promoter single nucleotide variants (SNVs). Most of the variants are predicted to have a mild effect, but there are some hotspots, especially at the TSS (position 100), where several possible mutations are predicted to either hyperactivate or to kill the TERT promoter. As an example, Figure 4.12c shows both the wild type TERT promoter, and the variant (p101:A>C) with highest predicted promoter probability, and the sequence patterns the CNN model recognized and used to make the prediction. This visualization, made using DeepLIFT [86], shows that the mutation would create a binding site for an ETS-class TF (TTCCGG), that the model predicts would lead to a highly active promoter.

Further analysis of the 14 recurring TERT promoter mutations observed in patient samples [169] showed that the CNN model trained on the STARR-seq TSS-aligned promoters correctly predicted the direction of the mutation effect in 13 out of 14 cases (see Publication II). Interestingly, the variant (p101:A>C) obtaining the highest predicted promoter probability for the CNN model was not within these recurring TERT variants. Figure 4.13 illustrates how the CNN model has learned to correctly predict the effects of three most commonly observed cancer-associated mutations at the TERT promoter [34, 169, 170] - the CNN model correctly predicts that the mutations increase promoter activity by predicting a higher promoter probability for the mutants versus the wild type promoter. DeepLIFT visualization of the features used by the CNN in making the predictions highlights that the CNN has learned to recognize the ETS-class TF binding sites known to be created by these mutations. Taken together these results highlight that combining modern machine learning methods with large-scale, unbiased data from human gene regulatory element activities allows creating even more accurate models of gene regulation than studying only those sequences that appear in the human genome.

4.3 Novel deep learning model interpretation methods for genomics

In Publication II, we used deep learning based classifiers to learn the grammar recognized by transcription factors in human promoters and enhancers. Interpreting the features driving gene expression according to the deep learning models was one of the key questions and led into development of two novel approaches to interpretation of deep learning models in genomics. These approaches are discussed in the following.

4.3.1 *Testing if a convolutional neural network model uses similar features than a logistic regression model with designed features*

In addition to applying previously described interpretation methods described in Methods such as DeepLIFT, TF-MoDisCo or using the deep learning models to make predictions on synthetic sequences with designed features, we developed two novel model interpretation methods. In Publication II we describe how a convolutional neural network (CNN) based classifier outperforms logistic regression classifiers that use known TF binding motifs as

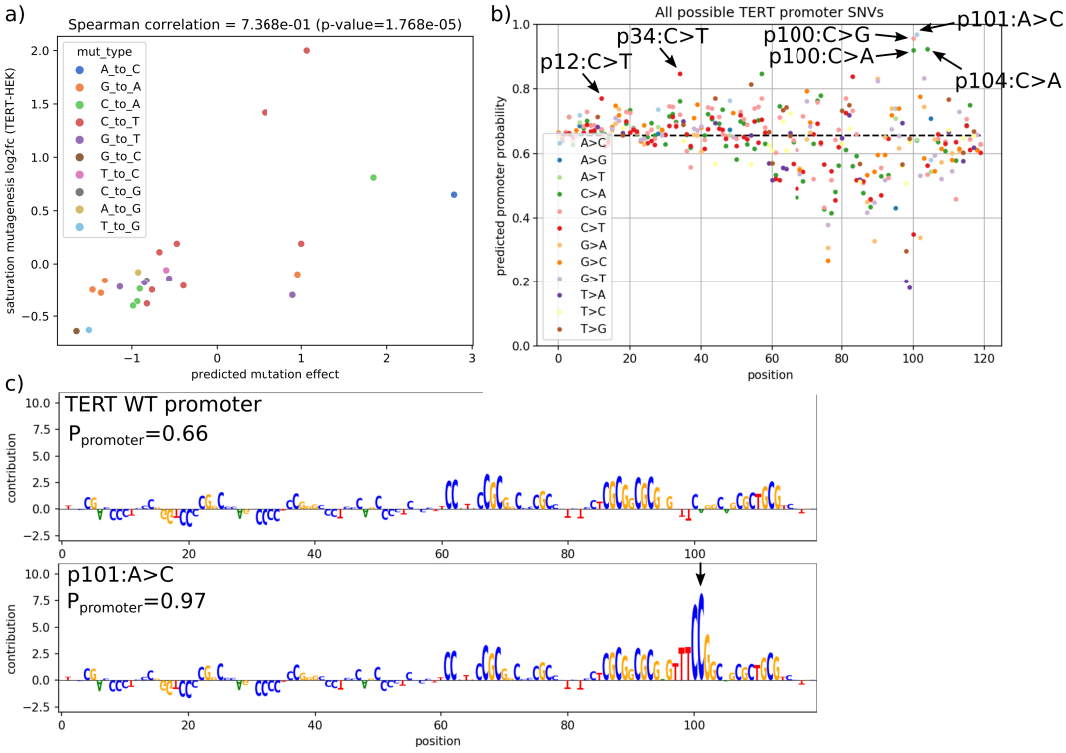

FIGURE 4.12: a) The effects of point mutations predicted by the CNN model trained on STARR-seq promoters selected from random input sequences (activity measured in GP5d colon cancer cells) correlate with the measured effect of the same mutations in a published saturation mutagenesis MPRA study of the human TERT promoter [128] in HEK293T cells. Correlation is shown between the predicted mutation effect (see Methods for derivation) and the measurement from the saturation mutagenesis measurements [128]. Only statistically significant mutations ($p<0.05$) in both predictions/measurements are shown. b) Predicted effect of each possible single nucleotide variant (SNV) at the TERT promoter from the CNN model trained on promoters enriched from random sequences. Colors indicate different mutations, x-axis shows the position along the promoter (TSS at position 100), and y-axis shows the predicted promoter probability. The dashed line shows the predicted promoter probability of the wild type TERT promoter. c) DeepLIFT [86] visualization of the positions that contribute most towards the CNN predictions for the wild type TERT promoter (top) and the most active variant (p101:A>C, bottom, arrow indicates the mutated position) according to the predicted promoter probability ($P_{promoter}$). Height of the letters indicates the importance of the nucleotide at that position towards the CNN prediction. Figure modified and adapted from publication II.

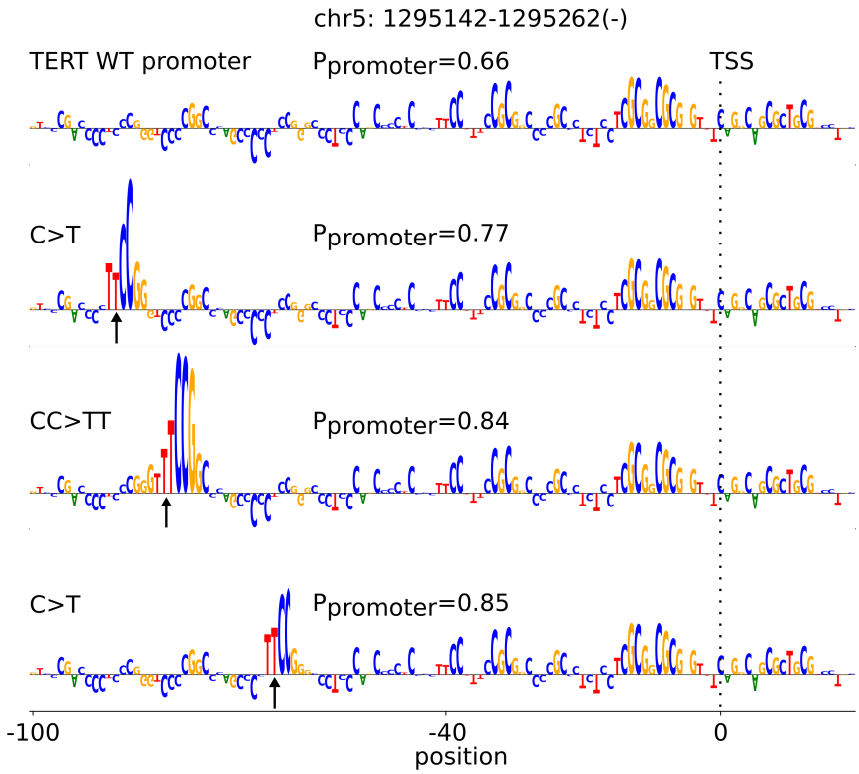

FIGURE 4.13: The CNN trained on the TSS-aligned promoters enriched from random sequences correctly identifies cancer-associated mutations in the TERT promoter. Top: predicted sequence determinants at the TERT promoter as determined by DeepLIFT [86] analysis of the CNN model. Bottom: the effect of three driver mutations [34, 169, 170] on the predicted activity of the promoter (mutated bases highlighted). $P_{promoter}$ is the predicted promoter probability and heights of the letters indicate the importance of the corresponding position towards the prediction verdict. Figure modified and adapted from publication II.

features in predicting which sequences are active enhancers and which not. To understand what different features the CNN model uses to achieve the better predictive performance, we developed the "Nsweep" algorithm (Algorithm 1), that can be used to determine if a deep learning model uses specific TF binding motifs in its predictions. Schematic presentation of the Nsweep algorithm is shown in Figure 4.14.

Shortly, the idea is to systematically score the consensus sequence of a PFM and all sequences one substitution (Hamming distance 1) away from it with the CNN model and record the effect of each substitution towards the prediction made by the CNN model. This was done by embedding the variants into random positions in different background sequences and recording the contribution of each mutated position, assessed using DeepLIFT ([86]), as described in Algorithm 1 and Figure 4.14. As discussed earlier, the main assumption in a PFM/PWM model is that the positions of the model are independent of each other and thus each position can be tested separately when assessing whether a CNN has learned a PWM-like feature. As an end result, the Nsweep algorithm will produce a sequence logo we call CNN Activity Contribution Weight Matrix (CACWM). If the CACWM looks similar to the input PFM, the CNN is using a feature similar to the original PFM.

Figure 4.15 shows an example from Publication II where a CNN classifier trained to separate active STARR-seq enhancers from inactive input library sequences was used to re-generate TF binding motifs used by a logistic regression classifier trained to perform the same task. The figure shows selected examples of cases where the CNN has learned parts of the HT-SELEX motifs used by the logistic regression classifier (FLI1, IRF3, TP63 & TCF7), where the CNN has not learned the HT-SELEX motif (HNF4A), or where the real feature in the data likely is something simple like a stretch of Cs that has predictive power but that is not present in the set of features available for the logistic regression model, so the logistic regression has picked up the TF binding motif that is closest to the stretch of Cs. The Nsweep analysis, together with other model interpretation analyses presented in Publication II indicated that the main difference between the features learned by the CNN and the features used by the logistic regression was that the CNN was able to learn TF binding motifs that are slightly different from the HT-SELEX motifs, and thus likely better optimized for the classification task.

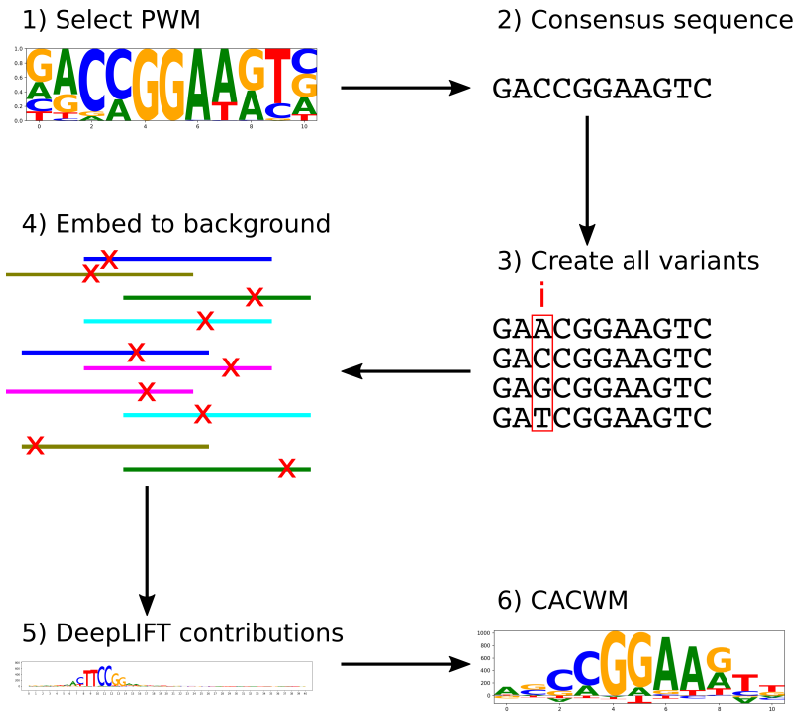

FIGURE 4.14: Schematic overview of the Nsweep algorithm for testing if a deep learning model has learned a given TF binding motif.

/*variables:;
PFM : Position Weight Matrix of TF of interest;
s_{PFM} : Consensus sequence of the PFM;
$CACWM$: CNN Activity Contribution Weight Matrix;
N_{pos} : Number of positions where each variant sequence is embedded;
N_{bg} : Number of different background sequences;
N_{ref} : Number of reference sequences for DeepLIFT;
*/;
foreach *column i in PFM* **do**
    Create all 4 possible single nucleotide variants of the consensus sequence;
    Embed each variant sequence separately to N_{pos} different positions in N_{bg} different background sequences;
    foreach *seq of the resulting* $4 \times N_{pos} \times N_{bg}$ *sequences* **do**
        Compute DeepLIFT contributions for each position of *seq* with the CNN model against N_{ref} dinucleotide-shuffled versions of *seq*;
        Add the resulting contribution at position *i* of the variant to position *i* of the $CACWM$ for the nucleotide *n* that is at position *i* in the variant embedded to *seq*;
    end
    Normalize the column *i* by dividing with the number of variant sequences created for each column and each variant ($N_{pos} \times N_{bg}$);
end

Algorithm 1: Pseudocode for the Nsweep algorithm for generating CNN Activity Contribution Weight Matrices (CACWMs).

4.3.2 *General machine learning model interpretation tool to highlight dependencies learned by the model*

One advantage of deep learning models is that they can in principle learn any kind of complex interactions. It is, however, not straightforward to determine if a deep learning model has learned interactions between features unless there are specific hypotheses one can test by embedding the interactions into synthetic input samples and scoring these samples containing the interactions with a pre-trained deep learning model. In Publication III, we developed a general machine learning model interpretation tool, PlotMI, that can intuitively visualize pairwise dependencies learned by a machine learning model trained on sequence data.

We use a pre-trained machine learning model to score a set of input samples, and select a subset of the input based on the machine learning model predictions for downstream analysis. We then compute mutual information (MI) ([130]) between position-specific k-mer distributions in this subset and visualize MI as a two-dimensional heat map as described in Figure 4.16a. PlotMI can highlight spacings between interacting features as well as positions of interacting features learned by a machine learning model. Figure 4.16b shows an example of a mutual information plot (MI-plot) visualizing dependencies between pairwise positional 3-mer distributions in a simulated dataset with two mutually exclusive embedded interactions. MI-plot offers an easily interpretable visualization of the dependencies in the data, even in presence of multiple interactions, unlike distance

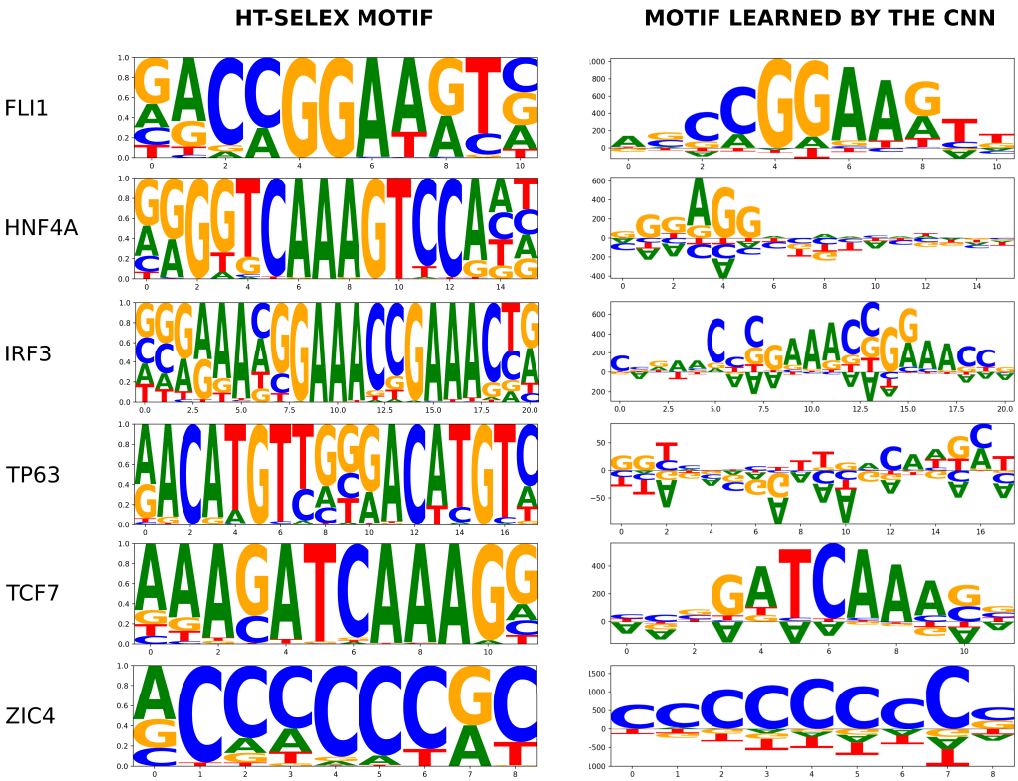

FIGURE 4.15: Example CNN activity contribution weight matrices (CACWMs) learned by the CNN model from the random enhancer STARR-seq data analyzed using the Nsweep approach. HT-SELEX PFMs are shown on the left and the corresponding Nsweep motif on the right. The CACWMs reproduce the parts of the HT-SELEX motifs that the CNN has learned to be important in predicting active enhancers. Motifs not reproduced by the Nsweep analysis have not been learned by the CNN. Figure modified and adapted from Publication II.

measures like Jensen-Shannon divergence that measure similarity between probability distributions (Figure 4.16c, see also Publication III).

As an example of using the MI-plot to visualize dependencies learned by a deep learning model, we applied PlotMI to interpret a CNN model trained to recognize human genomic promoters. The promoter sequences were downloaded from the Eukaryotic Promoter Database (EPD, [171]) and the CNN was trained to separate the real promoters from dinucleotide-shuffled versions of the promoter sequences. Figures 4.17a-b show that the CNN model has learned an interaction between the canonical TATA-box position approximately 30 bp before the transcription start site (TSS), and the TSS itself, highlighting the ability of PlotMI analysis to visualize dependencies learned by a deep learning model from real biological data.

A strength of PlotMI as a model interpretation tool is that it is not dependent on the machine learning model architecture as it only operates on the input sequences selected based on the model predictions. As an example of a non-deep learning model, Figure 4.17c shows PlotMI visualization of pairwise dependencies learned by N-score ([125]), a wavelet-based logistic regression classifier trained to predict which sequences bind to nucleosomes in yeast genome. The analysis of MI within high-scoring random sequences scored by N-score shows that the model has learned a dependency pattern, where each position interacts with positions equidistant from the middle position of the 131 bp long model. Separately plotting the maximum MI of each diagonal in the MI-plot (from main diagonal to bottom left corner) further illustrates the periodic interaction learned (Figure 4.17d).

In Publication III we also used PlotMI to visualize different protein fitness models trained to predict the insulin binding affinity of GB1 protein. The models were based on linear regression (LR), sequence convolutional neural network (CNN) and graph convolutional neural network (GCN) architectures, and were published in [126]. Only single and double mutants of the wild type GB1 sequence were seen by the models during training. This means that only a small fraction of the possible sequence space, and only very near the wild type sequence, was sampled. Thus the ability of the models to generalize to unseen mutations will be heavily tested when using them to score random amino acid sequences. Additionally, a three dimensional structure of GB1 protein domain was used in training of the GCN model, but not in training of the other models. Contact map showing the physical distances between each amino acid residue pair computed from this structure is shown in Figure 4.18a.

We scored 10 million random synthetic amino acid sequences with each of the three models to see what pairwise dependencies the models had learned. We visualized top 100,000 (1%) of the random sequences according to the predicted fitness with PlotMI (Figures 4.18b-d). MI was computed between position-specific 1-mer distributions. The GCN model, being the only model where structural information was utilized during training, had learned more of the important pairwise dependencies visible from the contact map than the other models. In [126] the authors report no significant difference in performance of the GCN and CNN models in predicting unseen GB1 variant (single or double mutant) fitnesses tested in deep mutational scanning experiments, even though the MI-plots show that the CNN has learned only a small fraction of the pairwise dependencies from the contact map. It seems that the structural information used in training of the GCN helps in assigning realistic fitness values for proteins that are further away from the wild type sequence than any training samples from the mutational scanning data. In contrast to the GCN and CNN model architectures, the LR model can only learn additive effects

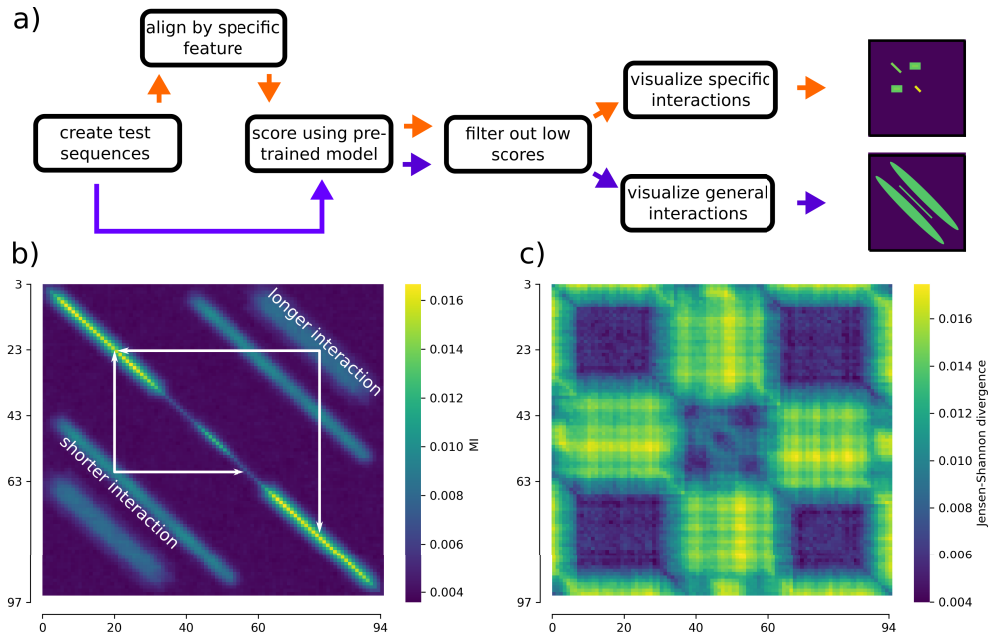

FIGURE 4.16: Mutual information (MI) based visualization reveals pairwise and positional dependencies in a set of sequences. a) Two possible visualization workflows using PlotMI. MI reveals the distance between interacting features if the visualized model of interest (MOI) is trained on unaligned data and the PlotMI input sequences are also unaligned (blue arrows). If either the MOI is trained on aligned data, or the PlotMI input sequences have been aligned, MI can reveal the exact positions of the interacting features (orange arrows). The axes indicate positions along the input sequences. b) Example MI visualization of a dataset where two different pairs of transcription factor binding motifs have been embedded with different spacings (30bp and 50bp) on DNA sequences drawn from uniform nucleotide distribution creating two mutually independent dependencies. MI was computed between position-specific 3-mer distributions. Signal on the main diagonal of the MI-plot highlights parts of the sequences where adjacent positions have dependencies with each other. Signal off the main diagonal corresponds to longer-range dependencies. Exact positions of the interacting pairs can be found by drawing lines parallel to the x- and y-axes towards the main diagonal, as illustrated with the white arrows. c) Importantly, measures of similarity of position-specific 3-mer distributions cannot be used to highlight multiple independent interactions from a dataset, as illustrated here using Jensen-Shannon divergence for the same data set visualized using MI in panel b. Figure modified and adapted from Publication III.

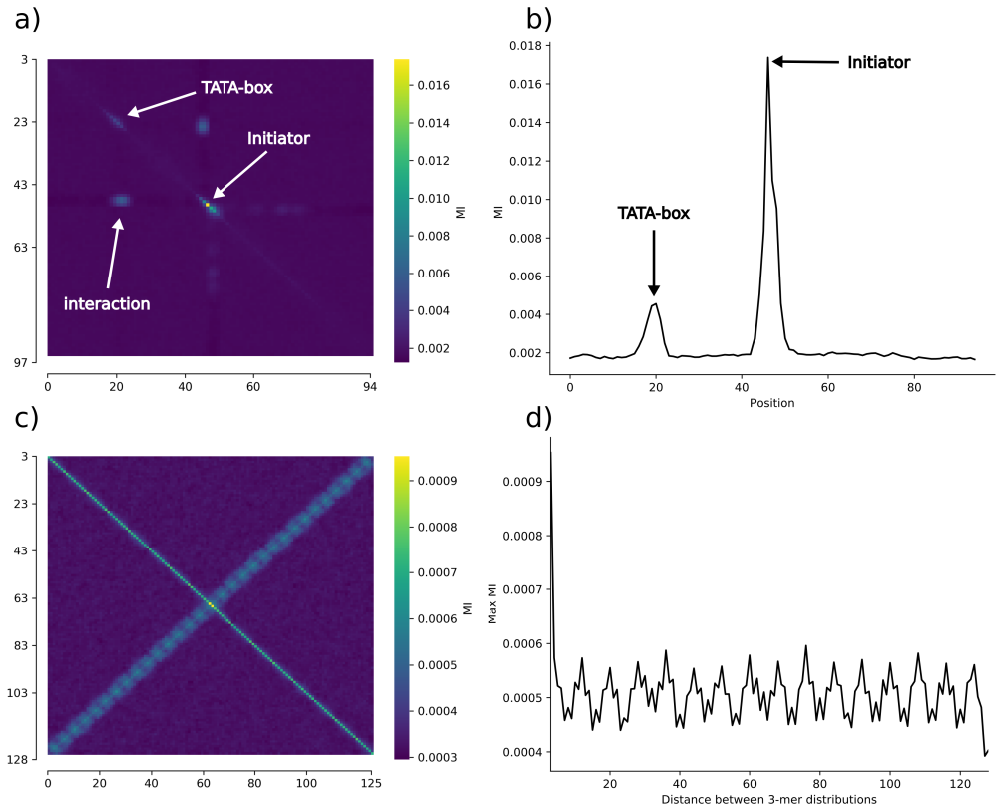

FIGURE 4.17: PlotMI-visualizations of machine learning models trained on genomic DNA sequences. a) The MI-plot shows that a CNN classifier trained on 100bp long sequences centered at human genomic transcription start site (TSS) positions has learned an interaction between the TSS and TATA-box region around 30 bp upstream from TSS. b) Mutual information (MI) of the main diagonal of the MI-plot from panel a. c) PlotMI visualization of N-score model trained to distinguish 131 bp long nucleosome binding DNA from non-nucleosomal DNA [125] in yeast genome shows that N-score has learned a periodic interaction pattern where 3-mer distributions at positions separated from each other by multiplicatives of a fixed period are dependent on each other. The learned interaction is symmetric relative to the middle position of the 131 bp sequences such that the strongest pairwise dependencies are observed between positions same distance away from the middle position, but to opposite directions. d) Maximum MI of each diagonal of the MI-plot shown in panel c. Figure modified and adapted from Publication III.

between positions [126]. Still the dependency pattern learned by the LR model is largely similar to the CNN. Based on the PlotMI-analysis, it seems that the limited sequence space available for these protein models during training prohibits the CNN and LR models from learning a realistic representation for the functional GB1 protein, even though the models can still make good predictions for mutational effects near the wild type sequence.

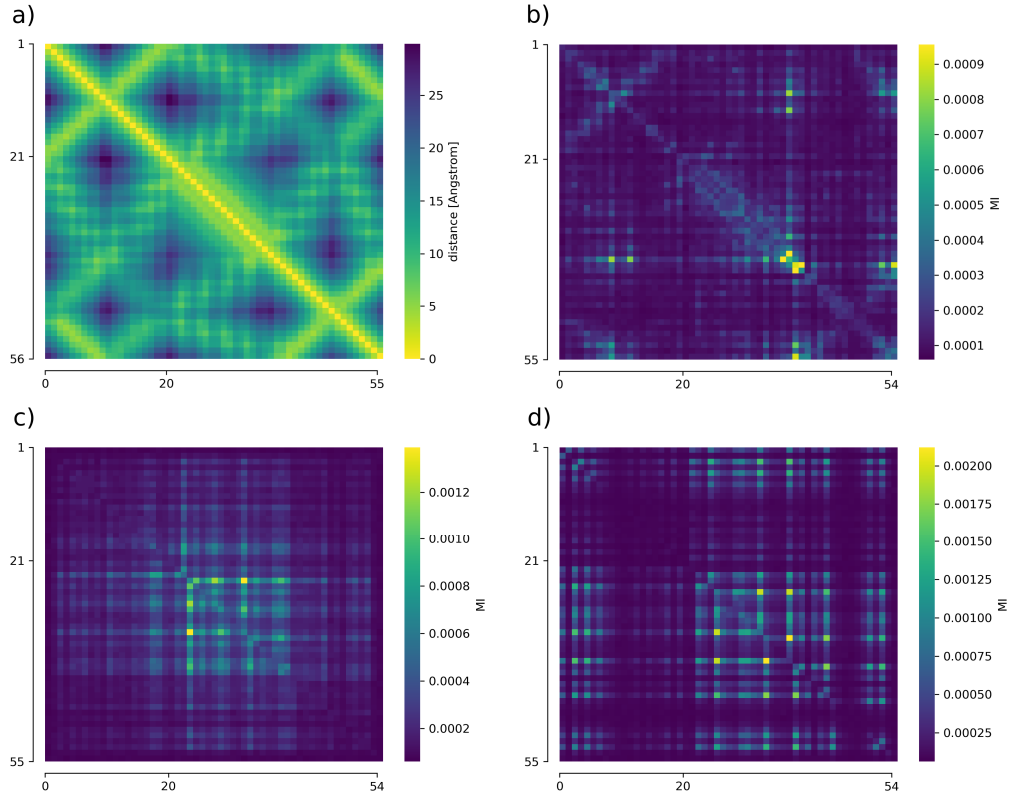

FIGURE 4.18: PlotMI-visualizations of models predicting protein fitness reveal different types of dependencies learned by models with different architectures. a) Contact map of GB1 protein domain derived from 2QMT PDB structure that is used in training of the GCN model but not in training of the CNN and LR models. Color indicates distance between each pair of α-carbons. Note that darker colors mean longer distance so that the heat map is easier to compare against the MI-plots, where darker colors mean lower pairwise MI. b-d) MI-plots created using position-specific 1-mer distributions showing the pairwise dependencies learned by the GCN model (b), the CNN model (c) and the linear regression model (d). Figure modified and adapted from Publication III.

5

CONCLUDING DISCUSSION

The rapid advancement of high-throughput sequencing based experiments has lead to explosion of available experimental data for scientists studying gene regulation. This has caused a situation where the major bottleneck in research is not anymore so much in experimentation and data generation, but in analysis and interpretation of the vast amounts of experimental data generated. In the work described in this thesis, we both developed new tools and applied existing state-of-the art methods to analyze and interpret data from modern high-throughput regulatory genomics experiments such as ChIP-exo/nexus and STARR-seq. Specifically, in Publication I we developed a software called PeakXus for accurate and unbiased determination of TF-DNA binding sites genome-wide from novel ChIP-exo and ChIP-nexus experiments. In Publication II, we performed highly complex massively parallel reporter gene assay (MPRA) experiments and integrated this data with both new and publicly available large genome-wide measurements of TF-DNA binding, chromatin accessibility and histone modification status. Modeling these experiments with state-of-the-art convolutional neural network (CNN) models allowed us to comprehensively characterize the sequence determinants of human enhancers and promoters and revealed unexpected lack of specific interactions between promoters and local classical enhancers. In Publication III, we presented a novel application of computing mutual information (MI) between position-specific k-mer distributions to interpret dependencies learned by any machine learning model, but specifically deep learning models, trained on sequence data.

To allow the scientific community to better build on the results presented in this thesis, most of the generated new data and software tools have been released for public use. The algorithms developed in Publications I and III are available in GitHub as described in the publications, and the essential custom code and the pre-trained machine learning models along with the respective training, validation and test data sets from Publications II and III are archived to Zenodo. The raw experimental data generated in Publication II is archived to GEO.

The main design principle of the PeakXus peak calling algorithm presented in Publication I was to make as few assumptions about the shape of the signal from the ChIP-exo/nexus experiments as possible. The only major assumption we made stems from the fact that the λ-exonuclease treatment causes 5′ ends of reads overlapping a binding site to pile at same bases adjacent to the bound TF. We showed that PeakXus achieves similar accuracy in localizing the TF-DNA binding sites than previous methods, and reports more binding events that overlap with the TF-specific binding sequence. Importantly, PeakXus achieves this while not fitting the shape of the expected binding signal to the locations with highest overall signal genome-wide, a feature that likely makes PeakXus more flexible in detecting different types of signals that could originate for example from multiple TFs binding together or in very close proximity from each other.

In Publication I, we also demonstrated how the improved resolution of ChIP-exo and ChIP-nexus compared to ChIP-seq can be leveraged to more accurately study allele specific binding (ASB) of TFs. We presented a novel algorithm for studying ASB which improved and simplified duplicate read filtering by the use of unique molecular identifiers (UMIs, [96]. We also applied, to our knowledge first time, the Audic-Claverie test [163] to measure ASB

in such a way that the uncertainty in the fraction of reads mapping to the reference allele in whole genome sequencing (WGS) is taken into account.

Despite major efforts in measuring TF-DNA binding specificities *in vitro* [3, 6–8, 172] and in determining the TF-DNA binding positions *in vivo* [4, 9, 10], most sequence features that drive the activity of human promoters and enhancers remain poorly understood. TFs regulate gene expression by binding DNA at promoters and enhancers [9, 117, 173], and many features correlate with promoters and enhancers genome-wide. Both promoters and enhancers correlate with RNA transcription [77, 174], open chromatin [99, 175, 176] and histone H3 lysine 27 acetylation (H3K27ac) [104, 177], whereas promoters are marked by trimethylation and enhancers by monomethylation of histone H3 lysine 4 (H3K4) [178]. All these features can be used to predict promoter and enhancer positions and activities, but they do not establish what are the atomic DNA sequence elements required for promoter or enhancer activity.

The picture of transcriptional regulation in humans revealed by interrogation of the gene regulatory machinery of the cell by massively parallel reporter gene assays in Publication II was surprisingly simple. We trained CNN models capable of learning interactions between virtually any type of sequence features between promoters and classical enhancers enriched from random sequence, but no specific interactions were found. Counting pairwise enrichment of TFs between promoters and enhancers supported this result - enhancer motifs and promoter motifs enriched independent of each other. Non-specificity of most interactions between promoters and enhancers was also observed in a very recent independent report [179].

We probed the sequence features required for human enhancer activity by studying features enriched from completely random sequences that had enhancer activity. This revealed only a few TF binding motifs are highly active per cell type, and that motif grammar in these enhancers is weak on the level of spacing and orientation preferences of specific TFs, although some active heterodimer motifs were found. Machine learning models trained on the enhancers enriched from random sequences revealed that only a handful of TF binding motifs are needed for optimal classification between active and inactive enhancer sequences. Furthermore, no beyond additive interactions between TF binding motifs were found within the enhancers enriched from random sequences. These results are consistent with a recent report showing that independent actions of TFs can explain over 92% of the transcriptional activity measured from random yeast promoters [21].

Large-scale integrative analysis of enhancer activities measured with genomic STARR-seq and measurements of TF binding, histone modification status and chromatin accessibility genome-wide revealed two other types of active enhancers in addition to classical enhancers. We call these chromatin-dependent enhancers and closed chromatin enhancers. Here, classical enhancers are defined according to their original functional definition as sequences that can trans-activate a promoter regardless of position and orientation [117]. We found that chromatin-dependent enhancers are characterized by motifs of forkhead family TFs, binding of Mediator and p300 proteins, and strong signal for H3K27 acetylation. A lasso regression predictor using features defined based on the different regulatory element classes revealed that the presence of chromatin-dependent enhancers is strongly predictive of tissue-specific gene expression. The third type of active enhancers, closed chromatin enhancers, are located in regions with only moderate or no signal for ATAC-seq and are not silenced by CpG methylation. The closed chromatin enhancers appear to consist of only a single TF, such as p53, or a set of closely bound TFs that fit between or associate directly with well-ordered nucleosomes [57]. Importantly, all three active enhancer classes

had independent predictive power according to the gene expression predictor suggesting that they all play a role in regulation of transcription.

The finding in Publication II that most of the TF motif activities are similar between the tested cell types in the enhancers enriched from random sequences contrasts with the known tissue-specificity of some human enhancers *in vivo* [180, 181]. Also, based on analysis described in Publication II, the level of conservation of many human genomic enhancers appears to be higher than the information content of active elements selected in the random enhancer STARR-seq assay. The simplest explanation for these observations would be, that *in vivo* it is more difficult for enhancers to evolve to become specific than active. And as the STARR-seq assay is designed on purpose to reveal the sequence features determining transcriptional activity, the possible logic needed for tissue specificity will not be found from these enhancers selected from random sequences. Specificity will naturally require specific TF combinations, and also fine-tuning using motif number, spacing, orientation and affinity (see, for example [27, 182, 183].

In Publication II, analysis of features present in promoters enriched from random sequences led to discovery of a novel G-rich element downstream of the TSS. This element interacts with the TSS, potentially positioning RNA polymerase II independently of the TATA box. Also a very strict positioning preference of the YY motif right after the TSS was observed. A CNN model trained on the promoters enriched from random sequences predicted the active TSS positions in the genome better than a similar model trained on the human genomic promoters. This model was also able to predict the effects and mechanisms of action of known cancer-associated mutations in the human TERT promoter. These results demonstrate the usefulness of the approach where transcriptionally active sequences are selected from a random pool of input in learning models of gene regulation. Moreover, the ability of random STARR-seq experiments to interrogate a larger sequence space than what is available in the human genome is advantageous in training machine learning models that are able to better generalize the rules of transcriptional regulation. After all, the human reference genome, the only input sequence used in training of most of the machine learning models of gene regulation, is only one sample from the ensemble of possibly equally well functioning human genomes.

Overall the study presented in Publication II showed that the transcriptional activities of TF binding motifs can be classified into three groups that are not mutually exclusive: TSS-position determining activity (e.g. TATA-box, YY), promoter specific activity (e.g. NRF1) and enhancing activity localized both at the promoter and at the enhancer position (most of the TFs). Notably, no enhancer-specific TF binding motifs were found.

Taken together, the lack of enhancer-specific motifs and the lack of specific interactions between the promoters and the enhancers points towards a rather simple mechanism of action, where the activities of individual TFs bound to an enhancer are integrated irrespective of the specific TFs in question, and their total activity then activates the promoter. These results are consistent with the least specific type of molecular interaction, steric hindrance. The simplest mechanism of enhancer action would thus rely upon the size difference between TFs ($\sim$ 50 kDA) and the proteins of the transcriptional machinery such as Mediator and RNA-polymerases ($\sim 1-3$ MDa), along the lines suggested in [184]. Briefly, in this model the binding of small TFs to DNA in condensed chromatin could "lock" the regulatory regions near the surface of condensed chromatin domains as these regions of DNA with multiple bound TFs could not anymore penetrate the condensed chromatin due to steric hindrance, unlike smaller individual TFs. This would then allow the larger protein complexes of the transcriptional machinery to access the regulatory regions locked at the

surface of the condensed chromatin domains. The simple model describe above cannot, however, be the complete picture of enhancer action, as it has been previously shown that in the highly evolved genomic context, more specific interactions do exist between enhancers and particular promoters, as reported in cases such as multi-chromosome structures that control the expression of the repertoire of olfactory receptor genes [185] or the complex regulatory landscape of the HOX genes [186]. The results presented in this work suggest that specific TF-TF interactions allowing enhancers to act selectively at a long range would be associated with the chromatin-dependent and not the classical enhancers. However, experiments with longer distance between the tested promoter and enhancer pairs are needed for better understanding the effect of distance separating the elements. It is worth noting, that the observations presented in Publication II are not in conflict with the recently proposed formation of super enhancers via liquid-liquid phase separation [30] mediated by non-specific interactions between low complexity domains of TFs [31, 187].

The enrichment of active promoter and enhancer sequences from random input library, the approach used to study the sequence determinants of gene regulatory elements in Publication II, is a *bottom-up* style approach studying how easy it is to select functional gene regulatory elements from completely random pool of input sequences. A complementary *top-down* study would systematically mutate complex genomic enhancers, for example using CRISPR genome editing, that are too complex to be enriched from totally random input, and study the effect of these mutations on gene expression. Repetition of the STARR-seq experiments described here in additional cell types would further validate how much of the regulatory activities of TFs are shared between cell types. With modern deep learning methods, data from the top-down and bottom-up experiments in multiple cell types could be integrated together to train more accurate and generalizeable models of gene expression, for example using transfer learning approaches that have been recently shown to boost prediction of TF binding affinities for TFs with little experimental data available [188].

Training of accurate enhancer models able to predict the effects of mutations on gene expression from enhancers enriched from random input is more difficult than training such promoter models, partly because enhancers lack general "anchor" features such as the TSS or TATA-box, that help learning position-specific effects at promoters. Accurate and unbiased modeling of the effect of mutations at enhancers is of great importance, as only a small fraction of non-coding disease associated mutations occur at very close proximity of the TSS and can be covered with promoter models such as the one presented in Publication II. In the near future, comprehensive models of gene regulation, able to predict the effects of non-coding mutations accurately, can probably be achieved by carefully training machine learning models able to handle very long-range interactions (such as the Enformer [81]) on data from different types of experiments combining the top-down and bottom-up approaches as well as activities measured from the genomic regulatory elements and activities of designed or random sequences.

The PlotMI tool presented in Publication III tackles one of the main problems in applying state-of-the-art deep learning methods to study sequence-based problems - model interpretability. Previously, several methods for interpreting the individual features learned by sequence-based deep learning methods have been developed (e.g. [42, 43, 86]), but to our knowledge a general and easy-to-use tool for interpreting pairwise dependencies and positional preferences learned by a deep learning model has been missing. We show that using a pre-trained sequence-based machine learning model to filter random input based on model predictions, and then computing pairwise mutual information (MI) between positional k-mer distributions from the filtered sequences produces a visualization that

reveals pairwise dependencies learned by the machine learning model. The resulting visualization shows qualitatively the pairwise interactions learned by the model, but could be in the future extended to provide quantitative information of the effect sizes of the observed interaction patterns by integrating with the recently introduced global importance analysis [189].

Model interpretation using PlotMI is independent of the machine learning model architecture, and was demonstrated to work for binary classification and regression tasks as well as with DNA, RNA and amino acid sequence based models. The implementation of PlotMI is agnostic of the sequence alphabet used in model training and thus the tool can be applied also to interpreting models trained with very different types of sequence data, including, but not limited to, temporal sequences. The caveat is that in order to have an interpretation for the distance between interacting features, an interpretation must exist for distances within the samples used in model training.

The general idea behind PlotMI, feeding random input to a pre-trained model and filtering using model predictions, can also be applied to other types of visualizations in the future. Essentially, this idea is very similar to some high-throughput biological experiments such as HT-SELEX [47], where an initial library of random sequences is exposed to a target ligand such as a specific TF, and the sequences bound by the TF are selected and analyzed for common patterns. With this approach, we basically perform *in silico* experiments on the machine learning model. The model can be fed completely random input which is then filtered based on model predictions as in majority of the analyses in Publication II, but the input can also be designed to test a specific hypothesis. For example in Publication II, we took a *data driven* approach to studying regulation of gene expression by studying the transcriptional activities of completely random synthetic DNA sequences. We then trained deep learning models on this data and posed the questions about specific hypotheses, such as presence of specific interactions between promoters and enhancers to these models. Deep learning models can also be used to generate specific hypotheses for testing in validation experiments, for example deep neural network hallucination has been shown to generate *de novo* proteins "designed" by the neural network that fold into monomeric stable structures *in vivo* [190]. Following a similar strategy, one could validate the activities of deep learning model generated "optimal" human promoters and enhancers not present in the human genome by editing them into genomes of cell lines or model organisms.

This thesis describes computational tools that can help the genomics research community to understand the effects of variation in the regulatory genome to gene expression in disease. We described how high-resolution ChIP-exo/nexus experiments can be used to study allele specificity of TF-DNA binding, and trained machine learning models that can predict and explain the effects of promoter variants to gene expression. We also developed new ways to interpret deep learning models that hopefully help researchers in the future better understand how interactions learned by deep learning models affect their predictions. By interrogating the gene regulatory activities of a vast collection of DNA sequences, and integrating this data with genomic measurements of chromatin accessibility, TF-DNA binding, histone modification status etc. using machine learning and other approaches, we obtained novel insights into regulation of transcription in humans. We for example learned that TFs can be divided into mutually non-exclusive classes of TSS-positioning, promoter-specific, and enhancing activities. We also learned that local interactions between promoters and enhancers seem to be additive and non-specific. Importantly, we show that measuring the transcriptional activities of random synthetic DNA sequences in human

cells and modeling these data using modern deep learning methods is a powerful approach that can outperform models trained on genomic data and measurements only.

ACKNOWLEDGMENTS

The work presented in this thesis was carried out in the laboratory of professor Jussi Taipale at the Medical Faculty of University of Helsinki and at the Department of Biochemistry in University of Cambridge during 2014-2021. I would like to thank my supervisors professor Jussi Taipale and docent Teemu Kivioja for giving me the opportunity to work with several exciting projects and with high-quality data. This work would not have been possible without their continuous support and ideas. Special thanks to Teemu for his continous help with the day-to-day scientific work and for patiently answering my questions, even on the 15th time I asked him to explain the template switch experiment for capturing transcription start site position in the STARR-seq random promoters.

I would like to thank the members of my thesis committee, professor Harri Lähdesmäki and docent Rainer Lehtonen. I would also like to thank the reviewers of this thesis professor Veli Mäkinen and Dr. Markus Heinonen for taking time to give their expert opinions on my thesis, as well as Dr. Julia Zeitlinger for agreeing to be the official opponent in the thesis defense.

I have had a pleasure of working with talented colleagues in the laboratory of professor Jussi Taipale both in Helsinki and in Cambridge, as well as virtually with the scientists working at the Taipale lab in Karolinska Institute. I would especially like to thank the co-authors in the research included in this thesis, Drs. Biswajyoti Sahu, Päivi Pihlajamaa, Kashyap Dave, Bei Wei, Fangjie Zhu and Eevi Kaasinen. Also all the other current and former members of the Taipale lab in Helsinki have all contributed to my PhD journey. Especially I would like to thank Drs. Anna Vähärautio, Norman Zielke and Daniela Ungureanu, the "Co-PIs" of the Helsinki lab during my time there, for always providing support and keeping the lab running. Many thanks to all the current and former colleagues in the Helsinki lab: Drs. Kimmo Palin, Mikko Turunen, Maria Sokolova, Ping Chen and Pratyush Kumar Das, and Matias Kinnunen, Tuomas Lohi, Tomi Leung, Anu Luoto, Mika Pruikkonen, Kaisu Jussila and Katariina Sarin.

During my PhD studies I had the great opportunity to work at the Taipale lab in University of Cambridge, UK. I would like to thank the colleagues from that time in the Cambridge lab: Drs. Minna Taipale, Yin Lin, Fangjie Zhu, Otto Kauko and Yimeng Yin for your company and scientific discussions and advice.

I would like to extend my thanks to all former and current colleagues in the Taipale lab at Karolinska Institute and current members of the Taipale lab in Cambridge, especially Margareta Kling-Pilström and Emma Inns for all the help with practical matters.

I would also like to thank all collaborators in the research published during my PhD or yet unpublished. Especially I would like to thank everyone who has participated in the frequent "SELEX-meetings" over the years for stimulating scientific discussions and for help in my projects: professor Esko Ukkonen, and Drs. Jarkko Toivonen, Leena Salmela and John Davies, among others already mentioned above. Many thanks to all past and present members of the Academy of Finland Center of Excellences our lab has been a part of during my PhD journey.

I would like to extend heartfelt collective thanks to all the fantastic people I have had the privilege to meet via the activities of the ILS doctoral programme. Your company and peer support has been crucial during this journey. Especially I would like to thank the people

from the ILS student council: Alok, Behnam, Darshan, Elli, Elina, Geri, Heidi, Illida, Isabel, Jarno, Johanna, Jurgita, Kornelia, Kul, Maarja, Markku, Mridul, Sawan, Siggi and others. I would also like to thank the ILS doctoral programme for funding the early years of my PhD work, and the past and present coordinators of the programme, especially Erkki, for their help with my studies.

I would also like to acknowledge the funding from the Academy of Finland Center of Excellence program, and from Emil Aaltonen foundation for my research visit to Cambridge.

Some of the people already mentioned were also part of the TEDxHelsinkiUniversity organizing team. Putting those events together was one of the best experiences during my PhD years. I would like to thank the whole team, and especially the main organizers Chiara and Shishir!

I wish to thank my friends for their company and for giving me other things to do and think about than work, especially Andrew, Ida, Joonas, Julia, Jussi, Lauri, Simo, Suvi and Vappu. Special thanks also to Elina, Juho, Lauri and Liia for lunch break company in Meilahti during the years.

And lastly, the most special thanks to Krista, and my family, mom, dad, Tuukka and Tuuli.

A

APPENDIX

A.1 DERIVATION OF OCCUPANCY PROBABILITIES OF DNA SEQUENCES BY INDIVIDUAL TFS, OR PAIRS OF TFS

This section describes in detail the derivation and assumptions in modeling the occupancy probabilities of enhancer sequences by TFs and pairs of TFs in the logistic regression analysis conducted in Publication II.

A.1.1 *Converting PWM scores into free energies of binding*

For each variable (each variable corresponds to one TF binding motif, or later a pair of TF binding motifs) a single score per each 170 bp long STARR-seq sequence is calculated. This score estimates the probability that a given sequence is occupied by a given TF or TF-pair. The score derivation follows closely [110]. We start by noting that the equilibrium dissociation constant between a TF X and DNA can be approximated using TF-DNA affinity measurements. The equilibrium dissociation constant is defined as:

$$K_{d,X} = \exp\left(\frac{-\Delta G_X}{RT}\right), \tag{A.1}$$

where ΔG_X is the free energy of binding between the TF and DNA, R is the molar gas constant and T is temperature. Now if we make the assumption that each DNA base in a binding site of a TF makes an independent contribution to the total free energy of binding between the TF and DNA, ΔG_X can be calculated using the PWM describing the binding specificity of TF X. Denoting the frequency of base b at position x with $f_{b,x}$ and the background frequency of base b with p_b we can write

$$\Delta G_X = RT \sum_x \ln\left(\frac{f_{b,x}}{p_b}\right), \tag{A.2}$$

where the summation runs over the length of the PWM. With this we can calculate the equilibrium dissociation constant for a binding site starting at position i in a sequence:

$$K_{d,X,i} = \exp\left(\frac{-\Delta G_{X,i}}{RT}\right). \tag{A.3}$$

A.1.2 *TF-DNA binding of single TF*

Now assuming that the binding between the TF X and DNA is non-cooperative, meaning that binding of one molecule of X does not change the affinity of other possible nearby binding sites, we can write the probability that a given binding site i is occupied according to the Hill equation (originally described in [191]):

$$P_i = \frac{[X]}{K_{d,X,i} + [X]}, \tag{A.4}$$

where $[X]$ is the free concentration of TF X. The free concentration is difficult to measure in practice, so we follow the convention used in [110] and set it to equal the equilibrium dissociation constant of the optimal binding site of X (the consensus sequence) for now. The probability of at least one molecule of X binding a given DNA sequence s is obtained by considering all binding sites within the sequence:

$$P_{X,s} \approx 1 - \prod_{i=1}^{N_{sites}} \left(\frac{1}{1 + K_{a,X,i}[X]} \right), \tag{A.5}$$

where N_{sites} is the number of binding sites of TF X on the sequence with affinity > 2 and $K_{a,X,i} = 1/K_{d,X,i}$ is the equilibrium association constant for binding site i. A minimum threshold for a binding site is used to speed up the calculations instead of going through each possible binding site position along each sequence. Notice that in case we suspect that the binding is cooperative in nature, the equation would include the Hill coefficient n [192] as an additional parameter:

$$P_{X,s} \approx 1 - \prod_{i=1}^{N_{sites}} \left(\frac{1}{1 + (K_{a,X,i}[X])^n} \right). \tag{A.6}$$

However, in the following we follow the example of [110] and assume the TF-DNA interaction is non-cooperative. Denoting the PWM match score of TF X at position i with $S_{X,i}$ and the match score of the consensus sequence as $S_{consensus}$, the final score for each sequence is calculated as:

$$P_{X,s} = 1 - \prod_{i=1}^{N_{sites}} \left(\frac{1}{1 + \exp\left(S_{X,i} - S_{consensus}\right)} \right). \tag{A.7}$$

A.1.3 *Cooperative TF-DNA binding of TF-pairs*

In [110], cooperative interactions between TF-pairs are introduced via the notion that binding of XY complex to DNA is thermodynamically equivalent to X binding DNA with higher affinity if Y is already bound. Thus we can calculate the occupancy probability of the XY dimer for a sequence by accounting for all binding site pairs on the sequence:

$$P_{XY,s} \approx 1 - \prod_{i=1}^{N_{sites}} \left(\prod_{j=1}^{M_{sites}} \left(\frac{1}{1 + K_{a,XY,ij}[XY]} \right) \right). \tag{A.8}$$

Here $K_{a,XY,ij}$ is the equilibrium association constant for the XY dimer binding to sites i and j, respectively. In [110] it is defined as the product of the individual equilibrium association constants of X and Y multiplied by a weight function $\kappa_{C,ij}$ that can for example favor adjacent sites over distant ones.

$$K_{a,XY,ij} = \kappa_{C,ij} K_{a,X,i} K_{a,Y,j}. \tag{A.9}$$

For simplicity we chose $\kappa_{C,ij} = 1$ meaning that the interaction between two binding sites is similar regardless of how they are positioned within the 170 bp long random enhancer sequences, although in general this is not true since we know that certain TFs form dimers with specific spacings (see e.g. [56]). The concentration of the dimer XY can be calculated if the dimerization constant is known:

$$[XY] = \frac{[X][Y]}{K_{d,dimer}} = [X][Y]K_{a,dimer}. \tag{A.10}$$

The choice in [110] is to set $K_{a,dimer} = K_{a,Y} = \frac{1}{[Y]}$, which we follow, leading to

$$[XY] = [X]. \tag{A.11}$$

Thus the final score per sequence for a variable considering a pair of TFs X and Y is

$$P_{XY,s} = 1 - \prod_{i=1}^{N_{sites}} \left(\prod_{j=1}^{M_{sites}} \left(\frac{1}{1 + \exp\left(S_{X,i} + S_{Y,j} - S_{X,consensus}\right)} \right) \right). \tag{A.12}$$

The interpretation of this equation is that in the presence of a binding site for Y (and the protein Y), X prefers to bind as a dimer by a factor that is relative to the affinity of the site j for factor Y. As long as the affinity of site j for factor Y is > 0, X binds as a dimer XY more likely to the sites i and j than as a monomer X to site i.

A.1.4 *Determining the values of free concentration [X]*

In [110] the free concentration of each TF was set to equal the k_d of their consensus sequences. In our case this turned out to be problematic since for some TFs with a long PWM (such as p53) matches to the exact consensus sequence are quite rare. This means that setting the scoring as described above will drive the occupancy scores of many functional binding sites to 0 thus greatly reducing the variance of the scores of the variables corresponding to these TFs.

To overcome this problem we decided to use a normalization that is motivated by the fact that generally speaking TFs have approximately $10,000 - 30,0000$ active binding sites in the human genome. Thus we defined the free concentration of each protein to correspond to the strength of the binding site that corresponds to the $10,000$th strongest binding site in the human genome. This means that if a sequence has a single binding site with affinity that would give $10,000$ hits in human genome, this sequence will get occupancy probability of 0.5. To express this mathematically, let us denote with Y the target number of binding sites in the human genome (in our case this is $10,000$), l_x the length of the PWM of TF X, L the length of the input sequences, N the number of sequences in the input library and g the genome size. With these we can calculate how many binding sites need to be calculated from the input library for the strength of the binding site to correspond to the strength of Yth binding site in the human genome:

$$M_X = \frac{(L - n_x)N}{g} Y. \tag{A.13}$$

This rank is calculated from the input library specifically because the binding of the TFs to the enhancer sequences happens in the conditions and concentrations of the input library. Now when we calculate the PWM hits for TF X and sort them, the M_Xth score ($S_{X,Y}$) is the one that corresponds to the affinity of the Yth strongest hit in the genome. With this we can define

$$[X] = \exp\left(-S_{X,Y}\right). \tag{A.14}$$

In reality going through all possible binding sites for all TFs becomes computationally very expensive when we are dealing with over ten million 170 bp long sequences. To make

the computation of the PWM scores faster we decided to consider only the binding sites that give occupancy values of $P_{occ} > 0.01$ and set the scores of weaker binding sites to zero. We can calculate the affinity threshold corresponding to this cut-off, S_{0_X} for each TF separately from:

$$P_0 = 1 - \frac{1}{1 + \exp\left(S_{0_X} - S_{M_X}\right)}, \tag{A.15}$$

where $P_0 = 0.01$ is the minimum occupancy considered. Solving this equation we get

$$S_{0_X} = \ln\left(-\frac{P_0 \exp\left(S_{M_X}\right)}{P_0 - 1}\right). \tag{A.16}$$

With these, the final equation for the occupancy probability of sequence s with TF X is

$$P_{X,s} = 1 - \prod_{i=1}^{N_{sites}} \left(\frac{1}{1 + \exp\left(S_{X,i} - S_{X,10000}\right)}\right), \tag{A.17}$$

where the product runs over all binding sites of X with affinity stronger than S_{0_X}. Similarly, the occupancy probability of s by a TF-TF pair X and Y becomes

$$P_{XY,s} = 1 - \prod_{i=1}^{N_{sites}} \left(\prod_{j=1}^{M_{sites}} \left(\frac{1}{1 + \exp\left(S_{X,i} + S_{Y,j} - S_{X,10000}\right)}\right)\right). \tag{A.18}$$

BIBLIOGRAPHY

1. Regev, A., Teichmann, S., Lander, E., Amit, I., Benoist, C., Birney, E., Bodenmiller, B., Campbell, P., Carninci, P., Clatworthy, M., Clevers, H., Deplancke, B., Dunham, I., Eberwine, J., Eils, R., Enard, W., Farmer, A., Fugger, L., Göttgens, B., Hacohen, N., Haniffa, M., Hemberg, M., Kim, S., Klenerman, P., Kriegstein, A., Lein, E., Linnarsson, S., Lundberg, E., Lundeberg, J., Majumder, P., Marioni, J., Merad, M., Mhlanga, M., Nawijn, M., Netea, M., Nolan, G., Pe'er, D., Phillipakis, A., Ponting, C., Quake, S., Reik, W., Rozenblatt-Rosen, O., Sanes, J., Satija, R., Schumacher, T., Shalek, A., Shapiro, E., Sharma, P., Shin, J., Stegle, O., Stratton, M., Stubbington, M., Theis, F., Uhlen, M., van Oudenaarden, A., Wagner, A., Watt, F., Weissman, J., Wold, B., Xavier, R., Yosef, N. & Human Cell Atlas Meeting Participants. Science forum: the human cell atlas. *elife* **6**, e27041 (2017).
2. Lambert, S. A., Jolma, A., Campitelli, L. F., Das, P. K., Yin, Y., Albu, M., Chen, X., Taipale, J., Hughes, T. R. & Weirauch, M. T. The human transcription factors. *Cell* **172**, 650 (2018).
3. Jolma, A., Yan, J., Whitington, T., Toivonen, J., Nitta, K. R., Rastas, P., Morgunova, E., Enge, M., Taipale, M., Wei, G., Palin, K., Vaquerizas, J., Vincentelli, R., Luscombe, N., Hughes, T., Lemaire, P., Ukkonen, E., Kivioja, T. & Taipale, J. DNA-binding specificities of human transcription factors. *Cell* **152**, 327 (2013).
4. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57 (2012).
5. Zhang, F. & Lupski, J. R. Non-coding genetic variants in human disease. *Human molecular genetics* **24**, R102 (2015).
6. Badis, G., Berger, M. F., Philippakis, A. A., Talukder, S., Gehrke, A. R., Jaeger, S. A., Chan, E. T., Metzler, G., Vedenko, A., Chen, X., Kuznetsov, H., Wang, C., Coburn, D., Newburger, D., Morris, Q., Hughes, T. & ML, B. Diversity and complexity in DNA recognition by transcription factors. *Science* **324**, 1720 (2009).
7. Schmitges, F. W., Radovani, E., Najafabadi, H. S., Barazandeh, M., Campitelli, L. F., Yin, Y., Jolma, A., Zhong, G., Guo, H., Kanagalingam, T., Dai, W., Taipale, J., Emili, A., Greenblatt, J. & Hughes, T. Multiparameter functional diversity of human C2H2 zinc finger proteins. *Genome research* **26**, 1742 (2016).
8. Yin, Y., Morgunova, E., Jolma, A., Kaasinen, E., Sahu, B., Khund-Sayeed, S., Das, P. K., Kivioja, T., Dave, K., Zhong, F., Nitta, K., Taipale, M., Popov, A., Ginno, P., Domcke, S., Yan, J., Schübeler, D., Vinson, C. & Taipale, J. Impact of cytosine methylation on DNA binding specificities of human transcription factors. *Science* **356** (2017).
9. Yan, J., Enge, M., Whitington, T., Dave, K., Liu, J., Sur, I., Schmierer, B., Jolma, A., Kivioja, T., Taipale, M. & Taipale, J. Transcription factor binding in human cells occurs in dense clusters formed around cohesin anchor sites. *Cell* **154**, 801 (2013).

10. Partridge, E. C., Chhetri, S. B., Prokop, J. W., Ramaker, R. C., Jansen, C. S., Goh, S.-T., Mackiewicz, M., Newberry, K. M., Brandsmeier, L. A., Meadows, S. K., Messer, C., Hardigan, A., Coppola, C., Dean, E., Jiang, S., D, S., Mortazavi, A., Wold, B., Myers, R. & Mendenhall, E. Occupancy maps of 208 chromatin-associated proteins in one human cell type. *Nature* **583**, 720 (2020).

11. He, Q., Johnston, J. & Zeitlinger, J. ChIP-nexus enables improved detection of in vivo transcription factor binding footprints. *Nature biotechnology* **33**, 395 (2015).

12. Rhee, H. S. & Pugh, B. F. Comprehensive genome-wide protein-DNA interactions detected at single-nucleotide resolution. *Cell* **147**, 1408 (2011).

13. Arnold, C. D., Gerlach, D., Stelzer, C., Boryń, Ł. M., Rath, M. & Stark, A. Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science* **339**, 1074 (2013).

14. Vockley, C. M., D'Ippolito, A. M., McDowell, I. C., Majoros, W. H., Safi, A., Song, L., Crawford, G. E. & Reddy, T. E. Direct GR binding sites potentiate clusters of TF binding across the human genome. *Cell* **166**, 1269 (2016).

15. Liu, Y., Yu, S., Dhiman, V. K., Brunetti, T., Eckart, H. & White, K. P. Functional assessment of human enhancer activities using whole-genome STARR-sequencing. *Genome biology* **18**, 1 (2017).

16. Van Arensbergen, J., FitzPatrick, V. D., de Haas, M., Pagie, L., Sluimer, J., Bussemaker, H. J. & van Steensel, B. Genome-wide mapping of autonomous promoter activity in human cells. *Nature biotechnology* **35**, 145 (2017).

17. Wang, X., He, L., Goggin, S. M., Saadat, A., Wang, L., Sinnott-Armstrong, N., Claussnitzer, M. & Kellis, M. High-resolution genome-wide functional dissection of transcriptional regulatory regions and nucleotides in human. *Nature communications* **9**, 1 (2018).

18. Muerdter, F., Boryń, Ł. M., Woodfin, A. R., Neumayr, C., Rath, M., Zabidi, M. A., Pagani, M., Haberle, V., Kazmar, T., Catarino, R. R., Schernhuber, K., Arnold, C. & Stark, A. Resolving systematic errors in widely used enhancer activity assays in human cells. *Nature methods* **15**, 141 (2018).

19. Johnson, G. D., Barrera, A., McDowell, I. C., D'Ippolito, A. M., Majoros, W. H., Vockley, C. M., Wang, X., Allen, A. S. & Reddy, T. E. Human genome-wide measurement of drug-responsive regulatory activity. *Nature communications* **9**, 1 (2018).

20. Yuan, Y., Guo, L., Shen, L. & Liu, J. S. Predicting gene expression from sequence: a reexamination. *PLoS computational biology* **3**, e243 (2007).

21. De Boer, C. G., Vaishnav, E. D., Sadeh, R., Abeyta, E. L., Friedman, N. & Regev, A. Deciphering eukaryotic gene-regulatory logic with 100 million random promoters. *Nature biotechnology* **38**, 56 (2020).

22. Lieberman-Aiden, E., Van Berkum, N. L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B. R., Sabo, P. J., Dorschner, M. O., Sandstrom, R., Bernstein, B., Bender, M., Groudine, M., Gnirke, A., Stamatoyannopoulos, J., Mirny, L., Lander, E. & Dekker, J. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *science* **326**, 289 (2009).

23. Fullwood, M. J., Liu, M. H., Pan, Y. F., Liu, J., Xu, H., Mohamed, Y. B., Orlov, Y. L., Velkov, S., Ho, A., Mei, P. H., Chew, E., Huang, P., Welboren, W., Han, Y., Ooi, H., Ariyaratne, P., Vega, V., Luo, Y., Tan, P., Choy, P., Wansa, K., Zhao, B., Lim, K., Leow, S., Yow, J., Joseph, R., Li, H., Desai, K., Thomsen, J., Lee, Y., Karuturi, R., Herve, T., Bourque, G., Stunnenberg, H., Ruan, X., Cacheux-Rataboul, V., Sung, W., Liu, E., Wei, C., Cheung, E. & Ruan, Y. An oestrogen-receptor-α-bound human chromatin interactome. *Nature* **462**, 58 (2009).

24. Dixon, J. R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J. S. & Ren, B. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**, 376 (2012).

25. Ghavi-Helm, Y., Jankowski, A., Meiers, S., Viales, R. R., Korbel, J. O. & Furlong, E. E. Highly rearranged chromosomes reveal uncoupling between genome topology and gene expression. *Nature genetics* **51**, 1272 (2019).

26. Spielmann, M., Lupiáñez, D. G. & Mundlos, S. Structural variation in the 3D genome. *Nature Reviews Genetics* **19**, 453 (2018).

27. Panne, D., Maniatis, T. & Harrison, S. C. An atomic model of the interferon-β enhanceosome. *Cell* **129**, 1111 (2007).

28. Grossman, S. R., Zhang, X., Wang, L., Engreitz, J., Melnikov, A., Rogov, P., Tewhey, R., Isakova, A., Deplancke, B., Bernstein, B. E., Mikkelsen, T. & Lander, E. Systematic dissection of genomic features determining transcription factor binding and enhancer function. *Proceedings of the National Academy of Sciences* **114**, E1291 (2017).

29. Weingarten-Gabbay, S., Nir, R., Lubliner, S., Sharon, E., Kalma, Y., Weinberger, A. & Segal, E. Systematic interrogation of human promoters. *Genome research* **29**, 171 (2019).

30. Hnisz, D., Shrinivas, K., Young, R. A., Chakraborty, A. K. & Sharp, P. A. A phase separation model for transcriptional control. *Cell* **169**, 13 (2017).

31. Sabari, B. R., Dall'Agnese, A., Boija, A., Klein, I. A., Coffey, E. L., Shrinivas, K., Abraham, B. J., Hannett, N. M., Zamudio, A. V., Manteiga, J. C., Li, C., Guo, Y., Day, D., Schuijers, J., Vasile, E., Malik, S., Hnisz, D., Lee, T., Cisse, I., Roeder, R., Sharp, P., Chakraborty, A. & Young, R. Coactivator condensation at super-enhancers links phase separation and gene control. *Science* **361** (2018).

32. Cho, W.-K., Spille, J.-H., Hecht, M., Lee, C., Li, C., Grube, V. & Cisse, I. I. Mediator and RNA polymerase II clusters associate in transcription-dependent condensates. *Science* **361**, 412 (2018).

33. Boija, A., Klein, I. A., Sabari, B. R., Dall'Agnese, A., Coffey, E. L., Zamudio, A. V., Li, C. H., Shrinivas, K., Manteiga, J. C., Hannett, N. M., Abraham, B., Afeyan, L., Guo, Y., Rimel, J., Fant, C., Schuijers, J., Lee, T., Taatjes, D. & Young, R. Transcription factors activate genes through the phase-separation capacity of their activation domains. *Cell* **175**, 1842 (2018).

34. Horn, S., Figl, A., Rachakonda, P. S., Fischer, C., Sucker, A., Gast, A., Kadel, S., Moll, I., Nagore, E., Hemminki, K., Schadendorf, D. & Kumar, R. TERT promoter mutations in familial and sporadic melanoma. *Science* **339**, 959 (2013).

35. Mansour, M. R., Abraham, B. J., Anders, L., Berezovskaya, A., Gutierrez, A., Durbin, A. D., Etchin, J., Lawton, L., Sallan, S. E., Silverman, L. B., Loh, M., Hunger, S., Sanda, T., Young, R. & Look, A. An oncogenic super-enhancer formed through somatic mutation of a noncoding intergenic element. *Science* **346**, 1373 (2014).

36. Sabarinathan, R., Mularoni, L., Deu-Pons, J., Gonzalez-Perez, A. & López-Bigas, N. Nucleotide excision repair is impaired by binding of transcription factors to DNA. *Nature* **532**, 264 (2016).

37. Katainen, R., Dave, K., Pitkänen, E., Palin, K., Kivioja, T., Välimäki, N., Gylfe, A. E., Ristolainen, H., Hänninen, U. A., Cajuso, T., Kondelin, J., Tanskanen, T., Mecklin, J., Järvinen, H., Renkonen-Sinisalo, L., Lepistö, A., Kaasinen, E., Kilpivaara, O., Tuupanen, S., Enge, M., Taipale, J. & Aaltonen, L. CTCF/cohesin-binding sites are frequently mutated in cancer. *Nature genetics* **47**, 818 (2015).

38. Zhang, D., Mishra, S., Brynjolfsson, E., Etchemendy, J., Ganguli, D., Grosz, B., Lyons, T., Manyika, J., Niebles, J. C., Sellitto, M., Shoham, Y., Clark, J. & Perrault, R. The AI index 2021 annual report. *arXiv preprint arXiv:2103.06312* (2021).

39. Cireşan, D., Meier, U., Masci, J. & Schmidhuber, J. *A committee of neural networks for traffic sign classification* in *The 2011 international joint conference on neural networks* (2011), 1918.

40. Ciregan, D., Meier, U. & Schmidhuber, J. *Multi-column deep neural networks for image classification* in *2012 IEEE conference on computer vision and pattern recognition* (2012), 3642.

41. Ioffe, S. & Szegedy, C. *Batch normalization: Accelerating deep network training by reducing internal covariate shift* in *International conference on machine learning* (2015), 448.

42. Alipanahi, B., Delong, A., Weirauch, M. T. & Frey, B. J. Predicting the sequence specificities of DNA-and RNA-binding proteins by deep learning. *Nature biotechnology* **33**, 831 (2015).

43. Kelley, D. R., Snoek, J. & Rinn, J. L. Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. *Genome research* **26**, 990 (2016).

44. Senior, A. W., Evans, R., Jumper, J., Kirkpatrick, J., Sifre, L., Green, T., Qin, C., Žıdek, A., Nelson, A. W., Bridgland, A., Penedones, H., Petersen, S., Simonyan, K., Crossan, S., Kohli, P., Jones, D., Silver, D., Kavukcuoglu, K. & Hassabis, D. Improved protein structure prediction using potentials from deep learning. *Nature* **577**, 706 (2020).

45. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žıdek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S., Ballard, A., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., Petersen, S., Reiman, D., Clancy, E., Zielinski, M., Steinegger, M., Pacholska, M., Berghammer, T., Bodenstein, S., Silver, D., Vinyals, O., Senior, A., Kavukcuoglu, K., Kohli, P. & Hassabis, D. Highly accurate protein structure prediction with AlphaFold. *Nature*, 1 (2021).

46. Bulyk, M. L., Huang, X., Choo, Y. & Church, G. M. Exploring the DNA-binding specificities of zinc fingers with DNA microarrays. *Proceedings of the National Academy of Sciences* **98**, 7158 (2001).

47. Jolma, A., Kivioja, T., Toivonen, J., Cheng, L., Wei, G., Enge, M., Taipale, M., Vaquerizas, J. M., Yan, J., Sillanpää, M. J., Bonke, M., Palin, K., Talukder, S., Hughes, T., Luscombe, N., Ukkonen, E. & Taipale, J. Multiplexed massively parallel SELEX for characterization of human transcription factor binding specificities. *Genome research* **20**, 861 (2010).

48. Barski, A., Cuddapah, S., Cui, K., Roh, T.-Y., Schones, D. E., Wang, Z., Wei, G., Chepelev, I. & Zhao, K. High-resolution profiling of histone methylations in the human genome. *Cell* **129**, 823 (2007).

49. Skene, P. J., Henikoff, J. G. & Henikoff, S. Targeted in situ genome-wide profiling with high efficiency for low cell numbers. *Nature protocols* **13**, 1006 (2018).

50. Oliphant, A. R., Brandl, C. J. & Struhl, K. Defining the sequence specificity of DNA-binding proteins by selecting binding sites from random-sequence oligonucleotides: analysis of yeast GCN4 protein. *Molecular and cellular biology* **9**, 2944 (1989).

51. Pease, A. C., Solas, D., Sullivan, E. J., Cronin, M. T., Holmes, C. P. & Fodor, S. Light-generated oligonucleotide arrays for rapid DNA sequence analysis. *Proceedings of the National Academy of Sciences* **91**, 5022 (1994).

52. Schena, M., Shalon, D., Davis, R. W. & Brown, P. O. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, 467 (1995).

53. Mukherjee, S., Berger, M. F., Jona, G., Wang, X. S., Muzzey, D., Snyder, M., Young, R. A. & Bulyk, M. L. Rapid analysis of the DNA-binding specificities of transcription factors with DNA microarrays. *Nature genetics* **36**, 1331 (2004).

54. Kinzler, K. W. & Vogelstein, B. The GLI gene encodes a nuclear protein which binds specific sequences in the human genome. *Molecular and cellular biology* **10**, 634 (1990).

55. Tuerk, C. & Gold, L. Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. *science*, 505 (1990).

56. Jolma, A., Yin, Y., Nitta, K. R., Dave, K., Popov, A., Taipale, M., Enge, M., Kivioja, T., Morgunova, E. & Taipale, J. DNA-dependent formation of transcription factor pairs alters their binding specificity. *Nature* **527**, 384 (2015).

57. Zhu, F., Farnung, L., Kaasinen, E., Sahu, B., Yin, Y., Wei, B., Dodonova, S. O., Nitta, K. R., Morgunova, E., Taipale, M., Cramer, P. & Taipale, J. The interaction landscape between transcription factors and the nucleosome. *Nature* **562**, 76 (2018).

58. Gilmour, D. S. & Lis, J. T. Detecting protein-DNA interactions in vivo: distribution of RNA polymerase on specific bacterial genes. *Proceedings of the National Academy of Sciences* **81**, 4275 (1984).

59. Solomon, M. J., Larsen, P. L. & Varshavsky, A. Mapping proteinDNA interactions in vivo with formaldehyde: Evidence that histone H4 is retained on a highly transcribed gene. *Cell* **53**, 937 (1988).

60. Blat, Y. & Kleckner, N. Cohesins bind to preferential sites along yeast chromosome III, with differential regulation along arms versus the centric region. *Cell* **98**, 249 (1999).

61. Meyer, C. A. & Liu, X. S. Identifying and mitigating bias in next-generation sequencing methods for chromatin biology. *Nature Reviews Genetics* **15**, 709 (2014).

62. Baranello, L., Kouzine, F., Sanford, S. & Levens, D. ChIP bias as a function of cross-linking time. *Chromosome Research* **24**, 175 (2016).

63. Stormo, G. D., Schneider, T. D., Gold, L. & Ehrenfeucht, A. Use of the 'Perceptron'algorithm to distinguish translational initiation sites in E. coli. *Nucleic acids research* **10**, 2997 (1982).

64. Stormo, G. D., Schneider, T. D. & Gold, L. Quantitative analysis of the relationship between nucleotide sequence and functional activity. *Nucleic acids research* **14**, 6661 (1986).

65. Siebert, M. & Söding, J. Bayesian Markov models consistently outperform PWMs at predicting motifs in nucleotide sequences. *Nucleic acids research* **44**, 6055 (2016).

66. Toivonen, J., Das, P. K., Taipale, J. & Ukkonen, E. MODER2: first-order Markov modeling and discovery of monomeric and dimeric binding motifs. *Bioinformatics* **36**, 2690 (2020).

67. Ching, T., Himmelstein, D. S., Beaulieu-Jones, B. K., Kalinin, A. A., Do, B. T., Way, G. P., Ferrero, E., Agapow, P.-M., Zietz, M., Hoffman, M. M., Xie, W., Rosen, G., Lengerich, B., Israeli, J., Lanchantin, J., Woloszynek, S., Carpenter, A., Shrikumar, A., Xu, J., Cofer, E., Lavender, C., Turaga, S., Alexandari, A., Lu, Z., Harris, D., DeCaprio, D., Qi, Y., Kundaje, A., Peng, Y., Wiley, L., Segler, M., Boca, S., Swamidass, S., Huang, A., Gitter, A. & Greene, C. Opportunities and obstacles for deep learning in biology and medicine. *Journal of The Royal Society Interface* **15**, 20170387 (2018).

68. Zeng, H., Edwards, M. D., Liu, G. & Gifford, D. K. Convolutional neural network architectures for predicting DNA–protein binding. *Bioinformatics* **32**, i121 (2016).

69. Krizhevsky, A., Sutskever, I. & Hinton, G. E. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems* **25**, 1097 (2012).

70. Le, Q. V. *Building high-level features using large scale unsupervised learning* in *2013 IEEE international conference on acoustics, speech and signal processing* (2013), 8595.

71. Trabelsi, A., Chaabane, M. & Ben-Hur, A. Comprehensive evaluation of deep learning architectures for prediction of DNA/RNA sequence binding specificities. *Bioinformatics* **35**, i269 (2019).

72. Hirschberg, J. & Manning, C. D. Advances in natural language processing. *Science* **349**, 261 (2015).

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

74. Zhou, J., Theesfeld, C. L., Yao, K., Chen, K. M., Wong, A. K. & Troyanskaya, O. G. Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk. *Nature genetics* **50**, 1171 (2018).

75. Umarov, R. K. & Solovyev, V. V. Recognition of prokaryotic and eukaryotic promoters using convolutional deep learning neural networks. *PloS one* **12**, e0171410 (2017).

76. Min, X., Chen, N., Chen, T. & Jiang, R. *DeepEnhancer: Predicting enhancers by convolutional neural networks* in *2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* (2016), 637.

77. Andersson, R., Gebhard, C., Miguel-Escalada, I., Hoof, I., Bornholdt, J., Boyd, M., Chen, Y., Zhao, X., Schmidl, C., Suzuki, T., *et al.* An atlas of active enhancers across human cell types and tissues. *Nature* **507**, 455 (2014).

78. Kelley, D. R., Reshef, Y. A., Bileschi, M., Belanger, D., McLean, C. Y. & Snoek, J. Sequential regulatory activity prediction across chromosomes with convolutional neural networks. *Genome research* **28**, 739 (2018).

79. Yu, F. & Koltun, V. Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122* (2015).

80. Avsec, Ž., Weilert, M., Shrikumar, A., Krueger, S., Alexandari, A., Dalal, K., Fropf, R., McAnany, C., Gagneur, J., Kundaje, A. & Zeitlinger, J. Base-resolution models of transcription-factor binding reveal soft motif syntax. *Nature Genetics* **53**, 354 (2021).

81. Avsec, Z., Agarwal, V., Visentin, D., Ledsam, J. R., Grabska-Barwinska, A., Taylor, K. R., Assael, Y., Jumper, J., Kohli, P. & Kelley, D. R. Effective gene expression prediction from sequence by integrating long-range interactions. *bioRxiv* (2021).

82. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. & Polosukhin, I. *Attention is all you need* in *Advances in neural information processing systems* (2017), 5998.

83. Koo, P. K. & Eddy, S. R. Representation learning of genomic sequence motifs with convolutional neural networks. *PLoS computational biology* **15**, e1007560 (2019).

84. Nair, S., Shrikumar, A. & Kundaje, A. fastISM: Performant in-silico saturation mutagenesis for convolutional neural networks. *bioRxiv* (2020).

85. Ancona, M., Ceolini, E., Öztireli, C. & Gross, M. Towards better understanding of gradient-based attribution methods for deep neural networks. *arXiv preprint arXiv:1711.06104* (2017).

86. Shrikumar, A., Greenside, P. & Kundaje, A. *Learning important features through propagating activation differences* in *International Conference on Machine Learning* (2017), 3145.

87. Lundberg, S. & Lee, S.-I. A unified approach to interpreting model predictions. *arXiv preprint arXiv:1705.07874* (2017).

88. Shrikumar, A., Tian, K., Avsec, Ž., Shcherbina, A., Banerjee, A., Sharmin, M., Nair, S. & Kundaje, A. Technical note on transcription factor motif discovery from importance scores (TF-MoDISco) version 0.5. 6.5. *arXiv preprint arXiv:1811.00416* (2018).

89. Finnegan, A. & Song, J. S. Maximum entropy methods for extracting the learned features of deep neural networks. *PLoS computational biology* **13**, e1005836 (2017).

90. Liu, G., Zeng, H. & Gifford, D. K. Visualizing complex feature interactions and feature sharing in genomic deep neural networks. *BMC bioinformatics* **20**, 1 (2019).

91. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *bioinformatics* **25**, 1754 (2009).

92. Aird, D., Ross, M. G., Chen, W.-S., Danielsson, M., Fennell, T., Russ, C., Jaffe, D. B., Nusbaum, C. & Gnirke, A. Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome biology* **12**, 1 (2011).

93. Pfeiffer, F., Gröber, C., Blank, M., Händler, K., Beyer, M., Schultze, J. L. & Mayer, G. Systematic evaluation of error rates and causes in short samples in next-generation sequencing. *Scientific reports* **8**, 1 (2018).

94. Wong, K. M., Suchard, M. A. & Huelsenbeck, J. P. Alignment uncertainty and genomic analysis. *Science* **319**, 473 (2008).

95. Thomas, R., Thomas, S., Holloway, A. K. & Pollard, K. S. Features that define the best ChIP-seq peak calling algorithms. *Briefings in bioinformatics* **18**, 441 (2017).

96. Kivioja, T., Vähärautio, A., Karlsson, K., Bonke, M., Enge, M., Linnarsson, S. & Taipale, J. Counting absolute numbers of molecules using unique molecular identifiers. *Nature methods* **9**, 72 (2012).

97. Robinson, J. T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G. & Mesirov, J. P. Integrative genomics viewer. *Nature biotechnology* **29**, 24 (2011).

98. Palin, K., Pitkänen, E., Turunen, M., Sahu, B., Pihlajamaa, P., Kivioja, T., Kaasinen, E., Välimäki, N., Hänninen, U. A., Cajuso, T., *et al.* Contribution of allelic imbalance to colorectal cancer. *Nature communications* **9**, 1 (2018).

99. Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nature methods* **10**, 1213 (2013).

100. Zhang, Y., Liu, T., Meyer, C. A., Eeckhoute, J., Johnson, D. S., Bernstein, B. E., Nusbaum, C., Myers, R. M., Brown, M., Li, W. & Liu, X. Model-based analysis of ChIP-Seq (MACS). *Genome biology* **9**, 1 (2008).

101. Yan, F., Powell, D. R., Curtis, D. J. & Wong, N. C. From reads to insight: a hitchhiker's guide to ATAC-seq data analysis. *Genome biology* **21**, 1 (2020).

102. Amemiya, H. M., Kundaje, A. & Boyle, A. P. The ENCODE blacklist: identification of problematic regions of the genome. *Scientific reports* **9**, 1 (2019).

103. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nature methods* **9**, 357 (2012).

104. Creyghton, M. P., Cheng, A. W., Welstead, G. G., Kooistra, T., Carey, B. W., Steine, E. J., Hanna, J., Lodato, M. A., Frampton, G. M., Sharp, P. A., *et al.* Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proceedings of the National Academy of Sciences* **107**, 21931 (2010).

105. Ogryzko, V. V., Schiltz, R. L., Russanova, V., Howard, B. H. & Nakatani, Y. The transcriptional coactivators p300 and CBP are histone acetyltransferases. *Cell* **87**, 953 (1996).

106. Delvecchio, M., Gaucher, J., Aguilar-Gurrieri, C., Ortega, E. & Panne, D. Structure of the p300 catalytic core and implications for chromatin targeting and HAT regulation. *Nature structural & molecular biology* **20**, 1040 (2013).

107. Tropberger, P., Pott, S., Keller, C., Kamieniarz-Gdula, K., Caron, M., Richter, F., Li, G., Mittler, G., Liu, E. T., Bühler, M., Margueron, R. & Schneider, R. Regulation of transcription through acetylation of H3K122 on the lateral surface of the histone octamer. *Cell* **152**, 859 (2013).

108. Hebiri, M. & Lederer, J. How correlations influence lasso prediction. *IEEE Transactions on Information Theory* **59**, 1846 (2012).

109. Korhonen, J., Martinmäki, P., Pizzi, C., Rastas, P. & Ukkonen, E. MOODS: fast search for position weight matrix matches in DNA sequences. *Bioinformatics* **25**, 3181 (2009).

110. Granek, J. A. & Clarke, N. D. Explicit equilibrium modeling of transcription-factor binding and gene regulation. *Genome biology* **6**, 1 (2005).

111. Erhard, F. Estimating pseudocounts and fold changes for digital expression measurements. *Bioinformatics* **34**, 4054 (2018).

112. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. & Duchesnay, E. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **12**, 2825 (2011).

113. Glorot, X., Bordes, A. & Bengio, Y. *Deep sparse rectifier neural networks* in *Proceedings of the fourteenth international conference on artificial intelligence and statistics* (2011), 315.

114. Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I. & Salakhutdinov, R. Improving neural networks by preventing co-adaptation of feature detectors. *CoRR* **abs/1207.0580** (2012).

115. Holschneider, M., Kronland-Martinet, R., Morlet, J. & Tchamitchian, P. in *Wavelets* 286 (Springer, 1990).

116. Shensa, M. J. The discrete wavelet transform: wedding the a trous and Mallat algorithms. *IEEE Transactions on signal processing* **40**, 2464 (1992).

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

118. Chollet, F. *et al. Keras* `https://github.com/fchollet/keras`.

119. Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mane, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, Vasudevan, V., Viegas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y. & Zheng, X. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467* (2016).

120. He, K., Zhang, X., Ren, S. & Sun, J. *Delving deep into rectifiers: Surpassing human-level performance on imagenet classification* in *Proceedings of the IEEE international conference on computer vision* (2015), 1026.

121. Ghandi, M., Lee, D., Mohammad-Noori, M. & Beer, M. A. Enhanced regulatory sequence prediction using gapped k-mer features. *PLoS computational biology* **10**, e1003711 (2014).

122. Lee, D. LS-GKM: a new gkm-SVM for large-scale datasets. *Bioinformatics* **32**, 2196 (2016).

123. Santosa, F. & Symes, W. W. Linear inversion of band-limited reflection seismograms. *SIAM Journal on Scientific and Statistical Computing* **7**, 1307 (1986).

124. Bourgon, R., Gentleman, R. & Huber, W. Independent filtering increases detection power for high-throughput experiments. *Proceedings of the National Academy of Sciences* **107**, 9546 (2010).

125. Yuan, G.-C. & Liu, J. S. Genomic sequence is highly predictive of local nucleosome depletion. *PLoS computational biology* **4**, e13 (2008).

126. Gelman, S., Fahlberg, S. A., Heinzelman, P., Romero, P. A. & Gitter, A. Neural networks to learn protein sequence-function relationships from deep mutational scanning data. *bioRxiv*, 2020 (2021).

127. Bailey, T. L. STREME: accurate and versatile sequence motif discovery. *Bioinformatics*. btab203 (2021).

128. Kircher, M., Xiong, C., Martin, B., Schubach, M., Inoue, F., Bell, R. J., Costello, J. F., Shendure, J. & Ahituv, N. Saturation mutagenesis of twenty disease-associated regulatory elements at single base-pair resolution. *Nature communications* **10**, 1 (2019).

129. Fernandes, A. D. & Gloor, G. B. Mutual information is critically dependent on prior assumptions: would the correct estimate of mutual information please identify itself? *Bioinformatics* **26**, 1135 (2010).

130. Shannon, C. E. A mathematical theory of communication. *The Bell system technical journal* **27**, 379 (1948).

131. 1000 Genomes Project Consortium, Auton, A., Brooks, L., Durbin, R., Garrison, E., Kang, H., Korbel, J., Marchini, J., McCarthy, S., McVean, G. & Abecasis, G. A global reference for human genetic variation. *Nature* **526**, 68 (2015).

132. Fornes, O., Castro-Mondragon, J. A., Khan, A., Van der Lee, R., Zhang, X., Richmond, P. A., Modi, B. P., Correard, S., Gheorghe, M., Baranašić, D., Santana-Garcia, W., Tan, G., Chéneby, J., Ballester, B., Parcy, F., Sandelin, A., Lenhard, B., Wasserman, W. & Mathelier, A. JASPAR 2020: update of the open-access database of transcription factor binding profiles. *Nucleic acids research* **48**, D87 (2020).

133. Bucher, P. Weight matrix descriptions of four eukaryotic RNA polymerase II promoter elements derived from 502 unrelated promoter sequences. *Journal of molecular biology* **212**, 563 (1990).

134. Jin, V. X., Singer, G. A., Agosto-Pérez, F. J., Liyanarachchi, S. & Davuluri, R. V. Genome-wide analysis of core promoter elements from conserved human and mouse orthologous pairs. *BMC bioinformatics* **7**, 1 (2006).

135. Frericks Schmidt, H. L., Sperling, L. J., Gao, Y. G., Wylie, B. J., Boettcher, J. M., Wilson, S. R. & Rienstra, C. M. Crystal polymorphism of protein GB1 examined by solid-state NMR spectroscopy and X-ray diffraction. *The Journal of Physical Chemistry B* **111**, 14362 (2007).

136. Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. & Bourne, P. E. The protein data bank. *Nucleic acids research* **28**, 235 (2000).

137. Dreos, R., Ambrosini, G., Groux, R., Cavin Périer, R. & Bucher, P. The eukaryotic promoter database in its 30th year: focus on non-vertebrate organisms. *Nucleic acids research* **45**, D51 (2017).

138. Cooper, G. M., Stone, E. A., Asimenos, G., Green, E. D., Batzoglou, S. & Sidow, A. Distribution and intensity of constraint in mammalian genomic sequence. *Genome research* **15**, 901 (2005).

139. Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S. & Mesirov, J. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences* **102**, 15545 (2005).

140. Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nature biotechnology* **34**, 525 (2016).

141. Pimentel, H., Bray, N. L., Puente, S., Melsted, P. & Pachter, L. Differential analysis of RNA-seq incorporating quantification uncertainty. *Nature methods* **14**, 687 (2017).

142. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G. & Durbin, R. The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078 (2009).

143. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841 (2010).

144. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. journal* **17**, 10 (2011).

145. Bailey, T. L., Boden, M., Buske, F. A., Frith, M., Grant, C. E., Clementi, L., Ren, J., Li, W. W. & Noble, W. S. MEME SUITE: tools for motif discovery and searching. *Nucleic acids research* **37**, W202 (2009).

146. Wei, B., Jolma, A., Sahu, B., Orre, L. M., Zhong, F., Zhu, F., Kivioja, T., Sur, I., Lehtiö, J., Taipale, M. & Taipale, J. A protein activity assay to measure global transcription factor activity reveals determinants of chromatin accessibility. *Nature biotechnology* **36**, 521 (2018).

147. Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y. C., Laslo, P., Cheng, J. X., Murre, C., Singh, H. & Glass, C. K. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Molecular cell* **38**, 576 (2010).

148. Whyte, W. A., Orlando, D. A., Hnisz, D., Abraham, B. J., Lin, C. Y., Kagey, M. H., Rahl, P. B., Lee, T. I. & Young, R. A. Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell* **153**, 307 (2013).

149. Magoč, T. & Salzberg, S. L. FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics* **27**, 2957 (2011).

150. Daley, T. & Smith, A. D. Predicting the molecular complexity of sequencing libraries. *Nature methods* **10**, 325 (2013).

151. Zorita, E., Cusco, P. & Filion, G. J. Starcode: sequence clustering based on all-pairs search. *Bioinformatics* **31**, 1913 (2015).

152. Li, Q., Brown, J. B., Huang, H. & Bickel, P. J. Measuring reproducibility of high-throughput experiments. *The annals of applied statistics* **5**, 1752 (2011).

153. Frith, M. C., Valen, E., Krogh, A., Hayashizaki, Y., Carninci, P. & Sandelin, A. A code for transcription initiation in mammalian genomes. *Genome research* **18**, 1 (2008).

154. McDonald, J. H. *Handbook of biological statistics* (sparky house publishing Baltimore, MD, 2009).

155. Benjamini, Y. & Yekutieli, D. The control of the false discovery rate in multiple testing under dependency. *Annals of statistics*, 1165 (2001).

156. Bardet, A. F., Steinmann, J., Bafna, S., Knoblich, J. A., Zeitlinger, J. & Stark, A. Identification of transcription factor binding sites from ChIP-seq data at high resolution. *Bioinformatics* **29**, 2705 (2013).

157. Wang, L., Chen, J., Wang, C., Uusküla-Reimand, L., Chen, K., Medina-Rivera, A., Young, E. J., Zimmermann, M. T., Yan, H., Sun, Z., Zhang, Y., Wu, S., Huang, H., Wilson, M., Kocher, J. & Li, W. MACE: model based analysis of ChIP-exo. *Nucleic acids research* **42**, e156 (2014).

158. Albert, I., Wachi, S., Jiang, C. & Pugh, B. F. GeneTrack—a genomic data processing and visualization framework. *Bioinformatics* **24**, 1305 (2008).

159. Bailey, T. L. & Elkan, C. Fitting a mixture model by expectation maximization to discover motifs in bipolymers (1994).

160. Waszak, S. M., Kilpinen, H., Gschwind, A. R., Orioli, A., Raghav, S. K., Witwicki, R. M., Migliavacca, E., Yurovsky, A., Lappalainen, T., Hernandez, N., Reymond, A., Dermitzakis, E. & Deplancke, B. Identification and removal of low-complexity sites in allele-specific analysis of ChIP-seq data. *Bioinformatics* **30**, 165 (2014).

161. Degner, J. F., Marioni, J. C., Pai, A. A., Pickrell, J. K., Nkadori, E., Gilad, Y. & Pritchard, J. K. Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data. *Bioinformatics* **25**, 3207 (2009).

162. Bailey, S. D., Virtanen, C., Haibe-Kains, B. & Lupien, M. ABC: a tool to identify SNVs causing allele-specific transcription factor binding from ChIP-Seq experiments. *Bioinformatics* **31**, 3057 (2015).

163. Audic, S. & Claverie, J.-M. The significance of digital gene expression profiles. *Genome research* **7**, 986 (1997).

164. Baran, Y., Subramaniam, M., Biton, A., Tukiainen, T., Tsang, E. K., Rivas, M. A., Pirinen, M., Gutierrez-Arcelus, M., Smith, K. S., Kukurba, K. R., Zhang, R., Eng, C., Torgerson, D., Urbanek, C., GTEx Consortium, Li, J., Rodriguez-Santana, J., Burchard, E., Seibold, M., MacArthur, D., Montgomery, S., Zaitlen, N. & Lappalainen, T. The landscape of genomic imprinting across diverse adult human tissues. *Genome research* **25**, 927 (2015).

165. Lubliner, S., Regev, I., Lotan-Pompan, M., Edelheit, S., Weinberger, A. & Segal, E. Core promoter sequence in yeast is a major determinant of expression level. *Genome research* **25**, 1008 (2015).

166. Arnold, C. D., Zabidi, M. A., Pagani, M., Rath, M., Schernhuber, K., Kazmar, T. & Stark, A. Genome-wide assessment of sequence-intrinsic enhancer responsiveness at single-base-pair resolution. *Nature biotechnology* **35**, 136 (2017).

167. Heinz, S., Romanoski, C. E., Benner, C. & Glass, C. K. The selection and function of cell type-specific enhancers. *Nature reviews Molecular cell biology* **16**, 144 (2015).

168. Nielsen, A. L., Oulad-Abdelghani, M., Ortiz, J. A., Remboutsika, E., Chambon, P. & Losson, R. Heterochromatin formation in mammalian cells: interaction between histones and HP1 proteins. *Molecular cell* **7**, 729 (2001).

169. Zehir, A., Benayed, R., Shah, R. H., Syed, A., Middha, S., Kim, H. R., Srinivasan, P., Gao, J., Chakravarty, D., Devlin, S. M., Hellmann, M., Barron, D., Schram, A., Hameed, M., Dogan, S., Ross, D., Hechtman, J., DeLair, D., Yao, J., Mandelker, D., Cheng, D., Chandramohan, R., Mohanty, A., Ptashkin, R., Jayakumaran, G., Prasad, M., Syed, M., Rema, A., Liu, Z., Nafa, K., Borsu, L., Sadowska, J., Casanova, J., Bacares, R., Kiecka, I., Razumova, A., Son, J., Stewart, L., Baldi, T., Mullaney, K., Al-Ahmadie, H., Vakiani, E., Abeshouse, A., Penson, A., Jonsson, P., Camacho, N., Chang, M., Won, H., Gross, B., Kundra, R., Heins, Z., Chen, H., Phillips, S., Zhang, H., Wang, J., Ochoa, A., Wills, J., Eubank, M., Thomas, S., Gardos, S., Reales, D., Galle, J., Durany, R., Cambria, R., Abida, W., Cercek, A., Feldman, D., Gounder, M., Hakimi, A., Harding, J., Iyer, G., Janjigian, Y., Jordan, E., Kelly, C., Lowery, M., Morris, L., Omuro, A., Raj, N., Razavi, P., Shoushtari, A., Shukla, N., Soumerai, T., Varghese, A., Yaeger, R., Coleman, J., Bochner, B., Riely, G., Saltz, L., Scher, H., Sabbatini, P., Robson, M., Klimstra, D., Taylor, B., Baselga, J., Schultz, N., Hyman, D., Arcila, M., Solit, D.,

Ladanyi, M. & Berger, M. Mutational landscape of metastatic cancer revealed from prospective clinical sequencing of 10,000 patients. *Nature medicine* **23**, 703 (2017).

170. Huang, F. W., Hodis, E., Xu, M. J., Kryukov, G. V., Chin, L. & Garraway, L. A. Highly recurrent TERT promoter mutations in human melanoma. *Science* **339**, 957 (2013).

171. Dreos, R., Ambrosini, G., Périer, R. C. & Bucher, P. The Eukaryotic Promoter Database: expansion of EPDnew and new promoter analysis tools. *Nucleic acids research* **43**, D92 (2015).

172. Berger, M. F., Philippakis, A. A., Qureshi, A. M., He, F. S., Estep, P. W. & Bulyk, M. L. Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. *Nature biotechnology* **24**, 1429 (2006).

173. Gasperini, M., Hill, A. J., McFaline-Figueroa, J. L., Martin, B., Kim, S., Zhang, M. D., Jackson, D., Leith, A., Schreiber, J., Noble, W. S., Trapnell, C., Ahituv, N. & Shendure, J. A genome-wide framework for mapping gene regulation via cellular genetic screens. *Cell* **176**, 377 (2019).

174. Takahashi, H., Lassmann, T., Murata, M. & Carninci, P. 5 end–centered expression profiling using cap-analysis gene expression and next-generation sequencing. *Nature protocols* **7**, 542 (2012).

175. Boyle, A. P., Davis, S., Shulha, H. P., Meltzer, P., Margulies, E. H., Weng, Z., Furey, T. S. & Crawford, G. E. High-resolution mapping and characterization of open chromatin across the genome. *Cell* **132**, 311 (2008).

176. Core, L. J., Martins, A. L., Danko, C. G., Waters, C. T., Siepel, A. & Lis, J. T. Analysis of nascent RNA identifies a unified architecture of initiation regions at mammalian promoters and enhancers. *Nature genetics* **46**, 1311 (2014).

177. Juven-Gershon, T. & Kadonaga, J. T. Regulation of gene expression via the core promoter and the basal transcriptional machinery. *Developmental biology* **339**, 225 (2010).

178. Heintzman, N. D., Stuart, R. K., Hon, G., Fu, Y., Ching, C. W., Hawkins, R. D., Barrera, L. O., Van Calcar, S., Qu, C., Ching, K. A., Wang, W., Weng, Z., Green, R., Crawford, G. & Ren, B. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nature genetics* **39**, 311 (2007).

179. Bergman, D., Jones, T., Liu, V., Siraj, L., Kang, H., Nasser, J., Nguyen, T., Grossman, S., Fulco, C., Lander, E. & Engreitz, J. Compatibility logic of human enhancer and promoter sequences (2021).

180. Thurman, R. E., Rynes, E., Humbert, R., Vierstra, J., Maurano, M. T., Haugen, E., Sheffield, N. C., Stergachis, A. B., Wang, H., Vernot, B., Garg, K., John, S., Sandstrom, R., Bates, D., Boatman, L., Canfield, T., Diegel, M., Dunn, D., Ebersol, A., Frum, T., Giste, E., Johnson, A., Johnson, E., Kutyavin, T., Lajoie, B., Lee, B., Lee, K., London, D., Lotakis, D., Neph, S., Neri, F., Nguyen, E., Qu, H., Reynolds, A., Roach, V., Safi, A., Sanchez, M., Sanyal, A., Shafer, A., Simon, J., Song, L., Vong, S., Weaver, M., Yan, Y., Zhang, Z., Zhang, Z., Lenhard, B., Tewari, M., Dorschner, M., Hansen, R., Navas, P., Stamatoyannopoulos, G., Iyer, V., Lieb, J., Sunyaev, S., Akey, J., Sabo, P., Kaul, R., Furey, T., Dekker, J., Crawford, G. & Stamatoyannopoulos, J. The accessible chromatin landscape of the human genome. *Nature* **489**, 75 (2012).

181. Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., Ziller, M. J., Amin, V., Whitaker, J., Schultz, M., Ward, L., Sarkar, A., Quon, G., Sandstrom, R., Eaton, M., Wu, Y., Pfenning, A., Wang, X., Claussnitzer, M., Liu, Y., Coarfa, C., Harris, R., Shoresh, N., Epstein, C., Gjoneska, E., Leung, D., Xie, W., Hawkins, R., Lister, R., Hong, C., Gascard, P., Mungall, A., Moore, R., Chuah, E., Tam, A., Canfield, T., Hansen, R., Kaul, R., Sabo, P., Bansal, M., Carles, A., Dixon, J., Farh, K., Feizi, S., Karlic, R., Kim, A., Kulkarni, A., Li, D., Lowdon, R., Elliott, G., Mercer, T., Neph, S., Onuchic, V., Polak, P., Rajagopal, N., Ray, P., Sallari, R., Siebenthall, K., Sinnott-Armstrong, N., Stevens, M., Thurman, R., Wu, J., Zhang, B., Zhou, X., Beaudet, A., Boyer, L., De Jager, P., Farnham, P., Fisher, S., Haussler, D., Jones, S., Li, W., Marra, M., McManus, M., Sunyaev, S., Thomson, J., Tlsty, T., Tsai, L., Wang, W., Waterland, R., Zhang, M., Chadwick, L., Bernstein, B., Costello, J., Ecker, J., Hirst, M., Meissner, A., Milosavljevic, A., Ren, B., Stamatoyannopoulos, J., Wang, T. & Kellis, M. Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317 (2015).

182. Carey, M. The enhanceosome and transcriptional synergy. *Cell* **92**, 5 (1998).

183. Crocker, J., Abe, N., Rinaldi, L., McGregor, A. P., Frankel, N., Wang, S., Alsawadi, A., Valenti, P., Plaza, S., Payre, F., Mann, R. & Stern, D. Low affinity binding site clusters confer hox specificity and regulatory robustness. *Cell* **160**, 191 (2015).

184. Maeshima, K., Kaizu, K., Tamura, S., Nozaki, T., Kokubo, T. & Takahashi, K. The physical size of transcription factors is key to transcriptional regulation in chromatin domains. *Journal of Physics: Condensed Matter* **27**, 064116 (2015).

185. Monahan, K., Horta, A. & Lomvardas, S. LHX2-and LDB1-mediated trans interactions regulate olfactory receptor choice. *Nature* **565**, 448 (2019).

186. De Laat, W. & Duboule, D. Topology of mammalian developmental enhancers and their regulatory landscapes. *Nature* **502**, 499 (2013).

187. Chong, S., Dugast-Darzacq, C., Liu, Z., Dong, P., Dailey, G. M., Cattoglio, C., Heckert, A., Banala, S., Lavis, L., Darzacq, X. & Tjian, R. Imaging dynamic and selective low-complexity domain interactions that control gene transcription. *Science* **361** (2018).

188. Novakovsky, G., Saraswat, M., Fornes, O., Mostafavi, S. & Wasserman, W. W. Biologically relevant transfer learning improves transcription factor binding prediction. *Genome biology* **22**, 1 (2021).

189. Koo, P. K., Majdandzic, A., Ploenzke, M., Anand, P. & Paul, S. B. Global importance analysis: An interpretability method to quantify importance of genomic features in deep neural networks. *PLoS computational biology* **17**, e1008925 (2021).

190. Anishchenko, I., Chidyausiku, T. M., Ovchinnikov, S., Pellock, S. J. & Baker, D. De novo protein design by deep network hallucination. *bioRxiv* (2020).

191. Hill, A. V. The possible effects of the aggregation of the molecules of haemoglobin on its dissociation curves. *j. physiol.* **40**, 4 (1910).

192. Stefan, M. I. & Le Novère, N. Cooperative binding. *PLoS computational biology* **9**, e1003106 (2013).