Learning genome-wide interactions of intrinsically disordered proteins with DNA using U-DisCo

Hongwei Tu

CMU-CS-24-154
November 2024

Computer Science Department
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

Thesis Committee:
Jian Ma, Chair
Lei Li

Submitted in partial fulfillment of the requirements for the degree of Masters of Science in Computer Science.

Keywords: intrinsically disordered proteins, protein-DNA interactions, deep learning

Abstract

Proteins are essential regulators of cellular processes. Intrinsically disordered proteins (IDPs), despite lacking stable tertiary structures under physiological conditions, play crucial yet often underexplored roles in biological processes. With recent experimental advances like DisP-seq for probing IDP-DNA binding, there is a pressing need for efficient, interpretable computational methods to identify sequence determinants of IDP-DNA interactions and analyze their cooperative effects on gene regulation. To address this, we develop U-DisCo, a novel deep learning model that predicts base-resolution IDP-DNA binding profiles directly from DNA sequences. Leveraging a U-Net architecture, U-DisCo captures both local base-level interactions and long-range dependencies up to 20 kilobases with high accuracy and computational efficiency, outperforming the baseline BPNet. By incorporating ATAC-seq data, U-DisCo enables robust cross-cell type predictions as a multimodal framework. U-DisCo identified key IDP-binding motifs, revealing distinct interaction patterns and cooperative behaviors across different IDPs. Interestingly, we observed short-range interactions for motifs like AP-2 and EWS-FLI1 (single GGAA motif), while others exhibited independent, enhancer-like functions. Further analysis revealed that some IDPs favored certain strand orientations, suggesting their involvement in specific regulatory mechanisms. Overall, U-DisCo is the first computational approach to explore multiple IDPs within a single cell type, offering a versatile framework for studying IDP-mediated gene regulation and genome-wide regulatory elements.

Acknowledgments

I would like to express my deepest gratitude to my thesis advisor, Professor Jian Ma, for his invaluable guidance, insight, and encouragement throughout this research journey. His mentorship has been instrumental in shaping this work, and his support has greatly enhanced my academic and research skills. I am also grateful to Professor Lei Li for his constructive feedback and expertise, which have contributed significantly to the development of this thesis.

I would like to extend special thanks to Yang Zhang in the Ma Lab for helping me formulate research ideas, providing thoughtful feedback, and assisting with revisions of the thesis. His support has been indispensable in bringing this project to completion.

My sincere appreciation goes to Professor Ruben Martins, Professor Dave Eckhardt, and Angy Malloy for their organizational and logistical support throughout the MSCS program. Their dedication to running the program smoothly has created an environment in which students can thrive.

Contents

1 Introduction **1**
- 1.1 Background . . . 1
- 1.2 Limitations of sequencing methods . . . 1
- 1.3 Key challenges with machine learning methods . . . 2
- 1.4 Contributions . . . 3

2 Related machine learning approaches **5**
- 2.1 Early machine learning methods . . . 5
- 2.2 Deep learning . . . 5
  - 2.2.1 Convolutional neural networks . . . 6
  - 2.2.2 Transformers . . . 6
- 2.3 Taxonomy of machine learning models . . . 7
  - 2.3.1 Binary models . . . 7
  - 2.3.2 Profile models . . . 7

3 The U-DisCo framework **9**
- 3.1 Overview of U-DisCo . . . 10
- 3.2 Data collection . . . 10
- 3.3 U-DisCo model architecture and training . . . 10
- 3.4 Baseline models, training, and benchmarking . . . 12
- 3.5 Peak intersection ratios . . . 12
- 3.6 Results: performance evaluation across cell lines . . . 13

4 Identification of IDP binding motifs **17**
- 4.1 Model interpretation and importance scores . . . 17
- 4.2 Motif identification . . . 17
- 4.3 Chromatin state analysis . . . 18
- 4.4 State association analysis . . . 18
- 4.5 Results: IDP-binding motifs and their regulatory roles . . . 18

5 Cooperative intercations between IDPs **23**
- 5.1 Co-occurrence analysis . . . 23
- 5.2 Neighborhood density analysis . . . 24
- 5.3 Strand-specific preference analysis . . . 24

5.4 Results 24
5.4.1 Cooperative behaviors between IDPs 24
5.4.2 IDP binding preferences for orientation and spacing 26

6 Conclusion **27**

Bibliography **29**

List of Figures

3.1 U-DisCo model architecture. 9
3.2 Detailed illustration of the U-Net architecture. 11
3.3 Performance evaluation of U-DisCo against baselines across cell lines. 14
3.4 Example predictions with low PCCs but high peak intersection ratios. 15

4.1 Identification of IDP-binding motifs. 19
4.2 PONDR and metapredict V2 IDR predictions. 20
4.3 Continuation of PONDR and metapredict V2 IDR predictions. 21

5.1 Co-occurrence analysis of cooperative behaviors and binding preferences. 25

Chapter 1

Introduction

1.1 Background

Proteins perform diverse functions in organisms [1–3], yet most of our knowledge is limited to proteins with stable and well-defined structures [4–8]. Increasing evidence suggests that those proteins lacking well-defined structures, known as intrinsically disordered proteins (IDPs) [9, 10], also play crucial roles in processes such as transcription, RNA processing, signaling, and cell cycle control [11–17]. The structural plasticity of IDPs enables them to adopt multiple conformations [18, 19] and interact with various targets under different physiological conditions [20, 21]. For instance, the H1 linker histone undergoes disorder-to-order transitions upon binding to target molecules [22]. However, the full role of IDPs in gene regulation remains incomplete due to limited experimental and computational methods for probing their DNA binding profiles.

As such, this work aims to develop an efficient computational framework to explore genome-wide interactions between IDPs and DNA. By training and interpreting our deep learning model using a recent IDP-DNA binding assay, we enable the identification of key DNA sequence determinants and regulatory mechanisms underlying IDP-mediated gene regulation.

1.2 Limitations of sequencing methods

Traditional sequencing and probing methods, such as ChIP-seq (Chromatin Immunoprecipitation Sequencing) and CUT&RUN (Cleavage Under Targets and Release Using Nuclease) [23–25], have been widely used to map the binding sites of transcription factors (TFs) and other DNA-associated proteins. ChIP-seq, for instance, uses a specific antibody to target the protein of interest, allowing the isolation of DNA regions bound by that protein. The DNA is then sequenced, providing a high-resolution map of protein-DNA interactions across the genome. Similarly, CUT&RUN combines an antibody-guided approach with targeted cleavage by nuclease, allowing for more precise mapping with less background noise compared to ChIP-seq.

While these techniques have been successful in mapping well-characterized and structured proteins with available antibodies, they are significantly limited for studying IDPs, primarily due to the reliance on specific antibodies to recognize target proteins. IDPs often exhibit weak immune responses and low-affinity binding due to their flexible, unstructured nature [26, 27].

Unlike the stable, high-affinity interactions needed for antibody recognition, IDPs rely on transient interactions for rapid signaling [12, 28]. As a result, antibody-based methods may struggle to capture the genome-wide interaction patterns of IDPs.

To address this challenge, disordered protein precipitation followed by DNA sequencing (DisP-seq) was recently developed [29], an antibody-independent assay that allows for simultaneous mapping of multiple DNA-associated disordered proteins and quantifies their cooperative behaviors in gene regulation. Nevertheless, no sequence-based computational methods currently exist to directly investigate the impact of DNA sequence determinants on IDP binding or their cooperative interactions.

1.3 Key challenges with machine learning methods

Machine learning models, particularly convolutional neural networks (CNNs) in deep learning, have been applied in the prediction of protein-DNA interactions and chromatin profiles [30–36]. Early CNN-based methods, with their sequential, layer-by-layer architectures, were designed to predict either binary binding labels or low-resolution continuous signals averaged across genomic bins of 100 to 200 base pairs (bp) [30, 31, 35, 36]. While these models could offer insights into protein-DNA interactions, their low resolution limits their ability to capture finer binding details for IDP-DNA interactions that may vary at base-level resolution.

More recent methods, such as BPNet [33], have achieved higher precision by extending CNNs to predict base-resolution profiles; however, their local receptive fields restrict them to relatively short sequences (around 1 kb), limiting their ability to model long-range interactions across tens of kilobases that are crucial for understanding IDP-mediated gene regulation. Since IDPs have flexible binding patterns and interact with distal regulatory regions, these models may underperform in capturing the dynamic chromatin environment influenced by IDPs.

Transformers, with their self-attention mechanisms, allow models to attend to positions across an entire sequence and capture long-range interactions more effectively [34, 37]. While transformers have shown promise in genomics by expanding the feasible sequence length for interaction modeling, they are highly computationally intensive when applied to large-scale genomic data comprising numerous long DNA sequences. The memory requirements for self-attention scale quadratically with the input length ($O(N^2)$), making transformers prohibitively costly when handling genome-wide data at base-level resolution.

To date, no machine learning models have explored genome-wide interactions between IDPs and DNA. One reason is the lack of labeled IDP binding data. Before DisP-seq, an antibody-independent assay for simultaneous mapping of IDPs, most available datasets of protein-DNA interactions focused on well-structured transcription factors. There were no IDP-specific benchmark datasets to guide model development or assess performance. Moreover, addressing this gap requires models that can accurately predict base-resolution IDP-DNA binding profiles, model long-range dependencies, and operate with computational efficiency for large-scale analyses. Achieving these goals will demand innovative architectural designs, improved data representations, and efficient analysis approaches.

1.4 Contributions

To bridge these gaps, we combine deep learning with the DisP-seq assay. We introduce U-DisCo, a novel deep learning approach predicting base-resolution DisP-seq profiles from DNA sequences using a U-Net architecture. Unlike conventional CNNs, U-Net combines downsampling and upsampling paths with skip connections in its U-shaped design, enabling U-DisCo to efficiently capture local base-level interactions and long-range dependencies up to 20 kilobases. This capability is particularly useful for identifying complex sequence patterns necessary for understanding IDP functions in gene regulation. We comprehensively evaluated U-DisCo against baseline models on DisP-seq datasets from three cell lines and ChIP-nexus data from mouse embryonic stem cells (mESCs). Our key contributions are as follows:

- U-DisCo is the first computational approach to explore protein-DNA interactions for multiple IDPs;
- Leveraging U-Net, U-DisCo captures both base-resolution and long-range interactions with computational efficiency, outperforming baseline models such as BPNet;
- By incorporating ATAC-seq data as an optional input, U-DisCo functions as a multimodal framework generalizing across cell types, achieving performance on par with biological replicates;
- Importantly, U-DisCo identifies key IDP-binding motifs, revealing genome-wide regulatory roles of IDPs, including spatial distribution, cooperative interactions, and binding preferences.

Together, U-DisCo provides a versatile method for exploring IDP-mediated gene regulation and regulatory mechanisms genome-wide.

Chapter 2

Related machine learning approaches

In this chapter, we provide a comprehensive overview of machine learning approaches that have been applied to study protein-DNA interactions and chromatin profiles. We start with early machine learning models and then explore deep learning models, including convolutional neural networks (CNNs) and transformers. Finally, we categorize these approaches based on their prediction objectives: binary methods for classifying binding sites and profile methods for predicting continuous chromatin signals.

2.1 Early machine learning methods

Early machine learning methods relied on manually engineered features to model protein-DNA interactions. Feature engineering was necessary because these models required structured representations of DNA sequences or protein-binding patterns as inputs. Common features included k-mer frequencies and position weight matrices (PWMs) of known DNA motifs, used as inputs to traditional machine learning models like support vector machines (SVMs) and random forests. For example, the gkm-SVM approach was developed to predict transcription factor (TF) binding sites [38]. By transforming raw DNA sequences into feature vectors based on gapped k-mer frequencies, SVMs could then learn to classify regions as either bound or unbound by specific proteins. Similarly, an integrative approach trained random forests to predict TF binding sites, using features such as nucleotide positional dependencies, DNA structure, and PWMs [39].

However, the reliance on feature engineering led to several limitations. Manually extracted features often fail to capture the complex, high-dimensional interactions underlying protein-DNA binding. Moreover, the choice of features could bias the model toward known binding motifs, limiting its ability to discover novel interactions or binding patterns. Nevertheless, early machine learning models demonstrated the potential of computational approaches for identifying potential DNA sequence determinants, paving the way for more complex models.

2.2 Deep learning

Unlike traditional machine learning methods, deep learning models are capable of learning hierarchical representations directly from raw sequence data, eliminating the need for manual feature

engineering. Deep learning has facilitated the development of models that can capture complex patterns in DNA sequences and chromatin profiles.

2.2.1 Convolutional neural networks

Convolutional neural networks (CNNs) use convolutional layers to scan DNA sequences, identifying motifs and patterns within short, contiguous regions. CNNs are effective at capturing local dependencies in DNA like transcription factor binding motifs. The typical structure of a CNN consists of sequential convolutional layers, each applying a set of filters to detect sequence features. For example, an initial convolutional layer might identify nucleotide motifs, while subsequent layers may learn to recognize more complex patterns by combining lower-level features. Pooling layers often follow convolutional layers, reducing the spatial dimensions of the data and focusing the model on the most prominent features.

Early CNN-based models, such as DeepBind [30] and Basset [31], were applied to protein-DNA interaction prediction. DeepBind used CNNs to predict the binary presence or absence of transcription factor binding, while Basset extended this approach to simultaneously predict chromatin accessibility scores in 164 cell types. Both models demonstrated that CNNs could automatically learn biologically relevant motifs from raw sequence data, outperforming traditional machine learning models that required feature engineering. Later, models like BPNet [33] further refined CNN architectures to predict base-resolution binding profiles, enabling high-resolution mapping of protein-DNA interactions.

CNNs are well-suited to predict binding interactions and chromatin accessibility within short or medium sequences. However, CNNs' reliance on local receptive fields limits their ability to capture long-range dependencies, especially when modeling regulatory interactions that span several kilobases.

2.2.2 Transformers

Transformers [37] are capable of capturing long-range dependencies that CNNs struggle to model. Transformers use self-attention mechanisms, which allow each position in a sequence to attend to every other position. The self-attention mechanism operates by computing a set of attention weights for each position in the sequence, determining which other positions are relevant for that position's representation. Therefore, transformers can model dependencies extending across the entire sequence regardless of length. This enables transformers to model interactions over long DNA sequences and study distal regulatory elements and complex gene regulation patterns.

For instance, Enformer [34] is a transformer-based model to predict various gene expression and chromatin profiles. By leveraging self-attention, Enformer can integrate long-range interactions across DNA sequences up to 100 kb. Built on top of Enformer, Borzoi [40] predicts RNA-seq coverage from input DNA sequences. Borzoi accounts for interactions across 524 kb sequences by combining convolutional layers and transformer layers.

While transformers can capture long-range interactions, the computational cost of the self-attention mechanism scales quadratically with input length. This poses a challenge when analyzing genome-wide sequence data that are often several kb long.

2.3 Taxonomy of machine learning models

2.3.1 Binary models

Binary models predict binary labels for genomic regions (e.g., bound or unbound by a protein or factor), providing a straightforward modeling approach for binding potential. The raw output of these models is typically a probability score indicating the likelihood of binding at a specific location. The majority of models fall into this category [30, 31, 35, 36] and allow researchers to identify regions of interest within the genome. However, they do not benefit from profile shape information and thus are particularly prone to overfitting compared to models that predict profile shapes [41].

2.3.2 Profile models

Profile models predict continuous signals, representing binding intensity or chromatin accessibility across a genomic region. Unlike binary models, profile models capture the quantitative nature of protein-DNA interactions, enabling high-resolution mapping of chromatin features. These models can learn the strength or frequency of binding across each base pair, as well as the spatial distribution of interactions within a region.

Recently, models in this category have been developed, such as BPNet [33], Enformer [34], and Borzoi [40], which can finely track sequence patterns along peak regions, allowing them to accurately locate motifs. These models have proven valuable in studying dynamic chromatin landscapes, as they provide a detailed view of binding and accessibility patterns.

Chapter 3

The U-DisCo framework

We present U-DisCo, a novel deep learning framework to explore IDP-DNA interactions genome-wide. Based on a U-Net architecture, U-DisCo accurately and efficiently predicts base-resolution DisP-seq profiles directly from DNA sequences, capturing both local and long-range dependencies. By incorporating ATAC-seq data, U-DisCo functions as a multimodal framework and enables cross-cell line predictions. Here, we detail the data collection, model architecture, training, and evaluation of the U-DisCo framework. In the next chapters, we demonstrate that the interpretation of U-DisCo predictions reveals distinct interaction patterns and cooperative behaviors across different IDPs, providing insights into IDP-mediated gene regulation.

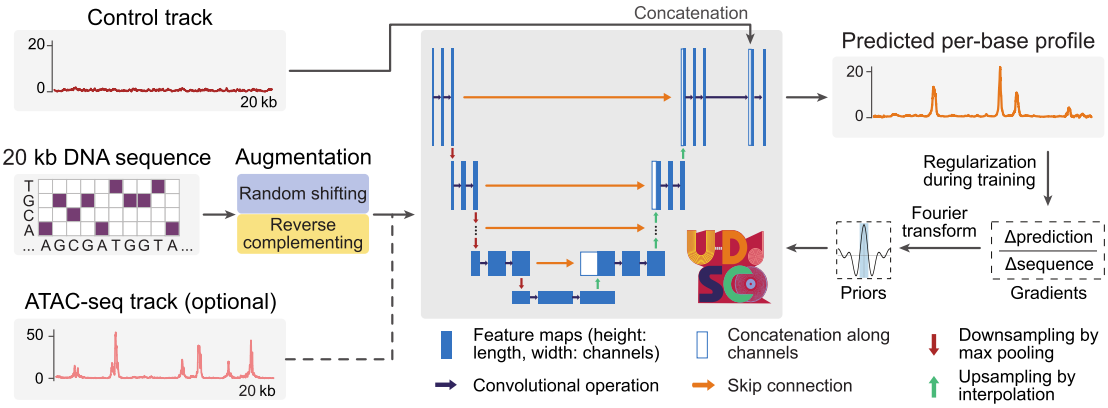

Figure 3.1: U-DisCo model architecture. The inputs to U-DisCo include a 20 kb DNA sequence, a control track of the same length, and an optional ATAC-seq signal track for multimodal learning. Data augmentation techniques such as random shifting and reverse complementing are employed during the training step. U-DisCo is based on a U-Net architecture, incorporating convolutional blocks and skip connections to predict base-resolution DisP-seq profiles. Fourier-based priors are used to regularize gradients during training to improve interpretability.

3.1 Overview of U-DisCo

An illustration of the U-DisCo model is shown in **Fig.** 3.1. The inputs to U-DisCo include a 20 kb DNA sequence centered on the peak, a control track for the same region, and an optional ATAC-seq signal track. The control track corrects for biases and reduces noise during learning, as established in [33, 41], while the optional ATAC-seq input enables U-DisCo to perform cross-cell line predictions. U-DisCo outputs a base-resolution DisP-seq profile matching the length of the input DNA sequence. Based on a U-Net architecture, U-DisCo uses a series of convolutional blocks for both downsampling and upsampling. The network contains eight downsampling and eight upsampling blocks, with skip connections transferring feature maps between corresponding blocks to preserve information at different length scales (see **Fig.** 3.2 for a detailed illustration). During training, multinomial negative log-likelihood loss is employed, and data augmentation techniques such as random shifting and reverse complementing improve generalization. To prevent the model from focusing on short bursts along DNA sequences, we use Fourier-based priors to penalize high-frequency gradient components [41].

3.2 Data collection

The DisP-seq assay identified 22,633 peaks in SKNMC cells [29]. We selected 22,632 peaks with sufficient margins to accommodate a 20 kb window around each peak center, averaged across two biological replicates. This window length was chosen to ensure full coverage of each peak. Chromosomes 1, 8, and 21 were held out for evaluation, with the remaining chromosomes split into training (80%) and validation (20%) sets. We also included two additional cell lines, H446 and MRC5, and processed raw sequencing reads following the pipeline from [29], identifying 20,334 and 19,372 peaks, respectively. ATAC-seq data for all cell lines were downloaded from the ENCODE project [42], with DNase-seq data for IMR90 used as a proxy for MRC5. DNA sequences from the hg19 reference genome were one-hot encoded (A: $[1, 0, 0, 0]$, C: $[0, 1, 0, 0]$, G: $[0, 0, 1, 0]$, T: $[0, 0, 0, 1]$). Control data served as background signals to correct for biases and noise in DisP-seq profiles.

Additionally, we collected ChIP-nexus data from mouse embryonic stem cells (mESCs) reported in [33]. Peaks with maximum signal intensities above the third quartile across all peaks were retained to reduce noise, and 1 kb windows were extracted around peak centers. Chromosomes 1 and 8 were held out for evaluation, with the remainder split into training (80%) and validation (20%) sets. DNase-seq data for mESC E14 from ENCODE were used as a proxy for ATAC-seq, and DNA sequences from the mm10 reference genome were one-hot encoded. PATCH–CAP data were used as control tracks.

3.3 U-DisCo model architecture and training

U-DisCo is based on a U-Net architecture [43] with double convolutional blocks for downsampling and upsampling. Each block contains two convolutional layers (kernel size 3, padding 1), batch normalization, ReLU activation, and dropout (rate 0.1). Max-pooling (stride 2) is used for

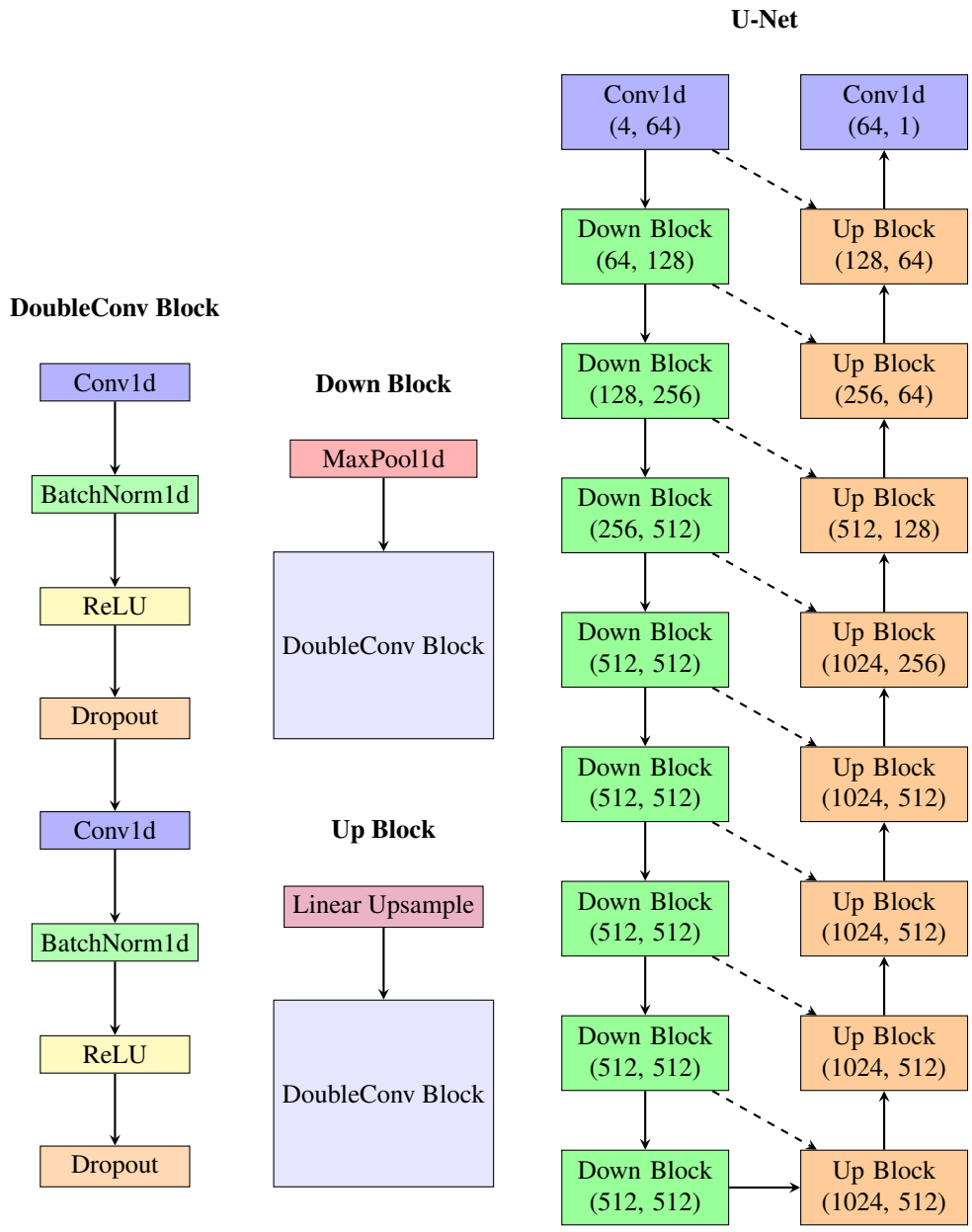

Figure 3.2: Illustration of the U-DisCo model, showing the building blocks and U-Net architecture. Dashed arrows represent skip connections. Input and output channels are specified in parentheses as (in_channels, out_channels). All convolutional layers (Conv1d) use padding set to 'same' to preserve sequence length. Concatenation with the bias track is not shown in the U-Net for simplicity. DoubleConv Block, double convolutional block. Down Block, downsampling block. Up Block, upsampling block.

downsampling, while linear upsampling (factor 2) is used in the upsampling path. The network includes eight downsampling and eight upsampling blocks, with channels doubling from 64 to 512 during downsampling, and halving symmetrically during upsampling. Skip connections

transfer feature maps between the downsampling and upsampling paths.

A one-hot encoded DNA sequence, optionally concatenated with an ATAC-seq track, is first processed through a convolutional layer (kernel size 25) to extract coarse-grained motif features. After passing through the U-Net, the feature maps are reduced via another convolutional layer (kernel size 25), concatenated with a control track, and passed through a final convolution (kernel size 1) to generate predicted DisP-seq profiles. A detailed depiction of the U-DisCo architecture is shown in **Fig.** 3.2.

The model was trained using a multinomial negative log-likelihood (NLL) loss:

$$\text{NLL} = -\log(N!) + \sum_{i=1}^{L} \log(x_i!) - \sum_{i=1}^{L} x_i \log(p_i), \tag{3.1}$$

where N is the sum of all DisP-seq signals in the profile, L is the profile length, x_i is the observed signal at position i, and p_i is the predicted probability at position i.

We applied several data augmentation techniques, including random shifts up to 1 kb and reverse complementing with a 20% probability to enhance generalization. Fourier-based priors [41] were used to penalize high-frequency gradient components with a frequency limit of 3000, a softness of 0.2, and Gaussian smooth sigma of 3, as detailed in [41]. The regularization loss term was weighted by 15,000, approximately half of the converged NLL loss.

The model was trained using an AdamW optimizer (learning rate 0.0005, batch size 32, 200 epochs), with exponential moving averages (EMA, decay rate 0.99) of model weights for stabilization during evaluation. The model with the lowest validation loss was used for the final evaluation on the held-out chromosomes.

3.4 Baseline models, training, and benchmarking

U-DisCo was benchmarked against BPNet [33] and LightGBM [44]. BPNet employs convolutional blocks (kernel size 3, padding 1) with dilation rates doubling per block, 64 channels, and residual connections. LightGBM, a gradient boosting algorithm, took k-mer frequencies ($k = 2$, 31 bp windows centered at each position) and ATAC-seq data as input features. It was trained with L^1 loss, a learning rate of 0.05, 2^{12} leaves, a minimum of 50 data points per leaf, and a maximum depth of 14 for 2000 epochs, with early stopping after 10 epochs.

All models were trained and evaluated on the same training and validation sets. Both U-DisCo and BPNet were trained using the same procedure, with data augmentation techniques and Fourier-based priors for regularization. The models with the lowest validation loss were chosen for evaluation on the held-out chromosomes. We evaluated models using Pearson correlation coefficients (PCCs) between predicted and observed windows.

3.5 Peak intersection ratios

To investigate the overlap between predicted and observed DisP-seq peak regions, we designed peak intersection ratios. This metric quantifies the agreement between predicted and observed

peak sets instead of signal intensity, measuring the model's ability to recognize peak locations. For an observed or predicted DisP-seq region, we defined its peak set by selecting areas with DisP-seq signals above the median. The peak intersection ratio between two DisP-seq regions, with peak sets P_1 and P_2, was calculated as:

$$\text{precision} = \frac{|P_1 \cap P_2|}{|P_1|}, \qquad \text{recall} = \frac{|P_1 \cap P_2|}{|P_2|}, \tag{3.2}$$

$$\text{peak intersection ratio} = \frac{2}{\text{precision}^{-1} + \text{recall}^{-1}}, \tag{3.3}$$

where $|\cdot|$ represents the cardinality of a set.

3.6 Results: performance evaluation across cell lines

We trained and evaluated U-DisCo using 22,632 DisP-seq peak regions from SKNMC cells, averaged across two biological replicates, with a window length of 20 kb. U-DisCo was benchmarked against biological replicates from DisP-seq and baseline models, including BPNet and LightGBM. Pearson correlation coefficients (PCCs) between the two biological replicates and between predicted and observed DisP-seq profiles on held-out chromosomes (chr1, chr8, and chr21) are shown in **Fig.** 3.3a, with mean PCCs for five training runs and per-peak PCCs. U-DisCo outperformed all baselines, achieving PCCs comparable to biological replicate reproducibility, highlighting U-DisCo's reliability in predicting DisP-seq profiles.

We further plotted U-DisCo's PCCs against biological replicates for each peak region, with density plots grouped by DisP island state (island/non-island) and number of summits (one or more) (**Fig.** 3.3b). DisP islands are large DisP-seq clusters defined in [29]. U-DisCo achieved high PCCs consistent with biological replicates, with higher performance on non-island regions and peaks with single summits. To examine peaks with lower PCCs, we analyzed peak intersection ratios between predicted and observed peak regions (**Fig.** 3.3c). For U-DisCo, these ratios were grouped by PCC thresholds (≥ 0.5 and < 0.5), while for biological replicates, they were calculated between two replicates. Notably, U-DisCo consistently achieved higher intersection ratios than biological replicates, suggesting low PCCs were due to mismatched profile shapes rather than incorrect peak locations (examples in **Fig.** 3.4).

To further assess U-DisCo's performance and versatility across data types, we compared U-DisCo with BPNet and LightGBM using DisP-seq datasets from two additional cell lines (H446 and MRC5), and ChIP-nexus in mESCs originally used to train BPNet [33]. The ChIP-nexus data provide base-resolution binding profiles for four pluripotency TFs (Oct4, Sox2, Nanog, and Klf4) on the positive and negative strands. Models were trained to predict binding profiles for all TFs on both strands using multi-task learning, making the task challenging. Mean PCCs across held-out peaks for all cell lines are presented in **Fig.** 3.3d, where U-DisCo consistently outperformed the baselines, particularly in mESCs.

We further tested U-DisCo's multimodal capability by integrating additional ATAC-seq inputs in three distinct settings: (i) within-line, where training and testing occurred on the same cell line; (ii) cross-line, where models were trained on SKNMC using its ATAC-seq data and tested on a different cell line using ATAC-seq data from the target cell line; and (iii) mixed,

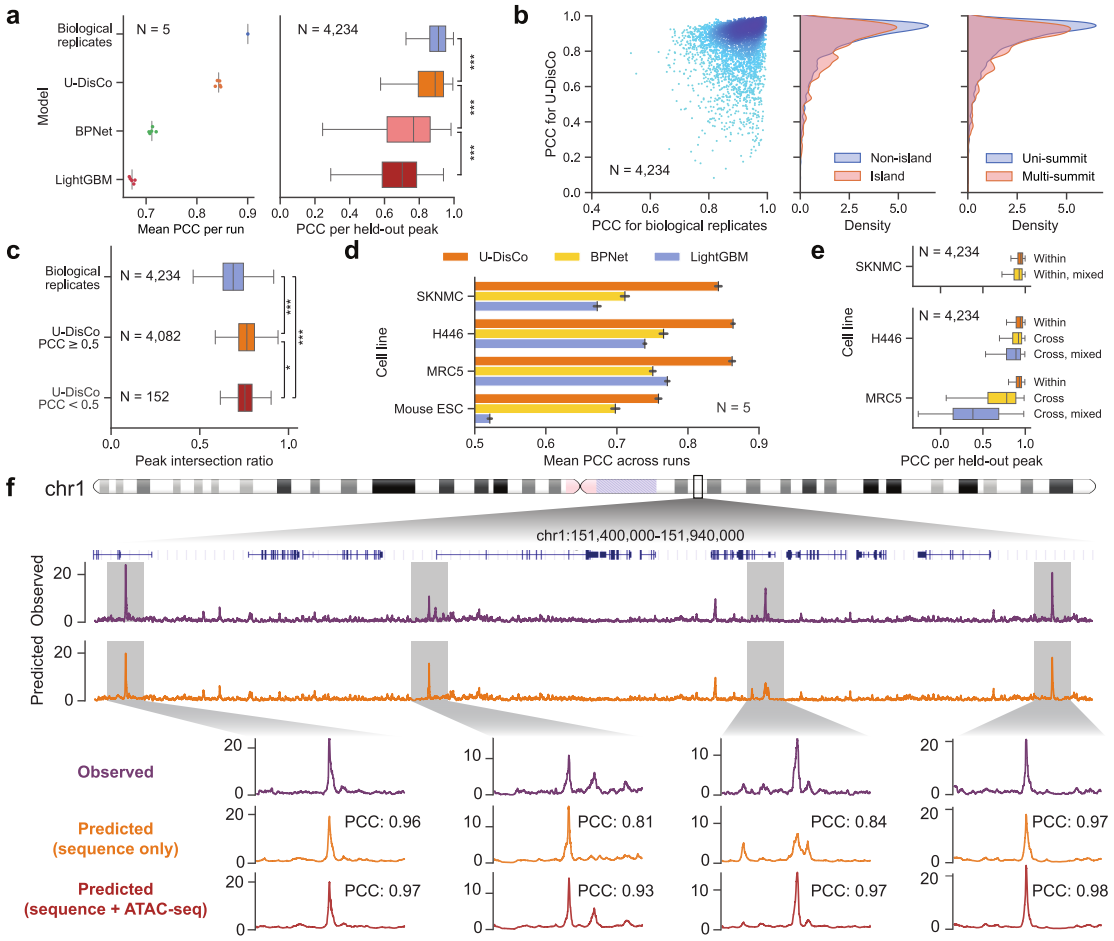

Figure 3.3: Performance evaluation of U-DisCo against baselines across cell lines. **a**. Pearson correlation coefficients (PCCs) between predicted and observed DisP-seq profiles on held-out peaks for U-DisCo and baselines, and PCCs between two biological replicates. Mean PCCs across peaks for five runs are provided, along with per-peak PCCs. **b**. Scatter plot of U-DisCo's PCCs vs. biological replicates, with density plots grouped by DisP island state and number of summits. **c**. Peak intersection ratios (predicted vs. observed peak overlap) for U-DisCo, grouped by PCC thresholds, and between both biological replicates. **d**. Mean PCCs across held-out peaks for U-DisCo and baselines on four cell lines. Error bars represent the standard deviation across five training runs. **e**. PCCs for multimodal U-DisCo across cell lines: within-line (trained/tested on same cell line), cross-line (trained on SKNMC, tested on another cell line), and mixed (trained on SKNMC using its ATAC-seq data, tested using the average ATAC-seq track from SKNMC and H446). **f**. Example predictions for a chr1 region in SKNMC cells, showing observed/predicted DisP-seq profiles and PCCs with/without ATAC-seq input. $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$ (Mann-Whitney U test for significance in distribution differences between two groups).

where training was on SKNMC using its ATAC-seq data, and testing used the average ATAC-seq track from both SKNMC and H446 rather than the target cell line. Boxplots in **Fig.** 3.3e show the per-peak PCCs across these settings. U-DisCo performed well within each cell line,

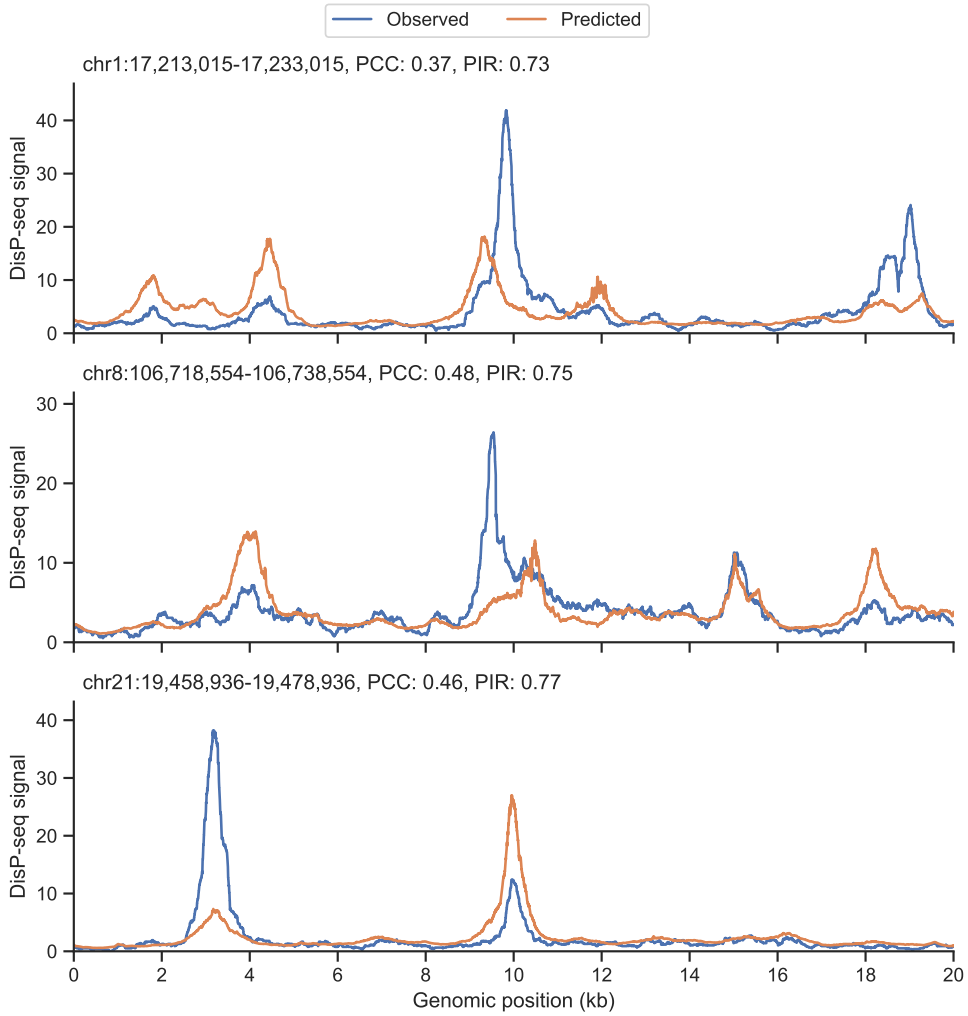

Figure 3.4: Observed and predicted DisP-seq profiles on held-out peaks, where U-DisCo achieved low PCCs but high peak intersection ratios. PIR, peak intersection ratio.

generalized effectively to H446, and reasonably well to MRC5, demonstrating its ability to learn cell type-specific patterns by incorporating ATAC-seq data. A genome browser view of sample predictions on SKNMC is shown in **Fig.** 3.3f, where U-DisCo's predictions achieved high PCCs and closely aligned with observed profiles. Incorporating ATAC-seq further improved prediction accuracy, particularly in regions with complex profile shapes.

Overall, U-DisCo demonstrated robust and reliable performance compared to state-of-the-art methods in predicting base-resolution genomic signals. This advancement paves the way for uncovering the sequence syntax underpinning IDP-DNA interactions.

Chapter 4

Identification of IDP binding motifs

Having trained U-DisCo on DisP-seq data from SKNMC cells and established its predictive power, we next sought to interpret the model's learned patterns and identify potential DNA determinants driving genome-wide IDP-DNA interactions. In this chapter, we describe the identification of IDP-binding motifs, highlight the functional importance of these motifs, and explore their potential roles in gene regulation.

4.1 Model interpretation and importance scores

To interpret U-DisCo's predictions, we used integrated gradients with SHapley Additive exPlanations (SHAP) [45] to assign importance scores to each nucleotide (A, C, G, T) in the input sequence. For each 20 kb sequence, a reference set of 100 randomly shuffled versions of the input sequence was generated, preserving dinucleotide frequencies as recommended in [46]. For control tracks, the reference set consisted of 100 copies of them. 200 samples were drawn for each interpretation.

We processed the model output by subtracting the mean and converting it to log-probability space. Following prior works [41], this output was then weighted by post-softmax probabilities (detached from the computation graph) to ensure that high-probability positions received exponentially higher weights, and vice versa. The resulting scalar value, obtained by summing across all positions, was then explained to generate importance scores for each base. This approach allowed us to quantify the contribution of each genomic base to the model's predictions, identifying motif patterns most influential in determining the DisP-seq profiles.

4.2 Motif identification

Normalized importance scores (mean-subtracted along the one-hot encoded dimensions) were used with TF-MoDISco [47] for motif discovery, which identifies recurring patterns with high importance scores. TF-MoDISco was configured with a maximum of 1,000,000 seqlets and a 20,000 bp window. Identified motifs were compared against known motifs from the MEME Suite to validate their biological significance.

4.3 Chromatin state analysis

To understand the regulatory roles of DisP-seq peak regions, we annotated each genomic bin with one of five chromatin states (transcription start site, active enhancer, weak enhancer, transcribed region, and quiescent region) using ChromHMM v1.25 [48]. Histone modification ChIP-seq datasets (H3K27ac, H3K9ac, H3K4me1, H3K9me3, H3K4me3, and H3K36me3) for SKNMC cells were downloaded from ENCODE and used as inputs to ChromHMM. The dominant chromatin state for each peak was determined by overlapping DisP-seq peaks with ChromHMM-defined states using bedtools v2.31.0 [49].

4.4 State association analysis

To assess the association of motifs with DisP island states or chromatin states, odds ratios were calculated by comparing the observed count of a particular state to the expected count after shuffling motif instances across peaks. A contingency matrix was constructed as:

$$\mathbf{C} = \begin{bmatrix} \text{peak count w/ motif w/o state (shuffled)} & \text{peak count w/ motif w/o state} \\ \text{peak count w/ motif and state (shuffled)} & \text{peak count w/ motif and state} \end{bmatrix}, \tag{4.1}$$

and the odds ratio was calculated as:

$$\text{odds ratio} = \frac{\text{peak count w/ motif and state} \times \text{peak count w/ motif w/o state (shuffled)}}{\text{peak count w/ motif w/o state} \times \text{peak count w/ motif and state (shuffled)}}. \tag{4.2}$$

An odds ratio greater than 1 indicates a significant association with the state, while a value less than 1 suggests a negative association. Pearson's chi-squared test was applied to the contingency matrix to determine the statistical significance of the association.

4.5 Results: IDP-binding motifs and their regulatory roles

By applying SHAP and TF-MoDISco, we assigned importance scores to DNA sequences and uncovered recurring motif patterns with high importance scores. These motifs, predictive of IDP activity, are biologically meaningful and less prone to false positives than traditional de novo approaches like HOMER [50], which rely on statistically over-represented sequences matched to position weight matrices (PWMs) [33, 36].

Key motifs identified include AP-2, NFI, EWS-FLI1 (single GGAA and GGAA repeat), C/EBP, TWIST, POU, and HOX, listed in decreasing order of importance (**Fig.** 4.1a). Notably, the disordered fusion protein EWS-FLI1 – a driver of Ewing sarcoma (the second most common pediatric bone cancer in SKNMC cells [51, 52]), was identified. HOMER analysis conducted in the DisP-seq assay [29] reported the top identified motifs: AP-2, NFI, and both EWS-FLI1 motifs (single GGAA and GGAA repeat), listed in increasing order of p-value, suggesting that U-DisCo, while focusing on the predictive power of DNA sequences, also successfully captured statistically meaningful motifs. Moreover, their importance scores aligned with their statistical significance (p-values), further corroborating our results. EWS-FLI1 has been

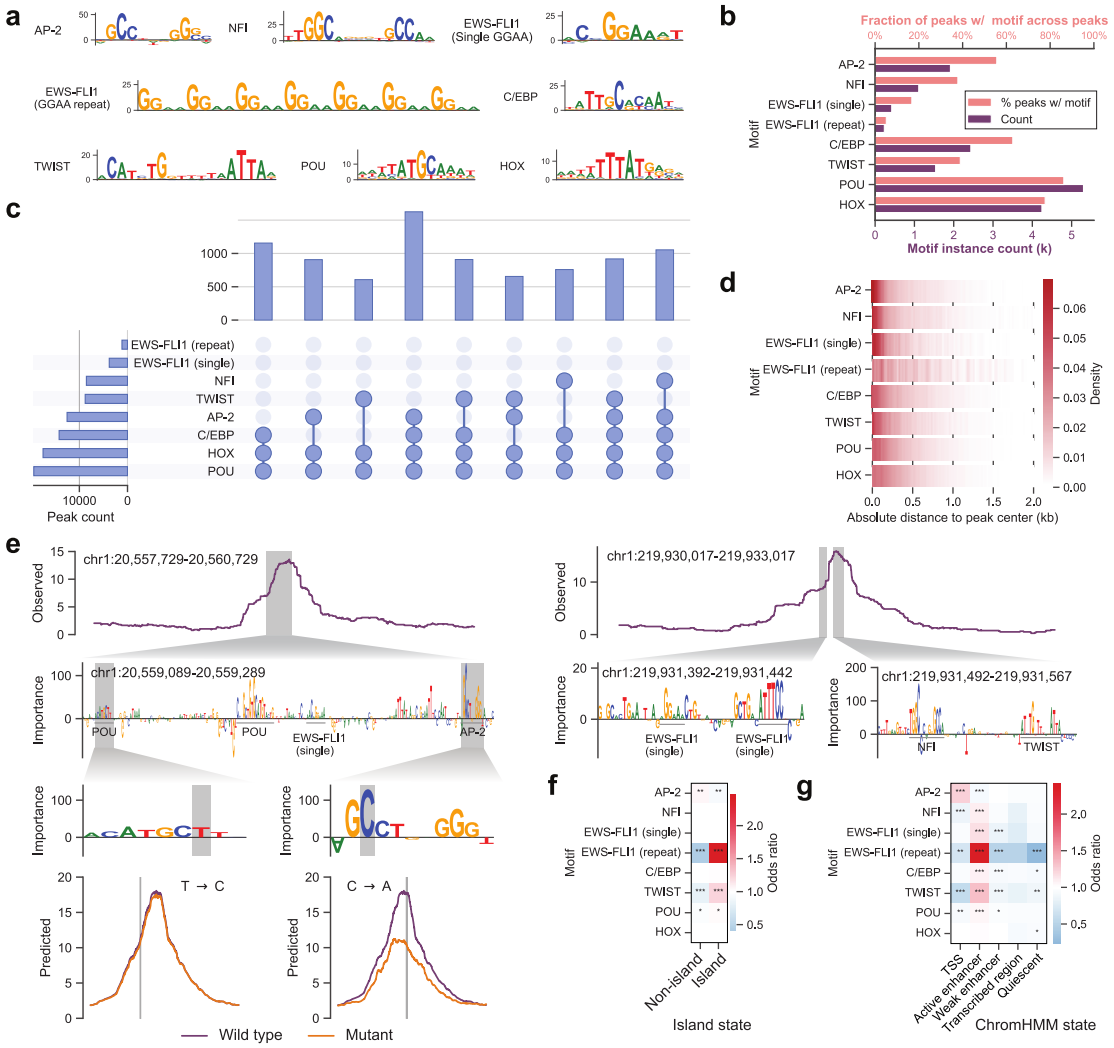

Figure 4.1: Identification of IDP-binding motifs driving genome-wide interactions. **a.** Top motifs with the highest importance scores, where the height of each letter in the sequence logos reflects its importance. **b.** Percentage of peaks containing each motif and their total counts. **c.** Upset plot showing motif combinations within peaks, with low-frequency combinations filtered out. **d.** Distribution of motif instances relative to peak centers. **e.** Example motif instances on chr1, showing single nucleotide mutagenesis effects on model predictions. **f.** Odds ratios of DisP island states for each motif, where values above 1 indicate significant association, and below 1 indicate negative association. **g.** Odds ratios for chromatin states, labeled by ChromHMM, associated with each motif, with the same interpretation as in panel **f**. TSS, transcription start site. $***p < 0.001$, $**p < 0.01$, $*p < 0.05$ (chi-squared test).

experimentally validated as an intrinsically disordered protein [51, 52], while the other proteins were predicted to contain significant intrinsically disordered regions (IDRs) by PONDR [53] and metapredict V2 [54] (see **Fig.** 4.2 and 4.3).

We calculated the percentage of peaks containing each motif, along with the total number of motif instances (**Fig.** 4.1b). AP-2, NFI, and EWS-FLI1, despite their high importance, were

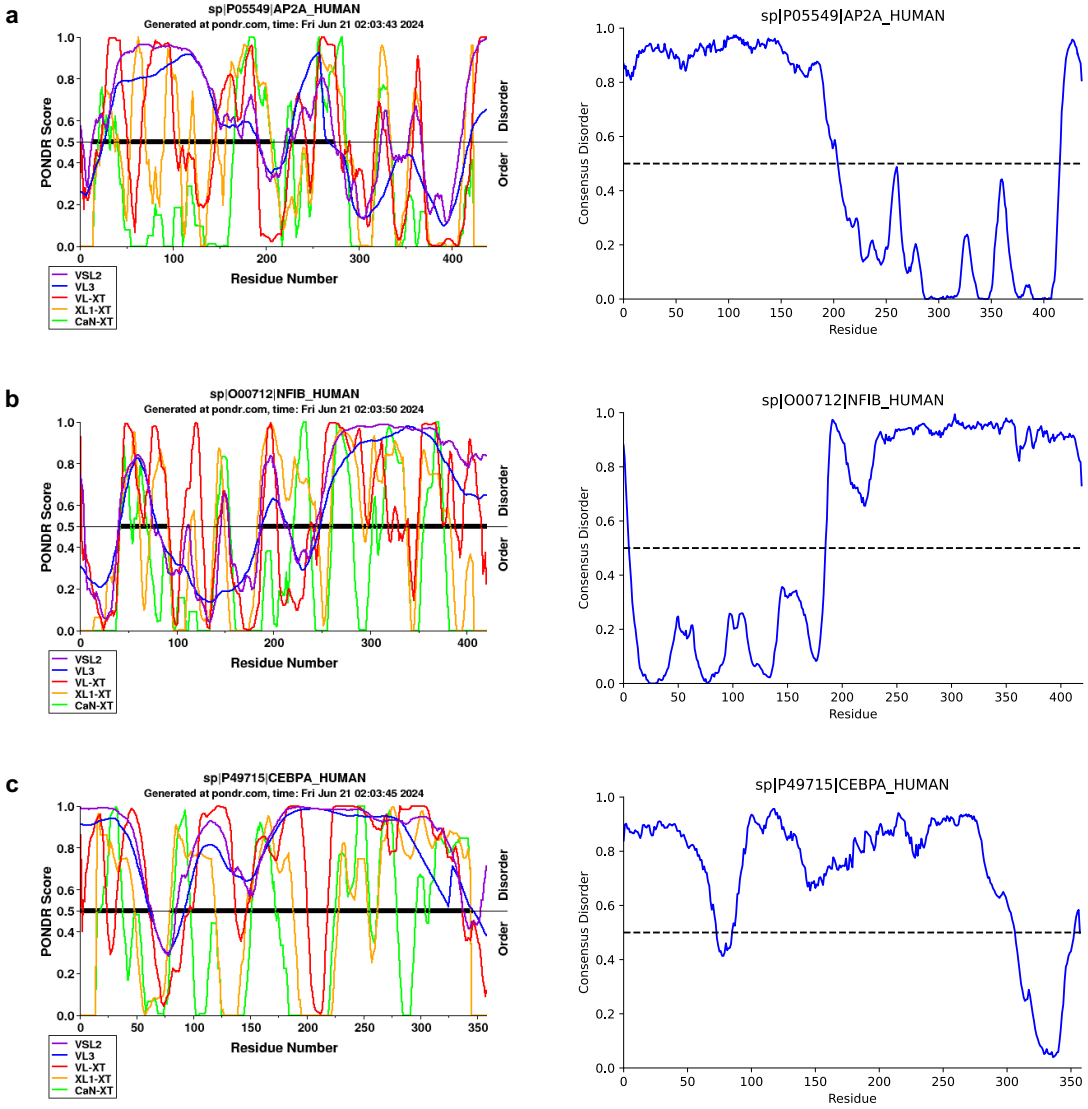

Figure 4.2: PONDR and metapredict V2 IDR predictions for proteins identified by U-DisCo. One representative protein isoform was selected for prediction in each case. **A**, AP-2. **B**, NFI. **C**, C/EBP.

present in only a fraction of peaks, suggesting their roles in specific regulatory contexts. In contrast, C/EBP, POU, and HOX were more prevalent, indicating broader regulatory functions. Motif combinations within peaks were visualized using an upset plot (**Fig.** 4.1c), where low-frequency combinations were filtered out. AP-2, NFI, and TWIST frequently co-occurred with C/EBP, POU, and HOX, whereas EWS-FLI1 motifs were less common. The distribution of motif instances relative to peak centers (**Fig.** 4.1d) showed that AP-2, NFI, and EWS-FLI1 (single GGAA) had the highest concentrations, while EWS-FLI1 (GGAA repeat) motifs were more dispersed, suggesting a role in distal regulation, potentially as enhancers.

Fig. 4.1e shows the identified motif instances under two example peaks on chr1, where many

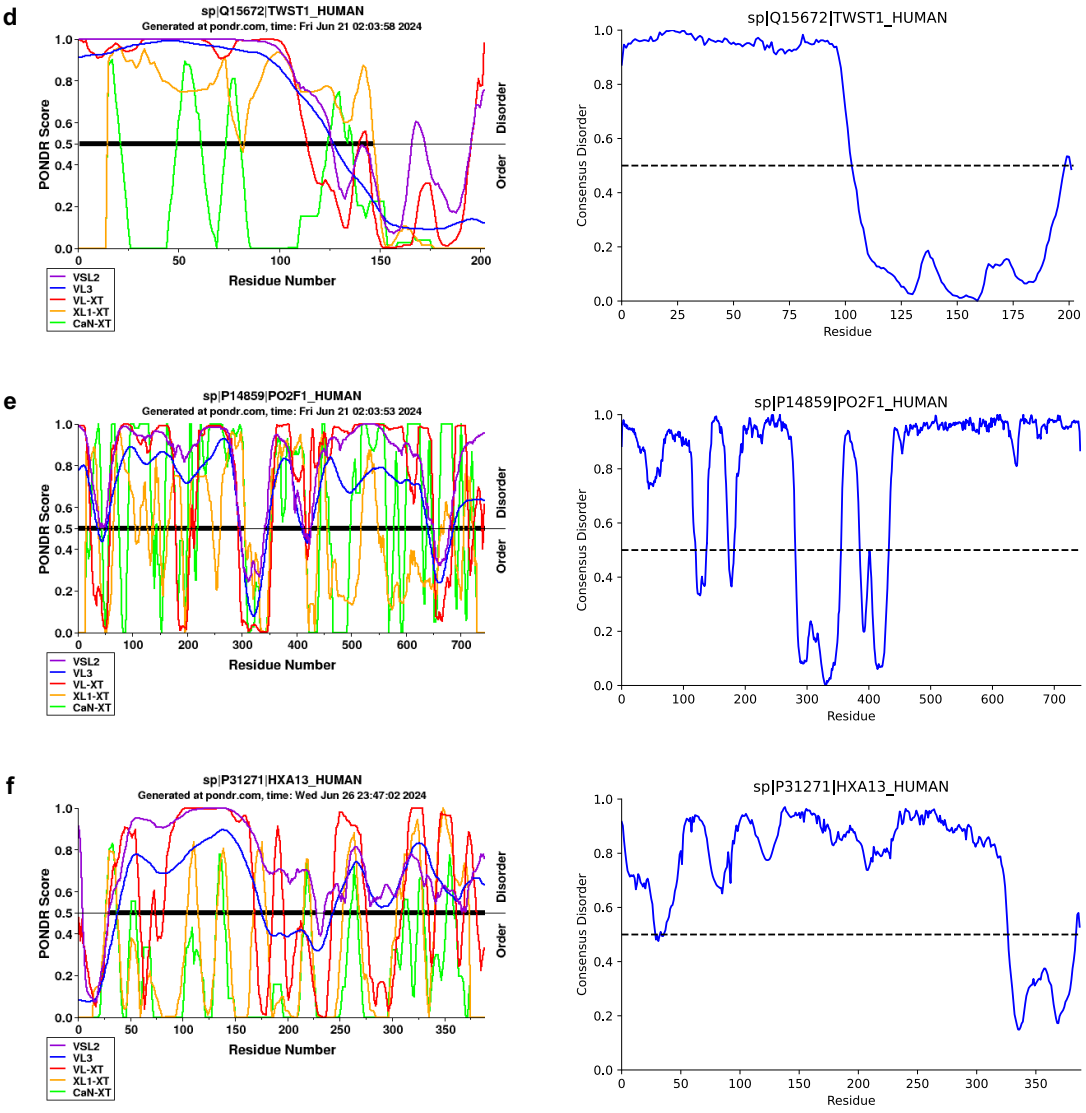

Figure 4.3: Continuation of PONDR and metapredict V2 IDR predictions for proteins identified by U-DisCo. One representative protein isoform was selected for prediction in each case. **D**, TWIST. **E**, POU. **F**, HOX.

instances are clustered near peak summits. To further investigate their importance, we performed single nucleotide mutagenesis experiments. For instance, mutating C to A at a high-importance site within an AP-2 instance resulted in a noticeable decrease in the predicted signals. In contrast, mutating a less important nucleotide T to C within a POU instance had minimal impact on the predicted profile.

To assess motif associations with DisP-seq clusters, we calculated odds ratios for DisP island states (**Fig.** 4.1f). AP-2 was weakly associated with non-island regions, while EWS-FLI1 (GGAA repeat) was strongly associated with islands. TWIST also showed a weak association with islands, while other motifs displayed no significant relationships. Finally, we computed

odds ratios for chromatin states, labeled by ChromHMM, associated with each motif (**Fig.** 4.1g). AP-2 showed a weak association with transcription start sites (TSS). NFI, EWS-FLI1 (single GGAA), C/EBP, TWIST, and POU were weakly associated with active enhancers, while EWS-FLI1 (GGAA repeat) was strongly linked to active enhancers. This suggests that these motifs are involved in enhancer-mediated gene regulation, particularly EWS-FLI1 (GGAA repeat), consistent with its dispersed distribution (**Fig.** 4.1d).

Chapter 5

Cooperative intercations between IDPs

After identifying the binding motifs of IDPs in SKNMC cells, we aimed to understand how these motifs interact with each other. We conducted a co-occurrence analysis to assess whether motif pairs were co-occurring due to cooperative interactions or random chance. We also investigated the strand-specific preferences and spacing between motif instances to uncover subtle patterns. Our results revealed cooperative behaviors between IDPs and their binding preferences, providing fine-grained insights into the IDP-mediated gene regulation in SKNMC cells.

5.1 Co-occurrence analysis

To investigate the cooperative behaviors of IDP-binding motifs, we analyzed their co-occurrences to assess proximity and potential interactions. Four distance ranges between motif instances were considered: less than 150 bp, 150-300 bp, 300-500 bp, and greater than 500 bp. 100 shuffled versions of the motif instances were generated by randomly permuting their positions within chromosomes, preserving their chromosomal distribution. For each distance range, all motif pairs were evaluated using k-d trees [55] for efficiency. Counts of motif pairs falling inside and outside the range were obtained for both the original and shuffled motif instances, where the counts from the 100 shuffled versions were averaged. Following [33], a contingency matrix was constructed, where Pearson's chi-squared test was applied to determine the significance:

$$\mathbf{C} = \begin{bmatrix} \text{outside count (shuffled)} & \text{outside count} \\ \text{inside count (shuffled)} & \text{inside count} \end{bmatrix}, \tag{5.1}$$

and the odds ratio was calculated as:

$$\text{odds ratio} = \frac{\text{inside count} \times \text{outside count (shuffled)}}{\text{outside count} \times \text{inside count (shuffled)}}. \tag{5.2}$$

An odds ratio greater than 1 indicates significant co-occurrence, while a value less than 1 suggests fewer co-occurrences than expected by random chance.

5.2 Neighborhood density analysis

To investigate whether co-occurrence within 150 bp indicates cooperative interactions, we grouped motif instances by the number of neighboring instances of the paired motif within 150 bp. For each motif pair A and B, where A and B may be the same, instances of motif A were grouped by the number of neighboring B instances, using the following thresholds: isolated (less than 10% quantile), normal (10-90% quantile), and gregarious (greater than 90% quantile). Similarly, motif B was grouped by the number of neighboring A instances. Mann-Whitney U test was performed to assess the significance of differences in importance scores between adjacent groups (isolated vs. normal, normal vs. gregarious).

5.3 Strand-specific preference analysis

We extended the co-occurrence analysis to include strand-specific preferences and more granular distance ranges to uncover subtle patterns. Four strand orientation combinations for motif A and motif B were considered: (i) $\Rightarrow\rightarrow$ (A to B, parallel); (ii) $\rightarrow\Rightarrow$ (B to A, parallel); (iii) $\leftarrow\Rightarrow$ (tail-to-tail); and (iv) $\Rightarrow\leftarrow$ (head-to-head). Although there are eight possible orientations, symmetry allows them to be condensed into these four categories. When A and B refer to the same motif, categories (i) and (ii) collapse into one. The distance ranges were binned every 25 bp, extending from 0 up to 600 bp. For each distance range and strand orientation combination, odds ratios were calculated for motif pairs falling within the specified parameters, as described in Section 5.1, with error bars representing the standard deviation across 100 shuffled versions. Paired permutation test was conducted to assess the significance of strand orientation preferences, where the odds ratios for each orientation were paired with those from other orientations at the same distance, followed by shuffling across orientations. The variance of the mean odds ratios from each shuffled orientation was used as the test statistic.

5.4 Results

5.4.1 Cooperative behaviors between IDPs

The co-occurrence analysis revealed distinct patterns of potential cooperative interactions (**Fig.** 5.1a). Short-range interactions (less than 150 bp) showed strong co-occurrence between motifs such as AP-2, NFI, EWS-FLI1 (single GGAA), and TWIST, suggesting they may participate in larger regulatory complexes. Notably, AP-2 and EWS-FLI1 (single GGAA) exhibited homotypic interactions, indicating possible homotypic binding or cooperative formation of regulatory domains. As the distance increased to medium ranges (150-300 bp and 300-500 bp), the odds ratios for co-occurrence gradually declined. In contrast, long-range interactions (greater than 500 bp) revealed high odds ratios for EWS-FLI1 (GGAA repeat), suggesting that this high-importance motif operates independently at longer distances, potentially focusing on distal regulatory elements critical for gene expression programs in Ewing sarcoma. NFI also exhibited strong homotypic co-occurrence at these distances, though it showed less interaction with itself at shorter ranges.

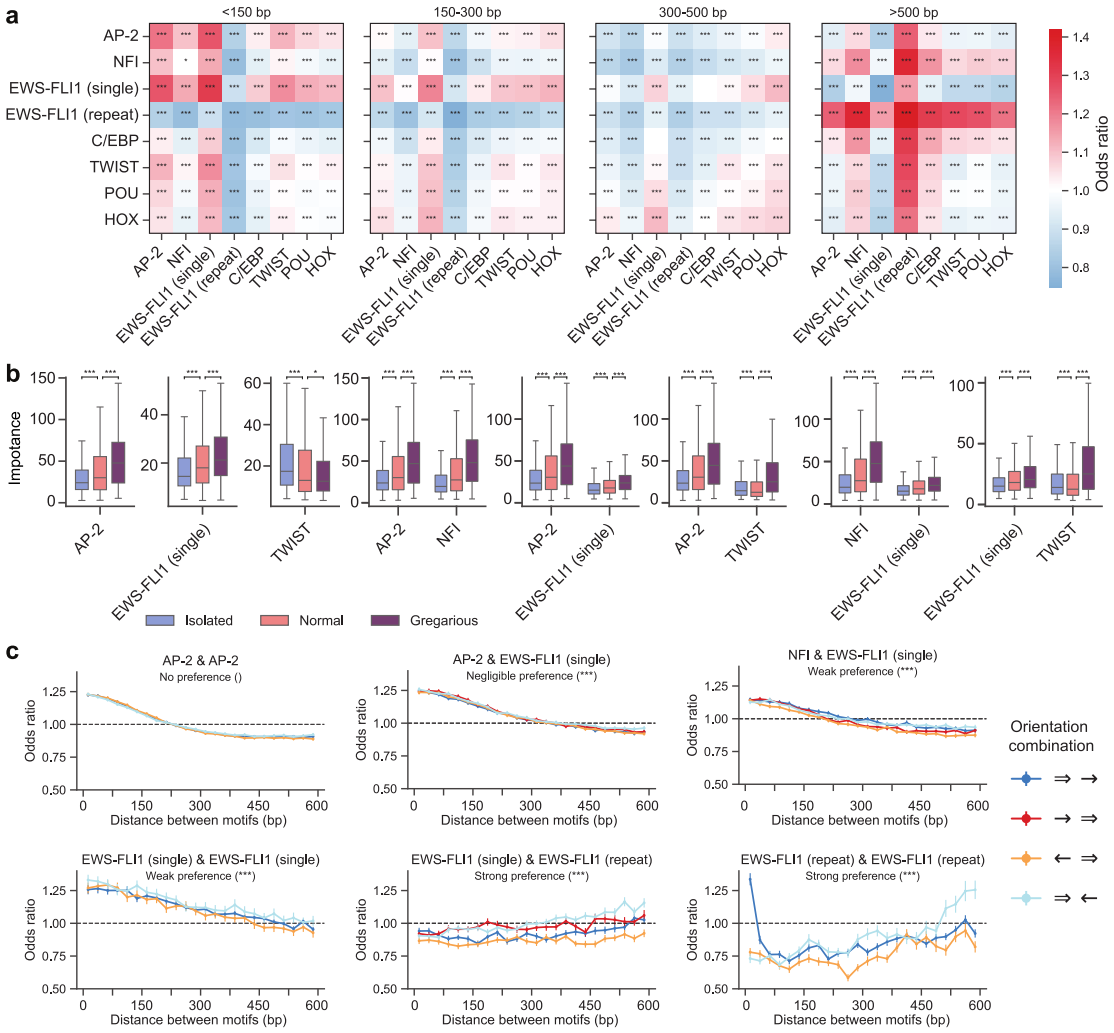

Figure 5.1: Co-occurrence analysis revealing cooperative behaviors between IDPs and their binding preferences. **a.** Odds ratios for motif pair co-occurrence across distance ranges. Odds ratios above 1 indicate co-occurrence, and below 1 suggest fewer co-occurrences than expected by chance. $***p < 0.001$, $**p < 0.01$, $*p < 0.05$ (chi-squared test). **b.** Importance scores for motif pairs grouped by the density of their paired motif instances within 150 bp. A boxplot with a single motif represents that motif paired with itself. Boxplots with two motifs represent those motifs paired together. $***p < 0.001$, $**p < 0.01$, $*p < 0.05$ (Mann-Whitney U test for significance in distribution differences between adjacent groups). **c.** Odds ratios for motif pair co-occurrence grouped by strand orientation across 25 bp distance bins. Interpretation of odds ratios follows panel **a**. Error bars represent the standard deviation across 100 simulations. $***p < 0.001$, $**p < 0.01$, $*p < 0.05$ (paired permutation test for significance of strand orientation preferences).

To confirm that co-occurrence within 150 bp indicates cooperative interactions rather than mere proximity, we analyzed the importance scores of motif pairs based on the density of neighboring motifs (**Fig.** 5.1b). Overall, motifs with more neighbors had significantly higher im-

portance scores, indicating cooperative behavior. One exception was TWIST, which showed decreasing importance scores with more TWIST neighbors, explaining its weak homotypic co-occurrence (**Fig.** 5.1a). These findings suggest that co-occurrence within 150 bp is generally indicative of cooperative interactions between motifs.

Importantly, while HOMER can identify sequence motifs under DisP-seq peaks for co-occurrence analysis, its statistical approach cannot reveal base-level contributions to DisP-seq signals, lacking fine-grained attribution for each motif and potentially leading to false positives. In contrast, U-DisCo provides simultaneous importance scores for all motifs, identifying instances specific to SKNMC and predictive of DisP-seq signals, enabling a biologically meaningful co-occurrence analysis.

5.4.2 IDP binding preferences for orientation and spacing

To further investigate motif co-occurrences, we incorporated strand orientation combinations and finer distance intervals (binned every 25 bp up to 600 bp) and calculated odds ratios for motif pair occurrences (**Fig.** 5.1c). This finer granularity revealed co-occurrence patterns that were not apparent previously. The results aligned with previous analyses (**Fig.** 5.1a), showing that co-occurring motif pairs generally had higher odds ratios at closer distances. Three types of orientation preference emerged: none, weak, and strong. AP-2 & AP-2, and AP-2 & EWS-FLI1 (single GGAA) displayed no or negligible preference for any orientation, suggesting flexible functionality. NFI & EWS-FLI1 (single GGAA), and EWS-FLI1 (single GGAA) & EWS-FLI1 (single GGAA) showed a weak preference for parallel and head-to-head orientations. In contrast, EWS-FLI1 (single GGAA) & EWS-FLI1 (GGAA repeat) demonstrated a strong, distinct preference for different orientations depending on the distance. At distances greater than 350 bp, this pair favored head-to-head orientations, suggesting potential mechanisms for long-range interactions. EWS-FLI1 (GGAA repeat) & EWS-FLI1 (GGAA repeat) showed a strong preference for parallel orientations within 50 bp, indicating a tendency to cluster and potentially form extended GGAA sequences. At distances beyond 500 bp, their head-to-head orientation preference hints at the formation of higher-order structures.

Chapter 6

Conclusion

This work presents the first computational approach to explore and quantify multiple IDPs within a single cell type, rather than in isolation. U-DisCo, our deep learning model, achieved high accuracy and computational efficiency in predicting base-resolution IDP-DNA binding profiles from DNA sequences and demonstrated robust generalization across cell lines. Beyond identifying key IDP-binding motifs, U-DisCo revealed functional syntax, cooperative behaviors, and binding preferences of IDPs in SKNMC cells. This approach can be adapted to other cell types and protein-DNA binding profiles to investigate regulatory mechanisms across diverse biological contexts.

IDPs play a central role in complex regulatory networks, making it crucial to understand their interactions with DNA. Although previous sequencing efforts have explored IDP-DNA interactions, these are limited to one protein type at a time due to reliance on targeted antibodies. The antibody-independent DisP-seq assay overcomes this limitation by mapping multiple disordered proteins simultaneously. However, no prior computational approaches could investigate IDP-DNA interactions directly from DNA sequence data. Frequency-based methods like k-mer [56, 57] and dictionary-based algorithms [58, 59] often miss low-frequency motifs, such as GGAA repeats, or overestimate highly frequent motifs like HOX and POU. Deep learning has been applied to protein-DNA binding predictions, but achieving base-resolution interpretability for IDPs has remained challenging. U-DisCo addresses this gap by combining deep learning with DisP-seq data, using a U-Net architecture to analyze long-range genomic contexts while preserving interpretability.

Despite these advances, U-DisCo currently relies primarily on DNA sequence data. Future iterations could integrate protein structural information predicted by methods like RoseTTAFold [60], enabling deeper exploration of the fine-tuned binding mechanisms at the protein-DNA interface. While U-DisCo is computationally efficient in analyzing 20 kb input DNA sequences, it may face limitations when handling distal interactions spanning several megabases. Transformer-based methods, known for handling long-range dependencies, could complement U-DisCo in exploring such large-scale interactions, broadening the scope of IDP-DNA dynamics across regulatory domains. The lower performance of U-DisCo in cross-cell line evaluations on MRC5 likely stems from using DNase-seq data from IMR90 cells as a proxy, due to the absence of MRC5-specific ATAC-seq data. However, this reduced performance may also indicate inherent cell type-specific differences in IDP-DNA interactions, a topic for further

study. Additionally, the motif syntax identified in this study may not be exhaustive. Further investigations could reveal additional motifs and their regulatory roles. Experimental validation of motif interactions is essential for a comprehensive understanding of the regulatory networks mediated by IDPs. Overall, these findings lay a foundation for future research into genome-wide regulatory mechanisms, enhancing our understanding of IDP-mediated gene regulation.

Bibliography

[1] Ardito, F., Giuliani, M., Perrone, D., Troiano, G. & Lo Muzio, L. The crucial role of protein phosphorylation in cell signaling and its use as targeted therapy. *International journal of molecular medicine* **40**, 271–280 (2017). 1.1

[2] Yang, X. & Qian, K. Protein o-glcnacylation: emerging mechanisms and functions. *Nature reviews Molecular cell biology* **18**, 452–465 (2017).

[3] Hetz, C., Zhang, K. & Kaufman, R. J. Mechanisms, regulation and functions of the unfolded protein response. *Nature reviews Molecular cell biology* **21**, 421–438 (2020). 1.1

[4] Anfinsen, C. B. Principles that govern the folding of protein chains. *Science* **181**, 223–230 (1973). 1.1

[5] Tompa, P. The interplay between structure and function in intrinsically unstructured proteins. *FEBS letters* **579**, 3346–3354 (2005).

[6] Redfern, O. C., Dessailly, B. & Orengo, C. A. Exploring the structure and function paradigm. *Current opinion in structural biology* **18**, 394–402 (2008).

[7] Papoian, G. A. Proteins with weakly funneled energy landscapes challenge the classical structure–function paradigm. *Proceedings of the National Academy of Sciences* **105**, 14237–14238 (2008).

[8] Pearce, R. & Zhang, Y. Toward the solution of the protein structure prediction problem. *Journal of Biological Chemistry* **297** (2021). 1.1

[9] Oldfield, C. J. & Dunker, A. K. Intrinsically disordered proteins and intrinsically disordered protein regions. *Annual review of biochemistry* **83**, 553–584 (2014). 1.1

[10] Piovesan, D. *et al.* Mobidb: 10 years of intrinsically disordered proteins. *Nucleic acids research* **51**, D438–D444 (2023). 1.1

[11] Habchi, J., Tompa, P., Longhi, S. & Uversky, V. N. Introducing protein intrinsic disorder. *Chemical reviews* **114**, 6561–6588 (2014). 1.1

[12] Wright, P. E. & Dyson, H. J. Intrinsically disordered proteins in cellular signalling and regulation. *Nature reviews Molecular cell biology* **16**, 18–29 (2015). 1.2

[13] Calabretta, S. & Richard, S. Emerging roles of disordered sequences in rna-binding proteins. *Trends in biochemical sciences* **40**, 662–672 (2015).

[14] Ozdilek, B. A. *et al.* Intrinsically disordered rgg/rg domains mediate degenerate specificity in rna binding. *Nucleic acids research* **45**, 7984–7996 (2017).

[15] Staller, M. V. *et al.* A high-throughput mutational scan of an intrinsically disordered acidic transcriptional activation domain. *Cell systems* **6**, 444–455 (2018).

[16] Brodsky, S. *et al.* Intrinsically disordered regions direct transcription factor in vivo binding specificity. *Molecular cell* **79**, 459–471 (2020).

[17] Bondos, S. E., Dunker, A. K. & Uversky, V. N. Intrinsically disordered proteins play diverse roles in cell signaling. *Cell Communication and Signaling* **20**, 20 (2022). 1.1

[18] Uversky, V. N. A protein-chameleon: conformational plasticity of α-synuclein, a disordered protein involved in neurodegenerative disorders. *Journal of Biomolecular Structure and Dynamics* **21**, 211–234 (2003). 1.1

[19] Menon, S. & Mondal, J. Conformational plasticity in α-synuclein and how crowded environment modulates it. *The Journal of Physical Chemistry B* **127**, 4032–4049 (2023). 1.1

[20] Malaney, P., Pathak, R. R., Xue, B., Uversky, V. N. & Davé, V. Intrinsic disorder in pten and its interactome confers structural plasticity and functional versatility. *Scientific reports* **3**, 2035 (2013). 1.1

[21] Uversky, V. N. Intrinsic disorder, protein–protein interactions, and disease. *Advances in protein chemistry and structural biology* **110**, 85–121 (2018). 1.1

[22] Sridhar, A., Orozco, M. & Collepardo-Guevara, R. Protein disorder-to-order transition enhances the nucleosome-binding affinity of h1. *Nucleic acids research* **48**, 5318–5331 (2020). 1.1

[23] Park, P. J. Chip–seq: advantages and challenges of a maturing technology. *Nature reviews genetics* **10**, 669–680 (2009). 1.2

[24] Kasinathan, S., Orsi, G. A., Zentner, G. E., Ahmad, K. & Henikoff, S. High-resolution mapping of transcription factor binding sites on native chromatin. *Nature methods* **11**, 203–209 (2014).

[25] Skene, P. J. & Henikoff, S. An efficient targeted nuclease strategy for high-resolution mapping of dna binding sites. *elife* **6**, e21856 (2017). 1.2

[26] Dunker, A. K., Brown, C. J., Lawson, J. D., Iakoucheva, L. M. & Obradović, Z. Intrinsic disorder and protein function. *Biochemistry* **41**, 6573–6582 (2002). 1.2

[27] Uversky, V. N. Unusual biophysics of intrinsically disordered proteins. *Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics* **1834**, 932–951 (2013). 1.2

[28] Kwong, P. D. *et al.* Hiv-1 evades antibody-mediated neutralization through conformational masking of receptor-binding sites. *Nature* **420**, 678–682 (2002). 1.2

[29] Xing, Y.-H. *et al.* Disp-seq reveals the genome-wide functional organization of dna-associated disordered proteins. *Nature Biotechnology* **42**, 52–64 (2024). 1.2, 3.2, 3.6, 4.5

[30] 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**, 831–838 (2015). 1.3, 2.2.1, 2.3.1

[31] 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**, 990–999 (2016). 1.3, 2.2.1, 2.3.1

[32] Kim, D. S. *et al.* The dynamic, combinatorial cis-regulatory lexicon of epidermal differentiation. *Nature genetics* **53**, 1564–1576 (2021).

[33] Avsec, Ž. *et al.* Base-resolution models of transcription-factor binding reveal soft motif syntax. *Nature genetics* **53**, 354–366 (2021). 1.3, 2.2.1, 2.3.2, 3.1, 3.2, 3.4, 3.6, 4.5, 5.1

[34] Avsec, Ž. *et al.* Effective gene expression prediction from sequence by integrating long-range interactions. *Nature methods* **18**, 1196–1203 (2021). 1.3, 2.2.2, 2.3.2

[35] Janssens, J. *et al.* Decoding gene regulation in the fly brain. *Nature* **601**, 630–636 (2022). 1.3, 2.3.1

[36] de Almeida, B. P., Reiter, F., Pagani, M. & Stark, A. Deepstarr predicts enhancer activity from dna sequence and enables the de novo design of synthetic enhancers. *Nature genetics* **54**, 613–624 (2022). 1.3, 2.3.1, 4.5

[37] Vaswani, A. *et al.* Attention is all you need. *Advances in neural information processing systems* **30** (2017). 1.3, 2.2.2

[38] 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). 2.1

[39] Hooghe, B., Broos, S., Van Roy, F. & De Bleser, P. A flexible integrative approach based on random forest improves prediction of transcription factor binding sites. *Nucleic acids research* **40**, e106–e106 (2012). 2.1

[40] Linder, J., Srivastava, D., Yuan, H., Agarwal, V. & Kelley, D. R. Predicting rna-seq coverage from dna sequence as a unifying model of gene regulation. *Biorxiv* 2023–08 (2023). 2.2.2, 2.3.2

[41] Tseng, A., Shrikumar, A. & Kundaje, A. Fourier-transform-based attribution priors improve the interpretability and stability of deep learning models for genomics. *Advances in Neural Information Processing Systems* **33**, 1913–1923 (2020). 2.3.1, 3.1, 3.3, 4.1

[42] Kagda, M. S. *et al.* Data navigation on the encode portal. *arXiv preprint arXiv:2305.00006* (2023). 3.2

[43] Ronneberger, O., Fischer, P. & Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, 234–241 (Springer, 2015). 3.3

[44] Ke, G. *et al.* Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems* **30** (2017). 3.4

[45] Lundberg, S. M. & Lee, S.-I. A unified approach to interpreting model predictions. *Advances in neural information processing systems* **30** (2017). 4.1

[46] Shrikumar, A., Greenside, P. & Kundaje, A. Learning important features through propa-

gating activation differences. In *International conference on machine learning*, 3145–3153 (PMlR, 2017). 4.1

[47] Shrikumar, A. *et al.* Technical note on transcription factor motif discovery from importance scores (tf-modisco) version 0.5. 6.5. *arXiv preprint arXiv:1811.00416* (2018). 4.2

[48] Ernst, J. & Kellis, M. Chromhmm: automating chromatin-state discovery and characterization. *Nature methods* **9**, 215–216 (2012). 4.3

[49] Quinlan, A. R. & Hall, I. M. Bedtools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010). 4.3

[50] Heinz, S. *et al.* Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and b cell identities. *Molecular cell* **38**, 576–589 (2010). 4.5

[51] Grünewald, T. G. *et al.* Ewing sarcoma. *Nature reviews Disease primers* **4**, 5 (2018). 4.5

[52] Riggi, N., Suvà, M. L. & Stamenkovic, I. Ewing's sarcoma. *New England Journal of Medicine* **384**, 154–164 (2021). 4.5

[53] Peng, K., Radivojac, P., Vucetic, S., Dunker, A. K. & Obradovic, Z. Length-dependent prediction of protein intrinsic disorder. *BMC bioinformatics* **7**, 1–17 (2006). 4.5

[54] Emenecker, R., Griffith, D. & Holehouse, A. Metapredict v2: An update to metapredict, a fast, accurate, and easy-to-use predictor of consensus disorder and structure. biorxiv 2022.06. 06.494887 (2022). 4.5

[55] Maneewongvatana, S. & Mount, D. M. Analysis of approximate nearest neighbor searching with clustered point sets. *arXiv preprint cs/9901013* (1999). 5.1

[56] Guo, Y., Tian, K., Zeng, H., Guo, X. & Gifford, D. K. A novel k-mer set memory (ksm) motif representation improves regulatory variant prediction. *Genome research* **28**, 891–900 (2018). 6

[57] Kirk, J. M. *et al.* Functional classification of long non-coding rnas by k-mer content. *Nature genetics* **50**, 1474–1482 (2018). 6

[58] Wang, G., Yu, T. & Zhang, W. Wordspy: identifying transcription factor binding motifs by building a dictionary and learning a grammar. *Nucleic acids research* **33**, W412–W416 (2005). 6

[59] Sharov, A. A. & Ko, M. S. Exhaustive search for over-represented dna sequence motifs with cisfinder. *DNA research* **16**, 261–273 (2009). 6

[60] Baek, M. *et al.* Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **373**, 871–876 (2021). 6