Mechanisms of Genome Organization across Scales: From Chromosomal Communication to DNA–Protein and Drug Interactions

A Dissertation Presented for the
Doctor of Philosophy
Degree
The University of Tennessee, Knoxville

Yingjie Xu
May 2026

Copyright © 2026 by Yingjie Xu.
All rights reserved.

DEDICATION

To my parents, who taught me the value of curiosity and perseverance, and who have always supported my choices.

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to my advisor, Dr. Tongye Shen, for his invaluable guidance, encouragement, and patience throughout my doctoral journey. His mentorship has shaped my scientific thinking in profound ways, teaching me how to approach complex problems with clarity, creativity, and rigor. I am especially grateful for his steady support during challenging moments in my research, and for the confidence he placed in me. His commitment to high standards, his thoughtful feedback, and his genuine care for my academic growth have made this dissertation possible and have deeply influenced the way I will conduct science in the future.

I am also sincerely grateful to the members of my dissertation committee, Dr. Albrecht von Arnim, Dr. Jaan Mannik, and Dr. Rachel Patton McCord, for their time, constructive feedback, and insightful discussions. I appreciate Dr. von Arnim, who first introduced me to the GST program and consistently supported my academic development, even including helping arrange my TA position and graduate funding. I am also deeply appreciative of Dr. McCord for her expertise in molecular biology and cancer research, whose guidance has shaped the biological direction of my work. I also thank Dr. Mannik for his perspective on physical modeling and quantitative approaches, which strengthened the rigor and conceptual foundation of my structural analyses. Their combined expertise has broadened my perspective and substantially strengthened this dissertation.

Special thanks go to both the former and current members of our laboratory, including David Foutch, Michael Harnish, and Joy Fasae, whose collaboration, technical support, and camaraderie made the research process both productive and enjoyable.

I am further indebted to the University of Tennessee and the Bredesen Center staff, whose tireless and selfless assistance at every stage of this process provided the foundation that made my work possible.

Finally, I extend my heartfelt appreciation to my family and friends for their unwavering support and encouragement, which gave me the strength to persevere through challenges and to celebrate achievements along the way.

ABSTRACT

The three-dimensional organization of the genome is central to gene regulation and cellular identity. To explore how chromatin architecture, protein–DNA interactions, and small molecules collectively influence nuclear function, I carried out four complementary studies spanning genome to molecular scales using human cell systems. First, I developed chromosome structure networks (CSN) by representing Hi-C contact matrices as graphs, with chromosomal segments as nodes and contacts as edges. Local clustering coefficients (LCC3, LCC4) more effectively captured structural and functional features than global centrality, distinguishing wild-type, condensin II mutant, and inverted nuclei, and correlating with lamina-associated domains (LADs) and chromatin states. Second, I extended this framework to interchromosomal communication (CSNI). New measurements (ΔL3, ΔL4, L4E) quantified cross-chromosome connectivity and revealed selective interactions among chromosomes 17, 19, and 22, with patterns differing across genetic perturbations. These measures negatively correlated with LADs and heterochromatic state while showing positive associations with transcriptionally active states. Third, I examined the hinge region of thyroid hormone receptor α (TRα) using long-timescale molecular dynamics and machine learning–based contact analysis. DNA binding induced a helix-to-coil conformational transition in the hinge, enabling minor-groove interactions and DNA bending, and promoting a closed-to-open conformational transition of the receptor that underlies hinge-mediated allosteric regulation. Finally, I studied curaxins, low-genotoxic anticancer drugs, in complex with DNA. Docking and simulations revealed preferential binding to the major groove of bent B-DNA, with active analogs stabilized by polar arms that form base-specific hydrogen bonds. These findings suggest that curaxins, through DNA shape selectivity, may modulate transcription factor activity, including the chromatin remodeling complex Facilitates Chromatin Transcription (FACT) and the architectural transcription factor CCCTC-binding factor (CTCF). Together, these studies integrate structural modeling, statistical analysis, and molecular simulations to provide a multiscale perspective on genome organization and regulation.

TABLE OF CONTENTS

Introduction 1
The Structural Basis of Biological Regulation Across Scales 1
Three-Dimensional Genome Structure 3
Chromosome Territories and Large-Scale Nuclear Organization 4
Compartments, Domains, and Local Chromatin Structure 4
Experimental Approaches for Studying Genome Architecture 5
Interchromosomal Interactions and Nuclear Communication 6
Chromatin as a Network: A Graph-Based Perspective 7
Motivation for Chromosome Structure Network Study 8
Protein–DNA Interaction and Structural Dynamics of Transcription Factors 8
Structure of Transcription Factors 9
Intrinsically Disordered Regions and Their Role in DNA Recognition 9
Molecular Dynamics Simulation of Protein–DNA Complex 10
Machine-Learning Approach for Analyzing Conformational Ensembles 11
Motivation for the Machine-Learning Analysis of TRα Dynamics 11
Mechanistic Insights into Curaxin–DNA Interaction 12
DNA Structural Features Relevant to Small-Molecule Interactions 12
Historical Context of DNA-Targeting Anticancer Agents 13
Curaxins as Low-Genotoxicity Anticancer Agents 13
Structural Consequences of Curaxin Binding 15
Knowledge Gaps and Motivation 16
REFERENCES 18
CHAPTER I Node Features of Chromosome Structure Networks and Their Connections to Genome Annotation 27
Abstract 28
Introduction 29
Methods and Systems 34
Contact network construction from chromosome structure information 34
Network properties 36

Systems and data 39
Results and Discussion 42
Effects of coverage and resolution on chromosome structure network properties ... 42
Network properties display distinct structural features of chromosomes for different cell types 45
The connection between network properties and genomic annotations 50
REFERENCES 55
APPENDIX 61
CHAPTER II Visualizing Interchromosomal Interactions at Sub-Megabasepair Resolution Using Network Clustering Coefficients 69
Abstract: 70
Introduction 71
Method and systems 76
Results 80
Construction of networks using Hi-C contact maps. 80
Assessment of node-based metrics for interchromosomal interaction strength 81
Annotation of LAD and the association of chromosome and nuclear lamina. 88
Effects of condensin II deletion 89
Concluding remarks 90
REFERENCES 92
APPENDIX 98
CHAPTER III Conformational Transition of a Polycationic Hinge Domain contributes to DNA Binding 104
Abstract 105
Introduction 105
Systems and Methods 109
Modeling the protein-DNA complex 109
Atomistic molecular dynamics simulation 112
Statistical analysis of conformational dynamics 113
Results and Discussion 115

The hinge–DNA interaction assists DBD–DNA association..... 115
Allostery and interdomain interaction modulated by DNA binding..... 117
Helix-to-unwound transition of hinge segments upon DNA binding..... 120
Effects of VAE parameters and sampling on Latent Representations of Hinge Dynamics 124
Concluding Remarks..... 129
REFERENCES 130
APPENDIX..... 139
CHAPTER IV Computational Assessment of the association between curaxin and selected DNA structures in genome..... 141
Abstract:..... 142
Introduction..... 142
Methods and systems:..... 147
Results and Discussion 149
Molecular docking 149
Molecular dynamics simulation and binding energy calculation 151
Genomic Analysis of Curaxin-Mediated Modulation of CTCF Binding 152
REFERENCES 157
Conclusion overall 188
VITA..... 191

LIST OF TABLES

Table S2. 1 Chromatin state coverage and correlation with interaction strength on chromosome 19. For each chromatin state (CS1–CS15), the table shows the fraction of bins with non-zero appearance and the Spearman correlation (ρ) between Q and the consensus of communication metrics (PC1 of ΔC4 vs. C4E)..... 98

LIST OF FIGURES

Figure I. 1 (a) Cytotoxicity of curaxins to cultured human diploid fibroblasts (Wi38) and fibrosarcoma (HT1080) cells (mean of three replicates ± SD). (b) Regrowth of RCC45 and NKE-hTERT cells after curaxin treatment for 3 hours. Methylene blue–stained cells were counted 5 days later. Adapted from ref. 17. 14

Figure 1. 1 (a) Hi-C contact map of IMR90 (lung fibroblasts cell) chr21 (region 15.25–48.25 Mb) at 50 kb resolution (b) The corresponding network adjacency matrix representation with a cutoff of 40% coverage (c) The corresponding force-directed network graph representation of panel (b). Numbers within nodes indicate the corresponding genomic bin indices, edges represent contacts, and node color encodes betweenness centrality (BC). (d) Network feature (LCC3) output (red) and LAD annotation (gray) as functions of node (genomic bin) index. 30

Figure 1. 2 (a) Pedagogical schematic of 3 types of network properties. Red and orange highlights schematically indicate the nodes and edges, respectively, emphasized by each metric. The first subpanel shows path-based centrality measures, including betweenness centrality (BC) and closeness centrality (CC), which quantify how shortest paths traverse or terminate at a node. The second subpanel shows adjacency-based centrality measures, including degree centrality (DC) and eigenvector centrality (EC), which are determined by direct neighbor connectivity and global spectral properties, respectively. The third subpanel shows higher-order clustering measures, including the local triangle clustering coefficient (LCC3) and square clustering coefficient (LCC4), which quantify closed local motifs involving three or four nodes. (b) The coverage (normalized rank of contact) p^* shown as a function of contact strength a^*. Δp indicates the coverage margin. (c) Cumulative fraction curves for a label being present (black), absent (red), and reference (green). ROC is further defined based on associated concepts TPR, TNR, FPR, and FNR. . 32

Figure 1. 3 (a) LCC3 value as a function of node (genomic bin) index, under different coverage thresholds p^* (max, 80%, 60%, 40%, 20%) at 50 kb resolution for IMR90, chr21:15.25–48.25 Mb. (b) IMR90 chromosome 21 (same region as above) 40% p^* cutoff EC and LCC3 values at different resolution (50 kb vs 250 kb bin

sizes). (c) Comparison of different network properties including BC, CC, EC, LCC3 and LCC4 across the same IMR90 chr21 region as in a and b. Arrow indicates a notable region of divergent values between network properties at a given genomic location. 43
Figure 1. 4 Network properties (EC and LCC3/4) as functions of network node index for cancer cell type K562 are shown in (a) and for haploid cell HAP1 in (b). The corresponding results for lymphoblast GM12878 are shown as a comparison. Data of region 15.25–48.25 Mb for chr21 (50 kb resolution) is used. Constant coverage of 40% is used. (c) Same comparison as (b) but HAP1 coverage is set to 33.4%, so that $p*/pc$ is kept constant. 46
Figure 1. 5 The structures of chr11 of mouse rod photoreceptor cell wildtype (WT) and LBR mutant are characterized using network properties at a resolution of 200 kb. (a) Heatmap comparison. Above the diagonal is LBR -/- cell. Below the diagonal line is WT cell. (b) BC value comparison between two the above cells. (c) CC value comparison. (d) EC value comparison. (e) LCC3 value comparison. (f) LCC4 value comparison. 49
Figure 1. 6 LAD cumulative fraction (CF) plots of network properties and MI-PCA for a region of IMR90 chr21. Here classifier performance concepts using CFP (black), CFN (red), CFR (green dash) are shown for EC in (a), LCC4 in (b) and PCA in (c). (d) ROC curves of six different network properties for the LAD classification. 51
Figure 1. 7 Using network properties discern selected six chromatin state labels of chr21 GM12878 for LCC3 (a) and LCC4 (b). The corresponding distribution of network properties of all 15 different chromatin states are shown in (c) and (d). The resolution is 50 kb. 54
Figure S1. 1 (a–f) Scatter plots of IMR90 chr21 (15.25–48.25 Mb) at 50 kb resolution showing network properties versus MIPCA (A/B compartment). (g–h) Correlation between different network properties using the same system as above. (i–j) Network properties versus MIPCA (A/B compartment) for IMR90 chr10 (0.0–17.3 Mb) at 50 kb resolution. (k) Correlation between LCC3 and LCC4 in mouse thymocyte (WT) cells, chr11 at 200 kb resolution. 61

Figure S1. 2 (a) Scatter plot of LCC3 and LCC4 for GM12878 chr21. (b) A/B compartments of GM12878 chr21 classified by MIPCA. (c) Ratio of LCC3/LCC4, with two slope values similar to those in panel (a) marked with dashed lines. All panels use GM12878 chr21 (15.25–48.25 Mb) at 50 kb resolution. 62
Figure S1. 3 (a) Network properties under different coverage levels for HAP1 cells, chr21 (15.25–48.25 Mb) at 50 kb resolution. (b) EC as a function of genomic node index for K562 compared with HAP1. Chromosome region and resolution are the same as above, with a constant coverage of 40%..... 63
Figure S1. 4 The comparison of chromosome structure networks between HAP1 wild type (WT) and a mutant (DCAP-H2) for chr21: 15.2–46.65 Mb at 50 kb resolution. (a) Contact network constructed both using median value as cutoff. (b) Comparison of EC, LCC3 and LCC4 values. (c) Difference in LCC values (DLCC = $LCC_{DCAP\text{-}H2}$ – LCC_{WT})..... 64
Figure S1. 5 The comparison of contact networks of different human blood cells at different stages of development (chr10: 0–17 Mb, 250 kb resolution): (a) The comparison of the contact network pattern of the three types of cells: erythroblast, megakaryocyte, and neutrophil (coverage at 49%). (b) Blood cell lineage showing the developmental relationship of these cells. (c) The comparison of network properties between these three cell types (color-coded: black, erythroblast; red, megakaryocyte; green, neutrophil). (d-e) The comparison between a pair of "close relative" cell types using B-cell and GM12878. The chromosome region, resolution and coverage are the same as in panel (a)..... 65
Figure S1. 6 Panel 1/2: The contact matrices and chromosome structure network properties (EC, LCC3, and LCC4) of chicken chromosome 21 during mitosis. The bin size is 40kb. The medians of non-zero values are used as the cutoff values for CSN. Time points (t = 0, 2, 5, 7, 10, 30, and 60 min) starting from G2 phase to late prometaphase are labeled on the left..... 66
Figure S1. 7 The cumulative fraction curves of chromosome sub-compartment as a function of the LCC values. Here, three cell lines (IMR90, GM12878, and K562) and sub-compartments obtained from different methods (Rao et al., 2014 and Xiong

& Ma, 2019) are shown. Here, 15.25–48.25 Mb of chr21 at 50 kb resolution is used. 68
Figure 2. 1 A simple network display of contacts within chromosome 21 (GM12878) Only the Hi-C interaction resolved region (15.25-48.25Mb) is used at the 50-kb resolution (orange dots). Here the top 40% (ranked by strength) of the contacts are shown (down sampled to 1% for a better visualization). Collectively, network information (node-based network centralities and clustering coefficients) provides a picture of structure organization of single chromosomes. With cross-chromosomal contacts being considered, network analysis can assist us to understand the interaction patterns between chromosomes. 74
Figure 2. 2 (a) The conversion of contact heatmap of a pair of chromosomes to a chromosome structure network shown. Three individual threshold values are used to ensure internal contact networks of chr j, of chr k, and the interchromosomal network. (b) A cartoon illustration and the corresponding matrix representation of network where the full network can be divided by intra-contact subnetwork (blue, cis-contact links) and inter-contact subnetwork (green, trans-contact links). The green subnetwork is a bipartite network. (c) The ranked contact strength as a function of its ranking (normalized by the size of contact matrix). Here “cis” indicates the intra-chromosomal contact while “trans” indicates the inter-chromosomal contacts. (d) The corresponding ranked ICE balanced contact value shows the inverse of the distribution function for chr 17 and/or chr 19 systems. The dashed line shows the cutoff position for constructing the inter-chromosomal contact network (the median of all non-zero values). 77
Figure 2. 3 (a) The top left image shows the force distribution layout of nodes in chromosomes 17-19 system, and the other panels show two chromosomes separately. A randomly sampled 1% of the top 50% interactions (edges) are shown to improve visual clarity. The rightward arrows point to chromosomes 17 and 19 with nodes colored by their genomic positions. Below, from left to right, nodes are colored by C4E, ΔC4, ΔC3, and the intensity value of the chromatin state CS13

signal (heterochromatin). (b) Th scatter plot of ΔC4 versus ΔC3 of the same Chr17:Chr19 system. (c) The corresponding scatter plot of ΔC4 versus C4E. 83
Figure 2. 4 Scatter plots of communication metrics of a specific chromosome colored by chromatin state annotation. (a) The scatter plot shows how chr19 interacts with the two other chromosomes (x-axis with chr17, y-axis with chr22), expressed by C4E and color coded by the chromatin state CS13 (intensity of the Heterochromatin signal of the fragment). Each node represents a 250kb genomic bin from chr19. Similarly, we display other situations with a combination of different chromosome, communication metric, and chromatin state in (b–f). 84
Figure 2. 5 (a) Concurrence parameter shows the percentage of the chromosome nodes that are simultaneously involved in interactions (judged by node-based metrics) with two other chromosomes. Here A indicates a subset of nodes that does not interact with either of two other chromosomes. B and C indicate sets that interact with only one other chromosome. D indicates a set that interacts with both. We define concurrence parameter = D/(B +C +D) = D/(CD + BD – D). The top 20% values of C4E or ΔC4 were selected as the cutoff value for contact. The percentages in brackets indicate the proportion of the D region to the entire chromosome. (b) The normalized overlap as a function of normalized cutoff. (c) Nodes of chromosome 19 isolated from the chromosomes 17-19 layout colored by bin index. (d) Nodes of chromosome 19 that interact with chromosomes 17 and 22 through different regions. Red indicates contact only with chr17. Blue indicates contact only with chr22. Purple indicates contact with both. The top 20% values of C4E were selected as the cutoff value for contact. (e) Same setup as panel (d) except ΔL4 is used as the interaction strength..... 87
Figure 2. 6 Cumulative distribution fraction (CDF) plots of network metrics for chromosome networks. Here classifier performance concepts using CFP (black), CFN (red), CFR (green) are shown for chr21 using feature C4A in (a), for chr17-chr19 using C4E in (b), for chr17-chr19 using DC4 in (c). interchromosomal network C4 in (d), ΔC4 in (e) and intrachromosomal network C4 in (f). 89

Figure S2. 1 Comparison of LCC3 and LCC4 values of different contact networks in Chr17:Chr19 system. Bin index 0-324 is chromosome 17 (0 – 81,195,210 bp), 325-561 is chromosome 19 (0 – 59,128,983 bp), resolution 250kb. 99
Figure S2. 2 (a-b) Force distribution layout of the (a) chromosome 17–22 and (b) chromosome 19–22 contact network system in HAP1 wild-type and ΔCAPH2 mutant cells, at 250-kb resolution. Green lines represent inter-chromosomal interactions. To improve clarity, only randomly sampled 1% of the top 50% strongest interactions (edges) are shown. (c-d) Scatter plots showing the correlation between C4E and ΔC4 in two chromosomal systems: (c) chromosome 17-22 system, and (d) chromosome 19-22 system. 100
Figure S2. 3 Network display using the top 1% links for a comparions of the random 1% of top 50% used in Main manuscript Figure 2.3a. 101
Figure S2. 4 The comparison of chr17:19 interchromosomal structure networks between HAP1 wild type (WT) and a mutant (ΔCAP-H2). Bin index 0-324 is chromosome 17 (0 – 81,195,210 kb), 325-561 is chromosome 19 (0 – 59,128,983 kb), resolution 250kb. (a) Heatmap from Hi-C data and contact network with 50% non-zero values as cutoff. (b) Comparison of LCC3 values in the intra-chromosomal contact network of chromosome 21, at 50-kb resolution. (c-e) Comparison of LCC4, ΔL3, and ΔL4 values in the inter-chromosomal contact networks of chromosomes 17–19 at 250kb resolution. 102
Figure S2. 5 Force distribution layout of the chromosome 17–19 contact network system in HAP1 wild-type and ΔCAPH2 mutant cells, at 250-kb resolution. Green lines represent inter-chromosomal interactions, and blue lines represent intra-chromosomal interactions. To improve clarity, only randomly sampled 1% of the top 50% strongest interactions (edges) are shown. 103
Figure 3. 1 (a) A contact heatmap of protein-DNA interaction with strong interaction regions labeled. The standard index is used for TR and an internal index for DNA duplex (strand I: 1-24, strand II: 25-48). (b) The cartoon representation of DBD:DNA binding. The DNA-binding box is highlighted in yellow. All residue indices are of chicken TR. The polycationic hinge distorts the dsDNA backbone

where the distances between neighboring phosphates are shortened due to the inserted arginine sidechains. 116
Figure 3. 2 (a) The contact heatmaps the internal contacts of TR. Here the upper triangle is TR with DNA and the lower one is TR without DNA. The dash lines separate three domains (DBD, hinge, and LBD). (b) A comparison of contact heatmap features (without DNA - red, with DNA - blue) using selected node-based network properties, EC, LCC3, and LCC4. (c) Flexibility expressed as calculated B-factors as a function of residue index. Again, red indicates TR without DNA and blue for TR in the presence of DNA. All residue labels are from the internal index, which can be converted to the chicken TR index by adding 48. 118
Figure 3. 3 (a) A focused view of hinge contact heatmap. (b) iPCA reveals subdomain structure of the hinge (red for TR without DNA and blue for TR with DNA). (c) Population of secondary structural elements as a function of residue index for the hinge region. All residue labels are in the internal index, which can be converted to the cTRα index by adding 48. 123
Figure 3. 4 The dynamics of hinge domain expressed using contact degrees of freedom. (a) Individual conformations projected by two latent variables using VAE scheme. Red for TR in the absence of DNA and blue for TR in the presence of DNA. Two representative contact maps and corresponding 3D structures (N-termini at upper-left) are shown. (b) The latent space of this VAE scheme. (c) The dynamic evolution of the hinge region with and without DNA. 125
Figure 3. 5 Schematic of the variational auto-encoder (VAE) workflow. The hinge contact matrix derived from residues 76–112 (37×37) is expanded to a 40×40 input by padding 1 row/column at the top and 2 at the bottom. Convolutional operations are performed with a kernel size of 3 and a stride of 2 throughout the encoder..... 126
Figure 3. 6 Effect of removing TR protein near-diagonal backbone contacts on VAE latent-space organization. $|i–j| \geq k$ indicates that only contacts at least k residues away from the matrix diagonal (the backbone region) are retained. 128

Figure 3. 7 Effect of removing Hinge region of TR near-diagonal backbone contacts on VAE latent-space organization. |i–j| ≥ k indicates that only contacts at least k residues away from the matrix diagonal (the backbone region) are retained. 128
Figure S3. 1 Overall interaction strength within and between domains indicated by block summation of the mean contact matrix. 139
Figure S3. 2 Mean distances between TR domains with and without DNA. 139
Figure S3. 3 Root-mean-square deviation of TR structure as a function of time. 140
Figure 4. 1 Docking results from AutoDock Vina for curaxins and other DNA-binding compounds interacting with different DNA conformations. Curaxins with anticancer activity are marked with "+". For each curaxin–DNA pairing, the best-scoring pose is shown. 150
Figure 4. 2 Two representative binding modes of CBL0137 predicted by AutoDock Vina. In two-dimensional interaction diagram, blue halos indicate solvent exposure, and green arrows represent hydrogen-bond donor to acceptor interactions. 151
Figure 4. 3 The motif sites in HT1080 cells treated with CBL0137 that showed increased and decreased CTCF binding. The upper motif shows increased CTCF binding, while the lower motif shows decreased CTCF binding. N represents the number of detected peaks, and the corresponding CTCF zinc finger reference binding sites are shown below. Each CTCF zinc finger recognizes approximately three base pairs; however, the green boxes are shown as schematic guides rather than exact base-pair assignments. Due to motif degeneracy and alignment across multiple binding sites, individual zinc fingers may not align precisely with a fixed base-pair position, and not all zinc fingers necessarily make direct base-specific contacts in every motif context. The blue box indicates the main differences between the two motifs. The blue and red markings indicate the two zinc finger positions we are interested in. 153
Figure 4. 4 CTCF motif variants and their genomic occurrence in HT1080 cells treated with CBL0137. Motif variants were designed to test nucleotide identity at specific zinc-finger contact positions, including AT- versus CG-rich variants at the ZF4 position, thymine versus non-thymine variants at the ZF8 position, and the U-motif. For each motif, the table reports the total number of detected motif instances across

the genome (#) and the average number of motif instances per CTCF peak (#/Peak highlighted by orange). Each motif search uses its own most stringent p-value, ensuring consistency at every position. 154

INTRODUCTION

The Structural Basis of Biological Regulation Across Scales

Biological systems rely fundamentally on the intimate relationship between molecular structure and biological function. From the three-dimensional folding of chromosomes in the nucleus to the dynamic conformational changes of transcription factors and the physical remodeling of DNA by small-molecule drugs, structural organization governs nearly every layer of gene regulation (1-3). Although modern molecular biology traditionally studies these scales separately, a growing body of evidence suggests that chromatin architecture, protein–DNA interactions, and small-molecule perturbations are deeply interconnected (3-6). Understanding this interplay requires an integrative view that spans from megabase-scale chromosome folding to atomic-level conformational transitions within individual macromolecules.

At the largest scale, the genome is not a random polymer but a hierarchically organized structure that governs gene expression, developmental programs, and cellular identity (7). Chromosomes occupy distinct territories, segregate into A/B compartments, form topologically associating domains (TADs), and establish long-range interactions both within and across chromosomes (8,9). These spatial arrangements influence enhancer–promoter communication, transcriptional co-regulation, and the stability of cell-type–specific gene expression programs (7). Genome-wide conformation capture technologies, such as Hi-C and Micro-C, have enabled quantitative maps of these interactions, yet the functional interpretation of such maps remains challenging (9,10). High-dimensional interaction matrices capture contact probabilities but do not directly reveal mechanistic principles, prompting the need for new analytic frameworks that can translate genomic architecture into meaningful structural descriptors.

One promising avenue is the application of network science to genomic data. Treating chromatin as a graph—where genomic loci serve as nodes and Hi-C interactions serve as weighted edges—allows the use of topological metrics to quantify structural roles of regions that traditional matrix-based approaches overlook. Clustering coefficients, degree distributions, interchromosomal connectivity patterns, and other graph-theoretical

properties offer insights into how local and global genome structure collectively support functional outputs such as transcription, replication timing, and nuclear compartmentalization (11). Despite advances, many aspects of interchromosomal communication and its biological implications remain poorly characterized, particularly at the node-local scale.

At an intermediate scale, transcriptional regulation depends on the behavior of transcription factors (TFs), whose structural dynamics enable precise DNA recognition and gene regulatory control (12). Nuclear receptors, such as thyroid hormone receptor α (TRα), exemplify proteins whose function emerges from a delicate interplay between structured domains and intrinsically disordered regions (IDRs). The DNA-binding domain (DBD) recognizes specific response elements, while the hinge region—an IDR—modulates flexibility, orientation, and allosteric communication between domains (13). Unlike static crystallographic structures, which capture only one conformation, molecular dynamics (MD) simulations reveal that IDRs populate large conformational ensembles that cannot be adequately described by a single fold (14). Recently, machine-learning approaches have become powerful tools for interpreting these ensembles by extracting latent structural states from thousands of simulation frames (15). Such methods offer unprecedented opportunities to quantify disorder-to-order transitions, domain rearrangements, and DNA-induced conformational shifts that shape transcription factor function.

At the smallest molecular scale, small-molecule ligands can reshape DNA structure and thereby modulate gene expression. DNA has long served as a therapeutic target, with early anticancer agents relying on DNA damage or replication stress (16). More recent work has uncovered non-genotoxic mechanisms in which small molecules alter DNA shape, groove width, twist, or mechanical properties while less likely to cause breaks or mutations (17). Curaxins represent a particularly striking example: these small molecules bind DNA in a noncovalent, non-genotoxic manner, disrupt nucleosome stability, and interfere with transcriptional regulation by altering the physical properties of DNA rather than damaging it (18). Their ability to selectively kill cancer cells highlights the importance of DNA structural plasticity in cellular homeostasis and raises compelling

questions about how small changes in DNA architecture propagate into large-scale functional consequences (17).

Together, these three scales—chromosome architecture, transcription factor dynamics, and small-molecule–DNA interactions—form a continuum in which physical structure shapes biological outcomes. A comprehensive understanding of gene regulation and cellular behavior requires integrating these perspectives rather than treating them as isolated domains. This dissertation adopts such an integrative approach, using computational modeling and structural analysis to examine how genome topology, protein–DNA interactions, and small-molecule–induced DNA perturbations influence each other and regulate function.

This introduction synthesizes current knowledge across these scales, highlights major technological and conceptual advances, and identifies key gaps that motivate the four lines of research examined in this dissertation. By contextualizing chromosome networks, MD simulations of transcription factor dynamics, and small-molecule DNA targeting within a unified structural–functional framework, this chapter establishes the conceptual foundation for understanding how molecular architecture governs biological regulation from the level of chromosomes down to individual atoms.

Three-Dimensional Genome Structure

The three-dimensional organization of the genome is a fundamental determinant of gene regulation, cellular identity, and genomic stability. Chromatin does not exist as a linear sequence folded randomly within the nucleus. Instead, it adopts a hierarchically organized set of structures that reflect both physical constraints and functional requirements. Over the past decade, advances in genome-wide conformation capture techniques, imaging methods, and computational modeling have reshaped our understanding of how chromosomes fold, how regulatory elements find each other, and how nuclear architecture contributes to biological outcomes. This section provides an integrated overview of the major principles governing genome architecture, the experimental methods used to study chromatin conformation, and the emerging recognition of interchromosomal communication as a biologically meaningful feature of

nuclear organization. These topics form the conceptual and methodological foundation for the network-based analyses developed in this dissertation.

Chromosome Territories and Large-Scale Nuclear Organization

At the broadest level, chromosomes occupy discrete and nonrandom territories within the nucleus (19). Imaging studies and modern high-resolution microscopy have consistently shown that each chromosome is confined to a preferred region rather than becoming extensively intermixed with others (20). The spatial arrangement of chromosome territories correlates with gene density, transcriptional activity, and the epigenomic state of the underlying DNA (8,9). Gene-rich chromosomes tend to be closer to the nuclear interior, whereas gene-poor chromosomes and heterochromatic regions frequently associate with nuclear periphery. This spatial arrangement has functional implications, influencing gene accessibility, susceptibility to DNA damage, and interactions with nuclear structures such as lamina or nucleoli (21,22). Although chromosome territories impose broad spatial segregation, they do not isolate chromosomes completely. Instead, they provide a structural framework within which finer organizational features arise.

Compartments, Domains, and Local Chromatin Structure

Within chromosome territories, the genome further segregates into two major compartments that reflect broad patterns of chromatin activity and nuclear positioning. Compartment A is associated with open chromatin, high gene expression, and early replication timing. Compartment B contains regions enriched for heterochromatin, lamina-associated domains, and repressive epigenetic marks (9,23). This compartmentalization was first revealed through eigenvector analysis of Hi-C contact matrices, which demonstrated that the genome displays a striking checkerboard pattern consistent with large-scale segregation of active and inactive chromatin (9). Compartment identity correlates with multiple regulatory features and remains one of the most robust and conserved aspects of genome organization (24).

At a finer scale, the genome is segmented into topologically associating domains, which are contiguous genomic intervals characterized by enriched internal contacts and insulated boundaries (8). TADs play important roles in restricting enhancer–promoter communication, stabilizing long-range regulatory loops, and organizing chromatin into

modular structural units (25). Although TADs cannot be fully described by a single mechanistic model, the loop extrusion hypothesis provides a compelling explanation for many observed patterns (26). According to this model, cohesin complexes extrude chromatin loops until encountering boundary elements, often marked by CTCF in a specific orientation (26,27). The model accounts for the dynamic yet structured nature of TADs and loops, as well as their sensitivity to perturbations of cohesin or CTCF (28). Together, compartments and TADs form a nested hierarchy that organizes chromatin across megabase and sub-megabase scales.

Experimental Approaches for Studying Genome Architecture

The development of chromosome conformation capture technologies has been essential for characterizing genome architecture. Among these methods, Hi-C has become the most widely used approach for quantifying chromatin interactions genome-wide (9). In Hi-C, chromatin is first crosslinked in vivo to preserve physical contacts, digested with a restriction enzyme, and subjected to proximity ligation so that fragments that were spatially near each other become joined into chimeric DNA molecules. Sequencing these junctions yields pairs of genomic coordinates whose ligation frequency serves as a proxy for spatial proximity (9,29). Aggregated across millions of cells, these counts form a genome-wide contact matrix in which each entry approximates the interaction frequency between two loci (9,29). This matrix provides a quantitative, population-averaged representation of the three-dimensional folding of chromosomes and has enabled the discovery of multiple hierarchical features of genome organization.

Despite its power, Hi-C datasets pose analytical challenges. Contact frequencies represent ensemble averages and do not capture cell-to-cell variation in bulk Hi-C (30). The technique is also affected by technical biases related to restriction site distribution, GC content, mappability, and sequencing depth, which require normalization strategies such as iterative correction and eigenvector decomposition (ICE) or Knight–Ruiz (KR) balancing (31). Furthermore, Hi-C does not directly measure spatial distance, nor does it differentiate between transient encounters and stable interactions (32). These limitations complicate the interpretation of contact matrices and highlight the need for computational

frameworks that can extract meaningful structural information from high-dimensional, noisy data.

Complementary methods extend the capabilities of Hi-C. Techniques such as SPRITE capture multi-way interactions that are not accessible to pairwise assays (33). Genome Architecture Mapping (GAM) infers contacts based on the co-segregation of genomic fragments in nuclear slices (34). Imaging-based approaches, including DNA FISH and live-cell microscopy, provide orthogonal validation and reveal dynamic behaviors that occur at the level of single cells (35,36). Together, these methods emphasize that genome architecture is both structured and dynamic, with a diversity of conformations contributing to the observed population averages.

Interchromosomal Interactions and Nuclear Communication

Although intrachromosomal interactions dominate Hi-C matrices, the genome also exhibits numerous interchromosomal contacts that carry functional significance. Early analyses often regarded trans interactions as stochastic or artifactual because their contact frequencies are low and their detection sensitive to noise (9). However, a growing body of evidence has demonstrated that regions located on different chromosomes can co-localize in shared nuclear subdomains and that such interactions contribute to coordinated gene regulation (37,38).

Examples of functionally important interchromosomal interactions have emerged from several biological systems. In olfactory sensory neurons, the expression of a single olfactory receptor gene requires the formation of a multi-chromosomal hub in which numerous receptor loci cluster together, thereby enabling the selection of only one active gene through competition for shared, limiting transcriptional resources (39). In pluripotent stem cells, regulatory loci involved in maintaining pluripotency have been shown to co-localize across chromosomes (40). Studies of transcription factories have revealed that actively transcribed genes from different chromosomes can converge at shared transcriptional foci (41). These findings suggest that interchromosomal interactions, although less frequent than cis interactions, are integral to the regulation of gene expression and cellular identity.

Despite this growing recognition, the analysis of interchromosomal organization remains challenging. Trans contacts occur at lower frequencies, making them difficult to detect and interpret statistically. Their biological significance can vary across cell types and developmental states, and their relationship to nuclear bodies and chromatin states is complex. The conventional matrix-based representation of Hi-C data places strong emphasis on cis interactions, which can obscure the contribution of individual loci to interchromosomal communication. These challenges motivate the need for alternative analytical frameworks capable of capturing the structural significance of trans interactions at the locus-specific level.

Chromatin as a Network: A Graph-Based Perspective

Representing the genome as a network offers an effective way to address many of the analytical difficulties inherent in high-dimensional contact matrices (42). In this framework, genomic loci are treated as nodes and the contact frequencies between them as weighted edges. This representation enables the application of graph-theoretical tools to study the topological and structural properties of chromatin (42). Network representations make it possible to evaluate the local and global connectivity of each locus, identify regions that serve as structural hubs, and quantify participation in higher-order communities (43). In conventional contact matrices, high-frequency, distance-dependent cis interactions dominate the signal, resulting in sparse and low-intensity trans interactions that are often underrepresented or treated as background. Unlike matrix-based approaches, a network model places cis and trans interactions on equal footing, enabling direct comparison of intra- and interchromosomal contributions to nuclear architecture.

The network perspective also facilitates integration with genomic and epigenomic annotations. By linking node-level features to chromatin state, lamina association, transcriptional activity, or structural protein occupancy, researchers can examine how architectural and functional properties intersect (44). Moreover, network models allow systematic comparison of genome structure across conditions or perturbations. For example, the deletion of condensin II alters global chromosome compaction and

interterritory interactions (45). A node-level network analysis can quantify these effects by evaluating how specific loci gain or lose connectivity.

Motivation for Chromosome Structure Network Study

Although substantial progress has been made in characterizing genome architecture, many questions remain unresolved. Interchromosomal interactions are understudied relative to their potential importance in gene regulation. The structural contribution of individual loci to nuclear organization is difficult to quantify using traditional matrix-based approaches. The integration of chromatin topology with functional annotation requires analytical tools that can distill high-dimensional data into biologically interpretable features. Chromosome structure network study directly addresses these challenges by constructing chromosomal structure networks and analyzing node-level metrics that capture both local and global features of genome architecture. This framework provides new insights into interchromosomal communication and enables systematic comparison of nuclear organizations across different cell types and perturbations.

Protein–DNA Interaction and Structural Dynamics of Transcription Factors

Protein–DNA interactions constitute one of the central mechanisms through which cells regulate transcriptional programs, respond to extracellular signals, and maintain lineage-specific regulatory states (46). Transcription factors rely on a combination of structured domains and intrinsically disordered regions to recognize DNA sequences, recruit co-regulatory proteins, and modulate chromatin structure (12). This duality of order and disorder endows transcription factors with the flexibility needed to navigate the complex chromatin environment of the nucleus, yet it also introduces substantial structural heterogeneity that challenges conventional methods of structural characterization. Understanding the molecular logic of transcription factor–DNA recognition therefore requires an integrative perspective that incorporates structural biology, molecular dynamics, and computational approaches. This section focuses particularly on nuclear receptors and the thyroid hormone receptor alpha (TRα), whose modular architecture and

flexible hinge region illustrate the interconnected roles of structured and disordered elements in DNA binding. These considerations establish the conceptual foundation for the molecular dynamics and machine-learning analyses developed in TR-DNA interaction research.

Structure of Transcription Factors

Transcription factors typically possess a modular organization in which distinct domains contribute specialized structural and regulatory functions (47). The DNA-binding domain, which is usually well folded and evolutionarily conserved, mediates recognition of specific sequence motifs along the genome and provides the initial anchor for regulatory complex formation (48). In contrast, the transactivation domains, interaction interfaces, and linker regions that connect structured domains often exhibit substantial intrinsic disorder (49). These disordered segments allow transcription factors to sample a broad array of conformations, interact with diverse co-regulators, and integrate contextual information encoded in the chromatin landscape. Such modularity enables transcription factors to adapt to varied regulatory environments while maintaining specificity in DNA recognition (50-52).

Nuclear receptors exemplify this modularity. They regulate gene expression in response to hormonal signals and are involved in processes ranging from development to metabolism (53). Their canonical architecture includes a DNA-binding domain at the N-terminus, a flexible hinge region, and a ligand-binding domain at the C-terminus (54). The DNA-binding domain recognizes hormone response elements arranged in characteristic orientations, while the ligand-binding domain mediates hormone-dependent recruitment of co-regulators (54-56). Although these structured elements have been studied extensively, the hinge region has historically received far less attention due to its intrinsic disorder and the difficulty of resolving its conformation experimentally.

Intrinsically Disordered Regions and Their Role in DNA Recognition

Intrinsic disorder is pervasive among transcription factors and plays an essential role in their regulatory behavior. Disordered regions do not adopt stable secondary or tertiary structures in isolation. Instead, they form dynamic ensembles of rapidly interconverting conformations that respond sensitively to environmental and regulatory cues (57,58).

These ensembles facilitate target search by enabling flexible exploration of DNA, allow transient structural elements to form upon binding, and mediate long-range communication between structured domains (59,60). They also contribute to the assembly of transcriptional complexes by providing adaptable interaction surfaces for co-regulator binding (61).

The hinge region of TRα is a representative example of such functional disorder (62). Positioned between the DNA-binding and ligand-binding domains, it plays a crucial role in modulating the global conformation of the receptor (62). The hinge influences the relative positioning of the structured domains, affects the accessibility of interaction surfaces, and responds to DNA binding by adopting different conformational states (63). Despite its importance, the structural and dynamic properties of the hinge remain poorly characterized. Given the increasing recognition that IDRs contribute to allosteric regulation in multidomain proteins, understanding the behavior of the hinge is essential for elucidating how TRα transmits regulatory signals across its domains.

Molecular Dynamics Simulation of Protein–DNA Complex

Traditional structural techniques such as X-ray crystallography and cryo-electron microscopy provide valuable snapshots of protein–DNA complexes but are limited in their ability to capture the dynamic processes that underlie transcription factor function (64,65). Molecular dynamics simulations overcome this limitation by generating time-resolved trajectories that reveal how proteins and DNA fluctuate under thermal motion (66). These simulations allow researchers to characterize the full conformational ensemble of multidomain transcription factors, quantify the range of motions accessible to disordered regions, and identify transient structural intermediates that cannot be observed experimentally (66,67).

For TRα, previous simulation studies indicate that the hinge region is highly flexible and samples a broad range of conformations (68). Its dynamic behavior is influenced by DNA association, which can shift the balance among accessible hinge states and alter how the DNA-binding and ligand-binding domains are positioned relative to one another. These observations point to the possibility of long-range allosteric communication within the receptor, where DNA recognition affects ligand responsiveness and downstream

regulatory activity (69). In this context, molecular dynamics provides a unique means of probing how multidomain transcription factors—particularly those containing substantial intrinsic disorder—explore and reorganize their conformational landscapes.

Machine-Learning Approach for Analyzing Conformational Ensembles

The high dimensionality of molecular dynamics trajectories presents significant challenges for quantitative interpretation. Each frame of a simulation represents a complex arrangement of many atoms, and simple geometric descriptors often fail to capture the subtle yet functionally meaningful differences between conformations. One strategy for addressing this challenge is to transform three-dimensional structures into two-dimensional contact maps that encode pairwise proximities between residues. This representation preserves essential structural information while presenting it in a format suitable for computational analysis.

Machine-learning methods, particularly deep-learning models such as variational autoencoders, can extract latent features from contact maps and identify major modes of structural variation across simulations (70). These latent representations often correspond to meaningful structural parameters such as hinge compactness, interdomain angles, or changes in contact density (71). When applied to TRα, these methods can distinguish DNA-bound from unbound ensembles, reveal how hinge dynamics shift upon DNA engagement, and illuminate structural states that contribute to transcriptional regulation. Machine-learning approaches therefore provide a powerful complement to molecular dynamics simulations by enabling systematic comparison of conformational ensembles and by identifying features that govern protein–DNA recognition (72).

Motivation for the Machine-Learning Analysis of TRα Dynamics

Although considerable progress has been made in understanding transcription factor structure and function, the role of intrinsic disorder, especially within multi-domain nuclear receptors, remains insufficiently explored. The hinge region of TRα exemplifies this gap. Its conformational ensemble has not been thoroughly described under physiological conditions, and the mechanism by which DNA binding reorganizes the hinge is not fully understood. Furthermore, the extent to which hinge dynamics influence long-range communication between the DNA-binding and ligand-binding domains

remains unclear. Machine-learning methods offer new opportunities for quantifying these effects but have not been widely applied to intrinsically disordered regions in protein–DNA complexes. Project 3 addresses these gaps by combining long-timescale molecular dynamics simulations with image-based machine-learning techniques to characterize the conformational consequences of DNA binding on the TRα hinge region. This integrative computational approach provides new insight into how flexible protein segments contribute to allosteric regulation and transcriptional control.

Mechanistic Insights into Curaxin–DNA Interaction

DNA is not merely a passive template for genetic information. Its physical properties—such as helical twist, groove width, bendability, and torsional flexibility—play active roles in regulating transcription, replication, chromatin organization, and the recruitment of DNA-binding proteins (73,74). Because these structural and mechanical attributes influence essential biological processes, small molecules that interact with DNA can exert wide-ranging effects through mechanisms that do not rely solely on direct enzymatic inhibition or DNA damage (75). Historically, many anticancer agents acted by inducing DNA lesions, interfering with replication forks, or creating covalent adducts that disrupt genome integrity (76-78). In contrast, a growing class of small molecules, including the curaxin family, modulates DNA structure without causing significant genotoxicity in certain cases(17). This less-damaging mode of action has generated considerable interest because it raises the possibility of altering transcriptional and chromatin regulatory programs by reshaping the physical properties of DNA itself. This section examines the unique properties of curaxins as non-genotoxic anticancer agents and discusses potential mechanisms by which their interactions with DNA may selectively impair cancer cell viability.

DNA Structural Features Relevant to Small-Molecule Interactions

The double helix possesses a rich set of structural parameters that define its shape and flexibility. These include major and minor groove widths, backbone torsion angles, base pair step parameters, and overall helical curvature. Protein–DNA recognition often depends on these physical features as much as on the underlying sequence (79). Transcription factors, nucleosomes, and regulatory complexes frequently rely on DNA

shape preferences that facilitate binding, stabilize specific conformations, or promote cooperative interactions among regulatory elements (80).

Small molecules may interact with DNA through intercalation, groove binding, or major structural perturbation of the helical backbone. Some intercalators insert themselves between base pairs, increasing helical twist and modifying stacking interactions (81). Groove-binding agents typically exploit electrostatic complementarity or hydrophobic interactions to associate with the DNA surface (82). These interactions can modulate the accessibility of DNA-binding proteins, alter nucleosome stability, or shift local DNA curvature. Such structural effects provide avenues for altering transcriptional programs independently of DNA breakage.

Historical Context of DNA-Targeting Anticancer Agents

Traditional chemotherapeutic drugs illustrate the diversity of ways in which small molecules can influence DNA structure and cellular viability. Agents such as cisplatin induce covalent cross-links that distort the helix, block replication, and activate DNA damage responses (77). Topoisomerase inhibitors (such as doxorubicin) stabilize enzyme–DNA cleavage complexes, producing strand breaks during replication or transcription (73). Anthracyclines, such as doxorubicin, combine intercalation with interference in topoisomerase activity (83). Although these compounds are effective in killing rapidly proliferating cells, their mechanisms rely heavily on DNA damage, which contributes to toxicity and limits therapeutic windows.

The pursuit of non-genotoxic DNA-targeting compounds emerged from the need to separate DNA structural modulation from DNA damage. The idea that altering DNA mechanics and chromatin organization could modulate transcriptional programs without inducing breaks represented a major conceptual shift. Curaxins are among the most studied members of this emerging class because they modulate chromatin structure in a manner that selectively affects cancer cells while sparing normal tissues (17).

Curaxins as Low-Genotoxicity Anticancer Agents

Curaxins constitute a family of small-molecule anticancer agents characterized by tricyclic structures that facilitate strong yet non-covalent interactions with DNA. Unlike classical genotoxic chemotherapeutics, curaxins do not form covalent adducts, and do not

rely on replication stress to exert their effects. Instead, they bind DNA in a way that alters its shape, changes its electrostatic environment, and perturbs chromatin organization. These structural perturbations trigger a cascade of downstream effects that ultimately impair cancer cell viability (17,18). Importantly, these effects are preferentially manifested in cancer cells. Curaxins have been shown to selectively impair cancer cell viability while exhibiting comparatively low toxicity toward normal cells (Figure I.1).

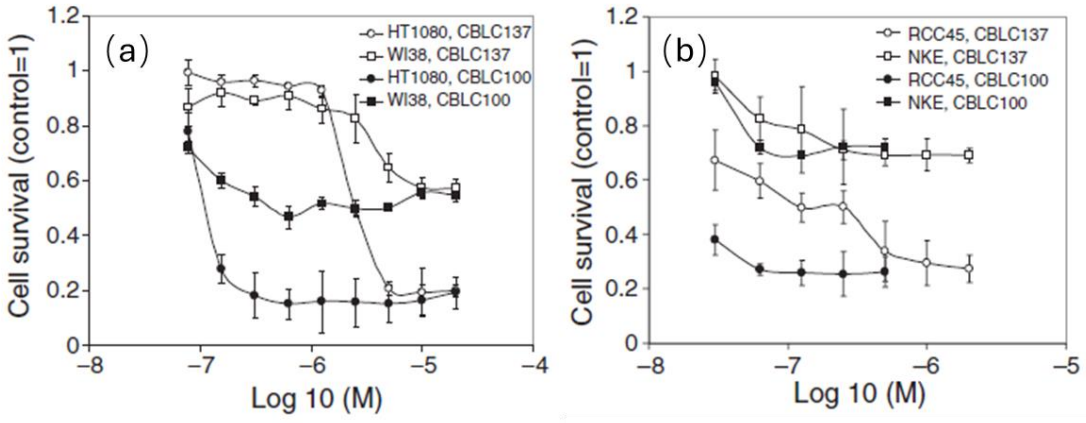


Figure I. 1 (a) Cytotoxicity of curaxins to cultured human diploid fibroblasts (Wi38) and fibrosarcoma (HT1080) cells (mean of three replicates ± SD). (b) Regrowth of RCC45 and NKE-hTERT cells after curaxin treatment for 3 hours. Methylene blue–stained cells were counted 5 days later. Adapted from ref. 17.

One of the most extensively characterized effects of curaxins is the destabilization of nucleosomes. By binding DNA and modifying its mechanical properties, curaxins interfere with the wrapping of DNA around histone cores (84-86). This destabilization leads to the release of histones and the exposure of regulatory elements that are normally shielded by nucleosome structures (84,85). The destabilization of nucleosomes has significant consequences for transcriptional regulation because it alters chromatin accessibility globally and disrupts the balance between transcriptional activation and repression (85,86). While nucleosome destabilization is expected to occur in both normal and cancer cells, its downstream consequences are highly context dependent. Cells with elevated transcriptional activity and altered chromatin organization, such as cancer cells, are more sensitive to disruptions in nucleosome stability. In these contexts, curaxin-induced chromatin perturbations may more readily lead to widespread transcriptional

imbalance and loss of cellular viability, whereas normal cells are better able to tolerate or compensate for such changes (84,85).

Another major effect of curaxins is the trapping of the Facilitates Chromatin Transcription (FACT) complex on chromatin. FACT is involved in remodeling nucleosomes during transcription, replication, and DNA repair (87). Curaxin-induced DNA structural changes enhance the affinity of FACT for chromatin, leading to its sequestration in inactive complexes (88). This sequestration prevents FACT from participating in essential regulatory processes, thereby compromising transcriptional elongation and DNA replication (86). The functional inactivation of FACT contributes to the selective toxicity of curaxins toward cancer cells, which often rely heavily on FACT due to their increased transcriptional and replicative demands (17,86).

In addition to FACT, curaxins also weaken the binding of CCCTC-binding factor (CTCF), a key architectural regulator that requires precise major-groove geometry and long-range DNA rigidity for stable occupancy (89). Whereas FACT responds to curaxin-induced DNA distortions with enhanced binding, CTCF exhibits reduced affinity, reflecting distinct structural dependencies among transcription factors (90). This differential sensitivity may contribute to curaxins' selective toxicity toward cancer cells, which rely heavily on both FACT-mediated transcriptional output and CTCF-dependent genome organization.

Structural Consequences of Curaxin Binding

Although curaxins interact with DNA non-covalently, they induce measurable and functionally significant changes in DNA geometry. Structural and biophysical studies indicate that curaxin binding can alter helical twist, modify minor-groove width, and induce local curvature or torsional stress on both naked and nucleosomal DNA (84,90). These distortions propagate beyond the immediate binding site, influencing DNA bendability, twist flexibility, and the conformational landscape available for protein binding.

Because many transcription factors rely on DNA-shape readout, even subtle curaxin-induced geometric changes can influence their binding affinity (79). Proteins such as

FACT, which interact extensively with distorted DNA, exhibit enhanced binding in the presence of curaxins. Moreover, changes in DNA flexibility and curvature influence nucleosome assembly and stability, providing a mechanistic basis for the chromatin-opening effects observed upon curaxin exposure (91). The combined impact on DNA geometry, transcription factor interactions, nucleosome positioning, and chromatin compaction reshapes higher-order chromatin organization and the transcriptional regulatory landscape.

Knowledge Gaps and Motivation

Despite substantial progress, important questions remain regarding the structural mechanisms through which curaxins modulate DNA and chromatin architecture. The molecular determinants governing curaxin–DNA binding specificity—particularly across diverse sequence and structural contexts—are not fully understood. Likewise, the extent to which curaxin-induced alterations in helical twist, groove geometry, and DNA stiffness influence protein–DNA recognition remains unclear.

A central unanswered question concerns why curaxins exhibit strong anticancer activity while showing comparatively lower toxicity toward normal cells, a phenomenon that is not yet fully understood at the mechanistic level. Current evidence suggests that curaxin-induced DNA distortions do not act uniformly across the genome but instead interact with sequence context, DNA physical properties, nucleosome positioning, and higher-order chromatin architecture (9,84). Because cancer cells typically harbor altered base composition, disrupted CTCF boundary insulation, increased chromatin accessibility, elevated transcriptional output, and greater torsional stress, their genomes may be intrinsically more sensitive to DNA-shape perturbations (92-95). Small differences in local twist, groove geometry, or DNA stiffness could therefore propagate more strongly in cancer genomes, destabilizing nucleosomes, weakening architectural factors such as CTCF, and amplifying transcriptional dysregulation. Understanding how these multi-scale structural effects differ between normal and malignant genomes is essential for explaining how curaxins reprogram gene expression without inducing canonical DNA damage.

Our study of curaxin–DNA interactions aim to understand how curaxins interact with structurally altered DNA and chromatin environments characteristic of cancer cells. Specifically, we examine how curaxins bind to DNA with different sequences and structural contexts, how they change DNA shape and mechanical properties in various chromatin environments, and how these changes affect the binding of important DNA-associated proteins. By combining computational modeling with structural analysis, this work seeks to clarify the basic mechanism of curaxin action and explain why curaxins show low genotoxicity while selectively inhibiting cancer cell viability.

REFERENCES

1. Deng, S., Feng, Y. and Pauklin, S. (2022) 3D chromatin architecture and transcription regulation in cancer. *J Hematol Oncol*, **15**, 49.
2. Henley, M.J. and Koehler, A.N. (2021) Advances in targeting 'undruggable' transcription factors with small molecules. *Nat Rev Drug Discov*, **20**, 669-688.
3. Bushweller, J.H. (2019) Targeting transcription factors in cancer - from undruggable to reality. *Nat Rev Cancer*, **19**, 611-624.
4. Vercellone, F., Chiariello, A.M., Esposito, A., Conte, M., Abraham, A., Fontana, A., Di Pierno, F., Tafuri, F., Guha, S., Kundu, S. *et al.* (2024) A Multiscale Perspective on Chromatin Architecture through Polymer Physics. *Physiology*, **40**, 291-300.
5. Conery, A.R., Rocnik, J.L. and Trojer, P. (2022) Small molecule targeting of chromatin writers in cancer. *Nat Chem Biol*, **18**, 124-133.
6. Hnisz, D., Day, D.S. and Young, R.A. (2016) Insulated Neighborhoods: Structural and Functional Units of Mammalian Gene Control. *Cell*, **167**, 1188-1200.
7. Bonev, B. and Cavalli, G. (2016) Organization and function of the 3D genome. *Nat Rev Genet*, **17**, 772.
8. Dixon, J.R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J.S. and Ren, B. (2012) Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, **485**, 376-380.
9. 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. *et al.* (2009) Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, **326**, 289-293.
10. Hsieh, T.H., Weiner, A., Lajoie, B., Dekker, J., Friedman, N. and Rando, O.J. (2015) Mapping Nucleosome Resolution Chromosome Folding in Yeast by Micro-C. *Cell*, **162**, 108-119.
11. Bernenko, D., Lee, S.H., Stenberg, P. and Lizana, L. (2023) Mapping the semi-nested community structure of 3D chromosome contact networks. *PLoS Comput Biol*, **19**, e1011185.

12. Wright, P.E. and Dyson, H.J. (2015) Intrinsically disordered proteins in cellular signalling and regulation. *Nat Rev Mol Cell Biol*, **16**, 18-29.
13. Tambones, I. and le Maire, A. (2024) Structural Insights Into Thyroid Hormone Receptors. *Endocrinology*, **166**.
14. Degiacomi, M.T. (2019) Coupling Molecular Dynamics and Deep Learning to Mine Protein Conformational Space. *Structure*, **27**, 1034-1040 e1033.
15. Noé, F., Tkatchenko, A., Müller, K.-R. and Clementi, C. (2020) Machine Learning for Molecular Simulation. *Annual Review of Physical Chemistry*, **71**, 361-390.
16. Pommier, Y., Leo, E., Zhang, H. and Marchand, C. (2010) DNA Topoisomerases and Their Poisoning by Anticancer and Antibacterial Drugs. *Chemistry & Biology*, **17**, 421-433.
17. Gasparian, A.V., Burkhart, C.A., Purmal, A.A., Brodsky, L., Pal, M., Saranadasa, M., Bosykh, D.A., Commane, M., Guryanova, O.A., Pal, S. *et al.* (2011) Curaxins: anticancer compounds that simultaneously suppress NF-kappaB and activate p53 by targeting FACT. *Sci Transl Med*, **3**, 95ra74.
18. Lu, K., Liu, C., Liu, Y., Luo, A., Chen, J., Lei, Z., Kong, J., Xiao, X., Zhang, S., Wang, Y.Z. *et al.* (2021) Curaxin-Induced DNA Topology Alterations Trigger the Distinct Binding Response of CTCF and FACT at the Single-Molecule Level. *Biochemistry*, **60**, 494-499.
19. Cremer, T. and Cremer, C. (2001) Chromosome territories, nuclear architecture and gene regulation in mammalian cells. *Nat Rev Genet*, **2**, 292-301.
20. Bolzer, A., Kreth, G., Solovei, I., Koehler, D., Saracoglu, K., Fauth, C., Müller, S., Eils, R., Cremer, C., Speicher, M.R. *et al.* (2005) Three-Dimensional Maps of All Chromosomes in Human Male Fibroblast Nuclei and Prometaphase Rosettes. *PLOS Biology*, **3**, e157.
21. van Steensel, B. and Belmont, A.S. (2017) Lamina-Associated Domains: Links with Chromosome Architecture, Heterochromatin, and Gene Repression. *Cell*, **169**, 780-791.
22. Meaburn, K.J. and Misteli, T. (2007) Cell biology: chromosome territories. *Nature*, **445**, 379-781.

23. Fortin, J.-P. and Hansen, K.D. (2015) Reconstructing A/B compartments as revealed by Hi-C using long-range correlations in epigenetic data. *Genome Biology*, **16**, 180.
24. Rowley, M.J. and Corces, V.G. (2018) Organizational principles of 3D genome architecture. *Nat Rev Genet*, **19**, 789-800.
25. Rowley, M.J. and Corces, V.G. (2018) Organizational principles of 3D genome architecture. *Nature Reviews Genetics*, **19**, 789-800.
26. Fudenberg, G., Imakaev, M., Lu, C., Goloborodko, A., Abdennur, N. and Mirny, Leonid A. (2016) Formation of Chromosomal Domains by Loop Extrusion. *Cell Reports*, **15**, 2038-2049.
27. Sanborn, A.L., Rao, S.S.P., Huang, S.-C., Durand, N.C., Huntley, M.H., Jewett, A.I., Bochkov, I.D., Chinnappan, D., Cutkosky, A., Li, J. *et al.* (2015) Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. *Proceedings of the National Academy of Sciences*, **112**, E6456-E6465.
28. Nora, E.P., Goloborodko, A., Valton, A.L., Gibcus, J.H., Uebersohn, A., Abdennur, N., Dekker, J., Mirny, L.A. and Bruneau, B.G. (2017) Targeted Degradation of CTCF Decouples Local Insulation of Chromosome Domains from Genomic Compartmentalization. *Cell*, **169**, 930-944 e922.
29. Rao, Suhas S.P., Huntley, Miriam H., Durand, Neva C., Stamenova, Elena K., Bochkov, Ivan D., Robinson, James T., Sanborn, Adrian L., Machol, I., Omer, Arina D., Lander, Eric S. *et al.* (2014) A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. *Cell*, **159**, 1665-1680.
30. Nagano, T., Lubling, Y., Stevens, T.J., Schoenfelder, S., Yaffe, E., Dean, W., Laue, E.D., Tanay, A. and Fraser, P. (2013) Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. *Nature*, **502**, 59-64.
31. Imakaev, M., Fudenberg, G., McCord, R.P., Naumova, N., Goloborodko, A., Lajoie, B.R., Dekker, J. and Mirny, L.A. (2012) Iterative correction of Hi-C data reveals hallmarks of chromosome organization. *Nature Methods*, **9**, 999-1003.

32. Dekker, J. and Mirny, L. (2016) The 3D Genome as Moderator of Chromosomal Communication. *Cell*, **164**, 1110-1121.
33. Quinodoz, S.A., Ollikainen, N., Tabak, B., Palla, A., Schmidt, J.M., Detmar, E., Lai, M.M., Shishkin, A.A., Bhat, P., Takei, Y. *et al.* (2018) Higher-Order Inter-chromosomal Hubs Shape 3D Genome Organization in the Nucleus. *Cell*, **174**, 744-757.e724.
34. Beagrie, R.A., Scialdone, A., Schueler, M., Kraemer, D.C., Chotalia, M., Xie, S.Q., Barbieri, M., de Santiago, I., Lavitas, L.M., Branco, M.R. *et al.* (2017) Complex multi-enhancer contacts captured by genome architecture mapping. *Nature*, **543**, 519-524.
35. Cardozo Gizzi, A.M., Cattoni, D.I., Fiche, J.B., Espinola, S.M., Gurgo, J., Messina, O., Houbron, C., Ogiyama, Y., Papadopoulos, G.L., Cavalli, G. *et al.* (2019) Microscopy-Based Chromosome Conformation Capture Enables Simultaneous Visualization of Genome Organization and Transcription in Intact Organisms. *Mol Cell*, **74**, 212-222 e215.
36. Quinodoz, S.A., Ollikainen, N., Tabak, B., Palla, A., Schmidt, J.M., Detmar, E., Lai, M.M., Shishkin, A.A., Bhat, P., Takei, Y. *et al.* (2018) Higher-Order Inter-chromosomal Hubs Shape 3D Genome Organization in the Nucleus. *Cell*, **174**, 744-757 e724.
37. Delaneau, O., Zazhytska, M., Borel, C., Giannuzzi, G., Rey, G., Howald, C., Kumar, S., Ongen, H., Popadin, K., Marbach, D. *et al.* (2019) Chromatin three-dimensional interactions mediate genetic effects on gene expression. *Science*, **364**.
38. Clowney, E.J., LeGros, Mark A., Mosley, Colleen P., Clowney, Fiona G., Markenskoff-Papadimitriou, Eirene C., Myllys, M., Barnea, G., Larabell, Carolyn A. and Lomvardas, S. (2012) Nuclear Aggregation of Olfactory Receptor Genes Governs Their Monogenic Expression. *Cell*, **151**, 724-737.
39. Apostolou, E., Ferrari, F., Walsh, R.M., Bar-Nur, O., Stadtfeld, M., Cheloufi, S., Stuart, H.T., Polo, J.M., Ohsumi, T.K., Borowsky, M.L. *et al.* (2013) Genome-wide chromatin interactions of the Nanog locus in pluripotency, differentiation, and reprogramming. *Cell Stem Cell*, **12**, 699-712.

40. Sutherland, H. and Bickmore, W.A. (2009) Transcription factories: gene expression in unions? *Nat Rev Genet*, **10**, 457-466.
41. Di Stefano, M., Paulsen, J., Lien, T.G., Hovig, E. and Micheletti, C. (2016) Hi-C-constrained physical models of human chromosomes recover functionally-related properties of genome organization. *Sci Rep*, **6**, 35985.
42. Sarkar, S. and Dong, A. (2011) Community detection in graphs using singular value decomposition. *Phys Rev E Stat Nonlin Soft Matter Phys*, **83**, 046114.
43. Shokraneh, N., Arab, M. and Libbrecht, M. (2023) Integrative chromatin domain annotation through graph embedding of Hi-C data. *Bioinformatics*, **39**.
44. Bauer, C.R., Hartl, T.A. and Bosco, G. (2012) Condensin II promotes the formation of chromosome territories by inducing axial compaction of polyploid interphase chromosomes. *PLoS Genet*, **8**, e1002873.
45. Ferraz, R.A.C., Lopes, A.L.G., da Silva, J.A.F., Moreira, D.F.V., Ferreira, M.J.N. and de Almeida Coimbra, S.V. (2021) DNA-protein interaction studies: a historical and comparative analysis. *Plant Methods*, **17**, 82.
46. Zhu, L. and Huq, E. (2011) Mapping functional domains of transcription factors. *Methods Mol Biol*, **754**, 167-184.
47. Latchman, D.S. (1997) Transcription factors: an overview. *Int J Biochem Cell Biol*, **29**, 1305-1312.
48. Brodsky, S., Jana, T., Mittelman, K., Chapal, M., Kumar, D.K., Carmi, M. and Barkai, N. (2020) Intrinsically Disordered Regions Direct Transcription Factor In Vivo Binding Specificity. *Mol Cell*, **79**, 459-471 e454.
49. Wright, P.E. and Dyson, H.J. (2015) Intrinsically disordered proteins in cellular signalling and regulation. *Nature Reviews Molecular Cell Biology*, **16**, 18-29.
50. Malaney, P., Benitez, O., Zhang, X. and Post, S.M. (2022) Assessing the role of intrinsic disorder in RNA-binding protein function: hnRNP K as a case study. *Methods*, **208**, 59-65.
51. Dyson, H.J. and Wright, P.E. (2016) Role of Intrinsic Protein Disorder in the Function and Interactions of the Transcriptional Coactivators CREB-binding Protein (CBP) and p300. *J Biol Chem*, **291**, 6714-6722.

52. Weikum, E.R., Liu, X. and Ortlund, E.A. (2018) The nuclear receptor superfamily: A structural perspective. *Protein Sci*, **27**, 1876-1892.
53. Sar, P. (2023) Nuclear receptor: Structure and function. *Prog Mol Biol Transl Sci*, **196**, 209-227.
54. Sun, Z. and Xu, Y. (2020) Nuclear Receptor Coactivators (NCOAs) and Corepressors (NCORs) in the Brain. *Endocrinology*, **161**.
55. Li, X., Kimbrel, E.A., Kenan, D.J. and McDonnell, D.P. (2002) Direct interactions between corepressors and coactivators permit the integration of nuclear receptor-mediated repression and activation. *Mol Endocrinol*, **16**, 1482-1491.
56. Uversky, V.N. (2019) Intrinsically Disordered Proteins and Their "Mysterious" (Meta)Physics. *Frontiers in Physics*, **Volume 7 - 2019**.
57. Tompa, P. (2005) The interplay between structure and function in intrinsically unstructured proteins. *FEBS Lett*, **579**, 3346-3354.
58. Fuxreiter, M., Simon, I. and Bondos, S. (2011) Dynamic protein–DNA recognition: beyond what can be seen. *Trends in Biochemical Sciences*, **36**, 415-423.
59. Yang, J., Gao, M., Xiong, J., Su, Z. and Huang, Y. (2019) Features of molecular recognition of intrinsically disordered proteins via coupled folding and binding. *Protein Sci*, **28**, 1952-1965.
60. Miao, J. and Chong, S. (2025) Roles of intrinsically disordered protein regions in transcriptional regulation and genome organization. *Curr Opin Genet Dev*, **90**, 102285.
61. Brent, G.A. (2012) Mechanisms of thyroid hormone action. *J Clin Invest*, **122**, 3035-3043.
62. Nascimento, A.S., Dias, S.M., Nunes, F.M., Aparicio, R., Ambrosio, A.L., Bleicher, L., Figueira, A.C., Santos, M.A., de Oliveira Neto, M., Fischer, H. *et al.* (2006) Structural rearrangements in the thyroid hormone receptor hinge domain and their putative role in the receptor function. *J Mol Biol*, **360**, 586-598.
63. Smyth, M.S. and Martin, J.H. (2000) x ray crystallography. *Mol Pathol*, **53**, 8-14.
64. Nogales, E. and Scheres, S.H. (2015) Cryo-EM: A Unique Tool for the Visualization of Macromolecular Complexity. *Mol Cell*, **58**, 677-689.

65. Dror, R.O., Dirks, R.M., Grossman, J.P., Xu, H. and Shaw, D.E. (2012) Biomolecular simulation: a computational microscope for molecular biology. *Annu Rev Biophys*, **41**, 429-452.
66. Vettoretti, G., Moroni, E., Sattin, S., Tao, J., Agard, D.A., Bernardi, A. and Colombo, G. (2016) Molecular Dynamics Simulations Reveal the Mechanisms of Allosteric Activation of Hsp90 by Designed Ligands. *Sci Rep*, **6**, 23830.
67. Martínez, L., Sonoda, M.T., Webb, P., Baxter, J.D., Skaf, M.S. and Polikarpov, I. (2005) Molecular Dynamics Simulations Reveal Multiple Pathways of Ligand Dissociation from Thyroid Hormone Receptors. *Biophysical journal*, **89**, 2011-2023.
68. Harnish, M.T., Pham, B., Arons, A.B., Xu, Y., Fernandez, E.J. and Shen, T. (2026) Conformational transition of a polycationic hinge domain contributes to DNA binding. *Biophys Chem*, **328**, 107531.
69. Wayment-Steele, H.K. and Pande, V.S. (2018) Note: Variational encoding of protein dynamics benefits from maximizing latent autocorrelation. *J Chem Phys*, **149**, 216101.
70. Matsumoto, S., Ishida, S., Terayama, K. and Okuno, Y. (2023) Quantitative analysis of protein dynamics using a deep learning technique combined with experimental cryo-EM density data and MD simulations. *Biophys Physicobiol*, **20**, e200022.
71. Durairaj, J., de Ridder, D. and van Dijk, A.D.J. (2023) Beyond sequence: Structure-based machine learning. *Comput Struct Biotechnol J*, **21**, 630-643.
72. Widom, J. (2001) Role of DNA sequence in nucleosome stability and dynamics. *Q Rev Biophys*, **34**, 269-324.
73. Liu, L.F. and Wang, J.C. (1987) Supercoiling of the DNA template during transcription. *Proc Natl Acad Sci U S A*, **84**, 7024-7027.
74. Hurley, L.H. (2002) DNA and its associated processes as targets for cancer therapy. *Nature Reviews Cancer*, **2**, 188-200.
75. Deans, A.J. and West, S.C. (2011) DNA interstrand crosslink repair and cancer. *Nat Rev Cancer*, **11**, 467-480.

76. Dasari, S. and Tchounwou, P.B. (2014) Cisplatin in cancer therapy: molecular mechanisms of action. *Eur J Pharmacol*, **740**, 364-378.
77. de Sousa Cavalcante, L. and Monteiro, G. (2014) Gemcitabine: metabolism and molecular mechanisms of action, sensitivity and chemoresistance in pancreatic cancer. *Eur J Pharmacol*, **741**, 8-16.
78. Rohs, R., West, S.M., Sosinsky, A., Liu, P., Mann, R.S. and Honig, B. (2009) The role of DNA shape in protein-DNA recognition. *Nature*, **461**, 1248-1253.
79. Kaplan, N., Moore, I.K., Fondufe-Mittendorf, Y., Gossett, A.J., Tillo, D., Field, Y., LeProust, E.M., Hughes, T.R., Lieb, J.D., Widom, J. *et al.* (2009) The DNA-encoded nucleosome organization of a eukaryotic genome. *Nature*, **458**, 362-366.
80. Lerman, L.S. (1961) Structural considerations in the interaction of DNA and acridines. *J Mol Biol*, **3**, 18-30.
81. Neidle, S. (2001) DNA minor-groove recognition by small molecules. *Nat Prod Rep*, **18**, 291-309.
82. Tewey, K.M., Rowe, T.C., Yang, L., Halligan, B.D. and Liu, L.F. (1984) Adriamycin-induced DNA damage mediated by mammalian DNA topoisomerase II. *Science*, **226**, 466-468.
83. Volokh, O.I., Sivkina, A.L., Moiseenko, A.V., Popinako, A.V., Karlova, M.G., Valieva, M.E., Kotova, E.Y., Kirpichnikov, M.P., Formosa, T., Studitsky, V.M. *et al.* (2022) Mechanism of curaxin-dependent nucleosome unfolding by FACT. *Front Mol Biosci*, **9**, 1048117.
84. Kantidze, O.L., Luzhin, A.V., Nizovtseva, E.V., Safina, A., Valieva, M.E., Golov, A.K., Velichko, A.K., Lyubitelev, A.V., Feofanov, A.V., Gurova, K.V. *et al.* (2019) The anti-cancer drugs curaxins target spatial genome organization. *Nat Commun*, **10**, 1441.
85. Nesher, E., Safina, A., Aljahdali, I., Portwood, S., Wang, E.S., Koman, I., Wang, J. and Gurova, K.V. (2018) Role of Chromatin Damage and Chromatin Trapping of FACT in Mediating the Anticancer Cytotoxicity of DNA-Binding Small-Molecule Drugs. *Cancer Res*, **78**, 1431-1443.
86. Reinberg, D. and Sims, R.J., 3rd. (2006) de FACTo nucleosome dynamics. *J Biol Chem*, **281**, 23297-23301.

87. Safina, A., Cheney, P., Pal, M., Brodsky, L., Ivanov, A., Kirsanov, K., Lesovaya, E., Naberezhnov, D., Nesher, E., Koman, I. *et al.* (2017) FACT is a sensor of DNA torsional stress in eukaryotic cells. *Nucleic Acids Res*, **45**, 1925-1945.
88. Ong, C.T. and Corces, V.G. (2014) CTCF: an architectural protein bridging genome topology and function. *Nat Rev Genet*, **15**, 234-246.
89. Lu, K., Liu, C., Liu, Y., Luo, A., Chen, J., Lei, Z., Kong, J., Xiao, X., Zhang, S., Wang, Y.-Z. *et al.* (2021) Curaxin-Induced DNA Topology Alterations Trigger the Distinct Binding Response of CTCF and FACT at the Single-Molecule Level. *Biochemistry*, **60**, 494-499.
90. Struhl, K. and Segal, E. (2013) Determinants of nucleosome positioning. *Nat Struct Mol Biol*, **20**, 267-273.
91. Flavahan, W.A., Drier, Y., Liau, B.B., Gillespie, S.M., Venteicher, A.S., Stemmer-Rachamimov, A.O., Suva, M.L. and Bernstein, B.E. (2016) Insulator dysfunction and oncogene activation in IDH mutant gliomas. *Nature*, **529**, 110-114.
92. Corces, M.R., Granja, J.M., Shams, S., Louie, B.H., Seoane, J.A., Zhou, W., Silva, T.C., Groeneveld, C., Wong, C.K., Cho, S.W. *et al.* (2018) The chromatin accessibility landscape of primary human cancers. *Science*, **362**.
93. Bradner, J.E., Hnisz, D. and Young, R.A. (2017) Transcriptional Addiction in Cancer. *Cell*, **168**, 629-643.
94. Naughton, C., Avlonitis, N., Corless, S., Prendergast, J.G., Mati, I.K., Eijk, P.P., Cockroft, S.L., Bradley, M., Ylstra, B. and Gilbert, N. (2013) Transcription forms and remodels supercoiling domains unfolding large-scale chromatin structures. *Nat Struct Mol Biol*, **20**, 387-395.

CHAPTER I

NODE FEATURES OF CHROMOSOME STRUCTURE NETWORKS AND THEIR CONNECTIONS TO GENOME ANNOTATION

A version of this chapter was originally published by Yingjie Xu, Priyojit Das, Rachel P. McCord, and Tongye Shen:

Yingjie Xu, Priyojit Das, Rachel P. McCord, and Tongye Shen. “Node Features of Chromosome Structure Networks and Their Connections to Genome Annotation.” Computational and Structural Biotechnology Journal, 2024, 23:2240-2250.

The dissertation author, Yingjie Xu, was the first author, responsible for developing the CSN framework, performing computational analyses, and drafting the manuscript. Dr. Tongye Shen provided supervision, conceptual guidance, and overall oversight. Dr. Rachel P. McCord contributed to biological interpretation, discussion of genome annotation, and critical manuscript review. Priyojit Das assisted with data processing, network construction pipelines, and preliminary analyses.

Abstract

The 3D conformations of chromosomes can encode biological significance, and the implications of such structures have been increasingly appreciated recently. Certain chromosome structural features, such as A/B compartmentalization, are frequently extracted from Hi-C pairwise genome contact information (physical association between different regions of the genome) and compared with linear annotations of the genome, such as histone modifications and lamina association. We investigate how additional properties of chromosome structure can be deduced using an abstract graph representation of the contact heatmap and describe specific network properties that can have a strong connection with some of these biological annotations. We constructed chromosome structure networks (CSNs) from bulk Hi-C data and calculated a set of site-resolved (node-based) network properties. These properties are useful for characterizing certain aspects of chromosomal structure. We examined the ability of network properties to differentiate several scenarios, such as haploid vs diploid cells, partially inverted nuclei vs conventional architecture, depletion of chromosome architectural proteins, and structural changes during cell development. We also examined the connection between network properties and a series of other linear annotations, such as histone modifications and chromatin states including poised promoter and enhancer labels. We found that semi-local network properties, including the triangle-based local clustering coefficient (LCC3)

and the square-based local clustering coefficient (LCC4), exhibit greater capability in characterizing genome annotations compared to diffusive or ultra-local node features. For example, the LCC4 can be a strong classifier of lamina-associated domains (LADs), whereas a path-based network property, closeness centrality, does not vary concordantly with LAD status. We demonstrated that network properties can be useful for highlighting large-scale chromosome structure differences that emerge in different biological situations.

Introduction

The chromosomes of eukaryotes are arranged inside the nucleus with a multi-scale architecture of complex folds which can be important for their biological functions (24). The long-range spatial structures of chromosomes can contribute to the regulation of gene expression by fostering enhancer-promoter contacts locally through loop extrusion or over longer distances (and between chromosomes) through spatial compartmentalization (7,96). Variations in chromosome structure can both be an important source of functional diversity and a cause of certain pathological traits (97-99). Chromosome structure capture methods have enabled the study of chromosome organization at the genome scale (100,101). Hi-C is a high-throughput chromosome 3D structure capture method that can reveal population-based (bulk) chromosome structures expressed by pairwise contact interactions (102,103). A typical Hi-C analysis uses a two-dimensional contact map that indicates the frequency of contacts between all possible pairs of genomic positions (genomic bins), as illustrated in Fig. 1.1. However, this two-dimensional matrix may not be intuitive to grasp. It can be difficult to visualize the correlation between the spatial structure and biological annotations, which are often associated with individual genomic regions. Therefore, extracting one-dimensional (1D) indicators from a two-dimensional (2D) contact map can be a useful way to characterize chromosome structural features and compare them to other known linear genomic features, and to discern subtle differences between cell types or upon perturbations.

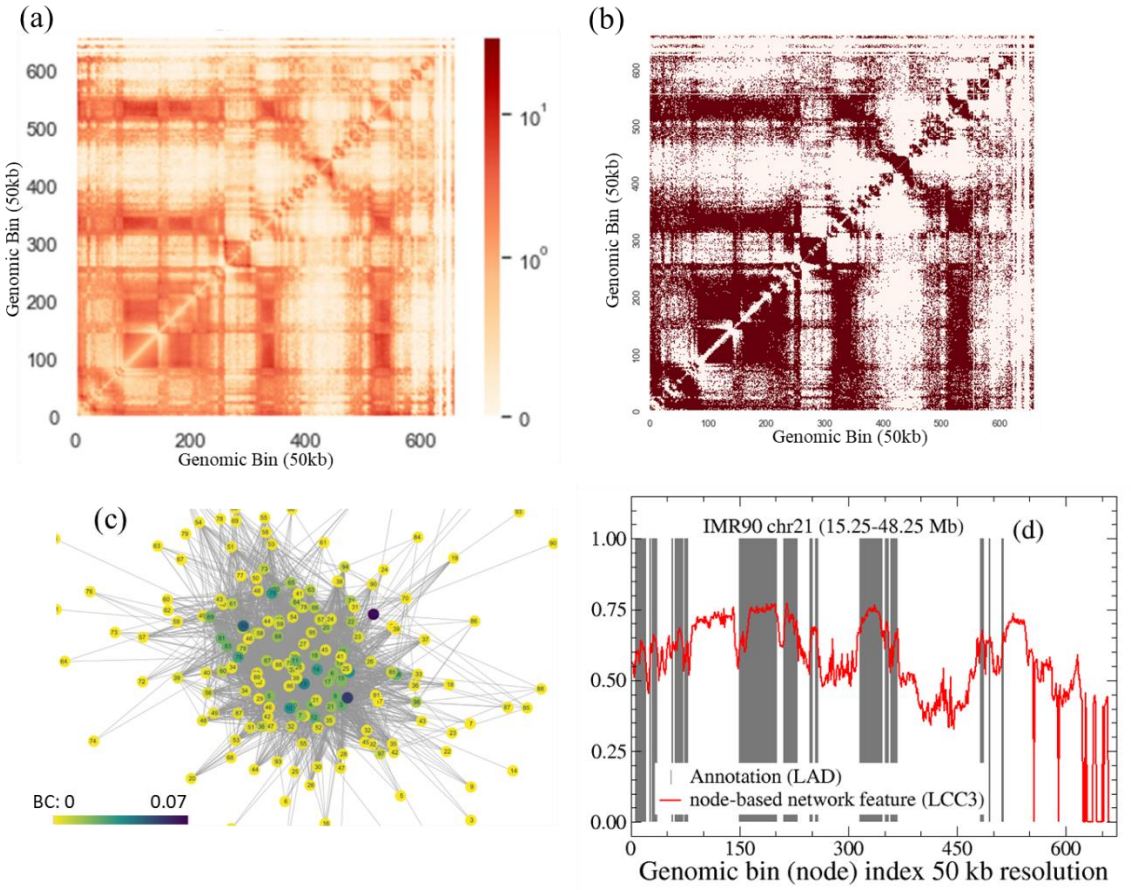


Figure 1. 1 (a) Hi-C contact map of IMR90 (lung fibroblasts cell) chr21 (region 15.25–48.25 Mb) at 50 kb resolution (b) The corresponding network adjacency matrix representation with a cutoff of 40% coverage (c) The corresponding force-directed network graph representation of panel (b). Numbers within nodes indicate the corresponding genomic bin indices, edges represent contacts, and node color encodes betweenness centrality (BC). (d) Network feature (LCC3) output (red) and LAD annotation (gray) as functions of node (genomic bin) index.

One commonly adopted 1D indicator is A/B compartment strength (which is often further discretized to a binary A/B classification) for each genomic bin position(102). A/B compartment strength is obtained from the principal component analysis (PCA) of bulk Hi-C-derived contact matrices. Other methods exist for rendering 2D contact matrices to 1D genomic bin-based information, such as the insulation score and the directionality index to define topologically associating domain (TAD) boundaries (8,104); the gene association domain (GAD) score (105), which measures the degree of spatial clustering among gene-rich regions and reflects coordinated transcriptional organization; the distal-

to-local ratio (DLR) for inferring the compaction of chromatin region, which captures the balance between long-range and short-range chromatin contacts and serves as an indicator of chromatin compaction; and interchromosomal fraction (ICF) (106), which have been used to measure dynamic properties such as Loss of Structure (LOS) after perturbation (107). Here, we explore a set of systematic methods for extracting information using abstract graph representation and network analysis. We aim to assess how these mathematical concepts relate to each other as well as the connection between the underlying genome structure and the biological linear annotation.

Discretizing spatial connection (contact interaction strength) and constructing graph representations of biomolecular structures facilitate the application of a family of node-based network properties (Fig. 1.2a), such as centralities and clustering coefficients, to understand biological functions. Node-based network properties can measure relationships ranging from two-body interaction (such as degree centrality and betweenness centrality) to higher-order ones including three-body (such as local clustering coefficient LCC3), four-body (such as square clustering coefficient LCC4), and many-body interactions (such as eigenvector centrality). At a relatively small scale, network analysis is applied to study chemical structures (108) and the internal interaction of macromolecules such as proteins. Protein structure networks have been extensively discussed and shown to be useful for studying protein structure and dynamics, (109-116). Beyond single molecules, abstract graph theories have been used to describe the structure and reaction dynamics of molecular networks of chemicals, from the oxidation of oil paintings to the sol-gel phase transition of polymers (117,118). It is interesting to examine how such abstract graph representations of structures can be connected to biological function and biophysical aspects of complex and heterogeneous chromosome structures.

Previously, many network analysis for genome biology were molecular interaction networks and co-expression networks (119-121). There is a variety of investigations using the concept of chromosome network (122), such as chromatin interaction networks(122-125). In the current study, we emphasize two unique aspects of our approach. First, our construction of network focuses on purely physical and structure

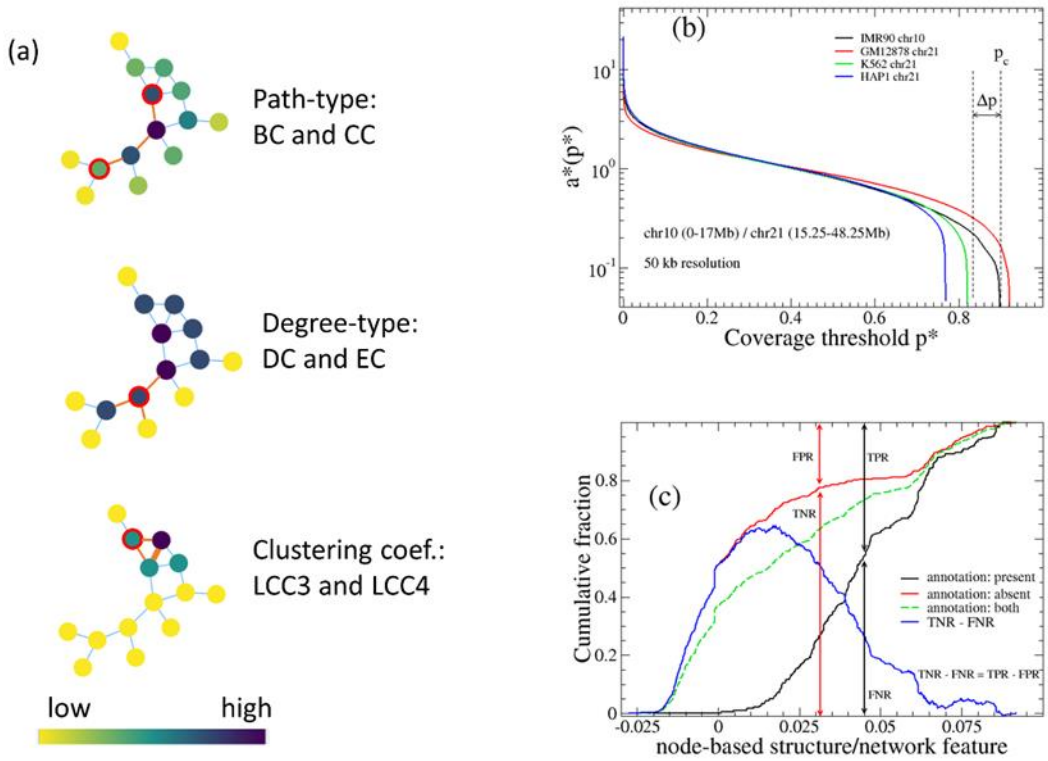


Figure 1. 2 **(a) Pedagogical schematic of 3 types of network properties. Red and orange highlights schematically indicate the nodes and edges, respectively, emphasized by each metric. The first subpanel shows path-based centrality measures, including betweenness centrality (BC) and closeness centrality (CC), which quantify how shortest paths traverse or terminate at a node. The second subpanel shows adjacency-based centrality measures, including degree centrality (DC) and eigenvector centrality (EC), which are determined by direct neighbor connectivity and global spectral properties, respectively. The third subpanel shows higher-order clustering measures, including the local triangle clustering coefficient (LCC3) and square clustering coefficient (LCC4), which quantify closed local motifs involving three or four nodes. (b) The coverage (normalized rank of contact) p^* shown as a function of contact strength a^*. Δp indicates the coverage margin. (c) Cumulative fraction curves for a label being present (black), absent (red), and reference (green). ROC is further defined based on associated concepts TPR, TNR, FPR, and FNR.**

aspects of chromosome, with node and edge definitions directly and uniformly derived from Hi-C contact matrices, without including any additional biological annotation, supervision, or inferences regarding the structure or sequence information. In contrast, numerous previous approaches have pre-selected “actively transcribed” nodes or promoter-enhancer focused networks. Additionally, we explore a set of node-based features that reflect various ranges of the interaction between a given chromosome region with its spatial neighbors (and neighbors’ neighbors). Some of the metrics are quite local, such as the degree centrality and local clustering coefficients, while others are quite global, such as closeness and betweenness centralities (114,126). It is interesting to examine whether specific features of networks reflect strongly a specific biological feature and to what extent network analysis can reveal changes in structure due to cell type variation and environmental influences.

The focus of this work is to construct the structure network of individual chromosomes and compare the network features with selected linear genomic features across different cell types under diverse physiological and pathological conditions. One example of the influence of the chromosome territory environment on chromosome internal structures is found when one compares wild-type (WT) thymocytes, lamin B receptor (LBR) mutant thymocytes, and rod cells. Interphase nuclei of WT thymocytes represent a conventional architecture: heterochromatin regions are primarily located at the nuclear periphery, whereas euchromatin (typically, regions that are gene-rich and largely active) resides in the nuclear interior. FISH experiments (127) have revealed that the structure is inverted in rod photoreceptor cells of nocturnal mammals, and likewise partially inverted in LBR -/- mutant thymocytes (128). The previous direct comparison of Hi-C data between such inverted cells and the normal counterpart found straightforward differences between rod cells and WT thymocytes. In rod cells, A/B compartment organization is inverted relative to conventional nuclei, with heterochromatin localized centrally and euchromatin peripherally (35). But the difference between LBR mutant and WT thymocytes is subtle and the A/B compartment representation was unchanged despite the partial inversion. Here, we ask whether distinct features can be extracted by network properties and how a network viewpoint enhances our understanding of this type of biological comparison.

Another application of graph representation is the influence of cell differentiation on chromosome structure. Here, we investigate how chromosome structural features are altered between progenitor cells and differentiated cells using a set of blood cell types (129). Additional applications include how chromosome structures of haploid cell HAP1 differ from those of ΔCAP-H2 mutant (a condensing II disruption of HAP1) (130) and how HAP1 differs from diploid cell. We also explore the effect of cell cycle on chromosome structures (131), the differences between a cancer cell type and its normal counterpart, and how the nuclear lamina-chromosome interaction is reflected in the chromosome structure network (CSN).

Methods and Systems

Contact network construction from chromosome structure information

An abstract graph is a useful way of representing structural components and their relationship. For each graph (network), two types of elements, vertices and edges, are present. The basic structural unit is termed a vertex, which is also called a node in computational sciences or a site in physical sciences. An edge that connects two vertices is termed a link in computational sciences or a bond in physical sciences. We will use these terms interchangeably. In this study, each node is a chromosome region (genomic bin) of size 50 kb unless specified otherwise, whereas each link indicates two such regions are close spatially and their association (contact interaction) can be detected in Hi-C experiments.

The bulk Hi-C data can be represented as a contact matrix A, where a_{jk} indicates the number of contacts recorded between regions j and k, assuming that each region is indexed from 1 to N. There are a total of N nodes in the system. Though self-edges are typically included in a graph representation and are used for downstream analyses, those contacts (diagonal elements) are not considered for the current analysis. Essentially, we would introduce a cutoff value to discretize the values of a_{jk} to either 1 when $a_{jk} \geq a^*$ or 0 otherwise. Here 1 indicates a link (edge) is formed between nodes j and k. Such discretized contact matrix is termed the adjacency matrix of a network. The discretization and construction of graph representation helps us simplify the relationship between nodes. However, one needs to be careful when choosing a threshold. Multiple methods of

defining the threshold exist and each has a different emphasis. For example, the FitHiC method uses a random polymer model as a reference and considers a link formed if the contact value is more than the reference value (132,133). Such a method might provide different answers depending on the data resolution and it may convert to an overly saturated network in practical cases, especially when the resolution is low and bin size is large.

In this study, we determined the threshold by selecting the median value of non-zero contact strengths, thereby maximizing the Shannon Entropy of edge formation within the reference random network. Essentially, 50% of the $N(N-1)/2$ pair of nodes are connected. In a binary classification, the two extreme cases are an "all-linked" complete graph and a "none-linked" empty graph. Selecting half linked ensures the sharpest contrast with a random network under this constraint. Thus, threshold p = ½ maximizes the entropy of the corresponding random network, S=(1-p) ln (1-p) + p ln p. which enhances the signal-noise ratio. Note that this threshold selection is sensitive to the data quality, i.e., the normalized number of nonzero a_{jk}. A properly selected threshold value a^* sensitively depends on experimental conditions, such as sampling size. Since the meaning of cutoff value a^* is not intuitive for selecting links, we define a concept termed link saturation (coverage) p^*, a quantitative comparison with the complete network, where all edges are formed. With a running cutoff, we can plot the function $a^*(p^*)$ as a curve to connect these two concepts. As shown in Fig. 1.2b, the j*th* ranked contact with strength a_j contributes to a point $(p^*, a^*)= (\frac{j}{\left[\frac{N(N-1)}{2}\right]}, a_j)$ on the curve $a^*(p^*)$. When a^* is 0, all links are formed and one obtains a complete graph where there are total $N(N-1)/2$ links in the graph. With increasing a^*, some of the links are removed. Parameter p^* can be defined as the ratio between the number of total links of a graph and the link number of a complete graph. We can use a cutoff link saturation (p^*) to specify a^* and further define the contact network, which means only the largest contacts (p^* ×100 %) were considered formed. The proper cutoff value ensures that network properties maintain structure information while, at the same time, the cutoff filters out other contacts deemed insignificant.

Note that practically, p^* may not always be set at a high value that is close to 1, as it is affected by how many zero elements the contact matrix has, which in turn is affected by several factors. Certain regions of the chromosome are highly repetitive and cannot be easily resolved. As a result, when there are only total N_c nonzero values, we can define $p_c = 2\,N_c/[N(N-1)]$. Since p_c represents the maximum link saturation at a contact threshold of zero, clearly p^* should be less than p_c. Other factors that affect the sampling of a contact map include crosslinking duration and sequence depth. These factors create another layer of uncertainty. Practically, we can define the number of elements a_{jk} that only have a single "hit", N_1, and a safety margin parameter $\Delta_p = 2\,N_1/[N(N-1)]$. Together we have $p^* < p_c - \Delta_p$. Ideally one could choose $p^* = \frac{p_c - \Delta_p}{2}$. Note that the raw (discrete) number of contact counts is only used for selecting proper threshold p^*, whereas a^* requires iterative correction (ICE balancing (134)) to adjust biases in Hi-C data collection before one can render a network representation. Here N_1 and N_c indicate the quality of the sampling. For example, for the whole chr21 of the IMR90 dataset (135) at 250 kb resolution, $p_c = 53.1\%$ and $\Delta_p = 1.2\%$. Thus, one should not choose p^* greater than 50% in such a case. In Fig. 1.2b, we plot the threshold value a^* vs cutoff link saturation p^* for chr10 (region 0-17Mb). One can see that a small (tolerant) a^* threshold will include all the nonzero elements of the contact matrix and results in a near maximum coverage p^*. Note that p^* does not necessarily achieve 100% as not all genome bin interactions are detected, especially at higher resolutions. For the case of IMR90 chr10, $p_c = 89.9\%$ and the coverage margin parameter $\Delta_p = 6.6\%$. Ideally one should choose p^* at the flattest region of the $a^*(p^*)$ curve since such cutoff will make the selection least sensitive to the choice of a*. According to the results, it is better to choose a cutoff of 40% under 250 kb resolution. Practically, we use p^* =40% unless specified otherwise.

Network properties

Once the discretization of the contact matrix is achieved and an abstract graph is constructed (Fig. 1.1a-c), we can further calculate site-based network properties. For a given network, there are a range of node-based network properties one can construct, from properties that are quite local and reflect only the connectivity of how a genomic

bin with its linked neighbors, to those that are global and collective. Here, we mainly study centrality properties, clustering properties, and a hybrid of centrality and clustering properties. As a comparative overview, the two centralities (closeness centrality CC and betweenness centrality BC) that are defined from paths are quite global, while the two eigensystem-based properties, eigenvector centrality (EC) and A/B strength (MI-PCA), are semi-local. Local clustering coefficients (LCCs) are a step more local, and finally degree centrality (DC) is the most local. The order (from global to local node features) is: CC and BC, EC and MI-PCA, LCC4, LCC3, and DC.

Generally speaking, betweenness centrality (BC) measures how often a node appears on the shortest paths between two nodes, while closeness centrality (CC) focuses on nodes being the ends of a shortest path (Fig. 1.2a). Both properties are path-based parameters and emphasize the global features of the network. In contrast, degree centrality (DC) and eigenvector centrality (EC) measure local properties. Both DC and EC count the number of neighbors of a node, but the main difference is that EC considers a self-consistent weight assigned to each node, and it is more global than DC. Highly connected nodes weigh more than less connected nodes. The local clustering coefficient (LCC) is another way of associating network features to node-based values. LCC is a way of indicating how dense the links are around a node, and in comparison, it is more local than some of the centralities. Specifically, the LCC value for node i is given by $LCC3_i = 2T_i/[D_i(D_i-1)]$ (126). The term T_i represents the count of triangles that include vertex i and $2/[D_i(D_i-1)]$, the total possible triangles given the degree of ith node, D_i, which defines the number of neighbors of vertex v. Using adjacency matrix, A, one can generalize LCC3 and describe completed neighboring squares for ith node (136,137), $LCC4_i = \sum_{jk} A_{ij}A_{jk}A_{ki}/[D_i(D_i-1)]$ where $D_i = \sum_j A_{ij}$ and $A_{ij} = 1$ when a link exists between i and j, 0 otherwise. As we will demonstrate below, LCC4 (LCC-even, in general) is largely independent from LCC3 (LCC-odd) and especially useful for studying certain types of networks such as bipartite networks (138). Both provide distinct perspectives of network features. Besides the loop-based generalization we adopted, there are some other expansions of definitions, such as clique-based LCCs (139) that can provide an additional insight on the architecture of networks in general.

C. Statistical analysis

Although site-based contact PCA is not deduced from the properties of a network, it is a popular way of reducing 2D contact map information to site-based values (140). For MI-PCA and downstream discretization of A/B compartment analysis, each element x_j of the top eigenvector (one with the largest eigenvalue) is associated with each genomic bin j, and the A/B compartment is assigned to each bin depending on the sign of the eigenvector element at that bin. As one group of elements is often associated with gene rich, while the other gene poor, one can assign the signs so that j belongs to the A compartment when $x_j > 0$ and belong to B when $x_j < 0$. The absolute value $|x_j|$ also is used to symbolize the strength of the compartment.

We would like to quantitatively compare local, genomic bin-associated network properties and linear annotations of linear genome information, which are often expressed as discrete (even binary) states. The conventional scatter plot and correlation coefficients may not work in such cases. Instead, we report the area under the cumulative fraction curves (CFCs), which are essentially an integration of the raw scatter plot function.

For each pair of network properties and binary annotation information, one can construct three cumulative fraction curves, CF_P (black), CF_N (red), and CF_R (green dashes), as illustrated in Fig. 1.2c. Each genomic bin position has a network value, e.g., LCC3 and a binary annotation, e.g., LAD or non-LAD. For CF_P, the cumulative fraction curve examines the positive response and increases by a fraction of $dy = 1/N_p$ at position x = $LCC3_i$, when i*th* genomic bin is a LAD and $dy = 0$ when it is not a LAD. Here N_p is the total number of nodes with LAD status. Conversely, CF_N examines the negative response and increases when a bin is not a LAD by $dy = 1/N_n$. Here, parameter N_n is the total number of non-LAD nodes. The reference curve CF_R increases by $dy = 1/(N_n + N_p)$ at $LCC3_i$ regardless of LAD status. When we make a hypothesis that larger values of LCC3 are associated LAD status, the value of CF_P is the false negative rate (FNR) while 1-FNR is true positive rate (TPR). Similarly, one can obtain the true negative rate (TNR) and false positive rate (FPR) from the CF_N curve. We can further define TNR-FNR (=TRP-FRP) and find an optimal threshold value to classify these two states (LAD or not) from

LCC3 values. Note that the positive likelihood ratio LR+ =TNR/FNP is a similar method of discerning these two curves. Further, the widely applied receiver operating characteristic (ROC) curve is plotted (x, y) = (FPR, TPR). The area under the ROC curve (AUC) is a value between 0 and 1, where a larger value implies a stronger performance measurement for the classification of an annotation from a given network property.

Systems and data

1. Hi-C data of human cell lines**:** All structural data of chromosomes used in this study came from bulk Hi-C data. In this study, we used publicly available (www.ncbi.nlm.nih.gov/geo) Hi-C data from the following human cell types: lymphoblast GM12878 (GSE63525) (135), fibroblast IMR90 (GSE63525) (135), leukemia cell line K562 (GSE63525) (135), HAP1 (GSE95014) (130) – a haploid cell derived from K562 (141), condensin II deletion in HAP1 cells (GSE163625) (130) and Hi-C data from different blood cell types: erythroblast, neutrophil, and megakaryocyte, with permission from the PCHI-C Consortium (129).

Hi-C raw data is expressed as a contact frequency between regions of the genome at a specific resolution (genomic bin). With proper procedures, one can describe the information as a contact matrix. Unless otherwise specified, all Hi-C data came through MAPQGE30-filtering and was further normalized by ICE balancing (134). Furthermore, we apply a sequence distance normalization, i.e., $a_{ij} \to a_{ij}/b_k$ with $b_k = \sum_{ij} u_{ij} \times \delta_{|i-j|-k} / \sum_{ij} \delta_{|i-j|-k} = \sum_i u_{i,i+k}/(N-k)$, where Kronecker delta selects $\mathrm{k} = |\mathrm{i} - \mathrm{j}|$, as stated in Ref. (140). Previously, we showed that unlike ICE balancing which is a necessary site-based correction, both distance-normalized and untreated versions can be useful, where the former is focused on interactions that occur more often than expected in a random polymer, and therefore perhaps are mediated by specific biological mechanisms, while the latter represents how often two regions actually come in contact in a nucleus, which has implications for biochemical reactions (140). Given our focus on studying structure patterns, we opted for distance normalization for all data presented, except in the case of chicken mitotic chromosome, where we compared both.

We focus on intra-chromosomal interactions in this study, and human chromosome 21 is examined by default. Since a part of the chr21 is poorly covered by Hi-C experiments due

to its repetitive nature, we used the 15–48 Mb region. To generalize our conclusions, we also examined another (larger) chromosome, chr10 (~ 0–17 Mb) for a comparison and obtained consistent results.

There are many interesting linear annotation data on chromosomes available (some are direct measurements and others are derived): chromosome histone modification marks, lamina-associated domains (LADs), gene expression, and derived information such as chromosome sub-compartments and chromatin states. We focus specifically on examining chromatin state, LADs, and compartment subtype in this work. We utilized these biological properties since they are closely associated with the structural properties of chromosomes and thus may manifest as network properties or show biological significance.

2. Lamina-Associated Domain: Lamin is a nuclear inner membrane protein that is critical for various biological processes within the nucleus, such as chromatin organization, DNA repair, and gene expression. In mammals, both A- and B-type lamins at the nuclear periphery interact with hundreds of large chromatin domains known as lamin-associated domains (LADs) (142). LADs might be closely tied to geometrical properties of chromosome structure that are therefore reflected in a graph representation. LADs also change their spatial localization in response to cell type-specific gene expression during differentiation and development (143). For this study, LAD data came from Tig3 fibroblasts (48) for comparison to the similar IMR90 fibroblast Hi-C data.

3. Chromatin state: Chromatin states (CSs) map epigenomic marks such as histone modifications, histone variants, open chromatin regions, and other associated marks to their likely functional roles in gene regulation (144). Each region of the chromosome is assigned to only one of the 15 CS states according to their combination of epigenetic marks: CS1=Active Promoter, CS2=Weak Promoter, CS3=Poised Promoter, CS4=Strong Enhancer 1, CS5 =Strong Enhancer 2, CS6 =Weak Enhancer 1, CS7=Weak Enhancer 2, CS8=Insulator, CS9 = Transcription Transition, CS10= Transcription Elongation, CS11=Weak Transcription, CS12=Repressed, CS13=Heterochromatin/low signal, CS14=Repetitive/ Copy Number Variation, CS15=Repetitive/ Copy Number Variation. We used CSs of GM12878 in this study.

Note that the annotation state of a genome sequence, such as chromatin state, exists at a different (and often higher) resolution than that of the network nodes. In the current study, CS data has a much higher resolution of 200 bp. When we examine cumulative fractions and make comparisons between chromatin states and network properties, we operate at the lower resolution (e.g., 50 kb) of the two datasets and directly assign the annotation state of the exact genome position of the node (the lower bound of the corresponding genome bin). Our straightforward definition works well for correlation calculations. However, in our attempt to illustrate the statistical features of network properties for a particular chromatin state using a distribution (e.g. a violin plot), such as in the case of CS3 (poised promoter), we encountered a challenge since this annotation is typically very small and is not always assigned to the genomic bin node. To resolve this issue, we practically operate at the higher resolution (200 bp) mode and interpolate by assigning the same network property value to all bins that fall within each 50 kb range (one structure node). This procedure does not imply that a sparse chromatin state occupies the entire 50 kb region, but instead provides a weighted statistical association between high-resolution chromatin annotations and node-level network properties. Thus, when constructing the violin plots, multiple chromatin state labels can be assigned to the same node and quantitative weights ensured.

4. Subcompartments**:** In addition to the binary classification of A/B compartment (largely correlated with euchromatin vs heterochromatin) (145), additional studies classify regions into multiple compartment subtypes (sub-compartments) using a range of data, including interchromosomal contacts, strength of PCA eigenvectors, gene expression, and epigenetic information (135,146-150). Here, we compared our network properties across several cell types using two sub-compartment definitions, which extends the A/B definition to six states, A2, B1, B2, B3 and B4 (135,145).

5. Mouse Interphase nuclei**:** A partially inverted architecture was observed when lamin B receptor is deleted from thymocytes. Compared to WT thymocytes, these LBR mutant cells cannot anchor lamin associated domains onto the nuclear membrane (128). Though it is relatively simple to discern the differences in Hi-C maps from WT thymocytes

(conventional) and rod cells (inverted) because these are quite different cell types, it is difficult to distinguish between LBR -/- (partially inverted) and WT (conventional) architecture using Hi-C contact maps directly. Therefore, we examine whether network properties can distinguish these two nuclear organizations based on their subtle changes in Hi-C data. We used the gene-dense mouse chr11 for this study. The original data resolution is 20 kb, and we use 200 kb for better sampling in the network analysis.

6. Chicken mitotic chromosome: Chromosomes go through dramatic structure transformations during the cell cycle. We study CSNs derived from Hi-C data of mitotic chicken chromosome 21 during the transition from G2 into prophase (time = 0, 2, 5, 7, 10, 30, 60min) from Ref. (131). ICE balance and distance normalization were applied. The resolution of the contact matrices is 40 kb.

Results and Discussion

Effects of coverage and resolution on chromosome structure network properties

We first report how resolution and threshold may affect the analysis using chr21 of IMR90 (region 15.25–48.25 Mb) as an example. By varying coverage parameter p^*, we found that a higher cutoff value such as 80% results in constant high values of the node-based network property LCC3, as the nodes of the CSN are almost fully connected. Conversely, if the threshold p^* is as low as 20%, the resulting curve has larger variation with some nodes being zero, as the CSN has fewer links and is broken into disconnected smaller subnetworks (Fig. 1.3a). These results suggest a p^* threshold of 40%, the same as chosen based on observations from Fig. 1.2b, is also appropriate to yield network properties that vary meaningfully across the chromosome.

Resolution also has an impact on network properties, as shown in Fig. 1.3b. For example, there is a potentially spurious high peak for LCC3 values near 37.5 Mb (bin index 89). However, the peak disappears if the resolution is 50 kb (bin index 445), when the genomic bin size is decreased from 250 kb to 50 kb. We observe that the resolution has a relatively small effect on EC, which indicates eigenvector centrality is less sensitive to resolution selection since EC is more of a semi-global property compared to local properties such as LCC3. Generally, a smaller bin size provides a high-resolution description only if there are sufficient samples in each bin.

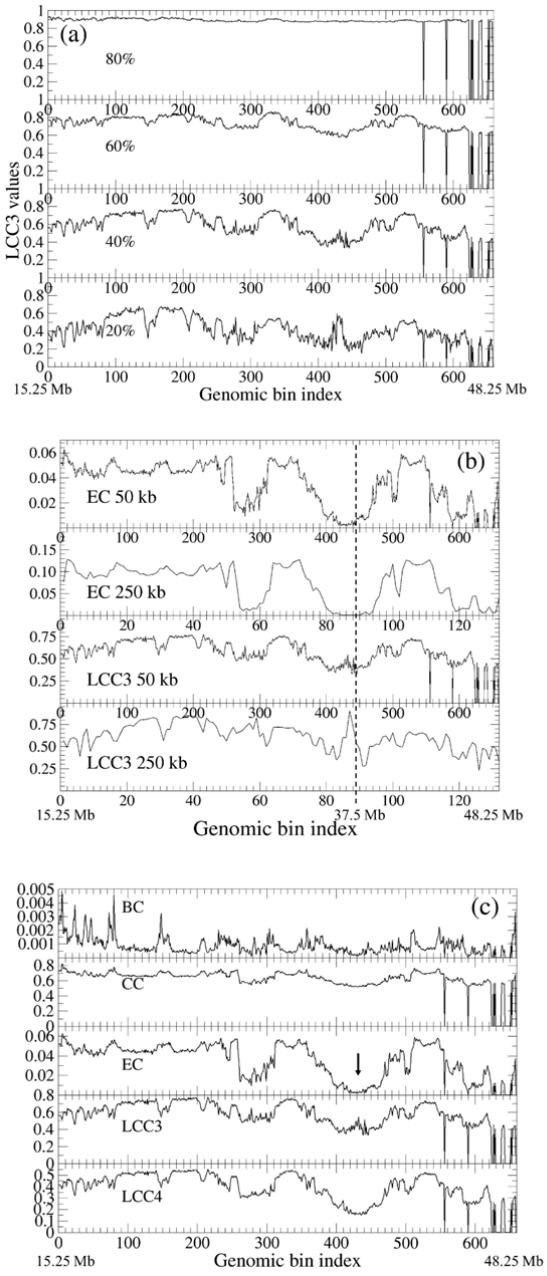


Figure 1. 3 (a) LCC3 value as a function of node (genomic bin) index, under different coverage thresholds p^* (max, 80%, 60%, 40%, 20%) at 50 kb resolution for IMR90, chr21:15.25–48.25 Mb. (b) IMR90 chromosome 21 (same region as above) 40% p^* cutoff EC and LCC3 values at different resolution (50 kb vs 250 kb bin sizes). (c) Comparison of different network properties including BC, CC, EC, LCC3 and LCC4 across the same IMR90 chr21 region as in a and b. Arrow indicates a notable region of divergent values between network properties at a given genomic location.

We next compare the similarities and differences in the quantified network properties, LCC3, EC, BC, CC, and LCC4, using the same parameters (250 kb bins and p^*= 40%) (Fig. 1.3c). BC exhibits the most frequent variations, while the other four properties show changes more gradually across the genomic region. Overall, BC values are nearly constantly low while CC values are constantly high. EC and LCC3 have a wider range of variation. However, their peaks and valleys are not correlated, which indicates that different network properties reflect distinct features of the CSN. Using several different network properties may provide a more comprehensive understanding of chromosome structure. For example, at positions around 35,000~37,500 kb (arrow, Fig. 1.3c), EC is almost zero, the LCC4 value is around 0.3, and the LCC3 value is high. The low CC and EC suggests that the region has few neighbors, while the contrast between LCC3 and LCC4 values indicates that there is a short-distance ring structure in the area. Thus, this region is like an isolated island, likely to form links within itself rather than with others. The network properties provide different measurements of the structure features. One can further define a derived network property as a combination of basic network properties: BC, CC, EC, and LCCs. A previous study pointed out possible correlations between derivative network properties, such as BC vs DC $\times$ (1 $-$ LCC3) (151). We directly compared the correlation between different network properties and between network properties and A/B compartment strength in Supplementary information Fig. S1.1. Among all the basic network properties studied, the majority show little correlations, indicating that they are largely independent metrics. Only a few pairs of properties show correlation in restricted regions. For example, LCC4 and A/B compartment strength show correlation at high value regions whereas EC and A/B correspond at low value regions. We found only one strongly correlated pair: EC and DC $\times$ LCC3. Additionally, as shown in Fig. S1.2, a derived network property, the LCC4/LCC3 ratio, is correlated to the A/B compartment classification. This ratio compares the relatively longer connectivity (neighbors of nearest neighbors) to short-range connectivity (nearest neighbors). Nodes with high LCC4/LCC3 ratios are likely a part of the inactive region (B compartment) and vice versa. This distinction further underscores the utility of graph properties in characterizing of biologically related structure features.

Network properties display distinct structural features of chromosomes for different cell types

There are several epigenetic differences between chromosomes of different cell types. We evaluated whether network properties and CSN can discern the structural differences that result from epigenetic differences. We first compared the CSN of chr21 between a normal cell type (GM12878, non-cancerous lymphoblast) and a cancer counterpart (K562; leukemia). As shown in Fig. 1.4a, the overall fluctuation of LCC3 and LCC4 values of K562 is much more subdued than the corresponding fluctuation of normal cells (GM12878). This plateaued pattern, which does not vary significantly between sequence neighbors, suggests that many nodes on a chromosome have a similar (and even) degree of connectivity with other nodes in the network. This loss of network features in bulk cancer cell data may reflect the loss of functional organization of the chromosome. Therefore, unlike normal cells which have highly organized chromosome structure, the degree of chromosome organization in cancer cell K562 is weakened. Although most of the remaining peak and valley positions of K562 are similar to that of GM12878 (especially for EC), there are exceptions. For example, K562 is missing the peak at about 37 Mb for LCC3, potentially a site of more dramatic chromosome structural difference. In addition to changes in chromosomal structure in cancer, we also examined the impact of interactions between homologous chromosomes on genome structure. In this case, haploid cells HAP1 vs. diploid cells GM12878 were compared. HAP1 has only a single copy of each chromosome except chromosome 15 (the second copy of chr15 is only an incomplete fragment). Therefore, chromosomes of haploid cells may lose some of the interchromosomal interactions of the corresponding diploid cells. We constructed contact networks for HAP1 using the same cutoff ($p^* = 40\%$) and displayed the EC and LCC3 results in Fig. 1.4b. Note that there is a dispairty in the coverage of Hi-C sequencing data, where GM12878 has a higher p_c than HAP1 (p_c =92.5% vs 77.2%). Another aspect to consider when interpreting HAP1 and K562 results is the translocations that occur in these cell lines (152). For example, chr21 and chr12 exhibit translocations in the region

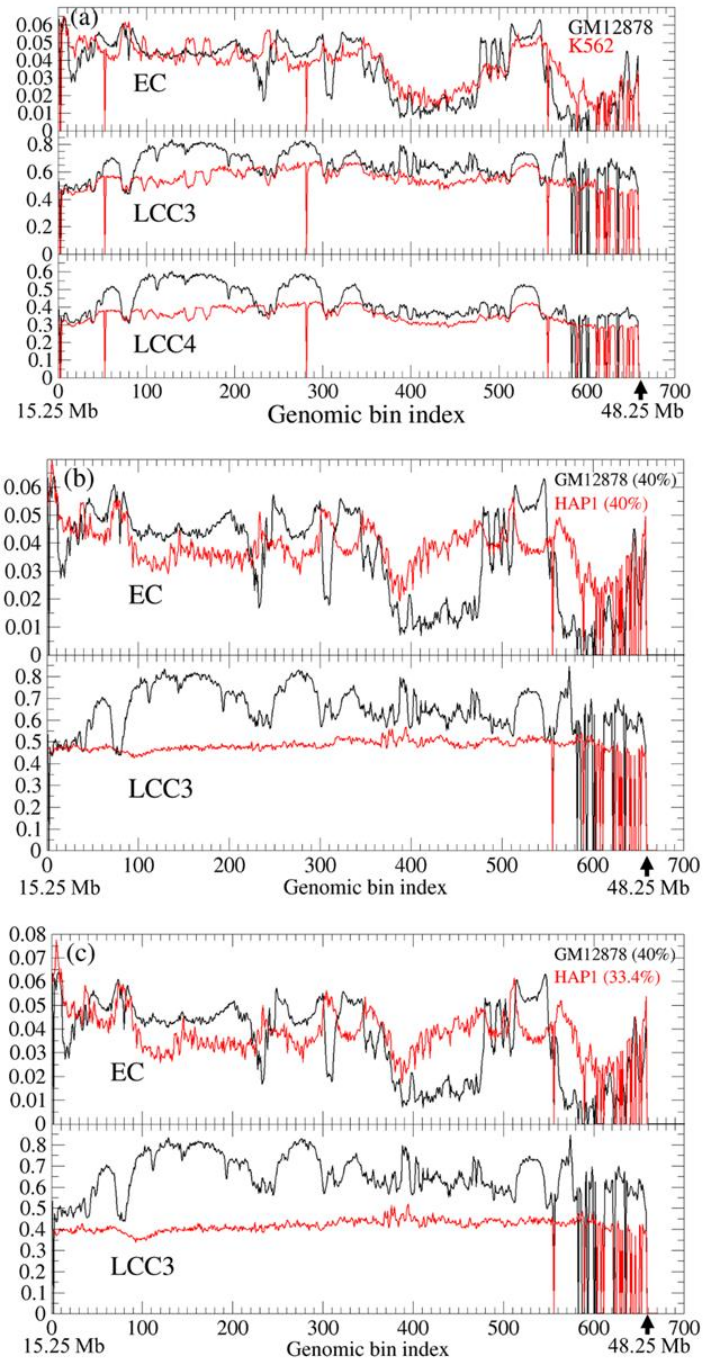


Figure 1. 4 Network properties (EC and LCC3/4) as functions of network node index for cancer cell type K562 are shown in (a) and for haploid cell HAP1 in (b). The corresponding results for lymphoblast GM12878 are shown as a comparison. Data of region 15.25–48.25 Mb for chr21 (50 kb resolution) is used. Constant coverage of 40% is used. (c) Same comparison as (b) but HAP1 coverage is set to 33.4%, so that p^*/p_c is kept constant.

near chr21:27Mb and chr12:23Mb in K562. These genomic rearrangements can affect contacts near the edges of the translocation region.

When using the same cutoff p^*, the system with lower p_c has fewer dispensable edges to choose and thus leads to a structure network that demonstrates an ensemble average of configurations, an evenly connected global network. Since p_c (data coverage) influences the effect of p^*, we also used an alternative definition, a scaled p^* to compare different systems. Specifically, instead of using a constant p^* , we use different p^* for different systems while keep $\frac{p^*}{p_c}$ = constant. For this case, we make sure that $\frac{p^*_{HAP}}{p_{c_{HAP}}} = \frac{p^*_{GM}}{p_{c_{GM}}}$. As shown in Fig. 1.4c, we set p^*=33.4% to construct the network of HAP1, to ensure 33.4%/77.2% = 40%/92.5%. When we compare Fig. 1.4b and 1.4c, the results are not sensitive to the selection of p^*, i.e., lowering the cutoff value of HAP1 cells, did not significantly affect LCC3 and EC values. Similar to the cancer cell K562, from which HAP1 is derived, the LCC3 values are almost constant, and each position has a similar number of random contacts with its surrounding network. Subsequent tests showed that even if p^* is lowered to 20%, the LCC3 value of HAP1 would still be largely constant with slightly more random fluctuation (Fig. S1.3). Therefore, this phenomenon is robust and not an effect of p^*, which suggests the lack of well-defined chromosome structure features in HAP1. Since HAP1 is both haploid cell and originates from cancer cell line, it is unclear whether one or both factors contribute to the phenomenon. However, upon examining the EC values, we observe that HAP1 exhibits further attenuation of the organizational structure (for example, in the 19.5–24 Mb region, SI Fig. S1.3) compared to that of K562. This structural changes might be caused by the haploid nature of the HAP1.

We also studied a HAP1 mutation (DCAP-H2) (130), as shown in SI Fig. S1.4. CAP-H2 is a subunit of Condensin II, which is involved in chromosome looping and condensation. Previous research reported large scale chromosome structure changes and notable changes in interchromosomal contacts with CAP-H2 deletion (46). However, local changes are difficult to discern from the Hi-C contact maps alone (SI Fig. S1.4). Although condensin II primarily affects regions near centromeres, our selected region (q-

arm of chr21) still exhibits clear effects of this mutation using network analysis. We observed that notable changes occur in LCC3 and LCC4 while EC is largely perserved. Their values decreased across the entire chromosome, indicating a general loss of intermediate-scale connectivity along the chromosome, consistent with the loss of loops formed by condensin. Additionally, the CAP-H2 mutant cells exhibited greater variations in LCC3 and LCC4 compared to the WT, indicating that certain genomic regions are more strongly influenced by the loss of condensin than others. Therefore, network properties can reveal differences in decondensation after condensin deletion, which may not be readily apparent from the 2D contact map alone.

Rod cells in mice have a unique inverted nuclear structure. It is difficult to visualize the difference between the "inverted" and "non-inverted" structures from the Hi-C contact map directly. While distance-based methods, such as image-based chromatin tracing (153), may be better at discerning this structure difference than Hi-C, we demonstrate that CSN metrics can still assist in revealing differences in the Hi-C information,. We analyzed the differences between WT type and LBR mutation using the chromosome structure networks. As shown in Fig. 1.5, BC, CC and EC show a similar distribution pattern of high values at both ends and low in the middle. Conversely, LCCs are higher in the middle than at both ends. This suggests that the nodes in the central region of mouse chr11 have fewer neighbors but are more locally connected, whereas the nodes at the ends are more urban. A scenario of long-distance enhancer-promoter interaction at the ends and highly correlated regulation complexes in the middle can be a realization of such CSNs. Compared with the wild type, the value of network properties fluctuated less in the LBR mutant. Similarly, the LCC values of LBR mutant are lower than that of wild type, while BC and CC do not change much, which suggests that the LBR mutation has weaker short-distance contacts and a loss of some local organization. Thus, chromosomes of LBR mutant cells have weakened short range contacts compared to those of WT cells.

As a further comparison of related cell types, we compared three cell types from different branches of the same hematopoietic lineage: erythroblasts, neutrophils and megakaryocytes.

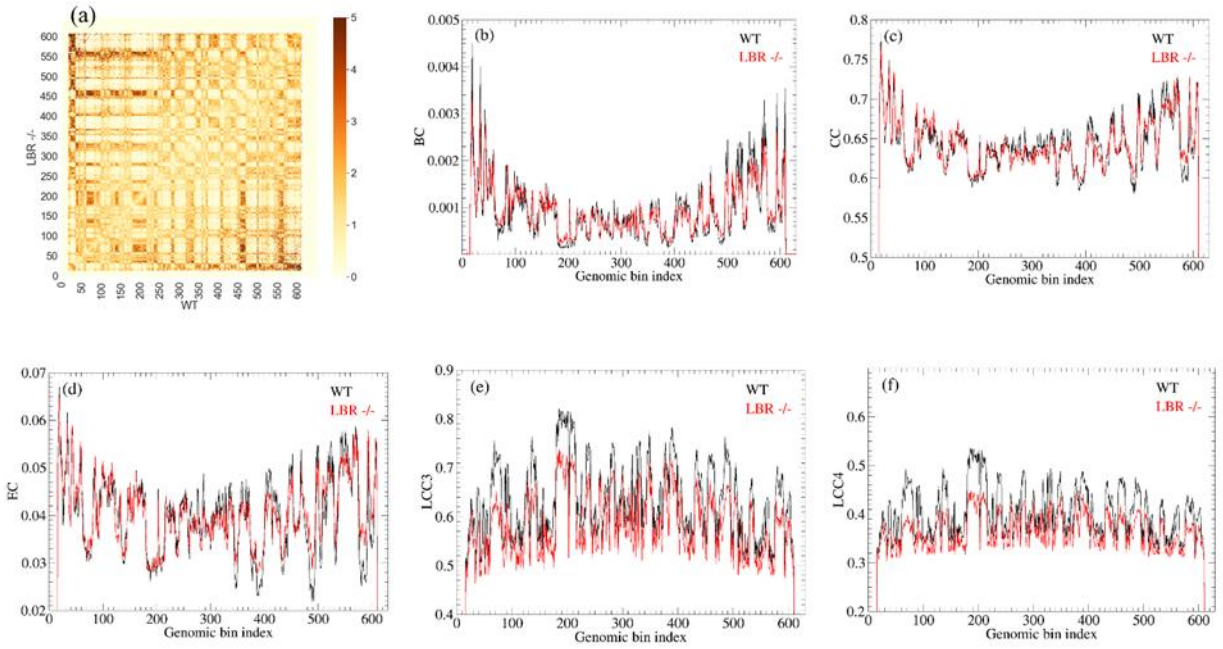


Figure 1. 5 The structures of chr11 of mouse rod photoreceptor cell wildtype (WT) and LBR mutant are characterized using network properties at a resolution of 200 kb. (a) Heatmap comparison. Above the diagonal is LBR -/- cell. Below the diagonal line is WT cell. (b) BC value comparison between two the above cells. (c) CC value comparison. (d) EC value comparison. (e) LCC3 value comparison. (f) LCC4 value comparison.

As indicated in SI Fig. S1.5, the network properties of erythroblast and megakaryocyte cells are relatively close, whereas neutrophils are far from both erythroblasts and megakaryocytes. For example, the BC and LCC values for neutrophil structure have extremely high peaks around chr10:12.5 Mb (50 bin index). Such a relationship is consistent with the differentiation path of the cell types. Both erythroblast and megakaryocyte cells can further give rise to additional cell types (such as erythrocytes and platelets) in processes that involve the loss of the nucleus, while neutrophils are at a terminal stage (38). This difference may be reflected in the increased structure specialization of the neutrophil chromosomes. These selected three cell types exhibit significant overall differences from each other, raising the question of how effectively network properties can distinguish between more closely related cell types. To explore

this question further, we also compared a pair of more closely related cell types, B cell (59) and GM12878 in SI Figure S1.5 (d) and (e). We observe that node features are much closer to each other than those in (a-c), yet certain notable differences persist.

To evaluate how CSN metrics perform with dramatic shape changes in entire chromosomes, we examined the transformation of chromosome structure from interphase through prophase to metaphase. We used Hi-C data from chicken chromosome 21 at different stages of chromosome condensation (time = 0, 2, 5, 7, 10, 30, and 60 min) (131) to illustrate the large scale chromosome dynamics. As shown in SI Figure S1.6, we compared both distance decay normalized (panel 1) and nonnormalized (panel 2). We found that the node features of CSN do not simply disappear over time, even for the normalized cases. Instead, several types of structure rearrangements occur over time. For example, EC and LCCs indicate a type of structural reorganization occurs at 5 min and another one at 30 min. However, creating a distinct network from uniformly condensed mitotic chromosomes (at later timepoints) becomes more difficultand the discretized contact maps at these stages show few distinctive features. The corresponding results without normalization show that the EC metric can capture certain global features of the structural changes as chromosomes progress to condensed metaphase stages. Overall, we found that LCCs did not reveal hidden features overlooked by a direct inspection of the 2D contact maps. This might be a limitation of the network metrics which focus on the connectivity of only a few nodes and not sensitive to drastic global changes.

The connection between network properties and genomic annotations

Having observed that quantitative CSN properties differ between cell types and conditions, we also evaluated whether these properties of the structural network relate to other biologically significant linear annotations along the genome. A major advantage of CSN is the potential for data reduction. Various node-based network properties render the complex structure information (two-dimensional contact matrix originated from Hi-C) into different one-dimensional functions of positions (genomic bins). In this way, it is convenient to compare 1D node properties with known genomic linear annotations. Indeed, previous work has observed that Hi-C network properties such as node degree can vary between promoter and enhancer regions (154).

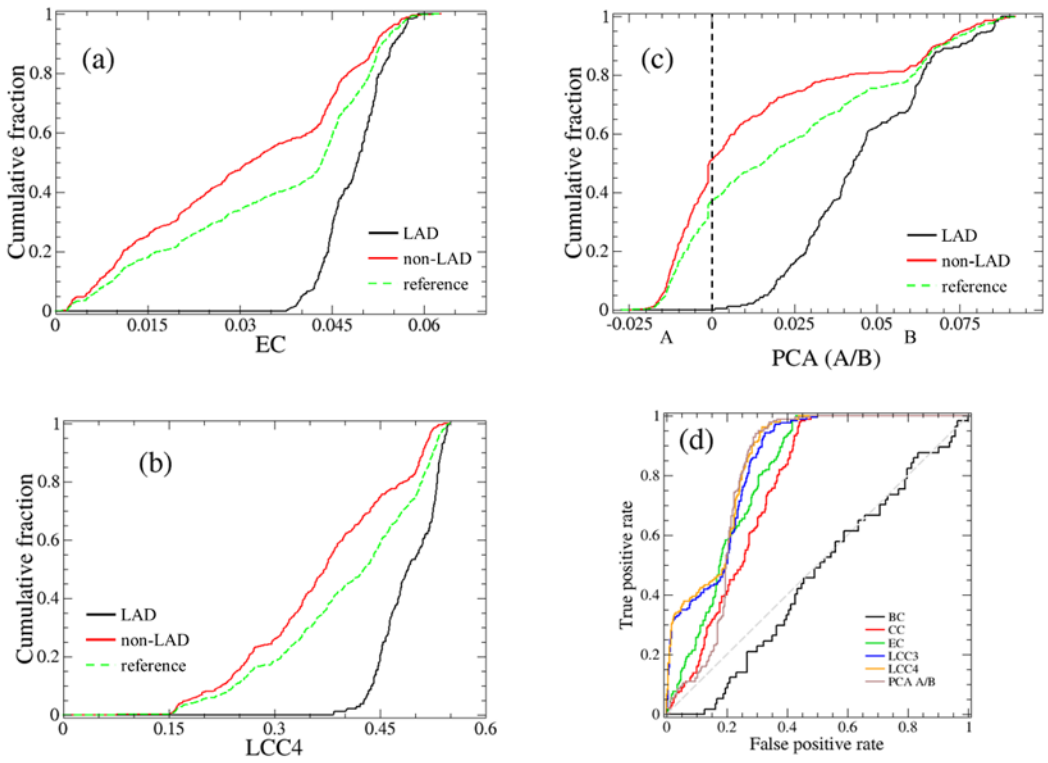


Figure 1. 6 LAD cumulative fraction (CF) plots of network properties and MI-PCA for a region of IMR90 chr21. Here classifier performance concepts using CFP (black), CFN (red), CFR (green dash) are shown for EC in (a), LCC4 in (b) and PCA in (c). (d) ROC curves of six different network properties for the LAD classification.

In our study, we used LAD status and chromatin states as examples to study the relationship between genomic linear annotations and network properties of the structure (LCC3, LCC4, EC, BC, and CC), focusing on IMR90 cells, as shown in Fig. 1.6. We also compared LAD annotation with A/B compartment strength (top eigenvector MI-PCA of Hi-C contact (140), up to a sign, Fig. 1.6c), where we observe that LAD labels are more likely associated with compartment B. Note that we did not fix the sign, and compartment B is positive (and A is negative) for this top eigenvector of MI-PCA. Though the A/B compartment value is not a node-based network property, it nevertheless provides a value associated with each genomic position. Researchers have routinely used it to classify compartments, while more advanced methods of classification have been developed in

recent years (149). As described in the Methods, we used cumulative fraction curves (CFCs) to discern the separation of LAD states (LAD versus non-LAD) by different node network properties. As shown in Fig. 1.6a, EC can clearly separate LADs from non-LADs. The genomic bins associated with LAD labels are more likely to have high EC values. Similar results can also be observed for the CFC of LCC4 in Fig. 1.6b. In contrast, although MI-PCA can also distinguish LAD and non-LAD regions at low values, the three CFC curves representing LAD, non-LAD and reference overlap at high values (A compartment) (Fig. 1.6c). Biologically, these results suggest that lamin-associated domains are characterized by strong global and higher-order structural connectivity rather than by simple local compaction alone. The preferential association of LADs with high eigenvector centrality implies that LAD regions tend to interact with other structurally important regions that are themselves highly connected, reflecting coordinated anchoring to the nuclear lamina. In contrast, the reduced discriminative power of A/B compartment values at high ranges indicates that compartment-level measures capture broad chromatin segregation but do not fully reflect the network-level structural coherence that distinguishes LADs from non-LAD regions.

The comparison of how well each network property can distinguish LADs from non-LADs can be seen from the ROC graph in Fig. 1.6d. When an ROC is high (thus giving a larger AUC value), there is a strong separation of LAD vs. non-LAD by this metric. LCC4 values (AUC = 0.852) are the strongest LAD predictor, followed by EC and MI-PCA, while LAD regions show no distinction based on BC values. Except BC, all five other network properties are associated with differences in LAD classification. Both LCC3 and LCC4 outperform others, including the A-B compartment signal, in resolving LAD classification especially at the small false positive rate region. This suggests that LADs regions have characteristic modes of local clustering within the chromosome structure that are even more distinctive to LADs than B compartment association in general. This observation confirms the importance of quantifying different network properties to study chromosome structure and biological annotation. By examining various network properties that range from global to local, we find that LADs seem to have more distinctive local rather than global network properties.

Besides comparing LAD and MI-PCA (A/B compartment strengths), we also examined the correlation between A/B compartment subtypes and network properties. The difference between A1 and A2 sub-compartments is that A1 has higher CG content and shorter genes. B1, B2, and B3 differ in their association with histone modifications. According to the available data, chr21 of IMR90, GM12878, and K562 cells contains up to four sub-compartments A1, A2, B1, and B2. By comparing LCC3 and LCC4 with A1, A2, B1 and B2 classifications, we found that network property LCCs can distinguish different compartment subtypes in IMR90 (SI Fig. S1.6). For both subcompartment definition approaches for IMR90 (135,145), LCC3 and LCC4 values show significant distinctions between A1 and B2 regions, with A2 and B1 regions showing intermediate LCC3 and LCC4 values. This pattern is also evident for LCC4 values in GM12878 and K562 cells , although for GM12878 cells LCC3 values are much more similar across all subcompartment types.

Additionally, we compared chromatin states with network properties. Chromatin states classify chromatin regions by properties such as their epigenetic marks and transcription state to identify their specific functions in transcriptional regulation (155). The chromatin state data of GM12878 chr21 (GEO:GSM936082) was used in this comparison. As shown in Fig.1.7a and 1.7b, we show cumulative curves with respect to LCC values for the six most data-rich chromatin states. Despite the similar nature of the definitions, the accumulation curves of chromatin states in LCC3 and LCC4 are distinct in discerning chromatin state classifications. LCC4 can better discriminate heterochromatin (CS13) versus other states. The distribution of either LCC metric for poised promoter state classification (CS3) is sharply peaked and there are no zero values (Fig. 1.7c and 1.7d). This result suggests that the chromatin state label CS3 may have a highly specialized structure motif. This observation may be connected to previous findings suggesting that the promoter-based subnetwork, as opposed to the rest, has a signature of positive assortativity (125). In addition, we found that some chromatin states, such as heterochromatin and repetitive/CNVs, had more than one aggregation peak at LCC4, which may indicate that there are subtypes of chromosomal structures associated with these CS labels. As previous literature on heterochromatin structure suggests, there is a

possible phase separation of heterochromatin (156). It is an interesting hypothesis that the two sub-states observed for CS15 (heterochromatin) could represent one relatively condensed phase (positive assortativity) and one relatively open structure (negative assortativity) as depicted in Figure 3c of Ref. (156).

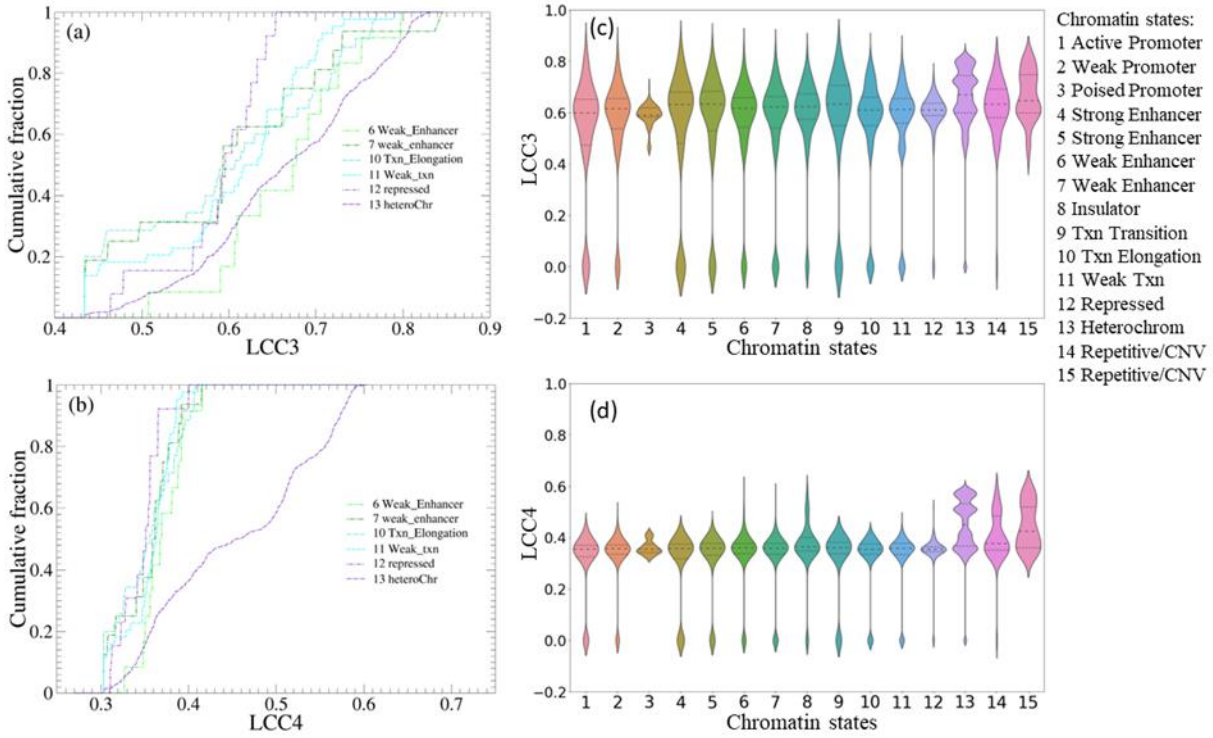


Figure 1. 7 Using network properties discern selected six chromatin state labels of chr21 GM12878 for LCC3 (a) and LCC4 (b). The corresponding distribution of network properties of all 15 different chromatin states are shown in (c) and (d). The resolution is 50 kb.

REFERENCES

1. Rowley, M.J. and Corces, V.G. (2018) Organizational principles of 3D genome architecture. Nat Rev Genet, 19, 789-800.
2. Bonev, B. and Cavalli, G. (2016) Organization and function of the 3D genome. Nat Rev Genet, 17, 772.
3. Robson, M.I., Ringel, A.R. and Mundlos, S. (2019) Regulatory Landscaping: How Enhancer-Promoter Communication Is Sculpted in 3D. Molecular Cell, 74, 1110-1122.
4. Liu, H., Tsai, H., Yang, M., Li, G., Bian, Q., Ding, G., Wu, D. and Dai, J. (2023) Three-dimensional genome structure and function. MedComm (2020), 4, e326.
5. da Costa-Nunes, J.A. and Noordermeer, D. (2023) TADs: Dynamic structures to create stable regulatory functions. Curr Opin Struct Biol, 81, 102622.
6. Bruckner, D.B., Chen, H., Barinov, L., Zoller, B. and Gregor, T. (2023) Stochastic motion and transcriptional dynamics of pairs of distal DNA loci on a compacted chromosome. Science, 380, 1357-1362.
7. McCord, R.P., Kaplan, N. and Giorgetti, L. (2020) Chromosome Conformation Capture and Beyond: Toward an Integrative View of Chromosome Structure and Function. Mol Cell, 77, 688-708.
8. Corsi, F., Rusch, E. and Goloborodko, A. (2023) Loop extrusion rules: the next generation. Curr Opin Genet Dev, 81, 102061.
9. 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. et al. (2009) Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. Science, 326, 289-293.
10. Zhang, B. and Wolynes, P.G. (2015) Topology, structures, and energy landscapes of human chromosomes. Proceedings of the National Academy of Sciences, 112, 6062-6067.
11. Dixon, J.R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J.S. and Ren, B. (2012) Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature, 485, 376-380.

12. Crane, E., Bian, Q., McCord, R.P., Lajoie, B.R., Wheeler, B.S., Ralston, E.J., Uzawa, S., Dekker, J. and Meyer, B.J. (2015) Condensin-driven remodelling of X chromosome topology during dosage compensation. Nature, 523, 240-244.
13. Zhang, C., Xu, Z., Yang, S., Sun, G., Jia, L., Zheng, Z., Gu, Q., Tao, W., Cheng, T., Li, C. et al. (2020) tagHi-C Reveals 3D Chromatin Architecture Dynamics during Mouse Hematopoiesis. Cell Rep, 32, 108206.
14. Heinz, S., Texari, L., Hayes, M.G.B., Urbanowski, M., Chang, M.W., Givarkes, N., Rialdi, A., White, K.M., Albrecht, R.A., Pache, L. et al. (2018) Transcription Elongation Can Affect Genome 3D Structure. Cell, 174, 1522-+.
15. Belaghzal, H., Borrman, T., Stephens, A.D., Lafontaine, D.L., Venev, S.V., Weng, Z., Marko, J.F. and Dekker, J. (2021) Liquid chromatin Hi-C characterizes compartment-dependent chromatin interaction dynamics. Nat Genet, 53, 367-378.
16. García-Domenech, R., Gálvez, J., de Julián-Ortiz, J.V. and Pogliani, L. (2008) Some New Trends in Chemical Graph Theory. Chemical Reviews, 108, 1127-1169.
17. Atilgan, C., Okan, O.B. and Atilgan, A.R. (2012) In Rees, D. C. (ed.), Annual Review of Biophysics, Vol 41, Vol. 41, pp. 205-225.
18. Amitai, G., Shemesh, A., Sitbon, E., Shklar, M., Netanely, D., Venger, I. and Pietrokovski, S. (2004) Network analysis of protein structures identifies functional residues. Journal of Molecular Biology, 344, 1135-1146.
19. Di Paola, L. and Giuliani, A. (2015) Protein contact network topology: a natural language for allostery. Current Opinion in Structural Biology, 31, 43-48.
20. Doshi, U., Holliday, M.J., Eisenmesser, E.Z. and Hamelberg, D. (2016) Dynamical network of residue–residue contacts reveals coupled allosteric effects in recognition, catalysis, and mutation. Proceedings of the National Academy of Sciences, 113, 4735-4740.
21. Daily, M.D., Upadhyaya, T.J. and Gray, J.J. (2008) Contact rearrangements form coupled networks from local motions in allosteric proteins. Proteins-Structure Function and Bioinformatics, 71, 455-466.
22. Albert, R. and Barabási, A.-L. (2002) Statistical mechanics of complex networks. Reviews of Modern Physics, 74, 47-97.

23. Yao, X.Q., Momin, M. and Hamelberg, D. (2018) Elucidating Allosteric Communications in Proteins with Difference Contact Network Analysis. J Chem Inf Model, 58, 1325-1330.
24. Sethi, A., Tian, J., Vu, Dung M. and Gnanakaran, S. (2012) Identification of Minimally Interacting Modules in an Intrinsically Disordered Protein. Biophysical Journal, 103, 748-757.
25. Kryven, I., Duivenvoorden, J., Hermans, J. and Iedema, P.D. (2016) Random Graph Approach to Multifunctional Molecular Networks. Macromolecular Theory and Simulations, 25, 449-465.
26. Stauffer, D. (1985) Introduction to Percolation Theory. Taylor & Francis.
27. Zhang, B. and Horvath, S. (2005) A general framework for weighted gene co-expression network analysis. Stat Appl Genet Mol Biol, 4, Article17.
28. Hao, D., Ren, C. and Li, C. (2012) Revisiting the variation of clustering coefficient of biological networks suggests new modular structure. BMC Syst Biol, 6, 34.
29. Panditrao, G., Bhowmick, R., Meena, C. and Sarkar, R.R. (2022) Emerging landscape of molecular interaction networks: Opportunities, challenges and prospects. Journal of Biosciences, 47, 24.
30. Pancaldi, V. (2023) Network models of chromatin structure. Current Opinion in Genetics & Development, 80, 102051.
31. Botta, M., Haider, S., Leung, I.X.Y., Lio, P. and Mozziconacci, J. (2010) Intra‐and inter‐chromosomal interactions correlate with CTCF binding genome wide. Molecular Systems Biology, 6, 426.
32. Sandhu, K.S., Li, G., Poh, H.M., Quek, Y.L., Sia, Y.Y., Peh, S.Q., Mulawadi, F.H., Lim, J., Sikic, M., Menghi, F. et al. (2012) Large-scale functional organization of long-range chromatin interaction networks. Cell Rep, 2, 1207-1219.
33. Pancaldi, V., Carrillo-de-Santa-Pau, E., Javierre, B.M., Juan, D., Fraser, P., Spivakov, M., Valencia, A. and Rico, D. (2016) Integrating epigenomic data and 3D genomic structure with a new measure of chromatin assortativity. Genome Biology, 17, 152.
34. Newman, M. (2018) Networks. OUP Oxford.

35. Solovei, I., Kreysing, M., Lanctot, C., Kosem, S., Peichl, L., Cremer, T., Guck, J. and Joffe, B. (2009) Nuclear architecture of rod photoreceptor cells adapts to vision in mammalian evolution. Cell, 137, 356-368.
36. Falk, M., Feodorova, Y., Naumova, N., Imakaev, M., Lajoie, B.R., Leonhardt, H., Joffe, B., Dekker, J., Fudenberg, G., Solovei, I. et al. (2019) Heterochromatin drives compartmentalization of inverted and conventional nuclei. Nature, 570, 395-399.
37. Javierre, B.M., Burren, O.S., Wilder, S.P., Kreuzhuber, R., Hill, S.M., Sewitz, S., Cairns, J., Wingett, S.W., Varnai, C., Thiecke, M.J. et al. (2016) Lineage-Specific Genome Architecture Links Enhancers and Non-coding Disease Variants to Target Gene Promoters. Cell, 167, 1369-1384 e1319.
38. Hoencamp, C., Dudchenko, O., Elbatsh, A.M.O., Brahmachari, S., Raaijmakers, J.A., van Schaik, T., Sedeño Cacciatore, Á., Contessoto, V.G., van Heesbeen, R., van den Broek, B. et al. (2021) 3D genomics across the tree of life reveals condensin II as a determinant of architecture type. Science, 372, 984-989.
39. Gibcus, J.H., Samejima, K., Goloborodko, A., Samejima, I., Naumova, N., Nuebler, J., Kanemaki, M.T., Xie, L., Paulson, J.R., Earnshaw, W.C. et al. (2018) A pathway for mitotic chromosome formation. Science, 359, eaao6135.
40. Kaul, A., Bhattacharyya, S. and Ay, F. (2020) Identifying statistically significant chromatin contacts from Hi-C data with FitHiC2. Nat Protoc, 15, 991-1012.
41. Ay, F., Bailey, T.L. and Noble, W.S. (2014) Statistical confidence estimation for Hi-C data reveals regulatory chromatin contacts. Genome Res, 24, 999-1011.
42. Imakaev, M., Fudenberg, G., McCord, R.P., Naumova, N., Goloborodko, A., Lajoie, B.R., Dekker, J. and Mirny, L.A. (2012) Iterative correction of Hi-C data reveals hallmarks of chromosome organization. Nat Methods, 9, 999-1003.
43. Rao, S.S., Huntley, M.H., Durand, N.C., Stamenova, E.K., Bochkov, I.D., Robinson, J.T., Sanborn, A.L., Machol, I., Omer, A.D., Lander, E.S. et al. (2014) A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell, 159, 1665-1680.
44. Caldarelli, G., Pastor-Satorras, R. and Vespignani, A. (2004) Structure of cycles and local ordering in complex networks. The European Physical Journal B, 38, 183-186.

45. Fronczak, A., Hołyst, J.A., Jedynak, M. and Sienkiewicz, J. (2002) Higher order clustering coefficients in Barabási–Albert networks. Physica A: Statistical Mechanics and its Applications, 316, 688-694.
46. Lind, P.G., González, M.C. and Herrmann, H.J. (2005) Cycles and clustering in bipartite networks. Physical Review E, 72, 056127.
47. Yin, H., Benson, A.R. and Leskovec, J. (2018) Higher-order clustering in networks. Physical Review E, 97, 052306.
48. Das, P., Golloshi, R., McCord, R.P. and Shen, T. (2020) Using contact statistics to characterize structure transformation of biopolymer ensembles. Physical Review E, 101, 012419.
49. Haarhuis, J.H.I., van der Weide, R.H., Blomen, V.A., Yanez-Cuna, J.O., Amendola, M., van Ruiten, M.S., Krijger, P.H.L., Teunissen, H., Medema, R.H., van Steensel, B. et al. (2017) The Cohesin Release Factor WAPL Restricts Chromatin Loop Extension. Cell, 169, 693-707 e614.
50. Kind, J., Pagie, L., de Vries, S.S., Nahidiazar, L., Dey, S.S., Bienko, M., Zhan, Y., Lajoie, B., de Graaf, C.A., Amendola, M. et al. (2015) Genome-wide Maps of Nuclear Lamina Interactions in Single Human Cells. Cell, 163, 134-147.
51. Das, P., San Martin, R. and McCord, R.P. (2023) Differential contributions of nuclear lamina association and genome compartmentalization to gene regulation. Nucleus, 14, 2197693.
52. Ernst, J., Kheradpour, P., Mikkelsen, T.S., Shoresh, N., Ward, L.D., Epstein, C.B., Zhang, X., Wang, L., Issner, R., Coyne, M. et al. (2011) Mapping and analysis of chromatin state dynamics in nine human cell types. Nature, 473, 43-49.
53. Xiong, K. and Ma, J. (2019) Revealing Hi-C subcompartments by imputing inter-chromosomal chromatin interactions. Nat Commun, 10, 5069.
54. Wen, Z., Zhang, W., Zhong, Q., Xu, J., Hou, C., Qin, Z.S. and Li, L. (2022) Extensive Chromatin Structure-Function Associations Revealed by Accurate 3D Compartmentalization Characterization. Front Cell Dev Biol, 10, 845118.
55. Nichols, M.H. and Corces, V.G. (2021) Principles of 3D compartmentalization of the human genome. Cell Rep, 35, 109330.

56. Yaffe, E. and Tanay, A. (2011) Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture. Nat Genet, 43, 1059-1065.
57. Kariti, H., Feld, T. and Kaplan, N. (2023) Hypothesis-driven probabilistic modelling enables a principled perspective of genomic compartments. Nucleic Acids Res, 51, 1103-1119.
58. Dodero-Rojas, E., Mello, M.F., Brahmachari, S., Oliveira Junior, A.B., Contessoto, V.G. and Onuchic, J.N. (2023) PyMEGABASE: Predicting Cell-Type-Specific Structural Annotations of Chromosomes Using the Epigenome. J Mol Biol, 435, 168180.
59. Fatima, U., Hina, S. and Wasif, M. (2023) A novel global clustering coefficient-dependent degree centrality (GCCDC) metric for large network analysis using real-world datasets. Journal of Computational Science, 70, 102008.
60. Malod-Dognin, N., Pancaldi, V., Valencia, A. and Pržulj, N. (2020) Chromatin network markers of leukemia. Bioinformatics, 36, i455-i463.
61. Su, J.H., Zheng, P., Kinrot, S.S., Bintu, B. and Zhuang, X. (2020) Genome-Scale Imaging of the 3D Organization and Transcriptional Activity of Chromatin. Cell, 182, 1641-1659.e1626.
62. Thibodeau, A., Márquez, E.J., Shin, D.-G., Vera-Licona, P. and Ucar, D. (2017) Chromatin interaction networks revealed unique connectivity patterns of broad H3K4me3 domains and super enhancers in 3D chromatin. Scientific Reports, 7, 14466.
63. Ernst, J. and Kellis, M. (2017) Chromatin-state discovery and genome annotation with ChromHMM. Nat Protoc, 12, 2478-2492.
64. Abdulla, A.Z., Salari, H., Tortora, M.M.C., Vaillant, C. and Jost, D. (2023) 4D epigenomics: deciphering the coupling between genome folding and epigenomic regulation with biophysical modeling. Current Opinion in Genetics & Development, 79, 102033.

APPENDIX

Supplementary Materials

Besides the SI figures, we have added examples of network analysis for calculating chromosome structure network features (.ipynb format) and deposited them in GitHub (https://github.com/JackXu10469/ chromosome-structure-network). They have been tested and can be run on Google Colab. The code is thoroughly annotated which helps users understand and apply it to their own systems.

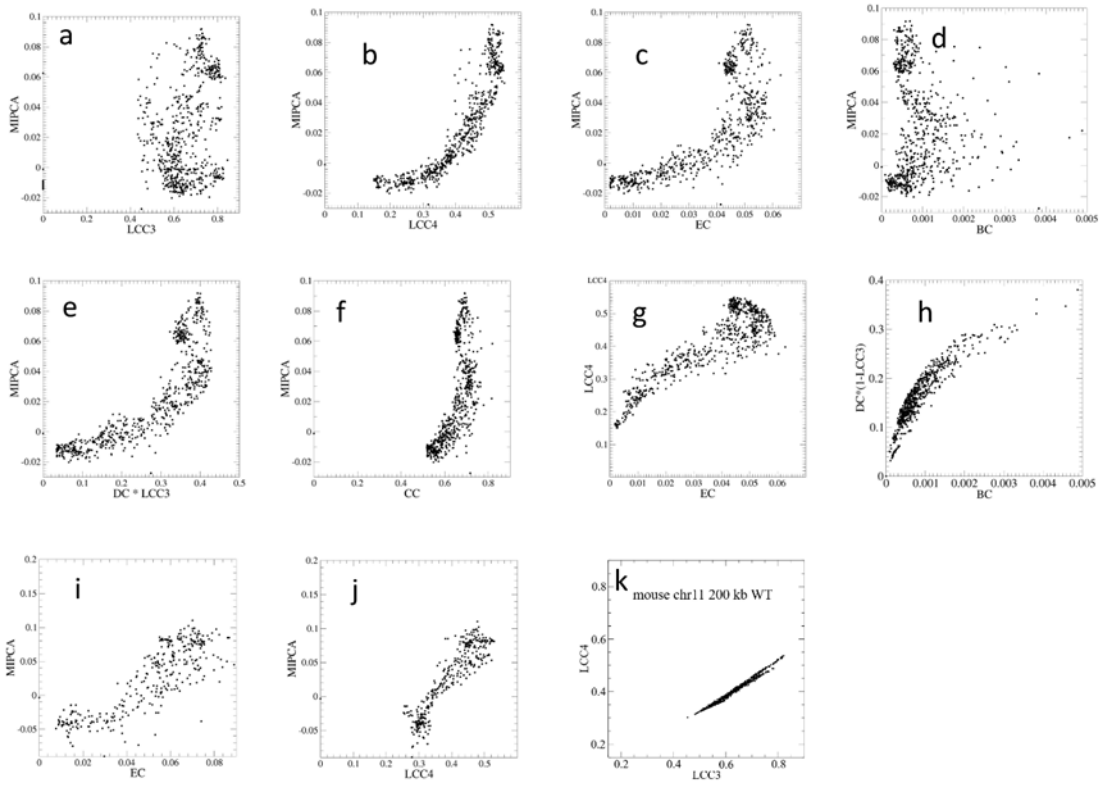


Figure S1. 1 (a–f) Scatter plots of IMR90 chr21 (15.25–48.25 Mb) at 50 kb resolution showing network properties versus MIPCA (A/B compartment). (g–h) Correlation between different network properties using the same system as above. (i–j) Network properties versus MIPCA (A/B compartment) for IMR90 chr10 (0.0–17.3 Mb) at 50 kb resolution. (k) Correlation between LCC3 and LCC4 in mouse thymocyte (WT) cells, chr11 at 200 kb resolution.

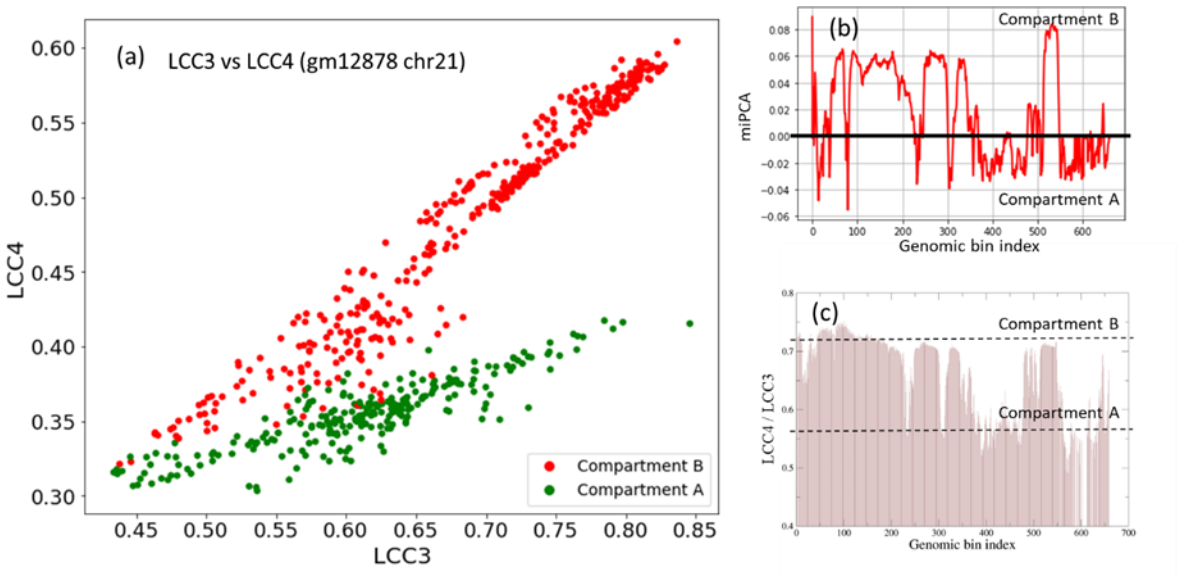


Figure S1. 2 (a) Scatter plot of LCC3 and LCC4 for GM12878 chr21. (b) A/B compartments of GM12878 chr21 classified by MIPCA. (c) Ratio of LCC3/LCC4, with two slope values similar to those in panel (a) marked with dashed lines. All panels use GM12878 chr21 (15.25–48.25 Mb) at 50 kb resolution.

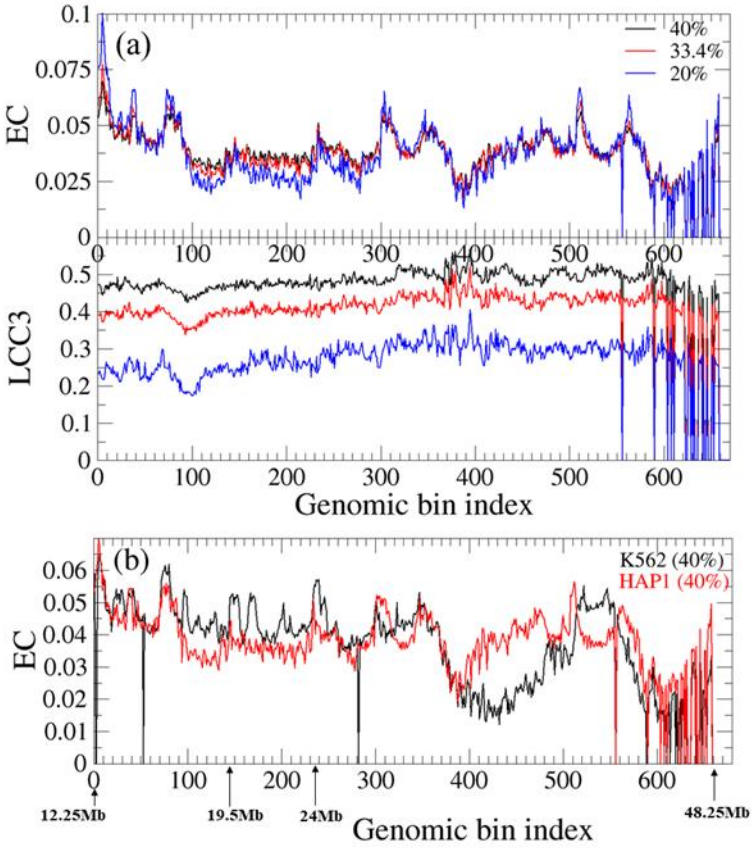


Figure S1. 3 (a) Network properties under different coverage levels for HAP1 cells, chr21 (15.25–48.25 Mb) at 50 kb resolution. (b) EC as a function of genomic node index for K562 compared with HAP1. Chromosome region and resolution are the same as above, with a constant coverage of 40%.

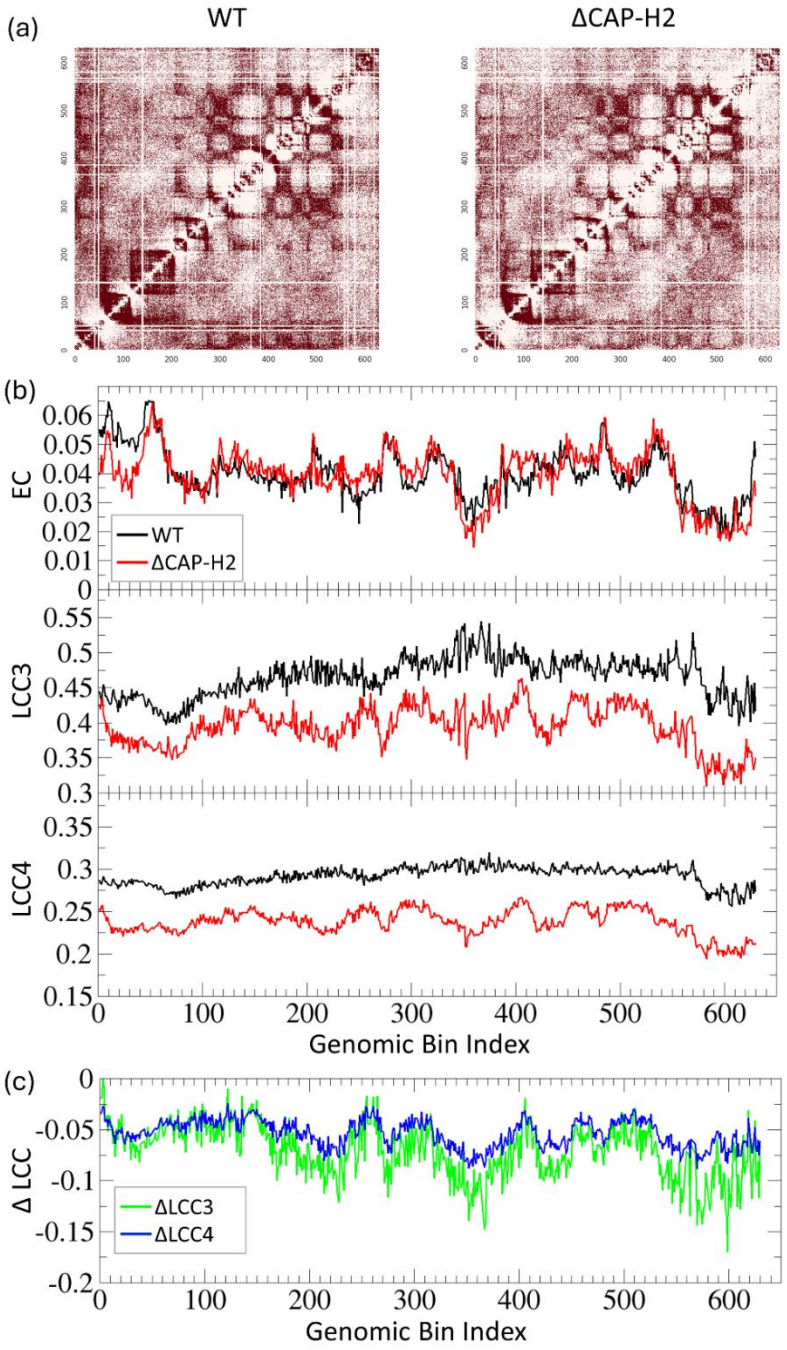


Figure S1. 4 The comparison of chromosome structure networks between HAP1 wild type (WT) and a mutant (DCAP-H2) for chr21: 15.2–46.65 Mb at 50 kb resolution. (a) Contact network constructed both using median value as cutoff. (b) Comparison of EC, LCC3 and LCC4 values. (c) Difference in LCC values (DLCC = $LCC_{DCAP\text{-}H2} - LCC_{WT}$).

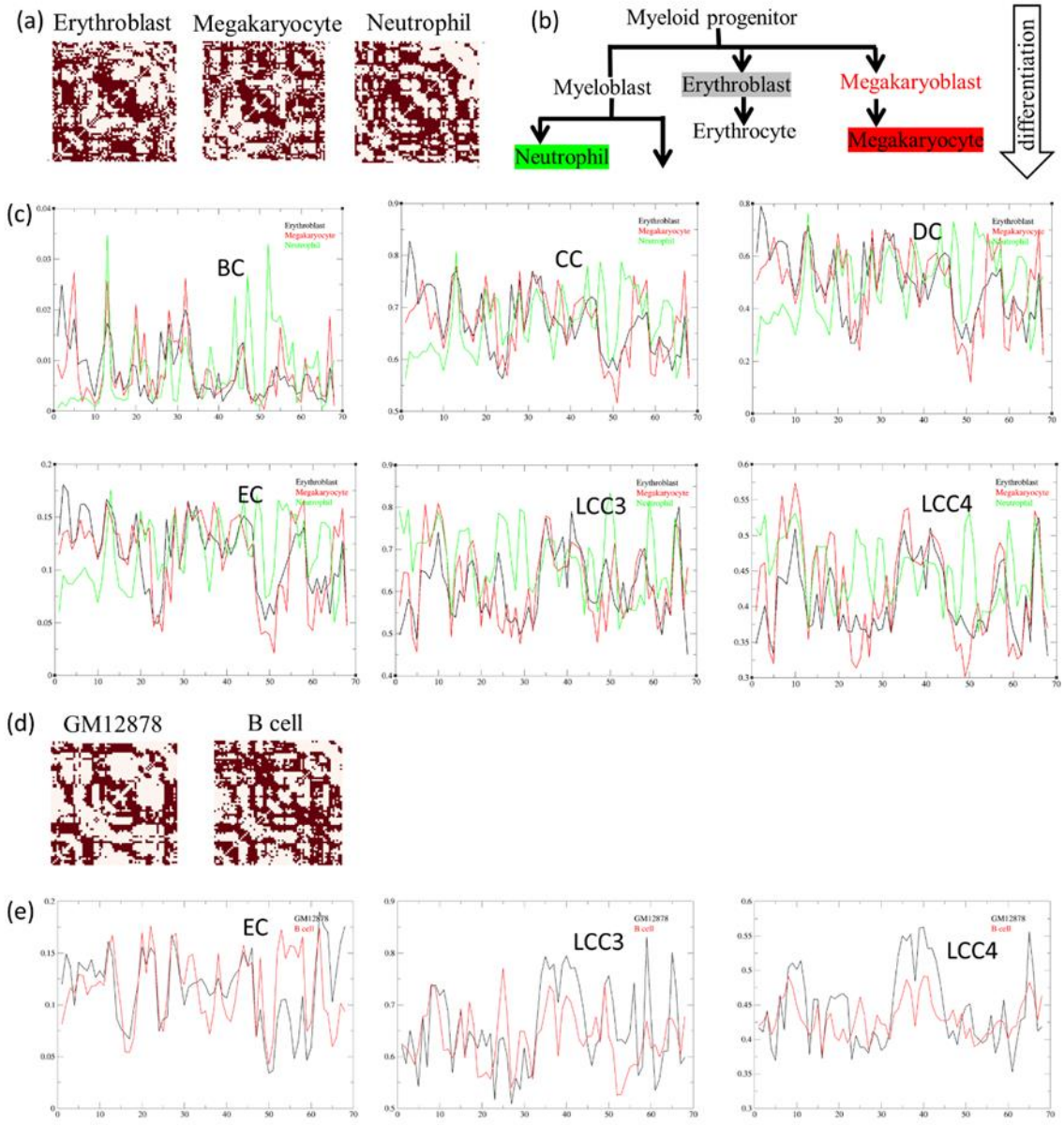


Figure S1. 5 The comparison of contact networks of different human blood cells at different stages of development (chr10: 0–17 Mb, 250 kb resolution): (a) The comparison of the contact network pattern of the three types of cells: erythroblast, megakaryocyte, and neutrophil (coverage at 49%). (b) Blood cell lineage showing the developmental relationship of these cells. (c) The comparison of network properties between these three cell types (color-coded: black, erythroblast; red, megakaryocyte; green, neutrophil). (d-e) The comparison between a pair of "close relative" cell types using B-cell and GM12878. The chromosome region, resolution and coverage are the same as in panel (a).

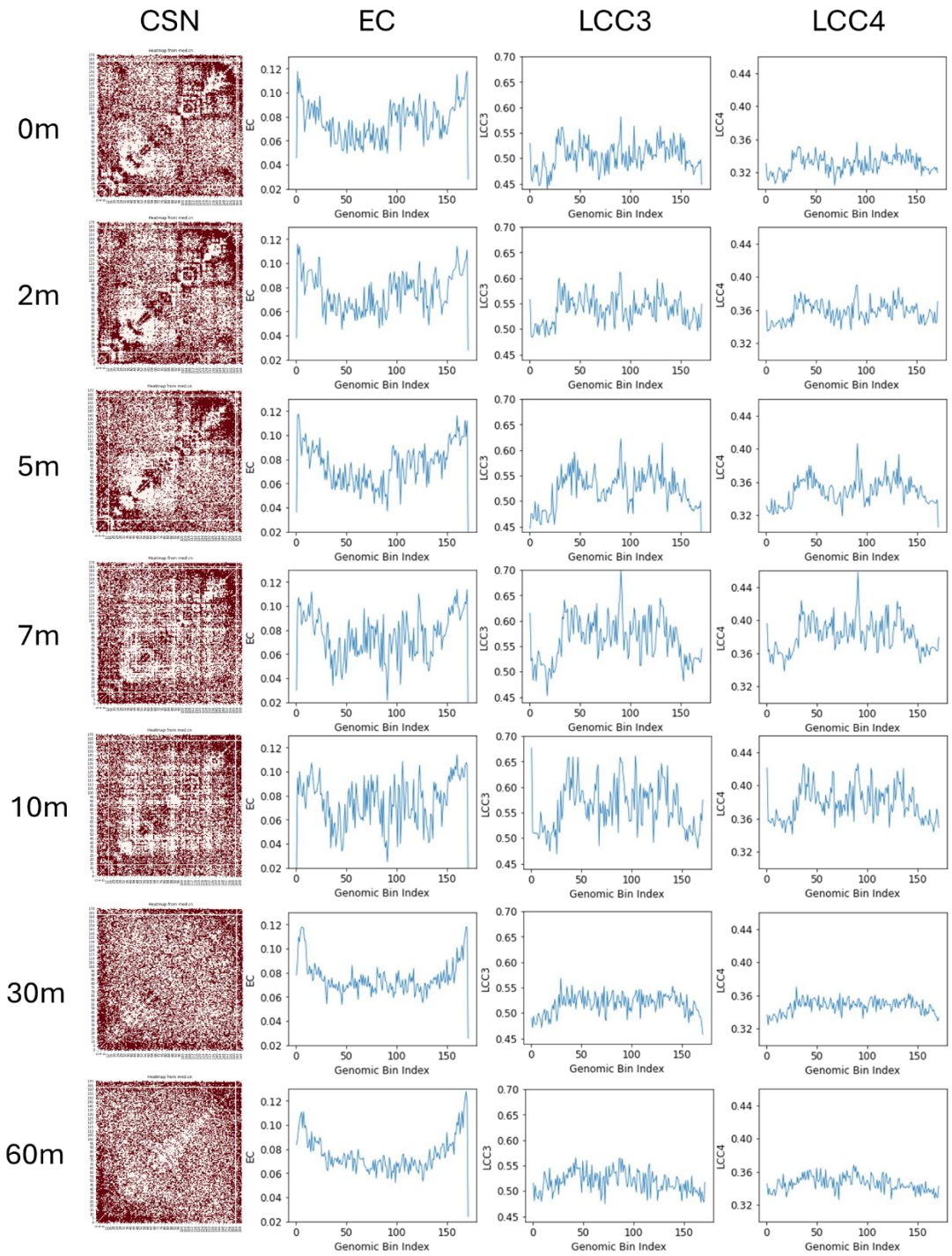


Figure S1. 6 Panel 1/2: The contact matrices and chromosome structure network properties (EC, LCC3, and LCC4) of chicken chromosome 21 during mitosis. The bin size is 40kb. The medians of non-zero values are used as the cutoff values for CSN. Time points (t = 0, 2, 5, 7, 10, 30, and 60 min) starting from G2 phase to late prometaphase are labeled on the left.

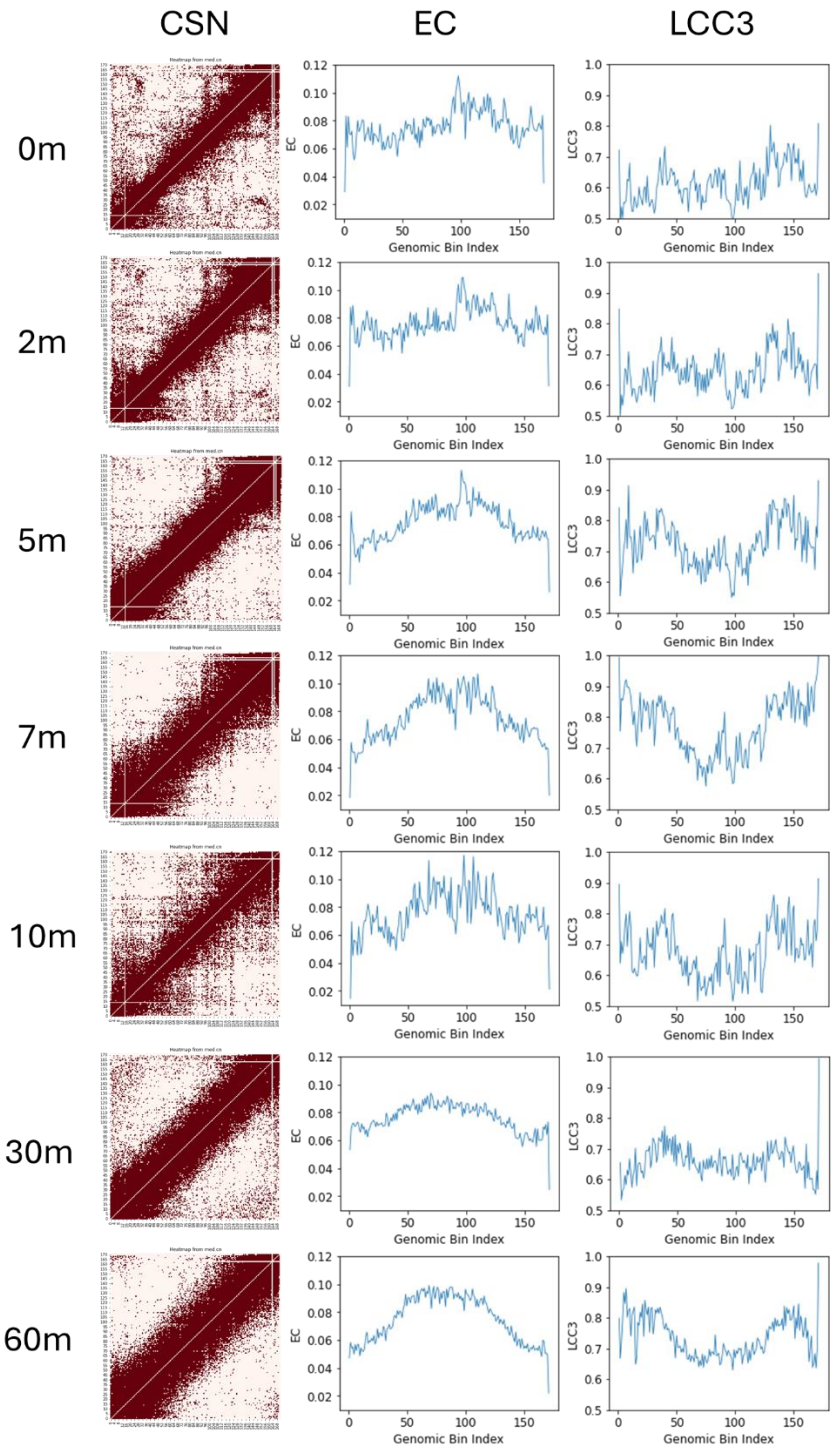


Figure S1.6 Panel 2/2 The corresponding results without distance decay normalization.

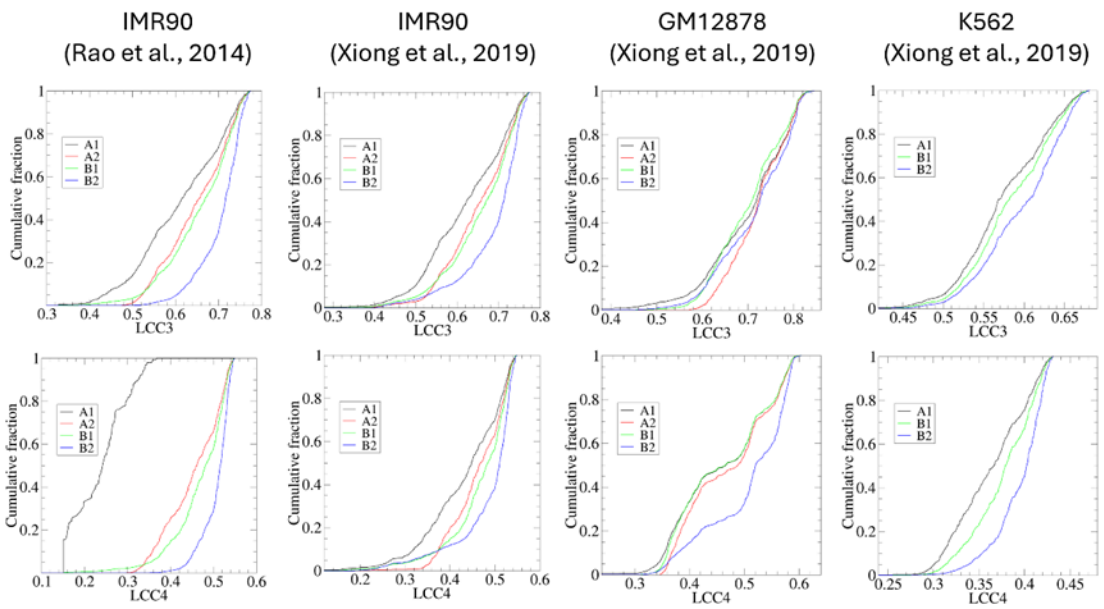


Figure S1. 7 The cumulative fraction curves of chromosome sub-compartment as a function of the LCC values. Here, three cell lines (IMR90, GM12878, and K562) and sub-compartments obtained from different methods (Rao et al., 2014 and Xiong & Ma, 2019) are shown. Here, 15.25–48.25 Mb of chr21 at 50 kb resolution is used.

CHAPTER II

VISUALIZING INTERCHROMOSOMAL INTERACTIONS AT SUB-MEGABASEPAIR RESOLUTION USING NETWORK CLUSTERING COEFFICIENTS

Abstract:

Specific interchromosomal interactions indicate direct and nonrandom physical interactions between pairs of genome positions on two different chromosomes. These contact interactions are direct communication between non-homologous chromosomes and can enable coordinated activities. It is useful to annotate these complex contact interaction patterns and render them to a property associated with a single genome position, both for a clean visualization of the patterns and for facilitating the comparison with linear genomic annotations and underpinning biological functions. We utilize abstract graphs to characterize interchromosomal interaction, as network analysis may succinctly summarize complex interaction structures. We built a graph representation of cross-chromosomal contact interactions derived from Hi-C data and implemented three network-based annotations which consistently indicate the interchromosomal interaction strength associated with specific genomic positions. Equipped with these metrics, we further investigated whether interchromosomal interactions are organized through shared structural hotspots. We found that half of the strong interaction positions of chromosome 19 are shared for interacting with chromosomes 17 and 22. We also discovered that positions with lamina-associated domain (LAD, propensity annotation for nuclear envelop association) tend to be shielded from cross-chromosomal interactions at the active gene region, which is complementary to the previous results of LAD and internal interactions indicating attraction between LAD and the inactive gene region of chromosomes. Overall, the network-based annotation framework reveals distinct chromosome regulation patches and provides insight into how chromosomes associate with each other and organize with respect to the nuclear envelope.

Introduction

The spatial organization of chromosomes within the cell nucleus may affect multiple aspects of gene regulation and determination of cell fate (157-159). Interactions between chromosomes play an important role in interchromosomal communication and cross-chromosomal regulation of genes (160-163). Interchromosomal interactions, also known as transchromosomal interactions (trans contacts), are defined as interactions between different chromosomes (161) whereas the more numerous internal contacts are termed cis contact interactions.

Although trans contacts happen far less frequently than cis contacts, they are not random and can contain specific interactions (164-166) which may assist researchers decode how chromosomes are spatially organized and how they communicate with each other (24,161,167). Interestingly, imaging techniques reveal that chromosomes from different species at specific points of the cell cycle have distinct contact interaction patterns, ranging from Rabl-like architecture to chromosomal territory (168). During interphase, one class of highly regulated chromosomes occupy a localized area within the nucleus known as the chromosomal territory (CT) (19,169). These territories reduce the nonspecific interactions between chromosomes, helping preserve the integrity and transcriptional activity of each chromosome (19,170). The spatial configuration of these territories both enables and restricts interactions between specific regions across different chromosomes (170-172).

At a higher resolution, interchromosomal interactions can be analyzed using advanced techniques such as Hi-C, a powerful method to explore interchromosomal interactions across the entire genome (173). Hi-C is a chromosome conformation capture technique that measures the interactions between chromosomal regions within the nucleus at a resolution of up to 1 kb (135). This method can quantify the frequency of physical association between regions, creating a detailed map of cis and trans contacts (9). Substantial efforts have been directed towards developing algorithms and computational models to extract information contained in the massive datasets generated by Hi-C and other technologies (174-179). These computational approaches are essential for

identifying key interaction hubs, uncovering functional connections between chromosomal regions, and elucidating the patterns of nuclear organization (133,180). By integrating experimental data and computational models, researchers can infer the biological implications of interchromosomal interactions and gain insights into processes such as transcription regulation, DNA replication, and repair (7,181). A key area of data analysis involves rendering the 2D Hi-C contact matrix into 1D linear annotations along the genome. One widely adopted method is a principal component analysis (PCA)-based A/B compartment analysis, where the sign of the compartment strength (the elements of the top eigenvector) is often associated with the active (A) versus inactive (B) genomic region (9). This PCA-based approach primarily focuses on revealing one global feature of intrachromosomal organization. However, extending these methods to analyze both intra- and interchromosomal contacts introduces challenges, such as imbalances in cis vs trans contacts. One can drastically reduce the resolution and study all interchromosomal contacts using the whole chromosome as a unit, as shown in Ref (169). Other combined linear and nonlinear metrics may better capture these contact interactions at a resolution of megabase or higher. In this work, we examine network metric-based features.

Abstract graph layouts (182-185) and network analysis (126,186,187) are powerful mathematical tools for characterizing various complex structures including biological structures. Here, sophisticated interactions between regions of genome are abstracted as links (edges) between nodes (vertices) (188,189). By discretizing spatial connections, quantifying contact interaction strengths, and constructing graph representations of chromosomal structure and interaction, researchers can apply a wide range of network properties to elucidate subtle biophysical and geometrical features contained in genome organization (42). Both local features (such as degree centrality, eigenvector centrality, local clustering coefficients, (126), and assortativity) and global properties (such as modularity and topological weights) of a network (190) can be used to characterize of the genome organization.

In this work, we focus chromosome structure network (CSN), which designates the linear genome units (such as transcription units) as nodes and the contact interaction between units (such as those indicated by Hi-C data) as “links”, as pointed out by several previous

work (122,189,191). Network models of genome structures at other resolutions also exist and each model emphasizes different aspects of genome organization. For instance, in a polymer network model at kilobase resolution, the links symbolize the physical connectivity and topology of dsDNA (190). In contrast, CSN highlights the noncovalent interaction aspects of the genome organization. A notable application used multigraph representation of interchromosomal interactions in plant cells to demonstrate Rabl-like configuration and the association between centromeres and distal regions of chromosomes interact (189). Another study analyzed assortativity and its relationship with specific chromatin states at finer resolution, focusing primarily on interactions within a single chromosome (125). In this study, we aim to have a quantitative description of the interchromosomal interaction at a resolution of 250 kb using clustering coefficients. Local clustering coefficients (triangle-based clustering coefficient C3 and square clustering coefficient C4) capture many-body interactions by measuring whether three to four nodes of the network form tightly connected clusters of multilateral (trilateral and quadrilateral) interactions.

The earlier network model of chromosomal structure (derived from cis contact only at 50 kb) (192) has already uncovered a set of intriguing properties unseen previously for single chromosomes. As shown in Figure 2.1, a CSN model of chromosome (based on chromosome 21) contains three parts: a core and two ends. The (network) core contains "urban" nodes that are high in degree centrality, betweenness centrality, closeness centrality, but low in clustering coefficients such as C3 and C4. "Urban" nodes have more neighbors, but they are not connected to each other. In contrast, the "suburban" nodes at the two ends are low in number of neighbors but high in connectivity as shown in higher clustering coefficients. Like a Janus particle, two ends of the structure have different properties. The "stem" end largely contains nodes of inactive genes manifested by high heterochromatin contents while the other end, the "cap", contain active nodes, low in C4, but high in C3. The stem also contains nodes that are interacting with nucleic envelop.

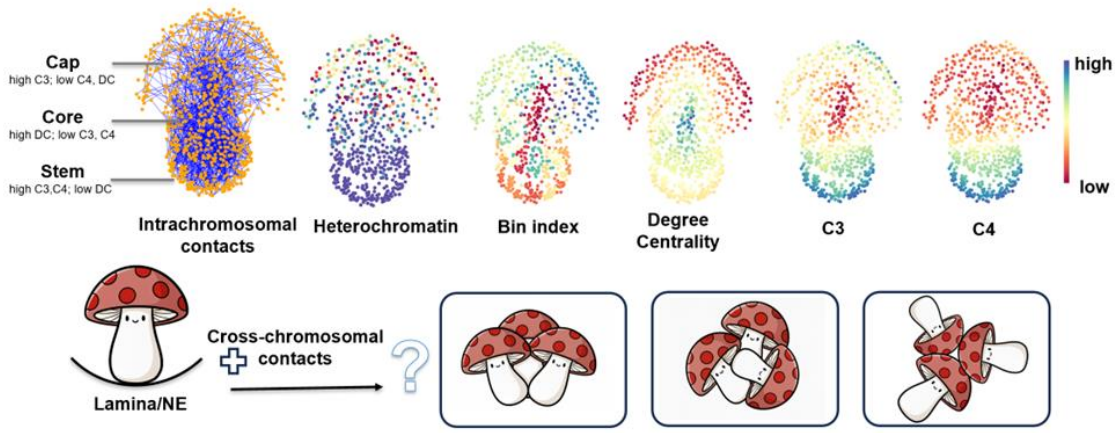


Figure 2. 1 A simple network display of contacts within chromosome 21 (GM12878) Only the Hi-C interaction resolved region (15.25-48.25Mb) is used at the 50-kb resolution (orange dots). Here the top 40% (ranked by strength) of the contacts are shown (down sampled to 1% for a better visualization). Collectively, network information (node-based network centralities and clustering coefficients) provides a picture of structure organization of single chromosomes. With cross-chromosomal contacts being considered, network analysis can assist us to understand the interaction patterns between chromosomes.

The present network model (including trans contact) represents a significant extension of the network approach to capture communication between chromosomes. Our goal is to use node-based network properties to identify key features of contact interactions occurring between different chromosomes. Because the interchromosomal subnetwork forms a bipartite network (138,186), the C3 metric cannot be directly applied to quantify node-level interaction strength. Instead, we show below that it is possible to develop one direct and two indirect approaches to characterize interchromosomal interactions. Applying abstract graph and network theory can also bridge the detailed structural features of chromosome territories to specific chromatin state labels (144) and other biophysical annotations such as lamina-associated domains (LADs) (193). Previously, using cis contact network, we have reported that semi-local network properties such as C4 are suitable descriptors that correlate with certain biological features and annotations, such as chromatin states and LAD (192). The effectiveness of C3 and C4 suggests that biologically relevant interactions can involve three- or four-body contacts, meaning proximity to three or four genomic regions is key for certain biological features. In

contrast, metrics such as eigenvector centrality and A/B compartment classifications rely on semi-long or long-range, many-body interactions. While these may provide stronger global indicators of overall structural organization, they can be less responsive to subtle yet biologically meaningful structural variations (194). When we extend the network to include trans contacts, the model can deepen our understanding of how chromosomes are organized and reveal the underlying patterns that govern interactions between genomic regions of different chromosomes. Applying abstract graph and network theory can also bridge the detailed structural features of chromosome territories to specific chromatin state labels (144) and other biophysical annotations such as lamina-associated domains (LADs) (193). Previously, using cis contact network, we have reported that semi-local network properties such as C4 are suitable descriptors that correlate with certain biological features and annotations, such as chromatin states and LAD (192). The effectiveness of C3 and C4 suggests that biologically relevant interactions can involve three- or four-body contacts, meaning proximity to three or four genomic regions is key for certain biological features. In contrast, metrics such as eigenvector centrality and A/B compartment classifications rely on semi-long or long-range, many-body interactions. While these may provide stronger global indicators of overall structural organization, they can be less responsive to subtle yet biologically meaningful structural variations (194). When we extend the network to include trans contacts, the model can deepen our understanding of how chromosomes are organized and reveal the underlying patterns that govern interactions between genomic regions of different chromosomes.

In this study, we use a resolution of 250 kilobase pair (kb), where each genomic bin corresponds to a single node in the chromosome structure network. Because chromosome structures are hierarchical structure, the corresponding network at one resolution might have structure patterns that are not visible at another (195,196). The 250 kb resolution represents a balance between maintaining sufficient contacts for reliable analysis and avoiding excessive averaging over distinct genes—on average, each node spans only a few genes, since human gene density is about 15 genes per megabase.

We demonstrate newly developed metrics for the communications via cis contacts between chr17, chr19, and chr22, which has been previously reported to exhibit high

frequency of interactions (37,169). We limit the analysis solely between pairs of chromosomes, e.g., the chr17 and chr19, and construct each network in a pairwise fashion as shown in Figure 2.2. This numerical simplification is used in place of directly modeling the multi-chromosomal contact network. However, it is straightforward to generalize this network approach to treat cases where three or more chromosomes form biologically important contact interaction hubs. The pairwise results can serve as a basis for testing how well the shapes of individual chromosomes are defined, and the extent of physical constraint and interaction that promote or inhibit the occurrence of three-body or many-body loci, i.e., simultaneous contacts between three or more chromosomes (30,32). It also should be noted that because our network is constructed from mean contacts, the direct multi-locus interactions observed in single-cell Hi-C and other related single-molecule methods (135,197) are not present in this work.(135,197) are not present in this work. Nevertheless, metrics like clustering coefficients still provide valuable insights into the multivalency of contact interactions through a "mean-field" framework based on bulk Hi-C data. Extending beyond the current bulk Hi-C framework, the network analysis on single-cell Hi-C data may provide methods to study the dynamic correlation of contacts.

Method and systems

Abstract graphs represent the essence of structural components of a complex structure and the interactions among components. Generally, such graphs consist of two elements: "node" and "link." In graph theory, these are referred to as "vertex" and "edge," while in the physical sciences, they are known as "sites" and "bonds." In this study of chromosomal structures, each node represents a chromosomal position, which is also called a genomic bin. Each node has a default size of 250 kilobases (kb), unless otherwise stated. The slightly large bin size is chosen as the trans contact is sparse. Converting from Hi-C heat map to network is equivalent of transforming contact matrix N_{jk} (where each element represents contact strength between nodes j and k) to adjacency matrix A_{jk} (where elements being Boolean). The links between nodes denote spatial proximity between chromosomal positions, corresponding to "contact hits" identified in Hi-C experiments. Such link formation can be defined based on a threshold of specific contact strength or frequency measured from Hi-C data, that is, $A_{ij} = 1$ if $N_{ij} > N_C$ and

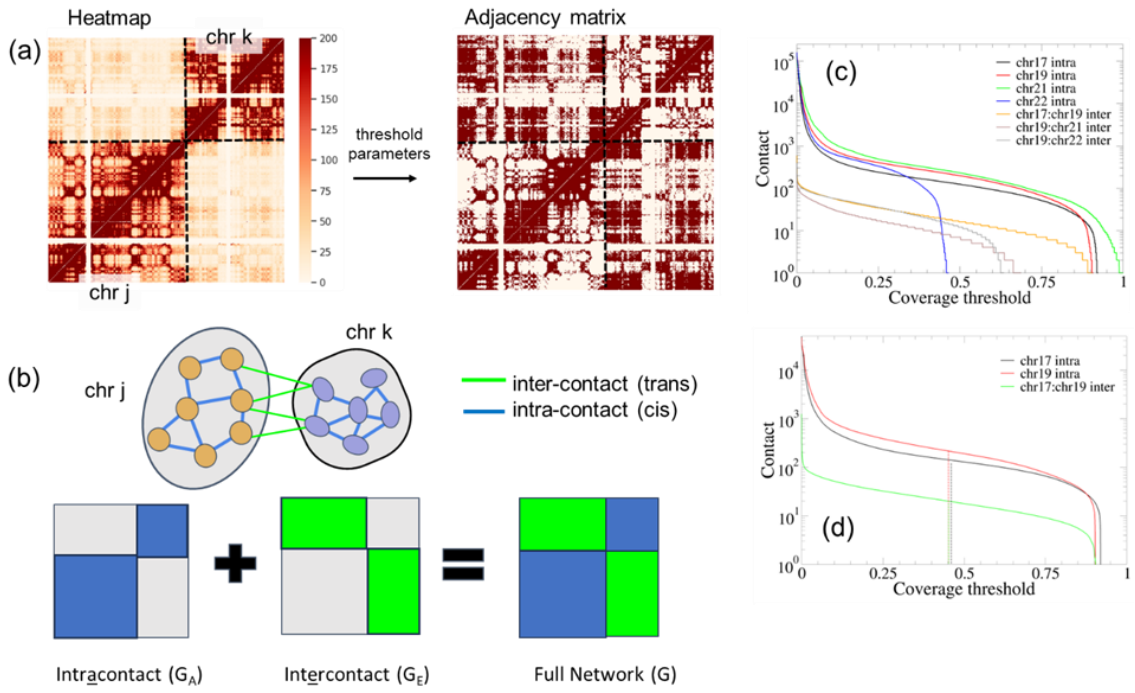


Figure 2. 2 (a) The conversion of contact heatmap of a pair of chromosomes to a chromosome structure network shown. Three individual threshold values are used to ensure internal contact networks of chr j, of chr k, and the interchromosomal network. (b) A cartoon illustration and the corresponding matrix representation of network where the full network can be divided by intra-contact subnetwork (blue, cis-contact links) and inter-contact subnetwork (green, trans-contact links). The green subnetwork is a bipartite network. (c) The ranked contact strength as a function of its ranking (normalized by the size of contact matrix). Here "cis" indicates the intra-chromosomal contact while "trans" indicates the inter-chromosomal contacts. (d) The corresponding ranked ICE balanced contact value shows the inverse of the distribution function for chr 17 and/or chr 19 systems. The dashed line shows the cutoff position for constructing the inter-chromosomal contact network (the median of all non-zero values).

$A_{ij} = 0$ otherwise. However, as was argued previously for the cis contact network, selecting a specific percentage of links, density=50%, to be formed provides more information for the bulk Hi-C contact matrices. Here, the density of the graph is defined as the number of links divided by the maximum possible number of links in the graph. One reason is that the mean contact matrix of chromosomes is highly resolution-dependent and sensitive to experimental procedures. As a result, it is challenging to establish a consistent and transferable threshold of contact strength when transforming a Hi-C contact map into a chromosome structure network (CSN). Instead, we define coverage parameters, which rank relative contact strengths and specify the percentage of links retained from the complete graph. These coverage parameters provide an intuitive connection to information theory and better reflect the characteristics of Hi-C data. This approach emphasizes the most significant and reproducible contact interactions. It is worth noting that other studies have proposed different criteria for defining "significant" contacts, typically to discern noisy interaction from stable ones (132,133). As discussed later in the Results section, our informatic link-percentage criteria are intentionally more stringent in comparison, aiming to minimize numerical noise introduced during network construction.

The link formation is similarly defined as the previous network of cis contacts. Specifically, the threshold of contact interaction strength (after ICE balance (179)) is selected to the median of nonzero elements of the contact matrix. As discussed in the previous work on cis CSN construction (192), this network construction yields a fixed coverage at 50% of maximum possible links, a level that captures the greatest amount of structural information relative a random network measured by Shannon entropy. For the intrachromosomal (cis) contact network, G_A, a single threshold to define link formation. However, for the whole network including G that include both G_A and interchromosomal (trans) contact network, G_E, the raw "hits" between chromosomes are much weaker (roughly only 10%) than the corresponding item within a single chromosome. Therefore, we employ a separate threshold for each subnetwork.

Using a pairwise system, such as the chr17-chr19 interaction (Hi-C data from ref (135)) as an example, the adjacency matrix of the full network contains three components (cis

chr17, cis chr19, and trans chr17-chr19), as illustrated in Figure 2.2. We have defined the full network G joined by two subnetworks G_E and G_A. Graph G_E represents purely direct trans interactions. Because G_E excludes all contacts within a single chromosome, it is a bipartite network. In such a network, such as a reader-book or an actor-film network (138,186,187), nodes are divided into two groups and all the links are formed between these groups and not within them.

Bipartite networks do not contain cycles of odd network lengths, resulting in strict zero values for all C3-based (triangle based) clustering coefficient. However, bipartite networks can still be characterized by even-length cycles, such as those quantified by the square clustering coefficient (C4). To incorporate triangle-based clustering information, we construct a reference graph G_A, containing only internal (intrachromosomal) interactions, and a full graph $G = G_E \cup G_A$. Both G_A and G yield nonzero C3 values. The difference between them presents the effect of G_E measured by triangle-based clustering coefficients. Finally, we denote clustering coefficient of the full network G as C3, that of G_A as C3A, and that of G_E as C3E. Similarly, the C4 values for G, G_A and G_E are denoted as C4, C4A, and C4E, respectively. We can then define $\Delta C3 = C3 - C3A$, representing an alternative measure of contact interaction. Additionally, we can also calculate $\Delta C4 = C4 - C4A$. In total, three node-based metrics, C4E, ΔC3, and ΔC4, were used to evaluate contact strength at the genomic-bin resolution. Among these, C4E directly quantifies interchromosomal contacts, whereas ΔC3 and ΔC4 capture cis-contact–mediated trans interactions (CMTI).

The rationale for studying chromosomes 17 and 19 is that they appear to play the most important roles in interchromosomal communication. Human chr19 has the highest gene density followed by chr17 being the second highest. In GM12878 cells, chr17 exhibits the strongest trans contact interactions, followed by chr19. The interaction between chr17 and chr19 is stronger than that of any other pairs. Therefore, our analysis centers on the interaction network formed between these two chromosomes. Previous contact PCA results has also demonstrated that the dominant eigenvector (PC1) has large loading from chromosomes 17, 19, 22, and 1. (169) The second dominant eigenvector (PC2) revealed an intriguing anti-synchronous pattern, where chr17 correlates with one group of

chromosomes and chr19 with another distinct group. Chromosomes that frequently contact chromosome 17 (indicated by positive PC2 values) tend to interact less with chromosome 19, and vice versa.

In this study, we examine several key linear biological annotations including lamina-associated domains (LADs) (198) and chromatin states (199). LADs are large chromatin regions anchored to the nuclear lamina, typically corresponding to transcriptionally repressed zones that influence chromosomal positioning and functional organization within the nucleus. Chromatin states, defined by distinct patterns of histone modifications, mark genomic regions of active or repressed gene expression and provide insight into regulatory mechanisms across the genome. Understanding the connection between these annotations and trans interaction strength is crucial for unraveling the complex network of chromosomal interactions that maintain cellular function and genomic integrity.

Results

Construction of networks using Hi-C contact maps.

The data of this study came from the Hi-C sequencing dataset from human lymphoblastoid cell line (GM12878) reported by Rao et al. (2014) at 250-kb resolution (135), unless otherwise specified.

As described in the methodology, two main factors influence link selection: (1) each contact must be statistically meaningful (i.e., exceed a significance threshold to avoid random "hits"), and (2) when a large fraction of contacts is meaningful, selecting approximately 50% of them minimizes information loss during the transformation from the continuous contact matrix to a discrete adjacency matrix. In this setup (chr17–chr19, 250 kb bin size), we find that a large fraction of elements in the contact matrix contain nonzero values. We first examine a ranking contact function $f(x)$, where x represents the normalized rank index (rank index divided by the total number of contact pairs, sorted from largest to smallest of contact strengths), and f returns the corresponding contact strength (n_{ij}, the number of raw contact hits). Mathematically, this ranking function

corresponds to the inverse of the cumulative distribution function F(n_{ij}), and its derivative yields the probability density of p(n_{ij}).

As shown in Figure 2.2a, ranking contact functions for seven networks are shown: intrachromosomal contact matrices for chr17, chr19, chr21, chr22 and interchromosomal contacts for chr17-chr19, chr17:chr22 and chr19:chr22. For the example of the chr17-chr19 system, the linear sizes chromosomes are N_{17} = 325 and N_{19} =237. Thus, the total number for intrachromosomal contact for chr17 is N_{17}´ (N_{17} -1)/2 = 325´324/2= 52,650 whereas that for chr19 is N_{19} ´ (N_{19} -1)/2=237´236/2 = 27,966. The total interchromosomal contact pairs N = N_{17}´ N_{19} = 77,025. For the cis contacts of chr17, the ranking function drops to zero at x= 0.923, indicating a maximum coverage of 92.3%. At this maximum coverage, even a single Hi-C hit is sufficient to form a link. The corresponding values for internal chr19 and chr17-chr19 are 91.5% and 90.4%, respectively. Both interchromosomal and intrachromosomal contacts exhibit roughly 90% maximum coverage. However, cis and trans contact strengths are quite different. The percentages of strength (represented by total raw contact hits) for cis chr17 contacts, cis chr19 contacts, and trans chr17-chr19 contacts are 55.7%, 42.1%, and 2.2%, respectively. The ratio of intra- vs inter-contact strength is approximately 50:1 for this specific sequencing data. Similar ranking curves can be plotted for chr19 and chr17–chr19. As observed, interchromosomal contact coverage is lower than intrachromosomal coverage, evident from the faster decay of its ranking curve toward zero. Instead of using the raw n_{ij} data directly, we apply ICE normalization (179), which transforms the raw contact counts n_{ij} into balanced values a_{ij}. As shown in Figure 2.2b, the contact strength a_{ij} for the chr17–chr19 system is plotted against the normalized rank index. The cutoff threshold is defined as the median of all nonzero contact values. To ensure that our observations are generalizable, we also studied chr22 (N_{22} = 205) and its interactions with chr17 and chr19.

Assessment of node-based metrics for interchromosomal interaction strength

As described in the Methods section, we constructed three node-based metrics (C3E, ΔC3, and ΔC4) designed to quantify the strength of interchromosomal interactions. We evaluated their ability to capture the strength of interchromosomal interactions (Figure

2.3). Because the interchromosomal network (G_E) is bipartite, C3E values are strictly zero. Thus, only C3 and C3A are nontrivial, whereas all three networks may exhibit nonzero values for C4.

All three metrics for chr17-chr19 are visualized on a force-directed network layout in Figure 2.3a. Additionally, these metrics plotted as functions of the genomic bin are shown in Supplementary Information (SI) Figure S2.1. Further, additional results for the corresponding chr17-chr22 and chr19-chr22 systems are provided in SI Figure S2.2. The network layouts were generated using the Kamada–Kawai algorithm (182) The genomic bin indices were plotted alongside the network to serve as reference coordinates. The strongest interchromosomal communication originates from the p-arm of chr19 and the distal regions of both the p- and q-arms of chr17. As shown in Figure 2.3a, nodes with high metric values coincide with regions exhibiting dense interchromosomal interactions, corresponding to the top 50% of contact strength. For clarity, not all top 50% of interactions are plotted; instead, 1% of these links were randomly selected and are shown explicitly as green lines. When only the top 1% of interactions are considered, there is less correlation between metrics such as C4E and link density (SI Figure S2.3). This observation suggests that the strongest direct interchromosomal contacts rarely participate in extensive multivalent interactions. Scatter plots (Figures 2.3b and 2.3c) illustrate the relationships among three interaction metrics: C4E, ΔC3, and ΔC4. The results reveal strong correlations among the three metrics. Therefore, all three network properties consistently symbolize the inter-chromosomal interaction strength in certain aspects. Deviations from linear correlations among these metrics may reflect specific genomic or structural features, as further demonstrated in Figures 2.4 and 2.5.

Hubs for interchromosomal interactions

After defining the metrics that represent the strength of cis interactions, we address one important question regarding cross-chromosomal interaction: does a given chromosome use the same genomic region(s) to interact with other chromosomes? To explore this, we generated scatter plots to examine whether a chromosome, such as chromosome 19, uses a set of shared nodes to interact with other chromosomes, including chr17 and chr22, as

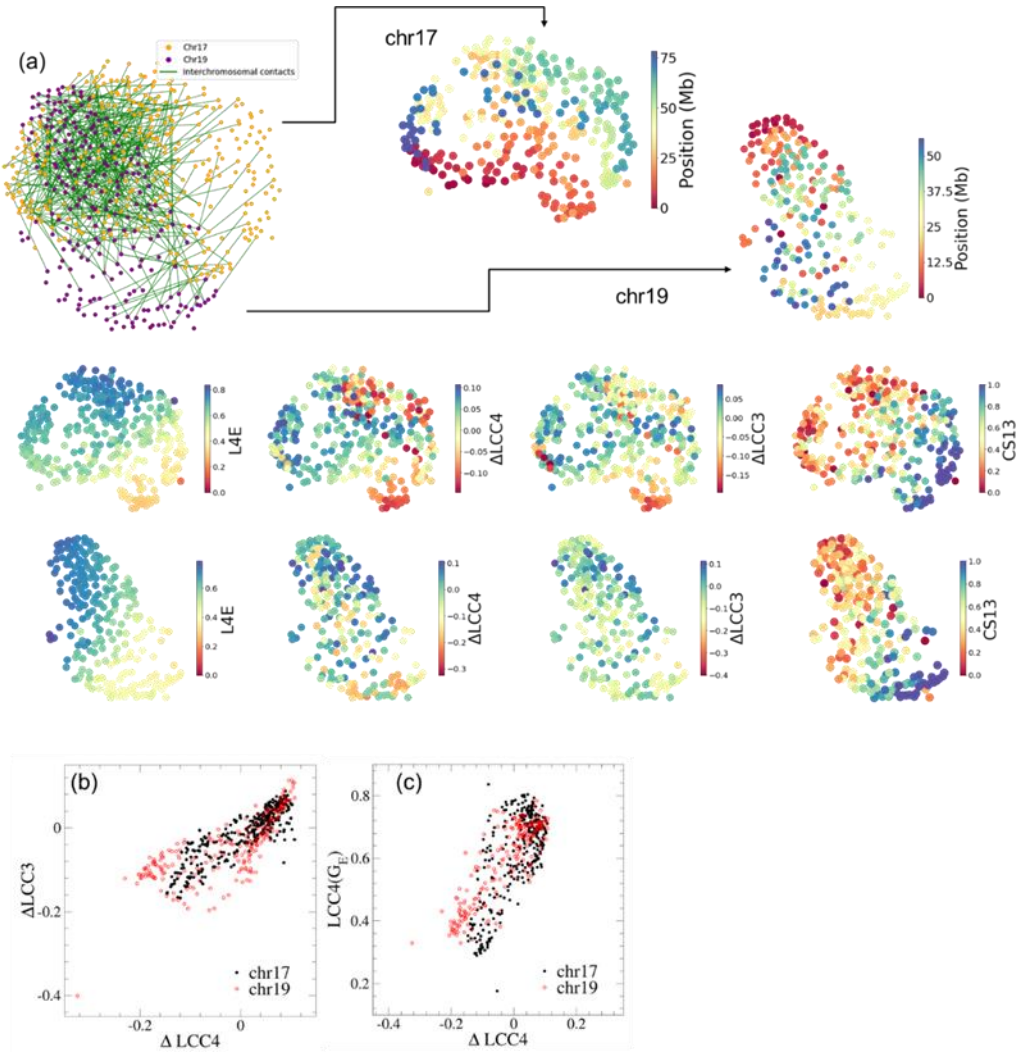


Figure 2. 3 (a) The top left image shows the force distribution layout of nodes in chromosomes 17-19 system, and the other panels show two chromosomes separately. A randomly sampled 1% of the top 50% interactions (edges) are shown to improve visual clarity. The rightward arrows point to chromosomes 17 and 19 with nodes colored by their genomic positions. Below, from left to right, nodes are colored by C4E, ΔC4, ΔC3, and the intensity value of the chromatin state CS13 signal (heterochromatin). (b) Th scatter plot of ΔC4 versus ΔC3 of the same Chr17:Chr19 system. (c) The corresponding scatter plot of ΔC4 versus C4E.

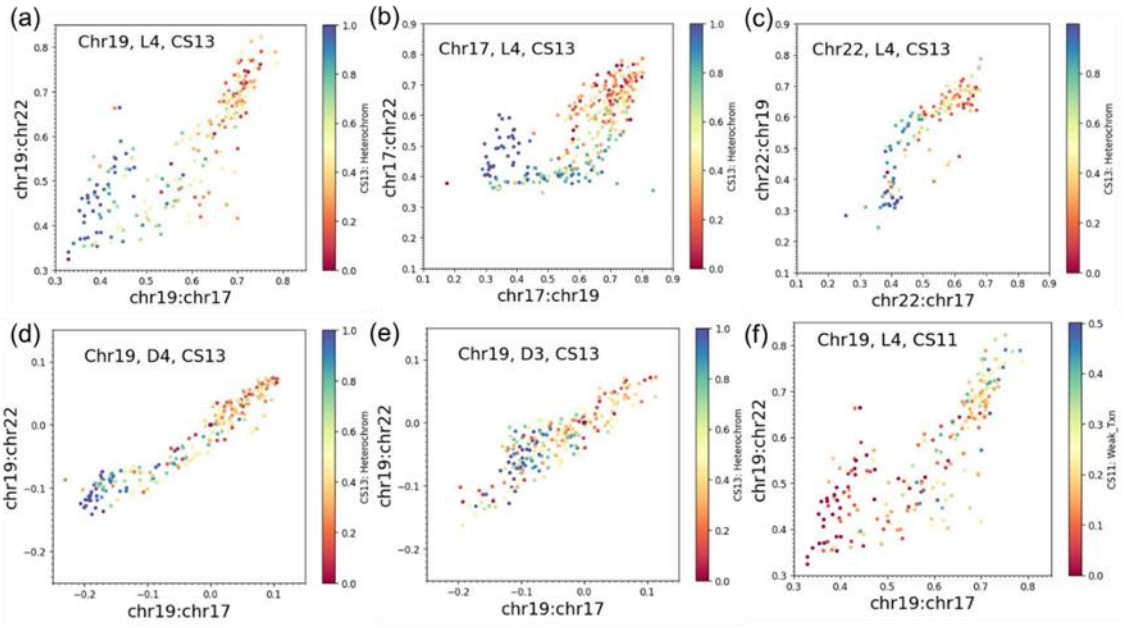


Figure 2. 4 Scatter plots of communication metrics of a specific chromosome colored by chromatin state annotation. (a) The scatter plot shows how chr19 interacts with the two other chromosomes (x-axis with chr17, y-axis with chr22), expressed by C4E and color coded by the chromatin state CS13 (intensity of the Heterochromatin signal of the fragment). Each node represents a 250kb genomic bin from chr19. Similarly, we display other situations with a combination of different chromosome, communication metric, and chromatin state in (b–f).

shown in Fig 2.4. In addition to plotting the interaction strength, we incorporated a third variable, chromatin-state appearance, by color-coding.

The chromatin state appearance, deoted Q_j, for the j-th nodes is defined as follows. We begin with labels of chromatin states assigned by Ernst et al. (ref. (199)) for GM12878 cells, where each position (resolution = 200 bp) of chromosome is assigned a single state CSX (X = 1 to 15). To quantify how frequently represented a specific chromatin state is, we assign j −th node the chromatin state appearance for X-th chromatin state, CSX, as $Q_{i,X}$ which is normalized to a value between 0 and1. Thus, appearance Q of a node is defined as the ratio of the total sequence length where a CS label is detected in a node to the total length of the node (L =250 kb), $Q = L_X/L$. If none of the specified CS intervals fall within the node, Q is strictly zero. We further denote the number of nodes with Q_i=0 as N_o. If $N_o > 90\% \times N_{19}$, we deem the specific CS state does not have enough

representation in chr19, and we omit such a CS from further analysis to avoid spurious correlation between chromatin state labels and network features. For chr19 of cell line GM12878, we found that CS13 (heterochromatin) has the highest presence among 15 CSs (94.9% of the nodes have non-zero values for CS13), followed by CS8 (86.9%) and CS11 (86.5%), whereas CS3 displays the lowest representation, 35.0% (SI Table S1). Thus, for this cutoff of 90%, all 15 CSs have sufficient hits for chr19 at 250 kb resolution.

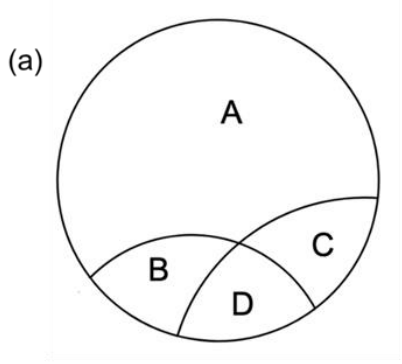
From the scatter plot, the interactions (measured by C4E) of chr19:chr17 and chr19:chr22 generally show a positive correlation. The nodes clustered on the diagonal line represent genomic bins on chr19 that interact similarly with both chr17 and chr22. However, certain off-diagonal features can be observed in the graph. The bins clustered to the upper left of the diagonal line show a stronger tendency to form contact with chr22 and less so with chr17. We observe that the nodes on chr19 that frequently contact chr17 are likely gene active according to the low CS13 appearance (red color nodes) on Fig2.4a. Similar patterns are observed for chr17 and chr22, as shown in Figure 2.4b and 2.4c, respectively. When we use ΔC4 instead of C4E, we also obtained consistent results, as shown in panels 4d and 4e, respectively. Notably, ΔC4 exhibits stronger correlations than C4E. Additionally, ΔC4 more clearly differentiates CS13 while than ΔC3, which lacks a color pattern for the scatter plot. We also show a few other chromatin states, such as CS11 (weak transcription), as shown in Figure 2.4f. The pattern of CS11 shows an opposite trend since CS11 labels active regions.

To quantify the extent to which chromosomal regions are simultaneously engaged in interactions with other chromosomes, we define a concurrence parameter based on the shared nodes with strong inter-chromosomal contacts (Figure 2.5a). Each node was categorized according to its interaction partners as follows: A – a collection of nodes that do not strongly interact with any other chromosome; B, C – a set that interacts with only one other chromosome; D – a set of nodes that strongly interact with both chromosomes B and C simultaneously. The concurrence parameter was calculated as $Concurrence = D / (B + C + D)$. This value represents the proportion of inter-chromosomal interactions that occur concurrently with two partners, rather than in isolation. To

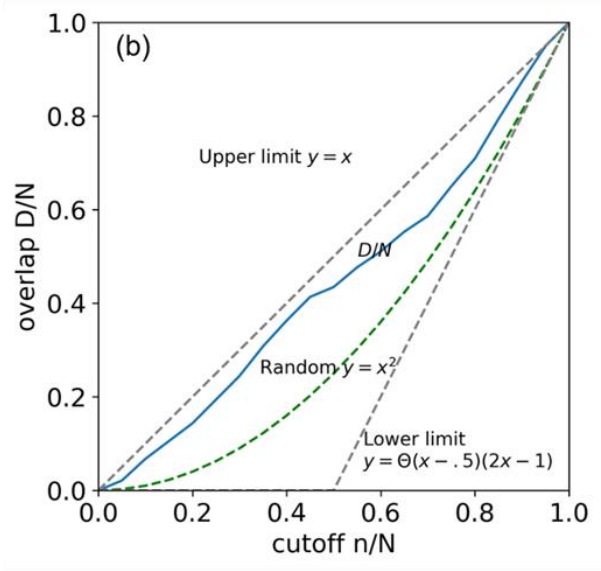
facilitate computation using full interaction sets between pairs of chromosomes, the parameter can also be expressed as $Concurrence = D / (BD + CD - D)$

Taking chromosome 19 as an example, BD is the total number of bins interacting with chromosome 17, CD is the number of bins interacting with chromosome 22, and D is the subset of regions involved in both interactions. Only the top 20% of strongly interacting nodes (ranked by network properties) were included in the analysis to reduce. As shown in Fig. 2.5, when filtering based on C4E values, 56.7% of the regions on chromosome 19 that interact with either chromosome 17 or chromosome 22 are concurrently in contact with both. These regions represent 14.3% of the entire chromosome 19. Figure 2.5b illustrates how these conclusions vary depending on the selected cutoff thresholds. The upper and lower bounds, and the uncorrelated reference plots are also shown. Figures 2.5c–e depict a distinctive spatial organization pattern of chromosome 19 in its interactions with chromosomes 17 and 22. Specifically, several nodes near the telomeric region of the p-arm on chromosome 19 appear to contact both chromosomes 17 and 22, suggesting the presence of shared contact domains that may function as hubs for interchromosomal communication. Extending beyond the current pairwise framework, multi-chromosomal interactions could potentially be detected through clustering coefficient of three chromosomes. For example, one can compare $C3E_h$ of chr19, which records heterogeneous triplet between chr17, chr19, and chr22 of the networks G and that of a corresponding random network G*(187) and inform us if shared nodes of chr19 makes chr17 and chr22 indirectly attracts ($C3E_h > C3E_h$*) or repels at these positions. In addition to these overlapping domains, our results also identify distinct regions on chromosome 19 that interact exclusively with either chromosome 17 or 22. Such exclusivity suggests a spatial and functional partitioning within chromosome 19, where some nodes maintain selective contact while others participate in shared interchromosomal interactions.

Furthermore, we quantified the connection between interchromosomal interaction strength (expressed by a linear fit model of ΔC4 and C4E) and the appearance of CS labels Q using a correlation coefficient r. The correlation coefficient r, bounded between –1 and 1, indicates the degree to which interaction strength (e.g., ΔC4) is correlated with



Top 20% nodes	19:17 vs 19:22	17:19 vs 17:22	22:17 vs 22:19
C4E	56.7% (14.3%)	41.3% (11.7%)	54.7% (14.2%)
ΔC4	74.1% (16.9%)	68.8% (16.4%)	54.7% (14.2%)



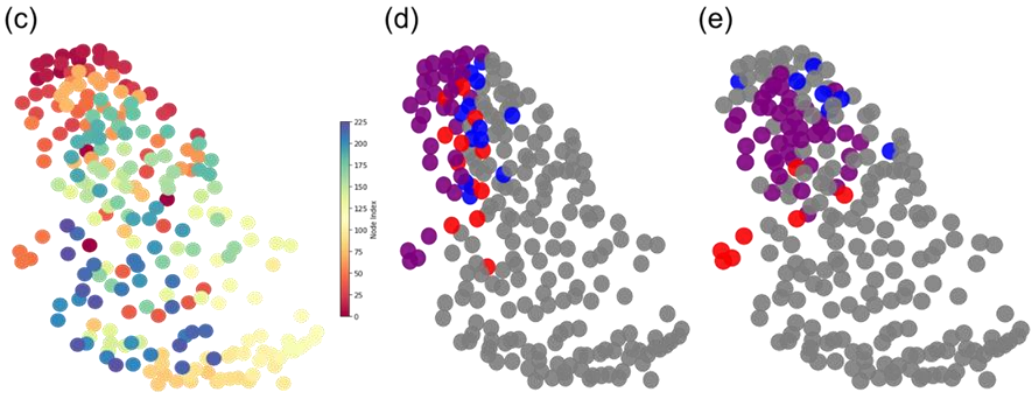


Figure 2. 5 (a) Concurrence parameter shows the percentage of the chromosome nodes that are simultaneously involved in interactions (judged by node-based metrics) with two other chromosomes. Here A indicates a subset of nodes that does not interact with either of two other chromosomes. B and C indicate sets that interact with only one other chromosome. D indicates a set that interacts with both. We define concurrence parameter = D/(B +C +D) = D/(CD + BD – D). The top 20% values of C4E or ΔC4 were selected as the cutoff value for contact. The percentages in brackets indicate the proportion of the D region to the entire chromosome. (b) The normalized overlap as a function of normalized cutoff. (c) Nodes of chromosome 19 isolated from the chromosomes 17-19 layout colored by bin index. (d) Nodes of chromosome 19 that interact with chromosomes 17 and 22 through different regions. Red indicates contact only with chr17. Blue indicates contact only with chr22. Purple indicates contact with both. The top 20% values of C4E were selected as the cutoff value for contact. (e) Same setup as panel (d) except ΔL4 is used as the interaction strength.

specific chromatin states (CSs). We first performed a principal component analysis (PCA) on the scatter plot of ΔC4 versus C4E, where the first principal component (PC1) defines the direction of consensus between the two node-based metrics. Next, we calculated Spearman's rank correlation coefficients between the PC1 projections and the chromatin-state appearance values (Q). Using chromosome 19 as a test case, we found that chromatin states associated with strong transcriptional activity, such as active promoters (CS1) and strong enhancers (CS4 and CS5), exhibited the highest positive correlations. In contrast, weaker regulatory states like poised promoters (CS3) showed lower, though still positive, correlations. Conversely, heterochromatin (CS13) was the only chromatin state showing a negative correlation with the interaction strength metrics (see SI Table S1).

Annotation of LAD and the association of chromosome and nuclear lamina.

In addition to chromatin state labels, lamina-associated domain (LAD) represents another key binary classification. LAD labels where chromosomes are associated with lamina (skeletal and anchoring proteins of the nuclear envelope) (193,200).(193,200). Because lamina association imposes mechanical constraints and defines nucleus morphology, we compare how LAD labels are correlated with network metrics. A previous study of LAD and cis contacts has already revealed that LAD are located at nodes with high C4A values and a high AUROC (area under the receiver operating characteristic curve) supports that C4A being a good feature for LAD classification. Here we would like to explore the connection between LAD and trans contacts.

As LAD classifier provides binary information for each node (either is associated with the nuclear lamina or not), we use cumulative distribution function (CDF) to evaluate the correlation. As shown in Figure 2.6, three sets of network node-based properties are evaluated to examine how well they are connected to the LAD labels. In each panel, a large separation between the CDF plots (black versus red) indicates the metric shown in x-axis is a good classifier for being LAD or not. Specifically, we compared C4A (for Chr21, a replot of the results of Reference (192)), C4E, and ΔC4. Among them, LAD is associated with lower C4E but higher C4A. Another inter-chromosomal strength indicator, ΔC4 , shows same trend as C4E. These findings suggest that chromosomal

regions engaged in interchromosomal contacts are less likely to be located at the nuclear lamina. In contrast, previous analysis of intrachromosomal contact networks using C4A revealed that highly connected nodes are more frequently associated with LADs.

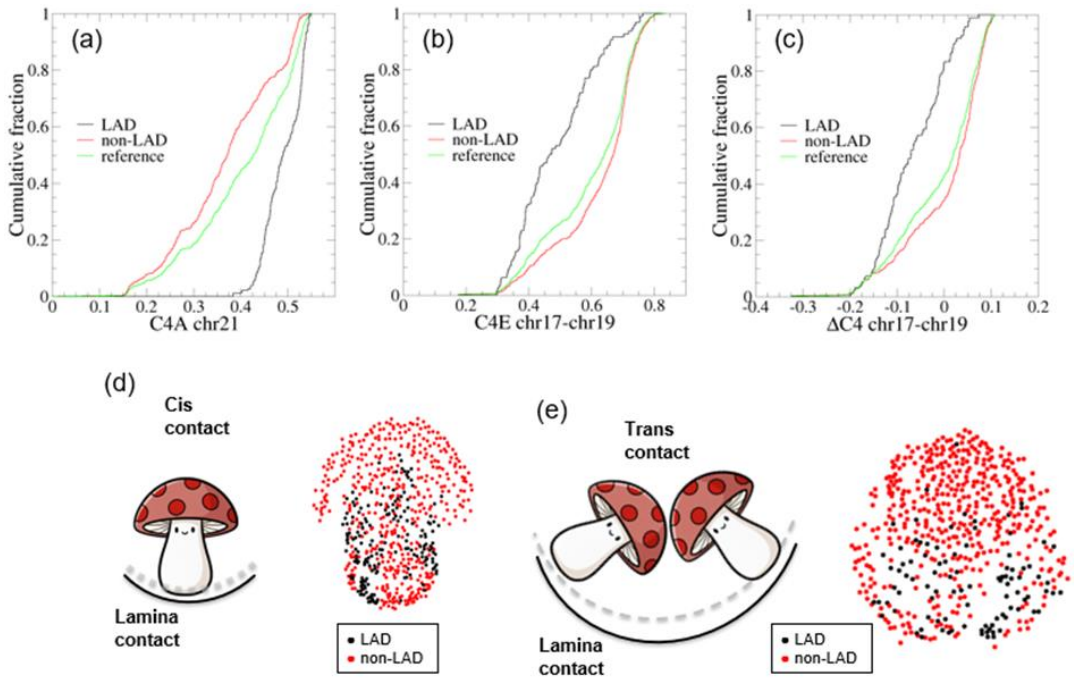


Figure 2. 6 Cumulative distribution fraction (CDF) plots of network metrics for chromosome networks. Here classifier performance concepts using CFP (black), CFN (red), CFR (green) are shown for chr21 using feature C4A in (a), for chr17-chr19 using C4E in (b), for chr17-chr19 using DC4 in (c). interchromosomal network C4 in (d), ΔC4 in (e) and intrachromosomal network C4 in (f).

Effects of condensin II deletion

Equipped with these three metrics, we further investigated how condensin II deletion mutant (ΔCAPH2) of HAP1 cells influences interchromosomal interactions. Condensin II is known for its importance in chromosome looping and compaction (201,202). The absence of CAP-H2 results in extensive changes to chromosome structure and notably affects interactions between chromosomes (203). The previous intra-chromosomal network analysis reveals increased fluctuation of C3 and C4 as functions of bin index (as shown in SI Fig. S2.4b and SI of reference (192)) in mutant cells, indicating region-

specific impacts of condensin II loss on chromosomal connectivity, consistent with the disruption of condensin-mediated loops. In the current study, we extended the analysis to trans interactions. We constructed the chr17–chr19 interaction network using structural data from both wild-type and ΔCAPH2 mutant HAP1 cells during mitosis. The comparison of network properties is shown in SI Figure S2.4, and the network display is presented in SI Figure S2.5.

Our analysis indicated that metrics ΔC3 and ΔC4 do not show substantial overall increases or decreases, and the general features persisted across conditions (SI Fig. S2.4c–e). However, in mutant cells, the structural features, represented by peaks and troughs, were noticeably dampened relative to wild-type cells; peak values were reduced, troughs elevated, and the overall curve appeared flattened. These results suggest that mutant cells exhibit a more homogenized chromosomal interaction pattern.

Compared to the internal changes, the changes between chromosomes presented here are less pronounced. Nevertheless, the result is consistent with the known function of condensin II, which primarily affects the ability of individual chromosomes to fold and organize themselves into compact territories. While its role in inter-chromosomal interactions is relatively smaller, the effects are still noticeable: the mutant ΔCAPH2 exhibits reduced heterogeneity of inter-chromosomal interaction compared to that of HAP1 cells. These findings indicate that condensin II contributes to the maintenance of a dynamic and heterogeneous nuclear architecture. Such a role is consistent with the reported function of condensin II in human cells (204,205). Notably, for the region near the centromere of chr17 (around 25Mb), a peak appeared in the mutant and indicated a possible change of interaction near the centromere. These findings further support the conclusion that condensin II both affects trans and cis interactions (206-208).

Concluding remarks

Through analysis of interchromosomal contact networks, we identified node-based parameters, such as ΔC4, as effective quantitative descriptors for annotating trans contact interactions. We found that interaction strengths, characterized by three metrics (C4E, ΔC3, and ΔC4), were strongly correlated with one another. The metrics exhibited a definitive association with chromatin state labels as well. Notably, heterochromatin

showed a strong negative correlation, whereas several active chromatin states displayed strong positive correlations. In addition, annotation for nuclear lamina affinity were negatively correlated with interchromosomal interaction strength, supporting a spatial chromatin organization model in which chromosome–nuclear periphery interactions suppress interchromosomal contacts—that is, chromosomes engage less with each other near the nuclear periphery. Using the "produce" model, chromosome "stems" interact with nuclear lamina whereas chromosome "caps" interact with each other. When these interaction metrics are applied to the ΔCAPH2 mutant, attenuated specific trans-interactions were observed compared to that of haploid HAP1 cells, which is consistent with the previous report that a reduced internal structural heterogeneity due to the mutation.

Overall, this study establishes a network-based framework for annotating interactions between chromosomes using clustering coefficients, offering new insights into chromatin states and chromosome-nuclear lamina interaction through the lens of interchromosomal transitivity. The methodologies and findings provide a foundation for future investigations into nuclear architecture and its implications in development, differentiation, and disease-related chromosome reorganization.

REFERENCES

1. Finn, E.H. and Misteli, T. (2019) Molecular basis and biological function of variability in spatial genome organization. *Science*, **365**.
2. Fraser, P. and Bickmore, W. (2007) Nuclear organization of the genome and the potential for gene regulation. *Nature*, **447**, 413-417.
3. Stadhouders, R., Filion, G.J. and Graf, T. (2019) Transcription factors and 3D genome conformation in cell-fate decisions. *Nature*, **569**, 345-354.
4. Tomikawa, J. (2024) Potential roles of inter-chromosomal interactions in cell fate determination. *Front Cell Dev Biol*, **12**, 1397807.
5. Zhao, Z., Tavoosidana, G., Sjolinder, M., Gondor, A., Mariano, P., Wang, S., Kanduri, C., Lezcano, M., Sandhu, K.S., Singh, U. *et al.* (2006) Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions. *Nat Genet*, **38**, 1341-1347.
6. Dekker, J., Marti-Renom, M.A. and Mirny, L.A. (2013) Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. *Nat Rev Genet*, **14**, 390-403.
7. Bashkirova, E. and Lomvardas, S. (2019) Olfactory receptor genes make the case for inter-chromosomal interactions. *Curr Opin Genet Dev*, **55**, 106-113.
8. Asamitsu, S. and Iwasaki, Y.W. (2025) Biomolecular liquid–liquid phase separation associated with repetitive genomic elements. *Polymer Journal*, **57**, 785-797.
9. Mehta, S. and Zhang, J. (2022) Liquid–liquid phase separation drives cellular function and dysfunction in cancer. *Nature Reviews Cancer*, **22**, 239-252.
10. Tong, X., Tang, R., Xu, J., Wang, W., Zhao, Y., Yu, X. and Shi, S. (2022) Liquid–liquid phase separation in tumor biology. *Signal Transduction and Targeted Therapy*, **7**, 221.
11. Maass, P.G., Barutcu, A.R. and Rinn, J.L. (2019) Interchromosomal interactions: A genomic love story of kissing chromosomes. *J Cell Biol*, **218**, 27-38.
12. Rowley, M.J. and Corces, V.G. (2018) Organizational principles of 3D genome architecture. *Nat Rev Genet*, **19**, 789-800.

13. Szalay, M.F., Majchrzycka, B., Jerković, I., Cavalli, G. and Ibrahim, D.M. (2024) Evolution and function of chromatin domains across the tree of life. *Nat Struct Mol Biol*, **31**, 1824-1837.

14. Cremer, T. and Cremer, C. (2001) Chromosome territories, nuclear architecture and gene regulation in mammalian cells. *Nat Rev Genet*, **2**, 292-301.

15. Das, P., Shen, T. and McCord, R.P. (2020) Inferring chromosome radial organization from Hi-C data. *BMC Bioinformatics*, **21**, 511.

16. Bickmore, W.A. and van Steensel, B. (2013) Genome architecture: domain organization of interphase chromosomes. *Cell*, **152**, 1270-1284.

17. Cremer, T., Cremer, M., Dietzel, S., Muller, S., Solovei, I. and Fakan, S. (2006) Chromosome territories--a functional nuclear landscape. *Curr Opin Cell Biol*, **18**, 307-316.

18. Misteli, T. (2007) Beyond the sequence: cellular organization of genome function. *Cell*, **128**, 787-800.

19. Belton, J.M., McCord, R.P., Gibcus, J.H., Naumova, N., Zhan, Y. and Dekker, J. (2012) Hi-C: a comprehensive technique to capture the conformation of genomes. *Methods*, **58**, 268-276.

20. Rao, S.S., Huntley, M.H., Durand, N.C., Stamenova, E.K., Bochkov, I.D., Robinson, J.T., Sanborn, A.L., Machol, I., Omer, A.D., Lander, E.S. *et al.* (2014) A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*, **159**, 1665-1680.

21. 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. *et al.* (2009) Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, **326**, 289-293.

22. Wolff, J., Backofen, R. and Gruning, B. (2022) Loop detection using Hi-C data with HiCExplorer. *Gigascience*, **11**.

23. Cameron, C.J., Dostie, J. and Blanchette, M. (2020) HIFI: estimating DNA-DNA interaction frequency from Hi-C data at restriction-fragment resolution. *Genome Biol*, **21**, 11.

24. Highsmith, M. and Cheng, J. (2021) VEHiCLE: a Variationally Encoded Hi-C Loss Enhancement algorithm for improving and generating Hi-C data. *Sci Rep*, **11**, 8880.
25. Lajoie, B.R., Dekker, J. and Kaplan, N. (2015) The Hitchhiker's guide to Hi-C analysis: practical guidelines. *Methods*, **72**, 65-75.
26. Lyu, H., Liu, E. and Wu, Z. (2020) Comparison of normalization methods for Hi-C data. *Biotechniques*, **68**, 56-64.
27. Servant, N., Varoquaux, N., Lajoie, B.R., Viara, E., Chen, C.J., Vert, J.P., Heard, E., Dekker, J. and Barillot, E. (2015) HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. *Genome Biol*, **16**, 259.
28. Ay, F., Bailey, T.L. and Noble, W.S. (2014) Statistical confidence estimation for Hi-C data reveals regulatory chromatin contacts. *Genome Res*, **24**, 999-1011.
29. Carty, M., Zamparo, L., Sahin, M., Gonzalez, A., Pelossof, R., Elemento, O. and Leslie, C.S. (2017) An integrated model for detecting significant chromatin interactions from high-resolution Hi-C data. *Nat Commun*, **8**, 15454.
30. Schmitt, A.D., Hu, M. and Ren, B. (2016) Genome-wide mapping and analysis of chromosome architecture. *Nat Rev Mol Cell Biol*, **17**, 743-755.
31. Bonev, B. and Cavalli, G. (2016) Organization and function of the 3D genome. *Nat Rev Genet*, **17**, 772.
32. Kamada, T. and Kawai, S. (1989) An algorithm for drawing general undirected graphs. *Information Processing Letters*, **31**, 7-15.
33. Fruchterman, T.M.J. and Reingold, E.M. (1991) Graph drawing by force-directed placement. *Software: Practice and Experience*, **21**, 1129-1164.
34. Martin, S., Brown, W.M., Klavans, R. and Boyack, K. (2011) *OpenOrd: an open-source toolbox for large graph layout*. SPIE.
35. Davidson, R. and Harel, D. (1996) Drawing graphs nicely using simulated annealing. *ACM Trans. Graph.*, **15**, 301–331.
36. Newman, M. (2018) *Networks*. OUP Oxford.
37. Zhang, P., Wang, J., Li, X., Li, M., Di, Z. and Fan, Y. (2008) Clustering coefficient and community structure of bipartite networks. *Physica A: Statistical Mechanics and its Applications*, **387**, 6869-6875.

38. Kingan, S. (2022). First edition. ed. Wiley.

39. Watts, D.J. and Strogatz, S.H. (1998) Collective dynamics of 'small-world' networks. *Nature*, **393**, 440-442.

40. Makai, D., Cseh, A., Sepsi, A. and Makai, S. (2022) A Multigraph-Based Representation of Hi-C Data. *Genes (Basel)*, **13**.

41. Di Stefano, M., Paulsen, J., Lien, T.G., Hovig, E. and Micheletti, C. (2016) Hi-C-constrained physical models of human chromosomes recover functionally-related properties of genome organization. *Sci Rep*, **6**, 35985.

42. Bonato, A., Chiang, M., Corbett, D., Kitaev, S., Marenduzzo, D., Morozov, A. and Orlandini, E. (2024) Topological Spectra and Entropy of Chromatin Loop Networks. *Phys Rev Lett*, **132**, 248403.

43. Xu, Y., Das, P., McCord, R.P. and Shen, T. (2024) Node features of chromosome structure networks and their connections to genome annotation. *Comput Struct Biotechnol J*, **23**, 2240-2250.

44. Pancaldi, V. (2023) Network models of chromatin structure. *Current Opinion in Genetics & Development*, **80**, 102051.

45. Pancaldi, V., Carrillo-de-Santa-Pau, E., Javierre, B.M., Juan, D., Fraser, P., Spivakov, M., Valencia, A. and Rico, D. (2016) Integrating epigenomic data and 3D genomic structure with a new measure of chromatin assortativity. *Genome Biology*, **17**, 152.

46. Xu, Y., Das, P., McCord, R.P. and Shen, T. (2024) Node features of chromosome structure networks and their connections to genome annotation. *Computational and Structural Biotechnology Journal*, **23**, 2240-2250.

47. Lind, P.G., González, M.C. and Herrmann, H.J. (2005) Cycles and clustering in bipartite networks. *Physical Review E*, **72**, 056127.

48. Ernst, J., Kheradpour, P., Mikkelsen, T.S., Shoresh, N., Ward, L.D., Epstein, C.B., Zhang, X., Wang, L., Issner, R., Coyne, M. *et al.* (2011) Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature*, **473**, 43-49.

49. Briand, N. and Collas, P. (2020) Lamina-associated domains: peripheral matters and internal affairs. *Genome Biol*, **21**, 85.

50. Foutch, D., Pham, B. and Shen, T. (2021) Protein conformational switch discerned via network centrality properties. *Computational and Structural Biotechnology Journal*, **19**, 3599-3608.
51. Lindsay, R.J., Pham, B., Shen, T. and McCord, R.P. (2018) Characterizing the 3D structure and dynamics of chromosomes and proteins in a common contact matrix framework. *Nucleic Acids Res*, **46**, 8143-8152.
52. Maisuradze, L., King, M.C., Surovtsev, I.V., Mochrie, S.G.J., Shattuck, M.D. and O'Hern, C.S. (2024) Identifying topologically associating domains using differential kernels. *PLoS Comput Biol*, **20**, e1012221.
53. Quinodoz, S.A., Ollikainen, N., Tabak, B., Palla, A., Schmidt, J.M., Detmar, E., Lai, M.M., Shishkin, A.A., Bhat, P., Takei, Y. *et al.* (2018) Higher-Order Inter-chromosomal Hubs Shape 3D Genome Organization in the Nucleus. *Cell*, **174**, 744-757 e724.
54. Nagano, T., Lubling, Y., Stevens, T.J., Schoenfelder, S., Yaffe, E., Dean, W., Laue, E.D., Tanay, A. and Fraser, P. (2013) Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. *Nature*, **502**, 59-64.
55. Dekker, J. and Mirny, L. (2016) The 3D Genome as Moderator of Chromosomal Communication. *Cell*, **164**, 1110-1121.
56. Fudenberg, G. and Imakaev, M. (2017) FISH-ing for captured contacts: towards reconciling FISH and 3C. *Nat Methods*, **14**, 673-678.
57. Kaul, A., Bhattacharyya, S. and Ay, F. (2020) Identifying statistically significant chromatin contacts from Hi-C data with FitHiC2. *Nat Protoc*, **15**, 991-1012.
58. Guelen, L., Pagie, L., Brasset, E., Meuleman, W., Faza, M.B., Talhout, W., Eussen, B.H., de Klein, A., Wessels, L., de Laat, W. *et al.* (2008) Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. *Nature*, **453**, 948-951.
59. Ernst, J. and Kellis, M. (2010) Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nat Biotechnol*, **28**, 817-825.

60. Kind, J., Pagie, L., de Vries, S.S., Nahidiazar, L., Dey, S.S., Bienko, M., Zhan, Y., Lajoie, B., de Graaf, C.A., Amendola, M. *et al.* (2015) Genome-wide maps of nuclear lamina interactions in single human cells. *Cell*, **163**, 134-147.
61. Hirano, T. (2012) Condensins: universal organizers of chromosomes with diverse functions. *Genes Dev*, **26**, 1659-1678.
62. Green, L.C., Kalitsis, P., Chang, T.M., Cipetic, M., Kim, J.H., Marshall, O., Turnbull, L., Whitchurch, C.B., Vagnarelli, P., Samejima, K. *et al.* (2012) Contrasting roles of condensin I and condensin II in mitotic chromosome formation. *J Cell Sci*, **125**, 1591-1604.
63. Hoencamp, C., Dudchenko, O., Elbatsh, A.M.O., Brahmachari, S., Raaijmakers, J.A., van Schaik, T., Sedeno Cacciatore, A., Contessoto, V.G., van Heesbeen, R., van den Broek, B. *et al.* (2021) 3D genomics across the tree of life reveals condensin II as a determinant of architecture type. *Science*, **372**, 984-989.
64. Ono, T., Fang, Y., Spector, D.L. and Hirano, T. (2004) Spatial and temporal regulation of Condensins I and II in mitotic chromosome assembly in human cells. *Mol Biol Cell*, **15**, 3296-3308.
65. Samejima, K., Booth, D.G., Ogawa, H., Paulson, J.R., Xie, L., Watson, C.A., Platani, M., Kanemaki, M.T. and Earnshaw, W.C. (2018) Functional analysis after rapid degradation of condensins and 3D-EM reveals chromatin volume is uncoupled from chromosome architecture in mitosis. *J Cell Sci*, **131**.
66. Dey, A., Shi, G., Takaki, R. and Thirumalai, D. (2023) Structural changes in chromosomes driven by multiple condensin motors during mitosis. *Cell Rep*, **42**, 112348.
67. Sakamoto, T., Sakamoto, Y., Grob, S., Slane, D., Yamashita, T., Ito, N., Oko, Y., Sugiyama, T., Higaki, T., Hasezawa, S. *et al.* (2022) Two-step regulation of centromere distribution by condensin II and the nuclear envelope proteins. *Nat Plants*, **8**, 940-953.
68. Hirano, T. (2016) Condensin-Based Chromosome Organization from Bacteria to Vertebrates. *Cell*, **164**, 847-857.

APPENDIX

Supplementary Information

Table S2. 1 Chromatin state coverage and correlation with interaction strength on chromosome 19. For each chromatin state (CS1–CS15), the table shows the fraction of bins with non-zero appearance and the Spearman correlation (ρ) between Q and the consensus of communication metrics (PC1 of ΔC4 vs. C4E).

Chr19 Chromatin State		Appearance Q (%)	ρ (ΔC4)	p-values	ρ (C4E)	p-values
CS_1	Active promoter	79.3	0.601	1.18E-24	0.782	1.89E-45
CS_2	Weak promoter	83.5	0.577	1.96E-22	0.709	7.99E-33
CS_3	Inactive/poised promoter	35.0	0.420	1.55E-11	0.382	6.79E-11
CS_4	Strong enhancer	63.7	0.670	3.86E-32	0.746	1.26E-44
CS_5	Strong enhancer	64.1	0.614	6.02E-26	0.672	1.68E-33
CS_6	Weak/poised enhancer	85.2	0.542	2.44E-30	0.776	6.07E-45
CS_7	Weak/poised enhancer	83.1	0.623	7.18E-27	0.707	2.20E-37
CS_8	Insulator	86.9	0.486	1.84E-15	0.476	8.39E-14
CS_9	Transcriptional transition	62.9	0.590	1.26E-23	0.613	4.34E-27
CS_10	Transcriptional elongation	76.8	0.388	6.08E-10	0.550	3.20E-18
CS_11	Weak transcribed	86.5	0.420	1.44E-11	0.672	8.14E-29
CS_12	Polycomb repressed	51.9	0.455	1.71E-13	0.384	2.40E-10
CS_13	heterochrom	94.9	-0.519	1.01E-17	-0.422	1.29E-13
CS_14	Rep/CNV	56.5	0.007	0.912537	0.169	0.024336
CS_15	Rep/CNV	39.2	0.108	0.098681	0.165	0.022638

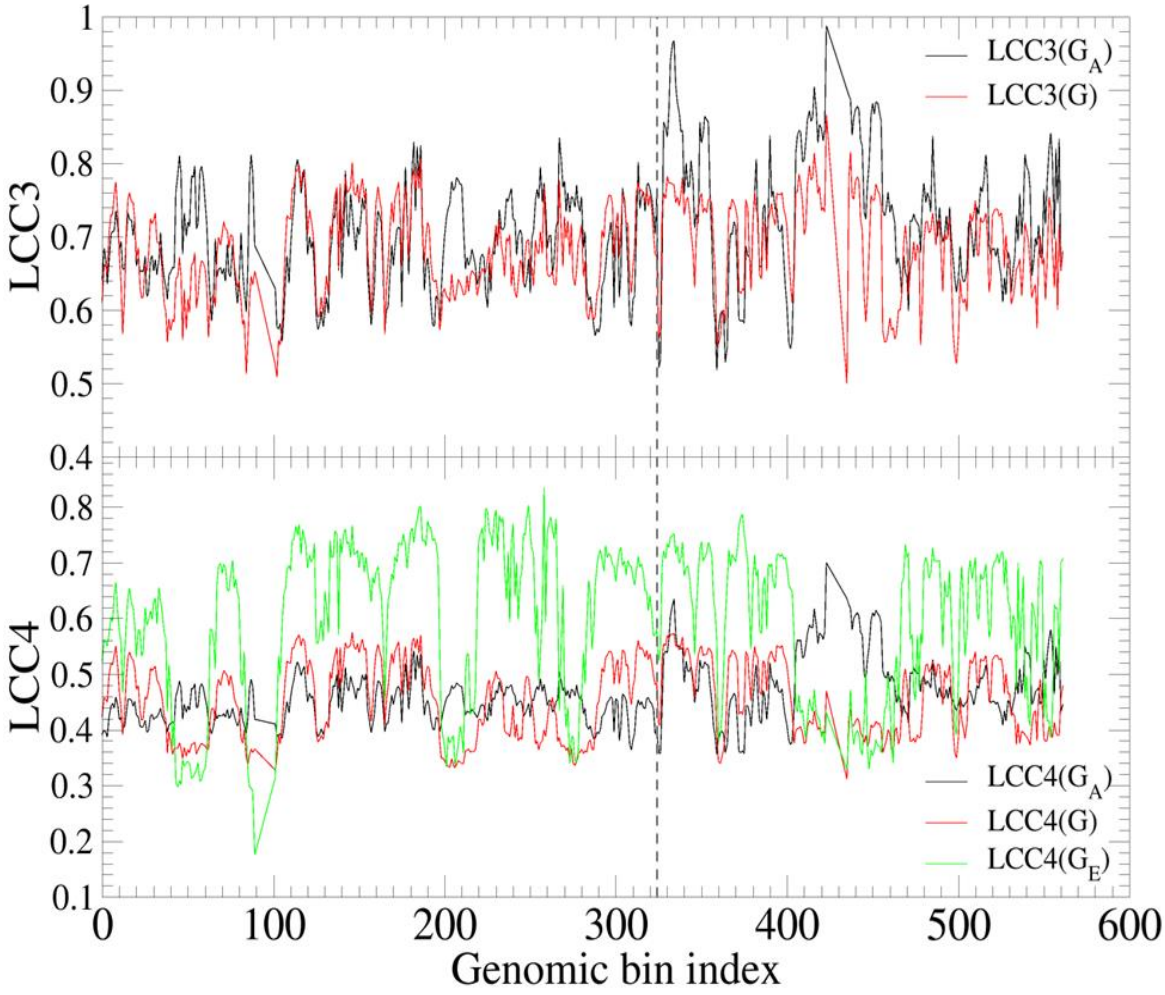


Figure S2. 1 Comparison of LCC3 and LCC4 values of different contact networks in Chr17:Chr19 system. Bin index 0-324 is chromosome 17 (0 – 81,195,210 bp), 325-561 is chromosome 19 (0 – 59,128,983 bp), resolution 250kb.

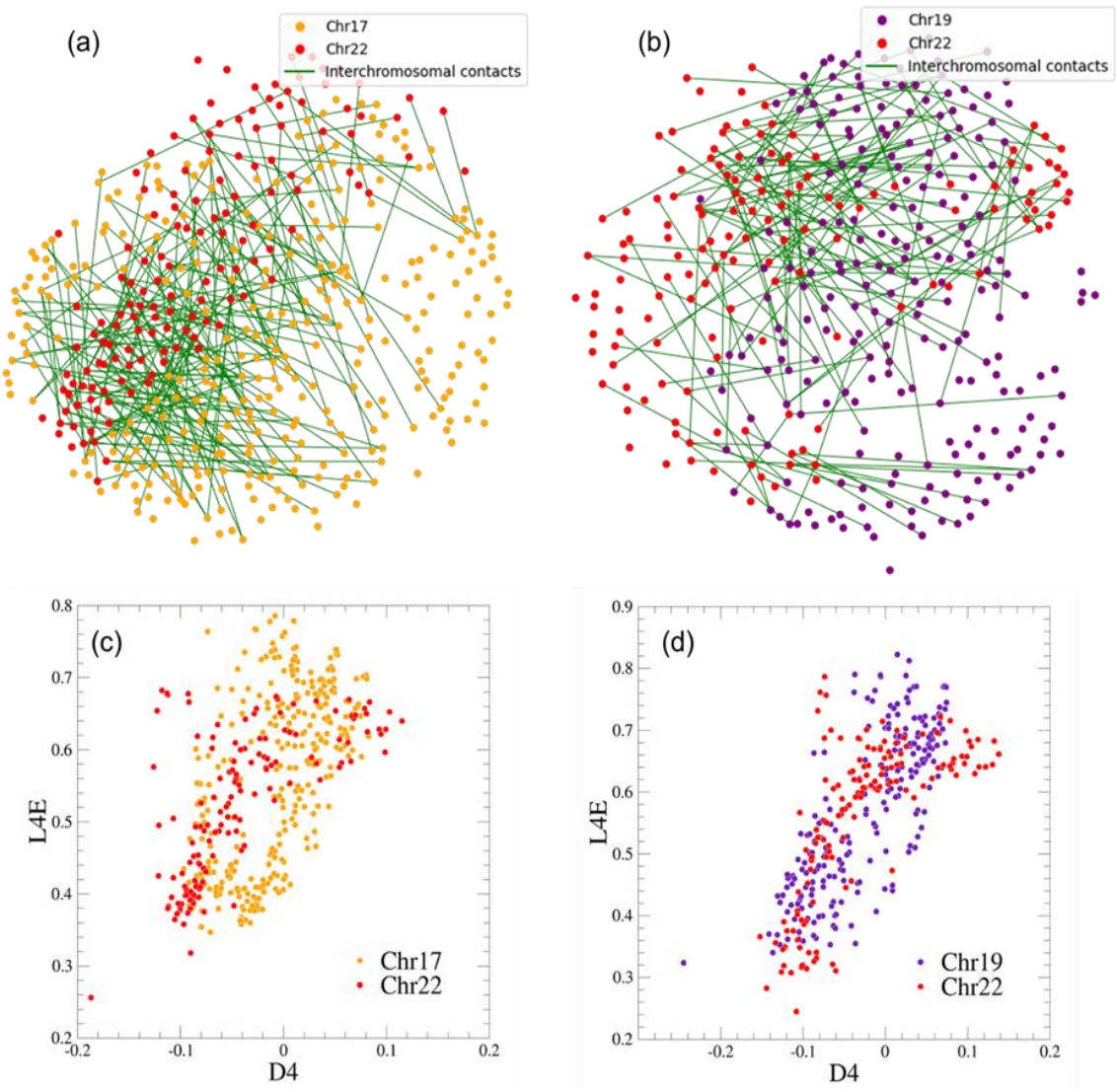


Figure S2. 2 (a-b) Force distribution layout of the (a) chromosome 17–22 and (b) chromosome 19–22 contact network system in HAP1 wild-type and ΔCAPH2 mutant cells, at 250-kb resolution. Green lines represent inter-chromosomal interactions. To improve clarity, only randomly sampled 1% of the top 50% strongest interactions (edges) are shown. (c-d) Scatter plots showing the correlation between C4E and ΔC4 in two chromosomal systems: (c) chromosome 17-22 system, and (d) chromosome 19-22 system.

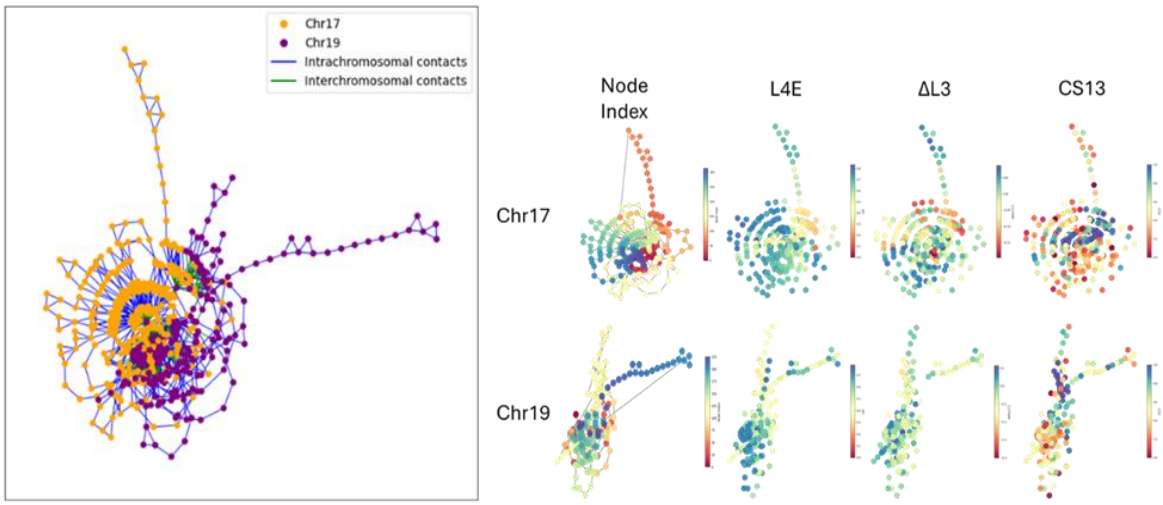


Figure S2. 3 Network display using the top 1% links for a comparions of the random 1% of top 50% used in Main manuscript Figure 2.3a.

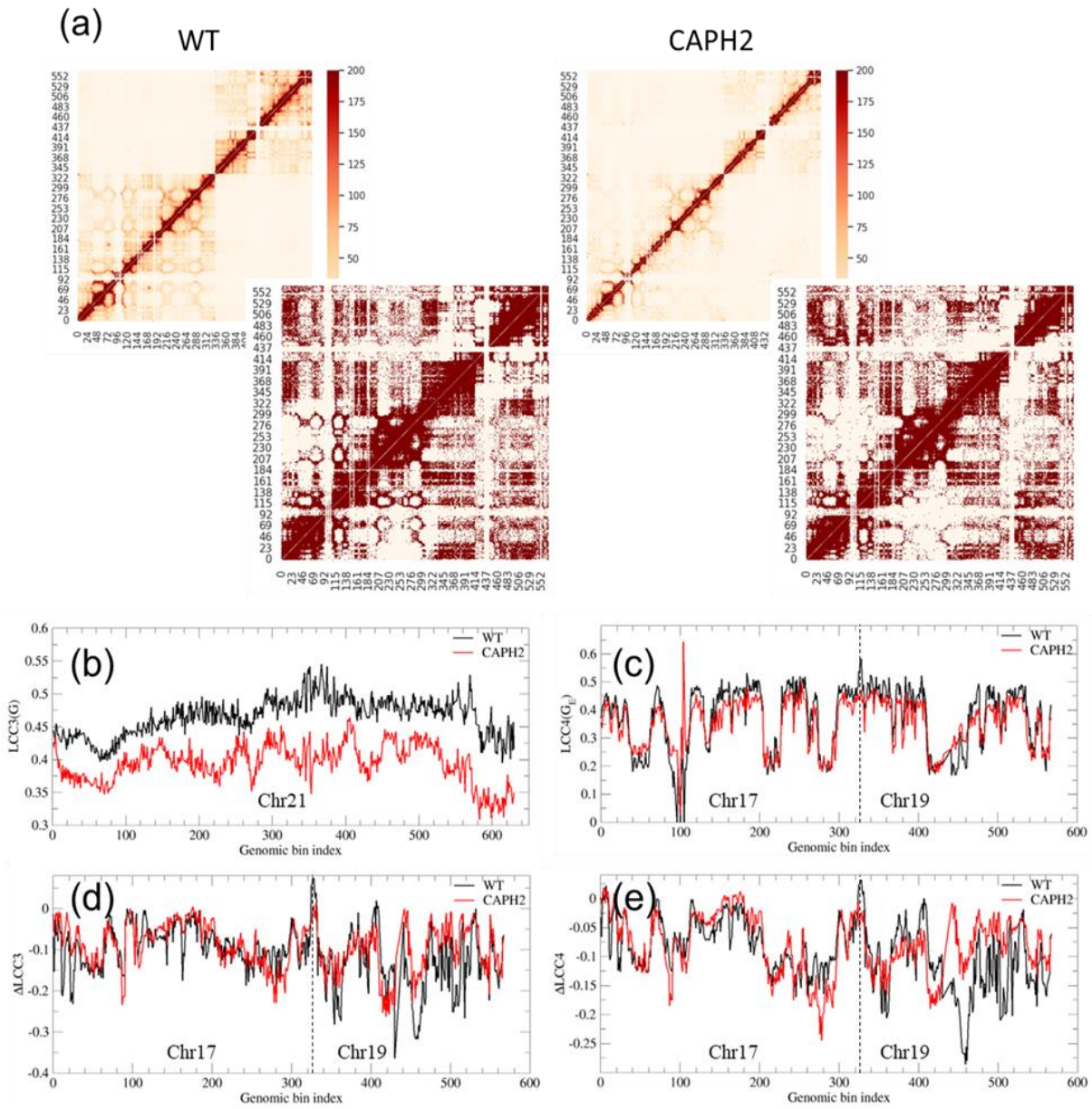


Figure S2. 4 The comparison of chr17:19 interchromosomal structure networks between HAP1 wild type (WT) and a mutant (ΔCAP-H2). Bin index 0-324 is chromosome 17 (0 – 81,195,210 kb), 325-561 is chromosome 19 (0 – 59,128,983 kb), resolution 250kb. (a) Heatmap from Hi-C data and contact network with 50% non-zero values as cutoff. (b) Comparison of LCC3 values in the intra-chromosomal contact network of chromosome 21, at 50-kb resolution. (c-e) Comparison of LCC4, ΔL3, and ΔL4 values in the inter-chromosomal contact networks of chromosomes 17–19 at 250kb resolution.

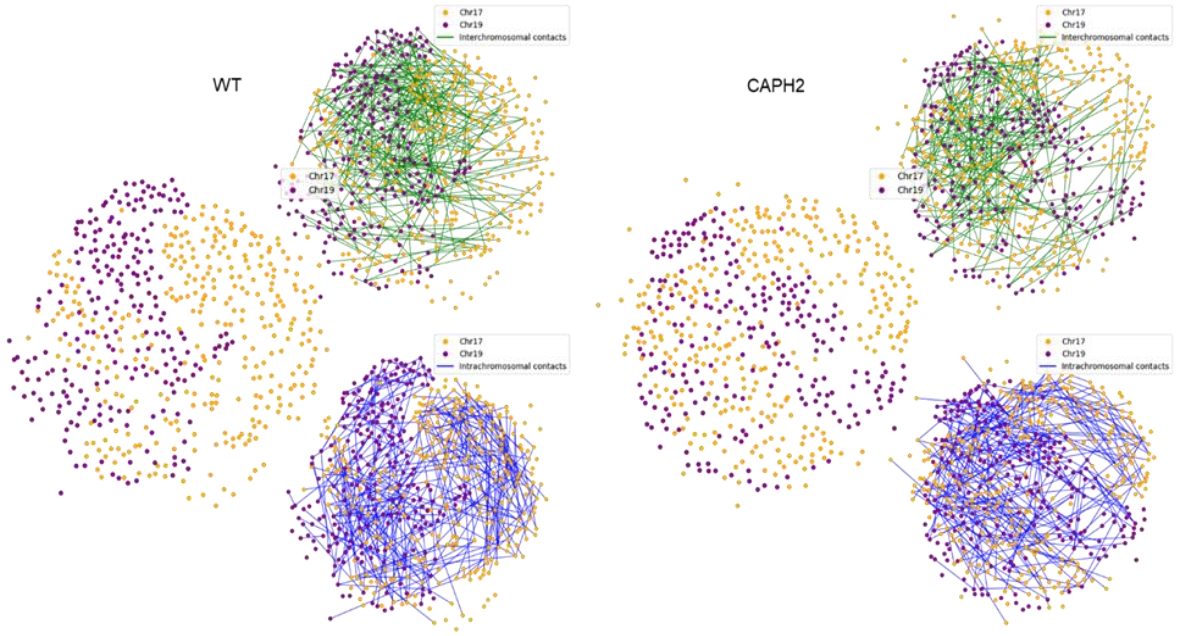


Figure S2. 5 Force distribution layout of the chromosome 17–19 contact network system in HAP1 wild-type and ΔCAPH2 mutant cells, at 250-kb resolution. Green lines represent inter-chromosomal interactions, and blue lines represent intra-chromosomal interactions. To improve clarity, only randomly sampled 1% of the top 50% strongest interactions (edges) are shown.

CHAPTER III

CONFORMATIONAL TRANSITION OF A POLYCATIONIC HINGE DOMAIN CONTRIBUTES TO DNA BINDING

Abstract

Nuclear receptors (NRs) are multidomain transcription factors that regulate gene expression through DNA binding and ligand sensing. While structured domains such as the DNA-binding domain (DBD) and ligand-binding domain (LBD) are well characterized, the functional role of intrinsically disordered linkers, such as the hinge between DBD and LBD, remains less understood. Here, we investigated the thyroid hormone receptor α (TRα) hinge using multi-microsecond molecular dynamics simulations of the DNA-bound and DNA-free states. To capture large-scale conformational changes beyond linear methods such as PCA, we applied a convolutional variational autoencoder (VAE) to contact matrices derived from the simulations. The VAE framework allowed us to explore the influence of different hyperparameters—including latent dimension, learning rate, and KL-divergence weighting—on the ability to resolve subtle conformational transitions. Our analysis revealed that DNA binding induces a helix-to-unwound coil transition in the hinge, coupled with DNA deformation and reduced DBD–LBD interactions. The latent space learned by the VAE provides a low-dimensional yet interpretable representation of hinge dynamics, linking structural disorder with functional activation. This study highlights the utility of machine learning approaches, particularly VAE parameterization, in probing the conformational ensembles of intrinsically disordered protein regions.

Introduction

Transcription factors consist of distinct domains with assigned functions. For example, a DNA-binding domain is responsible for recognizing a specific nucleic acid sequence while a regulatory domain senses specific environmental conditions. However, the action of the protein is more than a linear sum of individual activities from each modular domain. The interactions between domains appear to fine-tune the precise action of these modular proteins. A prominent example is the nuclear receptor (NR) family of transcription factors, which presents several longstanding puzzles related to activation mechanisms—including ligand sensing and DNA recognition (209,210). Notably, there are only a few canonical 6-base-pair DNA response element sequences that are targeted by the 48 classes of human NR families and hundreds of others across species. For

instance, thyroid hormone receptors, retinoid X receptors, and estrogen receptors all preferentially bind to the AGGTCA motif, while androgen, glucocorticoid, progesterone, and mineralocorticoid receptors recognize AGAACA(210-212). Several mechanisms have been proposed to explain this apparent DNA readout degeneracy, including DNA shape differences influenced by flanking sequences, recruitment of specific partner proteins, heterodimerization or oligomerization of NRs, and the involvement of additional domains that assist the canonical recognition site and contribute to binding specificity (211,213,214).

Overall, conformational dynamics and protein-polynucleotide interaction are recognized as being essential for multidomain signaling proteins to achieve their functions (215,216). The direct structural consequence and subsequent implications of protein-nucleic acid interaction have been increasingly appreciated in recent decades, especially in the areas of intrinsically disordered domains, multivalent interactions, folding-binding coupling, and cross-domain allostery (217-224). As a superfamily of ligand-dependent transcription factors, the NR superfamily senses various small molecules (from hormones, metabolites, to xenobiotics) and responds by turning on and off transcription of target genes (209). The current study focuses on the thyroid receptor (TR), which plays a crucial role in regulating several biological processes, including thyroid hormone-controlled metabolism, cell differentiation, and intercellular communication (225).

NRs share the overall modular domain arrangement of four domains, i) the disordered amino-terminal domain (NTD), ii) the DNA-binding domain (DBD), iii) the disordered hinge, iv) the ligand-binding domain (LBD) (226). Of these, the DBD and LBD have well-characterized globular structures. The LBD contains a ligand-binding pocket that activates the protein upon agonist binding. The NTD and the hinge domain are predicted to be more flexible and are less understood, although these disordered domains also play important biological roles (213,227,228).

For example, in addition to the direct activation mechanism by the binding of coactivators (229) and other transcription factors, studies indicate that allosteric conformational dynamics is important for NR activation due to hinge-mediated communication between LBD and DBD, which in turn affects the DNA-DBD

interactions (230,231). The mechanism of ligand sensing and allosteric activation has potential applications in drug discovery and protein engineering.

Intrinsically disordered regions (IDRs) within proteins may be an important form of regulation for transcription factors. In fact, about 40% of all eukaryotic proteins contain IDRs of 50 amino acids or more (223). The percentage for transcription factors is much higher, and in fact all human NRs, except CAR, have both disordered NTD and disordered hinge within their structural topology (232,233). The physiological functions of IDRs include inducing protein-rich transcriptomic condensates and liquid-liquid phase separation (LLPS) (165,166,234,235). However, for a subset of transcription factors, it is plausible that LLPS is not a manifestation of biological needs but is related to the fact that these IDRs contain malleable and transient structures that, upon various environmental cues, bind to distinct regulators. In addition, certain parts of the IDRs contain sequence motifs that mediate interactions with regulatory partners. Because these motifs often share similar physicochemical properties, such as charge or aromaticity, they can engage in multivalent weak interactions. In the absence of appropriate regulatory constraints, such interactions may promote self-association and aggregation.

In addition to conformational dynamics, allostery is an important mode of biomolecular interaction and regulation of functional activity within NRs. While ligand binding initiates conformational changes within the LBD that control interactions with coregulators, NR ligands also affect the interactions between DBD and DNA. Besides ligand allostery (230), there is also DNA-binding allostery suggesting that DBD interacts with LBD differently in response to DNA binding. Such reciprocal allosteric communication is not necessarily limited to changes in ligand-binding affinity, but may instead reshape the conformational energy landscape of the receptor, altering the population and dynamics of ligand-bound states, coregulator preferences, and transcriptional outcomes. The effects of DNA binding on NR conformations have been frequently debated in recent years (236-243). Previous studies have largely focused on the mean conformational changes upon DNA binding, while much less is understood about the associated dynamics and allostery.

Although both DBD and LBD are highly conserved in sequence and structure and have a well-defined role, the disordered hinge varies significantly between NRs sequence-wise(226). It is still unclear if they share the same function (63,244,245). The uncertainty of the hinge conformation also leads to a debate on the overall conformation of this modular protein, especially on how the DBD and LBD are spatially related to each other (236-239), since there are various suggestions on the association between the DBD and the LBD of NRs (240-243). One view suggests a tight DBD:LBD association, such as those reported in PDBs 3DZU (PPAR:RXR) (241) and 5UAN (RAR:RXR) (240), while another view reported fewer direct interactions between these two structural domains, such as EMD-1985 (VDR: RXR)(243), PDB 4UMM (USP:EcR) (246), and PDB 4NQA (LXR:RXR) (247). There is further indirect evidence on how an NR may change its conformation (particularly how the DBD and LBD interact) upon binding to DNA. Based on biophysical measurements using isothermal calorimetry (ITC), the hypothesis of a conformational switch triggered by DNA binding was also proposed (231).

The high variability of the hinge among NRs suggests that its exact action might be different for each receptor. The hinge might participate in several biophysical actions, from orienting DBD and LBD to modulating the DNA-protein interaction. We aim to explore the specific role of the TRα hinge and the DNA-binding aspect of this multidomain protein dynamics, as illustrated in Figure 1a. In addition to facilitating how the DBD and LBD interact in a specific and biologically meaningful way, the hinge also exhibits unstructured characteristics, which is known to play an important role in regulating transcription factors (248).

Previously, TRα hinge was reported to be crucial for the binding of agonist ligand, T3, at the LBD (245,249). In this work, we first constructed a structure model of the complex TRα(T3):DNA, where the agonist ligand T3 (3,5,3'-triiodothyronine) is bound to the LBD. TRα includes the DBD, the hinge and the LBD, as shown in Figure 1. We then conducted computer simulation and analyzed the dynamic trajectories. As a comparison, we also studied TRα (T3) in the absence of DNA. The simulation results suggest that DNA binding significantly affects the receptor conformational dynamics and vice versa, DNA dynamics is also affected by the protein-DNA interaction. We also observed that

protein-DNA interactions are not limited to DBD, since the hinge-DNA interaction is also prominent. Our results are applicable to both TRα and TRβ as the DBD-hinge-LBD sequences are highly conserved and the primary difference between TRα and TRβ lies in the NTD, which is not included in this study. The protein dynamics being studied can provide insight into the allosteric communication between three domains and how NRs, as multidomain transcription factors, are regulated in general.

Systems and Methods

Modeling the protein-DNA complex

To model the conformational changes induced by DNA binding, we constructed two systems (a protein-DNA complex and the protein alone), where the protein is in the same initial conformation with and without DNA. Currently, there are no experimental structures available for full-length TR. The TRα we modeled is the DBD-hinge-LBD of chicken TRα (cTRα, isoform α1, UniProt ID P04625). cTRα is used instead of hTRα (UniProt ID P10827) or TRβ (UniProt ID P10828) because it aligns with a previous experimental biophysical and structural study (231) and additionally, components of the cTRα models (LBD and ligand, DBD and NTD) were constructed previously when we examined in simulations the cross-LBD allostery (250) and NTD-DNA interaction(213), respectively.

Since structural information is missing for the full-length TRα, we used homology modeling based on the existing interaction of another NR system. For the protein-DNA complex, we need to consider four independent aspects since there is no direct structural information available for them: (1) the internal structure of DBD, (2) the relative position and orientation between the DBD and LBD domains, and (3) the conformation of the hinge region that connects the DBD and LBD, and (4) the DNA and DBD interaction. We first built DBD of cTRα using the structural template of human thyroid receptor beta hTRβ (PDB: 2NLL (251), chain B, of 89 residues, 90% identical and 94% positive between hTRβ-DBD and cTRα-DBD) for our target amino acid sequence of chicken TR (residues 49-136) (252). There are two zinc finger binding sites in the DBD, along with eight cystine residues (highlighted in Figure 1b). These cysteine residues were modeled in the deprotonated thiolate state to enable Zn^{2+} coordination within the zinc finger

motifs, whereas all other cysteines were kept in the protonated CYS state. The LBD structure of cTRα bound with the ligand *T3* was created using the existing structure (PDB: 3UVV, chain A), with missing loops added as described in a previous simulation study of the LBD dimer complex. We only used residues 161-405 of 3UVV, since the N-terminus of 3UVV is part of the hinge region and we did not use the conformation of 3UVV, as it does not align with the DBD/LBD conformation template.

To properly orient the DBD and LBD relative to each other, the corresponding orientation of DBD and LBD of the retinoic acid receptor (RAR) (PDB: 5UAN, chain B) was used as a template. Among the available structures, hRAR was chosen due to its high homology with cTRα (40% identical) and positioning of the two subunits was done using the structure alignment script in Pymol (253). The NTD (residues 1-48 using the standard cTRα index) was not included because the crystal structure of this region could not be resolved. Next, we modeled the hinge region (63) spanning residues 124-160 that joins two structured domains (DBD and LBD). We constructed a few structure models of this region using AlphaFold (254) and ModLoop (255). In comparison, AlphaFold2 or AlphaFold3 was unable to provide an acceptable starting conformation of this disordered hinge region for an obvious reason, as shown in the supporting information (SI) Figure S1. Nevertheless, it is interesting to note that AlphaFold2 hints that helical structure(s) persists at this disordered region when it is DNA-free. AlphaFold2 models consistently show a long α-helix (residues 126-154) that separates the DBD and LBD domains. This unconstrained α-helical conformation of the hinge domain prevents the two domains from forming contacts, which would be inconsistent with the crystal structures of 5UAN and some other structures.

In contrast, the model generated by ModLoop does not pose the same physical interaction problem. ModLoop successfully generated a flexible loop (residues 127-160) that allows for the DBD and LBD to form an interface in an orientation consistent with other homologous NRs. The reason for these differences is that AlphaFold is a data-driven structure predictor whereas ModLoop models disordered regions and relies more on local interactions. In this case, the PDB 2NLL, while providing a useful model of the DBD, contains a long α-helix in the hinge region that points away from where the LBD would

reside in our system. Similarly, the starting residues of the PDB 3UVV point away from the DBD of our model. Data driven models such as AlphaFold2/3 output models of the disordered, polycationic hinge region based on these two incompatible structures that would preclude the investigation of DBD-LBD interactions, thus making a local structural model such as Modloop more practical for our study. Protonation states for the LBD are inherited from Ref. (250) and we also checked the newly added DBD and the hinge region using H++. Conventional protonation states were assigned. In summary, the DBD of the cTRα is structurally homologous to hTR (PDB: 2NLL) and the LBD is identical to a large part of the PDB: 3UVV, while the relative orientation between the DBD and LBD is fitted from a template from RAR (PDB: 5UAN) and the linker region between is modeled for this study.

There are a few mobility assay-based experimental reports on high affinity dsDNA sequences for TR's DBD, the so-called TR-specific DNA response elements (DREs) (211,256,257). Although there are 48 types of human NRs, only a handful of DREs are being used, one so-called 6-bp "half-site" being 5'-AGGTCA-3'/5'-TGACCT-3' for TR (256,257). The half-site AGGTCA forms the well-known direct repeat motif AGGTCA NNN AGGTCA where "NNN" symbolizes a spacer of 1 to 10 bps of unspecified nucleotides for dimeric binding as the motif provides two AGGTCA sequences that interact with NR dimer complexes where each sequence binds to one NR. In addition, a 5'-flanking sequence TA (256) was added to the motif and finally we surrounded the core sequence with the filler of poly(A) nucleotides upstream and downstream. As shown in Figure 1, the result is a 24-bp dsDNA, with the top strand being 5'-AAATA AGGTC AGAAG GTCAG AAAA-3' and its complementary strand. This dsDNA sequence has a high affinity for the DBD of TRα1, which was also tested in biophysical experiments (231). Of the 24 bp, only 5 bp are different between these two dsDNA sequences. The DNA structure was constructed using AMBER tool NAB (258). Alignment of this DNA response element with DBD was done using the position of DNA in the crystal structure 2NLL as a template, where hTRβ binds the canonical thyroid response element (TRE), 5'-AGGTCA-3'.

The complete construction of the protein-DNA system is shown in Figure 1b. The TRα construct (total 357 residues) contains three domains using the standard index (or internal index): (i) DBD, 49-123 (internal indices1-75), (ii) hinge, 124-160 (76-112) and (iii) LBD, 161-405 (113-357). Using the internal index, the ligand *T3* is residue 358, two zinc ions are residues 359 and 360. One DNA strand spans residues 361-384, and the complementary strand spans residues 385-408 with the binding box at 366-371 and 398-403 (AGGTCA). Our structural model preserves several critical protein-DNA interactions between the recognition helix and the response element reported in (251), such as TRα Arg76:3G-4T (at the 3rd and 4th positions of the binding box, hydrogen bonding between the sidechain of Arg and the bases of G-T), which corresponds to the internal residue index 28:368-369, and to the DNA-binding structure template hTRβ (2NLL) indexing Arg328:515G-516T.

Atomistic molecular dynamics simulation

The first system, the protein-DNA complex system, was established through the LEaP program. The TRα-DNA complex was solvated in a 10-Å box of 33,396 water molecules. 49 Na^+ counterions were added to produce a neutral charge for the entire system with a total size of 107,527 atoms. For the TRα-alone setup there are a total of 74,776 atoms, including 23,003 water molecules and 3 Na^+ counterions. Finally, for the standalone dsDNA as the third system, the system has a total of 17,442 atoms, including 5,290 water molecules and 46 Na^+ counterions. A ligand (T3) associated with LBD and Zn^{2+} ions associated with a zinc-finger motif (259) are also included. The forcefield used matches the previous study of the heterodimeric system involving TR (T3), where amber14SB is used for protein and TIP3P for water. Additionally, the DNA forcefield used is bsc0. Two 5-microsecond computer simulations were performed on Anton2 and STAMPEDE2/3, one for the TRα-DNA complex and one for TRα alone. Continuous simulations were chosen instead of multiple replicas of short-time simulations. (260). This approach enables improved exploration of the dynamics and conformational space, which is particularly important for simulations involving IDRs, whereas advanced sampling methods (116) might not be practical for handling large systems. The convergence is discussed in the Results section below based on post hoc analysis.

Snapshots of the trajectory were stored every 0.1 ns during the simulation (total 50,000 frames for each system). All simulations used a Nose-Hoover thermostat with a set temperature of 300 K at a pressure of 1 bar. The time step was 2.5 fa for Anton2 simulations and 2.0 fs for STAMPEDE2/3. Our analysis in the following focuses on the microsecond timescale which we were able to approach with available resources.

Statistical analysis of conformational dynamics

To describe the internal conformational dynamics of the protein and the protein-DNA interaction, we focus on using residue-residue contact interactions in addition to conventional analyses. The starting point of contact analysis is to render each snapshot in a corresponding contact matrix, where a binary value of 1 or 0 is assigned to the interaction between two residues. We define that a contact is formed ($a_{ij} = 1$) when any atom in residue *i* is within a distance cutoff (4.2 Å) of any atom in residue *j*. Otherwise, no contact is formed ($a_{ij} = 0$). The distance cutoff value was selected based on previous studies (261,262). Note that the cutoff is used only for the analysis in the current study and we do not rely on the cutoff to construct the energy function or drive coarse-grained models, where the choice of the value can be more critical. In fact, a sensitivity study comparing a slightly more stringent 4.0 Å vs. a more tolerant 4.2 Å shows no significant differences (194). The contact matrix has a similar mathematical structure to the distance matrix, and the main difference is that the former discerns the contact energetics while the latter emphasizes long-distance connection and loses shorter-range sensitivity. An ensemble average will lead to the mean contact matrix which is also referred to as a contact heatmap. Furthermore, the collective variable analysis of these contact matrices through linear methods, such as principal component analysis (263), and nonlinear ones, such as variational autoencoder (264), lead to major dynamic modes of biomolecular structure and dynamics expressed by contact degrees of freedom (262).

As demonstrated and compared in previous studies, two types of contact-based PCA reveal distinct features. (140,265): Edge-based PCA (or explicit principal component analysis, ePCA) of proteins examines the dynamic fluctuations of individual contacts, such as conformational switch (266,267) and allosteric binding (250,268,269), while site-based PCA (or implicit principal component analysis, iPCA) focuses on the basic

consensus of structural features, such as structural elements and domains of largely disordered or partially folded proteins (140,270). In this work, we mainly used iPCA to characterize the residual structural consensus of the hinge with and without DNA.

As for the dynamic feature, instead of using ePCA, we opt for using a more versatile nonlinear analysis based on machine learning algorithms. While ePCA identifies collective modes as linear combinations of individual contacts, capturing concerted motions of contact formation and disruption, nonlinear methods such as neural networks with activation functions are better suited to accurately describing large-scale nonlinear motions exhibited in the disordered hinge. As a contact matrix shares the same mathematical structure as a (symmetrical) 2D image, where various advanced AI-based image processing algorithms are being developed, we use a specific convolutional neutral network (CNN) algorithm, and the architecture of the CNN is the widely used variational autoencoder (VAE) (264). The VAE scheme was used to compress the hinge contact matrices into a small number of collective variables that capture the dominant patterns of contact formation and disruption. This nonlinear dimensionality-reduction step allows us to better represent the large-scale motions of the disordered hinge compared with linear approaches such as ePCA. Because the VAE scheme contains both a recognition (discriminative) model for encoder and a generative model for decoder, the learned collective variables can be decoded back into representative contact maps, providing an interpretable link between the low-dimensional features and the underlying structural changes (SI Figure S2). The detailed parameters of our VAE model are listed in SI Figure S2. One advantage of VAE over many other nonlinear methods, especially those classification methods such as UMAP (271) and t-SNE (272), is that one can easily compute the average on the constructed collective variables and reverse them to a representative contact map using the decoder step, as illustrated in SI Figure S2.

In this work, we also used analysis tools that are typically used for characterizing abstract graphs. When we compare mean structural changes using mean contact matrices, sometimes a direct comparison is difficult to visualize and we may want to extract residue-residue contact interaction and express the information as residue-associated properties, which essentially is a 2D to 1D rendering. To achieve this, we use a set of

protein structure network analysis metrics that convert the two-dimensional contact interaction information to linear information associated with the residues. The concept of a protein structure network (PSN) (109,110) is based on the mean contact matrix. The formation of a link in PSN is decided by a threshold parameter δ and we construct a link between two residues (nodes) for PSN if the corresponding contact frequency is larger than $\delta\ (= 0.02)$ (194). Once the network (adjacency matrix) is defined, we apply additional node-based network metrics to extract the network features. The specific ones that we used here include eigenvalue centrality, clustering coefficient and square clustering coefficient (126,186).

Results and Discussion

The hinge–DNA interaction assists DBD–DNA association

A unique feature of the TRα hinge is its polyelectrolyte character, containing 13 charged residues within a short segment (132–147) immediately preceding the coactivator-like motif 148MIKS152L. This region exhibits strong DNA interaction, as shown by the contact heatmap (Figure 3.1a) and 3D representation (Figure 3.1b). In addition to the expected DBD–DNA binding (residues 49–123), the hinge region (124–160) contributes extensive contacts with DNA.

Within the DBD, the recognition helix (Arg76) interacts specifically with base 3G in the binding box with >90% frequency, while the scaffold helix and loop (Arg102–Arg109) contact the complementary strand and flanking cytosines through backbone hydrogen bonds. The arginine-rich segment of the hinge (Arg140–Arg144) engages the central binding box (2G–5C), strengthening DBD–DNA association.

As shown in Figure 3.1b, these hinge contacts induce DNA bending through insertion of arginine sidechains (Arg140–Arg143), reducing phosphate–phosphate spacing below 6.2 Å relative to the 6.5–7.0 Å in canonical B-DNA. This deformation suggests that hinge–DNA backbone interactions modulate local shape and complement sequence recognition by the DBD. The hinge loop (Lys150–Arg155) further inserts into the DNA minor groove (16G–20G, 31G–35T), while another loop (Arg144–Lys145) anchors nonspecifically to the backbone, collectively stabilizing the overall complex.

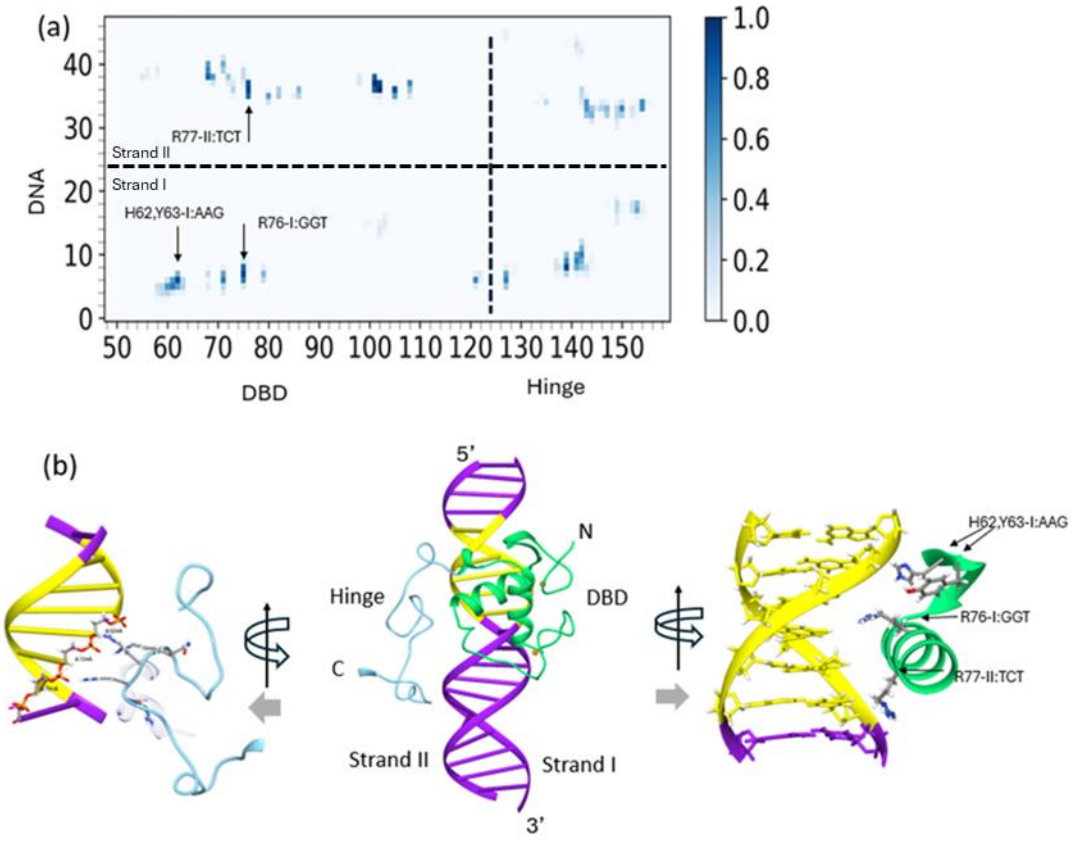


Figure 3. 1 (a) A contact heatmap of protein-DNA interaction with strong interaction regions labeled. The standard index is used for TR and an internal index for DNA duplex (strand I: 1-24, strand II: 25-48). (b) The cartoon representation of DBD:DNA binding. The DNA-binding box is highlighted in yellow. All residue indices are of chicken TR. The polycationic hinge distorts the dsDNA backbone where the distances between neighboring phosphates are shortened due to the inserted arginine sidechains.

Allostery and interdomain interaction modulated by DNA binding

We first present how the overall structure of the TRα conformation ensemble changes in response to DNA binding from a contact heatmap viewpoint, as shown in Figure 3a. Additional accompanying detailed information on both contact- and distance-based metrics can be found in SI Figure S3.1. As shown in Figure 3.2a (and SI Figure S3.1), the overall contacts between LBD and DBD decrease after DNA binding by roughly 50%, while the contacts between LBD and the hinge increase by approximately 30%. The 5-μs simulation also reveals a notable overall movement between the DNA-free and DNA-bound TRα. We note that, upon DNA association, TRα adopts a more open conformation (overall distance).

features are illustrated in SI Figure S3.2). In addition, the direct contact interaction between DBD and LBD also decreases significantly due to DNA binding, supported by the contact heatmaps of Figure 3.2a and SI Figure S3.1. This finding is also consistent with a previous study which reported that DNA binding reduces the DBD-LBD interactions (231).

Most structural changes revealed by our simulations are from interactions between domains and within the hinge, and only minor internal structural changes within the DBD or LBD exist. Although there was no comparison of conformations between the DNA-free and DNA-bound DBD for TRα, earlier studies examined the DNA-DBD interaction of other NRs such as the glucocorticoid receptor (GR) and the retinoid X receptor (RXR). The contact heatmaps for both structural domains maintain a consistent level of internal contacts. Our results (subtle direct change in DBD upon DNA binding) are consistent with a recent report on the crystal structure of the GR DBD-DNA complex, where the DBD of GR in the DNA-free state is similar to that of the DNA-bound form (273). However, earlier NMR studies performed on the DBD of GR present more dramatic conformational changes in the D loop region and mutations that mimic these DNA binding allostery were also observed (273,274). The difference might be tied to the oligomerization status, i.e., the monomer system (which lacks the dimerization interface), that is being simulated here.

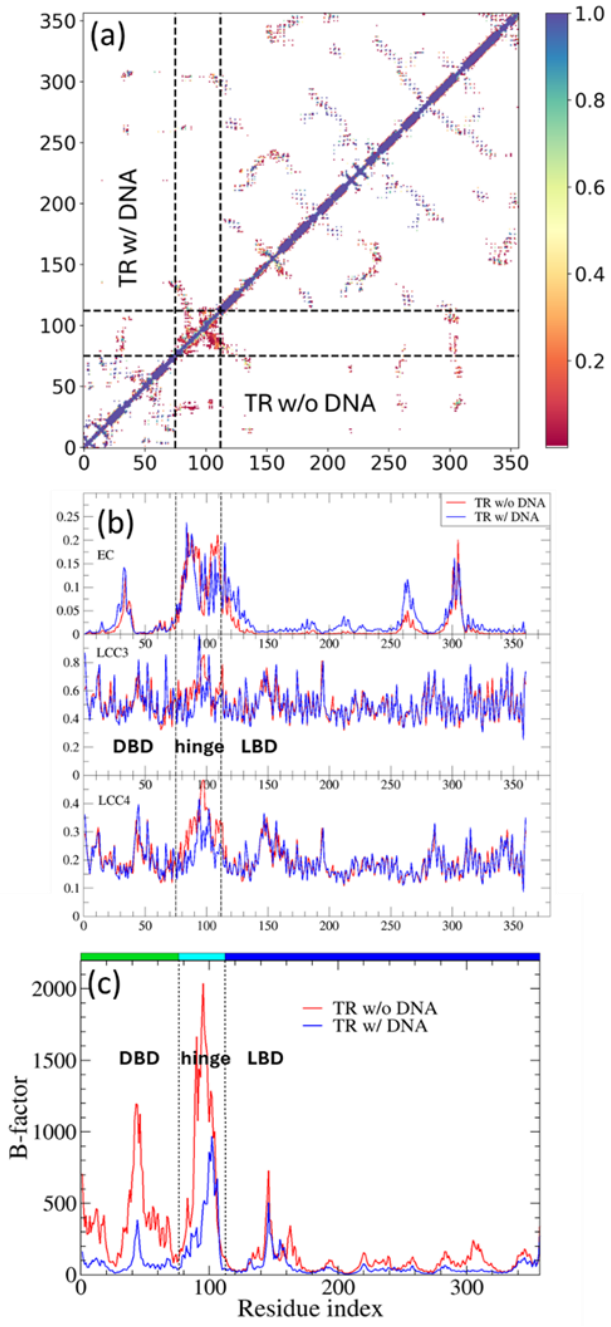


Figure 3. 2 (a) The contact heatmaps the internal contacts of TR. Here the upper triangle is TR with DNA and the lower one is TR without DNA. The dash lines separate three domains (DBD, hinge, and LBD). (b) A comparison of contact heatmap features (without DNA - red, with DNA - blue) using selected node-based network properties, EC, LCC3, and LCC4. (c) Flexibility expressed as calculated B-factors as a function of residue index. Again, red indicates TR without DNA and blue for TR in the presence of DNA. All residue labels are from the internal index, which can be converted to the chicken TR index by adding 48.

In addition to the 2D heatmap representation in Figure 3.2a, we rendered the 2D information into linear annotations at the residue level for easier comparison, as shown in Figure 3b. To achieve this, we applied network properties to extract structural changes at the per-residue level. As mentioned in the Method section, we first render the 2D heatmap into a protein structure network (PSN) and then calculate node-based metrics of the PSN. Here, we opt to use a single threshold PSN with a cut-off range [0.02, 1] instead of PSN* with a range [0.02, 0.98]. Although the previously defined PSN* (a dual threshold network that ignores the permanent contacts as well as those infrequent contacts) is more sensitive to conformational changes than PSN using centrality measurements (194), PSN is more connected, allowing for the calculation of well-defined local clustering coefficients (126,188). As shown in Figure 3.2b, node-based metrics, eigenvector centrality (EC) (126) and local clustering coefficients (LCC3 and LCC4) (186) are displayed as functions of residues. With DNA bound, the major changes in the LCC3 and LCC4 values are concentrated on the hinge, while the values in the DBD and LBD regions do not change significantly. For the hinge region, an LCC3 peak originally between 90 and 100 disappears and splits into two peaks. LCC4 confers the observations from LCC3, but in a more subdued display. All three analyses indicate that drastic conformational rearrangements upon TRα binding to DNA originate from the hinge region.

Additionally, to better understand how these contacts affect protein dynamics in response to DNA binding, we also examined protein flexibility throughout the 5-μs simulations of DNA-free and DNA-bound TRα, as shown in Figure 3.2c. DNA binding to the TRα generally increases the stability of the protein, as evidenced by the consistently lower B-factor values throughout most of the DNA-bound protein structure (compared to the DNA-free structure). In fact, the entire DBD and most of the LBD show diminished flexibility, with the exception of two functionally important regions for dynamic structural change. One of these regions is the loop region located after helix 1 of the LBD, which not only increases flexibility but also loses contact with the DBD when it binds to DNA. This is supported by the comparison of the contact map in Figure 3.2a, as we can see a delocalization and weakening of the contact between this loop and the DBD.

It becomes clear from the protein dynamics analysis that the increased flexibility of these regions can be important for mediating structural changes in response to DNA binding. Coupled with the information from the B-factor and contact dynamics, we can identify the major dynamic regions of the protein system. For the protein without DNA, there are two relatively dynamic regions: one is the DBD-LBD interface through the hinge. The second region is in a distal part of the LBD, which includes the ligand-binding pocket and the specific loop located after helix 1 of the LBD (residues 191-209). For the DNA-bound TRα, we also observe two major dynamics regions. One region is still at the LBD and DBD interface and is comparable to that of the DNA-free TRα. The second region is at the interface between the DNA recognition helix and the major groove of DNA.
Our results suggest that DNA binding decreases fluctuations at the ligand-binding site. In turn, this enhances the effectiveness of the ligand-binding domain in holding the ligand and consequently promotes the LBD active conformation and its ability to recruit coactivators. This finding is consistent with an earlier discovery that estradiol induces changes of binding kinetics and promotes the estrogen receptor-DNA interaction (230). Both demonstrate that there is an allosteric communication between the protein-DNA interface and the protein-ligand interface.

Helix-to-unwound transition of hinge segments upon DNA binding

Although there are subtle changes within the DBD and LBD, the hinge adopts different conformations when DNA is bound. Unlike highly structured domains, where the focus is often on capturing dominant conformational changes, the primary goal in analyzing a diverse structural ensemble—such as that of an intrinsically disordered region (IDR)—is to identify consensus structural elements. As shown in earlier studies, site-based principal component analysis (iPCA) is particularly suited for unstructured polymers like IDRs that contain loosely defined secondary structures, whereas edge-based PCA (ePCA) is more appropriate for revealing dominant dynamic fluctuations in well-folded, highly structured proteins (140,265). We usually only apply ePCA for highly structured proteins as there is no ambiguity about the native structure; whereas iPCA can be useful for largely unstructured polymers (but with residual structural features) such as genome folds or intrinsically disordered proteins (270).

The internal contact of the hinge is shown in Figure 3.3a. We further applied site-based contact PCA (iPCA) on this patch and obtained structural features as the dominant principal component PC1. The iPCA for IDP is analogous to the A/B compartment analysis performed on Hi-C data, which depicts structural elements in semiflexible polymer chains. In the IDP content, the top eigenvectors are shown as bar graphs in Figure 3.3b. Here, each strong peak or valley indicates a cluster of sequence neighboring residues for coherent contacts with other (unspecified) residues. Meanwhile, the nodal set (residues at the x intercept) can be viewed as the boundary between clusters. We found two subdomains (residues 76-89, 90-110 using the internal index) of the hinge when it is DNA-free. Upon DNA binding, the second subdomain further splits into two near residue 96 (internal index; corresponds to Gly144).

In addition to the analysis of the subdomain structure from the contact viewpoint, it is interesting to inspect the changes in the secondary structure from helix to unwound upon DNA binding, as shown in Figure 3.3c. Here, we sorted 8 DSSP-derived secondary structure states (275) into four larger classes: helix (green: G, H, and I), strand (red: E and B), turn/loop (black: S and T) and others. Note that we do not refer to this transformation as the classic helix-to-coil transition, since coil often refers to a random structure. Rather, we use “unwound” or “extended coil” to emphasize that DNA provides specific constraints on the conformation of the hinge region, leading to the unwinding and then the flattening of the initial helical structure of the hinge. The simulation results reveal that DNA binding induces an intricate conformational change at the secondary structure level for the hinge region. For example, segments ^{139}NRER^{143}R and ^{145}KEE^{148}M have a helical conformation in its DNA-free form and change into a disordered region upon DNA binding. Additionally, while the ^{149}IKSLQH^{155}R segment remains mostly helical, it also shows an increased propensity to form turns after DNA binding.

Our result in Figure 3.1a also indicates a high level of contact between this region and the minor groove. The ^{148}MIKS^{152}L region of TRα was previously reported to be helical under certain conditions and potentially involved in a significant protein-protein interaction that can lead to protein aggregation, since the region interacts with the AF2 region of another receptor’s LBD and triggers oligomerization or phase separation (252)

. Interestingly, the ^{148}MIKS^{152}L region is considered an LXXLL-like motif, and the AF2 region of RXR usually binds to the well-known LXXLL motif of a coactivator (276). Thus, this oligomerization can be viewed as an inhibition mechanism since the coactivator binding site at the AF2 is blocked. Thus, the change in conformation near this location upon DNA binding can have an implication on TR regulation. Specifically, DNA binding leads to a disordered hinge and reduces inter-protein contact, consequently, fewer oligomerizations occur, more receptors are open to coactivator interactions, and finally, transcription becomes more active.

In addition to the mean structure features revealed in Figure 3.3, we also study the structural fluctuations of these ensembles. Since the hinge is quite disordered, instead of looking for dynamic modes using linear methods such as ePCA, we directly apply a machine learning algorithm. As stated in the Method section, the variational autoencoder (VAE) is selected as it can both perform nonlinear transformations of input information (residue-residue contacts) to collective degrees of freedom (latent variables) and produce conformations based on latent variables using the decoder scheme. To construct the machine learning model, we used the conformations of both TR alone and TR in the complex, total 2 × 1,000 frames (at the frequency of 1 per 5 ns) as the training set, The snapshots projected onto the latent variable are shown in Figure 3.4a and a total of 2,000 are shown. The latent variable space is shown in Figure 3.4b. The dynamic evolution of the hinge is shown in Figure 3.4c with time being color-labeled. Because there are little significant internal changes of the DBD and LBD and most activities come from the hinge, the time-dependent plot Figure 3.4c represents the process of this closed-to-open dynamics. Advanced transition path-based methods could be used to further assess the dynamics quantitatively; however, the present projection provides a clear qualitative picture. Notably, without DNA, the conformations of the hinge appears to be at the upper-right corner of Figure 3.4a, and one can interpretate such conformations as the hinge having helical fragments in the unbound state (based on near-diagonal contacts formed in the upper-right corner of contact maps shown in latent space display, Figure 3.4b), which become unwound upon DNA binding, as reflected in the shift in the latent

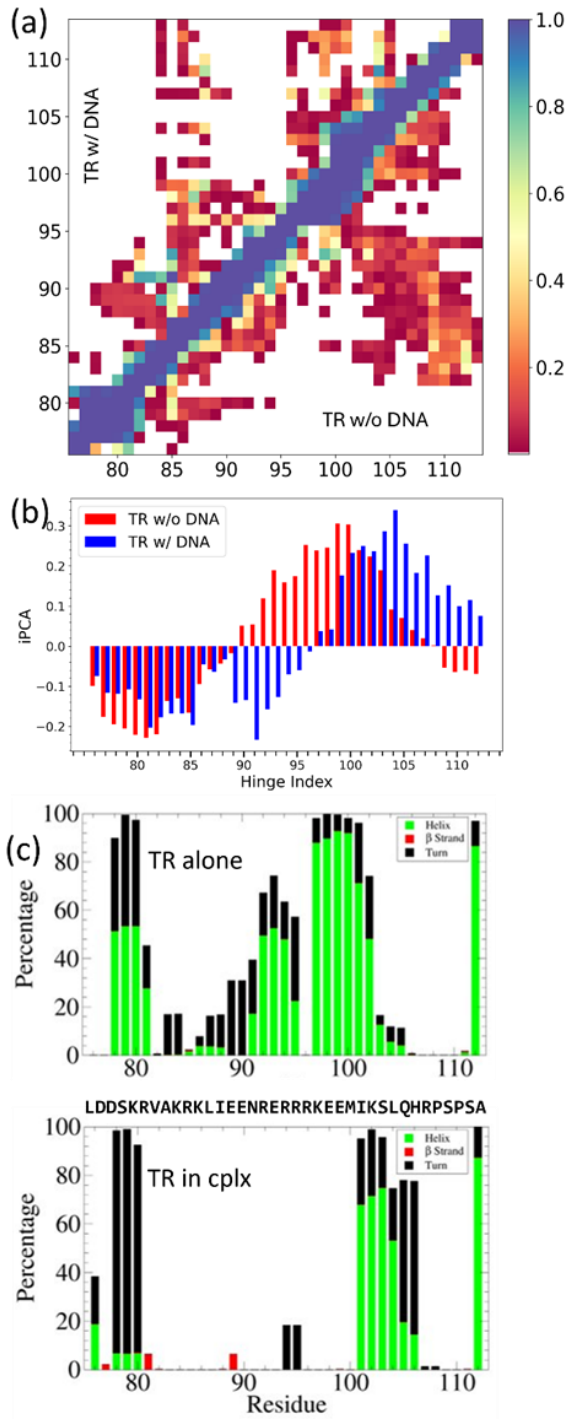


Figure 3. 3 (a) A focused view of hinge contact heatmap. (b) iPCA reveals subdomain structure of the hinge (red for TR without DNA and blue for TR with DNA). (c) Population of secondary structural elements as a function of residue index for the hinge region. All residue labels are in the internal index, which can be converted to the cTRα index by adding 48.

space. This transition is consistent with the helix-to-unwound transition described earlier in Figures 3.3b and 3.3c, reinforcing the observation that DNA binding induces structural disorder in the hinge.

Figure 3.4c also indicates that the time scale of this helical transition appears within one microsecond for the unbound TR hinge. From this and other explicitly time-dependent analyses, such as global root-mean-square plot rms(t) (SI Figure S3.3), one can conclude that each system is relaxed into a relatively stable state during the simulation. The system will eventually explore other conformations on a longer time scale, but they appear to be stable during this simulation period. Our report of DNA induced helix-to-unwound transition is relatively robust as we have discovered qualitatively similar transition of an alternative system where 5-μs simulation of the same protein with a different DNA response element was paired (data available upon request).

Although the current work on IDR is focused on the hinge region, it is interesting to make a connection to another IDR of nuclear receptors, the N-terminal domain (NTD). As the other well-known disordered domain for this superfamily of proteins, the NTD has been associated with ligand-independent activation (227,277). It was recently reported that NTD may affect the 5'-flanking sequence (further upstream at approximately 20 bp) binding with selectivity (213). This IDR-DNA interaction could be an important general theme for many transcription factors. It turns out that there are similar LXXLL-like motifs existing in both the NTD and the hinge regions for various NRs (227,276), and we may regard both of these IDR regions as one continuous, long IDR with a structured domain (DBD) embedded in the middle.

Effects of VAE parameters and sampling on Latent Representations of Hinge Dynamics

To characterize the conformational dynamics of the intrinsically disordered hinge region, we employed a convolutional variational autoencoder (VAE) trained on contact matrices derived from molecular dynamics simulations of DNA-bound and DNA-free TRα. The hinge segment (residues 76–112) produced contact matrices of size 37×37. For input standardization, each matrix was padded with zero values to reach a size of 40×40, ensuring compatibility with convolutional layers.

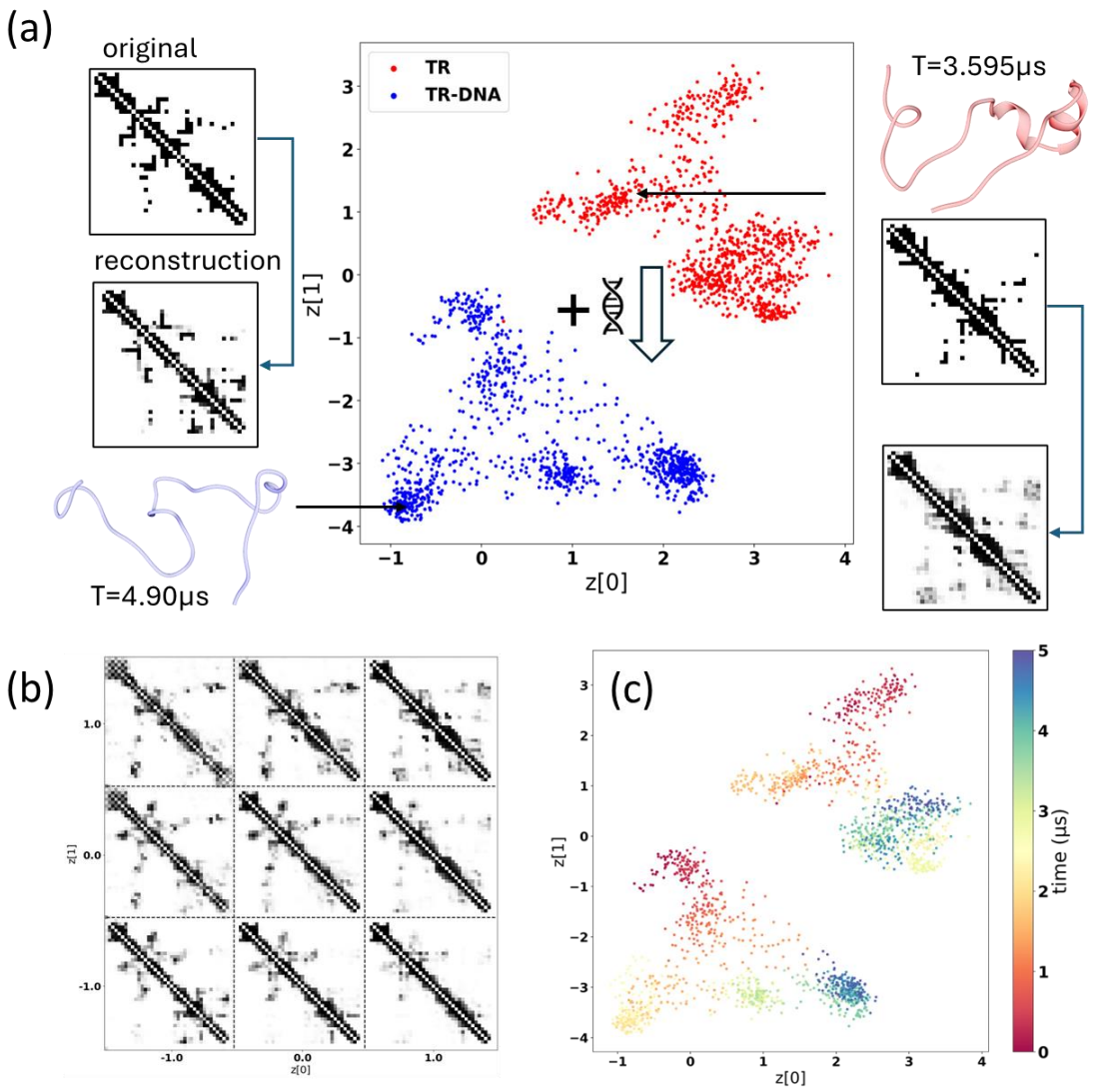
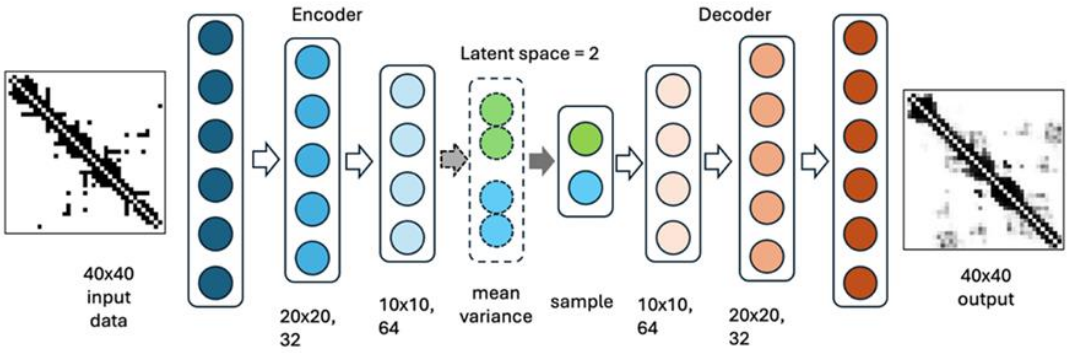


Figure 3. 4 The dynamics of hinge domain expressed using contact degrees of freedom. (a) Individual conformations projected by two latent variables using VAE scheme. Red for TR in the absence of DNA and blue for TR in the presence of DNA. Two representative contact maps and corresponding 3D structures (N-termini at upper-left) are shown. (b) The latent space of this VAE scheme. (c) The dynamic evolution of the hinge region with and without DNA.



Model: "encoder"

Layer (type)	Output Shape	Param #
input_layer (InputLayer)	(None, 40, 40, 1)	0
conv2d (Conv2D)	(None, 20, 20, 32)	320
conv2d_1 (Conv2D)	(None, 10, 10, 64)	18,496
flatten (Flatten)	(None, 6400)	0
dense (Dense)	(None, 32)	204,832
z_mean (Dense)	(None, 2)	66
z_log_var (Dense)	(None, 2)	66
sampling (Sampling)	(None, 2)	0

Model: "decoder"

Layer (type)	Output Shape	Param #
input_layer_1 (InputLayer)	(None, 2)	0
dense_1 (Dense)	(None, 6400)	19,200
reshape (Reshape)	(None, 10, 10, 64)	0
conv2d_transpose (Conv2DTranspose)	(None, 20, 20, 64)	36,928
conv2d_transpose_1 (Conv2DTranspose)	(None, 40, 40, 32)	18,464
conv2d_transpose_2 (Conv2DTranspose)	(None, 40, 40, 1)	289

Figure 3. 5 Schematic of the variational auto-encoder (VAE) workflow. The hinge contact matrix derived from residues 76–112 (37×37) is expanded to a 40×40 input by padding 1 row/column at the top and 2 at the bottom. Convolutional operations are performed with a kernel size of 3 and a stride of 2 throughout the encoder.

The VAE architecture (Figure 3.5) consisted of two convolutional neural network (CNN) layers in both encoder and decoder, each with kernel size 3×3 and stride 2. After the convolutional layers, a dense layer of 32 neurons was applied on each side, followed by a latent space containing two stochastic neurons. This compact latent representation allowed the model to capture the dominant modes of hinge motion while preserving interpretability.

The training dataset included 2,000 snapshots sampled every 5 ns (1,000 from the DNA-free simulation and 1,000 from the DNA-bound simulation). By balancing input ensembles, the VAE learned to compress high-dimensional contact patterns into low-dimensional latent variables while maintaining the ability to reconstruct representative contact maps through the decoder.

Using this setup, the latent space projection revealed a clear separation between DNA-free and DNA-bound ensembles. In DNA-free states, hinge conformations occupied a distinct region characterized by near-diagonal contacts, consistent with the presence of α-helical fragments. Upon DNA binding, the ensemble shifted to another region of latent space where these features diminished, reflecting a helix-to-unwound coil transition. Representative decoded contact maps corroborated this interpretation, showing loss of diagonal contact density and emergence of extended conformations. The temporal evolution of latent variables indicated that this transition occurs within approximately 1 μs, consistent with secondary structure analyses, and was observed across different DNA response elements, confirming the robustness of this transition.

I further investigated how the inclusion or exclusion of backbone contacts, represented by the near-diagonal elements of the contact matrix, affects the resolution of conformational states in the VAE latent space. When analyzing the entire protein rather than just the hinge, retaining all contacts caused the latent projection to collapse into a nearly linear distribution, where DNA-free TRα occupied the lower end of the z[0] axis. As shown in figure 3.6, after removing contacts within three residues along the diagonal ($|i–j| < 3$), the latent space projection remained essentially linear, suggesting that short-range backbone contacts dominated the representation and obscured global conformational differences. When a stricter cutoff was applied, excluding contacts within four residues of the diagonal ($|i–j| < 4$), the latent space began to spread out while still maintaining a clear separation between DNA-bound and DNA-free ensembles. At $|i–j| < 5$, the latent distribution organized into two approximate regression-like lines, one for DNA-free and one for DNA-bound states. These observations indicate that removing excessive backbone contributions improves the ability of the VAE to resolve large-scale conformational transitions by preventing local covalent geometry from overwhelming the signal. At the same time, retaining too few diagonal elements may reduce structural coherence. Thus, careful treatment of backbone contacts is critical for optimizing latent space interpretability, particularly when extending VAE analysis to whole-protein conformational ensembles.

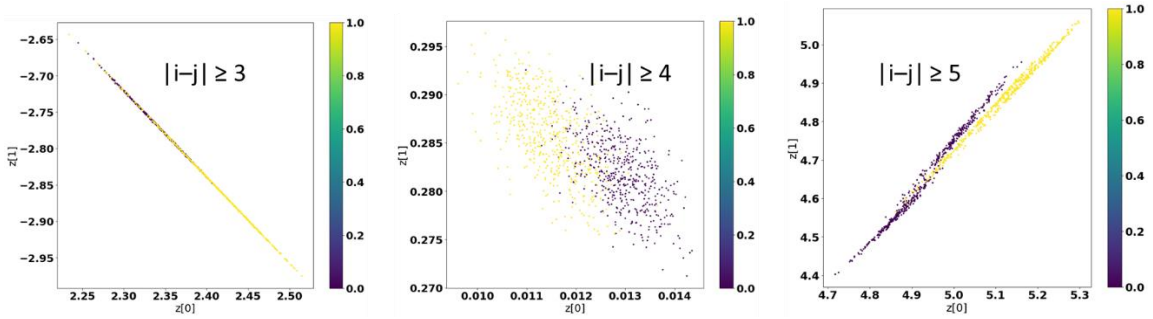


Figure 3. 6 Effect of removing TR protein near-diagonal backbone contacts on VAE latent-space organization. $|i - j| \geqslant k$ indicates that only contacts at least k residues away from the matrix diagonal (the backbone region) are retained.

However, when the analysis was restricted to the hinge region alone, the tolerance to backbone contacts had little effect on the latent space (Fig 3.7). The separation between DNA-bound and DNA-free ensembles remained robust regardless of whether near-diagonal elements were included or excluded. This suggests that for intrinsically disordered regions such as the hinge, the dominant conformational signatures captured by the VAE arise from nonlocal contact rearrangements, and are therefore less sensitive to backbone removal compared with whole-protein analyses.

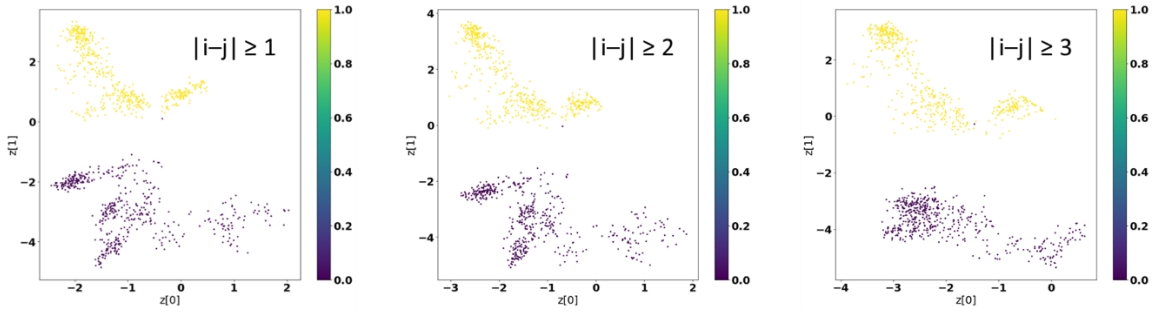


Figure 3. 7 Effect of removing Hinge region of TR near-diagonal backbone contacts on VAE latent-space organization. $|i–j| \geq k$ indicates that only contacts at least k residues away from the matrix diagonal (the backbone region) are retained.

Together, the results demonstrate that backbone contact density governs the expressive capacity of the VAE: eliminating near-diagonal elements is necessary to reveal large-scale conformational transitions in full-length TRα, but provides little benefit in the hinge, where intrinsic disorder causes nonlocal interactions to dominate the structural

landscape. Consequently, optimal VAE preprocessing must account for domain-specific contact patterns, with stricter backbone filtering required for structured regions but largely dispensable for flexible segments.

Concluding Remarks

Based on a set of 5-μs simulations of TRα (DBD-hinge-LBD) with DNA, we were able to sample how this multi-domain protein interacts with DNA. We found that the interaction is multivalent since both interactions (DBD–major groove and hinge–minor groove) make a significant contribution to the association, which in turn can be important for DNA recognition. Notably, both the recognition helix and the hinge region interact directly with nearly every base pair within the AGGTCA binding box of the TRE. At the global level, we found that the direct DBD–LBD contact interaction decreases upon DNA binding, which is consistent with previous experimental findings. The structural ensemble of hinge is strongly altered by DNA binding and the DNA structure is also distorted by the poly-Arg of the hinge. We report a helix-to-unwound transformation of the hinge upon DNA binding that can be significant for TR activation. In the absence of DNA, the hinge is more likely to adopt a helical conformation, which could favor interactions with the AF2 surface of another TR molecule. Such interactions may contribute to receptor oligomerization or higher-order assemblies, and potentially reduce the accessibility of AF2 for coactivator binding.

REFERENCES

1. Mangelsdorf, D.J., Thummel, C., Beato, M., Herrlich, P., Schütz, G., Umesono, K., Blumberg, B., Kastner, P., Mark, M., Chambon, P. *et al.* (1995) The Nuclear Receptor Superfamily: The Second Decade. *Cell*, **83**, 835-839.
2. Zilliacus, J., Carlstedt-Duke, J., Gustafsson, J.A. and Wright, A.P. (1994) Evolution of distinct DNA-binding specificities within the nuclear receptor family of transcription factors. *Proceedings of the National Academy of Sciences*, **91**, 4175-4179.
3. Sierk, M.L., Zhao, Q. and Rastinejad, F. (2001) DNA Deformability as a Recognition Feature in the RevErb Response Element. *Biochemistry*, **40**, 12833-12843.
4. Nelson, C.C., Hendy, S.C., Shukin, R.J., Cheng, H., Bruchovsky, N., Koop, B.F. and Rennie, P.S. (1999) Determinants of DNA Sequence Specificity of the Androgen, Progesterone, and Glucocorticoid Receptors: Evidence for Differential Steroid Receptor Response Elements. *Molecular Endocrinology*, **13**, 2090-2107.
5. Lohry, D.P., Stevens, T.A., Shen, T. and Fernandez, E.J. (2024) Hormone response elements for the thyroid receptor-α include specific distal 5′-flanking DNA. *Science Advances*, **10**, eadr1033.
6. Chen, X., Tsai, M.-Y. and Wolynes, P.G. (2022) The Role of Charge Density Coupled DNA Bending in Transcription Factor Sequence Binding Specificity: A Generic Mechanism for Indirect Readout. *Journal of the American Chemical Society*, **144**, 1835-1845.
7. Dickerson, R.E. and Chiu, T.K. (1997) Helix bending as a factor in protein/DNA recognition. *Biopolymers*, **44**, 361-403.
8. Rohs, R., Jin, X., West, S.M., Joshi, R., Honig, B. and Mann, R.S. (2010) Origins of Specificity in Protein-DNA Recognition. *Annual Review of Biochemistry*, **79**, 233-269.
9. Fernandez, E.J. (2018) Allosteric pathways in nuclear receptors - Potential targets for drug design. *Pharmacol Ther*, **183**, 152-159.
10. Baughman, H.E.R., Narang, D., Chen, W., Villagrán Suárez, A.C., Lee, J., Bachochin, M.J., Gunther, T.R., Wolynes, P.G. and Komives, E.A. (2022) An intrinsically disordered transcription activation domain increases the DNA binding affinity and reduces the specificity of NFκB p50/RelA. *J Biol Chem*, **298**, 102349.

11. Alberti, S. and Dormann, D. (2019) Liquid-Liquid Phase Separation in Disease. *Annu Rev Genet*, **53**, 171-194.

12. Fribourgh, J.L. and Partch, C.L. (2017) Assembly and function of bHLH–PAS complexes. *Proceedings of the National Academy of Sciences*, **114**, 5330-5332.

13. Oldfield, C.J. and Dunker, A.K. (2014) Intrinsically disordered proteins and intrinsically disordered protein regions. *Annu Rev Biochem*, **83**, 553-584.

14. Uversky, V.N., Oldfield, C.J. and Dunker, A.K. (2008) Intrinsically Disordered Proteins in Human Diseases: Introducing the D2 Concept. *Annual Review of Biophysics*, **37**, 215-246.

15. Basile, W., Salvatore, M., Bassot, C. and Elofsson, A. (2019) Why do eukaryotic proteins contain more intrinsically disordered regions? *PLoS Comput Biol*, **15**, e1007186-e1007186.

16. Geserick, C., Meyer, H.-A. and Haendler, B. (2005) The role of DNA response elements as allosteric modulators of steroid receptor function. *Molecular and Cellular Endocrinology*, **236**, 1-7.

17. Eldred, K.C., Hadyniak, S.E., Hussey, K.A., Brenerman, B., Zhang, P.-W., Chamling, X., Sluch, V.M., Welsbie, D.S., Hattar, S., Taylor, J. *et al.* (2018) Thyroid hormone signaling specifies cone subtypes in human retinal organoids. *Science*, **362**, eaau6348.

18. Papageorgiou, L., Shalzi, L., Efthimiadou, A., Bacopoulou, F., Chrousos, G.P., Eliopoulos, E. and Vlachakis, D. (2021) Conserved functional motifs of the nuclear receptor superfamily as potential pharmacological targets. *Int J Epigen*, **1**, 3.

19. Kumar, R. and Thompson, E.B. (2003) Transactivation functions of the N-terminal domains of nuclear hormone receptors: protein folding and coactivator interactions. *Mol Endocrinol*, **17**, 1-10.

20. Langlois, M.-F., Zanger, K., Monden, T., Safer, J.D., Hollenberg, A.N. and Wondisford, F.E. (1997) A Unique Role of the Beta-2 Thyroid Hormone Receptor Isoform in Negative Regulation by Thyroid Hormone. *Journal of Biological Chemistry*, **272**, 24927-24933.

21. Koenig, R.J. (1998) Thyroid hormone receptor coactivators and corepressors. *Thyroid*, **8**, 703-713.

22. Cheskis, B.J., Karathanasis, S. and Lyttle, C.R. (1997) Estrogen Receptor Ligands Modulate Its Interaction with DNA. *Journal of Biological Chemistry*, **272**, 11384-11391.

23. Putcha, B.-D.K. and Fernandez, E.J. (2009) Direct Interdomain Interactions Can Mediate Allosterism in the Thyroid Receptor. *Journal of Biological Chemistry*, **284**, 22517-22524.

24. Liu, J., Perumal, N.B., Oldfield, C.J., Su, E.W., Uversky, V.N. and Dunker, A.K. (2006) Intrinsic disorder in transcription factors. *Biochemistry*, **45**, 6873-6888.

25. Dunker, A.K., Silman, I., Uversky, V.N. and Sussman, J.L. (2008) Function and structure of inherently disordered proteins. *Current Opinion in Structural Biology*, **18**, 756-764.

26. Wang, J., Choi, J.-M., Holehouse, A.S., Lee, H.O., Zhang, X., Jahnel, M., Maharana, S., Lemaitre, R., Pozniakovsky, A., Drechsel, D. *et al.* (2018) A Molecular Grammar Governing the Driving Forces for Phase Separation of Prion-like RNA Binding Proteins. *Cell*, **174**, 688-699.e616.

27. Tong, X., Tang, R., Xu, J., Wang, W., Zhao, Y., Yu, X. and Shi, S. (2022) Liquid–liquid phase separation in tumor biology. *Signal Transduction and Targeted Therapy*, **7**, 221.

28. Thiyagarajan, T., Ponnusamy, S., Hwang, D.-J., He, Y., Asemota, S., Young, K.L., Johnson, D.L., Bocharova, V., Zhou, W., Jain, A.K. *et al.* (2023) Inhibiting androgen receptor splice variants with cysteine-selective irreversible covalent inhibitors to treat prostate cancer. *Proceedings of the National Academy of Sciences*, **120**, e2211832120.

29. Mehta, S. and Zhang, J. (2022) Liquid–liquid phase separation drives cellular function and dysfunction in cancer. *Nature Reviews Cancer*, **22**, 239-252.

30. Helsen, C. and Claessens, F. (2014) Looking at nuclear receptors from a new angle. *Molecular and Cellular Endocrinology*, **382**, 97-106.

31. Rastinejad, F., Ollendorff, V. and Polikarpov, I. (2015) Response to Moras et al. *Trends in Biochemical Sciences*, **40**, 290-292.

32. Moras, D., Billas, I.M.L., Rochel, N. and Klaholz, B.P. (2015) Structure–function relationships in nuclear receptors: the facts. *Trends in Biochemical Sciences*, **40**, 287-290.
33. Rastinejad, F., Ollendorff, V. and Polikarpov, I. (2015) Nuclear receptor full-length architectures: confronting myth and illusion with high resolution. *Trends Biochem Sci*, **40**, 16-24.
34. Chandra, V., Wu, D., Li, S., Potluri, N., Kim, Y. and Rastinejad, F. (2017) The quaternary architecture of RARβ–RXRα heterodimer facilitates domain–domain signal transmission. *Nature Communications*, **8**, 868.
35. Chandra, V., Huang, P., Hamuro, Y., Raghuram, S., Wang, Y., Burris, T.P. and Rastinejad, F. (2008) Structure of the intact PPAR-γ–RXR-α nuclear receptor complex on DNA. *Nature*, **456**, 350.
36. Rochel, N., Ciesielski, F., Godet, J., Moman, E., Roessle, M., Peluso-Iltis, C., Moulin, M., Haertlein, M., Callow, P., Mély, Y. *et al.* (2011) Common architecture of nuclear receptor heterodimers on DNA direct repeat elements with different spacings. *Nature Structural &Amp; Molecular Biology*, **18**, 564.
37. Orlov, I., Rochel, N., Moras, D. and Klaholz, B.P. (2012) Structure of the full human RXR/VDR nuclear receptor heterodimer complex with its DR3 target DNA. *The EMBO Journal*, **31**, 291-300.
38. Haelens, A., Tanner, T., Denayer, S., Callewaert, L. and Claessens, F. (2007) The Hinge Region Regulates DNA Binding, Nuclear Translocation, and Transactivation of the Androgen Receptor. *Cancer Research*, **67**, 4514-4523.
39. Miyamoto, T., Kakizawa, T., Ichikawa, K., Nishio, S., Takeda, T., Suzuki, S., Kaneko, A., Kumagai, M., Mori, J., Yamashita, K. *et al.* (2001) The role of hinge domain in heterodimerization and specific DNA recognition by nuclear receptors. *Mol Cell Endocrinol*, **181**, 229-238.
40. Nascimento, A.S., Dias, S.M., Nunes, F.M., Aparício, R., Ambrosio, A.L., Bleicher, L., Figueira, A.C., Santos, M.A., de Oliveira Neto, M., Fischer, H. *et al.* (2006) Structural rearrangements in the thyroid hormone receptor hinge domain and their putative role in the receptor function. *Journal of molecular biology*, **360**, 586-598.

41. Maletta, M., Orlov, I., Roblin, P., Beck, Y., Moras, D., Billas, I.M.L. and Klaholz, B.P. (2014) The palindromic DNA-bound USP/EcR nuclear receptor adopts an asymmetric organization with allosteric domain positioning. *Nature Communications*, **5**, 4139.

42. Lou, X., Toresson, G., Benod, C., Suh, J.H., Philips, K.J., Webb, P. and Gustafsson, J.-A. (2014) Structure of the retinoid X receptor α–liver X receptor β (RXRα–LXRβ) heterodimer on DNA. *Nature Structural & Molecular Biology*, **21**, 277-281.

43. van der Lee, R., Buljan, M., Lang, B., Weatheritt, R.J., Daughdrill, G.W., Dunker, A.K., Fuxreiter, M., Gough, J., Gsponer, J., Jones, D.T. *et al.* (2014) Classification of Intrinsically Disordered Regions and Proteins. *Chemical Reviews*, **114**, 6589-6631.

44. Lin, K.-H., Parkison, C., McPhie, P. and Cheng, S.-y. (1991) An Essential Role of Domain D in the Hormone-Binding Activity of Human β1 Thyroid Hormone Nuclear Receptor. *Molecular Endocrinology*, **5**, 485-492.

45. Johnson, Q.R., Lindsay, R.J., Nellas, R.B., Fernandez, E.J. and Shen, T. (2015) Mapping Allostery through Computational Glycine Scanning and Correlation Analysis of Residue–Residue Contacts. *Biochemistry*, **54**, 1534-1541.

46. Rastinejad, F., Perlmann, T., Evans, R.M. and Sigler, P.B. (1995) Structural determinants of nuclear receptor assembly on DNA direct repeats. *Nature*, **375**, 203-211.

47. Putcha, B., Wright, E., Brunzelle, J. and Fernandez, E. (2012) Structural basis for negative cooperativity within agonist-bound TR:RXR heterodimers. *Proceedings of the NAtional Academy of Sciences U.S.A.*, 6084-6087.

48. Schrodinger, LLC. (2015).

49. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A. *et al.* (2021) Highly accurate protein structure prediction with AlphaFold. *Nature*, **596**, 583-589.

50. Fiser, A., Do Richard Kinh, G. and Šali, A. (2008) Modeling of loops in protein structures. *Protein Science*, **9**, 1753-1773.

51. Katz, R.W. and Koenig, R.J. (1993) Nonbiased Identification of DNA Sequences That Bind Thyroid Hormone Receptor alpha-1 with High affinity. *The Journal of Biological Chemistry*, 1932-1937.

52. Koenig, R.J., Brent, G.A., Warne, R.L., Larsen, P.R. and Moore, D.D. (1987) Thyroid hormone receptor binds to a site in the rat growth hormone promoter required for induction by thyroid hormone. *Proc Natl Acad Sci U S A*, **84**, 5670-5674.

53. Macke, T.J., Svrcek-Seiler, W.A., Brown, R.A. and Case, D.A. (2006) *NAB (Nucleic Acid Builder) User's Manual*.

54. Cassandri, M., Smirnov, A., Novelli, F., Pitolli, C., Agostini, M., Malewicz, M., Melino, G. and Raschellà, G. (2017) Zinc-finger proteins in health and disease. *Cell Death Discovery*, **3**, 17071.

55. Landau, L.D. and Lifshitz, E.M. (2013) *Statistical Physics*. Butterworth-Heinemann.

56. Sethi, A., Tian, J., Vu, Dung M. and Gnanakaran, S. (2012) Identification of Minimally Interacting Modules in an Intrinsically Disordered Protein. *Biophysical Journal*, **103**, 748-757.

57. Kolinski, A., Godzik, A. and Skolnick, J. (1993) A general method for the prediction of the three dimensional structure and folding pathway of globular proteins: Application to designed helical proteins. *The Journal of Chemical Physics*, **98**, 7420-7433.

58. Johnson, Q.R., Lindsay, R.J. and Shen, T. (2018) CAMERRA: An analysis tool for the computation of conformational dynamics by evaluating residue–residue associations. *Journal of Computational Chemistry*, 1568-1578.

59. Foutch, D., Pham, B. and Shen, T. (2021) Protein conformational switch discerned via network centrality properties. *Computational and Structural Biotechnology Journal*, **19**, 3599-3608.

60. Brunton, S.L. and Kutz, J.N. (2019) *Data-Driven Science and Engineering: Machine Learning, Dynamical Systems, and Control*. Cambridge University Press.

61. Chollet, F. and Chollet, F. (2021) *Deep Learning with Python, Second Edition*. Manning.

62. Das, P., Golloshi, R., McCord, R.P. and Shen, T. (2020) Using contact statistics to characterize structure transformation of biopolymer ensembles. *Physical Review E*, **101**, 012419.
63. Lindsay, R.J., Pham, B., Shen, T. and McCord, R.P. (2018) Characterizing the 3D structure and dynamics of chromosomes and proteins in a common contact matrix framework. *Nucleic acids research*, **46**, 8143-8152.
64. Shen, T., Zong, C., Hamelberg, D., Andrew Mccammon, J. and Wolynes, P.G. (2005) The folding energy landscape and phosphorylation: modeling the conformational switch of the NFAT regulatory domain. *The FASEB Journal*, **19**, 1389-1395.
65. Johnson, Q.R., Lindsay, R.J., Nellas, R.B. and Shen, T. (2016) Pressure-induced conformational switch of an interfacial protein. *Proteins: Structure, Function, and Bioinformatics*, **84**, 820-827.
66. Pham, B., Lindsay, R.J. and Shen, T. (2019) Effector-Binding-Directed Dimerization and Dynamic Communication between Allosteric Sites of Ribonucleotide Reductase. *Biochemistry*, **58**, 697-705.
67. Clark, A.K., Wilder, J.H., Grayson, A.W., Johnson, Q.R., Lindsay, R.J., Nellas, R.B., Fernandez, E.J. and Shen, T. (2016) The Promiscuity of Allosteric Regulation of Nuclear Receptors by Retinoid X Receptor. *The Journal of Physical Chemistry B*, **120**, 8338-8345.
68. Lindsay, R.J., Mansbach, R.A., Gnanakaran, S. and Shen, T. (2021) Effects of pH on an IDP conformational ensemble explored by molecular dynamics simulation. *Biophysical Chemistry*, **271**, 106552.
69. McInnes, L., Healy, J. and Melville, J. (2020) UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv*, doi.org/10.48550/arXiv.1802.03426.
70. van der Maaten, L. and Hinton, G. (2008) Visualizing Data using t-SNE. *Journal of Machine Learning Research*, **9**, 2579-2605.
71. Atilgan, C., Okan, O.B. and Atilgan, A.R. (2012) In Rees, D. C. (ed.), *Annual Review of Biophysics, Vol 41*, Vol. 41, pp. 205-225.

72. Amitai, G., Shemesh, A., Sitbon, E., Shklar, M., Netanely, D., Venger, I. and Pietrokovski, S. (2004) Network analysis of protein structures identifies functional residues. *Journal of molecular biology*, **344**, 1135-1146.

73. Newman, M. (2018) *Networks*. OUP Oxford.

74. Zhang, P., Wang, J., Li, X., Li, M., Di, Z. and Fan, Y. (2008) Clustering coefficient and community structure of bipartite networks. *Physica A: Statistical Mechanics and its Applications*, **387**, 6869-6875.

75. Weikum, E.R., Liu, X. and Ortlund, E.A. (2018) The nuclear receptor superfamily: A structural perspective. *Protein Science*, **27**.

76. Holmbeck, S.M.A., Foster, M.P., Casimiro, D.R., Sem, D.S., Dyson, H.J. and Wright, P.E. (1998) High-resolution solution structure of the retinoid X receptor DNA-binding domainaaEdited by F. Cohen. *Journal of molecular biology*, **281**, 271-284.

77. El Hassan, M.A. and Calladine, C.R. (1998) Two distinct modes of protein-induced bending in DNA. *Journal of molecular biology*, **282**, 331-343.

78. Drew, H.R., Wing, R.M., Takano, T., Broka, C., Tanaka, S., Itakura, K. and Dickerson, R.E. (1981) Structure of a B-DNA dodecamer: conformation and dynamics. *Proceedings of the National Academy of Sciences*, **78**, 2179-2183.

79. Phan, T.Q., Jow, M.M. and Privalsky, M.L. (2010) DNA recognition by thyroid hormone and retinoic acid receptors: 3,4,5 rule modified. *Molecular and Cellular Endocrinology*, **319**, 88-98.

80. Kollman, P.A., Massova, I., Reyes, C., Kuhn, B., Huo, S., Chong, L., Lee, M., Lee, T., Duan, Y., Wang, W. *et al.* (2000) Calculating Structures and Free Energies of Complex Molecules: Combining Molecular Mechanics and Continuum Models. *Accounts of Chemical Research*, **33**, 889-897.

81. Frank, F., Okafor, C.D. and Ortlund, E.A. (2018) The first crystal structure of a DNA-free nuclear receptor DNA binding domain sheds light on DNA-driven allostery in the glucocorticoid receptor. *Scientific Reports*, **8**, 13497.

82. van Tilborg, M.A.A., Lefstin, J.A., Kruiskamp, M., Teuben, J.M., Boelens, R., Yamamoto, K.R. and Kaptein, R. (2000) *Mutations in the glucocorticoid receptor DNA-binding domain mimic an allosteric effect of DNA11Edited by P. E. Wright*.

83. Watts, D.J. and Strogatz, S.H. (1998) Collective dynamics of ‘small-world’ networks. *Nature*, **393**, 440-442.

84. Kabsch, W. and Sander, C. (1983) Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577-2637.

85. McInerney, E.M., Rose, D.W., Flynn, S.E., Westin, S., Mullen, T.M., Krones, A., Inostroza, J., Torchia, J., Nolte, R.T., Assa-Munt, N. *et al.* (1998) Determinants of coactivator LXXLL motif specificity in nuclear receptor transcriptional activation. *Genes Dev*, **12**, 3357-3368.

86. Tien, A.H. and Sadar, M.D. (2018) Order within a Disordered Structure. *Structure*, **26**, 4-6.

APPENDIX

Supplementary Information

TR w/o DNA	DBD	Hinge	LBD
DBD	428.6	-	-
Hinge	12.0	152.9	-
LBD	49.7	39.3	1539.8

TR w/ DNA	DBD	Hinge	LBD
DBD	424.3	-	-
Hinge	21.5	148.1	-
LBD	23.7	52.1	1528.8

Figure S3. 1 Overall interaction strength within and between domains indicated by block summation of the mean contact matrix.

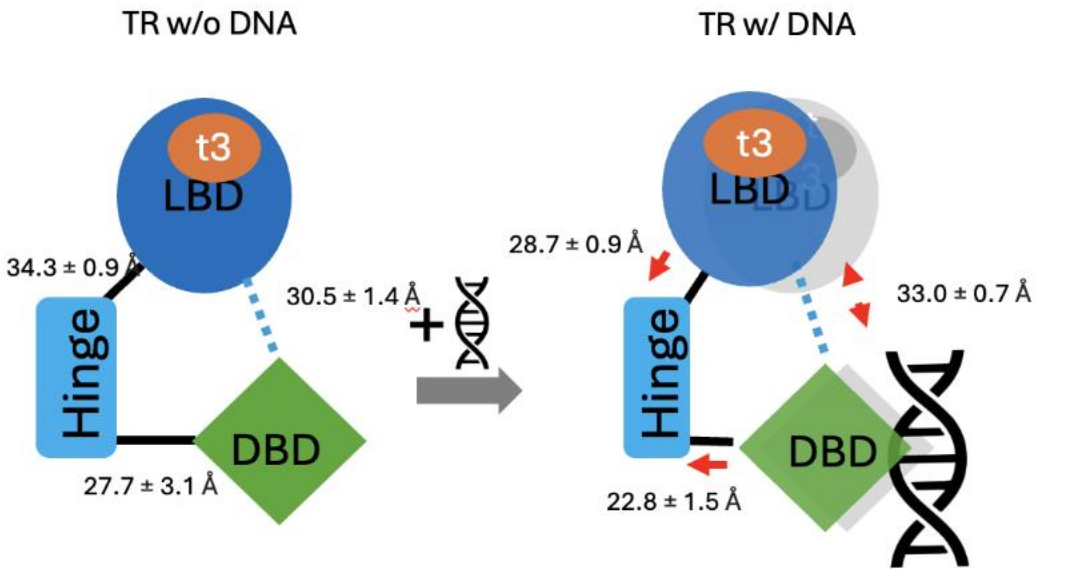


Figure S3. 2 Mean distances between TR domains with and without DNA.

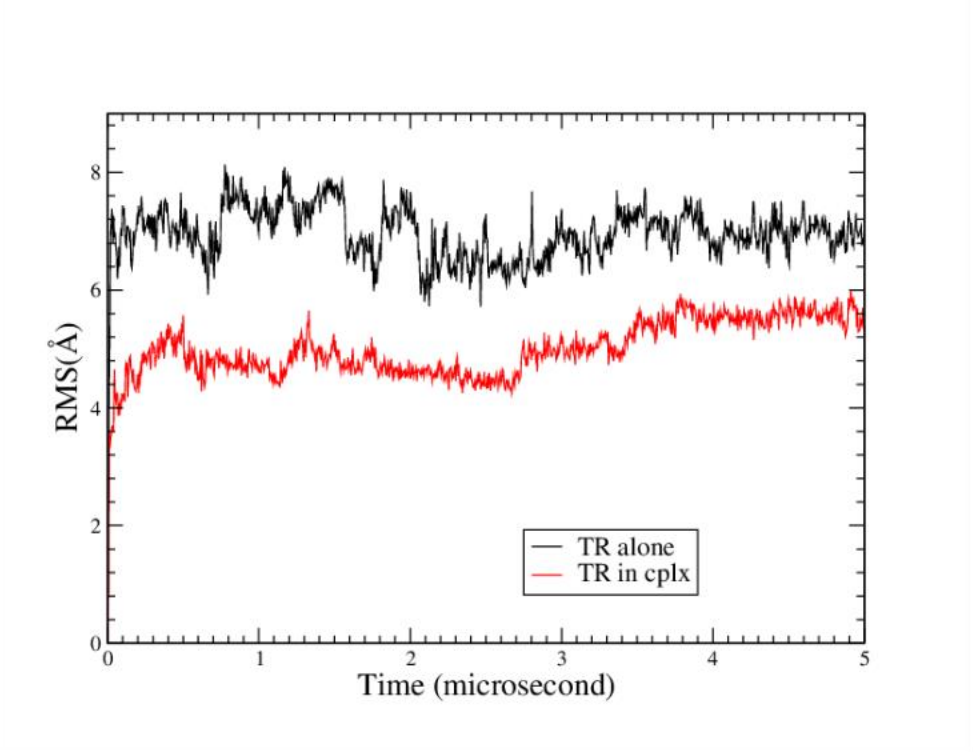


Figure S3. 3 Root-mean-square deviation of TR structure as a function of time.

CHAPTER IV

COMPUTATIONAL ASSESSMENT OF THE ASSOCIATION BETWEEN CURAXIN AND SELECTED DNA STRUCTURES IN GENOME

Abstract:

Curaxins are small chemicals that interact with the genome and slow down tumor growth. Numerous studies suggested curaxins are structurally selective DNA drugs that directly interact with DNA duplex. However, the detailed mechanism is still poorly understood, especially the issue of specificity. In this study, we compared the interaction between a set of curaxins of various levels of reported effectiveness (CBLC-137, CBLC-000, CBLC-100, CBLC-101, CBLC-120, CBLC-136, and quinacrine) and a few representative open dsDNA binding sites such as the minor groove of standard B-DNA and the major groove of a bent form of B-DNA. We first used molecular docking to explore possible binding modes. Selected binding poses were then subjected to atomistic molecular dynamics simulations and binding free energy calculations for high-resolution characterization. We identified specific curaxin-DNA binding modes including ones that are nucleobase-dependent. Both docking scores and energy calculations indicate that curaxins exhibit DNA shape selectivity, with a primary binding preference for the major groove of bent DNA, and a secondary preference for the minor groove of canonical B-DNA. From the ligand perspective, the main structural difference of the binding modes between relatively effective curaxins (such as CBLC-101 and CBLC-137) and ineffective ones (such as CBLC-136) might be the carboxyl containing "arms" of curaxins. The two arms of the ligand anchor between base pairs separated by a small number of base steps along the DNA helix. These polar arms form hydrogen bonds with specific nucleobases, potentially explaining the base-pair preferences reported in experimental studies. The results of curaxin-DNA interaction provide insights on how curaxin might interact with important genome structures and potentially how they might influence different transcription factors such as FACT complex (minor groove binder) and CTCF (major groove binder).

Introduction

While the bulk of small-molecule drugs target receptors and enzymes, there are a few cases that exploit the binding pockets presented by nucleic acid structures with various degrees of selectivity (278). Curaxins are a group of small synthetic chemicals that were designed to interrupt cancer cell activity at the level of gene regulation. They originated

from a focused structure-activity relationship study, where the structure of quinacrine was modified to create CBLC-000 and other analogs with enhanced activity in cell-based p53 and NF-κB reporter assays (279). Although the exact action of curaxin on transcription activities is still being studied, numerous evidence suggested a direct curaxin association with DNA which subsequently influences the interactions between DNA and transcription factors (TFs) (90). One important difference between curaxins and many other DNA-binding drugs is that curaxins are considered non-genotoxic (279). Other frequently mentioned DNA-binding anticancer drugs, such as doxorubicin (280), cyclophosphamide (281), and cisplatin (77) induce DNA damage, and researchers believe that inducing DNA damage is one of their anti-cancer mechanisms. These drugs can irreversibly modify DNA. In contrast, curaxins do not cause DNA damage, such as phosphodiester bond cleavage, covalent adduct formation, or chemical modification of nucleobases. Instead, curaxins are thought to interact with the DNA duplex in a mild and reversible manner, with sensitivity to local DNA sequence and structural context. Despite the absence of genotoxic effects, such interactions can stabilize distorted DNA conformations and/or introduce steric constraints, thereby modulating the binding of specific transcription factors to DNA.

One central characterization of conventional (Watson-Crick pairing) duplex DNA binding is how small chemicals and proteins position themselves relative to dsDNA. There are generally three types: major groove binders, minor groove binders, and even intercalators in the case of small ligands (282-284). The major groove and minor groove of dsDNA have distinct chemistry. The major groove potentially exposes the N7 positions of adenine and guanine, as well as the N4 and O4 atoms of cytosine and thymine, whereas the minor groove presents the N3 positions of adenine and guanine and the O2 atoms of cytosine and thymine. Interactions with the phosphate backbone and base-pair stacking during intercalation are largely nonspecific. Therefore, specific DNA base–drug recognition is likely mediated by polar interactions with characteristic nucleobase atoms exposed in the DNA grooves.

There is limited experimental or computational characterization of structure and binding energetics. Since intercalator are generally considered irreversible, curaxins are likely

groove binders. Which groove do curaxins bind to? The initial equilibrium dialysis assay did not resolve the specific binding site and reported binding in terms of an apparent association constant, $K_{\text{app}} = ([C_{\text{bound}}]/[C_{\text{free}}])/[NA]$, where $[NA]$represents the total concentration of nucleic acid base pairs (2,10). However, $K_{\text{app}}$is not equivalent to the true association constant K_a, because $[NA]$reflects the concentration of base pairs rather than the concentration of available binding sites on DNA, and the binding stoichiometry is unknown. A later introduced correction term ([*NA*]- [C_{bound}]) (285) still potentially overestimates the concentration of available binding sites as each ligand mostly definitely occupies more than 1 *NA* unit. Nevertheless, as a comparative method, this equilibrium dialysis approach consistent results compared to spectroscopy methods. The measured interaction strength of curaxin turns out to be comparable but weaker than intercalator ethidium(286). Meanwhile, the molecular docking study was limited to just one form of curaxin (CBLC-137) to just one type of binding site: minor groove of B-DNA (287). Additional docking studies focused on binding sites that are not a groove of DNA duplex, such as the ends of DNA or a quadruplex structure(288). However, one would need to compare different binding pockets and investigate the selectivity of DNA conformation besides sequencing specificity. Indeed, it remains puzzling how curaxin can function as reported with a relatively weak affinity but strong selectivity to cancer genome versus normal genome.

Based on the fact that curaxin has a strong effect on cancerous cell while does not affect normal cells, a more sensible assumption is that, regardless of the site identity, the targeting sites might not be too populated in normal DNAs and further can have a significant population difference between cancerous and normal cells whereas the usual sites such as the minor groove of standard B form of DNA is populated in both cell types and would not make a noticeable difference. An additional rationale comes from the facts on how curaxin affects preferentially to specific transcription factors such as FACT and CTCF. Meanwhile, protein complex FACT preferentially binds to distorted form of dsDNA to examine irregular conformations of DNA. Thus, it is important to compare curaxins and representative DNA duplex binding sites, such as the major groove side of a

bent DNA, where the alternative binding mode is not dissimilar to that of cisplatin, even though the latter leads to DNA adduct (289).

It is interesting to point out that curaxins have many potential targeting protein-DNA interactions. Besides the closed DNA state (nucleosome where DNA is complexed with histone proteins), there were reports on several types of open-DNA binding that are affected by curaxins. The most frequently reported are transcription factors Facilitates Chromatin Transcription (FACT) (290) and CCCTC-binding factor (CTCF) (291). Human FACT is composed of two subunits, Spt16 and SSRP1, and plays a key role in regulating gene expression within cells. Its primary function is to reorganize nucleosomes, keeping the chromatin structure open during transcription (290). Additionally, FACT also participates in the DNA damage response by interacting with other repair proteins such as RAD51 and BRCA1, contributing to the cell's mechanisms for repairing DNA damage (292-294). FACT levels are upregulated in many cancer cells (293,295). Genome-wide mapping of chromatin-bound FACT indicated that FACT's role in cancer most likely involves selective chromatin remodeling of genes that stimulate proliferation, inhibit cell death and differentiation, and regulate cellular stress responses (296). The intervention of curaxin will inhibit the function of FACT in cells(279,293,297). Except for less differentiated cells such as stem cells (298,299), normal cells are not sensitive to changes in FACT levels (300). Therefore, the inhibition of FACT function triggered by curaxin is more toxic to cancer cells than to normal cells. Having said that, how do curaxins inhibit FACT's function remains unknown. In vitro single-molecule experiments show that after curaxin binds to DNA, the affinity of FACT to DNA is enhanced. 2 μM CBLC-137 reduced the K_D of FACT-DNA from 120 nM to 25 nM. As a DNA-binding factor, increased affinity seems to contradict functional inhibition. There is a theory for this called c-trapping. Some researchers have found that curaxin converts DNA from the B form to the Z form. The FACT complex targets the Z-DNA and becomes trapped on it (88). However, this theory still believes that FACT recognizes and binds DNA due to the loss of histones, ignoring the change in the affinity between DNA and FACT itself. And considering that FACT has so many functions, even if their hypothesis is correct, this may not be the only pathway how FACT functions are

affected by curaxins. Fortunately, new molecular biology and clinical findings provide us with new ideas to explain the anti-cancer mechanism of curaxin when targeting FACT. This involves another important function of FACT, DNA damage repair. Research has shown that FACT interacts with cisplatin-modified DNA in vitro, an interaction that is contingent upon the high-mobility group (HMG) domain of the SSRP1 subunit (301). HMG group, also called HMG box, is a domain in the FACT complex specifically responsible for recognizing and binding DNA (302-304). It has been established that the HMG box can bind to the DNA helix at kinked regions (305), providing insight into how FACT can bind to DNA damaged by agents such as cisplatin and UV light. Recent studies further reveal that FACT accumulates at sites of various types of DNA damage, including oxidative and alkylated base damage (294,306), single-strand breaks (SSBs) and double-strand breaks (DSBs) (307,308), as well as UV-induced damage (309). CCCTC-binding factor (CTCF) is an essential protein that plays a critical role in maintaining the 3D architecture of chromatin, which in turn regulates gene expression and cellular differentiation(291). CTCF includes a long series of 11 zinc finger (ZF) motifs (310). Each zinc finger typically consists of approximately 30 amino acids and coordinates one zinc ion. CTCF functions as a multifunctional transcriptional regulator, known for its insulator activity, thereby preventing the interaction between enhancers and promoters when bound to DNA (89,311). Mutations in CTCF, particularly within its zinc finger (ZF) domain, have been linked to altered DNA binding affinities, resulting in disrupted chromatin structure and abnormal gene regulation, which are often observed in various cancers (312,313).

Curiously, curaxin CBLC-137 was reported to enhance the FACT activity and weakened CTCF interaction (90). Meanwhile, FACT is known to bind to minor groove while CTCF is known to be at major groove. Thus, a major groove binding model that is consistent with this information is that curaxin binds to major groove and thus support FACT's minor groove binding, just like cisplatin (314). On the other hand, if curaxin acts as a major groove binder will be a competitive inhibitor of CTCF binding and thus weakens CTCF binding. To quantitatively compare various modes of curaxin-DNA interaction, we conduct molecular docking and atomistic simulation to evaluate the interaction strength.

We also examined how DNA changes its conformation upon curaxin binding in selected modes. This information can provide insights on how DNA-binding drugs can affect DNA conformations and/or modulate protein-DNA interactions.

Methods and systems:

Previous computational studies on curaxin-DNA duplex interactions were limited to docking study of one form of curaxin (CBLC-137) with one form of DNA (minor groove of standard B-DNA). Here we compared a set of ligands and a set of duplex DNA conformation. We begin the discussion with the relatively simple task of listing ligands. The selection of relevant ligands mentioned in this study is shown in Table I. The rationale of selecting these ligands is the following. Ligand CBLC-137 is the most promising and widely tested curaxin and it is proved to be both more effective and less toxic for its tumor suppressing roles (279). Meanwhile, CBLC-000 and CBLC-100 also demonstrated potency compared to the original molecule, quinacrine (QC) (279). CBLC-100 shows one of the strongest affinities to DNA but was commonly used as it is deemed to be more toxic. Nevertheless, it is useful to study in order to reveal the binding mechanism. For a comparison, we also include a couple ineffective ones, such as CBLC-120 and CBLC-136 (279).

In contrast to the handful experimented curaxins, there are large variety of possible DNA binding site structures existing in genome. We certainly cannot test all possibilities and thus picked a commonly discussed set and limited our discussion to the ones that are must be Watson-Crick pairing. Certainly, more exotic and higher (free) energy structures, such as Hoogsteen paired structure as a binding site is possible in principle, though in order to compete with stronger affinity to curaxin, they must have a much stronger enthalpy of binding to compensate for the entropy disadvantage of being much less populated in genome structures. A list of DNA binding sites which we studied here are shown in Table II.

AutoDock Vina was used for molecular docking (315,316). The three-dimensional structures of the DNA receptors were obtained from the PDB database. The structures of all ligands including curaxins were built using the JMSE molecular editor (317) and CACTUS Online SMILES Translator. Before docking, the receptor and ligand structures

were preprocessed using AutoDockTools, including adding hydrogen atoms, assigning partial charges, and energy minimization (318,319). During docking, the grid box wraps around the DNA except for the head and tail. This is to prevent the ligand from binding to the open ends of the DNA to simulate the state of the ligand binding to the complete long-chain DNA in cells. The docking posture is evaluated using the scoring function of AutoDock Vina. The top-ranked poses were analyzed based on their binding energies. The most favorable pose was selected for further analysis.

For the molecular dynamics (MD) simulation, the amber force field was used. Here the nucleic acid force field is amber ff99bsc1 (320). For the unique ligands used in this work, quantum chemistry calculations for the force-field parameters of curaxins (CBLC-137 and CBLC-136) were performed using the AMBER's ANTECHAMBER module with the AMB-BCC charge method. Parmcheck2 was followed to construct "*.prep" and "*.frcmod" files. For counter ions, we choose to use K+ instead of Na+. There were reported specific ion effects in DNA systems (321), and K+ is more biologically relatable for genome properties.

NAMD was used for the MD simulation of curaxin and DNA complexes. The topology and initial coordinate files were generated using the TLEAP form AMBER20 tools. In solvating the system, the TIP3P water solvation model was used, and potassium (K+) counter ions (2 per bp) were added to neutralize the systems. After the minimization and heating steps, the production run (NPT) was carried out using the following parameters: the temperature of the biomolecular system was set at 300 K, the time step at 2.0 fs. The Langevin dynamics was used for the temperature control and a Langevin piston with a target pressure of 1.01325 bar. Snapshots were saved at a 1ps interval. All MD simulation was conducted in the Stampede3 supercomputer at the Texas Advanced Computing Centre (TACC). A total of 50 ns simulation was performed for each system. For the cases (to be discussed in the Result section) that we keep the DNA immobilized during simulation, a constraint .pdb file for the DNA was generated with the occupancy column to 1.00 for the DNA atoms and 0.00 for the other atoms. The fixed atom parameters were set at ON for the minimization and/or heating step.

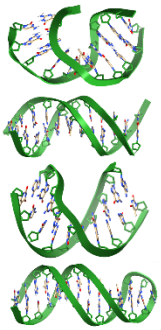
As we state in the result below that a single curaxin is not able to maintain the bent configuration without the presence of transcription factor (HMG), we have thus frozen the DNA during certain production runs that we explicitly state, and only allow the motions of the curaxin and the solvents when we sample how curaxin interacts with DNA. Note that we did not immobilize the DNA during the energy minimization procedure or other unspecified production runs such as sampling the DNA configurations without the ligand.

Implicit solvation model and MMGBSA was used to obtain the binding energy DG_MMGBSA between DNA and ligand. AMBER20 MMPBSA package was used to determine the binding free energy of the system. The modified GB model (igb=5) was used. The interior (DNA and curaxin) and exterior (implicit solvent) dielectric constants were set to 1.0 and 80.0, respectively. For additional DNA conformational analysis, such as the curvature and major and minor groove features, Curves+ webserver (322) was used to calculate the DNA geometry features.

Results and Discussion

Molecular docking

The docking results are summarized in Fig 4.1, where the best-scoring pose for each curaxin–DNA pairing is reported. Among the compounds tested, CBL0100 showed the highest predicted affinity, followed by CBL0137 and CBL0136 (CBL0100 > CBL0137 > CBL0136). DNA conformational state also had a influence on docking outcome. The curved B-form major-groove model consistently produced the most favorable scores, followed by the standard B-form minor groove and the hairpin model, with A-form and Hoogsteen conformations yielding weaker interactions (curved B > standard B > hairpin > A-form). These results indicate that both the identity of the curaxin compound and the underlying DNA topology contribute to binding energetics, offering structural insight into how curaxins may differentially engage and modulate DNA.



binding affinity (kcal/mol)	+ CBL0137	+ CBL0100	CBL0136	+ CBL000	CBL0101	CBL0102 Quinacrine	CBL0120	EB
distored-B	-7.352	-7.848	-7.049	-7.023	-6.994	-6.794	-8.255	-7.729
straight-B	-6.886	-7.327	-6.809	-6.634	-6.374	-6.049	-7.679	-6.742
A	-6.339	-6.603	-6.241	-5.956	-6.149	-5.934	-7.483	-6.936
hairpin	-6.721	-6.467	-6.749	-6.08	-6.229	-5.779	-7.634	-6.78

Figure 4. 1 Docking results from AutoDock Vina for curaxins and other DNA-binding compounds interacting with different DNA conformations. Curaxins with anticancer activity are marked with "+". For each curaxin–DNA pairing, the best-scoring pose is shown.

Besides docking scores, it is also important to study the structure features of the DNA-ligand association. We first characterize the consensus geometric features of curaxins before describing the curaxin-DNA complex structure. Overall, curaxins have three parts: central body, two arms, and a tail. For most of the curaxins, the central body is a carbazole (dibenzopyrrole) which is a tricyclic structure consisting of two six-member aromatic (benzene) rings fused to either side of a five-member nitrogen-containing (pyrrole) ring. C136 and C101 are exceptions, instead having β-carboline bodies which differ by containing a nitrogen atom located at position 2 of the tricyclic structure. The arms extend from positions 3 and 6 in symmetric curaxins and from position 3 only on asymmetric curaxins (C136 and C101). The arms are typically electron-withdrawing groups which are polar, or partially polar. For example, C137 and C000 present acetyl and nitro substituents respectively. The aminoalkyl tails extend from position N9 of the tricyclic body. Most of the curaxin tails are N-(ethyl)diethylamine. Exceptions are C120 and C137, whose tails terminate in a dimethylamine and isopropylamine, respectively.

We can largely classify the docking modes by which part(s) of the ligand (arms, tail, or body) and which parts of the DNA (base, sugar, phosphate) form interactions. It turns out that most contact interactions come from arms and/or tail with the base of DNAs. In a few cases, the arms interact with the phosphate backbone of DNA at the end of dsDNA. Here are the four populated binding modes we found: (i) Minor groove symmetric mode: two arms, symmetrical at minor groove, similar to dapi and distamycin (ii) Minor groove asymmetric mode: one arm and one tail (iii) Major groove symmetric mode: two arms, largely symmetric in major groove (iv) Major groove asymmetric mode: one arm and one tail. Two representative binding modes of CBL0137 are illustrated in Fig. 4.2. When interacting with linear B-form DNA, CBL0137 inserts either both arms or one arm together with the tail into the minor groove. In contrast, when binding to curved DNA, both arms and the tail can engage surrounding bases, allowing the ligand to establish multiple contacts with the DNA surface.

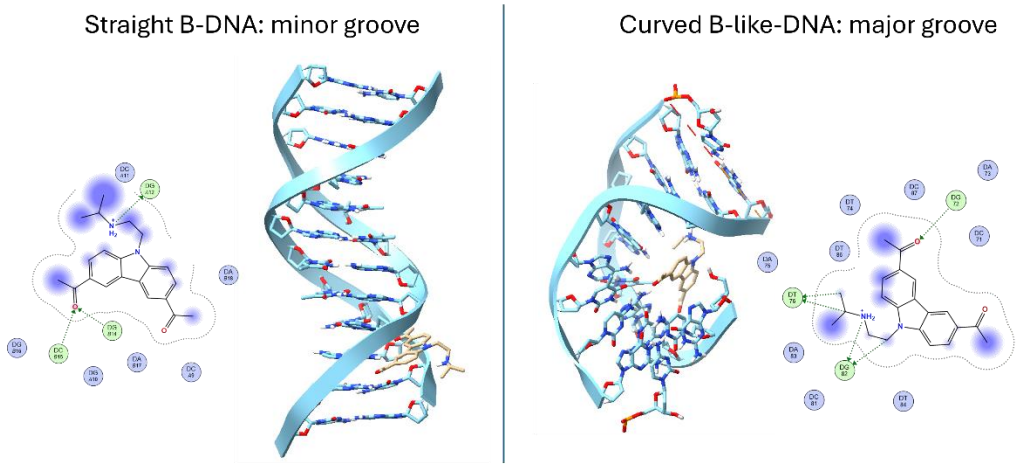


Figure 4. 2 Two representative binding modes of CBL0137 predicted by AutoDock Vina. In two-dimensional interaction diagram, blue halos indicate solvent exposure, and green arrows represent hydrogen-bond donor to acceptor interactions.

Molecular dynamics simulation and binding energy calculation

This work is still in progress. Besides using molecular docking of a set of curaxins to a set of DNA binding sites, we performed MD simulation in explicit solvents and followed free energy estimation for selected cases. The main difference between the docking scores and simulation results stems from several factors, such as the resolution of the

model (united vs all-atom), energy function (empirical vs quantum chemical-based), and solvation effect (implicit vs explicit).

Additionally, simulation includes the dynamics of DNA. It is important to point out that the dynamics of DNA duplex in MD simulations are relatively fast and sampling is not an issue when one tries to capture the relaxation dynamics of bent DNA to standard B-DNA in our 50 ns simulation. Only when researchers focus on more local and higher-energy states (such as base flipping), one may have concerns of convergence issues (323).

Genomic Analysis of Curaxin-Mediated Modulation of CTCF Binding

To investigate how curaxins modulate CTCF–DNA recognition at the sequence level, we performed a bioinformatic analysis using CTCF ChIP-seq profiles from HT1080 cells before and after CBL0137 treatment. We focused on CTCF sites located at TAD boundaries, as these regions represent strong architectural anchors that are particularly sensitive to perturbations in DNA structure. Motif discovery using MEME, combined with motif scanning via FIMO, were used to quantify the presence, strength, and positional distribution of CTCF binding motifs across regions exhibiting increased or decreased CTCF occupancy upon drug treatment.

To dissect the sequence determinants underlying curaxin-induced changes in CTCF binding, we analyzed a panel of motif variants centered on the CTCF zinc-finger recognition positions most strongly affected by drug treatment. Using MEME for de novo motif discovery and FIMO for genome-wide motif scanning, we identified two enriched consensus motifs (Figure 4.3) from HT1080 TAD-boundary CTCF peaks. From these motifs, we extracted a highly conserved core sequence that is shared between regions showing increased and decreased CTCF occupancy. We refer to this sequence (T8 = GCCMYCTRSTGG) as the baseline motif, as it represents the predominant binding configuration at untreated sites and provides a neutral reference for evaluating sequence deviations.

To test whether curaxins preferentially disrupt AT-rich motifs, we generated AT4 variants (WWGCCMYCTRSTGG and DHGCCMYCTRSTGG) and contrasting CG4 variants (SSGCCMYCTRSTGG and CGGCCMYCTRSTGG), which probe the influence

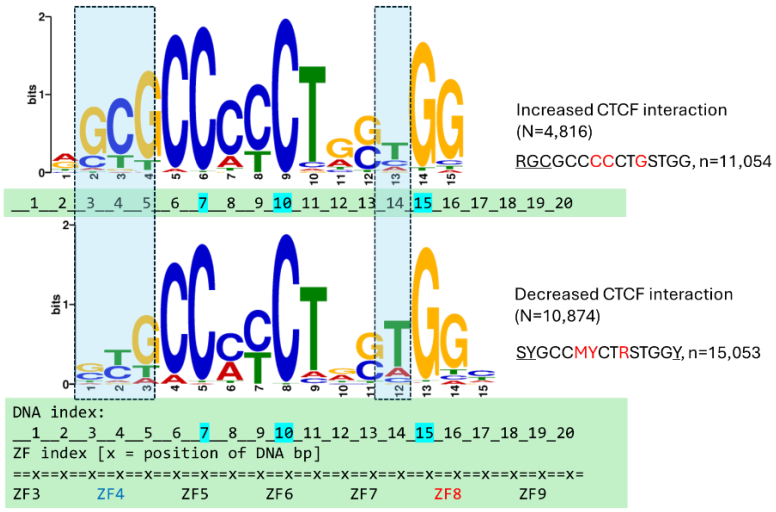


Figure 4. 3 The motif sites in HT1080 cells treated with CBL0137 that showed increased and decreased CTCF binding. The upper motif shows increased CTCF binding, while the lower motif shows decreased CTCF binding. N represents the number of detected peaks, and the corresponding CTCF zinc finger reference binding sites are shown below. Each CTCF zinc finger recognizes approximately three base pairs; however, the green boxes are shown as schematic guides rather than exact base-pair assignments. Due to motif degeneracy and alignment across multiple binding sites, individual zinc fingers may not align precisely with a fixed base-pair position, and not all zinc fingers necessarily make direct base-specific contacts in every motif context. The blue box indicates the main differences between the two motifs. The blue and red markings indicate the two zinc finger positions we are interested in.

of AT versus CG identity at the ZF4 contact position. We similarly compared T8 and NT8 (GCCMYCVRSTGG) motifs to assess whether thymine at the ZF8 position contributes to drug sensitivity. Finally, we examined the presence of the U-motif (NGGNACTGCA), previously proposed as a modulator of CTCF binding variability, to evaluate its potential involvement in curaxin responsiveness (Figure 4.4). Collectively, this curated set of motif variants enables a systematic assessment of how nucleotide identity at key zinc-finger positions shapes CTCF occupancy in response to curaxin treatment.

As shown in figure 4.4, Our analyses reveal several distinguishing features of how curaxins influence CTCF–DNA interactions. Regions showing decreased CTCF occupancy contain substantially more CTCF binding motifs than those showing increased occupancy, with ratios ranging from approximately 2:1 up to 10:1 depending on motif

strength and the applied cutoff threshold. Sequence-level examination suggests that nucleotide identity at key zinc finger positions play an important role. AT base pairs at

Motif	Increased peak (raw # | # per peak)		Decreased peak (raw # | # per peak)	
All	4,816		10,874	
Test1 (RGCGCCCCCTGSTGG)	11,053	2.30	11,708	1.08
Test2 (SYGCCMYCTRSTGGY)	9,482	1.97	15,053	1.38
Baseline (GCCMYCTRSTGG)	718	0.15	155	0.01
AT4 (WWGCCMYCTRSTGG)	33	0.01	169	0.02
CG4 (SSGCCMYCTRSTGG)	915	0.19	142	0.01
T8 (GCCMYCTRSTGG)	718	0.15	155	0.01
NT8 (GCCMYCVRSTGG)	15	0.003	21	0.002
Umotif (GCCMYCTRSTGGNN NNNNNN (5-6 N) NGGNACTGCA)	102	0.02	125	0.01

Figure 4. 4 CTCF motif variants and their genomic occurrence in HT1080 cells treated with CBL0137. Motif variants were designed to test nucleotide identity at specific zinc-finger contact positions, including AT- versus CG-rich variants at the ZF4 position, thymine versus non-thymine variants at the ZF8 position, and the U-motif. For each motif, the table reports the total number of detected motif instances across the genome (#) and the average number of motif instances per CTCF peak (#/Peak highlighted by orange). Each motif search uses its own most stringent p-value, ensuring consistency at every position.

the ZF4 recognition position are strongly associated with peaks that lose CTCF binding upon curaxin treatment, whereas CG base pairs at the same position tend to be enriched in regions showing increased binding. These observations support the hypothesis that curaxins preferentially interact with AT-rich sequences, thereby destabilizing local DNA structure and reducing CTCF association. Although thymine at the ZF7–8 positions show a weak trend toward decreased binding, the effect size is minimal and does not reach statistical significance. Furthermore, the U-motif, previously proposed as a determinant of CTCF binding variability, shows no detectable relationship to curaxin sensitivity in our dataset. Overall, the curaxin effect appears to be concentrated primarily at ZF4, with only minor and non-significant contributions from ZF8, highlighting a sequence-dependent mechanism through which curaxins can modulate CTCF–DNA interactions.

Concluding remarks

Although there's still a lot to do, across docking, molecular dynamics simulations, and genomic analyses, our study has provided some coherent, multiscale view of how curaxins interact with DNA and consequently modulate chromatin-associated processes. Molecular docking demonstrates that curaxin binding is highly sensitive to both ligand structure and DNA topology, with CBL0100 showing the strongest predicted affinity and curved B-form DNA providing the most favorable interaction environment. Structural inspection of docking poses reveals that curaxins engage DNA primarily through their aromatic arms and flexible aminoalkyl tails, adopting a small set of recurring binding modes distributed between major and minor grooves.

Molecular dynamics simulations extend these findings by capturing how curaxin–DNA complexes evolve under thermal motion. The trajectories highlight the rapid relaxation of bent DNA toward canonical B-form geometries, the dominant role of ligand reorientation in driving conformational equilibration, and the substantial increase in binding stability observed in later-stage conformations dominated by π–π stacking between the curaxin body and terminal base pairs. Free-energy estimates further support the formation of stable, energetically favorable complexes following initial structural rearrangement. DNA shape analysis indicates that curaxins induce measurable changes in groove width and local geometry, consistent with a mechanism in which physical deformation of DNA contributes to downstream regulatory effects.

Genomic analyses reveal a sequence-dependent component to curaxin activity, particularly in the context of CTCF binding. Regions that lose CTCF occupancy upon curaxin treatment contain substantially more CTCF motifs and are strongly enriched for AT base pairs at key zinc finger contact positions, especially ZF4. In contrast, CG-containing motifs at the same position are associated with increased or unchanged binding, suggesting that curaxins preferentially target AT-rich sequences in vivo. Additional zinc finger positions, such as ZF7–8, show only weak and statistically insignificant associations, and previously proposed sequence features such as the U-motif do not correlate with curaxin sensitivity. These observations indicate that curaxin-mediated disruption of CTCF binding is driven largely by selective interaction with AT-

rich DNA, which likely enhances local DNA deformability and reduces the stability of CTCF–DNA contacts.

REFERENCES

1. Deng, S., Feng, Y. and Pauklin, S. (2022) 3D chromatin architecture and transcription regulation in cancer. *J Hematol Oncol*, **15**, 49.
2. Henley, M.J. and Koehler, A.N. (2021) Advances in targeting 'undruggable' transcription factors with small molecules. *Nat Rev Drug Discov*, **20**, 669-688.
3. Bushweller, J.H. (2019) Targeting transcription factors in cancer - from undruggable to reality. *Nat Rev Cancer*, **19**, 611-624.
4. Vercellone, F., Chiariello, A.M., Esposito, A., Conte, M., Abraham, A., Fontana, A., Di Pierno, F., Tafuri, F., Guha, S., Kundu, S. *et al.* (2024) A Multiscale Perspective on Chromatin Architecture through Polymer Physics. *Physiology*, **40**, 291-300.
5. Conery, A.R., Rocnik, J.L. and Trojer, P. (2022) Small molecule targeting of chromatin writers in cancer. *Nat Chem Biol*, **18**, 124-133.
6. Hnisz, D., Day, D.S. and Young, R.A. (2016) Insulated Neighborhoods: Structural and Functional Units of Mammalian Gene Control. *Cell*, **167**, 1188-1200.
7. Bonev, B. and Cavalli, G. (2016) Organization and function of the 3D genome. *Nat Rev Genet*, **17**, 772.
8. Dixon, J.R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J.S. and Ren, B. (2012) Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, **485**, 376-380.
9. 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. *et al.* (2009) Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, **326**, 289-293.
10. Hsieh, T.H., Weiner, A., Lajoie, B., Dekker, J., Friedman, N. and Rando, O.J. (2015) Mapping Nucleosome Resolution Chromosome Folding in Yeast by Micro-C. *Cell*, **162**, 108-119.
11. Bernenko, D., Lee, S.H., Stenberg, P. and Lizana, L. (2023) Mapping the semi-nested community structure of 3D chromosome contact networks. *PLoS Comput Biol*, **19**, e1011185.

12. Wright, P.E. and Dyson, H.J. (2015) Intrinsically disordered proteins in cellular signalling and regulation. *Nat Rev Mol Cell Biol*, **16**, 18-29.
13. Tambones, I. and le Maire, A. (2024) Structural Insights Into Thyroid Hormone Receptors. *Endocrinology*, **166**.
14. Degiacomi, M.T. (2019) Coupling Molecular Dynamics and Deep Learning to Mine Protein Conformational Space. *Structure*, **27**, 1034-1040 e1033.
15. Noé, F., Tkatchenko, A., Müller, K.-R. and Clementi, C. (2020) Machine Learning for Molecular Simulation. *Annual Review of Physical Chemistry*, **71**, 361-390.
16. Pommier, Y., Leo, E., Zhang, H. and Marchand, C. (2010) DNA Topoisomerases and Their Poisoning by Anticancer and Antibacterial Drugs. *Chemistry & Biology*, **17**, 421-433.
17. Gasparian, A.V., Burkhart, C.A., Purmal, A.A., Brodsky, L., Pal, M., Saranadasa, M., Bosykh, D.A., Commane, M., Guryanova, O.A., Pal, S. *et al.* (2011) Curaxins: anticancer compounds that simultaneously suppress NF-kappaB and activate p53 by targeting FACT. *Sci Transl Med*, **3**, 95ra74.
18. Lu, K., Liu, C., Liu, Y., Luo, A., Chen, J., Lei, Z., Kong, J., Xiao, X., Zhang, S., Wang, Y.Z. *et al.* (2021) Curaxin-Induced DNA Topology Alterations Trigger the Distinct Binding Response of CTCF and FACT at the Single-Molecule Level. *Biochemistry*, **60**, 494-499.
19. Cremer, T. and Cremer, C. (2001) Chromosome territories, nuclear architecture and gene regulation in mammalian cells. *Nat Rev Genet*, **2**, 292-301.
20. Bolzer, A., Kreth, G., Solovei, I., Koehler, D., Saracoglu, K., Fauth, C., Müller, S., Eils, R., Cremer, C., Speicher, M.R. *et al.* (2005) Three-Dimensional Maps of All Chromosomes in Human Male Fibroblast Nuclei and Prometaphase Rosettes. *PLOS Biology*, **3**, e157.
21. van Steensel, B. and Belmont, A.S. (2017) Lamina-Associated Domains: Links with Chromosome Architecture, Heterochromatin, and Gene Repression. *Cell*, **169**, 780-791.
22. Meaburn, K.J. and Misteli, T. (2007) Cell biology: chromosome territories. *Nature*, **445**, 379-781.

23. Fortin, J.-P. and Hansen, K.D. (2015) Reconstructing A/B compartments as revealed by Hi-C using long-range correlations in epigenetic data. *Genome Biology*, **16**, 180.

24. Rowley, M.J. and Corces, V.G. (2018) Organizational principles of 3D genome architecture. *Nat Rev Genet*, **19**, 789-800.

25. Rowley, M.J. and Corces, V.G. (2018) Organizational principles of 3D genome architecture. *Nature Reviews Genetics*, **19**, 789-800.

26. Fudenberg, G., Imakaev, M., Lu, C., Goloborodko, A., Abdennur, N. and Mirny, Leonid A. (2016) Formation of Chromosomal Domains by Loop Extrusion. *Cell Reports*, **15**, 2038-2049.

27. Sanborn, A.L., Rao, S.S.P., Huang, S.-C., Durand, N.C., Huntley, M.H., Jewett, A.I., Bochkov, I.D., Chinnappan, D., Cutkosky, A., Li, J. *et al.* (2015) Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. *Proceedings of the National Academy of Sciences*, **112**, E6456-E6465.

28. Nora, E.P., Goloborodko, A., Valton, A.L., Gibcus, J.H., Uebersohn, A., Abdennur, N., Dekker, J., Mirny, L.A. and Bruneau, B.G. (2017) Targeted Degradation of CTCF Decouples Local Insulation of Chromosome Domains from Genomic Compartmentalization. *Cell*, **169**, 930-944 e922.

29. Rao, Suhas S.P., Huntley, Miriam H., Durand, Neva C., Stamenova, Elena K., Bochkov, Ivan D., Robinson, James T., Sanborn, Adrian L., Machol, I., Omer, Arina D., Lander, Eric S. *et al.* (2014) A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. *Cell*, **159**, 1665-1680.

30. Nagano, T., Lubling, Y., Stevens, T.J., Schoenfelder, S., Yaffe, E., Dean, W., Laue, E.D., Tanay, A. and Fraser, P. (2013) Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. *Nature*, **502**, 59-64.

31. Imakaev, M., Fudenberg, G., McCord, R.P., Naumova, N., Goloborodko, A., Lajoie, B.R., Dekker, J. and Mirny, L.A. (2012) Iterative correction of Hi-C data reveals hallmarks of chromosome organization. *Nature Methods*, **9**, 999-1003.

32. Dekker, J. and Mirny, L. (2016) The 3D Genome as Moderator of Chromosomal Communication. *Cell*, **164**, 1110-1121.
33. Quinodoz, S.A., Ollikainen, N., Tabak, B., Palla, A., Schmidt, J.M., Detmar, E., Lai, M.M., Shishkin, A.A., Bhat, P., Takei, Y. *et al.* (2018) Higher-Order Inter-chromosomal Hubs Shape 3D Genome Organization in the Nucleus. *Cell*, **174**, 744-757.e724.
34. Beagrie, R.A., Scialdone, A., Schueler, M., Kraemer, D.C., Chotalia, M., Xie, S.Q., Barbieri, M., de Santiago, I., Lavitas, L.M., Branco, M.R. *et al.* (2017) Complex multi-enhancer contacts captured by genome architecture mapping. *Nature*, **543**, 519-524.
35. Cardozo Gizzi, A.M., Cattoni, D.I., Fiche, J.B., Espinola, S.M., Gurgo, J., Messina, O., Houbron, C., Ogiyama, Y., Papadopoulos, G.L., Cavalli, G. *et al.* (2019) Microscopy-Based Chromosome Conformation Capture Enables Simultaneous Visualization of Genome Organization and Transcription in Intact Organisms. *Mol Cell*, **74**, 212-222 e215.
36. Bintu, B., Mateo, L.J., Su, J.H., Sinnott-Armstrong, N.A., Parker, M., Kinrot, S., Yamaya, K., Boettiger, A.N. and Zhuang, X. (2018) Super-resolution chromatin tracing reveals domains and cooperative interactions in single cells. *Science*, **362**.
37. Quinodoz, S.A., Ollikainen, N., Tabak, B., Palla, A., Schmidt, J.M., Detmar, E., Lai, M.M., Shishkin, A.A., Bhat, P., Takei, Y. *et al.* (2018) Higher-Order Inter-chromosomal Hubs Shape 3D Genome Organization in the Nucleus. *Cell*, **174**, 744-757 e724.
38. Delaneau, O., Zazhytska, M., Borel, C., Giannuzzi, G., Rey, G., Howald, C., Kumar, S., Ongen, H., Popadin, K., Marbach, D. *et al.* (2019) Chromatin three-dimensional interactions mediate genetic effects on gene expression. *Science*, **364**.
39. Clowney, E.J., LeGros, Mark A., Mosley, Colleen P., Clowney, Fiona G., Markenskoff-Papadimitriou, Eirene C., Myllys, M., Barnea, G., Larabell, Carolyn A. and Lomvardas, S. (2012) Nuclear Aggregation of Olfactory Receptor Genes Governs Their Monogenic Expression. *Cell*, **151**, 724-737.

40. Apostolou, E., Ferrari, F., Walsh, R.M., Bar-Nur, O., Stadtfeld, M., Cheloufi, S., Stuart, H.T., Polo, J.M., Ohsumi, T.K., Borowsky, M.L. *et al.* (2013) Genome-wide chromatin interactions of the Nanog locus in pluripotency, differentiation, and reprogramming. *Cell Stem Cell*, **12**, 699-712.

41. Sutherland, H. and Bickmore, W.A. (2009) Transcription factories: gene expression in unions? *Nat Rev Genet*, **10**, 457-466.

42. Di Stefano, M., Paulsen, J., Lien, T.G., Hovig, E. and Micheletti, C. (2016) Hi-C-constrained physical models of human chromosomes recover functionally-related properties of genome organization. *Sci Rep*, **6**, 35985.

43. Sarkar, S. and Dong, A. (2011) Community detection in graphs using singular value decomposition. *Phys Rev E Stat Nonlin Soft Matter Phys*, **83**, 046114.

44. Shokraneh, N., Arab, M. and Libbrecht, M. (2023) Integrative chromatin domain annotation through graph embedding of Hi-C data. *Bioinformatics*, **39**.

45. Bauer, C.R., Hartl, T.A. and Bosco, G. (2012) Condensin II promotes the formation of chromosome territories by inducing axial compaction of polyploid interphase chromosomes. *PLoS Genet*, **8**, e1002873.

46. Ferraz, R.A.C., Lopes, A.L.G., da Silva, J.A.F., Moreira, D.F.V., Ferreira, M.J.N. and de Almeida Coimbra, S.V. (2021) DNA-protein interaction studies: a historical and comparative analysis. *Plant Methods*, **17**, 82.

47. Zhu, L. and Huq, E. (2011) Mapping functional domains of transcription factors. *Methods Mol Biol*, **754**, 167-184.

48. Latchman, D.S. (1997) Transcription factors: an overview. *Int J Biochem Cell Biol*, **29**, 1305-1312.

49. Brodsky, S., Jana, T., Mittelman, K., Chapal, M., Kumar, D.K., Carmi, M. and Barkai, N. (2020) Intrinsically Disordered Regions Direct Transcription Factor In Vivo Binding Specificity. *Mol Cell*, **79**, 459-471 e454.

50. Wright, P.E. and Dyson, H.J. (2015) Intrinsically disordered proteins in cellular signalling and regulation. *Nature Reviews Molecular Cell Biology*, **16**, 18-29.

51. Malaney, P., Benitez, O., Zhang, X. and Post, S.M. (2022) Assessing the role of intrinsic disorder in RNA-binding protein function: hnRNP K as a case study. *Methods*, **208**, 59-65.
52. Dyson, H.J. and Wright, P.E. (2016) Role of Intrinsic Protein Disorder in the Function and Interactions of the Transcriptional Coactivators CREB-binding Protein (CBP) and p300. *J Biol Chem*, **291**, 6714-6722.
53. Weikum, E.R., Liu, X. and Ortlund, E.A. (2018) The nuclear receptor superfamily: A structural perspective. *Protein Sci*, **27**, 1876-1892.
54. Sar, P. (2023) Nuclear receptor: Structure and function. *Prog Mol Biol Transl Sci*, **196**, 209-227.
55. Sun, Z. and Xu, Y. (2020) Nuclear Receptor Coactivators (NCOAs) and Corepressors (NCORs) in the Brain. *Endocrinology*, **161**.
56. Li, X., Kimbrel, E.A., Kenan, D.J. and McDonnell, D.P. (2002) Direct interactions between corepressors and coactivators permit the integration of nuclear receptor-mediated repression and activation. *Mol Endocrinol*, **16**, 1482-1491.
57. Uversky, V.N. (2019) Intrinsically Disordered Proteins and Their "Mysterious" (Meta)Physics. *Frontiers in Physics*, **Volume 7 - 2019**.
58. Tompa, P. (2005) The interplay between structure and function in intrinsically unstructured proteins. *FEBS Lett*, **579**, 3346-3354.
59. Fuxreiter, M., Simon, I. and Bondos, S. (2011) Dynamic protein–DNA recognition: beyond what can be seen. *Trends in Biochemical Sciences*, **36**, 415-423.
60. Yang, J., Gao, M., Xiong, J., Su, Z. and Huang, Y. (2019) Features of molecular recognition of intrinsically disordered proteins via coupled folding and binding. *Protein Sci*, **28**, 1952-1965.
61. Miao, J. and Chong, S. (2025) Roles of intrinsically disordered protein regions in transcriptional regulation and genome organization. *Curr Opin Genet Dev*, **90**, 102285.
62. Brent, G.A. (2012) Mechanisms of thyroid hormone action. *J Clin Invest*, **122**, 3035-3043.
63. Nascimento, A.S., Dias, S.M., Nunes, F.M., Aparicio, R., Ambrosio, A.L., Bleicher, L., Figueira, A.C., Santos, M.A., de Oliveira Neto, M., Fischer, H. *et al.* (2006)

Structural rearrangements in the thyroid hormone receptor hinge domain and their putative role in the receptor function. *J Mol Biol*, **360**, 586-598.

64. Smyth, M.S. and Martin, J.H. (2000) x ray crystallography. *Mol Pathol*, **53**, 8-14.

65. Nogales, E. and Scheres, S.H. (2015) Cryo-EM: A Unique Tool for the Visualization of Macromolecular Complexity. *Mol Cell*, **58**, 677-689.

66. Dror, R.O., Dirks, R.M., Grossman, J.P., Xu, H. and Shaw, D.E. (2012) Biomolecular simulation: a computational microscope for molecular biology. *Annu Rev Biophys*, **41**, 429-452.

67. Vettoretti, G., Moroni, E., Sattin, S., Tao, J., Agard, D.A., Bernardi, A. and Colombo, G. (2016) Molecular Dynamics Simulations Reveal the Mechanisms of Allosteric Activation of Hsp90 by Designed Ligands. *Sci Rep*, **6**, 23830.

68. Martínez, L., Sonoda, M.T., Webb, P., Baxter, J.D., Skaf, M.S. and Polikarpov, I. (2005) Molecular Dynamics Simulations Reveal Multiple Pathways of Ligand Dissociation from Thyroid Hormone Receptors. *Biophysical journal*, **89**, 2011-2023.

69. Harnish, M.T., Pham, B., Arons, A.B., Xu, Y., Fernandez, E.J. and Shen, T. (2026) Conformational transition of a polycationic hinge domain contributes to DNA binding. *Biophys Chem*, **328**, 107531.

70. Wayment-Steele, H.K. and Pande, V.S. (2018) Note: Variational encoding of protein dynamics benefits from maximizing latent autocorrelation. *J Chem Phys*, **149**, 216101.

71. Matsumoto, S., Ishida, S., Terayama, K. and Okuno, Y. (2023) Quantitative analysis of protein dynamics using a deep learning technique combined with experimental cryo-EM density data and MD simulations. *Biophys Physicobiol*, **20**, e200022.

72. Durairaj, J., de Ridder, D. and van Dijk, A.D.J. (2023) Beyond sequence: Structure-based machine learning. *Comput Struct Biotechnol J*, **21**, 630-643.

73. Widom, J. (2001) Role of DNA sequence in nucleosome stability and dynamics. *Q Rev Biophys*, **34**, 269-324.

74. Liu, L.F. and Wang, J.C. (1987) Supercoiling of the DNA template during transcription. *Proc Natl Acad Sci U S A*, **84**, 7024-7027.

75. Hurley, L.H. (2002) DNA and its associated processes as targets for cancer therapy. *Nature Reviews Cancer*, **2**, 188-200.
76. Deans, A.J. and West, S.C. (2011) DNA interstrand crosslink repair and cancer. *Nat Rev Cancer*, **11**, 467-480.
77. Dasari, S. and Tchounwou, P.B. (2014) Cisplatin in cancer therapy: molecular mechanisms of action. *Eur J Pharmacol*, **740**, 364-378.
78. de Sousa Cavalcante, L. and Monteiro, G. (2014) Gemcitabine: metabolism and molecular mechanisms of action, sensitivity and chemoresistance in pancreatic cancer. *Eur J Pharmacol*, **741**, 8-16.
79. Rohs, R., West, S.M., Sosinsky, A., Liu, P., Mann, R.S. and Honig, B. (2009) The role of DNA shape in protein-DNA recognition. *Nature*, **461**, 1248-1253.
80. Kaplan, N., Moore, I.K., Fondufe-Mittendorf, Y., Gossett, A.J., Tillo, D., Field, Y., LeProust, E.M., Hughes, T.R., Lieb, J.D., Widom, J. *et al.* (2009) The DNA-encoded nucleosome organization of a eukaryotic genome. *Nature*, **458**, 362-366.
81. Lerman, L.S. (1961) Structural considerations in the interaction of DNA and acridines. *J Mol Biol*, **3**, 18-30.
82. Neidle, S. (2001) DNA minor-groove recognition by small molecules. *Nat Prod Rep*, **18**, 291-309.
83. Tewey, K.M., Rowe, T.C., Yang, L., Halligan, B.D. and Liu, L.F. (1984) Adriamycin-induced DNA damage mediated by mammalian DNA topoisomerase II. *Science*, **226**, 466-468.
84. Volokh, O.I., Sivkina, A.L., Moiseenko, A.V., Popinako, A.V., Karlova, M.G., Valieva, M.E., Kotova, E.Y., Kirpichnikov, M.P., Formosa, T., Studitsky, V.M. *et al.* (2022) Mechanism of curaxin-dependent nucleosome unfolding by FACT. *Front Mol Biosci*, **9**, 1048117.
85. Kantidze, O.L., Luzhin, A.V., Nizovtseva, E.V., Safina, A., Valieva, M.E., Golov, A.K., Velichko, A.K., Lyubitelev, A.V., Feofanov, A.V., Gurova, K.V. *et al.* (2019) The anti-cancer drugs curaxins target spatial genome organization. *Nat Commun*, **10**, 1441.
86. Nesher, E., Safina, A., Aljahdali, I., Portwood, S., Wang, E.S., Koman, I., Wang, J. and Gurova, K.V. (2018) Role of Chromatin Damage and Chromatin Trapping of

FACT in Mediating the Anticancer Cytotoxicity of DNA-Binding Small-Molecule Drugs. *Cancer Res*, **78**, 1431-1443.

87. Reinberg, D. and Sims, R.J., 3rd. (2006) de FACTo nucleosome dynamics. *J Biol Chem*, **281**, 23297-23301.

88. Safina, A., Cheney, P., Pal, M., Brodsky, L., Ivanov, A., Kirsanov, K., Lesovaya, E., Naberezhnov, D., Nesher, E., Koman, I. *et al.* (2017) FACT is a sensor of DNA torsional stress in eukaryotic cells. *Nucleic Acids Res*, **45**, 1925-1945.

89. Ong, C.T. and Corces, V.G. (2014) CTCF: an architectural protein bridging genome topology and function. *Nat Rev Genet*, **15**, 234-246.

90. Lu, K., Liu, C., Liu, Y., Luo, A., Chen, J., Lei, Z., Kong, J., Xiao, X., Zhang, S., Wang, Y.-Z. *et al.* (2021) Curaxin-Induced DNA Topology Alterations Trigger the Distinct Binding Response of CTCF and FACT at the Single-Molecule Level. *Biochemistry*, **60**, 494-499.

91. Struhl, K. and Segal, E. (2013) Determinants of nucleosome positioning. *Nat Struct Mol Biol*, **20**, 267-273.

92. Flavahan, W.A., Drier, Y., Liau, B.B., Gillespie, S.M., Venteicher, A.S., Stemmer-Rachamimov, A.O., Suva, M.L. and Bernstein, B.E. (2016) Insulator dysfunction and oncogene activation in IDH mutant gliomas. *Nature*, **529**, 110-114.

93. Corces, M.R., Granja, J.M., Shams, S., Louie, B.H., Seoane, J.A., Zhou, W., Silva, T.C., Groeneveld, C., Wong, C.K., Cho, S.W. *et al.* (2018) The chromatin accessibility landscape of primary human cancers. *Science*, **362**.

94. Bradner, J.E., Hnisz, D. and Young, R.A. (2017) Transcriptional Addiction in Cancer. *Cell*, **168**, 629-643.

95. Naughton, C., Avlonitis, N., Corless, S., Prendergast, J.G., Mati, I.K., Eijk, P.P., Cockroft, S.L., Bradley, M., Ylstra, B. and Gilbert, N. (2013) Transcription forms and remodels supercoiling domains unfolding large-scale chromatin structures. *Nat Struct Mol Biol*, **20**, 387-395.

96. Robson, M.I., Ringel, A.R. and Mundlos, S. (2019) Regulatory Landscaping: How Enhancer-Promoter Communication Is Sculpted in 3D. *Molecular Cell*, **74**, 1110-1122.

97. Liu, H., Tsai, H., Yang, M., Li, G., Bian, Q., Ding, G., Wu, D. and Dai, J. (2023) Three-dimensional genome structure and function. *MedComm (2020)*, **4**, e326.
98. da Costa-Nunes, J.A. and Noordermeer, D. (2023) TADs: Dynamic structures to create stable regulatory functions. *Curr Opin Struct Biol*, **81**, 102622.
99. Bruckner, D.B., Chen, H., Barinov, L., Zoller, B. and Gregor, T. (2023) Stochastic motion and transcriptional dynamics of pairs of distal DNA loci on a compacted chromosome. *Science*, **380**, 1357-1362.
100. McCord, R.P., Kaplan, N. and Giorgetti, L. (2020) Chromosome Conformation Capture and Beyond: Toward an Integrative View of Chromosome Structure and Function. *Mol Cell*, **77**, 688-708.
101. Corsi, F., Rusch, E. and Goloborodko, A. (2023) Loop extrusion rules: the next generation. *Curr Opin Genet Dev*, **81**, 102061.
102. 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. *et al.* (2009) Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. *Science*, **326**, 289-293.
103. Zhang, B. and Wolynes, P.G. (2015) Topology, structures, and energy landscapes of human chromosomes. *Proceedings of the National Academy of Sciences*, **112**, 6062-6067.
104. Crane, E., Bian, Q., McCord, R.P., Lajoie, B.R., Wheeler, B.S., Ralston, E.J., Uzawa, S., Dekker, J. and Meyer, B.J. (2015) Condensin-driven remodelling of X chromosome topology during dosage compensation. *Nature*, **523**, 240-244.
105. Zhang, C., Xu, Z., Yang, S., Sun, G., Jia, L., Zheng, Z., Gu, Q., Tao, W., Cheng, T., Li, C. *et al.* (2020) tagHi-C Reveals 3D Chromatin Architecture Dynamics during Mouse Hematopoiesis. *Cell Rep*, **32**, 108206.
106. Heinz, S., Texari, L., Hayes, M.G.B., Urbanowski, M., Chang, M.W., Givarkes, N., Rialdi, A., White, K.M., Albrecht, R.A., Pache, L. *et al.* (2018) Transcription Elongation Can Affect Genome 3D Structure. *Cell*, **174**, 1522-+.

107. Belaghzal, H., Borrman, T., Stephens, A.D., Lafontaine, D.L., Venev, S.V., Weng, Z., Marko, J.F. and Dekker, J. (2021) Liquid chromatin Hi-C characterizes compartment-dependent chromatin interaction dynamics. *Nat Genet*, **53**, 367-378.
108. García-Domenech, R., Gálvez, J., de Julián-Ortiz, J.V. and Pogliani, L. (2008) Some New Trends in Chemical Graph Theory. *Chemical Reviews*, **108**, 1127-1169.
109. Atilgan, C., Okan, O.B. and Atilgan, A.R. (2012) In Rees, D. C. (ed.), *Annual Review of Biophysics, Vol 41*, Vol. 41, pp. 205-225.
110. Amitai, G., Shemesh, A., Sitbon, E., Shklar, M., Netanely, D., Venger, I. and Pietrokovski, S. (2004) Network analysis of protein structures identifies functional residues. *Journal of molecular biology*, **344**, 1135-1146.
111. Di Paola, L. and Giuliani, A. (2015) Protein contact network topology: a natural language for allostery. *Current Opinion in Structural Biology*, **31**, 43-48.
112. Doshi, U., Holliday, M.J., Eisenmesser, E.Z. and Hamelberg, D. (2016) Dynamical network of residue–residue contacts reveals coupled allosteric effects in recognition, catalysis, and mutation. *Proceedings of the National Academy of Sciences*, **113**, 4735-4740.
113. Daily, M.D., Upadhyaya, T.J. and Gray, J.J. (2008) Contact rearrangements form coupled networks from local motions in allosteric proteins. *Proteins-Structure Function and Bioinformatics*, **71**, 455-466.
114. Albert, R. and Barabási, A.-L. (2002) Statistical mechanics of complex networks. *Reviews of Modern Physics*, **74**, 47-97.
115. Yao, X.Q., Momin, M. and Hamelberg, D. (2018) Elucidating Allosteric Communications in Proteins with Difference Contact Network Analysis. *J Chem Inf Model*, **58**, 1325-1330.
116. Sethi, A., Tian, J., Vu, Dung M. and Gnanakaran, S. (2012) Identification of Minimally Interacting Modules in an Intrinsically Disordered Protein. *Biophysical Journal*, **103**, 748-757.
117. Kryven, I., Duivenvoorden, J., Hermans, J. and Iedema, P.D. (2016) Random Graph Approach to Multifunctional Molecular Networks. *Macromolecular Theory and Simulations*, **25**, 449-465.

118. Stauffer, D. (1985) *Introduction to Percolation Theory*. Taylor & Francis.
119. Zhang, B. and Horvath, S. (2005) A general framework for weighted gene co-expression network analysis. *Stat Appl Genet Mol Biol*, **4**, Article17.
120. Hao, D., Ren, C. and Li, C. (2012) Revisiting the variation of clustering coefficient of biological networks suggests new modular structure. *BMC Syst Biol*, **6**, 34.
121. Panditrao, G., Bhowmick, R., Meena, C. and Sarkar, R.R. (2022) Emerging landscape of molecular interaction networks: Opportunities, challenges and prospects. *Journal of Biosciences*, **47**, 24.
122. Pancaldi, V. (2023) Network models of chromatin structure. *Current Opinion in Genetics & Development*, **80**, 102051.
123. Botta, M., Haider, S., Leung, I.X.Y., Lio, P. and Mozziconacci, J. (2010) Intra- and inter-chromosomal interactions correlate with CTCF binding genome wide. *Molecular Systems Biology*, **6**, 426.
124. Sandhu, K.S., Li, G., Poh, H.M., Quek, Y.L., Sia, Y.Y., Peh, S.Q., Mulawadi, F.H., Lim, J., Sikic, M., Menghi, F. *et al.* (2012) Large-scale functional organization of long-range chromatin interaction networks. *Cell Rep*, **2**, 1207-1219.
125. Pancaldi, V., Carrillo-de-Santa-Pau, E., Javierre, B.M., Juan, D., Fraser, P., Spivakov, M., Valencia, A. and Rico, D. (2016) Integrating epigenomic data and 3D genomic structure with a new measure of chromatin assortativity. *Genome Biology*, **17**, 152.
126. Newman, M. (2018) *Networks*. OUP Oxford.
127. Solovei, I., Kreysing, M., Lanctot, C., Kosem, S., Peichl, L., Cremer, T., Guck, J. and Joffe, B. (2009) Nuclear architecture of rod photoreceptor cells adapts to vision in mammalian evolution. *Cell*, **137**, 356-368.
128. Falk, M., Feodorova, Y., Naumova, N., Imakaev, M., Lajoie, B.R., Leonhardt, H., Joffe, B., Dekker, J., Fudenberg, G., Solovei, I. *et al.* (2019) Heterochromatin drives compartmentalization of inverted and conventional nuclei. *Nature*, **570**, 395-399.
129. Javierre, B.M., Burren, O.S., Wilder, S.P., Kreuzhuber, R., Hill, S.M., Sewitz, S., Cairns, J., Wingett, S.W., Varnai, C., Thiecke, M.J. *et al.* (2016) Lineage-Specific

Genome Architecture Links Enhancers and Non-coding Disease Variants to Target Gene Promoters. *Cell*, **167**, 1369-1384 e1319.

130. Hoencamp, C., Dudchenko, O., Elbatsh, A.M.O., Brahmachari, S., Raaijmakers, J.A., van Schaik, T., Sedeño Cacciatore, Á., Contessoto, V.G., van Heesbeen, R., van den Broek, B. *et al.* (2021) 3D genomics across the tree of life reveals condensin II as a determinant of architecture type. *Science*, **372**, 984-989.

131. Gibcus, J.H., Samejima, K., Goloborodko, A., Samejima, I., Naumova, N., Nuebler, J., Kanemaki, M.T., Xie, L., Paulson, J.R., Earnshaw, W.C. *et al.* (2018) A pathway for mitotic chromosome formation. *Science*, **359**, eaao6135.

132. Kaul, A., Bhattacharyya, S. and Ay, F. (2020) Identifying statistically significant chromatin contacts from Hi-C data with FitHiC2. *Nat Protoc*, **15**, 991-1012.

133. Ay, F., Bailey, T.L. and Noble, W.S. (2014) Statistical confidence estimation for Hi-C data reveals regulatory chromatin contacts. *Genome Res*, **24**, 999-1011.

134. Imakaev, M., Fudenberg, G., McCord, R.P., Naumova, N., Goloborodko, A., Lajoie, B.R., Dekker, J. and Mirny, L.A. (2012) Iterative correction of Hi-C data reveals hallmarks of chromosome organization. *Nat Methods*, **9**, 999-1003.

135. Rao, S.S., Huntley, M.H., Durand, N.C., Stamenova, E.K., Bochkov, I.D., Robinson, J.T., Sanborn, A.L., Machol, I., Omer, A.D., Lander, E.S. *et al.* (2014) A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*, **159**, 1665-1680.

136. Caldarelli, G., Pastor-Satorras, R. and Vespignani, A. (2004) Structure of cycles and local ordering in complex networks. *The European Physical Journal B*, **38**, 183-186.

137. Fronczak, A., Hołyst, J.A., Jedynak, M. and Sienkiewicz, J. (2002) Higher order clustering coefficients in Barabási–Albert networks. *Physica A: Statistical Mechanics and its Applications*, **316**, 688-694.

138. Lind, P.G., González, M.C. and Herrmann, H.J. (2005) Cycles and clustering in bipartite networks. *Physical Review E*, **72**, 056127.

139. Yin, H., Benson, A.R. and Leskovec, J. (2018) Higher-order clustering in networks. *Physical Review E*, **97**, 052306.

140. Das, P., Golloshi, R., McCord, R.P. and Shen, T. (2020) Using contact statistics to characterize structure transformation of biopolymer ensembles. *Physical Review E*, **101**, 012419.
141. Haarhuis, J.H.I., van der Weide, R.H., Blomen, V.A., Yanez-Cuna, J.O., Amendola, M., van Ruiten, M.S., Krijger, P.H.L., Teunissen, H., Medema, R.H., van Steensel, B. *et al.* (2017) The Cohesin Release Factor WAPL Restricts Chromatin Loop Extension. *Cell*, **169**, 693-707 e614.
142. Kind, J., Pagie, L., de Vries, S.S., Nahidiazar, L., Dey, S.S., Bienko, M., Zhan, Y., Lajoie, B., de Graaf, C.A., Amendola, M. *et al.* (2015) Genome-wide Maps of Nuclear Lamina Interactions in Single Human Cells. *Cell*, **163**, 134-147.
143. Das, P., San Martin, R. and McCord, R.P. (2023) Differential contributions of nuclear lamina association and genome compartmentalization to gene regulation. *Nucleus*, **14**, 2197693.
144. Ernst, J., Kheradpour, P., Mikkelsen, T.S., Shoresh, N., Ward, L.D., Epstein, C.B., Zhang, X., Wang, L., Issner, R., Coyne, M. *et al.* (2011) Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature*, **473**, 43-49.
145. Xiong, K. and Ma, J. (2019) Revealing Hi-C subcompartments by imputing inter-chromosomal chromatin interactions. *Nat Commun*, **10**, 5069.
146. Wen, Z., Zhang, W., Zhong, Q., Xu, J., Hou, C., Qin, Z.S. and Li, L. (2022) Extensive Chromatin Structure-Function Associations Revealed by Accurate 3D Compartmentalization Characterization. *Front Cell Dev Biol*, **10**, 845118.
147. Nichols, M.H. and Corces, V.G. (2021) Principles of 3D compartmentalization of the human genome. *Cell Rep*, **35**, 109330.
148. Yaffe, E. and Tanay, A. (2011) Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture. *Nat Genet*, **43**, 1059-1065.
149. Kariti, H., Feld, T. and Kaplan, N. (2023) Hypothesis-driven probabilistic modelling enables a principled perspective of genomic compartments. *Nucleic Acids Res*, **51**, 1103-1119.

150. Dodero-Rojas, E., Mello, M.F., Brahmachari, S., Oliveira Junior, A.B., Contessoto, V.G. and Onuchic, J.N. (2023) PyMEGABASE: Predicting Cell-Type-Specific Structural Annotations of Chromosomes Using the Epigenome. *J Mol Biol*, **435**, 168180.

151. Fatima, U., Hina, S. and Wasif, M. (2023) A novel global clustering coefficient-dependent degree centrality (GCCDC) metric for large network analysis using real-world datasets. *Journal of Computational Science*, **70**, 102008.

152. Malod-Dognin, N., Pancaldi, V., Valencia, A. and Pržulj, N. (2020) Chromatin network markers of leukemia. *Bioinformatics*, **36**, i455-i463.

153. Su, J.H., Zheng, P., Kinrot, S.S., Bintu, B. and Zhuang, X. (2020) Genome-Scale Imaging of the 3D Organization and Transcriptional Activity of Chromatin. *Cell*, **182**, 1641-1659.e1626.

154. Thibodeau, A., Márquez, E.J., Shin, D.-G., Vera-Licona, P. and Ucar, D. (2017) Chromatin interaction networks revealed unique connectivity patterns of broad H3K4me3 domains and super enhancers in 3D chromatin. *Scientific Reports*, **7**, 14466.

155. Ernst, J. and Kellis, M. (2017) Chromatin-state discovery and genome annotation with ChromHMM. *Nat Protoc*, **12**, 2478-2492.

156. Abdulla, A.Z., Salari, H., Tortora, M.M.C., Vaillant, C. and Jost, D. (2023) 4D epigenomics: deciphering the coupling between genome folding and epigenomic regulation with biophysical modeling. *Current Opinion in Genetics & Development*, **79**, 102033.

157. Finn, E.H. and Misteli, T. (2019) Molecular basis and biological function of variability in spatial genome organization. *Science*, **365**.

158. Fraser, P. and Bickmore, W. (2007) Nuclear organization of the genome and the potential for gene regulation. *Nature*, **447**, 413-417.

159. Stadhouders, R., Filion, G.J. and Graf, T. (2019) Transcription factors and 3D genome conformation in cell-fate decisions. *Nature*, **569**, 345-354.

160. Tomikawa, J. (2024) Potential roles of inter-chromosomal interactions in cell fate determination. *Front Cell Dev Biol*, **12**, 1397807.

161. Zhao, Z., Tavoosidana, G., Sjolinder, M., Gondor, A., Mariano, P., Wang, S., Kanduri, C., Lezcano, M., Sandhu, K.S., Singh, U. *et al.* (2006) Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions. *Nat Genet*, **38**, 1341-1347.
162. Dekker, J., Marti-Renom, M.A. and Mirny, L.A. (2013) Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. *Nat Rev Genet*, **14**, 390-403.
163. Bashkirova, E. and Lomvardas, S. (2019) Olfactory receptor genes make the case for inter-chromosomal interactions. *Curr Opin Genet Dev*, **55**, 106-113.
164. Asamitsu, S. and Iwasaki, Y.W. (2025) Biomolecular liquid–liquid phase separation associated with repetitive genomic elements. *Polymer Journal*, **57**, 785-797.
165. Mehta, S. and Zhang, J. (2022) Liquid–liquid phase separation drives cellular function and dysfunction in cancer. *Nature Reviews Cancer*, **22**, 239-252.
166. Tong, X., Tang, R., Xu, J., Wang, W., Zhao, Y., Yu, X. and Shi, S. (2022) Liquid–liquid phase separation in tumor biology. *Signal Transduction and Targeted Therapy*, **7**, 221.
167. Maass, P.G., Barutcu, A.R. and Rinn, J.L. (2019) Interchromosomal interactions: A genomic love story of kissing chromosomes. *J Cell Biol*, **218**, 27-38.
168. Szalay, M.F., Majchrzycka, B., Jerković, I., Cavalli, G. and Ibrahim, D.M. (2024) Evolution and function of chromatin domains across the tree of life. *Nat Struct Mol Biol*, **31**, 1824-1837.
169. Das, P., Shen, T. and McCord, R.P. (2020) Inferring chromosome radial organization from Hi-C data. *BMC Bioinformatics*, **21**, 511.
170. Bickmore, W.A. and van Steensel, B. (2013) Genome architecture: domain organization of interphase chromosomes. *Cell*, **152**, 1270-1284.
171. Cremer, T., Cremer, M., Dietzel, S., Muller, S., Solovei, I. and Fakan, S. (2006) Chromosome territories--a functional nuclear landscape. *Curr Opin Cell Biol*, **18**, 307-316.
172. Misteli, T. (2007) Beyond the sequence: cellular organization of genome function. *Cell*, **128**, 787-800.

173. Belton, J.M., McCord, R.P., Gibcus, J.H., Naumova, N., Zhan, Y. and Dekker, J. (2012) Hi-C: a comprehensive technique to capture the conformation of genomes. *Methods*, **58**, 268-276.

174. Wolff, J., Backofen, R. and Gruning, B. (2022) Loop detection using Hi-C data with HiCExplorer. *Gigascience*, **11**.

175. Cameron, C.J., Dostie, J. and Blanchette, M. (2020) HIFI: estimating DNA-DNA interaction frequency from Hi-C data at restriction-fragment resolution. *Genome Biol*, **21**, 11.

176. Highsmith, M. and Cheng, J. (2021) VEHiCLE: a Variationally Encoded Hi-C Loss Enhancement algorithm for improving and generating Hi-C data. *Sci Rep*, **11**, 8880.

177. Lajoie, B.R., Dekker, J. and Kaplan, N. (2015) The Hitchhiker's guide to Hi-C analysis: practical guidelines. *Methods*, **72**, 65-75.

178. Lyu, H., Liu, E. and Wu, Z. (2020) Comparison of normalization methods for Hi-C data. *Biotechniques*, **68**, 56-64.

179. Servant, N., Varoquaux, N., Lajoie, B.R., Viara, E., Chen, C.J., Vert, J.P., Heard, E., Dekker, J. and Barillot, E. (2015) HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. *Genome Biol*, **16**, 259.

180. Carty, M., Zamparo, L., Sahin, M., Gonzalez, A., Pelossof, R., Elemento, O. and Leslie, C.S. (2017) An integrated model for detecting significant chromatin interactions from high-resolution Hi-C data. *Nat Commun*, **8**, 15454.

181. Schmitt, A.D., Hu, M. and Ren, B. (2016) Genome-wide mapping and analysis of chromosome architecture. *Nat Rev Mol Cell Biol*, **17**, 743-755.

182. Kamada, T. and Kawai, S. (1989) An algorithm for drawing general undirected graphs. *Information Processing Letters*, **31**, 7-15.

183. Fruchterman, T.M.J. and Reingold, E.M. (1991) Graph drawing by force-directed placement. *Software: Practice and Experience*, **21**, 1129-1164.

184. Martin, S., Brown, W.M., Klavans, R. and Boyack, K. (2011) *OpenOrd: an open-source toolbox for large graph layout*. SPIE.

185. Davidson, R. and Harel, D. (1996) Drawing graphs nicely using simulated annealing. *ACM Trans. Graph.*, **15**, 301–331.

186. Zhang, P., Wang, J., Li, X., Li, M., Di, Z. and Fan, Y. (2008) Clustering coefficient and community structure of bipartite networks. *Physica A: Statistical Mechanics and its Applications*, **387**, 6869-6875.
187. Kingan, S. (2022). First edition. ed. Wiley.
188. Watts, D.J. and Strogatz, S.H. (1998) Collective dynamics of 'small-world' networks. *Nature*, **393**, 440-442.
189. Makai, D., Cseh, A., Sepsi, A. and Makai, S. (2022) A Multigraph-Based Representation of Hi-C Data. *Genes (Basel)*, **13**.
190. Bonato, A., Chiang, M., Corbett, D., Kitaev, S., Marenduzzo, D., Morozov, A. and Orlandini, E. (2024) Topological Spectra and Entropy of Chromatin Loop Networks. *Phys Rev Lett*, **132**, 248403.
191. Xu, Y., Das, P., McCord, R.P. and Shen, T. (2024) Node features of chromosome structure networks and their connections to genome annotation. *Comput Struct Biotechnol J*, **23**, 2240-2250.
192. Xu, Y., Das, P., McCord, R.P. and Shen, T. (2024) Node features of chromosome structure networks and their connections to genome annotation. *Computational and Structural Biotechnology Journal*, **23**, 2240-2250.
193. Briand, N. and Collas, P. (2020) Lamina-associated domains: peripheral matters and internal affairs. *Genome Biol*, **21**, 85.
194. Foutch, D., Pham, B. and Shen, T. (2021) Protein conformational switch discerned via network centrality properties. *Computational and Structural Biotechnology Journal*, **19**, 3599-3608.
195. Lindsay, R.J., Pham, B., Shen, T. and McCord, R.P. (2018) Characterizing the 3D structure and dynamics of chromosomes and proteins in a common contact matrix framework. *Nucleic Acids Res*, **46**, 8143-8152.
196. Maisuradze, L., King, M.C., Surovtsev, I.V., Mochrie, S.G.J., Shattuck, M.D. and O'Hern, C.S. (2024) Identifying topologically associating domains using differential kernels. *PLoS Comput Biol*, **20**, e1012221.
197. Fudenberg, G. and Imakaev, M. (2017) FISH-ing for captured contacts: towards reconciling FISH and 3C. *Nat Methods*, **14**, 673-678.

198. Guelen, L., Pagie, L., Brasset, E., Meuleman, W., Faza, M.B., Talhout, W., Eussen, B.H., de Klein, A., Wessels, L., de Laat, W. *et al.* (2008) Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. *Nature*, **453**, 948-951.
199. Ernst, J. and Kellis, M. (2010) Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nat Biotechnol*, **28**, 817-825.
200. Kind, J., Pagie, L., de Vries, S.S., Nahidiazar, L., Dey, S.S., Bienko, M., Zhan, Y., Lajoie, B., de Graaf, C.A., Amendola, M. *et al.* (2015) Genome-wide maps of nuclear lamina interactions in single human cells. *Cell*, **163**, 134-147.
201. Hirano, T. (2012) Condensins: universal organizers of chromosomes with diverse functions. *Genes Dev*, **26**, 1659-1678.
202. Green, L.C., Kalitsis, P., Chang, T.M., Cipetic, M., Kim, J.H., Marshall, O., Turnbull, L., Whitchurch, C.B., Vagnarelli, P., Samejima, K. *et al.* (2012) Contrasting roles of condensin I and condensin II in mitotic chromosome formation. *J Cell Sci*, **125**, 1591-1604.
203. Hoencamp, C., Dudchenko, O., Elbatsh, A.M.O., Brahmachari, S., Raaijmakers, J.A., van Schaik, T., Sedeno Cacciatore, A., Contessoto, V.G., van Heesbeen, R., van den Broek, B. *et al.* (2021) 3D genomics across the tree of life reveals condensin II as a determinant of architecture type. *Science*, **372**, 984-989.
204. Ono, T., Fang, Y., Spector, D.L. and Hirano, T. (2004) Spatial and temporal regulation of Condensins I and II in mitotic chromosome assembly in human cells. *Mol Biol Cell*, **15**, 3296-3308.
205. Samejima, K., Booth, D.G., Ogawa, H., Paulson, J.R., Xie, L., Watson, C.A., Platani, M., Kanemaki, M.T. and Earnshaw, W.C. (2018) Functional analysis after rapid degradation of condensins and 3D-EM reveals chromatin volume is uncoupled from chromosome architecture in mitosis. *J Cell Sci*, **131**.
206. Dey, A., Shi, G., Takaki, R. and Thirumalai, D. (2023) Structural changes in chromosomes driven by multiple condensin motors during mitosis. *Cell Rep*, **42**, 112348.

207. Sakamoto, T., Sakamoto, Y., Grob, S., Slane, D., Yamashita, T., Ito, N., Oko, Y., Sugiyama, T., Higaki, T., Hasezawa, S. *et al.* (2022) Two-step regulation of centromere distribution by condensin II and the nuclear envelope proteins. *Nat Plants*, **8**, 940-953.
208. Hirano, T. (2016) Condensin-Based Chromosome Organization from Bacteria to Vertebrates. *Cell*, **164**, 847-857.
209. Mangelsdorf, D.J., Thummel, C., Beato, M., Herrlich, P., Schütz, G., Umesono, K., Blumberg, B., Kastner, P., Mark, M., Chambon, P. *et al.* (1995) The Nuclear Receptor Superfamily: The Second Decade. *Cell*, **83**, 835-839.
210. Zilliacus, J., Carlstedt-Duke, J., Gustafsson, J.A. and Wright, A.P. (1994) Evolution of distinct DNA-binding specificities within the nuclear receptor family of transcription factors. *Proceedings of the National Academy of Sciences*, **91**, 4175-4179.
211. Sierk, M.L., Zhao, Q. and Rastinejad, F. (2001) DNA Deformability as a Recognition Feature in the RevErb Response Element. *Biochemistry*, **40**, 12833-12843.
212. Nelson, C.C., Hendy, S.C., Shukin, R.J., Cheng, H., Bruchovsky, N., Koop, B.F. and Rennie, P.S. (1999) Determinants of DNA Sequence Specificity of the Androgen, Progesterone, and Glucocorticoid Receptors: Evidence for Differential Steroid Receptor Response Elements. *Molecular Endocrinology*, **13**, 2090-2107.
213. Lohry, D.P., Stevens, T.A., Shen, T. and Fernandez, E.J. (2024) Hormone response elements for the thyroid receptor-α include specific distal 5′-flanking DNA. *Science Advances*, **10**, eadr1033.
214. Chen, X., Tsai, M.-Y. and Wolynes, P.G. (2022) The Role of Charge Density Coupled DNA Bending in Transcription Factor Sequence Binding Specificity: A Generic Mechanism for Indirect Readout. *Journal of the American Chemical Society*, **144**, 1835-1845.
215. Dickerson, R.E. and Chiu, T.K. (1997) Helix bending as a factor in protein/DNA recognition. *Biopolymers*, **44**, 361-403.
216. Rohs, R., Jin, X., West, S.M., Joshi, R., Honig, B. and Mann, R.S. (2010) Origins of Specificity in Protein-DNA Recognition. *Annual Review of Biochemistry*, **79**, 233-269.
217. Fernandez, E.J. (2018) Allosteric pathways in nuclear receptors - Potential targets for drug design. *Pharmacol Ther*, **183**, 152-159.

218. Baughman, H.E.R., Narang, D., Chen, W., Villagrán Suárez, A.C., Lee, J., Bachochin, M.J., Gunther, T.R., Wolynes, P.G. and Komives, E.A. (2022) An intrinsically disordered transcription activation domain increases the DNA binding affinity and reduces the specificity of NFκB p50/RelA. *J Biol Chem*, **298**, 102349.
219. Alberti, S. and Dormann, D. (2019) Liquid-Liquid Phase Separation in Disease. *Annu Rev Genet*, **53**, 171-194.
220. Fribourgh, J.L. and Partch, C.L. (2017) Assembly and function of bHLH–PAS complexes. *Proceedings of the National Academy of Sciences*, **114**, 5330-5332.
221. Oldfield, C.J. and Dunker, A.K. (2014) Intrinsically disordered proteins and intrinsically disordered protein regions. *Annu Rev Biochem*, **83**, 553-584.
222. Uversky, V.N., Oldfield, C.J. and Dunker, A.K. (2008) Intrinsically Disordered Proteins in Human Diseases: Introducing the D2 Concept. *Annual Review of Biophysics*, **37**, 215-246.
223. Basile, W., Salvatore, M., Bassot, C. and Elofsson, A. (2019) Why do eukaryotic proteins contain more intrinsically disordered regions? *PLoS Comput Biol*, **15**, e1007186-e1007186.
224. Geserick, C., Meyer, H.-A. and Haendler, B. (2005) The role of DNA response elements as allosteric modulators of steroid receptor function. *Molecular and Cellular Endocrinology*, **236**, 1-7.
225. Eldred, K.C., Hadyniak, S.E., Hussey, K.A., Brenerman, B., Zhang, P.-W., Chamling, X., Sluch, V.M., Welsbie, D.S., Hattar, S., Taylor, J. *et al.* (2018) Thyroid hormone signaling specifies cone subtypes in human retinal organoids. *Science*, **362**, eaau6348.
226. Papageorgiou, L., Shalzi, L., Efthimiadou, A., Bacopoulou, F., Chrousos, G.P., Eliopoulos, E. and Vlachakis, D. (2021) Conserved functional motifs of the nuclear receptor superfamily as potential pharmacological targets. *Int J Epigen*, **1**, 3.
227. Kumar, R. and Thompson, E.B. (2003) Transactivation functions of the N-terminal domains of nuclear hormone receptors: protein folding and coactivator interactions. *Mol Endocrinol*, **17**, 1-10.

228. Langlois, M.-F., Zanger, K., Monden, T., Safer, J.D., Hollenberg, A.N. and Wondisford, F.E. (1997) A Unique Role of the Beta-2 Thyroid Hormone Receptor Isoform in Negative Regulation by Thyroid Hormone. *Journal of Biological Chemistry*, **272**, 24927-24933.

229. Koenig, R.J. (1998) Thyroid hormone receptor coactivators and corepressors. *Thyroid*, **8**, 703-713.

230. Cheskis, B.J., Karathanasis, S. and Lyttle, C.R. (1997) Estrogen Receptor Ligands Modulate Its Interaction with DNA. *Journal of Biological Chemistry*, **272**, 11384-11391.

231. Putcha, B.-D.K. and Fernandez, E.J. (2009) Direct Interdomain Interactions Can Mediate Allosterism in the Thyroid Receptor. *Journal of Biological Chemistry*, **284**, 22517-22524.

232. Liu, J., Perumal, N.B., Oldfield, C.J., Su, E.W., Uversky, V.N. and Dunker, A.K. (2006) Intrinsic disorder in transcription factors. *Biochemistry*, **45**, 6873-6888.

233. Dunker, A.K., Silman, I., Uversky, V.N. and Sussman, J.L. (2008) Function and structure of inherently disordered proteins. *Current Opinion in Structural Biology*, **18**, 756-764.

234. Wang, J., Choi, J.-M., Holehouse, A.S., Lee, H.O., Zhang, X., Jahnel, M., Maharana, S., Lemaitre, R., Pozniakovsky, A., Drechsel, D. *et al.* (2018) A Molecular Grammar Governing the Driving Forces for Phase Separation of Prion-like RNA Binding Proteins. *Cell*, **174**, 688-699.e616.

235. Thiyagarajan, T., Ponnusamy, S., Hwang, D.-J., He, Y., Asemota, S., Young, K.L., Johnson, D.L., Bocharova, V., Zhou, W., Jain, A.K. *et al.* (2023) Inhibiting androgen receptor splice variants with cysteine-selective irreversible covalent inhibitors to treat prostate cancer. *Proceedings of the National Academy of Sciences*, **120**, e2211832120.

236. Helsen, C. and Claessens, F. (2014) Looking at nuclear receptors from a new angle. *Molecular and Cellular Endocrinology*, **382**, 97-106.

237. Rastinejad, F., Ollendorff, V. and Polikarpov, I. (2015) Response to Moras et al. *Trends in Biochemical Sciences*, **40**, 290-292.

238. Moras, D., Billas, I.M.L., Rochel, N. and Klaholz, B.P. (2015) Structure–function relationships in nuclear receptors: the facts. *Trends in Biochemical Sciences*, **40**, 287-290.
239. Rastinejad, F., Ollendorff, V. and Polikarpov, I. (2015) Nuclear receptor full-length architectures: confronting myth and illusion with high resolution. *Trends Biochem Sci*, **40**, 16-24.
240. Chandra, V., Wu, D., Li, S., Potluri, N., Kim, Y. and Rastinejad, F. (2017) The quaternary architecture of RARβ–RXRα heterodimer facilitates domain–domain signal transmission. *Nature Communications*, **8**, 868.
241. Chandra, V., Huang, P., Hamuro, Y., Raghuram, S., Wang, Y., Burris, T.P. and Rastinejad, F. (2008) Structure of the intact PPAR-γ–RXR-α nuclear receptor complex on DNA. *Nature*, **456**, 350.
242. Rochel, N., Ciesielski, F., Godet, J., Moman, E., Roessle, M., Peluso-Iltis, C., Moulin, M., Haertlein, M., Callow, P., Mély, Y. *et al.* (2011) Common architecture of nuclear receptor heterodimers on DNA direct repeat elements with different spacings. *Nature Structural &Amp; Molecular Biology*, **18**, 564.
243. Orlov, I., Rochel, N., Moras, D. and Klaholz, B.P. (2012) Structure of the full human RXR/VDR nuclear receptor heterodimer complex with its DR3 target DNA. *The EMBO Journal*, **31**, 291-300.
244. Haelens, A., Tanner, T., Denayer, S., Callewaert, L. and Claessens, F. (2007) The Hinge Region Regulates DNA Binding, Nuclear Translocation, and Transactivation of the Androgen Receptor. *Cancer Research*, **67**, 4514-4523.
245. Miyamoto, T., Kakizawa, T., Ichikawa, K., Nishio, S., Takeda, T., Suzuki, S., Kaneko, A., Kumagai, M., Mori, J., Yamashita, K. *et al.* (2001) The role of hinge domain in heterodimerization and specific DNA recognition by nuclear receptors. *Mol Cell Endocrinol*, **181**, 229-238.
246. Maletta, M., Orlov, I., Roblin, P., Beck, Y., Moras, D., Billas, I.M.L. and Klaholz, B.P. (2014) The palindromic DNA-bound USP/EcR nuclear receptor adopts an asymmetric organization with allosteric domain positioning. *Nature Communications*, **5**, 4139.

247. Lou, X., Toresson, G., Benod, C., Suh, J.H., Philips, K.J., Webb, P. and Gustafsson, J.-A. (2014) Structure of the retinoid X receptor α–liver X receptor β (RXRα–LXRβ) heterodimer on DNA. *Nature Structural & Molecular Biology*, **21**, 277-281.

248. van der Lee, R., Buljan, M., Lang, B., Weatheritt, R.J., Daughdrill, G.W., Dunker, A.K., Fuxreiter, M., Gough, J., Gsponer, J., Jones, D.T. *et al.* (2014) Classification of Intrinsically Disordered Regions and Proteins. *Chemical Reviews*, **114**, 6589-6631.

249. Lin, K.-H., Parkison, C., McPhie, P. and Cheng, S.-y. (1991) An Essential Role of Domain D in the Hormone-Binding Activity of Human β1 Thyroid Hormone Nuclear Receptor. *Molecular Endocrinology*, **5**, 485-492.

250. Johnson, Q.R., Lindsay, R.J., Nellas, R.B., Fernandez, E.J. and Shen, T. (2015) Mapping Allostery through Computational Glycine Scanning and Correlation Analysis of Residue–Residue Contacts. *Biochemistry*, **54**, 1534-1541.

251. Rastinejad, F., Perlmann, T., Evans, R.M. and Sigler, P.B. (1995) Structural determinants of nuclear receptor assembly on DNA direct repeats. *Nature*, **375**, 203-211.

252. Putcha, B., Wright, E., Brunzelle, J. and Fernandez, E. (2012) Structural basis for negative cooperativity within agonist-bound TR:RXR heterodimers. *Proceedings of the NAtional Academy of Sciences U.S.A.*, 6084-6087.

253. Schrodinger, LLC. (2015).

254. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A. *et al.* (2021) Highly accurate protein structure prediction with AlphaFold. *Nature*, **596**, 583-589.

255. Fiser, A., Do Richard Kinh, G. and Šali, A. (2008) Modeling of loops in protein structures. *Protein Science*, **9**, 1753-1773.

256. Katz, R.W. and Koenig, R.J. (1993) Nonbiased Identification of DNA Sequences That Bind Thyroid Hormone Receptor alpha-1 with High affinity. *The Journal of Biological Chemistry*, 1932-1937.

257. Koenig, R.J., Brent, G.A., Warne, R.L., Larsen, P.R. and Moore, D.D. (1987) Thyroid hormone receptor binds to a site in the rat growth hormone promoter required for induction by thyroid hormone. *Proc Natl Acad Sci U S A*, **84**, 5670-5674.

258. Macke, T.J., Svrcek-Seiler, W.A., Brown, R.A. and Case, D.A. (2006) *NAB (Nucleic Acid Builder) User's Manual*.

259. Cassandri, M., Smirnov, A., Novelli, F., Pitolli, C., Agostini, M., Malewicz, M., Melino, G. and Raschellà, G. (2017) Zinc-finger proteins in health and disease. *Cell Death Discovery*, **3**, 17071.

260. Landau, L.D. and Lifshitz, E.M. (2013) *Statistical Physics*. Butterworth-Heinemann.

261. Kolinski, A., Godzik, A. and Skolnick, J. (1993) A general method for the prediction of the three dimensional structure and folding pathway of globular proteins: Application to designed helical proteins. *The Journal of Chemical Physics*, **98**, 7420-7433.

262. Johnson, Q.R., Lindsay, R.J. and Shen, T. (2018) CAMERRA: An analysis tool for the computation of conformational dynamics by evaluating residue–residue associations. *Journal of Computational Chemistry*, 1568-1578.

263. Brunton, S.L. and Kutz, J.N. (2019) *Data-Driven Science and Engineering: Machine Learning, Dynamical Systems, and Control*. Cambridge University Press.

264. Chollet, F. and Chollet, F. (2021) *Deep Learning with Python, Second Edition*. Manning.

265. Lindsay, R.J., Pham, B., Shen, T. and McCord, R.P. (2018) Characterizing the 3D structure and dynamics of chromosomes and proteins in a common contact matrix framework. *Nucleic acids research*, **46**, 8143-8152.

266. Shen, T., Zong, C., Hamelberg, D., Andrew Mccammon, J. and Wolynes, P.G. (2005) The folding energy landscape and phosphorylation: modeling the conformational switch of the NFAT regulatory domain. *The FASEB Journal*, **19**, 1389-1395.

267. Johnson, Q.R., Lindsay, R.J., Nellas, R.B. and Shen, T. (2016) Pressure-induced conformational switch of an interfacial protein. *Proteins: Structure, Function, and Bioinformatics*, **84**, 820-827.

268. Pham, B., Lindsay, R.J. and Shen, T. (2019) Effector-Binding-Directed Dimerization and Dynamic Communication between Allosteric Sites of Ribonucleotide Reductase. *Biochemistry*, **58**, 697-705.

269. Clark, A.K., Wilder, J.H., Grayson, A.W., Johnson, Q.R., Lindsay, R.J., Nellas, R.B., Fernandez, E.J. and Shen, T. (2016) The Promiscuity of Allosteric Regulation of Nuclear Receptors by Retinoid X Receptor. *The Journal of Physical Chemistry B*, **120**, 8338-8345.

270. Lindsay, R.J., Mansbach, R.A., Gnanakaran, S. and Shen, T. (2021) Effects of pH on an IDP conformational ensemble explored by molecular dynamics simulation. *Biophysical Chemistry*, **271**, 106552.

271. McInnes, L., Healy, J. and Melville, J. (2020) UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv*, doi.org/10.48550/arXiv.1802.03426.

272. van der Maaten, L. and Hinton, G. (2008) Visualizing Data using t-SNE. *Journal of Machine Learning Research*, **9**, 2579-2605.

273. Frank, F., Okafor, C.D. and Ortlund, E.A. (2018) The first crystal structure of a DNA-free nuclear receptor DNA binding domain sheds light on DNA-driven allostery in the glucocorticoid receptor. *Scientific Reports*, **8**, 13497.

274. van Tilborg, M.A.A., Lefstin, J.A., Kruiskamp, M., Teuben, J.M., Boelens, R., Yamamoto, K.R. and Kaptein, R. (2000) *Mutations in the glucocorticoid receptor DNA-binding domain mimic an allosteric effect of DNA11Edited by P. E. Wright*.

275. Kabsch, W. and Sander, C. (1983) Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577-2637.

276. McInerney, E.M., Rose, D.W., Flynn, S.E., Westin, S., Mullen, T.M., Krones, A., Inostroza, J., Torchia, J., Nolte, R.T., Assa-Munt, N. *et al.* (1998) Determinants of coactivator LXXLL motif specificity in nuclear receptor transcriptional activation. *Genes Dev*, **12**, 3357-3368.

277. Tien, A.H. and Sadar, M.D. (2018) Order within a Disordered Structure. *Structure*, **26**, 4-6.

278. Drews, J. (2000) Drug Discovery: A Historical Perspective. *Science*, **287**, 1960-1964.

279. Gasparian, A.V., Burkhart, C.A., Purmal, A.A., Brodsky, L., Pal, M., Saranadasa, M., Bosykh, D.A., Commane, M., Guryanova, O.A., Pal, S. *et al.* (2011) Curaxins: anticancer compounds that simultaneously suppress NF-κB and activate p53 by targeting FACT. *Sci Transl Med*, **3**, 95ra74.
280. Thorn, C.F., Oshiro, C., Marsh, S., Hernandez-Boussard, T., McLeod, H., Klein, T.E. and Altman, R.B. (2011) Doxorubicin pathways: pharmacodynamics and adverse effects. *Pharmacogenet Genomics*, **21**, 440-446.
281. Zhang, Z., Pan, T., Liu, C., Shan, X., Xu, Z., Hong, H., Lin, H., Chen, J. and Sun, H. (2021) Cyclophosphamide induced physiological and biochemical changes in mice with an emphasis on sensitivity analysis. *Ecotoxicol Environ Saf*, **211**, 111889.
282. Morávek, Z., Neidle, S. and Schneider, B. (2002) Protein and drug interactions in the minor groove of DNA. *Nucleic Acids Research*, **30**, 1182-1191.
283. Reddy, B.S.P., Sondhi, S.M. and Lown, J.W. (1999) Synthetic DNA minor groove-binding drugs. *Pharmacology & Therapeutics*, **84**, 1-111.
284. Tabernero, L., Bella, J. and Alemán, C. (1996) Hydrogen Bond Geometry in DNA-Minor Groove Binding Drug Complexes. *Nucleic Acids Research*, **24**, 3458-3466.
285. Ragazzon, P.A., Garbett, N.C. and Chaires, J.B. (2007) Competition dialysis: a method for the study of structural selective nucleic acid binding. *Methods*, **42**, 173-182.
286. Ren, J. and Chaires, J.B. (2001) Rapid screening of structurally selective ligand binding to nucleic acids. *Methods Enzymol*, **340**, 99-108.
287. (!!! INVALID CITATION !!! (2)).
288. Dallavalle, S., Mattio, L.M., Artali, R., Musso, L., Avino, A., Fabrega, C., Eritja, R., Gargallo, R. and Mazzini, S. (2021) Exploring the Interaction of Curaxin CBL0137 with G-Quadruplex DNA Oligomers. *Int J Mol Sci*, **22**.
289. Hu, J., Lieb, J.D., Sancar, A. and Adar, S. (2016) Cisplatin DNA damage and repair maps of the human genome at single-nucleotide resolution. *Proceedings of the National Academy of Sciences*, **113**, 11507-11512.
290. Gurova, K., Chang, H.W., Valieva, M.E., Sandlesh, P. and Studitsky, V.M. (2018) Structure and function of the histone chaperone FACT - Resolving FACTual issues. *Biochim Biophys Acta Gene Regul Mech*.

291. Sun, X., Zhang, J. and Cao, C. (2022) CTCF and Its Partners: Shaper of 3D Genome during Development. *Genes (Basel)*, **13**.
292. Formosa, T. and Winston, F. (2020) The role of FACT in managing chromatin: disruption, assembly, or repair? *Nucleic Acids Res*, **48**, 11929-11941.
293. Bhakat, K.K. and Ray, S. (2022) The FAcilitates Chromatin Transcription (FACT) complex: Its roles in DNA repair and implications for cancer therapy. *DNA Repair (Amst)*, **109**, 103246.
294. Song, H., Zeng, J., Roychoudhury, S., Biswas, P., Mohapatra, B., Ray, S., Dowlatshahi, K., Wang, J., Band, V., Talmon, G. *et al.* (2020) Targeting Histone Chaperone FACT Complex Overcomes 5-Fluorouracil Resistance in Colon Cancer. *Mol Cancer Ther*, **19**, 258-269.
295. Hsieh, F.K., Kulaeva, O.I., Orlovsky, I.V. and Studitsky, V.M. (2011) FACT in Cell Differentiation and Carcinogenesis. *Oncotarget*, **2**, 830-832.
296. Garcia, H., Miecznikowski, J.C., Safina, A., Commane, M., Ruusulehto, A., Kilpinen, S., Leach, R.W., Attwood, K., Li, Y., Degan, S. *et al.* (2013) Facilitates chromatin transcription complex is an "accelerator" tumor transformation and potential marker and target of aggressive cancers. *Cell Rep*, **4**, 159-173.
297. Chang, H.W., Valieva, M.E., Safina, A., Chereji, R.V., Wang, J., Kulaeva, O.I., Morozov, A.V., Kirpichnikov, M.P., Feofanov, A.V., Gurova, K.V. *et al.* (2018) Mechanism of FACT removal from transcribed genes by anticancer drugs curaxins. *Sci Adv*, **4**, eaav2131.
298. Klein, D.C., Lardo, S.M., McCannell, K.N. and Hainer, S.J. (2023) FACT regulates pluripotency through proximal and distal regulation of gene expression in murine embryonic stem cells. *BMC Biology*, **21**, 167.
299. Shen, Z., Formosa, T. and Tantin, D. (2018) FACT Inhibition Blocks Induction But Not Maintenance of Pluripotency. *Stem Cells Dev*, **27**, 1693-1701.
300. Kolundzic, E., Ofenbauer, A., Bulut, S.I., Uyar, B., Baytek, G., Sommermeier, A., Seelk, S., He, M., Hirsekorn, A., Vucicevic, D. *et al.* (2018) FACT Sets a Barrier for Cell Fate Reprogramming in Caenorhabditis elegans and Human Cells. *Dev Cell*, **46**, 611-626 e612.

301. Yarnell, A.T., Oh, S., Reinberg, D. and Lippard, S.J. (2001) Interaction of FACT, SSRP1, and the high mobility group (HMG) domain of SSRP1 with DNA damaged by the anticancer drug cisplatin. *J Biol Chem*, **276**, 25736-25741.
302. Stros, M., Launholt, D. and Grasser, K.D. (2007) The HMG-box: a versatile protein domain occurring in a wide variety of DNA-binding proteins. *Cell Mol Life Sci*, **64**, 2590-2606.
303. Pfab, A., Gronlund, J.T., Holzinger, P., Langst, G. and Grasser, K.D. (2018) The Arabidopsis Histone Chaperone FACT: Role of the HMG-Box Domain of SSRP1. *J Mol Biol*, **430**, 2747-2759.
304. Murugesapillai, D., McCauley, M.J., Maher, L.J., 3rd and Williams, M.C. (2017) Single-molecule studies of high-mobility group B architectural DNA bending proteins. *Biophys Rev*, **9**, 17-40.
305. Malarkey, C.S. and Churchill, M.E. (2012) The high mobility group box: the ultimate utility player of a cell. *Trends Biochem Sci*, **37**, 553-562.
306. Charles Richard, J.L., Shukla, M.S., Menoni, H., Ouararhni, K., Lone, I.N., Roulland, Y., Papin, C., Ben Simon, E., Kundu, T., Hamiche, A. *et al.* (2016) FACT Assists Base Excision Repair by Boosting the Remodeling Activity of RSC. *PLoS Genet*, **12**, e1006221.
307. Gao, Y., Li, C., Wei, L., Teng, Y., Nakajima, S., Chen, X., Xu, J., Leger, B., Ma, H., Spagnol, S.T. *et al.* (2017) SSRP1 Cooperates with PARP and XRCC1 to Facilitate Single-Strand DNA Break Repair by Chromatin Priming. *Cancer Res*, **77**, 2674-2685.
308. Kari, V., Shchebet, A., Neumann, H. and Johnsen, S.A. (2011) The H2B ubiquitin ligase RNF40 cooperates with SUPT16H to induce dynamic changes in chromatin structure during DNA double-strand break repair. *Cell Cycle*, **10**, 3495-3504.
309. Dinant, C., Ampatziadis-Michailidis, G., Lans, H., Tresini, M., Lagarou, A., Grosbart, M., Theil, A.F., van Cappellen, W.A., Kimura, H., Bartek, J. *et al.* (2013) Enhanced chromatin dynamics by FACT promotes transcriptional restart after UV-induced DNA damage. *Mol Cell*, **51**, 469-479.

310. Yang, J., Horton, J.R., Liu, B., Corces, V.G., Blumenthal, R.M., Zhang, X. and Cheng, X. (2023) Structures of CTCF-DNA complexes including all 11 zinc fingers. *Nucleic Acids Res*, **51**, 8447-8462.

311. Kadauke, S. and Blobel, G.A. (2009) Chromatin loops in gene regulation. *Biochim Biophys Acta*, **1789**, 17-25.

312. Debaugny, R.E. and Skok, J.A. (2020) CTCF and CTCFL in cancer. *Curr Opin Genet Dev*, **61**, 44-52.

313. Bailey, C.G., Gupta, S., Metierre, C., Amarasekera, P.M.S., O'Young, P., Kyaw, W., Laletin, T., Francis, H., Semaan, C., Sharifi Tabar, M. *et al.* (2021) Structure-function relationships explain CTCF zinc finger mutation phenotypes in cancer. *Cell Mol Life Sci*, **78**, 7519-7536.

314. Ohndorf, U.-M., Rould, M.A., He, Q., Pabo, C.O. and Lippard, S.J. (1999) Basis for recognition of cisplatin-modified DNA by high-mobility-group proteins. *Nature*, **399**, 708-712.

315. Trott, O. and Olson, A.J. (2010) AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J Comput Chem*, **31**, 455-461.

316. Eberhardt, J., Santos-Martins, D., Tillack, A.F. and Forli, S. (2021) AutoDock Vina 1.2.0: New Docking Methods, Expanded Force Field, and Python Bindings. *J Chem Inf Model*, **61**, 3891-3898.

317. Bienfait, B. and Ertl, P. (2013) JSME: a free molecule editor in JavaScript. *J Cheminform*, **5**, 24.

318. El-Hachem, N., Haibe-Kains, B., Khalil, A., Kobeissy, F.H. and Nemer, G. (2017) AutoDock and AutoDockTools for Protein-Ligand Docking: Beta-Site Amyloid Precursor Protein Cleaving Enzyme 1(BACE1) as a Case Study. *Methods Mol Biol*, **1598**, 391-403.

319. Morris, G.M., Huey, R., Lindstrom, W., Sanner, M.F., Belew, R.K., Goodsell, D.S. and Olson, A.J. (2009) AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. *J Comput Chem*, **30**, 2785-2791.

320. Ivani, I., Dans, P.D., Noy, A., Pérez, A., Faustino, I., Hospital, A., Walther, J., Andrio, P., Goñi, R., Balaceanu, A. *et al.* (2016) Parmbsc1: a refined force field for DNA simulations. *Nat Methods*, **13**, 55-58.

321. Cheng, Y., Korolev, N. and Nordenskiöld, L. (2006) Similarities and differences in interaction of K + and Na + with condensed ordered DNA. A molecular dynamics computer simulation study. *Nucleic Acids Research*, **34**, 686-696.

322. Lavery, R., Moakher, M., Maddocks, J.H., Petkeviciute, D. and Zakrzewska, K. (2009) Conformational analysis of nucleic acids revisited: Curves+. *Nucleic Acids Res*, **37**, 5917-5929.

323. Galindo-Murillo, R., Roe, D.R. and Cheatham, T.E. (2015) Convergence and reproducibility in molecular dynamics simulations of the DNA duplex d(GCACGAACGAACGAACGC). *Biochimica et Biophysica Acta (BBA) - General Subjects*, **1850**, 1041-1058.

CONCLUSION OVERALL

Biological regulation emerges from a continuum of structural organization that spans from chromosome-scale topology to the dynamic behavior of transcription factors and the mechanical properties of the DNA double helix. The studies presented in this dissertation demonstrate that these scales operate within a unified physical framework: genome architecture shapes the regulatory environment, protein conformational dynamics interpret this environment, and DNA mechanics modulate the accessibility and stability of regulatory complexes. By integrating chromatin network analysis, molecular dynamics simulations, machine-learning–based structural characterization, and computational DNA mechanics, this work provides a multiscale perspective on how cellular regulation is encoded in molecular structure.

At the level of nuclear architecture, chromosome structure networks reveal that interchromosomal communication is a coordinated and biologically meaningful feature of genome organization. Representing genomic loci as nodes and chromatin contacts as weighted edges captures structural information that traditional matrix-based analyses overlook, allowing the identification of loci that act as connectivity hubs, link chromatin states, or exhibit sensitivity to architectural perturbations such as condensin II depletion. These findings demonstrate that the nuclear genome functions as an interconnected three-dimensional system in which both cis and trans interactions contribute to regulatory specificity. They highlight the importance of viewing genome architecture not as a collection of independent structures but as a distributed network that supplies the spatial context for gene regulation.

At the scale of protein–DNA interactions, studies of the thyroid hormone receptor α (TRα) illustrate how intrinsically disordered regions shape regulatory outcomes. Long-timescale molecular dynamics simulations reveal that the disordered hinge region does not behave as a passive linker but instead forms a broad ensemble of conformations whose distribution depends strongly on DNA binding. Binding to DNA reorganizes hinge dynamics, alters interdomain orientations, and establishes distinct structural pathways through which DNA influences receptor communication. Machine-learning models

applied to simulation-derived contact maps provide an essential analytical complement, enabling the extraction of latent structural states that would be difficult to isolate through conventional descriptors. Together, these approaches show that the TRα hinge acts as a dynamic, information-processing element that transmits structural and regulatory signals across domains—an example of how intrinsically disordered domain contributes to allostery in multidomain transcription factors.

At the molecular level of DNA itself, computational analyses of small-molecule interactions demonstrate that gene expression can be modulated through physical changes in DNA structure rather than through genotoxic mechanisms. Curaxins, which reshape groove geometry, twist, and local curvature without inducing DNA breaks, exemplify a class of therapeutics that target chromatin architecture through mechanical perturbation. These structural alterations destabilize nucleosomes, reconfigure chromatin accessibility, and influence transcriptional programs in ways that differ fundamentally from classical DNA-damaging chemotherapeutics. The results underscore the regulatory significance of DNA mechanical properties and highlight the potential of DNA-shape–modifying strategies for therapeutic intervention.

Across these scales, a unifying principle emerges: biological regulation is encoded in structure, and meaningful insight requires analytical frameworks that bridge the organizational hierarchy from chromosomes to proteins to DNA. The approaches developed in this dissertation—network analysis for chromatin topology, integrative simulations and machine learning for transcription factor dynamics, and DNA mechanics modeling for small-molecule interactions—collectively demonstrate the value of studying regulatory systems as interdependent components of a multiscale physical architecture.

This structural perspective suggests several promising directions for future research. Chromosome network methods applied to single-cell Hi-C could reveal the heterogeneity and cell-to-cell variability of interchromosomal communication. Integrative models linking genome structure with transcriptional and epigenomic dynamics may help distinguish structural drivers of regulation from downstream consequences. The MD + machine-learning framework developed for TRα can be extended to other transcription factors with large disordered regions, particularly those central to developmental

programs and disease. Finally, the concept of non-genotoxic modulation of DNA shape opens new possibilities for designing therapeutic agents that target chromatin regulation through physical rather than chemical pathways.

In conclusion, this dissertation shows that structural organization at multiple levels—from genome folding to protein conformations to the physical properties of DNA—works together to control gene regulation. By developing computational frameworks that connect these scales, this work advances our understanding of how molecular architecture encodes biological function and lays the conceptual foundation for future research aiming to interpret and manipulate regulation through the lens of structure.

VITA

Yingjie Xu was born in Wuxi, Jiangsu, China. He received the Bachelor of Science degree in Bioscience from Huazhong University of Science and Technology in 2018 and the Master of Science degree in Biology with a concentration in Molecular Biology from Adelphi University in 2021. He pursued doctoral studies at the University of Tennessee, Knoxville, in the Genome Science and Technology program, where his research focused on chromosome structure networks, interchromosomal communication, nuclear receptor dynamics, and drug–DNA interactions.

During his doctoral work, he served as a teaching assistant for undergraduate laboratory and lecture courses in biology and contributed to several collaborative projects in genomics and computational biology. His dissertation research resulted in a first-author publication in Computational and Structural Biotechnology Journal. He is expected to receive the Doctor of Philosophy degree in Genome Science and Technology in December 2025.