T.R.
GEBZE TECHNICAL UNIVERSITY
GRADUATE SCHOOL

INVESTIGATION OF ENHANCER-PROTEIN-PROMOTER INTERACTIONS WITHIN REGULATORY LOOPS

EMEL KÖKREK

A THESIS OF DOCTORATE
DEPARTMENT OF BIOENGINEERING
BIOINFORMATICS AND SYSTEMS BIOLOGY PROGRAM

ADVISOR: ASSIST.PROF. DR. PINAR PİR

JANUARY 2025

T.R.
GEBZE TECHNICAL UNIVERSITY
GRADUATE SCHOOL

INVESTIGATION OF ENHANCER-PROTEIN-PROMOTER INTERACTIONS WITHIN REGULATORY LOOPS

EMEL KÖKREK

A THESIS OF DOCTORATE
DEPARTMENT OF BIOENGINEERING
BIOINFORMATICS AND SYSTEMS BIOLOGY PROGRAM

ADVISOR: ASSIST.PROF. DR. PINAR PİR

JANUARY 2025

T.C.
GEBZE TEKNİK ÜNİVERSİTESİ
LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ

TRANSKRİPSİYON DÜZENLEYİCİ DÖNGÜLERDE ENHANCER-PROTEİN-PROMOTÖR ETKİLEŞİMLERİNİN İNCELENMESİ

EMEL KÖKREK

DOKTORA TEZİ
BİYOMÜHENDİSLİK ANABİLİM DALI
BİYOİNFORMATİK VE SİSTEM BİYOLOJİSİ PROGRAMI

DANIŞMAN: DR. ÖĞR. ÜYESİ PINAR PİR

OCAK 2025

DOCTORATE JURY APPROVAL FORM

A thesis submitted by Emel KÖKREK, defended on 20/01/2025 before the jury formed with the 13/01/2025 date and 2025/05 numbered decision of the GTU Graduate Administration Board, has been accepted as a DOCTOR of PHILOSOPHY thesis in the Department of Bioengineering, Bioinformatics and Systems Biology Program.

JURY

MEMBER (THESIS ADVISOR)	: Assist. Prof. Dr. Pınar PİR
MEMBER	: Prof. Dr. Tunahan ÇAKIR
MEMBER	: Assoc. Prof. Dr. Ogün ADEBALİ
MEMBER	: Assist. Prof. Dr. Nuray SÖĞÜNMEZ ERDOĞAN
MEMBER	: Assoc. Prof. Dr. Beste TURANLI

APPROVAL

Gebze Technical University Graduate Administration Board

…..…/…..…/..…..… date and ……/…….. numbered decision.

SIGNATURE/SEAL

To the afraid but dreaming child in all

ABSTRACT

Understanding the intricate regulatory mechanisms driving transcriptional activity is critical to deciphering cellular differentiation, immune aging, and pathological phenotypes. This study investigates transcriptional loops, focusing on the bidirectional interplay between enhancers and promoters, utilizing cutting-edge network analysis methods and integrating multi-omics data, primarily derived from human samples. A significant methodological focus addressed the low-resolution limitations of Hi-C data. Benchmarking various resolution-enhancement methods revealed their utility for TAD-calling but also highlighted challenges in defining high-resolution (<1 kb) interregional contacts, leading to a strategic focus on available high-resolution Hi-C datasets for downstream analyses.

To explore enhancer modes of action, chromatin interactions were represented within a network framework, facilitating the analysis of multi-way interactions and their relationship to gene expression during cardiomyocyte differentiation. Despite limited data availability, the application of this framework successfully elucidated regulatory dynamics, demonstrating the robustness of network-based approaches. Complementing this, the study examined transcriptional protein complexes in aging immune T cells, revealing critical insights into how enriched transcription factors interact with regulatory chromatin regions. The convergence of these findings with previous epigenomic and genomic state studies underscores the interplay between chromatin accessibility, regulatory complexes, and gene expression changes.

The results emphasize the role of enhancer multiplicity in cellular differentiation and identify potential transcriptional regulators underlying immune aging. By bridging high-resolution chromatin interaction data, network analysis, and regulatory protein complex dynamics, this research offers a novel computational methodology for studying 3D genome architecture and transcriptional regulation. These findings provide a foundation for further exploration of transcriptional loops, advancing our understanding of cellular and pathological processes and paving the way for future therapeutic strategies targeting enhancer-promoter interactions and transcriptional complexes.

Keywords: Transcriptional Regulation, Enhancer-Promoter Interactions, Multi-Omics Integration, Hi-C Resolution Enhancement, Network Analysis, Regulatory Protein Complexes

ÖZET

Transkripsiyonel aktiviteyi yönlendiren karmaşık düzenleyici mekanizmaların anlaşılması, hücresel farklılaşma, bağışıklık yaşlanması ve patolojik fenotiplerin deşifre edilmesi için kritik öneme sahiptir. Bu çalışma, transkripsiyonel döngüleri araştırmakta, enhancer ve promotör arasındaki karşılıklı etkileşime odaklanmakta, ağ analizi yöntemlerini kullanmakta ve öncelikle insan örneklerinden elde edilen multiomiks verilerini entegre etmektedir. Çalışmanın önemli bir metodolojik odak noktası da, Hi-C verilerinin düşük çözünürlük sınırlamalarını ele almıştır. Çeşitli çözünürlük artırma yöntemlerinin karşılaştırılması, bu yöntemlerin TAD tanımlama için faydalarını ortaya koymuş olmakla birlikte yüksek çözünürlüklü (<1 kb) bölgeler arası temasların tanımlanmasındaki zorlukları vurgulayarak, ileri analizler için mevcut yüksek çözünürlüklü Hi-C veri kümelerine yöneltmiştir.

Enhancer çalışma biçimlerini keşfetmek adına kromatin etkileşimleri bir ağ çerçevesinde temsil edilerek, böylelikle çok yönlü etkileşimlerin ve bunların kardiyomiyosit farklılaşması sırasında gen ifadesiyle ilişkilerinin analizini kolaylaştırmıştır. Sınırlı veriye rağmen, bu çerçevenin uygulanması düzenleyici dinamikleri başarılı bir şekilde aydınlatarak ağ tabanlı yaklaşımların kullanışlılığını göstermiştir. Bunu tamamlayan çalışma, yaşlanan bağışıklık T hücrelerindeki transkripsiyonel protein komplekslerini inceleyerek, transkripsiyon faktörlerinin düzenleyici kromatin bölgeleri ile nasıl etkileşime girdiğine dair kritik bilgiler ortaya koymuştur. Bu bulguların önceki epigenomik ve genom çalışmalarıyla benzerlik göstermesi, kromatin erişilebilirliği, düzenleyici kompleksler ve gen ifadesi değişiklikleri arasındaki etkileşimin altını çizmektedir.

Sonuçlar, hücresel farklılaşmada enhancer etkileşimlerinin rolünü vurgulamakta ve bağışıklık yaşlanmasının altında yatan potansiyel transkripsiyonel düzenleyicileri tanımlamaktadır. Yüksek çözünürlüklü kromatin etkileşim verileri, ağ analizi ve düzenleyici protein kompleksi dinamikleri arasında köprü kuran bu araştırma, üç boyutlu genom mimarisi ve transkripsiyonel düzenlemeyi incelemek için yeni bir hesaplamalı metodoloji sunmaktadır. Bu bulgular, transkripsiyonel döngülerin daha fazla araştırılması için bir temel oluşturmakta, hücresel ve patolojik süreçleri anlamamızı ilerletmekte ve enhancer-promotör etkileşimlerini ve transkripsiyonel kompleksleri hedefleyen gelecekteki terapötik stratejilerin önünü açmaktadır.

Anahtar Kelimeler: Transkripsiyonel Regülasyon, Enhancer-Promotör Etkileşimleri, Çoklu-Omik Entegrasyonu, Hi-C Çözünürlük Geliştirme, Ağ Analizi, Düzenleyici Protein Kompleksleri

ACKNOWLEDGEMENTS

First and foremost, my sincere gratitude goes to my supervisor, Assist.Prof. Pınar Pir, for her support and for granting me the flexibility and freedom to explore fields I am passionate about. Providing such freedom involves a certain level of risk, and her trust in my abilities to navigate those challenges has been both empowering and deeply motivating. Her encouragement and confidence allowed me to nurture my curiosity and pursue a research path that truly excites me.

Profound thanks are extended to my committee members for their significant contributions and unwavering support. Prof. Tunahan Çakır, whose dedication, kindness, and professionalism have always been a source of inspiration, has served as a role model not only as an academician but also as a person. The courses I took with him and the knowledge he shared have profoundly enriched my academic journey, for which I remain deeply grateful. Equally appreciated is Assoc.Prof. Ogün Adebali, who has been with us since the proposal stage and has continuously provided thoughtful guidance throughout this process. His detailed feedback on my progress reports and genuine involvement in my research made me feel like one of his own PhD students, an experience for which I am sincerely thankful.

Gratitude also goes to my friends, Jafar Khatibipour (PhD), İsa Yüksel, and Emre Taylan Duman, whose unique blend of madness, sharp intellect, humor, and naivety has made PhD life's challenges bearable and restored my faith in humanity. To Gülsüm Yiğit (PhD), who has kept me grounded amidst all the chaos and craziness, I extend my deepest thanks. My heartfelt appreciation goes to Elif Emanetçi Düz, who embarked on the GTÜ journey with me, patiently endured my complaints, odd questions, and lengthy, convoluted writings, and has always been a source of unwavering support. I would also like to extend my gratitude to the entire crew of the Bioinformatics and Systems Biology (BSB) program. Although I cannot name each of you individually, your collective support, collaboration, and camaraderie have played a crucial role in making this journey both fulfilling and enjoyable. Special thanks go to the *tez teşvikleri* Group led by my dearest friend Ayşe Meryem, whose motto, "Whether you write your thesis or not, you'll die anyway—enjoy the process," has been a constant source of motivation and humor throughout this journey.

Finally, to my family, my unshakable core, I extend a gratitude that words can barely hold. Seeking the poet's forgiveness, "Beş ben vardır bende benden içeru"—we are five, but our circle has grown, enriched by my sister's and brother's partners and their children, each a precious part of our family. Within me, all of you live and shape me—and, naturally, provide plenty of material for years of therapy. Your unwavering love, patience, and support have been my anchor through this journey, and my appreciation for you transcends the bounds of language.

TABLE OF CONTENTS

	Page
ABSTRACT	vi
ÖZET	vii
ACKNOWLEDGEMENT	viii
TABLE of CONTENTS	ix
LIST of ABBREVIATIONS and ACRONYMS	xi
LIST of FIGURES	xiii
LIST of TABLES	xx
1. INTRODUCTION	1
1.1. Literature Review and Related Work	9
1.1.1. Enhancement of Hi-C Resolution	9
1.1.2. Network Representation of Chromatin Interaction Data	13
1.1.3. Network-Based Regulatory Protein Complex Prediction	17
1.1.4. Aging T cells	19
2. METHODOLOGY	21
2.1. Benchmarking Resolution Enhancement Models	21
2.1.1. Comparison Scenarios	21
2.1.2. Metrics for Model Evaluation	23
2.2. Promoter-Enhancer Network Analysis in Cardiomyocyte Differentiation	27
2.2.1. Data Retrieval and Data Processing	27
2.2.2. Generation of Promoter-Based Networks	28
2.2.3. Computational Framework and Integration	34
2.2.4. Correlation Analysis of Network Connectivity and Gene Ex-pression	34
2.2.5. Interpretation of Correlation Patterns	35
2.3. Deregulated Protein Complex Analysis	36
2.3.1. Comparing scRNA-seq and Bulk RNA-seq Techniques	36
2.3.2. Dataset Details	37
2.3.3. Sample-Specific Interaction Network Generation	38
2.3.4. Transcriptional Protein Complex Prediction	38
2.3.5. Protein Complex Quantification and Statistical Comparison	39
2.3.6. Principal Component Analysis and Clustering	39
2.3.7. Network Analysis and Functional Annotation	39
2.3.8. Experimental Validation of Predicted Complexes	40
2.3.9. Identification of Regulatory Loops for the Significantly Deregulated Protein Complexes	40
3. RESULTS	42

3.1. Benchmarking the Resolution Enhancement Models 42
3.1.1. 16-Fold Enhancement of Matrices from the Same Cell Line: GM12878 43
3.1.2. 16-Fold Enhancement of Matrices Across Different Cell Lines: IMR90, K562, and HEK 56
3.1.3. Sensitivity to Initial Sequencing Depth 60
3.1.4. 16-Fold Enhancement of Matrices Across Organism: mESC 77
3.1.5. Conclusions on Benchmarking of Resolution Enhancement Tools 84
3.2 Promoter-Enhancer Networks 85
3.2.1. High Positive Group: Balanced Enhancer-Promoter Interactions 85
3.2.2. High Negative Group: Repressive Enhancer-Promoter Interactions 87
3.2.3. Opposite High Group: Inverse Dynamics in Regulatory Networks 88
3.2.4. Specific Group: Dynamic Enhancer Interactions 88
3.2.5. Synergistic/Competitive Group: Context-Dependent Regulation 95
3.3 Deregulated Transcriptional Protein Complexes 97
3.3.1. Comparison of scRNA-seq and Bulk RNA-seq for Complex Deregulation Analysis 97
3.3.2. Dataset 1: Lack of Significant Age-Associated Changes in PBMC Bulk RNA-seq 100
3.3.3. Dataset 2: Impact of Aging and COVID-19 on Transcriptional Protein Complexes in Naive T Cells 101
3.3.4. Dataset 3: Distinct Age-Related Patterns in nCD4T and nCD8T 120
4. CONCLUSION 136

REFERENCES 138
BIOGRAPHY 149
PUBLICATIONS WITHIN the SCOPE of THESIS STUDY 150
APPENDICES 151
A. Benchmarking Resolution Enhancement Models 151

LIST of ABBREVIATIONS and ACRONYMS

3C, 4C, 5C, Hi-C	:	Chromosome Conformation Capture Techniques
Ahern22	:	Dataset from Ahern et al., 2022
ARNT	:	Aryl Hydrocarbon Receptor Nuclear Translocator
ATAC-seq	:	Assay for Transposase-Accessible Chromatin using
sequencing ATF2	:	Activating Transcription Factor 2
BCL11A	:	BAF Chromatin Remodeling Complex Subunit 11A
BED	:	Browser Extensible Data
BLOC1	:	Biogenesis of Lysosome-Related Organelles Complex 1
BLOC2	:	Biogenesis of Lysosome-Related Organelles Complex 2
bp	:	Base Pairs
CASQ2	:	Calsequestrin 2
CH	:	Child
CNN	:	Convolutional Neural Network
COVID	:	Coronavirus Disease
CPM	:	Counts Per Million
CYP2E1	:	Cytochrome P450 Family 2 Subfamily E Member 1
CXB	:	Cufflinks Binary Format
CFC	:	Complex Fold Change
DAE	:	Denoising Autoencoder
DDIN	:	Domain-Domain Interaction Network
DDI	:	Domain-Domain Interaction
DEG	:	Differentially Expressed Gene
DISP1	:	Dispatched RND Transporter Family Member 1
DTNBP1	:	Dystrobrevin Binding Protein 1
E2F	:	E2 Promoter Binding Factor
ESCs	:	Embryonic Stem Cells
FDR	:	False Discovery Rate
FISH	:	Fluorescence In Situ Hybridization
FOSL2	:	Fos Like 2, AP-1 Transcription Factor Subunit
FPKM	:	Fragments Per Kilobase of transcript per Million
mapped reads GAN	:	Generative Adversarial Network
GEO	:	Gene Expression Omnibus
GM12878	:	Human B-Lymphocyte Cell Line
GTÜ	:	Gebze Teknik Üniversitesi
GWAS	:	Genome-Wide Association Studies
HiCARN	:	Hi-C Artifact Removal Network
HiCNN	:	High-Resolution Convolutional Neural Network
HiCSR	:	Hi-C Super-Resolution
HMGN4	:	High Mobility Group Nucleosomal Binding Domain 4
HOCOMOCO	:	Human Transcription Factor Database
HPS6	:	Hermansky-Pudlak Syndrome 6
IMR90	:	Human Fetal Lung Fibroblast Cell Line
KR	:	Knight-Ruiz Normalization
MACS2	:	Model-based Analysis of ChIP-Seq 2
MAML	:	Mastermind-like Coactivator

METTL16	:	Methyltransferase 16
MSE	:	Mean Squared Error
mESC	:	Mouse Embryonic Stem Cell
NHEK	:	Normal Human Epidermal Keratinocytes
Nkx2.5	:	NK2 Homeobox 5
NS	:	Not Significant
OA	:	Older Adult
ORFs	:	Open Reading Frames
PBMCs	:	Peripheral Blood Mononuclear Cells
PCA	:	Principal Component Analysis
PL	:	Perceptual Loss
PPIN	:	Protein-Protein Interaction Network
PPI	:	Protein-Protein Interaction
PPM1D	:	Protein Phosphatase, Mg2+/Mn2+ Dependent 1D
PSNR	:	Peak Signal-to-Noise Ratio
PTK2	:	Protein Tyrosine Kinase 2
RBPJ	:	Recombining Binding Protein Suppressor of Hairless
RNA-seq	:	RNA sequencing
RORC	:	RORγT (RAR-related orphan receptor gamma)
ROS	:	Reactive Oxygen Species
SAM	:	S-Adenosyl Methionine
SiComPre	:	Simulation of Complex Prediction
SRA	:	Sequence Read Archive
SRHiC	:	Super-Resolution Hi-C
SSIM	:	Structural Similarity Index
STAR	:	Spliced Transcripts Alignment to a Reference
TAD	:	Topologically Associated Domain
TBX21	:	T-bet (T-box transcription factor 21)
TF	:	Transcription Factor
TCF4	:	Transcription Factor 4
TRMT1	:	tRNA Methyltransferase 1
TV	:	Total Variation Loss
TSS	:	Transcription Start Site
UTP6	:	U3 Small Nucleolar Ribonucleoprotein 6
USP39	:	Ubiquitin Specific Peptidase 39
VAE	:	Variational Autoencoder
YA	:	Young Adult

LIST OF FIGURES

Page

Figure 1.1: Regulatory Loop and its Elements: Enhancer, Promoter, Transcription Factor(TF), Co-factor (Co-TF), Cohesin, Mediator Complex, RNA Polymerase II. 2
Figure 1.2: Proximity Ligation Based Chromatin Capture Method. 3
Figure 1.3: Hi-C Pipeline From Raw Reads to Contact Matrices. 4
Figure 1.4: Structural Patterns Detected on Contact Matrices. 5
Figure 1.5: Methodology to Investigate the 3D Regulatory Loop. 8

Figure 2.1: Training A Resolution Enhancement Model. 21
Figure 2.2: Pearson Correlation Coefficient calculated at different distances. MSE, PSNR and SSIM metrics compared within 1Mb submatrices. 25
Figure 2.3: TADCompare defined differential TAD categories: Merge, Split, Shifted, Strength Change, Complex. 26
Figure 2.4: Route 1 to Investigate Promoter-Interaction Networks. 28
Figure 2.5: Route 2 to Investigate Promoter-Interaction Networks. 29
Figure 2.6: Size distribution of elements across chromosomes. 31
Figure 2.7: Association of Regulatory Elements for the DEG ZNF143, promoter identification, determining the active elements by incorporating accessibility data. 33
Figure 2.8: Incorporating interaction data at each time point to the initial fully connected network of ZNF143 to make it stage-specific. 34
Figure 2.9: Spearman Correlation between the total number of edges in the network and expression level: CORRS_EDGES. Spearman Correlation between changes in network structure, measured by Hamming distance and changes in expression level: CORRS_CHANGES. 35

Figure 3.1: (a) MSE values of each model in 4 different human cell lines: GM12878 primary, IMR90, K562 and NHEK. Enhancement ratio is x16. (b) MSE values of each model (excluding SRHiC) in 4 different human cell lines: GM12878 primary, IMR90, K562 and NHEK. Enhancement ratio is x16. 44
Figure 3.2: (a) PSNR values of each model in 4 different human cell lines: GM12878 primary, IMR90, K562 and NHEK. Enhancement ratio is x16. (b) SSIM values of each model in 4 different human cell lines: GM12878 primary, IMR90, K562 and NHEK. Enhancement ratio is x16. 46
Figure 3.3: Heatmap of a 1 Mb region of the GM12878 primary cell line, corresponding to chromosome 13: 20,000,000–21,010,000. The heatmap shows the original high-resolution matrix, the downsampled low-resolution matrix, and the ×16 enhanced matrices generated by various models. 47

Figure 3.4: Heatmap of a 1 Mb region of the GM12878 primary cell line, corresponding to chromosome 16: 1,500,000–2,510,000. The heatmap shows the original high-resolution matrix, the down-sampled low-resolution matrix, and the ×16 enhanced matrices generated by various models. 48
Figure 3.5: Pearson Correlation values of each model in 4 different human cell lines: GM12878 primary, IMR90, K562 and NHEK. Enhancement ratio is x16. 49
Figure 3.6: (a) Differential Merge, Shifted, Split, Strength Change TADs (b) Differential Complex TADs in all human cell lines at enhancement ratio of x16. 51
Figure 3.7: Non-differential TADs in all human cell lines at enhancement ratio of x16. 53
Figure 3.8: Heatmap of a 1 Mb region of the IMR90 cell line, corresponding to chromosome 16: 1,500,000–2,510,000. The heatmap shows the original high-resolution matrix, the down-sampled low-resolution matrix, and the ×16 enhanced matrices generated by various models. 58
Figure 3.9: Heatmap of a 1 Mb region of the (a) K562 and (b) NHEK cell lines, corresponding to chromosome 16: 1,500,000–2,510,000. The heatmap shows the original high-resolution matrix, the down-sampled low-resolution matrix, and the ×16 enhanced matrices generated by various models. 59
Figure 3.10: MSE values of each model in 4 different human cell lines: GM12878 primary, IMR90, K562 and NHEK. Enhancement ratio is x25 61
Figure 3.11: (a) PSNR and (b) SSIM values of each model in 4 different human cell lines: GM12878 primary, IMR90, K562 and NHEK. Enhancement ratio is x25. 62
Figure 3.12: Pearson Correlation values of each model in 4 different human cell lines: GM12878 primary, IMR90, K562 and NHEK. Enhancement ratio is x25. 64
Figure 3.13: Non-differential TADs in all human cell lines at enhancement ratio of x25. 65
Figure 3.14: MSE values of each model in 4 different human cell lines: GM12878 primary, IMR90, K562 and NHEK. Enhancement ratio is x32. 67
Figure 3.15: (a) PSNR values of each model in 4 different human cell lines: GM12878 primary, IMR90, K562 and NHEK. Enhancement ratio is x32. (b) SSIM values of each model in 4 different human cell lines: GM12878 primary, IMR90, K562 and NHEK. Enhancement ratio is x32. 68
Figure 3.16: Pearson Correlation values of each model in 4 different human cell lines: GM12878 primary, IMR90, K562 and NHEK. Enhancement ratio is x32. 69
Figure 3.17: Non-differential TADs in all human cell lines at enhancement ratio of x32. 70
Figure 3.18: MSE values of each model in 4 different human cell lines: GM12878 primary, IMR90, K562 and NHEK. Enhancement ratio is x50. 71

Figure 3.19: (a) PSNR and (b) SSIM values of each model in 4 different human cell lines: GM12878 primary, IMR90, K562 and NHEK. Enhancement ratio is x50. 72
Figure 3.20: Pearson Correlation values of each model in 4 different human cell lines: GM12878 primary, IMR90, K562 and NHEK. Enhancement ratio is x50. 73
Figure 3.21: Non-differential TADs in all human cell lines at enhancement ratio of x50. 74
Figure 3.22: MSE values of each model in mouse embryonic stem cells (mESC). Enhancement ratio is x16. 78
Figure 3.23: (a) PSNR and (b) SSIM values of each model in mouse embryonic stem cells (mESC). Enhancement ratio is x16. 79
Figure 3.24: Pearson Correlation values of each model in mouse embryonic stem cells (mESC). Enhancement ratio is x16. 81
Figure 3.25: (a) Differential TADs (excluding Complex TADs) and (b) Complex TADs in mESCs at enhancement ratio of x16. 82
Figure 3.26: Non-differential TADs in mESCs at enhancement ratio of x16. 83
Figure 3.27: Enhancer Mode of Action Groups. 85
Figure 3.28: Change in Enhancer Network of ARNT (Promoter:GH01J150874) During Cardiomyocyte Differentiation. Elements shown in red are both promoters and enhancers whereas those shown in dark blue are enhancers. Element IDs are shortened in a way to follow common pattern: GH01J15. 89
Figure 3.29: Change in Enhancer Network of BCL11A (Promoter:GH02J060543) During Cardiomyocyte Differentiation. Elements shown in dark blue are enhancers. Element IDs are shortened in a way to follow common pattern: GH02J06. 90
Figure 3.30: Change in Enhancer Network of FOSL2 (Promoter:GH02J028389) During Cardiomyocyte Differentiation. Elements shown in dark blue are enhancers. Element IDs are shortened in a way to follow common pattern: GH02J02. 91
Figure 3.31: Enrichment of 195 Genes in the Specific Group Using TRANSFAC Database. 92
Figure 3.32: Number of Complexes Compared Between Cell Types in Each Study. 98
Figure 3.33: Number of Complexes Compared Between Studies for Each Cell Type. 100
Figure 3.34: Graphical Abstract of the Analysis Made with Dataset 2. 101
Figure 3.35: Number of mapped proteins, interactions and unique protein complexes predicted for nCD4T and nCD8T cells in 4 different conditions: Healthy Child, Healthy Adult, COVID Child, COVID Adult. Groups were compared by Wilcoxon rank sum exact test. NS: Not Significant (p-value > 0.05), *: Significant (p-value < 0.05), **: Significant (p-value < 0.01) ***: Significant (p-value < 0.001). 103

Figure 3.36: Intersection of deregulated protein complexes in age factor tests of nCD4T (A) and nCD8T (D) cells. Deregulated protein complexes in disease factor tests of nCD4T (B) and nCD8T (E) cells. Intersection of enriched transcription factors in deregulated protein complexes in age factor tests of nCD4T (C) and nCD8T (F) cells. 106

Figure 3.37: Top 25 deregulated regulatory protein complexes healthy children vs healthy adults comparison of nCD4T cells and their log2 fold change as a complex. Transcription factors in these complexes showing differential expression is represented as red or green dots and with corresponding gene name: Red dots representing the downregulation, green dots representing the upregulation. The position of the dot also shows the log2 fold change of the individual gene expression. 108

Figure 3.38: Top 25 deregulated regulatory protein complexes in COVID children vs COVID adults comparison of nCD4T cells and their log2 fold change as a complex. Transcription factors in these complexes showing differential expression is represented as red or green dots and with corresponding gene name: Red dots representing the downregulation, green dots representing the upregulation. The position of the dot also shows the log2 fold change of the individual gene expression. 109

Figure 3.39: Network representation and gene enrichment analysis of unique proteins involved in various protein complexes across different conditions of nCD4T cells: 25 proteins from 10 ATF2-complexes in the healthy age comparison, 29 proteins from 20 RBPJ- and TCF4-complexes in the COVID age comparison, and 34 proteins from 25 DDIT3- and ATF3-complexes in the pediatric infection comparison, analyzed using Reactome pathway terms. 110

Figure 3.40: Top 25 deregulated regulatory protein complexes healthy children vs healthy adults comparison of nCD8T cells and their log2 fold change as a complex. Transcription factors in these complexes showing differential expression is represented as red or green dots and with corresponding gene name: Red dots representing the downregulation, green dots representing the upregulation. The position of the dot also shows the log2 fold change of the individual gene expression. 111

Figure 3.41: Top 25 deregulated regulatory protein complexes in COVID children vs COVID adults comparison of nCD8T cells and their log2 fold change as a complex. Transcription factors in these complexes showing differential expression is represented as red or green dots and with corresponding gene name: Red dots representing the downregulation, green dots representing the upregulation. The position of the dot also shows the log2 fold change of the individual gene expression. 113

Figure 3.42: Network representation and gene enrichment analysis of unique proteins involved in various protein complexes across different conditions of nCD8T cells: 20 proteins from 10 BCL11A-complexes in the healthy age comparison, 27 proteins from 13 BCL11A-complexes in the COVID age comparison, and 29 proteins from 25 DDIT3- and ATF3-complexes in the pediatric infection comparison, analyzed using Reactome pathway terms. 114

Figure 3.43: Top 25 deregulated regulatory protein complexes in Healthy children vs COVID children comparison of nCD4T cells and their log2 fold change as a complex. Transcription factors in these complexes showing differential expression is represented as red or green dots and official gene name: Red dots representing the downregulation, green dots representing the upregulation. The position of the dot also shows the log2 fold change of the individual gene expression. 117

Figure 3.44: Top 25 deregulated regulatory protein complexes in Healthy children vs COVID children comparison of nCD8T cells and their log2 fold change as a complex. Transcription factors in these complexes showing differential expression is represented as red or green dots and official gene name: Red dots representing the downregulation, green dots representing the upregulation. The position of the dot also shows the log2 fold change of the individual gene expression. 118

Figure 3.45: Graphical Abstract of the Analysis Made with Dataset 3. 120

Figure 3.46: (A) Number of unique protein complexes predicted and compared for three age groups Child, Young Adult (Yadult), Older Adult (Oadult) in two cell types: naive CD4+ T cells (nCD4T) and naive CD8+ T cells (nCD8T). (B) Number of unique protein-protein interactions detected and compared for three age groups Child, Young Adult (Yadult), Older Adult (Oadult) in two cell types: naive CD4+ T cells (nCD4T) and naive CD8+ T cells (nCD8T). (C) Number of unique proteins mapped and compared for three age groups Child, Young Adult (Yadult), Older Adult (Oadult) in two cell types: naive CD4+ T cells (nCD4T) and naive CD8+ T cells (nCD8T). Wilcoxon rank sum exact test is used in all comparisons. NS: Not Significant p-value > 0.05, *: Significant p-value < 0.05, **: Significant p-value < 0.01 ***: Significant p-value < 0.001. 122

Figure 3.47: Number of up- and down-regulated protein complexes in three age group comparisons: Child vs Young Adult (CH_YA), Young Adult vs Older Adult (YA_OA), Child vs Older Adult (CH_OA) for two cell types: naive CD4+ T cells (nCD4T) and naive CD8+ T cells (nCD8T). 124

Figure 3.48: Representative deregulated protein complexes that show monotonous incline and decline across three age groups Child, Young Adult (Yadult), Older Adult (Oadult) and in two cell types: naive CD4+ T cells (nCD4T) and naive CD8+ T cells (nCD8T). Total number of complexes are shown at the top of each panel (A) Abundance values of 3 selected continuously up-regulated protein complexes: CD40 – CD40LG – TRAF3IP2 – PRKDC – STAT6; HDAC5 – SMAD3 – KLF4 – RUNX2 – HDAC7; ACTR3 – SMAD3 – JUND – FOS – ETS1. (B) Abundance values of 3 selected continuously down-regulated protein complexes: PHF20L1 – MTA1 – BCL11A – CHD4 – GATAD2B; BCL11B – BCL11A – KZF1 – PATZ1 – CRKL; CTR9 – SOX4. Wilcoxon rank sum exact test is used in all comparisons. NS: Not Significant p-value > 0.05, *: Significant p-value < 0.05, **: Significant p-value < 0.01 ***: Significant p-value < 0.001. 125

Figure 3.49: PCA of abundance data of (A) 70 deregulated trajectory protein complexes, (B) 39 SMAD3-containing up-regulated trajectory protein complexes and (C) 13 BCL11A-containing down-regulated trajectory protein complexes in three age groups Child, Young Adult (Yadult), Older Adult (Oadult) of naive CD4+ T cells represented on the first two principal components out of 20 computed. Results of K-means clustering represented on two principal components and calculated evaluation metrics: Accuracy, Recall, Precision and F1-score. 127

Figure 3.50: Network representation of 70 trajectory protein complexes of naive CD4+ T cells. MCD3 protein is indicated with a black arrowhead as the link between upper SMAD3 cluster and lower BCL11A cluster. The width of the edges reflects the frequency of the interaction in complexes. SMAD3 is the highest degree node of the network with 50 interactions of which the most frequent (5 times) partner is E3 SUMO-protein ligase PIAS3 (PIAS3). FOS comes the second in degree with 14 and Protein FosB (FOSB) is the most frequent (3 times) interacting partner. BCL11A is the third with 12 and it has equally (4 times) interacting partners Chromodomain-helicase-DNA-binding protein 4 (CHD4) and Metastasis-associated protein MTA1 (MTA1). SMAD3 and FOS appear in up-regulated complexes and BCL11A appears in the continuously down-regulated complexes. Colorless nodes represent proteins which are not detected at protein level and pink nodes represent proteins which are detected at protein level in naïve CD4+ T cells. 129

Figure 3.51: Regulatory Regions That Can be Bound by all Three TFs and their Target Genes. 133

Figure 3.52: Network visualizations of PKNOX1, RUNX1, and IFNAR2. 134

Figure A.1: Heatmap of a 1 Mb region of the GM12878 primary cell line, corresponding to chromosome 16: 3,150,000-4,160,000. The heatmap shows the original high-resolution matrix, the down-sampled low-resolution matrix, and the ×16 enhanced matrices generated by various models. 151
Figure A.2: Heatmap of a 1 Mb region of the IMR90 cell line, corresponding to chromosome 16: 3,150,000-4,160,000. The heatmap shows the original high-resolution matrix, the down-sampled low-resolution matrix, and the ×16 enhanced matrices generated by various models. 152
Figure A.3: Heatmap of a 1 Mb region of the K562 cell line, corresponding to chromosome 16: 3,150,000-4,160,000. The heatmap shows the original high-resolution matrix, the down-sampled low-resolution matrix, and the ×16 enhanced matrices generated by various models. 153
Figure A.4: Heatmap of a 1 Mb region of the NHEK cell line, corresponding to chromosome 16: 3,150,000-4,160,000. The heatmap shows the original high-resolution matrix, the down-sampled low-resolution matrix, and the ×16 enhanced matrices generated by various models. 154
Figure A.5: Heatmap of a 1 Mb region of the GM12878 primary cell line, corresponding to chromosome 16: 8,150,000-9,160,000. The heatmap shows the original high-resolution matrix, the down-sampled low-resolution matrix, and the ×16 enhanced matrices generated by various models. 155
Figure A.6: Heatmap of a 1 Mb region of the IMR90 cell line, corresponding to chromosome 16: 8,150,000-9,160,000. The heatmap shows the original high-resolution matrix, the down-sampled low-resolution matrix, and the ×16 enhanced matrices generated by various models. 156
Figure A.7: Heatmap of a 1 Mb region of the K562 cell line, corresponding to chromosome 16: 8,150,000-9,160,000. The heatmap shows the original high-resolution matrix, the down-sampled low-resolution matrix, and the ×16 enhanced matrices generated by various models. 157
Figure A.8: Heatmap of a 1 Mb region of the NHEK cell line, corresponding to chromosome 16: 8,150,000-9,160,000. The heatmap shows the original high-resolution matrix, the down-sampled low-resolution matrix, and the ×16 enhanced matrices generated by various models. 158

LIST OF TABLES

Page

Table 1.1: Summary of Deep Learning Models for Hi-C Matrix Super-Resolution, Including Architectures, Key Features, Loss Functions, and Trainable Weights 10

Table 2.1: Size distribution of elements across chromosomes. Includes mean, median, minimum, maximum, and standard deviation of element sizes, along with the number of promoters(P), enhancers(E), and promoter/enhancer(P/E) combinations. 32

Table 3.1: Pair-wise comparison significance values for Non-differential TADs in GM12878 cell line of x16 ratio. 53

Table 3.2: Pair-wise comparison significance values for Non-differential TADs in GM12878 cell line of x25 ratio. 66

Table 3.3: Pair-wise comparison significance values for Non-differential TADs in GM12878 cell line of x32 ratio. 70

Table 3.4: Pair-wise comparison significance values for Non-differential TADs in GM12878 cell line of x50 ratio. 74

Table 3.5: Pair-wise comparison significance values for Non-differential TADs in mESCs of x16 ratio. 83

Table 3.6: Genes in the High Positive Group 86

Table 3.7: Genes in the High Negative Group 87

Table 3.8: Genes in the Opposite High Group 88

Table 3.9: Genes in the Synergistic/Competitive Group 96

Table 3.10: Results of deregulated protein complex analyses, transcription factor enrichment, and differentially expressed gene analyses in different comparisons of naïve thymus-derived CD4+ (nCD4T) and CD8+ T (nCD8T) cells: AGE_HEALTHY represents the age factor tested in healthy individuals and AGE_COVID represents the age factor tested in COVID individuals. CHILD_DISEASE represents the infection factor tested in pediatric cases (COVID vs Healthy) and ADULT_DISEASE represents the infection factor tested in adults. Statistical test applied for complexes is Wilcoxon rank sum test (unpaired, non-parametric) and False Discovery Rate threshold is 0.05. 105

Table 3.11: Results of deregulated protein complex analyses and transcription factor enrichment in various comparisons of naïve CD4+ (nCD4T) and CD8+ (nCD8T) cells. Comparisons include: CH vs YA (children vs young adults), YA vs OA (young adults vs old adults), and CH vs OA (children vs old adults). Statistical testing for complexes was performed using the Wilcoxon rank-sum test (unpaired, non-parametric), with a False Discovery Rate (FDR) threshold of 0.05. 123

.

1. INTRODUCTION

Non-coding part of the genome holds many regulatory elements which play diverse roles in transcriptional, post-transcriptional and other processing stages of gene expression [1, 2]. Annotation of the non-coding genome holds a great significance in understanding the relationship between sequence variations in these elements and the associated phenotype [3]. The annotation mainly consists of the identification of these regions and their functional role in cellular processes. One major subcategory of non-coding regions is transcriptional regulators which includes promoters and enhancers.

Promoters are the upstream regulatory regions of associated genes; they mediate the initiation of the transcription. Different cut-offs exist to identify the extend of promoter sequence, but core promoter is defined as the immediate 100 bp long region covering the transcription start site (TSS) [4]. Enhancers are defined as several hundred bases long cis-regulatory (intra-chromosomal) regions that control the gene transcription [5]. Since this is a function-based definition rather than a chemistry-based definition, discovery and identification of enhancers depend on experimental studies measuring gene transcription rather than sequence-based analyses. Condition/tissue specific nature of enhancer activity and lack of a standard motif in their sequences (as in Open Reading Frames, ORFs) prevent creation of comprehensive lists without scanning through all conditions/cell types. Thus, first challenge in the annotation process, which is the identification of enhancer sequences, has been addressed with individual functional studies along with systematic ones where the omics data coming from a broad spectrum of cell lines/primary cell cultures is integrated. ENCODE and Roadmap Epigenomics projects present candidate enhancer regions which are shown to cover 7.9% of mappable genome [6, 7]. Thanks to the improvements mainly in sequencing technology, bioinformatic analyses helped to identify several molecular marks to associate with active enhancers; such as p300 binding, histone modifications, enhancer RNAs, genome accessibility, transcription factor (TF) binding etc. [8]

Enhancers, also known as distal regulatory elements, collaborate with core promoters of their target genes regardless of their own position, distance (ranging from a kilobase to tens of megabases), or orientation. In a study on enhancer-promoter interactions in

human cell lines for ENCODE pilot project regions, which cover 1% of the genome, only 7% of distal regulatory elements found to be targeting their closest promoters [9]. This finding necessitates an approach other than assigning the closest promoter as the target of the enhancer. The varying position and distance of enhancers within the chromosome makes the enhancer-target promoter pairing, the next step in the annotation process, very difficult. Yet, many bioinformatic tools adopting different statistical/mathematical methods such as correlation, regression, neural networks, score-based prioritization or a mix/ensemble of these and utilizing above mentioned omics data have been created for this promoter-enhancer pairing task in the recent years [10]. To give an example, correlation-based methods take advantage of the correlation between the gene expression and enhancer accessibility and/or enhancer activity (defined by histone marks or enhancer RNA expression) to distinguish true pairs. There are several models (which possibly co-exist) explaining how enhancers interact with core promoters but commonly accepted model states that distal elements come in proximity of core promoters through chromatin looping [4, 11]. Chromatin looping creates structural clusters or hubs where different regulatory regions come closer, their stability involves recruitment of various proteins such as Cohesin, CTCF, Transcription Factors etc. (Figure 1.1)

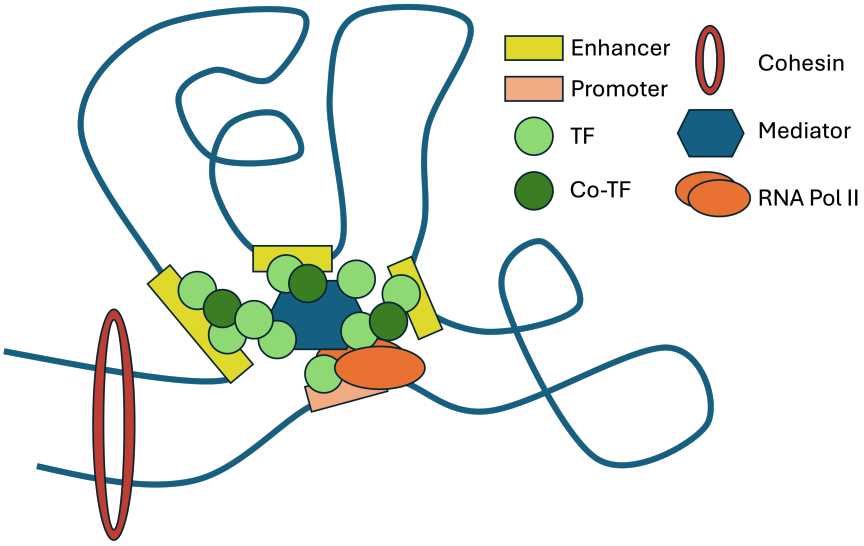


Figure 1.1: Regulatory Loop and its Elements: Enhancer, Promoter, Transcription Factor(TF), Co-factor (Co-TF), Cohesin, Mediator Complex, RNA Polymerase II.

Two major types of techniques provide experimental evidence and data for the looping model or proximity profile of the genomic regions: fluorescence in situ hybridization (FISH) and chromosome conformation capture methods (3C, 4C, 5C, Hi-C, etc.) [11]. FISH is a visualization technique relying on fluorescent probes recognizing complimentary sequences. However, it cannot be used to detect the genome wide pairwise interactions, it is limited to certain genomic loci. Capture methods rely on formaldehyde crosslinking and proximity ligation (Figure 1.2)[12]. The output data is typically presented as a contact matrix where interaction frequencies between two genomic regions over a population of cells are given (Figure 1.3). There are different adaptations of the technique. 3C, the early version, is described as "one-to-one" where specific pairwise chromosomal interactions are analyzed such as the one between a specific enhancer and promoter. In 4C, aka "one-to-many", a specific locus is chosen and all interactions from that "viewpoint" are studied. 5C is "many-to-many" where all pairwise interactions are investigated without choosing a specific "viewpoint" but within a limited region of interest. Finally, with Hi-C, "all-to-all", obtaining genome-wide high-throughput interaction profile is possible. Through utilizing Hi-C data, unbiased or hypothesis free analyses can be made about the genomic interactions [13].

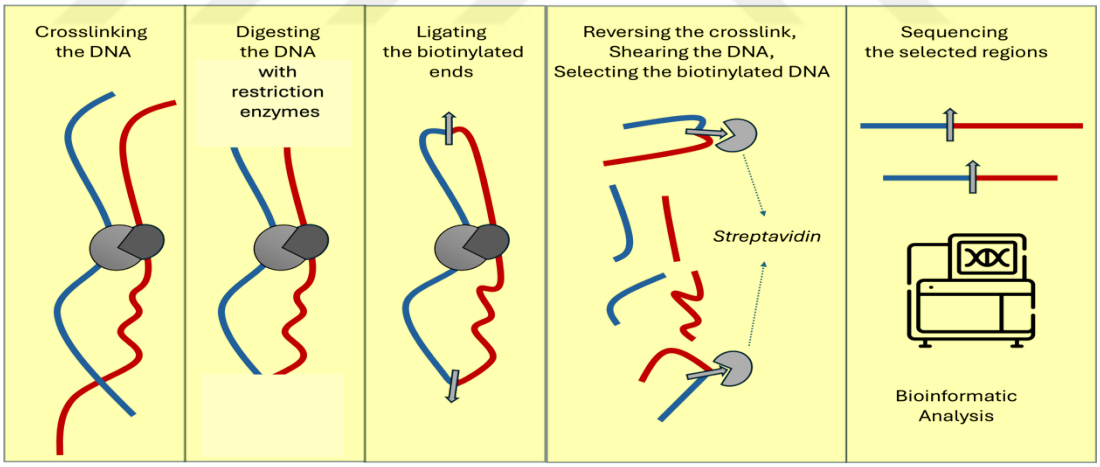


Figure 1.2: Proximity Ligation Based Chromatin Capture Method.

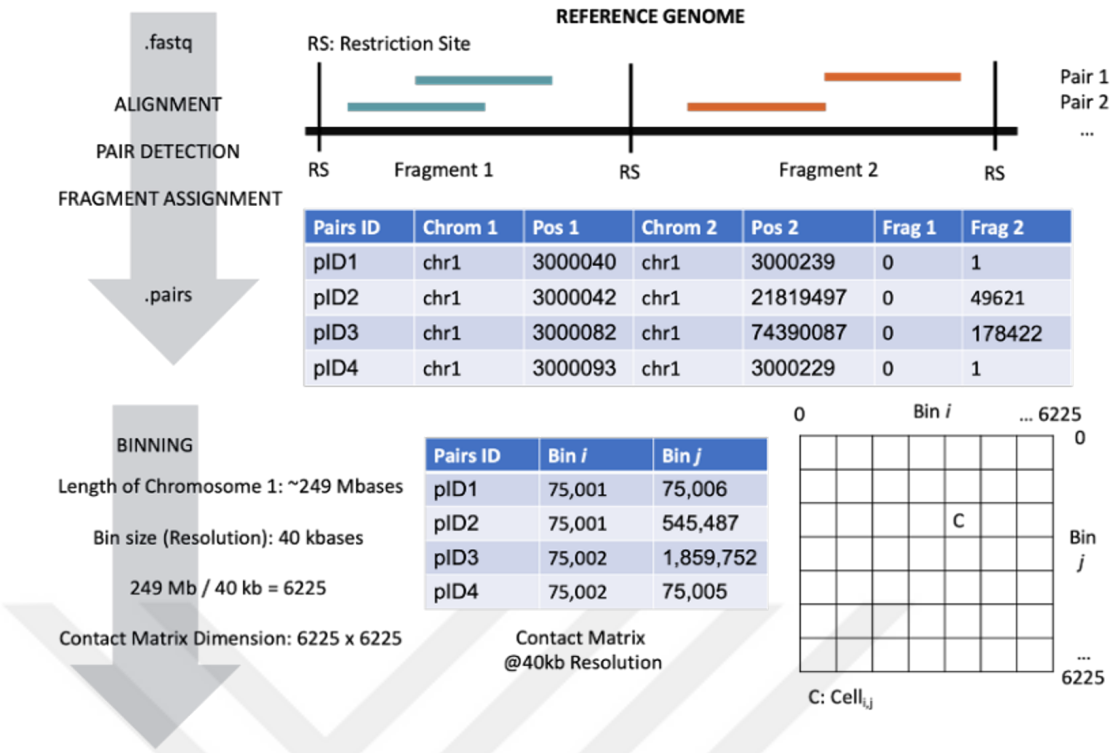


Figure 1.3: Hi-C Pipeline From Raw Reads to Contact Matrices.

Analysis of Hi-C data provides structural patterns within the genome (Figure 1.4)[14] such as alternating A/B compartments, topologically associated domains (TADs) and biologically significant chromatin interactions [15]. However, in order to study genome organization at a finer loop level, the resolution should be higher; the size of the bins in the contact matrix should be smaller such that within each bin, ideally not more than one regulatory element (promoter/enhancer) should reside. Typical Hi-C data coming from millions of cells and several hundred million of read pairs, produce contact matrices of 25-40 kb resolution. If the enhancer is assumed to be at a length of maximum 1000 bp, then this resolution of 25-40 kb gives a blurry interaction profile. If the bin size is decreased to a kilobase level, then the matrix becomes extremely sparse, meaning that each cell has a zero or one out of millions as interaction frequency value. The sparsity of contact matrix reduces the reliability and the reproducibility of comparisons. Experimental methods have been proposed to increase the resolution; however, without bioinformatic support, achievement of kb or sub-kb resolution is quite time intensiveand costly process [14].

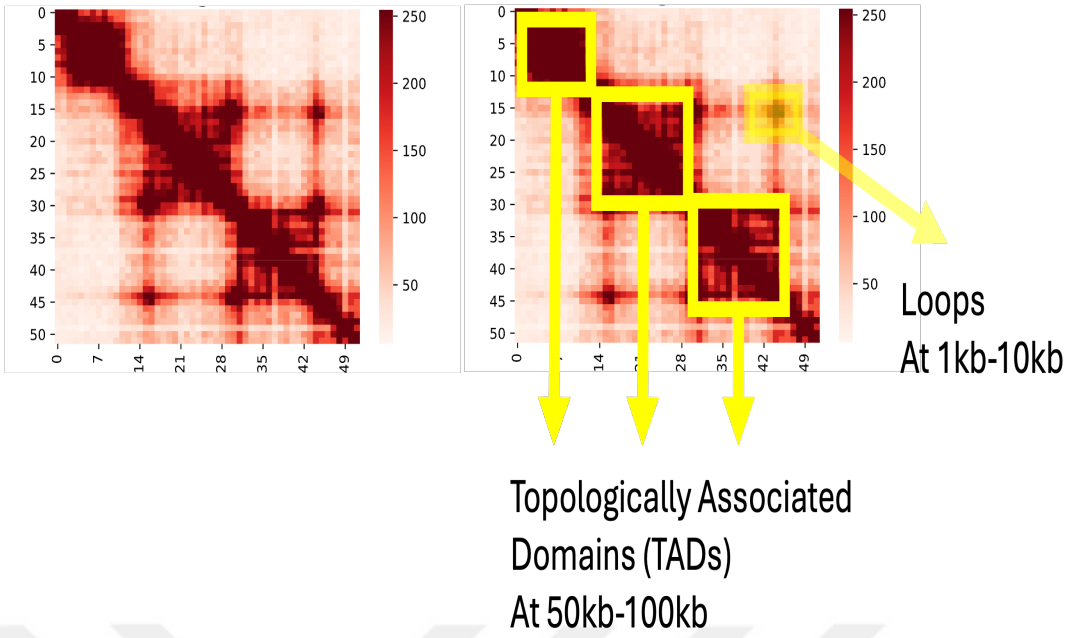


Figure 1.4: Structural Patterns Detected on Contact Matrices.

Recent years have witnessed development of tools and algorithms with regard to the low-resolution problem. Tools can be categorized into direct and indirect resolution enhancers [16, 17]. Direct methods increased the resolution from low-resolution Hi-C data using probabilistic and/or deep learning models. Indirect methods predicted the true interactions between regulatory regions (be it an enhancer-promoter, promoter-promoter, or enhancer-enhancer interaction) by taking advantage of other omics data types than Hi-C and computational algorithms, hence indirectly increased the resolution. In these indirect methods, the low resolution Hi-C data might be taken along with other data types and the binary interactions between selected regulatory elements (p-e, p-p, or e-e) are given as output instead of high resolution Hi-C data. Systematic comparisons within/between direct and indirect methods are necessary in terms of creating standard pipelines.

Apart from condition-specific activity and varying distance from their target, multi-targeting enhancers and redundant/cooperative enhancers complicate the functional annotation [5]. It is multiplicity in both directions: single enhancer to multiple targets and

multiple enhancers to single target. Due to the technical limitations and priorities, these problems were addressed only partially. For example, one-to-one physical interactions between regulatory regions cannot be studied genome-wide due to low-resolution of Hi-C data. Or many enhancers targeting single promoter have been studied through regression models, but they do not include the enhancer-enhancer or promoter-promoter interaction explicitly. However, with the above-mentioned resolution enhancement techniques and network analyses available now, it is possible to represent regulatory interactions in the nucleus as networks and study multiplicity computationally to guide experimental studies.

The functional and physical interactions within the regulatory loops involve many proteins which communicate through protein-protein or protein-DNA/RNA binding. When these bindings occur simultaneously to initiate/enhance the basic transcriptional machinery, the resulting structure can be defined as a protein complex. Thus, another layer of analysis involving regulatory protein complexes would help the understanding of enhancer-dependent transcription regulation. Protein-protein interaction (PPI) networks provide a framework for the identification of protein complexes bridging regulatory regions. Detecting dense/isolated subnetworks based on topological metrics within the PPI network is the baseline method for protein complex predictions regardless of their functional role [18]. Integration of biological data such as transcriptome/proteome, functional/component annotations of proteins, protein domain information, domain-domain interaction network etc. to PPI networks helps the accuracy in prediction algorithms [19].

Identification of cell-type-specific enhancers helped annotate disease-associated SNPs found with GWAS. It is shown that SNPs of psychiatric diseases are enriched in neuron-specific enhancers and Alzheimer's Disease-associated SNPs are enriched in microglia-specific enhancers [20]. Also, in cancer, analysis of large lineage-specific enhancer regions helped to uncover heterogeneity in diagnosis and therapy response [21, 22]. In parallel to that, it can be hypothesized that clarification of enhancer multiplicity by its DNA sequences (promoters, enhancers) and communicators (transcription factors) will help the clinical annotation process for many genetic variants. Coupled with epigenomic data such as histone modifications, enhancer RNAs and genome accessibility; distinguishing power of cooperative and competitive enhancers in clinical conditions

should be investigated. Graph theory and network representation can provide the appropriate framework for this investigation. Regulatory clusters/loops can be reconstructed through identification of functional regulatory partners and through connecting these cluster elements by proteins complexes.

Within the scope of this study, transcriptional loops were investigated at different levels to better understand the two-way multiplicity involving enhancers and promoters (Figure 1.5). Current network analysis methods were utilized, and omics data, primarily but not exclusively derived from human samples, were integrated. Initially, the low-resolution problem of Hi-C data was addressed by benchmarking current approaches to enhance Hi-C resolution. Various methods were tested in different scenarios, including same-cell, cross-cell, and cross-organism applications, as well as sensitivity to initial sequencing depth. However, the results indicated that while these methods could be helpful for enhancement and TAD-calling, they should be applied with caution. Specifically, when defining the existence of interregional contacts between regulatory regions of less than 1 kb, the use of such models tended to introduce more noise than reliable information. Consequently, the analyses focused on available high-resolution Hi-C datasets, which provided the necessary resolution for downstream integration. The network analysis aims to categorize and characterize enhancer mode of action during cellular differentiation.

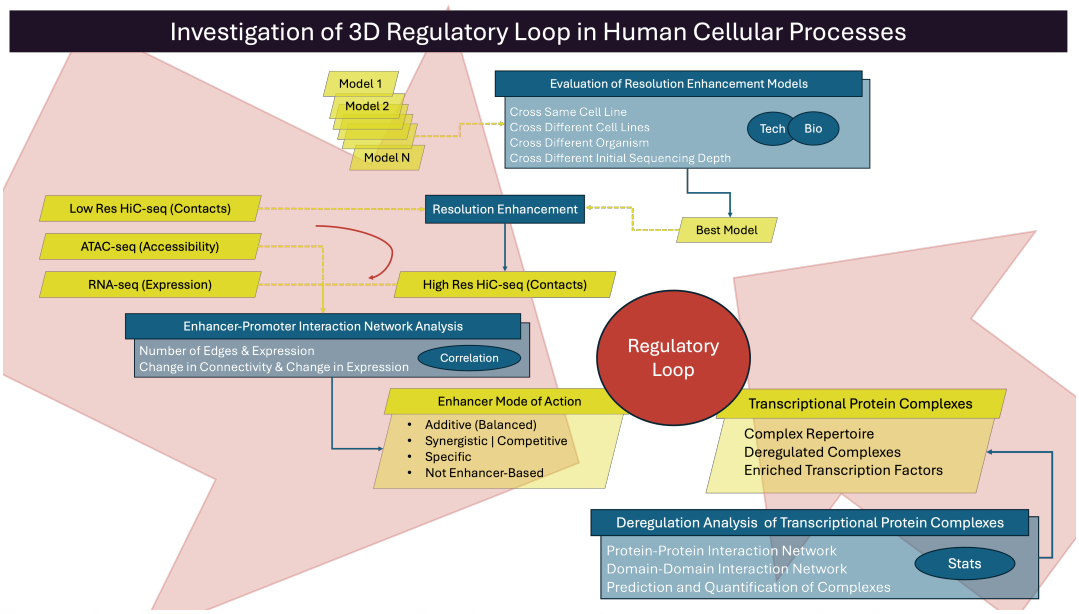


Figure 1.5: Methodology to Investigate the 3D Regulatory Loop.

By representing chromatin interactions in a network framework rather than a pairwise format, multi-way interactions and their correlation to gene expression changes can be effectively analyzed. In this context, promoter-enhancer networks during cardiomyocyte differentiation from human embryonic stem cells were successfully investigated, despite limited data availability. Regulatory protein complexes, which function as bridges between interacting chromatin regions, were also analyzed. However, due to sample size constraints, the quantitative analysis of protein complexes was conducted in a different biological system—aging immune T cells. This alternative analysis revealed promising results, showing that transcription factors enriched in altered complexes converge in regions similar to those identified by ATAC-seq studies. Finally, identifying the enhancer mode of action and the transcription factors driving regulation remains central to understanding the complex properties of pathological phenotypes. The insights from this research offer a valuable computational methodology for studying enhancer multiplicity based on 3D genome architecture, demonstrating its utility in diverse biological contexts. Having outlined the scope and methodology of this study, it is essential to contextualize it within existing research.

1.1. Literature Review and Related Work

The thesis project addresses three major tasks, each contributing to the clarification of enhancer and promoter multiplicity in regulatory regions: Enhancement of Hi-C resolution, Network Representation and Analysis of Chromatin Interaction Data During Cardiomyocyte Differentiation, and Network-Based Prediction of Regulatory Protein Complexes in Aging Immune T Cells. These tasks are framed within the broader goal of understanding transcriptional regulation and enhancer multiplicity, despite the challenges and deviations encountered during the study. A summary of the relevant publications and existing methodologies for each task is presented below, contextualizing this work within the current state of the field and identifying the gaps it seeks to address.

1.1.1. Enhancement of Hi-C Resolution

As ample amount of data is being generated in molecular context, researchers adopt big data-dependent neural network models which show great performance in computer vision, natural-language processing tasks. By drawing parallels to these tasks, bioinformaticians took advantage of neural network model architectures in their data-dependent models (Table 1.1).

Convolutional neural networks (CNNs) have demonstrated strong performance in single-image super-resolution tasks in computer vision. Drawing an analogy to the Hi-C resolution enhancement problem, Zhang et al. [17] proposed one of the first applications of CNNs in this domain: HiCPlus. This method employs a 3-layer CNN architecture to enhance the resolution of Hi-C contact matrices to 10 kb. HiCPlus optimizes based on the mean squared error (MSE) between predicted and true high-resolution matrices, and uses Pearson correlation as an evaluation metric. Its simplicity and effectiveness set a foundation for future studies to refine resolution enhancement techniques in chromatin interaction data.

Building on HiCPlus, T. Liu and Wang [23] introduced HiCNN, a significantly deeper CNN model with 54 layers. HiCNN's performance was evaluated across multiple scenarios, including training and testing on data from the same cell type, different cell types, and even different organisms. The model optimized the MSE loss to achieve

Table 1.1: Summary of Deep Learning Models for Hi-C Matrix Super-Resolution, Including Architectures, Key Features, Loss Functions, and Trainable Weights

Model	Architecture	Key Features	Loss Functions	Trainable Weights ($\times 10^3$)
HiCPlus	Three-layer CNN with 9×9, 1×1, 5×5 convolutions	Pixel-wise accuracy, Implicit residual learning	MSE	32
HiCNN	CNN with 54 layers for hierarchical refinement	Residual learning, Hierarchical feature refinement, Cross-cell type generalization	MSE	210
HiCNN2-1	HiCNN with feature fusion and weighted output layers	Feature fusion, Adaptive weighting for multiple outputs	MSE	450
HiCNN2-2	HiCNN combined with VDSR for enhanced resolution	Combines shallow and deep networks for hierarchical learning, Reproducibility metrics included	MSE, Adversarial Loss	550
HiCNN2-3	HiCNN2-2 with an additional HiCPlus-style network	Triple-network fusion, Dynamically learned weights for outputs	MSE, Adversarial Loss	700
DeepHiC	GAN with residual blocks in the Generator	Adversarial learning, Structural fidelity in low-resolution inputs (1% sampled data)	MSE, Adversarial Loss, SSIM, Perceptual Loss (PL), Total Variation Loss (TV)	2100
SRHiC	Multi-scale CNN with residual blocks for feature extraction	Multi-scale feature extraction, Preserves spatial resolution	MSE	350
HiCARN-1	Generative residual network with multi-scale feature learning	Residual learning, Lightweight architecture for efficiency	MSE, Perceptual Loss (PL), SSIM	1200
HiCARN-2	HiCARN-1 extended with adversarial training framework	Adversarial learning, Multi-objective optimization for fidelity	MSE, Adversarial Loss, SSIM	2500
HiCSR	GAN with DAE; 15 residual blocks in the Generator	Residual learning, Denoising Autoencoder for noise reduction, Focus on biological reproducibility	MSE, Adversarial Loss, SSIM	3100
hicGAN	GAN with 5 residual blocks in the generator	Adversarial optimization, Avoids oversmoothing with PSNR and SSIM metrics	MSE, Adversarial Loss, PSNR	1200
EnHiC	Multi-scale architecture with rank-1 matrix estimation and feature fusion	Rank-1 matrix decomposition, Multi-scale learning, Chromatin loop detection	MSE, Adversarial Loss, SSIM	3200
DFHiC	Dilated convolution layers for high-level feature extraction	Dilated convolutions, Feature map enhancement, TAD boundary accuracy	MAE	2400
VEHiCLE	GAN with 15 residual blocks, VAE module, and TAD regularization	Combines GAN and VAE, Multi-objective learning, TAD preservation via insulation loss, High interpretability	MSE, Adversarial Loss, VAE Loss, Insulation Loss, TAD Loss	9600

better visual fidelity, with Pearson and Spearman correlations used to evaluate accuracy. Furthermore, HiCNN increased the number of statistically significant chromatin interactions detected by the Fit-Hi-C algorithm, highlighting its utility in biological downstream analyses. Later in 2019, T. Liu and Wang [24] advanced their previous model with HiCNN2, which introduced three model architectures combining convolutional networks and generative adversarial networks (GANs). HiCNN2 optimized for multiple losses, including MSE and adversarial loss, while also evaluating reproducibility using HiC-Spector [25]. Among its three variants, HiCNN2-2 achieved the best results, particularly in reproducibility, outperforming both HiCPlus and HiCNN.

SRHiC (Super-Resolution Hi-C)[26] further refined the use of CNNs in resolution enhancement by incorporating ResNet-inspired architectures. SRHiC relied on MSE loss for optimization but introduced smaller convolutional kernels and residual blocks to improve computational efficiency and accuracy. The model outperformed HiCPlus and HiCNN in metrics such as SSIM and structural reproducibility, emphasizing the importance of architectural efficiency in large-scale genomic datasets.

hicGAN, proposed by Q. Liu, Lv, and Jiang [27], leveraged GANs to address the limitations of pixel-wise MSE, which often results in over-smoothing. In addition to MSE, hicGAN optimized for peak signal-to-noise ratio (PSNR) and structural similarity index (SSIM) to enhance the visual and structural fidelity of the predicted contact matrices. While it demonstrated significant improvements over HiCPlus and Gaussian smoothing, hicGAN did not yet incorporate reproducibility metrics, leaving this as a target for future enhancements.

DeepHiC, proposed by Hong et al. [28], further advanced GAN-based approaches by incorporating a hybrid loss function combining MSE, perceptual loss (PL), total variation loss (TV), adversarial loss, and SSIM. DeepHiC demonstrated robustness even with extremely low-resolution input data (1% down-sampled), enabling reliable identification of chromatin features such as topologically associated domains (TADs) and significant chromatin interactions. The comprehensive nature of its loss function allowed it to address both visual fidelity and biological consistency effectively.

HiCSR (High-Resolution Chromatin Structure Reconstruction) [29] introduced a deep autoencoder to extract meaningful chromatin features. The model optimized for mean

absolute error (MAE) loss while incorporating feature loss derived from the autoencoder. HiCSR demonstrated superior biological consistency in identifying chromatin loops and TADs compared to its predecessors.

ENHiC (Enhancer Hi-C) [30] employed a novel decomposition-reconstruction block architecture combined with subpixel convolutional layers. ENHiC's loss function integrated MSE, adversarial loss, and a biologically motivated rank-1 matrix decomposition loss, which enabled the model to focus on chromatin loop accuracy and biologically relevant features.

HiCARN [31] introduced a lightweight cascading residual network architecture, optimized using a combination of MSE, total variation loss (TV), and perceptual loss (PL). This approach allowed HiCARN to balance computational efficiency with high accuracy in visual metrics such as SSIM and reproducibility scores.

DFHiC (Dilated Fully Convolutional Network)[32] addressed limitations in global context capture by introducing dilated convolutions. DFHiC optimized for mean absolute error (MAE) loss, enabling the model to recover significant chromatin interactions and accurately identify TAD boundaries. By avoiding the segmentation of Hi-C ma- trices into smaller blocks, DFHiC preserved the full genomic context, outperforming HiCARN and DeepHiC in biological accuracy and computational efficiency.

Lastly in this group, VEHiCLE (Variationally Encoded Hi-C Loss Enhancer) [33] integrated multiple biologically informed loss functions into its training process. VEHiCLE's composite loss included adversarial loss, variational loss (to capture latent features), insulation loss (for TAD detection), and MSE loss. The insulation loss, a novel contribution, explicitly trained the network to align with biologically meaningful features such as TAD boundaries. VEHiCLE demonstrated strong performance across metrics like SSIM, TAD identification accuracy, and insulation vector fidelity, making it a biologically informed and interpretable model.

The progression in loss functions reflects a shift from simple MSE optimization in early models (HiCPlus, HiCNN) to complex, biologically relevant losses in later models (DeepHiC, VEHiCLE). GAN-based methods introduced adversarial loss, enhancing the structural realism of predictions, while biologically informed losses like insulation loss in VEHiCLE marked a significant advancement in integrating domain knowledge

into training. The adoption of hybrid loss functions combining MSE, perceptual, adversarial, and biologically driven terms represents the cutting edge in balancing visual fidelity with biological relevance.

Finally, an alternative probabilistic approach instead of neural networks was proposed by Cameron, Dostie, and Blanchette: HIFI[34]. Not only it differs in its use of probabilistic estimation methods like Kernel Density Estimation and Markov Random Fields, it is also the most ambitious method with regards to the resolution enhancement level. The authors claim to increase the resolution to the theoretical limit: restriction fragment size. And that is 0.5kb to 4kb depending on the restriction enzyme used. Since any Hi-C data is not available at that level to validate, cross validation and validation with 5C and ChIA-PET data are used to evaluate this model. They compare the model to HiCPlus and HMRFBayes [35] using Spearman correlation and stratum-adjusted correlation coefficient (SCC) which are proposed as a reproducibility measure by Yang et al. [36].

Although each article mentioned above provides some sort of comparison and validation, a systematic comparison of these methods is missing. The systematic comparison for Hi-C data processing, matrix normalization and reproducibility has been done previously [37–39]. As a significant improvement in the upstream steps of Hi-C data processing pipeline, resolution enhancement methodologies should be compared comprehensively, too.

1.1.2. Network Representation of Chromatin Interaction Data

Initial use of Hi-C data was to visualize it on a heatmap and identify A/B compartments and topologically associated domains (TADs) in different cell types/conditions. Recently this interaction data has started to be represented as networks. Several network representations and analyses, mainly based on Hi-C data, were published within 2019 during thesis proposal period.

Paulsen et al. [40] in their stem cell differentiation experiment, used Hi-C data to analyze long-range TAD-TAD interactions. They focused on TAD cliques, which are fully interacting TAD groups, and observed their dynamics throughout the cell differentiation process. They studied these TAD cliques in mouse embryonic stem-cell differen-

tiation and somatic cell reprogramming using single-cell high-resolution Hi-C data.

Song, Sharan, and Ovcharenko [41] used Hi-C data from 4 human cell lines [42] to construct regulatory networks consisting of promoters and enhancers. These regulatory elements were annotated as active enhancers and promoters using diverse type of data: histone marks (H3K27ac, H3K4me1, H3K27me3) and DNase I hypersensitivity. Regulatory networks for each active promoter are generated based on the Hi-C proximity data in 4 different cell types. Regulatory regions are represented as nodes and if they are caught to be proximal in 3D to each other by Hi-C data then an edge links these nodes. Using the regulatory network, they created "enhancer chains". Chains are starting from a promoter and they walk around the network one enhancer (node) at a time. The interesting finding applicable to all cell lines is that in an enhancer chain, closest enhancer to the promoter in 3D, is farther away to the promoter in genomic distance than the second enhancer is to the promoter. Based on that finding, they could hypothesize several scenarios on the role of the distal enhancer interacting with the promoter; the major one was that it safeguards the loop and creates a cluster for other proximal enhancers to modulate the gene transcription.

Thiel et al. [43] used ChIA-PET data for K562 human cell line to create regulatory networks. They annotated sequences (to become nodes in the network) as genes, enhancers, lncRNA, CTCF binding site, and unknown regions using chromHMM software from ENCODE. Using a clustering method developed by the same group, regulatory modules involving lncRNAs and gene pairs are found. Innovative side of the study is finding overlapping (fuzzy) clusters of lncRNA and gene pairs based on co-expression. Then lncRNAs in these modules are functionally annotated.

Chen et al. [44] used Hi-C interactions in a network framework to investigate enhancer-enhancer cooperation during human cardiac commitment. Major assumption in this investigation is that if two enhancers are in contact with the same promoter then they would co-regulate the gene and the enhancers are in contact, too. Network motifs in which multiple enhancers coregulating promoters are identified and compared during the cardiac commitment. This assumption should be tested with Hi-C data sampled from every step in cellular differentiation.

These studies highlight the utility of network-based approaches for analyzing chromatin

interactions and transcriptional regulation. However, the framework developed in this thesis offers a distinct perspective by integrating multiple data modalities—ATAC-seq, Hi-C, and RNA-seq—across all stages of cardiomyocyte differentiation. By doing so, our approach enables the prediction and characterization of enhancer modes of action, such as additive, synergistic, or specific, which have not been systematically explored in previous studies. Unlike prior works focusing on static interactions or individual data points/types, this comprehensive framework provides dynamic insights into the interplay between chromatin structure and transcriptional regulation during differentiation.

Cardiomyocyte differentiation from embryonic stem cells (ESCs) is a complex process that involves multiple stages and is regulated by various transcription factors. Understanding the differentiation stages and the roles of key transcription factors, such as Nkx2.5, GATA4, and Isl1, as well as the E2F family, is crucial for advancing regenerative medicine and developing therapies for heart diseases.

The differentiation of ESCs into cardiomyocytes typically involves several stages, starting from the formation of mesodermal progenitors to the maturation of functional cardiomyocytes. During the early stages, the expression of specific transcription factors is critical for guiding the differentiation process. Nkx2.5 is one of the earliest markers of cardiac progenitor cells and is essential for the specification of these cells from the mesoderm [45, 46]. Studies have shown that Nkx2.5 not only marks cardiac progenitors but also plays a pivotal role in the activation of downstream cardiac genes necessary for cardiomyocyte development [47, 48].

GATA4 is another crucial transcription factor that influences cardiomyocyte differentiation. It has been demonstrated that GATA4 is involved in the proliferation and survival of embryonic cardiomyocytes and acts as a direct activator of genes that promote cell cycle progression, such as cyclin D2 and Cdk4 [49]. Furthermore, GATA4 works in conjunction with Nkx2.5 to enhance cardiac gene expression, indicating a synergistic relationship between these transcription factors during cardiomyocyte differentiation [50](Nakajima et al., 2011). The interplay between GATA4 and Nkx2.5 is essential for the proper development of cardiac tissues, as disruptions in their expression can lead to impaired cardiomyocyte formation [49].

Isl1 is also significant in the differentiation of cardiomyocytes, particularly in the context of cardiac progenitor cells. It is known to be expressed in early cardiac progenitors and is critical for the development of both atrial and ventricular cardiomyocytes [48, 51]. The presence of Isl1 in progenitor cells indicates its role in the early stages of cardiomyocyte differentiation, where it helps to establish the cardiac lineage [48](Lee et al., 2017). The interaction between Isl1, Nkx2.5, and GATA4 highlights a network of transcription factors that coordinate the differentiation process, ensuring the proper specification and maturation of cardiomyocytes [47, 48].

In addition to these key transcription factors, the E2F family has been implicated in the regulation of cardiomyocyte differentiation. E2F transcription factors are known to control cell cycle progression and are involved in the transition from proliferation to differentiation in various cell types, including cardiomyocytes [52, 53]. The regulation of E2F activity during cardiomyocyte differentiation is crucial, as it influences the balance between cell proliferation and differentiation, thereby impacting the overall efficiency of cardiomyocyte generation from ESCs [52, 53].

Moreover, the differentiation process is influenced by various signaling pathways and microRNAs, which further modulate the expression of these transcription factors. For instance, microRNAs such as miR-27b and miR-23b have been shown to play roles in the regulation of cardiomyocyte differentiation by targeting specific transcription factors and signaling molecules [54, 55]. This regulatory network underscores the complexity of cardiomyocyte differentiation, where multiple layers of control ensure the precise development of cardiac cells from pluripotent stem cells.

In conclusion, cardiomyocyte differentiation from embryonic stem cells is a multifaceted process governed by a network of transcription factors, including Nkx2.5, Isl1, GATA4, TBX5, and MEF2C, alongside the E2F family. The interactions among these factors, along with the influence of signaling pathways and microRNAs, highlight the intricate regulatory mechanisms that guide the formation of functional cardiomyocytes. Understanding these processes is essential for advancing therapeutic strategies aimed at cardiac regeneration and repair.

1.1.3. Network-Based Regulatory Protein Complex Prediction

Known protein complexes [56, 57], although being incomplete, when mapped to protein-protein interaction (PPI) networks, appear as dense subnetworks[58]. Hence early algorithms used this topological property along with suitable clustering algorithms to predict protein complexes regardless of their function in the cell. However, the major difference between predicting protein complexes and the discovery of modules / clusters / subnetworks is the requirement of interactions occurring at the same time and the same place [58].

Therefore, initial approaches were disadvantageous in several aspects. Firstly, PPI networks are cumulative representations hence detecting mutually exclusive interactions is not possible based on the network alone. Secondly, interaction detection technique might be overlooked in the generation of networks; hence, the impossibility of finding the interacting proteins at the same place *in vivo* might be overlooked too. Stoichiometry of protein complexes cannot be analyzed on the current PPI network structures where each protein is a unique node. Finally, density-based clustering is not efficient for small complexes, such as complexes composed of less than 4 proteins. False positive rates increase significantly as the size of complexes decreases.

Since the first tool MCL proposed in 2000 [59], many algorithms/approaches have been published for the task of protein complex prediction. There is not a single best performing tool; however, there are best performers with respect to overcoming certain drawbacks mentioned above. There are network-based or topology-based algorithms that focus on improving the definition of density (cliqueness/isolation) metric in a network and the clustering/search algorithm (agglomerative or divisive) using that metric definition. Many algorithms integrated biological or functional information with the PPI network to increase the accuracy. The information includes transcriptome, proteome, Gene Ontology annotation terms (Biological Processes, Molecular Function, Cellular Component), protein domain information, domain-domain interaction network, evolutionary conservation, etc. Finally, there are specialized methods focusing on specific tasks such as prediction of small or sparse complexes [19].

Improvements in qualitative protein complex prediction algorithms led the researchers to quantitative complex predictions and differential analyses. Currently, there are three

quantitative complex prediction algorithms which differ in their complexity: ComplexBrowser [60], CompleXChange [61] and SiComPre [62]. ComplexBrowser relies on the assumption that subunits in the protein complexes are co-expressed and co-regulated. The algorithm takes proteome data and known protein complexes list as inputs. Known complexes needs either to be fetched from CORUM [56] and EBI Complex Portal databases [57] or they need to be submitted by the user in a specific format, as this tool does not make a qualitative protein complex prediction. Only complexes with 3 quantified subunits are taken into consideration for further analysis. Complex abundances are predicted by fast-FARMS algorithm which is a modified version of Bayesian factor analysis with an assumption of Gaussian noise. In short, the complex abundance is linearly summarized as the sum of weighted subunit abundances and noise. After complex abundances are predicted, complex fold change (CFC) is calculated for each complex in pairs of condition.

Will and Helms [61] proposed a pipeline to study protein complexes qualitatively and quantitatively. CompleXChange is the final step in this pipeline. The first tool is PPIXpress [63] which takes sample transcriptome data as input to generate sample-specific PPI and DDI networks. General PPI and DDI networks along with protein domain information are retrieved automatically by the command-line tool from specific databases. JDACO [64], the second tool, uses the sample specific networks to predict protein complexes. User submits a list of seed proteins which can be completely random list or functionally interesting/important group of gene products, like transcription factors. Finally, CompleXChange takes predicted protein complexes along with the same transcriptome data to determine differentially expressed protein complexes.

SiComPre is the most sophisticated algorithm where qualitative and quantitative predictions can be made. The underlying model is stochastic simulation of protein complexation and decomplexation reactions. PPI and DDI network, proteome data (protein abundances), protein domain information and GO functional terms are given as inputs. The cell can be modeled as a single compartment or multicompartment well-mixed volume where proteins diffuse and bind/unbind. The multicompartment case requires protein localization information related to these subcellular compartments. Proteome-level simulations for human samples using a typical size PPI network require very high computational power (CPU and GPU), memory and computational time requirements

make this tool less practical compared to its alternatives.

The characterization of the role and the communication between regulatory regions is a vital step in understanding the mechanism of enhancer-dependent transcription. Major challenges in this characterization process such as experimental cost/impracticality in measuring all data types to cover all interactions in all conditions/cell-types can be overcome with computational methods and frameworks. This study proposes a methodology to investigate regulatory loops. The methodology aims to improve the already present algorithms and combine them innovatively to study regulatory clusters of genomic loci and proteins. Solidifying the interactions between genomic loci with appropriate proteins and protein complexes adds another layer to the proposed analysis. Protein complex prediction algorithms are used rarely and in isolation from other analysis methods. Here, they are implemented in a way that provides both answers to biological problems and to the integration possibilities into the chromatin loop analysis by the improvements which deals with the predicted protein complexes in a functionally/locationally specific context.

1.1.4. Aging T cells

Age-related changes in the immune system, collectively termed immunosenescence, have significant clinical implications, including increased susceptibility to infections, inadequate vaccine responses, and heightened prevalence of autoimmune diseases [65]. Among immune cells, T cells of the adaptive immune system are particularly affected by aging, undergoing changes such as thymic involution, reduced T cell output, functional decline, and decreased T-cell receptor diversity [66–68]. These alterations are cell-type specific and significantly impact subsets such as CD4+ and CD8+ T cells, leading to shifts in differentiation patterns and ratios, further compounding immune dysfunction.

Recent research has focused on the molecular and cellular basis of these phenomena using omics approaches, including gene expression analyses, chromatin accessibility profiling, and studies of chromatin architecture reorganization [69–71]. For example, differential transcriptomic and epigenomic profiling has identified changes in regulatory mechanisms and transcription factors associated with aging in T cells. Transcrip-

tion factors from the basic leucine zipper (bZIP) family, particularly the AP-1 subfamily (Fos and Jun), have been reported as key regulators of aging-associated changes in T cells [69, 72–75]. Additionally, motif enrichment analyses have shown increased accessibility to Interferon Regulatory Factors (IRFs) and Kruppel-like factors (KLFs) binding motifs in aged individuals [76].

In child-adult comparisons and across different adult age groups (e.g., young adults 20–35, adults 35–65, elderly >65), transcriptional regulation appears to be mediated by distinct mechanisms. For example, children experience COVID-19 infections less severely than adults, a difference linked to immunosenescence [77, 78] . Studies on COVID-19 patients have revealed that age-related immune changes exacerbate susceptibility to severe outcomes, particularly in adults. Notably, T cells in juvenile and adult COVID-19 patients demonstrate distinct transcriptional responses, with transcription factors such as T-bet (TBX21) and RORγT (RORC) driving differential immune activity. In juvenile cases of severe COVID-19, enhanced expression of these factors was associated with increased Th1 and Th17 responses, which contribute to elevated levels of IFN-γ and IL-17, respectively [79, 80].

Despite advances in identifying these transcriptional regulators, critical gaps remain in understanding the interplay between these factors and their co-regulators, particularly in the context of protein complexes. Most studies to date have relied on bulk RNA-seq data or motif analyses that do not capture the interactions or functional units (e.g., protein complexes) driving transcriptional regulation. Network analyses incorporating protein-protein interaction data offer a promising approach to address these limitations by examining cell- or condition-specific regulatory dynamics.

2. METHODOLOGY

2.1. Benchmarking Resolution Enhancement Models

The deep learning models used in this study were trained to enhance low-resolution Hi-C contact matrices by learning the relationship between downsampled matrices and their corresponding high-resolution counterparts. Specifically, high-resolution matrices served as the labeled output, while their downsampled versions were used as the input during training. Given the limited availability of high-resolution Hi-C experiments, each chromosome contact matrix was divided into smaller non-overlapping submatrices. This approach significantly increased the sample size, enabling the models to generalize more effectively and learn detailed patterns of chromatin interactions. A schematic of this training framework is provided in Figure 2.1.

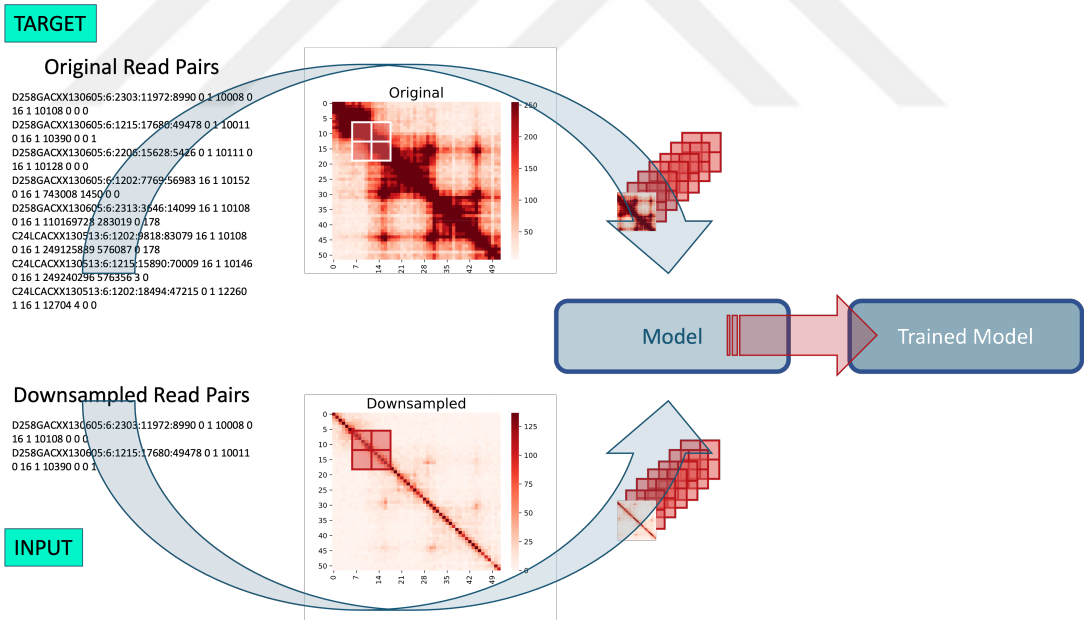


Figure 2.1: Training A Resolution Enhancement Model.

2.1.1. Comparison Scenarios

Enhancing low-resolution Hi-C contact matrices using deep learning models has become a critical advancement in genome architecture studies. In this work, ten deep

learning models (*HiCPlus*, *HiCNN*, *HiCNN21*, *HiCNN22*, *HiCNN23*, *DeepHiC*, *SRHiC*, *HiCARN1*, *HiCARN2*, and *HiCSR*) trained on the GM12878 cell line were benchmarked under four distinct scenarios to evaluate their generalizability, robustness, and accuracy in reconstructing high-resolution matrices. Each scenario was designed to highlight a specific aspect of model performance, providing a comprehensive assessment.

The first scenario evaluates the performance of the models on the GM12878 cell line at a downsampling factor of ×16, establishing a baseline for their ability to reconstruct data from the training cell line. This scenario assesses the ability of the models to reconstruct high-resolution Hi-C matrices for the GM12878 cell line, which was used for model training, at a downsampling factor of ×16. The primary objective is to establish baseline performance and evaluate the fidelity of the models when applied to familiar data.

The second scenario tests the generalizability of the models across different human cell lines, specifically IMR90, K562, and NHEK, which represent diverse chromatin landscapes, also at a downsampling factor of ×16. To evaluate the generalizability of the models, downsampled matrices from the IMR90, K562, and NHEK cell lines were enhanced using the trained models. These cell lines exhibit distinct chromatin interaction patterns, providing a robust comparison of model performance across diverse human cell types.

The third scenario explores the performance of the models across varying levels of data sparsity by analyzing GM12878 matrices downsampled at factors of ×25, ×32, and ×50, simulating different initial sequencing depths. In this scenario, GM12878 matrices downsampled at factors of ×25, ×32, and ×50 were used to examine the models' ability to handle data with progressively reduced resolution. This analysis simulates varying sequencing depths and evaluates model adaptability to different levels of data sparsity.

Finally, the fourth scenario assesses cross-species applicability by applying the models to mouse embryonic stem cell (mESC) matrices downsampled at ×16, thereby evaluating their robustness in a different genomic context. The final scenario evaluates the cross-species robustness of the models by applying them to mouse embryonic stem cell (mESC) matrices downsampled at a factor of ×16. This scenario tests the models' ca-

pability to enhance data from a different organism, highlighting potential limitations in handling species-specific genomic features.

To facilitate a localized performance analysis, each chromosome was divided into non-overlapping subregions of size 1Mb. A comprehensive set of metrics, including the Mean Squared Error (MSE), Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM), Pearson correlation coefficient, and biologically relevant TAD-based measures, was employed to ensure a multi-faceted evaluation of the models.

2.1.2. Metrics for Model Evaluation

Mean Squared Error (MSE)

The *Mean Squared Error (MSE)* quantified pixel-wise differences between the enhanced and high-resolution matrices:

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(x_i - y_i)^2$$

where x_i and y_i are pixel intensities in the enhanced and high-resolution matrices, respectively, and N is the total number of pixels. Lower MSE values indicate better enhancement quality.

Subregion-Based Comparison: MSE was calculated independently for each 1Mb subregion and averaged across all subregions within a chromosome.

Peak Signal-to-Noise Ratio (PSNR)

The *PSNR* measured fidelity in decibels (dB), inversely proportional to MSE:

$$PSNR = 10 \cdot log_{10}\left(\frac{MAX^2}{MSE}\right)$$

where MAX is the maximum possible pixel value. Higher PSNR values indicate better enhancement quality.

Subregion-Based Comparison: PSNR was computed for each 1Mb subregion and averaged across chromosomes.

Structural Similarity Index (SSIM)

The *SSIM* evaluated structural similarity between the enhanced and high-resolution

matrices by comparing luminance, contrast, and structure:

$$SSIM_{(x,y)} = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}$$

where μ, σ, and σ_{xy} are the mean, variance, and covariance of pixel intensities in x and y, and C_1, C_2 are constants. SSIM values range from 0 to 1, with higher values indicating greater structural similarity.

Subregion-Based Comparison: SSIM was calculated for each 1Mb subregion and averaged across chromosomes.

Pearson Correlation Coefficient

The *Pearson correlation coefficient (r)* was used to assess the preservation of long-range interaction decay patterns between enhanced and high-resolution matrices. Correlation was calculated for diagonals spanning distances up to 1Mb:

$$r_{(d)} = \frac{Cov(x_d, y_d)}{\sigma_{x_d}\sigma_{y_d}}$$

where x_d and y_d are pixel intensities at diagonal distance d, and Cov, σ_{x_d}, σ_{y_d}represent covariance and standard deviations.

Distance-Based Analysis: In order to calculate Pearson Correlation Coefficients between two chromosome contact matrices M1 and M2, the contact matrix is divided into distance arrays (Figure 2.2). Let M's be the n-by-n symmetrical chromosome contact matrices. The main diagonals (d0) represent the interactions between two loci where the distance between them is zero. If the diagonal array one bin right to the main diagonal array is taken (d1), that array shows the interactions between two loci where the distance between them is the bin size, in this case 10kb. Distance arrays are generated from compared matrices till the distance value of 1Mb is reached. Then Pearson Correlation Coefficient between associated d arrays are calculated. Correlation values were computed for all distances up to 1Mb and averaged to evaluate each model's abil- ity to preserve interaction patterns. Distance-specific trends were analyzed to identify performance variations.

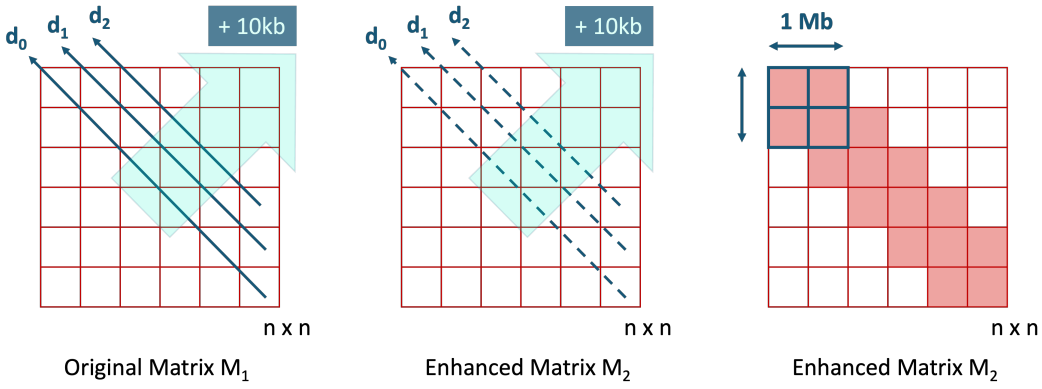


Figure 2.2: Pearson Correlation Coefficient calculated at different distances. MSE, PSNR and SSIM metrics compared within 1Mb submatrices.

Non-Differential TAD Proportion

Technical evaluation metrics help to obtain an opinion on mathematical performance of the methods; however, a biologically relevant metric should also be used. Hi-C contact matrices are used to determine A/B compartments which represent the actively transcribed and passive regions. Topologically associated domains (TADs) are another type of patterns called from a Hi-C contact matrix. TADs are the regions where intra-regional contacts are enhanced relative to inter-regional contacts. Finally, significant interactions, which are interactions probabilistically marginal compared to expected contact frequencies, can be detected from Hi-C contacts. TAD callers are more mature than the chromatin interaction callers. Hence, using TADs as a biological evaluation metric is reasonable.

Currently, there are tens of TAD callers among which Arrowhead is shown to be the most robust and biologically significant TAD finder [81]. However, number of TADs called with Arrowhead is limited compared to other tools. If Arrowhead is used to compare the model predictions, the difference might not be distinct enough. Instead, TADCompare tool is chosen which uses spectralTAD to call TADs and compares them quantitatively [82]. TADCompare tool compares two matrices and categorizes the TADs as Differential and Non-Differential. Further, the differential TADs are categorized into 5 classes (Figure 2.3): Complex, Merge, Split, Shifted, Strength Change. Complex differential TADs represent the complex class as the name implies and there is not a

specific pattern to represent the change in the structure of differential TADs. Merge class represents the TADs that lost a TAD boundary between two boundaries. Split is the reverse where a boundary appears between two TAD boundaries. Shifted represents a boundary shift and Strength Change represents the change in score that the TAD is called with. Here, we use the proportions of differential and non-differential TADs as biologically relevant evaluation metrics to assess the model predictions.

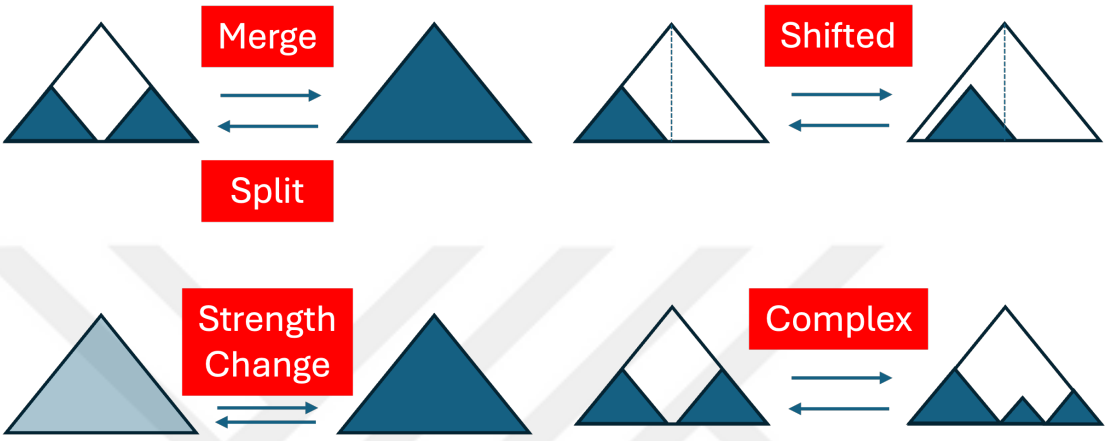


Figure 2.3: TADCompare defined differential TAD categories: Merge, Split, Shifted, Strength Change, Complex.

Proportion Calculation:

$$\text{Differential TAD Proportion} = \frac{\text{Number of Differential TADs}}{\text{Total Number of TADs in High-Resolution}}$$

$$\text{Non-Differential TAD Proportion} = \frac{\text{Number of Non-Differential TADs}}{\text{Total Number of TADs in High-Resolution}}$$

Statistical Analysis:

- *Pairwise Wilcoxon Rank-Sum Tests:* For each cell line and enhancement ratio, pairwise Wilcoxon rank-sum tests were conducted to compare non-differential TAD proportions across all models.
- *Significance Threshold:* A $p < 0.05$ threshold was used to determine significant differences.

- *Output:* Pairwise p-values were stored in symmetric matrices and summarized in comparison tables for interpretation.

Visualization:

- *Boxplots:* Distributions of differential and non-differential TAD proportions were visualized for each model, cell line, and enhancement ratio.

2.2. Promoter-Enhancer Network Analysis in Cardiomyocyte Differentiation

2.2.1. Data Retrieval and Data Processing

The raw sequencing reads were obtained from the Sequence Read Archive (SRA) under the accession number PRJNA417681 [83]. All analyses were conducted using the human reference genome (GRCh38) to ensure accurate alignment and downstream processing.

The RNA-seq data was processed using the pipeline described in the publication. Reads were aligned to the reference genome using STAR (version 2.5.2b) with default parameters. The resulting data was quantified and processed using the Cufflinks suite (version 2.2.1), including the generation of CXB files with cuffquant, FPKM and count tables with cuffnorm, and differential expression analysis with cuffdiff.

The ATAC-seq data was processed using a standardized pipeline (ENCODE Pipeline) that included alignment, filtering, and peak calling. Reads were aligned to the reference genome using Bowtie2, producing a tagAlign file, which contains read positions in gzipped BED format. A quality control report was generated. Peaks were identified using MACS2, which generated two key outputs: a signal fold change track in BigWig format and two peak sets, designated as optimal peaks and conservative peaks, both in BigBed format. The peak calling process utilized an overlap method to enhance reliability. A quality control report was linked to the BigWig output to evaluate the accuracy and robustness of the peak calling.

The Hi-C data was processed using a standardized pipeline (4DN Pipeline) that included alignment, filtering, and matrix aggregation and normalization steps. Hi-C reads were aligned to the GRCh38 reference genome using BWA (v0.7.17). To identify valid

Hi-C contacts, Pairtools (v0.2.2) was used for filtering. Unmapped portions shorter than 20 bp were ignored. The filtering workflow began with BAM files being converted to pairsam format using pairtools parse, which classified alignments and added necessary metadata. Next, the pairsam files were sorted by genomic coordinates (chr1-chr2-pos1-pos2) using pairtools sort. Multiple pairsam files were then combined in a sorted order using pairtools merge. Duplicates with the same 5' coordinates (±3 bp) were marked using pairtools dedup. Finally, valid contacts, classified as UU and UC, were retained with pairtools select. Hi-C contact matrices were generated in the **.hic** format using Juicer Tools (v1.8.9-cuda8) and normalized using KR (Knight-Ruiz) methods. Contact Matrices at 10kb were used for further analysis.

2.2.2. Generation of Promoter-Based Networks

To investigate promoter-based regulatory networks throughout cell-differentiation process, two alternative routes were designed. The first route (Figure 2.4) relied on Hi-C interaction profiles as the primary source, retaining unknown genomic regions within the network. This approach aimed to facilitate the discovery of candidate regulatory regions for future experimental validation. However, the first route was abandoned due to the insufficient resolution enhancement of deep learning models analyzed in the previous chapter. Instead, the second route (Figure 2.5) was prioritized, focusing on known regulatory regions and addressing the multiplicity problem.

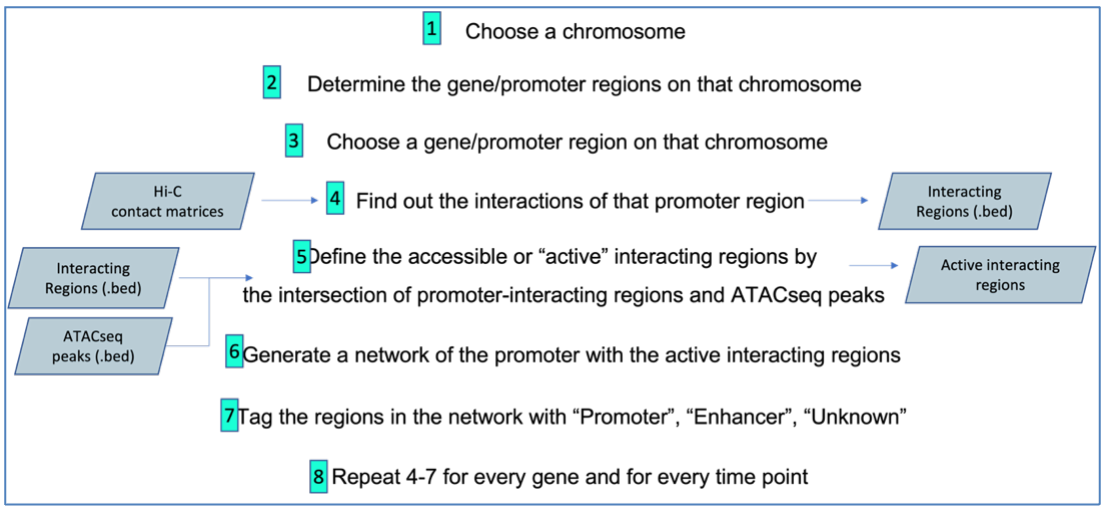


Figure 2.4: Route 1 to Investigate Promoter-Interaction Networks.

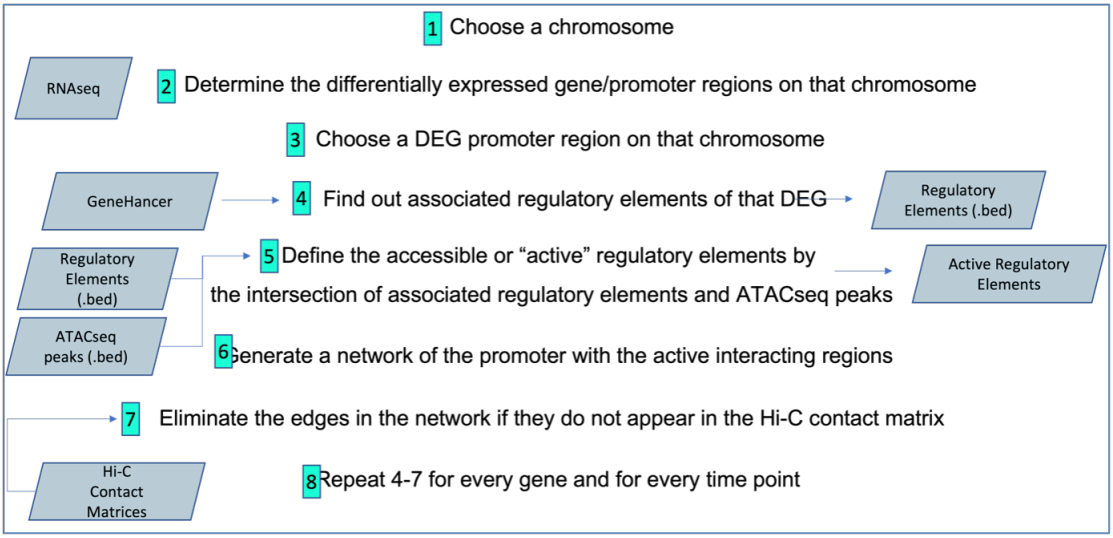


Figure 2.5: Route 2 to Investigate Promoter-Interaction Networks.

The second route is described below, focusing on a single chosen chromosome (e.g., Chr11) and leveraging differentially expressed genes (DEGs) identified across all time points (e.g., T0 vs. T1, T1 vs. T2, T0 vs. T3). The following steps were applied for each DEG on Chr11:

1. **Extraction of Regulatory Elements:** All regulatory elements ("DOUBLE ELITE" promoters and enhancers) located on Chr11 were downloaded (July 3rd,2022) from the GeneHancer database[84] using the UCSC Table Browser[85]. In the GeneHancer database, regulatory elements are classified into categories that re- flect the confidence level of their associations with target genes. GeneHancer identifications were created from more than one million regulatory elements ob- tained from seven genome-wide databases, including ENCODE project Z-Lab Enhancer-like regions (version v3), Ensembl regulatory build (version 92), FAN- TOM5 atlas of active enhancers, VISTA Enhancer Browser, dbSUPER super- enhancers, EPDnew promoters, and UCNEbase ultra-conserved noncoding ele- ments. Through an integration algorithm, approximately 250,000 candidate reg- ulatory elements (referred as GeneHancers) were obtained, where redundant ele- ments removed. Each GeneHancer is assigned a confidence score, with those de- rived from more than one information source classified as "elite" GeneHancers.

Additionally, gene-GeneHancer associations were generated using various sources of evidence linking regulatory elements to genes, such as expression quantitative trait loci (eQTLs) from GTEx (version v6p), Capture Hi-C promoter-enhancer long-range interactions, FANTOM5 eRNA-gene expression correlations, cross- tissue expression correlations between transcription factors interacting with a GeneHancer and candidate

target genes, and distance-based associations. Distance- based associations include nearest-neighbor relationships, overlaps with gene territories (intragenic), and proximity to transcription start sites (TSS) within 2 kb. Associations derived from more than one information source are defined as “elite” associations. Consequently, the “double elite” dataset refers to elite gene associations of elite GeneHancers, representing the highest confidence level in the GeneHancer database. By using double elite regulatory elements (Figure 2.6 and Table 2.1) and their interactions, the reliablity of the method increased.

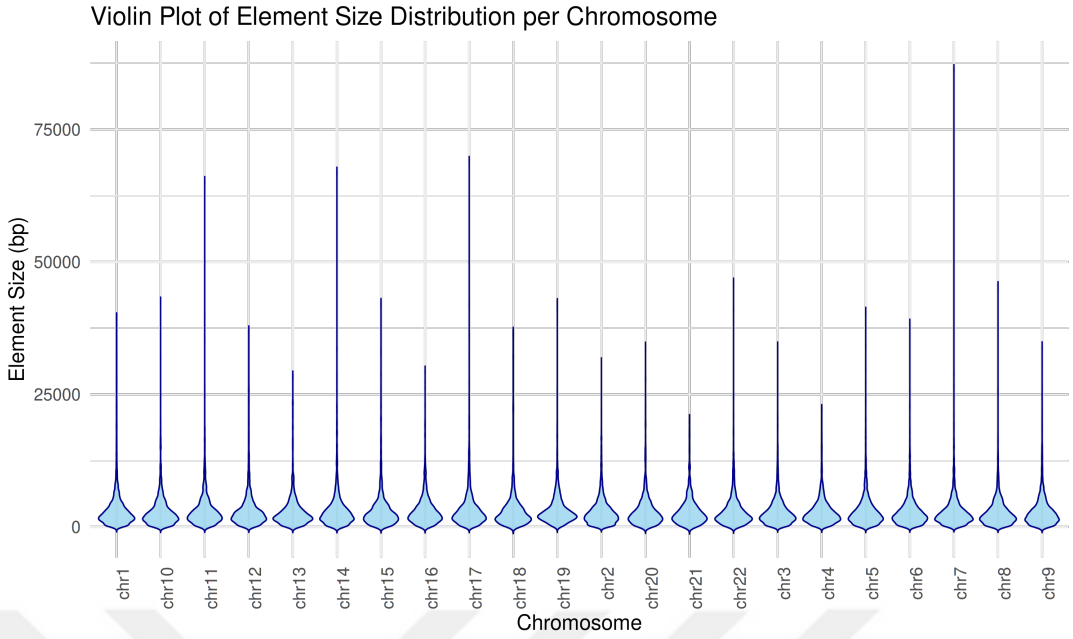


Figure 2.6: Size distribution of elements across chromosomes.

2. **Association of Regulatory Elements with DEGs:** Regulatory elements were mapped to their associated DEGs; e.g. 9 elements for ZNF143 (Figure 2.7). The promoter of each DEG was identified, leading to three possible scenarios:

- A single promoter associated with the DEG: This was designated as the primary promoter.
- Multiple promoters associated with the DEG: The primary promoter was defined as the one located within 500 bases upstream of the gene.
- No promoter associated with the DEG: These cases were marked as “miss- ing promoter.”

3. **Filtering Small Networks:** DEGs with fewer than two associated enhancers were excluded from the analysis, as their networks were too small to provide meaningful insights.

Table 2.1: Size distribution of elements across chromosomes. Includes mean, median, minimum, maximum, and standard deviation of element sizes, along with the number of promoters(P), enhancers(E), and promoter/enhancer(P/E) combinations.

Chr	Mean Size	Median Size	Min Size	Max Size	Standard Deviation	# of P	# of E	# of P/E
chr1	3061.4	2298.0	4	39411.0	2954.6	7	3877	1374
chr2	2916.7	2238.5	10	30916.0	2798.0	5	3308	867
chr3	2918.6	2211.0	6	33864.0	2789.6	3	2647	754
chr4	2780.1	2199.0	1	22034.0	2346.4	2	1752	509
chr5	3021.5	2250.0	13	40330.0	3034.3	3	1970	601
chr6	2887.3	2198.5	11	38148.0	2688.4	1	2415	722
chr7	2837.5	2142.0	11	86253.0	3057.0	1	2080	650
chr8	2726.5	2089.5	9	45199.0	2579.2	1	1914	471
chr9	2930.3	2234.0	2	33803.0	2902.2	-	1926	553
chr10	2931.0	2169.0	2	42275.0	3025.3	3	2182	518
chr11	3145.1	2333.0	5	65038.0	3245.5	2	2174	776
chr12	3089.9	2298.0	25	36844.0	3152.1	1	1940	688
chr13	2753.3	2050.0	16	28284.0	2726.2	-	1024	227
chr14	3108.7	2328.5	10	66724.0	3491.5	2	1359	427
chr15	2988.7	2205.0	9	41917.0	3156.5	1	1464	434
chr16	3093.2	2396.5	10	29169.0	2832.6	-	1394	582
chr17	3265.9	2371.0	6	68774.0	3496.6	2	1914	789
chr18	2875.3	2162.0	8	36371.0	2875.1	-	825	204
chr19	3261.6	2426.0	29	41981.0	3019.5	1	1157	1021
chr20	3027.0	2240.5	6	33597.0	3189.9	1	1164	359
chr21	2857.1	2192.0	1	19825.0	2566.6	1	645	151
chr22	3008.2	2147.0	6	45707.0	3218.6	-	1024	361

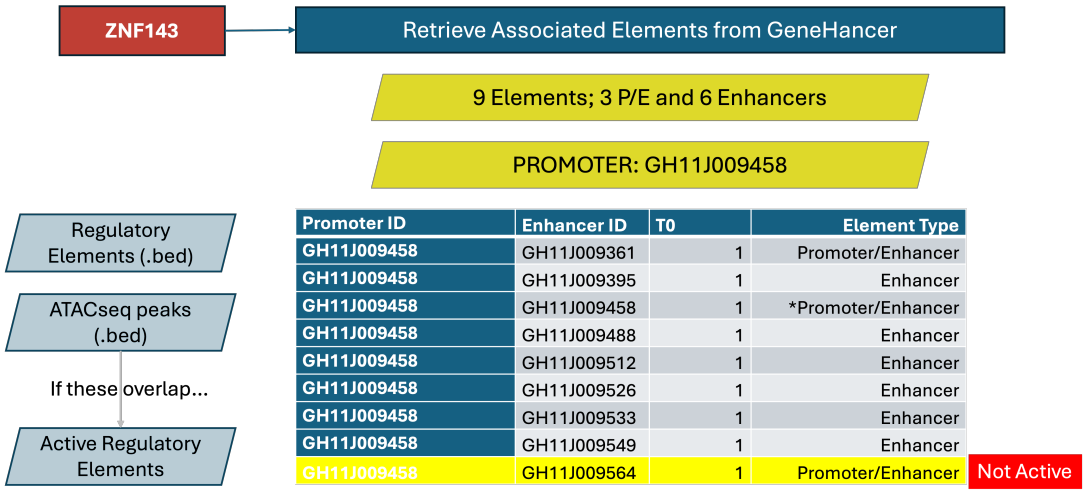


Figure 2.7: Association of Regulatory Elements for the DEG ZNF143, promoter identification, determining the active elements by incorporating accessibility data.

4. **Construction of Fully Connected Networks:** For each DEG with a primary promoter, a fully connected network was created, linking the promoter to its associated regulatory elements.

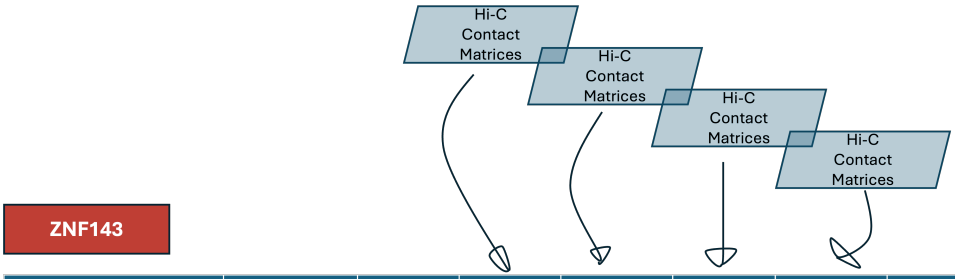
5. **Incorporating ATAC-seq and Hi-C Data:** The initial networks were refined at each time point using ATAC-seq and Hi-C data:

- Nodes (regulatory elements) were retained only if deemed "active" based on overlap with ATAC-seq peaks (Figure 2.7). Activity was determined us- ing a fixed intersection ratio/percentage defined in the tool (e.g., bedtools intersect).

- Edges (connections between regulatory elements) were preserved only if a Hi-C contact was detected in the 10kb resolution Hi-C interaction matrix (Figure 2.8).

6. **Network Validation:** Networks comprising only zeros or ones were eliminated, as they lack variability in connectivity and are thus uninformative regarding co-operative, competitive, or redundant regulatory behavior. Such networks cannot be linked to observed differential expression changes.

   Promoter-based networks were generated for each DEG in all chromosomes at every time point.



Promoter ID	Enhancer ID	T0	T1 ESC	T2 MES	T3 CP	T4 CM	Element Type
GH11J009458	GH11J009361	1	1	0	1	1	Promoter/Enhancer
GH11J009458	GH11J009395	1	1	1	1	1	Enhancer
GH11J009458	GH11J009458	1	1	1	1	1	*Promoter/Enhancer
GH11J009458	GH11J009488	1	1	1	1	1	Enhancer
GH11J009458	GH11J009512	1	1	1	1	0	Enhancer
GH11J009458	GH11J009526	1	1	1	1	0	Enhancer
GH11J009458	GH11J009533	1	1	1	1	0	Enhancer
GH11J009458	GH11J009549	1	1	1	0	1	Enhancer

Figure 2.8: Incorporating interaction data at each time point to the initial fully connected network of ZNF143 to make it stage-specific.

2.2.3. Computational Framework and Integration

The datasets and external resources required for network generation were integrated using the R programming language. Additional preprocessing tasks—such as data downloading, file format conversion, and identification of region overlaps—were performed using tools and platforms including UCSC Table Browser, UCSC utils toolkit, bedtools, juicertools[86], and custom Bash scripts.

2.2.4. Correlation Analysis of Network Connectivity and Gene Expression

To assess the relationship between network connectivity changes and gene expression, two types of correlations were calculated (Figure 2.9):

1. Spearman Correlation between the total number of edges in the network and expression level.
2. Spearman Correlation between changes in network structure, measured by Hamming distance (suitable for binary vectors), and changes in expression level.

Value	T1 ESC	T2 MES	T3 CP	T4 CM
# of edges	8	7	7	5
expression	21.173	150.158	134.968	100.927
Δ expression	0	128.985	-15.190	-34.041
Δ connectivity	0	1	2	4

CORRS_EDGES

CORRS_CHANGES

Figure 2.9: Spearman Correlation between the total number of edges in the network and expression level: CORRS_EDGES. Spearman Correlation between changes in network structure, measured by Hamming distance and changes in expression level: CORRS_CHANGES.

Spearman correlation coefficients were computed using the cor() function in R Studio. Network distances were calculated as follows:

- For the initial time point, the distance value was set to zero.
- Hamming distances were calculated between network vectors at successive time points (e.g., T2_MES and T1_ESC for the second time point). This yielded a distance vector: $[0, \text{dist}(T2 - T1), \text{dist}(T3 - T2), \text{dist}(T4 - T3)]$.

DEGs that are assigned to each group with absolute correlation coefficients above 0.9 in either correlation analysis were prioritized for further investigation and interpretation.

2.2.5. Interpretation of Correlation Patterns

The calculated correlation values were interpreted based on the following scenarios:

1. **Both Correlations Are High (≥0.9):** A strong and consistent relationship exists between enhancer-promoter interactions and gene expression. This suggests a balanced or additive mode of enhancer action, where the number and changes of interactions are in harmony with expression levels.

2. **High Correlation in Number of Edges (≥0.9), But Low in Network Changes (<0.5):** The static aspect (number of edges) correlates well with expression, but the dynamic aspect (network changes) does not. This might indicate that the overall level of connectivity is important, but specific changes in interactions over time are less critical. Such a scenario could suggest a complex or non-linear

mode of enhancer action, potentially synergistic or competitive.

3. **Low Correlation in Number of Edges (<0.5), But High in Network Changes (≥0.9):** The dynamic aspect of the network (changes in interactions) is more critical than the sheer number of interactions. This suggests that the timing and specific nature of interactions are crucial, rather than just the quantity.
4. **Both Correlations Are Low (<0.5):** This scenario indicates a more indirect or complex relationship between enhancer-promoter interactions and gene expression. Other regulatory mechanisms or factors might influence this relationship.

DEGs in each group were submitted to g:Profiler for functional enrichment using g:SCS correction. Significance thresholds were set at 0.05 for all enrichment analyses.

2.3. Deregulated Protein Complex Analysis

This section outlines the methodology employed to analyze deregulated transcriptional protein complexes. Protein complexes contained at least one transcription factor to ensure relevance to transcriptional regulation and hence named as "transcriptional protein complexes". The workflow comprised four main analyses: evaluating transcriptional complex prediction using scRNA-seq and bulk RNA-seq datasets, followed by detailed investigations into deregulated complexes in the context of aging T cells.

2.3.1. Comparing scRNA-seq and Bulk RNA-seq Techniques

The transcriptional protein complexes in classical and nonclassical monocytes were analyzed across four datasets: three scRNA-seq datasets (Yoshida22 [87], Ahern22 [88], Thomson23 [74]) and one bulk RNA-seq dataset (Williams17 [89]). This analysis aimed to identify differences in the number and composition of protein complexes between the two monocyte subsets, and to evaluate the capabilities of scRNA-seq versus bulk RNA-seq in capturing cell-type-specific transcriptional regulation. The following workflows were implemented:

Bulk RNA-seq Workflow

Raw sequencing reads were mapped to the reference genome using STAR, and gene-level read counts were generated with HTSeq. The counts were normalized to counts

per million (CPM) to address differences in library sizes. Differential gene expression analysis was conducted using DESeq2, and the processed data were used for protein complex prediction and quantification.

scRNA-seq Workflow

Single-cell expression matrices were retrieved from the GEO or CELLxGENE reposi- tories in *h5ad* format (AnnData version 0.8) and processed as follows:

- Quality control and preprocessing were performed using the *Scanpy* Python pack- age.
- Donor-specific pseudobulking was performed with *adpbulk*, employing mean aggregation.
- Cell-type-specific expression matrices for classical and nonclassical monocytes were generated for further analysis.

2.3.2. Dataset Details

Dataset 0: Datasets for Classical and Nonclassical Monocytes

Yoshida22 Dataset: scRNA-seq data from newborns to 14 years of age These datasets were used to explore cell-type-specific transcriptional protein complexes in classical and nonclassical monocytes.

Dataset 1: Aging in PBMC

Bulk RNA-seq data of peripheral blood mononuclear cells (PBMCs) were retrieved from SRA (Study Accession: SRP498208). Samples were categorized into three age groups: Young (n=11, avg age 27.5 ± 2.54), Middle (n=10, avg age 60.4 ± 2.22), and Old (n=14, avg age 80.5 ± 4.57). Raw data were processed as explained above in the Bulk RNA-seq Workflow and *DEseq2* used for differential analysis.

Dataset 2: Naive T Cells in Age and Disease Contexts

Processed scRNA-seq data of naive CD4+ and CD8+ T cells were retrieved from GEO (Series Accession: GSE168215 [87]), encompassing samples from healthy chil- dren (n=24), healthy adults (n=11), COVID-19 children (n=14), and COVID-19 adults (n=28). Consistent preprocessing steps were applied using *adpbulk* and *Scanpy*.

Dataset 3: Aging in Naive T Cells

ScRNA-seq data of naive T cells from children (11–13 yrs, n=16), young adults (20–35 yrs, n=16), and older adults (55–65 yrs, n=16) were retrieved from GEO (Series Accession: GSE214546 [74]). Additional validation data from Yoshida et al. Dataset 2 (Series Accession: GSE168215) were included.

- (n=15).
    - **Ahern22 Dataset**: scRNA-seq data from children aged 1–9 years (n=10).
    - **Thomson23 Dataset**: scRNA-seq data from children aged 11–13 years (n=16).
    - **Williams17 Dataset**: Bulk RNA-seq data from children aged 6 months–10 years (n=17).

2.3.3. Sample-Specific Interaction Network Generation

Sample-specific protein-protein interaction networks (PPINs) and domain-domain interaction networks (DDINs) were constructed using:

- Human protein-protein interaction network: PrePPI database (retrieved Dec 2022) [90].
- Domain-domain interaction network: 3did (release Aug 2022) [91].

The *PPIXpress* tool (version 1.23) [63] was used for network construction, with Ensembl annotation release 108.

2.3.4. Transcriptional Protein Complex Prediction

Protein complexes were predicted using the *JDACO* algorithm (version 1.03) [64] with the following key parameters:

- Seed proteins: 678 human transcription factors from HOCOMOCO

(version 11) [92].

- Complex size: 2–5 members.
- Input networks: Sample-specific PPINs and DDINs.

Each predicted complex contained at least one transcription factor to ensure relevance to transcriptional regulation.

2.3.5. Protein Complex Quantification and Statistical Comparison

Protein complex abundance was quantified using iterative distribution of expression values. Complexes detected in at least 75% of samples within a group were included for comparison. Statistical analyses were performed with *CompleXChange* (version 1.01)[61], employing Wilcoxon rank-sum tests with an FDR threshold of 0.05.

2.3.6. Principal Component Analysis and Clustering

Principal Component Analysis (PCA) and clustering were performed on deregulated protein complexes:

- PCA with 20 components used *Scikit-learn*'s StandardScaler and PCA modules.
- K-means clustering ($k = 3$) was applied to classify age groups, with performance evaluated using accuracy, precision, recall, and F1-score metrics.

2.3.7. Network Analysis and Functional Annotation

Protein complexes were visualized in Cytoscape (version 3.10.1) and subjected to func- tional and domain enrichment analyses:

- *g:Profiler* [93] for functional enrichment using g:SCS correction.
- *DAVID* [94] for domain enrichment of transcription factors.

Significance thresholds were set at 0.05 for all enrichment analyses.

2.3.8. Experimental Validation of Predicted Complexes

Key proteins were validated at the protein and complex levels:

- Protein-level validation: PeptideAtlas database (Human Build 2024.01)[95].
- Complex-level validation: hu.MAP (release 2.0)[96] and CORUM (release 4.1)[97] databases.

Complexes detected in both databases and containing experimentally validated proteins were prioritized for further analysis.

2.3.9. Identification of Regulatory Loops for the Significantly Deregulated Protein Complexes

To investigate the regulatory interactions of transcription factors in deregulated complexes, two complementary analyzes were performed. The first aimed to identify instances where all transcription factors bound to the same regulatory region and targeted the same gene, while the second examined cases where factors bound to different regulatory regions yet still converged on the same target gene.

In the first analysis, the dataset from GeneHancer was structured to group regulatory interactions by regulatory region and target gene. The presence of transcription factors at each region was aggregated, allowing the identification of genes regulated by all three factors within the same regulatory element. By filtering out cases where only one or two factors were present, the analysis focused specifically on co-regulated genes, providing insights into potential transcriptional complexes and cooperative regulation.

The second analysis expanded on this by exploring regulatory multiplicity, identifying target genes controlled through multiple distinct regulatory elements rather than a single shared region. Here, interactions were grouped by target gene, and the number of unique regulatory regions associated with each gene was counted. Only genes regulated through at least two distinct regulatory regions were retained for further study, ensuring that the analysis captured cases where transcription factors exert control through alternative binding sites. This approach highlights genes that may be under more complex regulatory control, potentially influenced by chromatin looping, enhancer interactions, or tissue-specific regulatory programs.

Together, these analyses provided a structured framework for understanding the binding specificity, co-regulation, and redundancy of the transcription factors. The findings

contribute to a deeper understanding of transcriptional networks, particularly in cases where regulatory elements serve as hubs for integrating multiple signals.

3. RESULTS

3.1. Benchmarking the Resolution Enhancement Models

To comprehensively evaluate the performance of deep learning models in enhancing low-resolution Hi-C contact matrices, four distinct scenarios were designed to capture different aspects of model robustness, generalizability, and accuracy. These scenarios form the framework for interpreting the results presented below.

The first scenario establishes baseline performance by assessing the models' ability to reconstruct high-resolution Hi-C matrices for the GM12878 cell line, which was used for model training. This evaluation provides a foundational reference for comparison across other conditions.

The second scenario explores the generalizability of the models to other human cell lines—IMR90, K562, and NHEK—each with distinct chromatin landscapes. This analysis highlights the models' capacity to enhance Hi-C data from diverse genomic environments beyond the training cell line.

In the third scenario, the models' adaptability to varying levels of data sparsity was tested by analyzing GM12878 matrices downsampled at factors of ×25, ×32, and ×50. This simulates different sequencing depths and assesses the models' robustness in handling progressively lower-resolution data.

Finally, the fourth scenario evaluates cross-species robustness by applying the models to mouse embryonic stem cell (mESC) matrices downsampled at ×16. This scenario investigates the models' ability to enhance Hi-C data from a different organism, shedding light on their limitations in addressing species-specific genomic features.

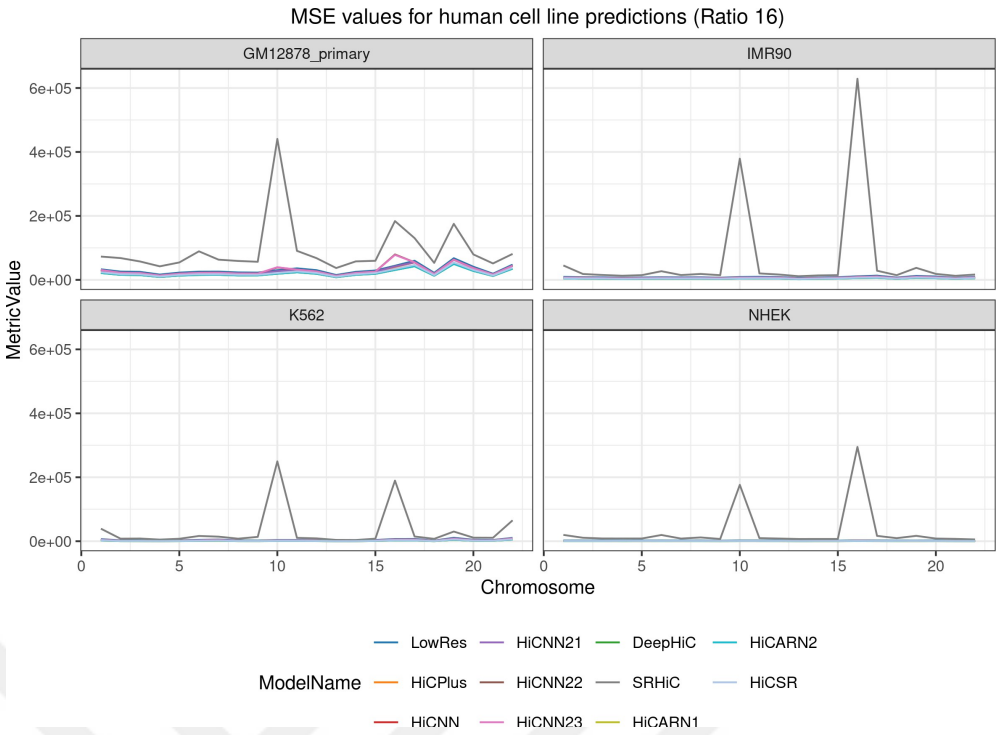
The results for each scenario are discussed in detail in the sections that follow, highlighting model performance across a range of metrics, including Mean Squared Error (MSE), Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM), Pearson correlation, and biologically relevant TAD-based measures. These results collectively provide a nuanced understanding of model capabilities under diverse conditions.

3.1.1.16-Fold Enhancement of Matrices from the Same Cell Line: GM12878

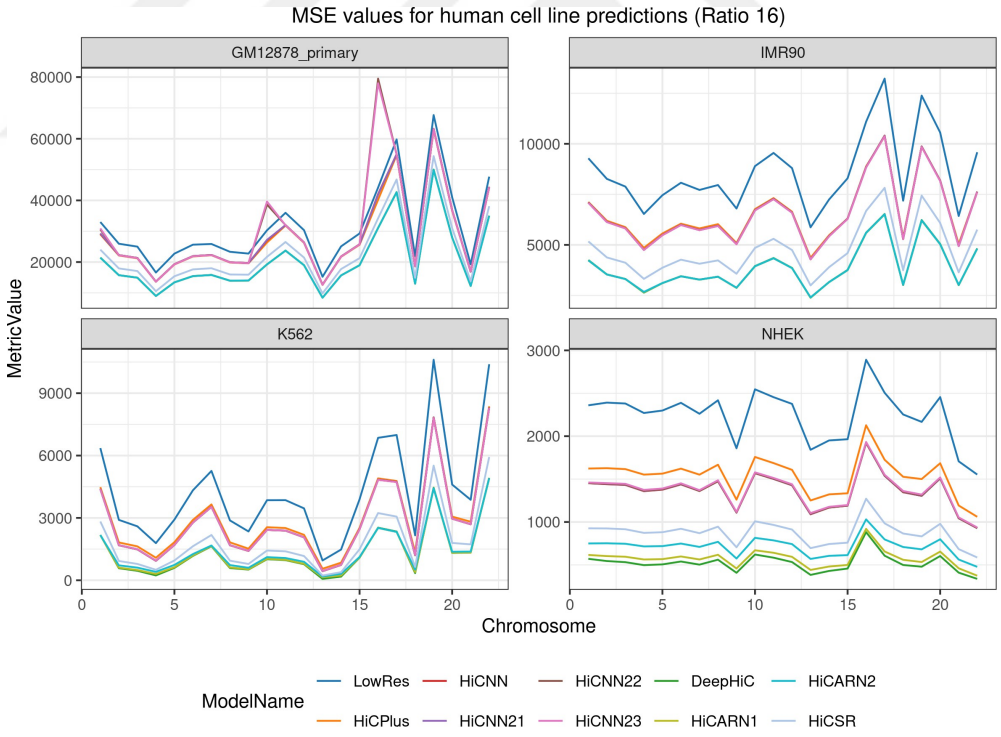
Evaluation of the ten deep learning models for the GM12878 primary cell line at an x16 enhancement ratio reveals significant variability in Mean Squared Error (MSE) performance (Figure 3.1). The low-resolution baseline matrix exhibited the highest average MSE value (31,300.17), confirming the loss of information due to downsampling. Among the models tested, DeepHiC achieved the lowest MSE (20,499.41), closely followed by HiCARN1 (20,465.24) and HiCARN2 (20,474.28), highlighting their superior ability to reconstruct high-resolution matrices.

HiCSR also showed notable performance with an MSE of 22,913.71, outperforming the baseline and other models such as HiCPlus (27,590.10), HiCNN (27,820.71), and HiCNN21 (27,827.28). In contrast, the HiCNN22 and HiCNN23 models showed less competitive results with MSE values of 30,114.69 and 30,068.52, respectively, remaining closer to the baseline than the top-performing methods. SRHiC recorded a higher MSE of 94,209.15, indicating significantly less competitive performance compared to the leading models and it has extreme peaks at chromosome 10 and 19 (Figure 3.1a).

These results emphasize the variability in model performance, with DeepHiC and the HiCARN models emerging as the most effective approaches for minimizing reconstruction error in GM12878 primary. The substantial improvement in MSE compared to the low-resolution matrix underscores the potential of these methods for resolution enhancement. SRHiC, however, demonstrates limitations in its ability to effectively reduce MSE.



(a)



(b)

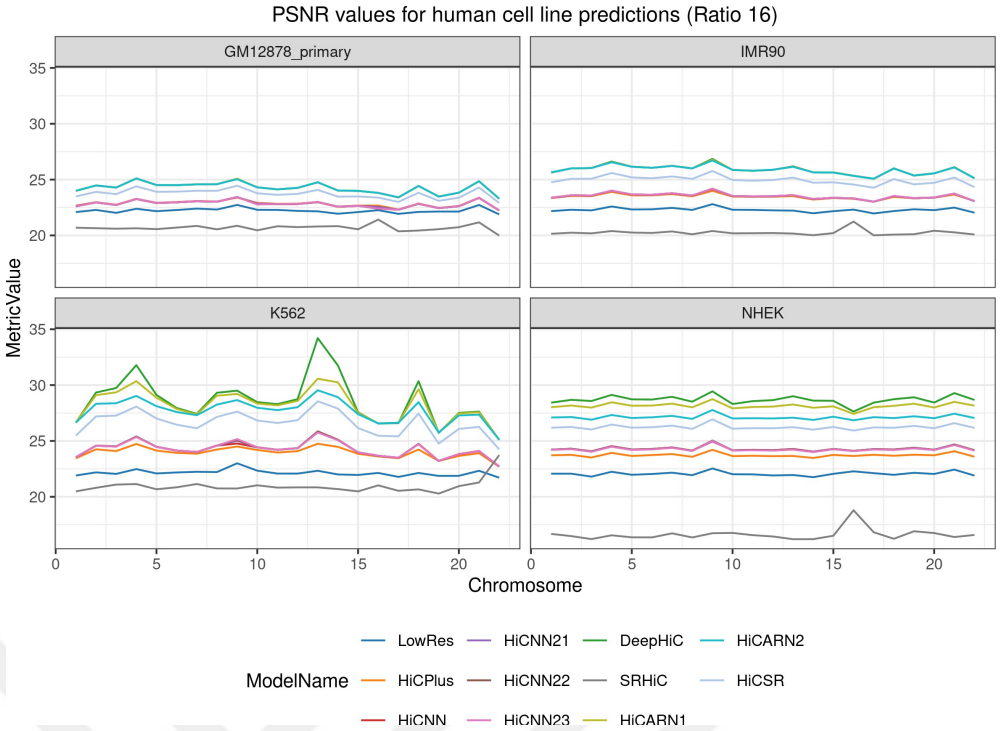
Figure 3.1: (a) MSE values of each model in 4 different human cell lines: GM12878 primary, IMR90, K562 and NHEK. Enhancement ratio is x16. (b) MSE values of each model (excluding SRHiC) in 4 different human cell lines: GM12878 primary, IMR90, K562 and NHEK. Enhancement ratio is x16.

The PSNR values for the GM12878 primary cell line highlight the performance of the models in reducing noise and enhancing resolution (Figure 3.2a). The low-resolution baseline exhibited a PSNR of 22.22, indicating poor signal quality. Among the enhancement models, HiCARN1 achieved the highest PSNR value (24.26), closely followed by HiCARN2 (24.26) and DeepHiC (24.25), showcasing their superior noise-reduction capabilities. HiCSR demonstrated moderate performance with a PSNR of 23.72, while HiCPlus (22.84) and the HiCNN family (22.81–22.82) delivered limited improvements over the baseline. SRHiC recorded a PSNR of 20.68, which was lower than all other tested models, indicating limited ability to enhance signal quality.

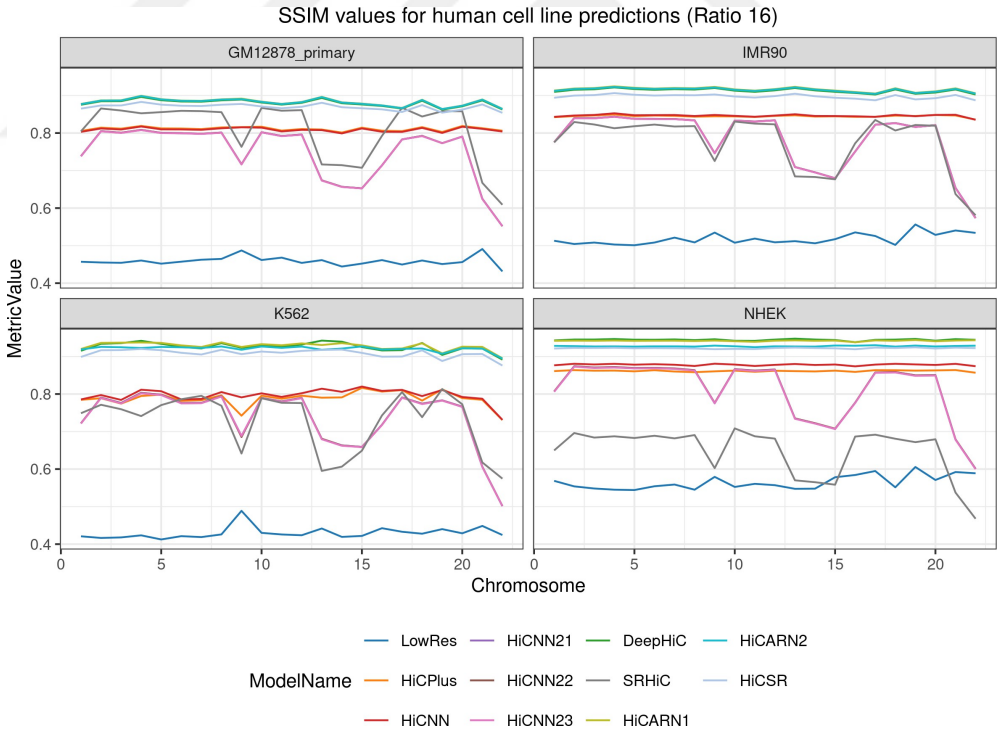
The average PSNR values over the chromosomes reinforce the trends observed for individual chromosomes, further highlighting the superior performance of the HiCARN and DeepHiC models in noise reduction and signal enhancement.

The SSIM metric, which evaluates the structural similarity between the enhanced and original high-resolution matrices, was assessed for the GM12878 primary cell line at an x16 enhancement ratio (Figure 3.2b). The low-resolution baseline matrix showed the lowest SSIM value (0.4588), highlighting the substantial structural degradation caused by downsampling. Among the tested models, HiCARN1 achieved the highest SSIM (0.8826), closely followed by HiCARN2 (0.8820) and DeepHiC (0.8802), indicating their strong ability to preserve structural fidelity during resolution enhancement.

HiCSR also demonstrated commendable performance with an SSIM value of 0.8696, outperforming most other models. HiCPlus (0.8110) and HiCNN (0.8089) displayed moderate SSIM values, indicating some level of structural recovery but falling short of the top-performing models. The HiCNN21, HiCNN22, and HiCNN23 models exhibited lower SSIM values, ranging between 0.7484 and 0.7487, suggesting limited improvement over the low-resolution baseline in terms of structural similarity. SRHiC demonstrated moderate performance with an SSIM value of 0.8089, which was better than the HiCNN family but significantly lower than the top-performing models.



(a)



(b)

Figure 3.2: (a) PSNR values of each model in 4 different human cell lines: GM12878 primary, IMR90, K562 and NHEK. Enhancement ratio is x16. (b) SSIM values of each model in 4 different human cell lines: GM12878 primary, IMR90, K562 and NHEK. Enhancement ratio is x16.

To complement these quantitative metrics, heatmaps were generated for specific regions in chromosomes 13 (Figure 3.3) and 16 (Figure 3.4, Figure A.1). These heatmaps revealed areas where certain models overcrowded regions, increasing noise and disrupting structural similarity. These qualitative visualizations serve as additional evidence supporting the trends observed in PSNR and SSIM values, particularly emphasizing the structural fidelity achieved by HiCARN1, HiCARN2, and DeepHiC. Such disruptions are notably less pronounced in these top-performing models, underscoring their ability to maintain structural integrity while enhancing resolution.

Overall, these results highlight the superior structural preservation capabilities of HiCARN1, HiCARN2, and DeepHiC for the GM12878 primary cell line, emphasizing their effectiveness in reconstructing high-resolution Hi-C matrices.

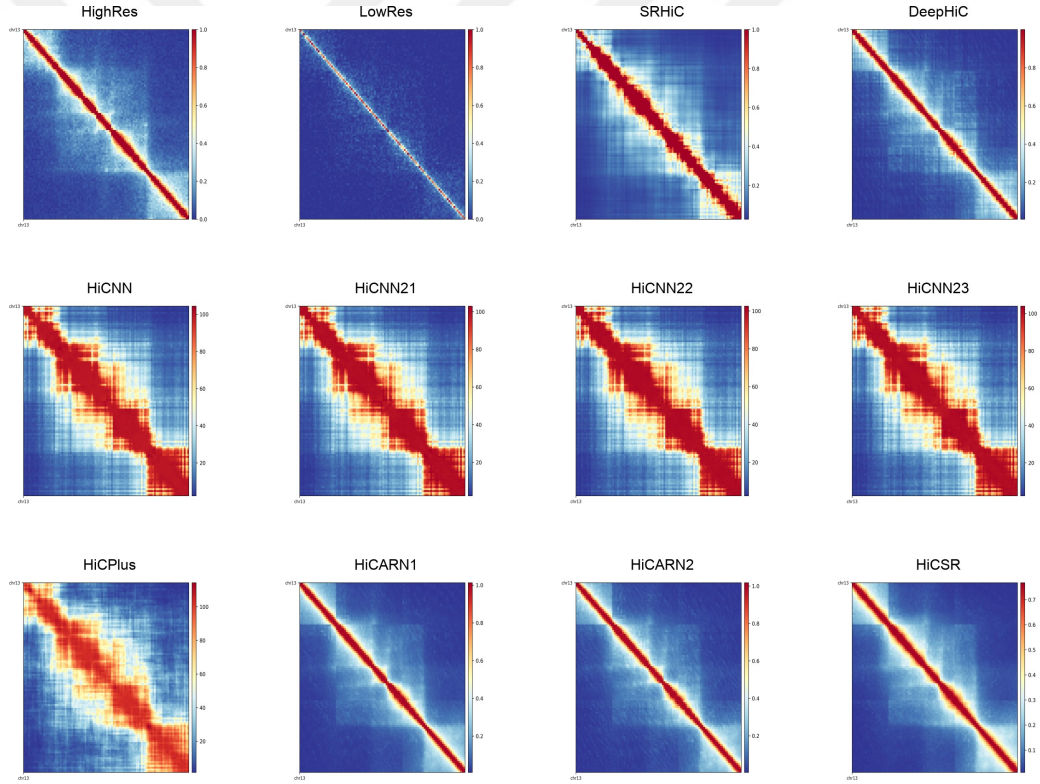


Figure 3.3: Heatmap of a 1 Mb region of the GM12878 primary cell line, corresponding to chromosome 13: 20,000,000–21,010,000. The heatmap shows the original high- resolution matrix, the downsampled low-resolution matrix, and the ×16 enhanced ma- trices generated by various models.

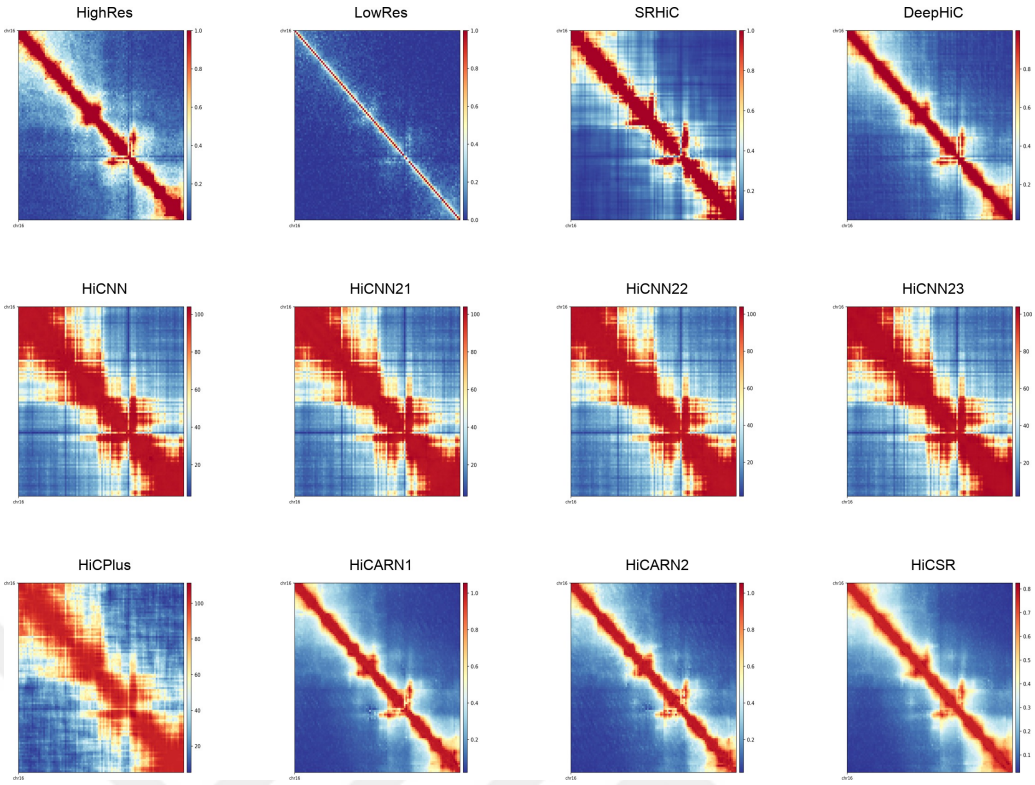


Figure 3.4: Heatmap of a 1 Mb region of the GM12878 primary cell line, corresponding to chromosome 16: 1,500,000–2,510,000. The heatmap shows the original high- resolution matrix, the downsampled low-resolution matrix, and the ×16 enhanced ma- trices generated by various models.

The Pearson correlation scores at the x16 enhancement ratio reveal considerable variability among models in maintaining similarity to high resolution diagonals (Figure 3.5). For the GM12878 primary cell line, HiCARN1 achieved the highest mean correlation score (0.8433), closely followed by HiCARN2 (0.8367) and HiCSR (0.8315). DeepHiC also showed strong performance, with a mean score of 0.8201. HiCPlus (0.6477), HiCNN family (0.6442–0.6459), and SRHiC (0.6863) showed moderate performance, while the low-resolution baseline had the lowest correlation (0.6397).

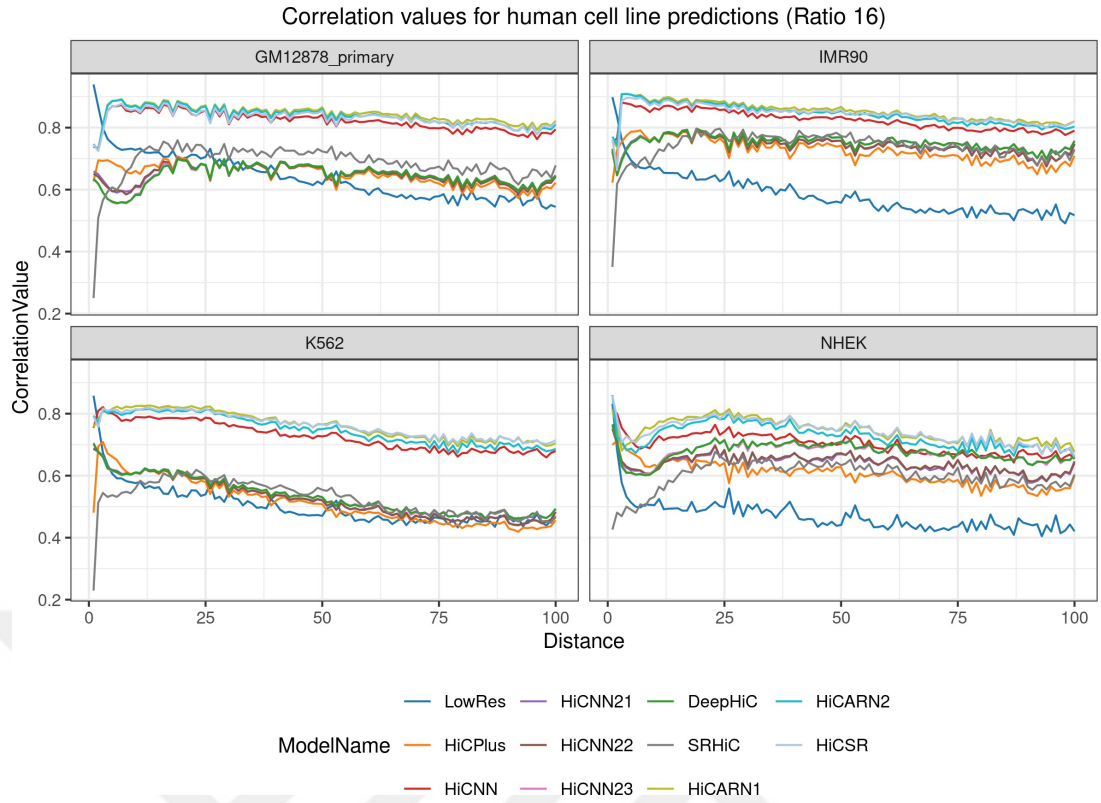


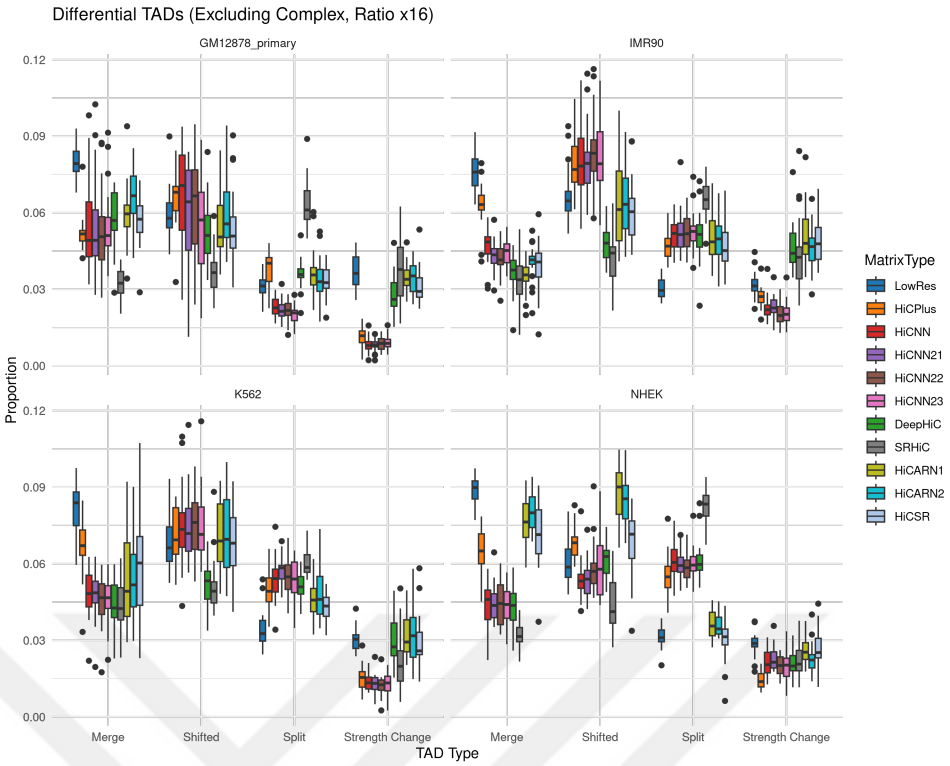
Figure 3.5: Pearson Correlation values of each model in 4 different human cell lines: GM12878 primary, IMR90, K562 and NHEK. Enhancement ratio is x16.

Regarding differential TADs (Figure 3.6a), such as merge TADs, the low-resolution baseline had the highest proportion, reflecting its tendency to oversmooth data and lose boundaries, while SRHiC achieved the lowest, showcasing its ability to preserve TAD boundaries effectively. In the case of shifted TADs, SRHiC once again performed best, with the lowest proportion, suggesting exceptional positional accuracy, whereas other models showed similar and higher proportions, reflecting slight inaccuracies in boundary placement.

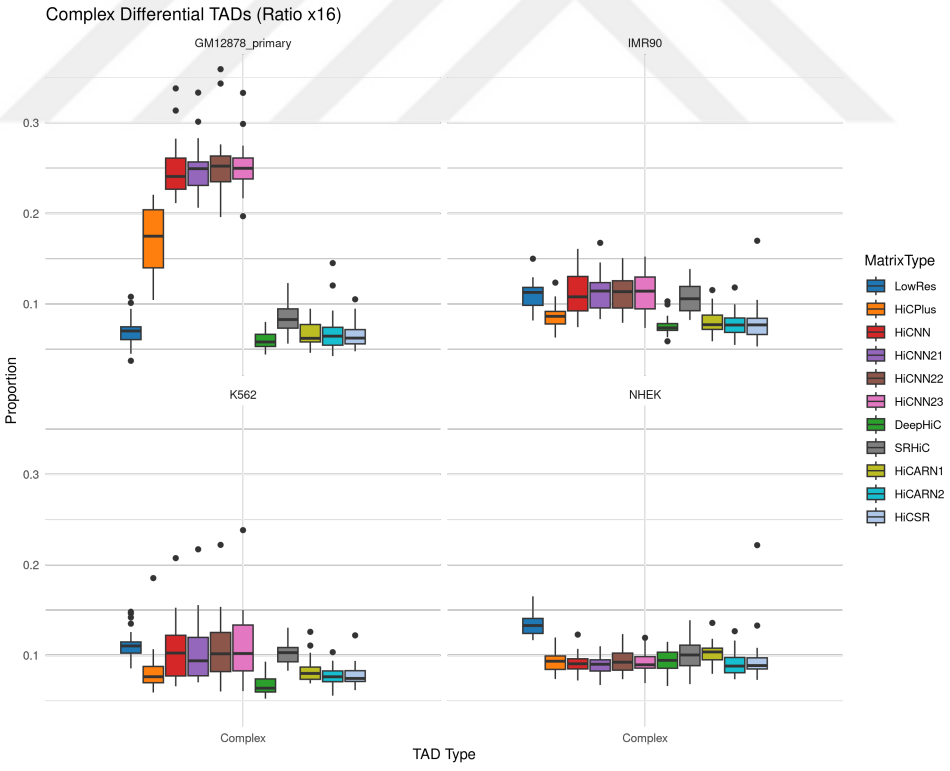
In contrast, for split TADs, SRHiC had the highest proportion, indicating a tendency to introduce spurious boundaries, while the HiCNN family had the lowest, reflecting a conservative approach to boundary detection. The remaining models displayed intermediate proportions, striking a balance between boundary introduction and preservation. Lastly, for strength change TADs, HiCPlus and the HiCNN family had the lowest proportions, indicating their robustness in maintaining boundary intensity, while other

models, including SRHiC, HiCARNs, and DeepHiC, showed higher and similar proportions, suggesting occasional variations in boundary strength.

For complex differential TADs (Figure 3.6b), the HiCNN family had the highest proportions, suggesting potential difficulties in preserving structural fidelity and a tendency to introduce artifacts, whereas models like SRHiC, HiCARNs, DeepHiC, and HiCSR exhibited much lower proportions, indicating stronger performance in maintaining well-defined TAD structures.



(a)



(b)

Figure 3.6: (a) Differential Merge, Shifted, Split, Strength Change TADs (b) Differential Complex TADs in all human cell lines at enhancement ratio of x16.

For the GM12878 primary cell line at the x16 enhancement ratio, the mean proportions of non-differential TADs reveal clear differences in model performance (Figure 3.7 and Table 3.1). The low-resolution baseline achieved a mean proportion of 0.7293, which is higher than that of HiCPlus (0.6668) and the HiCNN family (ranging from 0.6026 for HiCNN to 0.6137 for HiCNN23). This indicates that HiCPlus and HiCNN models perform poorly compared to the baseline.

DeepHiC emerges as the top-performing model, achieving the highest mean proportion of 0.7715, significantly outperforming both the low-resolution baseline and HiCPlus. HiCSR follows closely with a mean proportion of 0.7638, slightly trailing behind DeepHiC but still outperforming all other models. Among the HiCARN models, HiCARN1 and HiCARN2 also achieve strong results, with mean proportions of 0.7518 and 0.7429, respectively, exceeding the baseline and all HiCNN models.

SRHiC also performs well, with a mean proportion of 0.7488, surpassing the baseline and all HiCNN models but underperforming compared to DeepHiC, HiCSR, and HiCARN1. Meanwhile, the HiCNN family, with mean proportions ranging from 0.6026 to 0.6137, demonstrates the lowest performance across all models.

Overall, DeepHiC, HiCSR, and HiCARN1 stood out as the most robust models for preserving TAD structures, particularly excelling in non-differential TAD proportions. SRHiC demonstrated a mixed profile: while it excelled in positional accuracy and boundary preservation, its tendency to introduce new boundaries and other inconsistencies suggests that its performance may vary depending on the specific application. These findings emphasize the importance of selecting enhancement models based on the specific goals of the analysis, whether prioritizing large-scale structural preservation or fine-scale boundary accuracy.

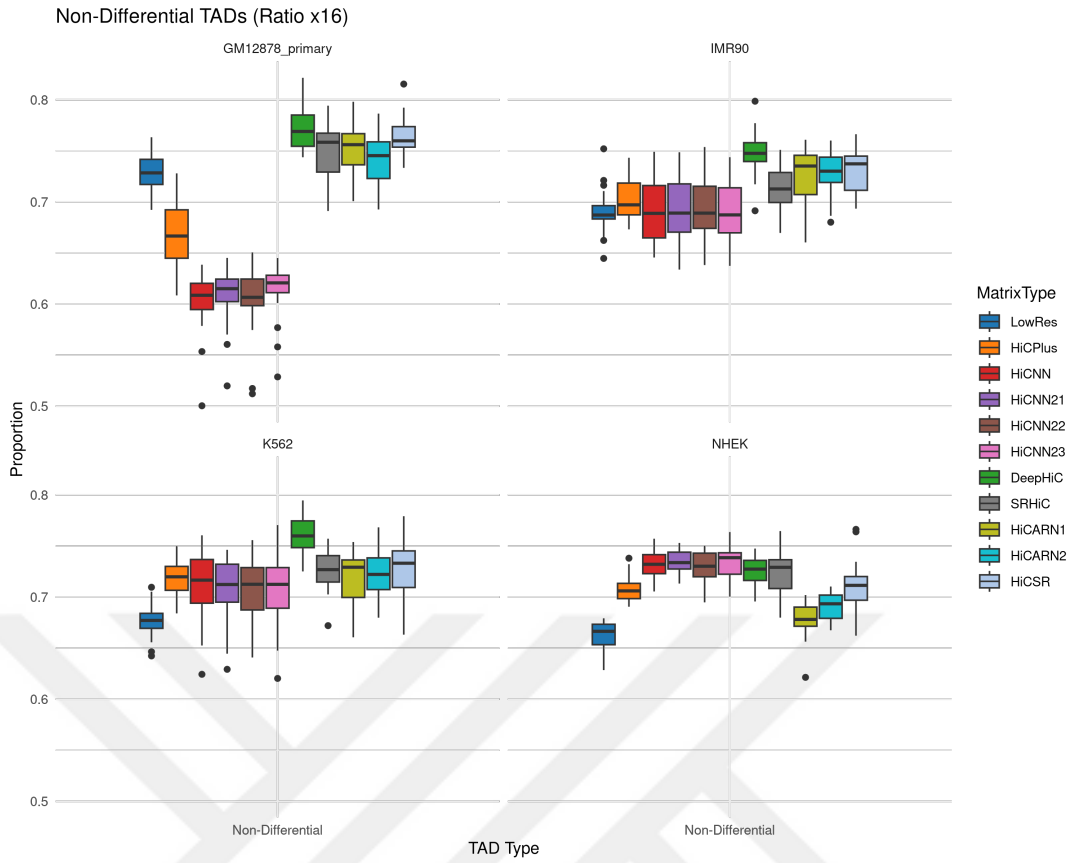


Figure 3.7: Non-differential TADs in all human cell lines at enhancement ratio of x16.

Table 3.1: Pair-wise comparison significance values for Non-differential TADs in GM12878 cell line of x16 ratio.

	LowRes	HiCPlus	HiCNN	HiCNN21	HiCNN22	HiCNN23	DeepHiC	SRHiC	HiCARN1	HiCARN2	HiCSR
LowRes	–	–	–	–	–	–	–	–	–	–	–
HiCPlus	**5.476e-09**	–	–	–	–	–	–	–	–	–	–
HiCNN	**9.505e-13**	**4.285e-09**	–	–	–	–	–	–	–	–	–
HiCNN21	**9.505e-13**	**1.349e-07**	ns	–	–	–	–	–	–	–	–
HiCNN22	**9.505e-13**	**2.174e-08**	ns	ns	–	–	–	–	–	–	–
HiCNN23	**9.505e-13**	**2.353e-07**	ns	ns	ns	–	–	–	–	–	–
DeepHiC	**3.336e-08**	**9.505e-13**	**9.505e-13**	**9.505e-13**	**9.505e-13**	**9.505e-13**	–	–	–	–	–
SRHiC	**4.270e-03**	**1.152e-09**	**9.505e-13**	**9.505e-13**	**9.505e-13**	**9.505e-13**	**1.223e-02**	–	–	–	–
HiCARN1	**1.289e-03**	**4.277e-11**	**9.505e-13**	**9.505e-13**	**9.505e-13**	**9.505e-13**	**1.612e-02**	ns	–	–	–
HiCARN2	ns	**1.152e-09**	**9.505e-13**	**9.505e-13**	**9.505e-13**	**9.505e-13**	**9.771e-04**	ns	ns	–	–
HiCSR	**3.368e-07**	**9.505e-13**	**9.505e-13**	**9.505e-13**	**9.505e-13**	**9.505e-13**	ns	ns	ns	**1.223e-02**	–

The results of the first scenario demonstrate a clear hierarchy in model performance, with HiCARN1, HiCARN2, and DeepHiC emerging as the most effective methods for enhancing resolution in the GM12878 primary cell line at a $\times 16$ enhancement ra- tio. These models leverage advanced architectural features such as residual learning, multi-scale feature extraction, and adversarial optimization, enabling them to achieve superior performance across multiple metrics, including MSE, PSNR, SSIM, and Pearson correlation. Their ability to minimize reconstruction error and preserve structural fidelity is expectable, given their sophisticated designs. However, there is still room for improvement to further enhance their effectiveness and adaptability across diverse scenarios.

HiCARN1 and HiCARN2, with their generative residual architectures and multi-scale learning capabilities, excelled in reconstructing chromatin interaction matrices with both high accuracy and structural fidelity [98, 99]. Their success highlights the importance of multi-scale learning in capturing both local and global patterns, as well as the value of residual connections in stabilizing training and ensuring robust gradient flow. Similarly, DeepHiC's GAN-based architecture, incorporating adversarial learning alongside residual blocks, demonstrates its strength in refining high-level structural features while maintaining fine-grained details. Despite their high performance, these models could be further improved by incorporating diverse biologically relevant loss functions, such as TAD boundary preservation loss or insulation loss, which would align model optimization with domain-specific goals. Additionally, the inclusion of attention mechanisms, such as self-attention layers, could enhance the models' ability to focus on critical chromatin interaction regions, potentially improving both accuracy and interpretability.

HiCSR demonstrated commendable performance, particularly in structural fidelity as reflected by its SSIM values, but it lagged behind the top-performing models in terms of reconstruction accuracy (MSE and PSNR). This outcome is consistent with its reliance on denoising autoencoders (DAEs) [100] and adversarial frameworks [101, 102], which prioritize noise reduction and biological reproducibility over pixel-level precision. To address these limitations, integrating multi-scale feature extraction could improve HiCSR's ability to capture diverse chromatin interaction patterns. Furthermore, refining its adversarial framework with perceptual loss or TAD-specific losses could

enhance its balance between accuracy and structural preservation.

SRHiC displayed mixed results, excelling in positional accuracy and boundary preservation while struggling with reconstruction accuracy and structural fidelity. These outcomes reflect the strengths and weaknesses of its multi-scale CNN architecture, which, while effective at certain scales, tends to amplify certain features disproportionately, resulting in spurious boundaries and over-enhancement in specific regions. Addressing these issues would require structural regularization techniques, such as constraint-based losses or biologically inspired penalties, to ensure that enhanced features align with biological expectations. Additionally, introducing residual attention mechanisms could dynamically adjust the emphasis on different chromosomal regions, reducing the likelihood of over-enhancement and improving overall performance.

In contrast, simpler models like HiCPlus and the HiCNN family struggled to compete, showing only marginal improvements over the low-resolution baseline. These models, characterized by their relatively shallow architectures and lack of advanced features such as residual learning, multi-scale refinement, or adversarial training, lack the complexity needed to effectively reconstruct high-resolution matrices. While their simplicity makes them computationally efficient, their underperformance highlights the need for structural enhancements. Increasing architectural depth, incorporating perceptual and SSIM losses, and experimenting with lightweight GAN frameworks could help these models become more competitive while retaining their efficiency. For instance, introducing hierarchical refinement layers, akin to those in HiCARN or HiCNN2-3, could significantly improve their ability to model complex chromatin interaction patterns.

Broader architectural innovations across all models could further elevate performance in this task. Attention mechanisms, such as those employed in transformer-based architectures, could enhance the models' focus on long-range chromatin interactions, improving their ability to generalize across cell lines and chromosomal regions [103–105]. Additionally, hybrid architectures that combine GANs with variational autoencoders (VAEs) or transformer components could provide a more robust framework for resolution enhancement by integrating complementary strengths. Domain-specific loss functions, such as TAD boundary accuracy, loop detection metrics, or insulation scores, could further improve the biological relevance and interpretability of the reconstructed

matrices.

In conclusion, while HiCARN1, HiCARN2, and DeepHiC represent the current state-of-the-art in reconstructing high-resolution Hi-C matrices, targeted improvements could further enhance their robustness, biological fidelity, and applicability to diverse datasets. For moderately performing models like HiCSR and SRHiC, integrating multi-scale learning and biologically constrained losses offers a clear pathway to improved performance. Simpler models such as HiCPlus and HiCNN, on the other hand, would benefit from deeper architectures and advanced training techniques to bridge the performance gap with more sophisticated methods. These findings emphasize the critical importance of aligning model design with specific biological goals and computational constraints to advance the field of genome architecture studies.

3.1.2. 16-Fold Enhancement of Matrices Across Different Cell Lines: IMR90, K562, and NHEK

The evaluation of x16 enhancement across three additional human cell lines—IMR90, K562, and NHEK—revealed trends consistent with the GM12878 primary cell line. DeepHiC and the HiCARN models (HiCARN1 and HiCARN2) consistently demonstrated superior performance across all metrics, including MSE, PSNR, and SSIM, confirming their robustness in reconstructing high-resolution Hi-C matrices.

For MSE, DeepHiC achieved the lowest values in all three cell lines, followed closely by HiCARN1 and HiCARN2. These models demonstrated substantial reductions in error compared to the low-resolution baseline, with improvements most pronounced for the K562 and NHEK cell lines. HiCSR consistently ranked as a mid-performing model, while HiCNN-based models and HiCPlus exhibited moderate improvements over the baseline but lagged significantly behind the top performers.

The PSNR and SSIM metrics further corroborated these findings. DeepHiC and HiCARN models consistently achieved the highest values, reflecting their ability to enhance signal quality and preserve structural integrity. Notably, NHEK exhibited the highest PSNR values across all models, with DeepHiC recording 28.68, suggesting slightly greater ease of resolution enhancement in this cell line. In the context of non-differential TAD proportions, DeepHiC again led the field, achieving mean proportions

of 0.7481, 0.7605, and 0.7248 for IMR90, K562, and NHEK, respectively. These results highlight the model's ability to preserve large-scale genomic structures, though the residual differential proportions underscore challenges in fully recovering all TADs. HiCSR and the HiCARN models performed similarly, while HiCNN and HiCPlus remained closer to the low-resolution baseline.

To complement the quantitative analysis, heatmaps of a representative 1 Mb region were generated for each cell line (for IMR90 Figures 3.8, A.2, A.6; for K562 Fig- ures 3.9a, A.3, A.7; and for NHEK Figures 3.9b, A.4, A.8). These visualiza- tions reveal critical differences in model performance. DeepHiC, HiCARN1, and HiCARN2 produce cleaner reconstructions with minimal noise, closely resembling the high-resolution ground truth. HiCSR exhibits moderate performance, with some visible artifacts that align with its mid-tier quantitative metrics. In contrast, SRHiC and HiCNN-based models overcrowd certain regions, introducing artificial peaks and disrupting structural similarity. For instance, SRHiC displays significant overcrowding in the IMR90 heatmap, correlating with its poor MSE and SSIM values.

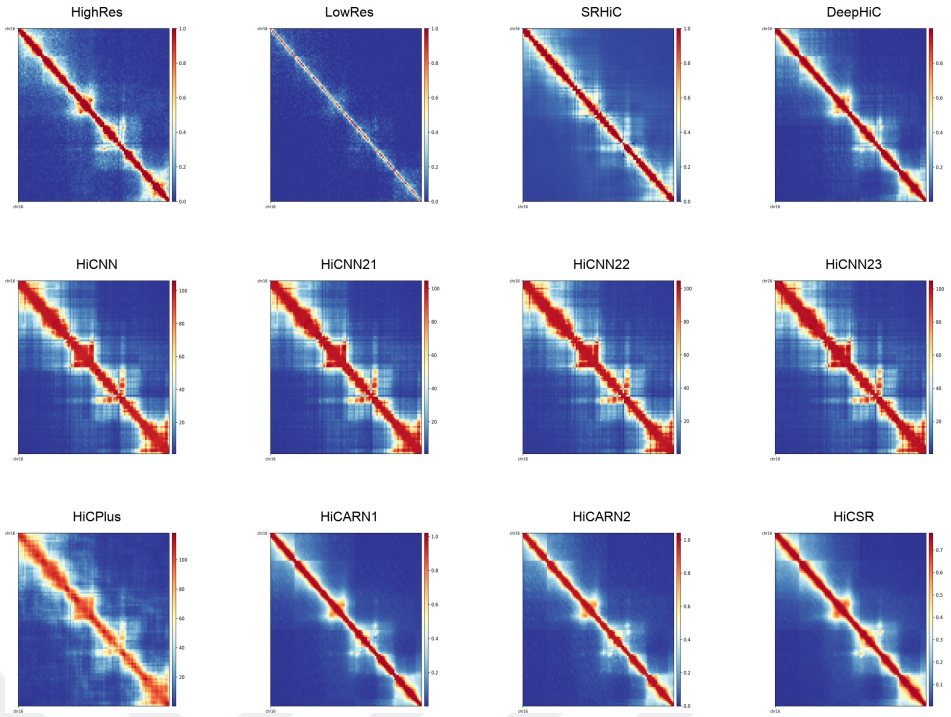
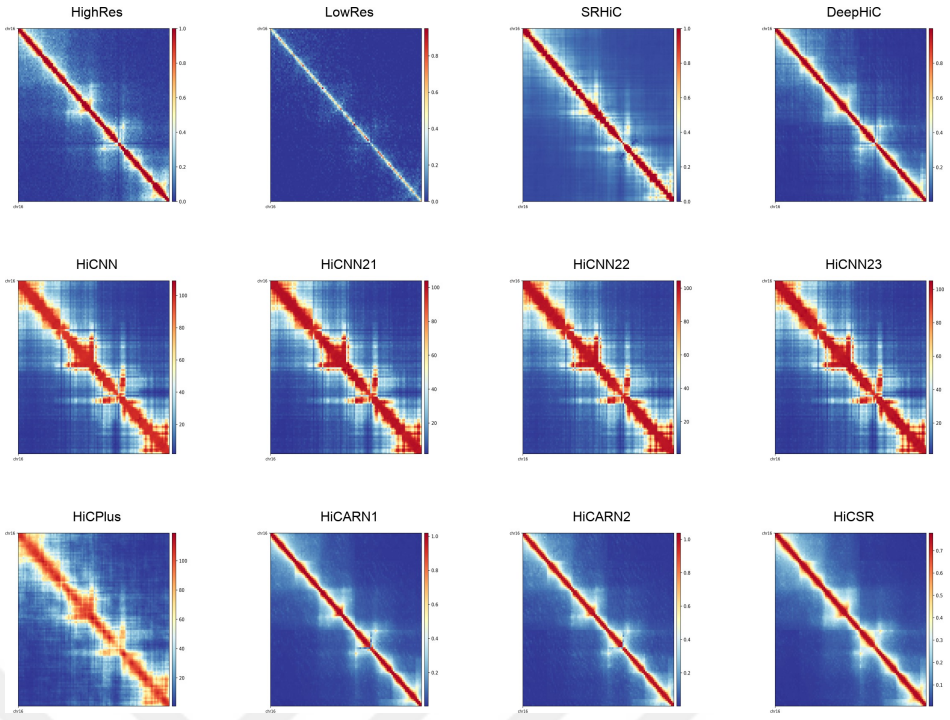
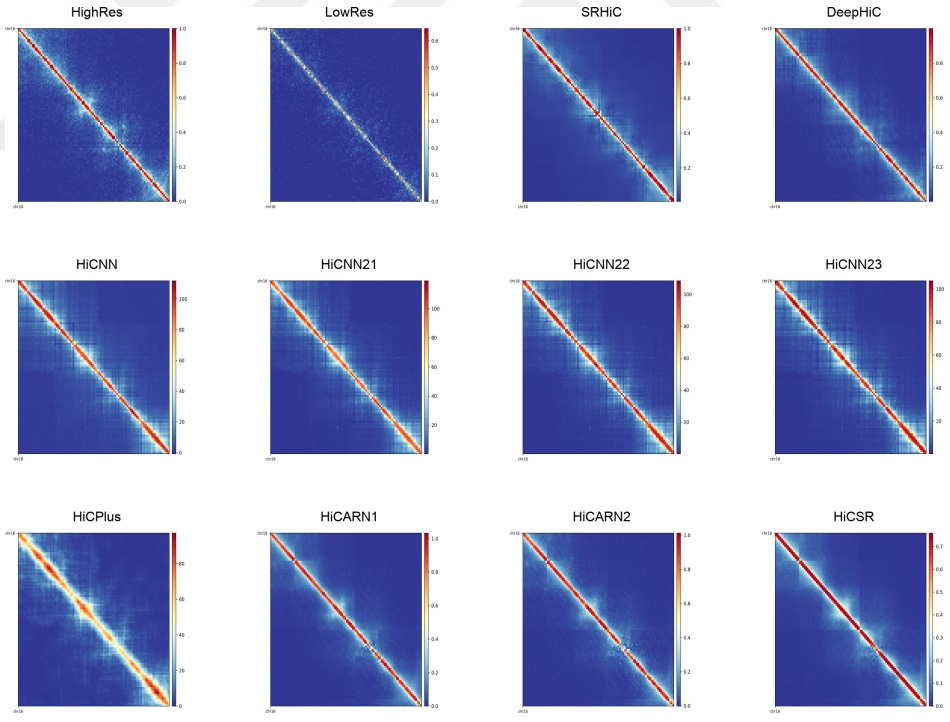


Figure 3.8: Heatmap of a 1 Mb region of the IMR90 cell line, corresponding to chromosome 16: 1,500,000–2,510,000. The heatmap shows the original high-resolution matrix, the down-sampled low-resolution matrix, and the ×16 enhanced matrices gen- erated by various models.

Across all four cell lines, the heatmaps consistently highlight the ability of DeepHiC and the HiCARN models to achieve smoother reconstructions with minimal noise artifacts. These visual observations, combined with the quantitative metrics, reinforce the generalizability of these models as the most effective tools for Hi-C resolution enhancement. However, while their performance in enhancing resolution and preserving structural integrity is robust, the limitations in finer-scale recovery, particularly for TADs, underscore the need for future methodological improvements to address these challenges and fully realize their potential for high-resolution genomic analyses.



(a)



(b)

Figure 3.9: Heatmap of a 1 Mb region of the (a) K562 and (b) NHEK cell lines, corresponding to chromosome 16: 1,500,000–2,510,000. The heatmap shows the original high-resolution matrix, the down-sampled low-resolution matrix, and the ×16 enhanced matrices generated by various models.

3.1.3. Sensitivity to Initial Sequencing Depth

The evaluation of deep learning models for Hi-C resolution enhancement revealed consistent trends across multiple metrics, including Mean Squared Error (MSE), Structural Similarity Index (SSIM), Peak Signal-to-Noise Ratio (PSNR), and non-differential TAD proportions. Analysis was conducted across three enhancement ratios (x25, x32, and x50) and four human cell lines (GM12878, IMR90, K562, and NHEK), providing a comprehensive perspective on model performance under varying degrees of downsampling.

Among the models, DeepHiC consistently emerged as the top performer. Across all enhancement ratios, it achieved the lowest MSE values, the highest SSIM scores, and the best PSNR results, underscoring its ability to balance fine-scale reconstruction, structural fidelity, and noise reduction. Importantly, DeepHiC also excelled in preserving large-scale genomic structures, as evidenced by the non-differential TAD proportions. For GM12878, it maintained TAD proportions of 0.7751, 0.7770, and 0.7650 at x25, x32, and x50 ratios, respectively. This robustness extended across all cell lines, with only minor declines in performance at higher enhancement ratios, demonstrating its suitability for applications requiring reliable structural preservation under sparse input conditions.

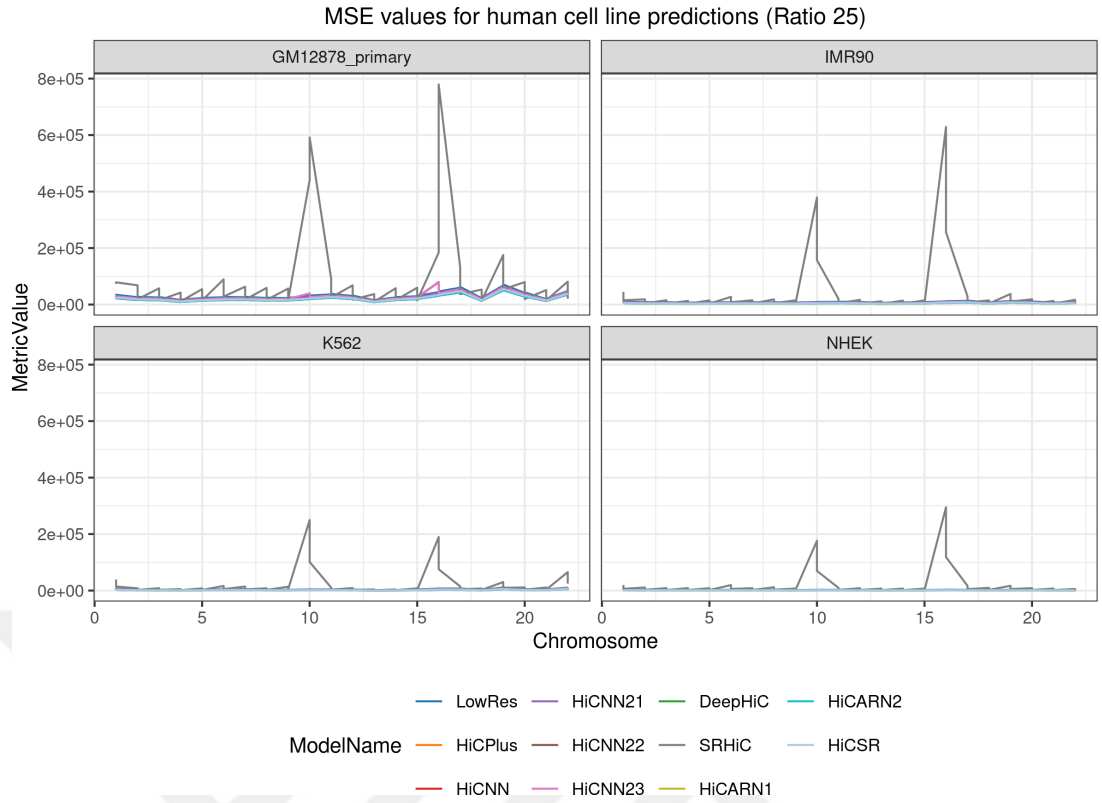
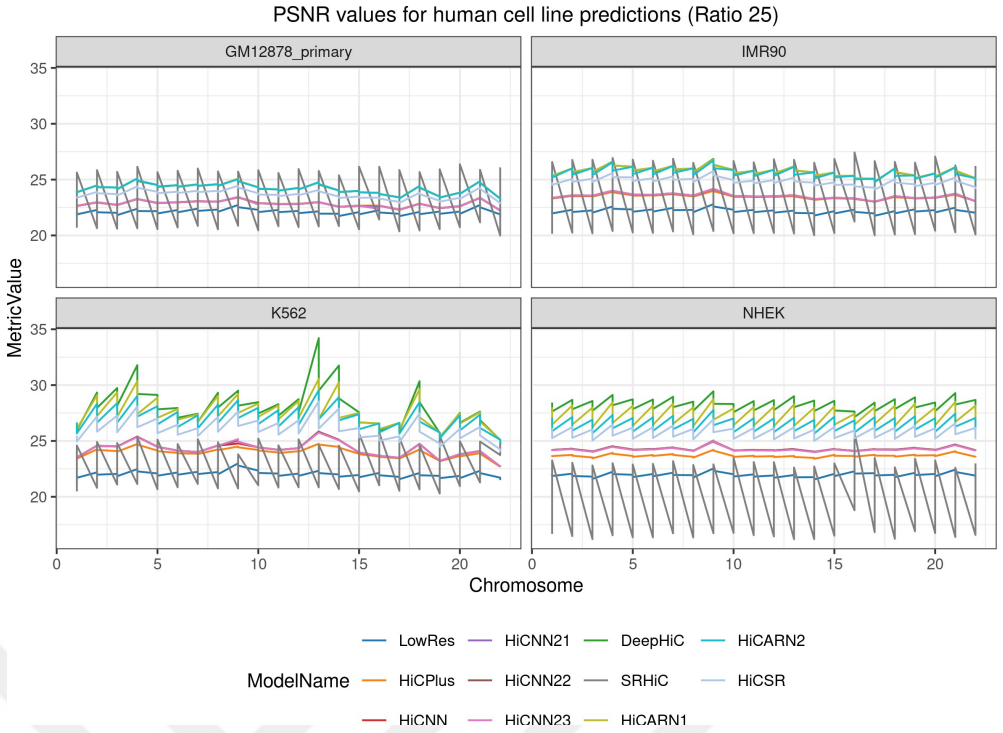
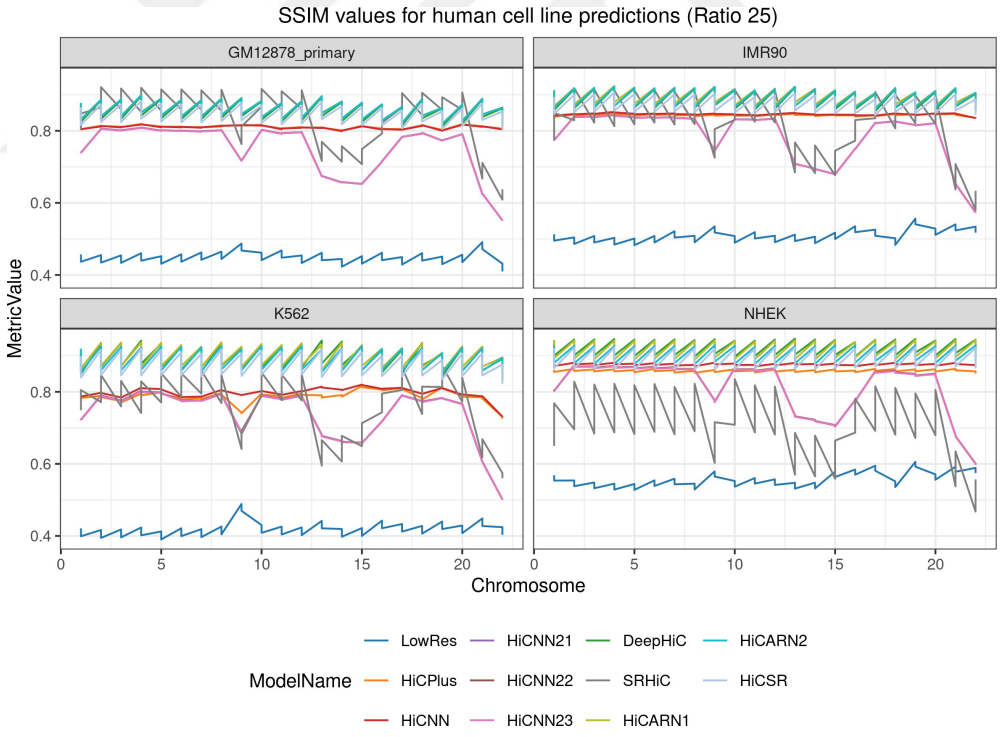


Figure 3.10: MSE values of each model in 4 different human cell lines: GM12878 primary, IMR90, K562 and NHEK. Enhancement ratio is x25.

The HiCARN models (HiCARN1 and HiCARN2) closely followed DeepHiC in overall performance. They consistently delivered low MSE and high SSIM and PSNR values, with non-differential TAD proportions comparable to DeepHiC in many cases. HiCARN1, for instance, achieved a proportion of 0.7559 for GM12878 at x32, slightly trailing DeepHiC but outperforming other models. While their performance remained robust across most cell lines, the HiCARN models exhibited slight declines at x50, particularly for NHEK, where TAD proportions dropped to 0.6551 for HiCARN1 and 0.6496 for HiCARN2. Despite these declines, HiCARN models demonstrated strong generalizability and a balance between fine-scale reconstruction and structural preservation.



(a)



(b)

Figure 3.11: (a) PSNR and (b) SSIM values of each model in 4 different human cell lines: GM12878 primary, IMR90, K562 and NHEK. Enhancement ratio is x25.

HiCSR displayed competitive performance across many metrics but revealed significant variability depending on the cell line and enhancement ratio. While its MSE, SSIM, and PSNR values placed it consistently in the mid-tier, its non-differential TAD proportions told a more nuanced story. For GM12878, HiCSR achieved 0.7687 at x25 and 0.7578 at x32, closely rivaling DeepHiC. However, its performance dropped sharply for NHEK at x50, where the proportion fell to 0.3042, highlighting a limitation in handling diverse genomic contexts under high downsampling. This variability was also reflected in its sensitivity to intrachromosomal variability, as evidenced by the emergence of zigzag patterns in PSNR and SSIM plots.

The zigzag patterns observed in PSNR and SSIM metrics provided additional insight into the models' performance. These patterns, characterized by a range of metric values across chromosomes, emerged most prominently at higher enhancement ratios (x32 and x50). Initially observed with SRHiC, the zigzags later appeared in DeepHiC and HiCARN models, reflecting their sensitivity to structural variability within chromosomes. The metrics were calculated using fixed-size (1Mb) subregions of the Hi-C matrix, which likely amplified the effects of intrachromosomal variability—the non-uniform organization of chromatin within chromosomes. Factors such as gene density, TAD complexity, and chromatin loop distribution likely contributed to these variations, challenging the models' ability to consistently reconstruct interactions across all subregions.

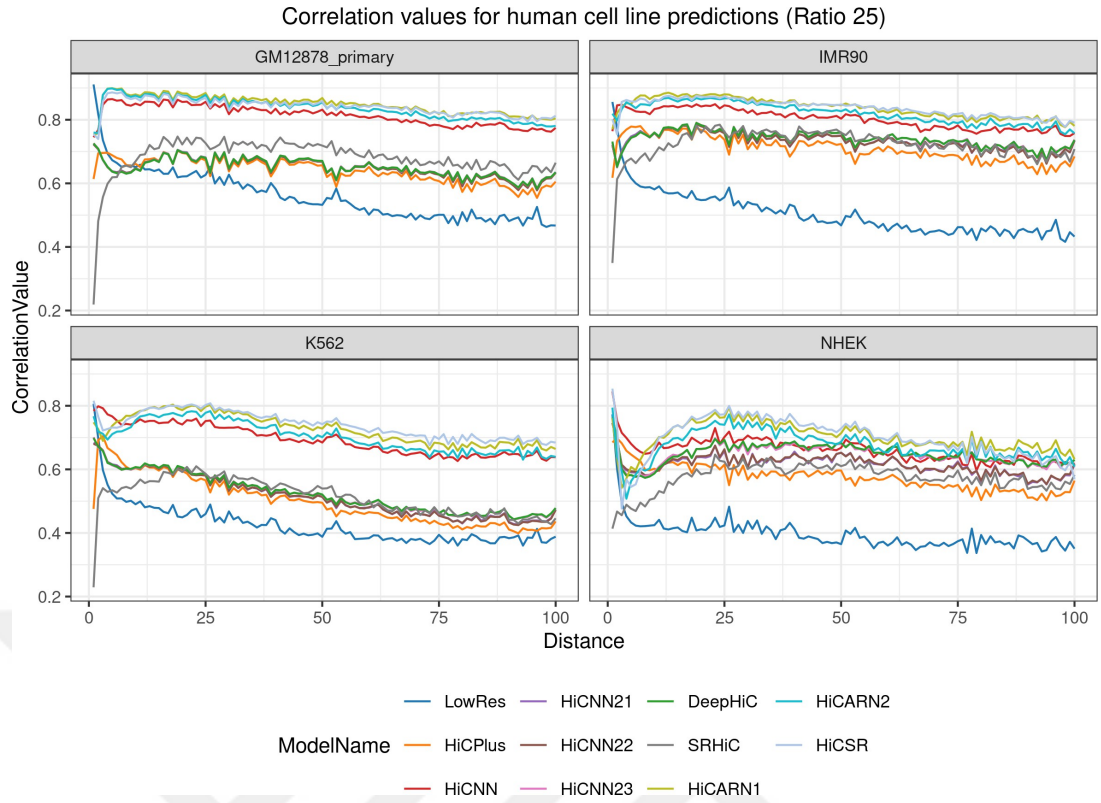


Figure 3.12: Pearson Correlation values of each model in 4 different human cell lines: GM12878 primary, IMR90, K562 and NHEK. Enhancement ratio is x25.

Models like SRHiC, which prioritize large-scale structures, exhibited zigzags earlier, reflecting their heightened sensitivity to structural heterogeneity. In contrast, the gradual emergence of zigzags in DeepHiC and HiCARN models suggests a more balanced approach, with these models attempting to reconcile fine-scale accuracy with structural fidelity. The lack of zigzags in low-resolution baselines and HiCNN-based models indicates their limited ability to capture nuanced structural features, resulting in more uniform but less accurate reconstructions.

SRHiC demonstrated notable strengths in preserving large-scale structures, as evidenced by its consistently high non-differential TAD proportions. For instance, SRHiC achieved 0.7440 for GM12878 at x50 and 0.7187 for NHEK at x32, often outperforming HiCARN models in this metric. However, its higher MSE values and lower SSIM and PSNR scores indicate a trade-off, with SRHiC prioritizing TAD-level fidelity at the expense of fine-scale reconstruction accuracy.

The HiCNN-based models and HiCPlus exhibited limited improvements across all metrics, consistently lagging behind DeepHiC, HiCARN models, and HiCSR. While these models demonstrated some structural recovery, their non-differential TAD proportions and other metrics highlighted their inability to handle higher enhancement ratios effectively. For example, HiCNN22 and HiCNN23 achieved slightly better proportions for NHEK at x50 (0.6989 and 0.7059, respectively) but failed to match the fine-scale reconstruction capabilities of the top models. These findings indicate that while HiCNN-based models and HiCPlus provide moderate performance, they lack the robustness required for high-resolution Hi-C analysis under extreme downsampling.

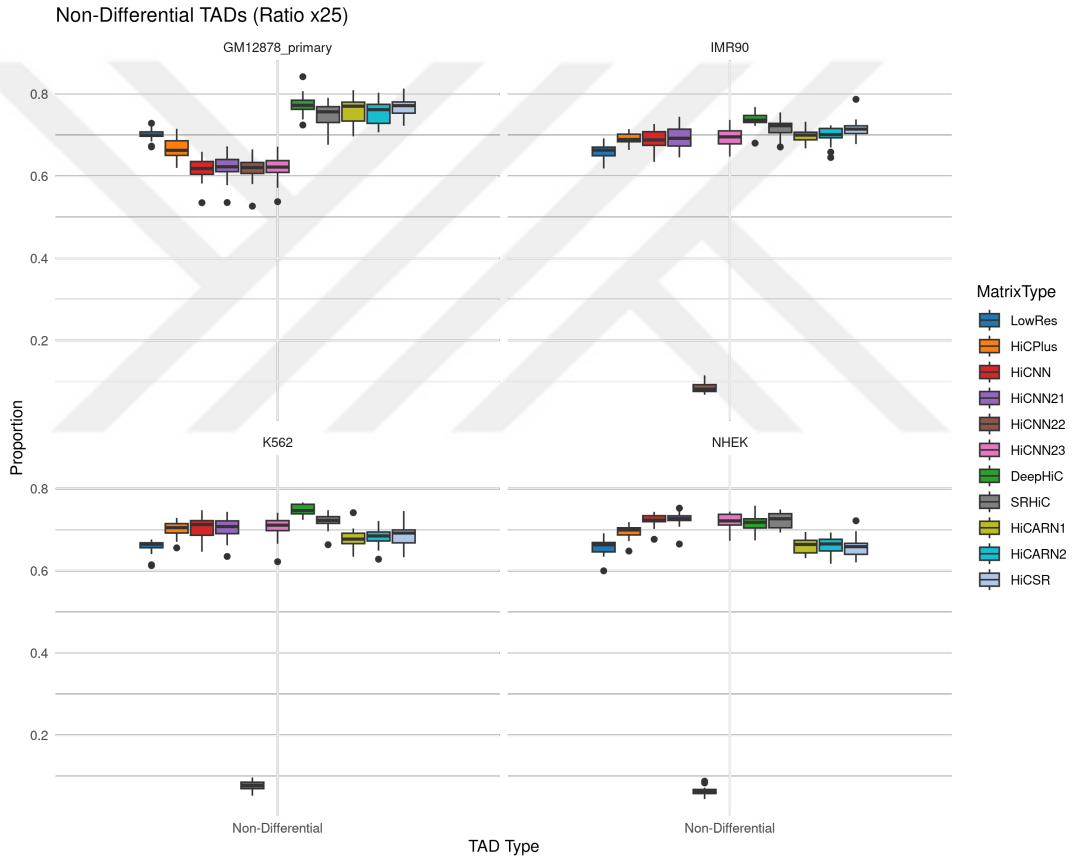


Figure 3.13: Non-differential TADs in all human cell lines at enhancement ratio of x25.

Table 3.2: Pair-wise comparison significance values for Non-differential TADs in GM12878 cell line of x25 ratio.

	LowRes	HiCPlus	HiCNN	HiCNN21	HiCNN22	HiCNN23	DeepHiC	SRHiC	HiCARN1	HiCARN2	HiCSR
LowRes	–	–	–	–	–	–	–	–	–	–	–
HiCPlus	2.426e-06	–	–	–	–	–	–	–	–	–	–
HiCNN	9.505e-13	4.016e-07	–	–	–	–	–	–	–	–	–
HiCNN21	1.901e-12	2.080e-06	ns	–	–	–	–	–	–	–	–
HiCNN22	9.505e-13	2.353e-07	ns	ns	–	–	–	–	–	–	–
HiCNN23	1.901e-12	1.035e-05	ns	ns	ns	–	–	–	–	–	–
DeepHiC	1.901e-12	9.505e-13	9.505e-13	9.505e-13	9.505e-13	9.505e-13	–	–	–	–	–
SRHiC	5.049e-08	9.220e-11	9.505e-13	9.505e-13	9.505e-13	9.505e-13	2.604e-03	–	–	–	–
HiCARN1	6.969e-09	3.802e-12	9.505e-13	9.505e-13	9.505e-13	9.505e-13	ns	ns	–	–	–
HiCARN2	1.984e-09	6.654e-12	9.505e-13	9.505e-13	9.505e-13	9.505e-13	ns	ns	ns	–	–
HiCSR	1.901e-12	9.505e-13	9.505e-13	9.505e-13	9.505e-13	9.505e-13	ns	3.911e-02	ns	ns	–

Overall, the analysis underscores the superior performance of DeepHiC and the HiCARN models, which consistently balanced fine-scale reconstruction and structural fidelity. The zigzag patterns in PSNR and SSIM plots highlight the impact of intrachromosomal variability on model performance, emphasizing the importance of selecting models that are robust across both genomic contexts and varying input sparsity conditions. These findings highlight the potential of advanced deep learning models like DeepHiC and HiCARN to advance genomic research while also identifying opportunities for further optimization to enhance model generalizability and consistency across diverse genomic landscapes.

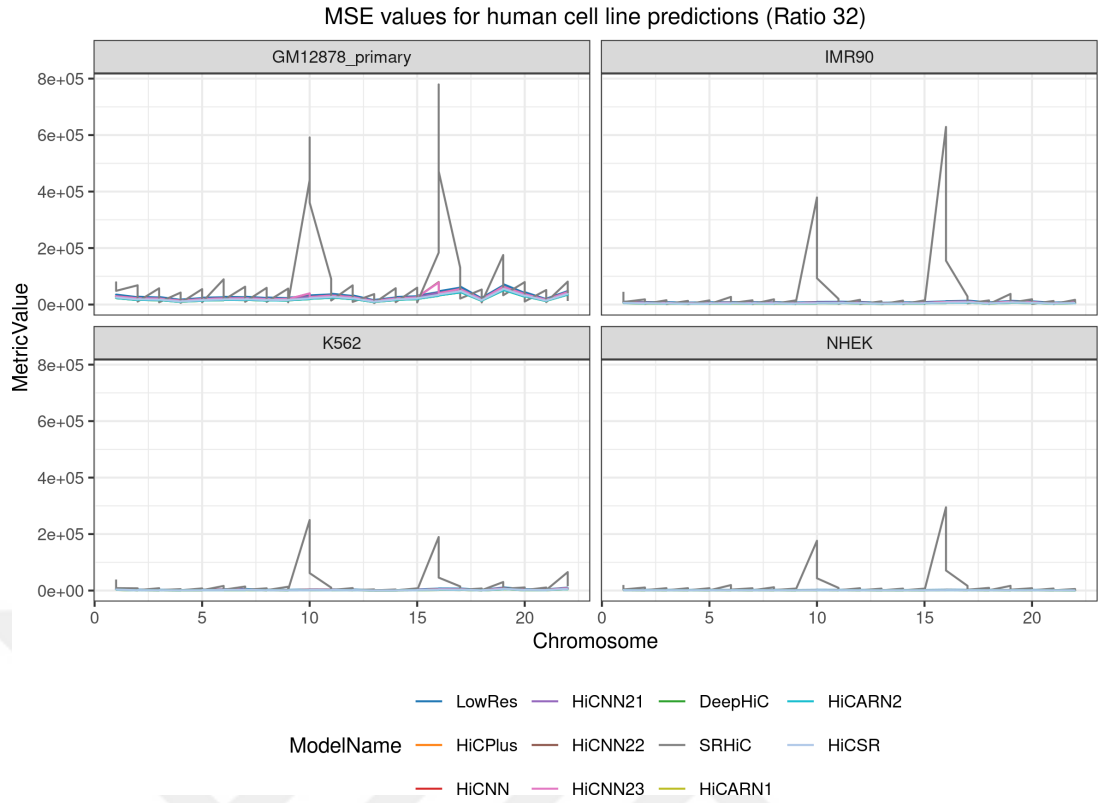
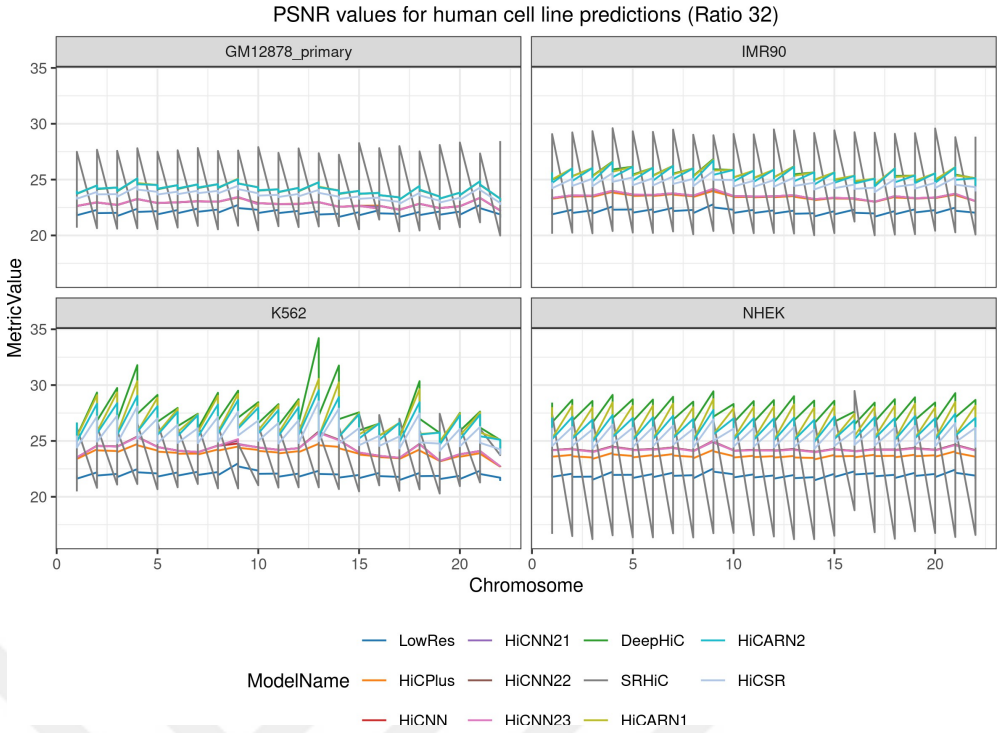
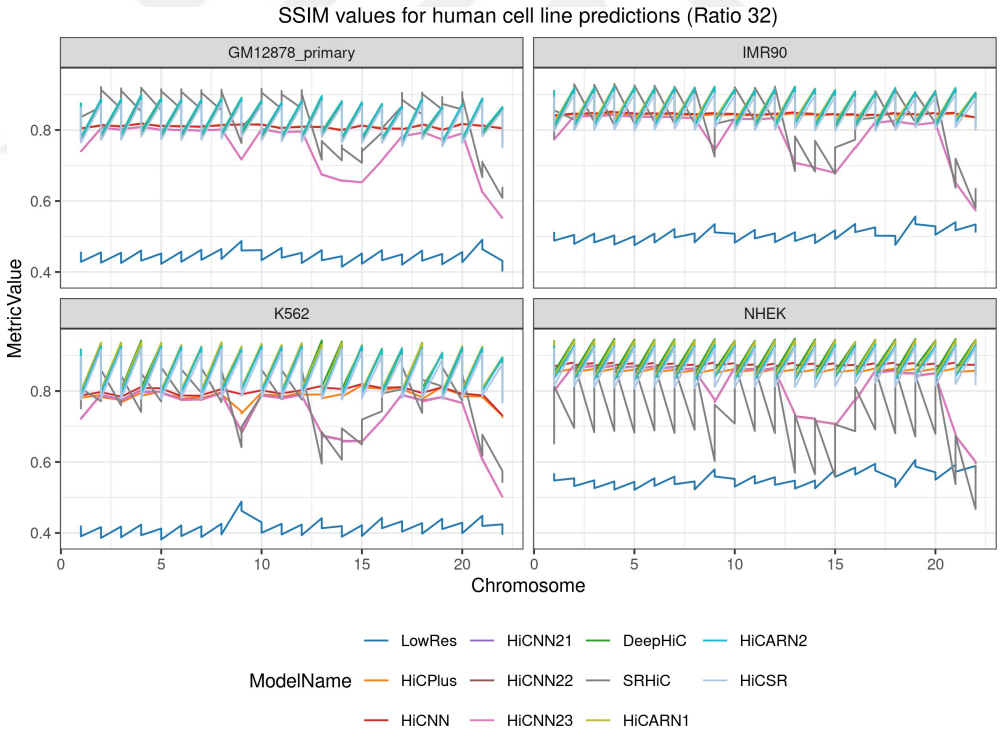


Figure 3.14: MSE values of each model in 4 different human cell lines: GM12878 primary, IMR90, K562 and NHEK. Enhancement ratio is x32.

The evaluation of deep learning models for Hi-C resolution enhancement revealed consistent trends across multiple metrics, including Mean Squared Error (MSE), Structural Similarity Index (SSIM), Peak Signal-to-Noise Ratio (PSNR), and non-differential TAD proportions. Analysis was conducted across three enhancement ratios (x25, x32, and x50) and four human cell lines (GM12878, IMR90, K562, and NHEK), providing a comprehensive perspective on model performance under varying degrees of downsampling.



(a)



(b)

Figure 3.15: (a) PSNR values of each model in 4 different human cell lines: GM12878 primary, IMR90, K562 and NHEK. Enhancement ratio is x32. (b) SSIM values of each model in 4 different human cell lines: GM12878 primary, IMR90, K562 and NHEK. Enhancement ratio is x32.

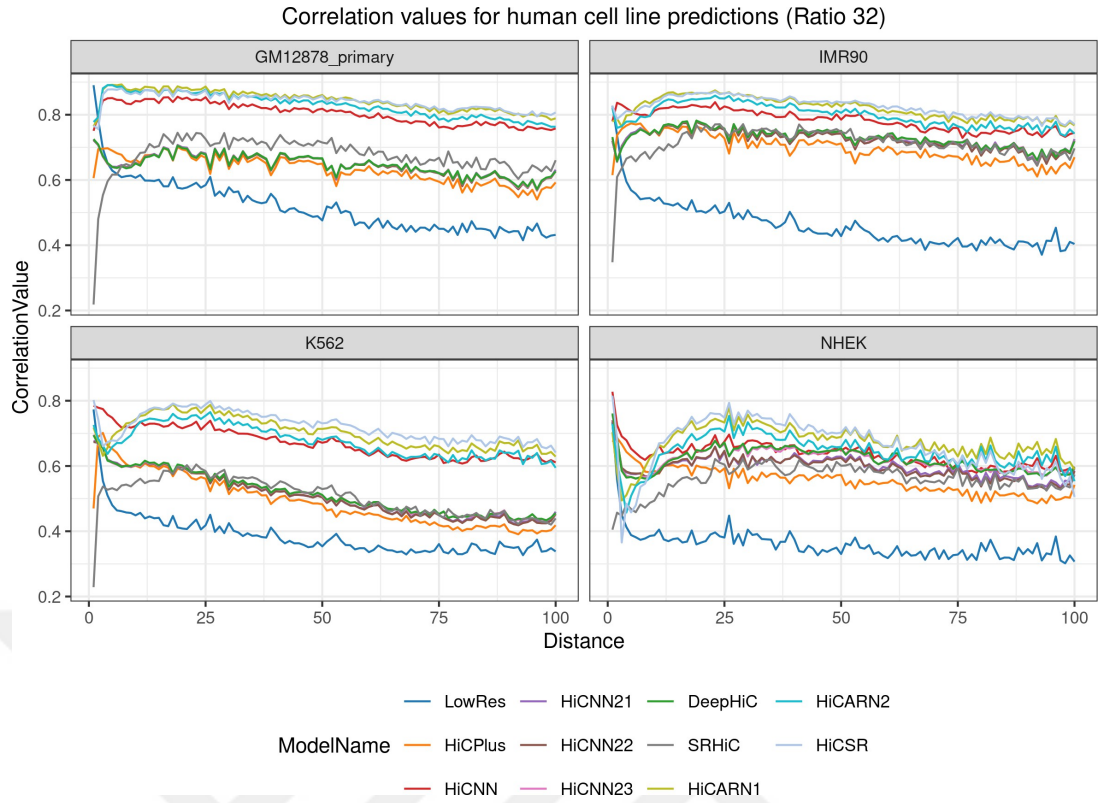


Figure 3.16: Pearson Correlation values of each model in 4 different human cell lines: GM12878 primary, IMR90, K562 and NHEK. Enhancement ratio is x32.

Among the models, DeepHiC consistently emerged as the top performer, achieving a balance between fine-scale reconstruction, structural fidelity, and noise reduction. The HiCARN models (HiCARN1 and HiCARN2) closely followed, demonstrating robust performance across metrics and cell lines. Models like SRHiC and HiCSR revealed specific strengths in preserving large-scale structures or noise reduction, respectively, though they exhibited variability in performance across genomic contexts. Meanwhile, HiCNN-based models and HiCPlus showed limited improvements over the low-resolution baseline.

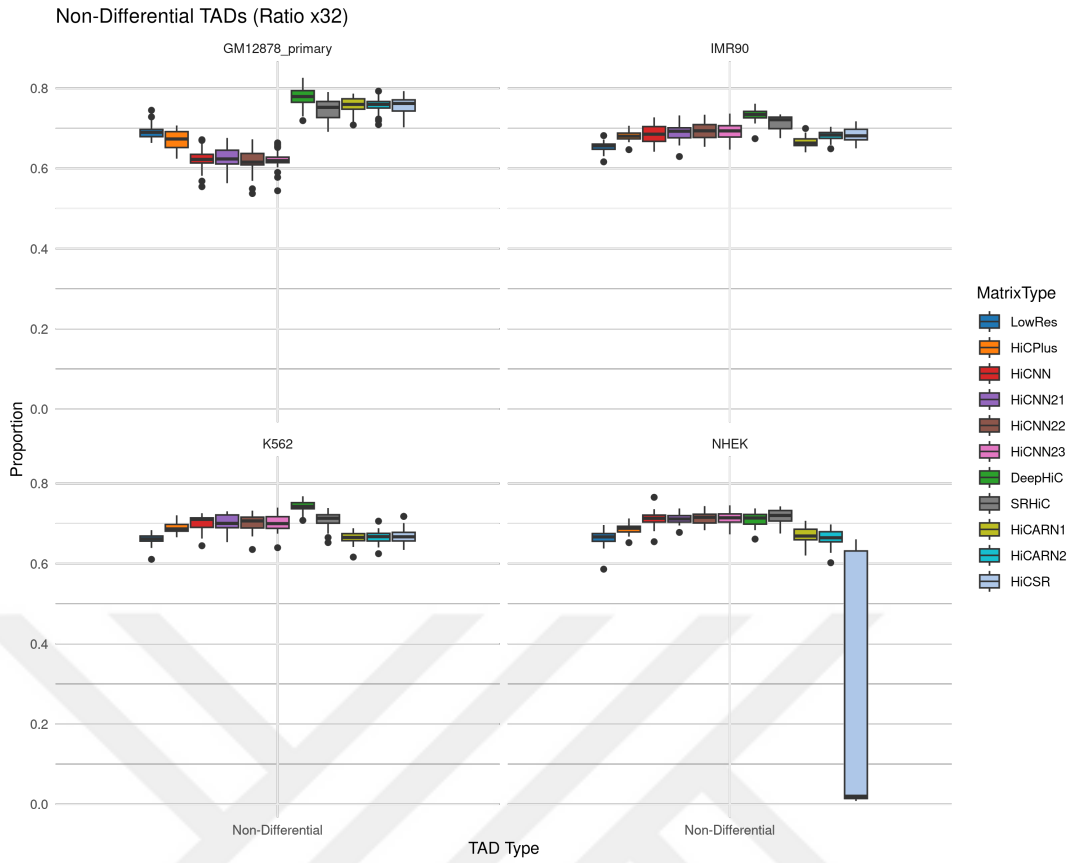


Figure 3.17: Non-differential TADs in all human cell lines at enhancement ratio of x32.

Table 3.3: Pair-wise comparison significance values for Non-differential TADs in GM12878 cell line of x32 ratio.

	LowRes	HiCPlus	HiCNN	HiCNN21	HiCNN22	HiCNN23	DeepHiC	SRHiC	HiCARN1	HiCARN2	HiCSR
LowRes	–	–	–	–	–	–	–	–	–	–	–
HiCPlus	9.874e-03	–	–	–	–	–	–	–	–	–	–
HiCNN	6.369e-11	2.819e-07	–	–	–	–	–	–	–	–	–
HiCNN21	1.141e-11	5.672e-07	ns	–	–	–	–	–	–	–	–
HiCNN22	1.141e-11	1.349e-07	ns	ns	–	–	–	–	–	–	–
HiCNN23	1.901e-12	1.114e-07	ns	ns	ns	–	–	–	–	–	–
DeepHiC	1.141e-11	9.505e-13	9.505e-13	9.505e-13	9.505e-13	9.505e-13	–	–	–	–	–
SRHiC	4.111e-08	6.369e-11	9.505e-13	9.505e-13	9.505e-13	9.505e-13	3.663e-04	–	–	–	–
HiCARN1	3.545e-10	9.505e-13	9.505e-13	9.505e-13	9.505e-13	9.505e-13	3.632e-03	ns	–	–	–
HiCARN2	1.854e-10	9.505e-13	9.505e-13	9.505e-13	9.505e-13	9.505e-13	4.625e-03	ns	ns	–	–
HiCSR	2.585e-10	3.802e-12	9.505e-13	9.505e-13	9.505e-13	9.505e-13	1.223e-02	ns	ns	ns	–

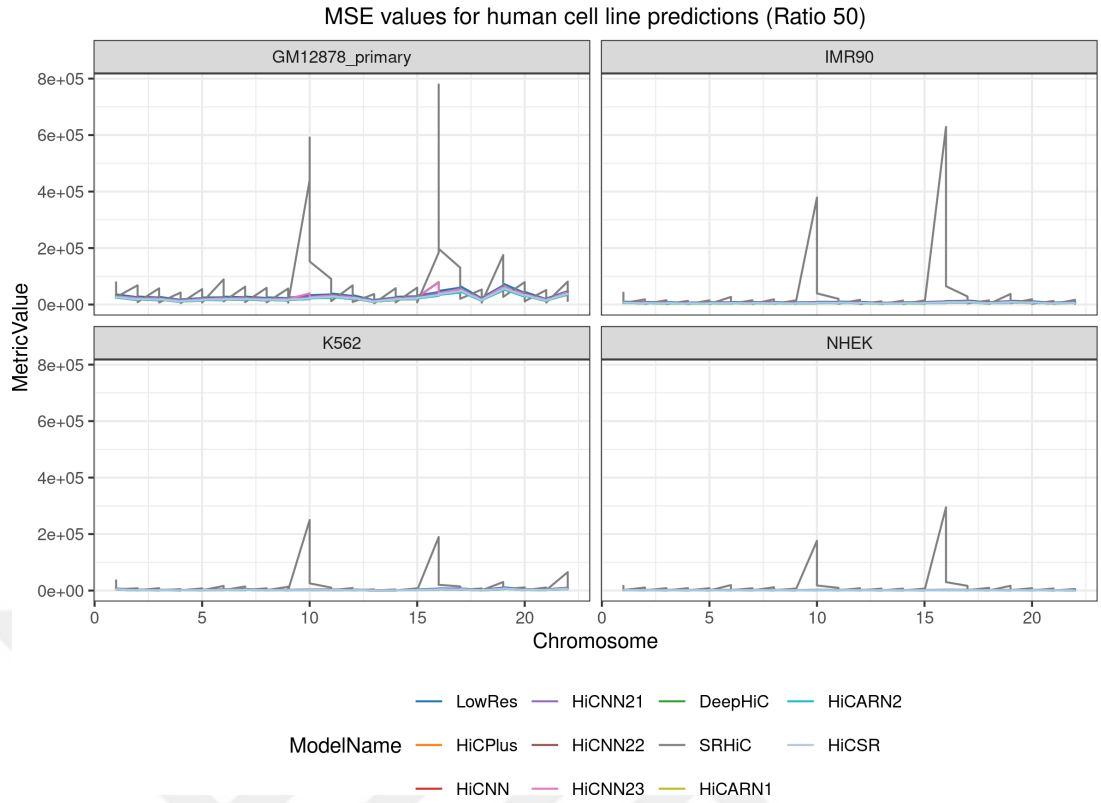
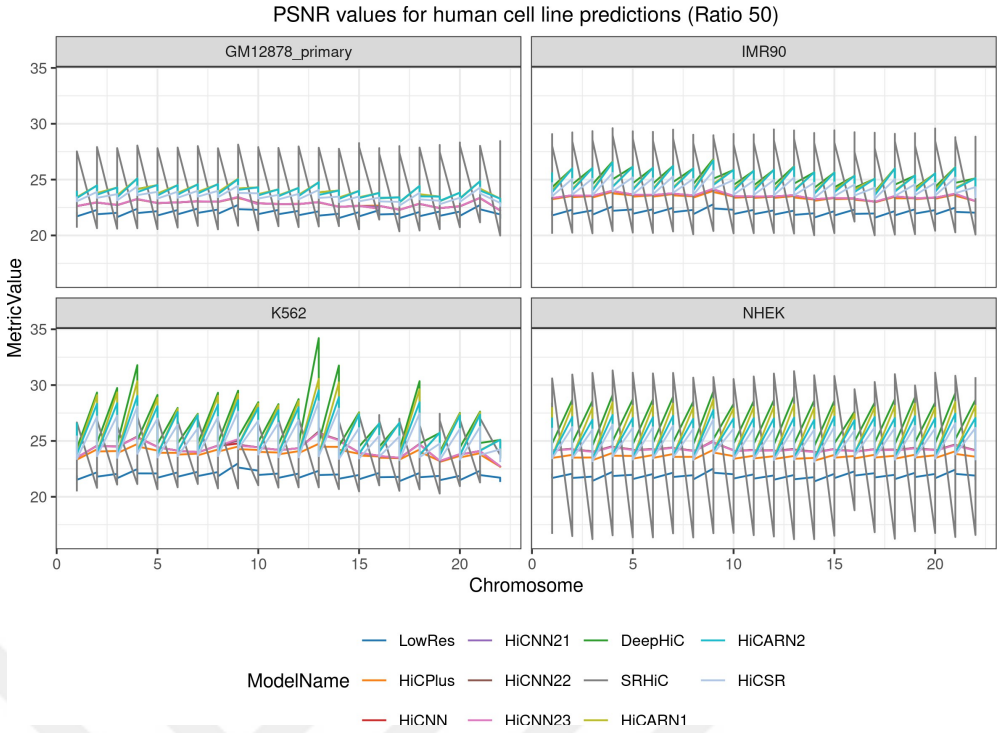
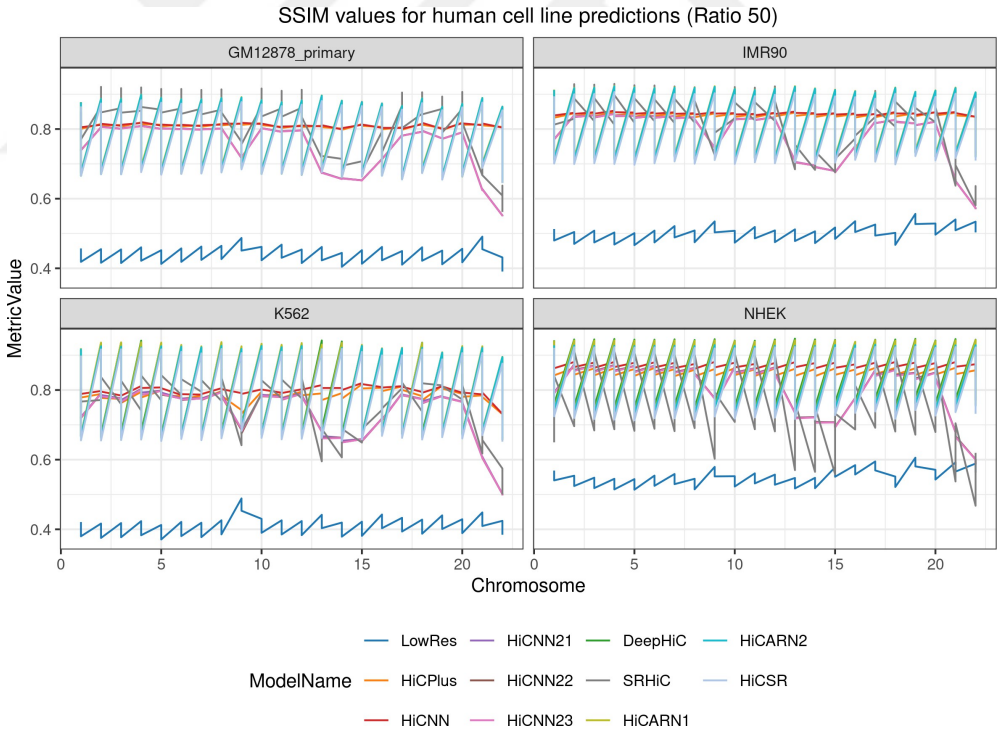


Figure 3.18: MSE values of each model in 4 different human cell lines: GM12878 primary, IMR90, K562 and NHEK. Enhancement ratio is x50.



(a)



(b)

Figure 3.19: (a) PSNR and (b) SSIM values of each model in 4 different human cell lines: GM12878 primary, IMR90, K562 and NHEK. Enhancement ratio is x50.

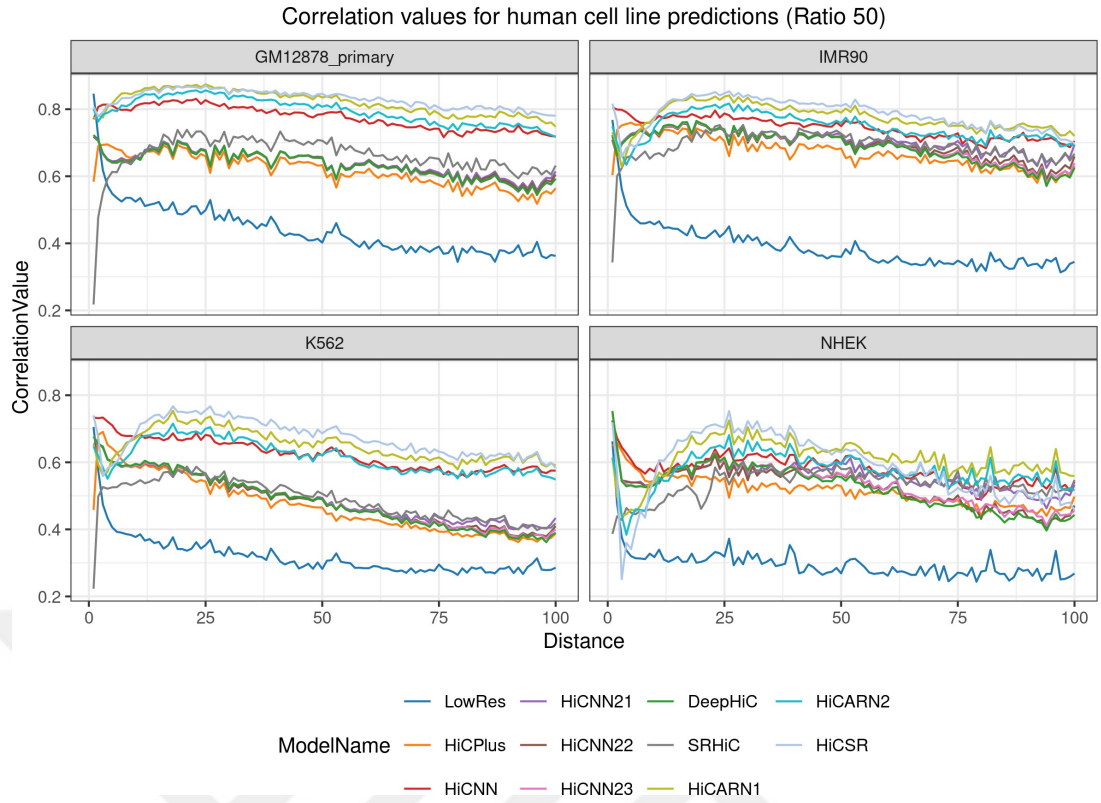


Figure 3.20: Pearson Correlation values of each model in 4 different human cell lines: GM12878 primary, IMR90, K562 and NHEK. Enhancement ratio is x50.

In addition to the primary metrics, a preliminary analysis of differential TAD proportions provided further insights into model performance, categorizing TAD variations into Complex, Merge, Shift, Split, and Strength Change groups. These categories represent diverse structural changes in chromatin, offering an additional lens through which to evaluate model capabilities. The analysis highlighted intriguing trends, such as SRHiC's strong preservation of large-scale TAD structures, reflected in its performance in Merge and Split categories, and the challenges faced by other models in capturing finer-scale changes, such as Strength Changes or Shifts.

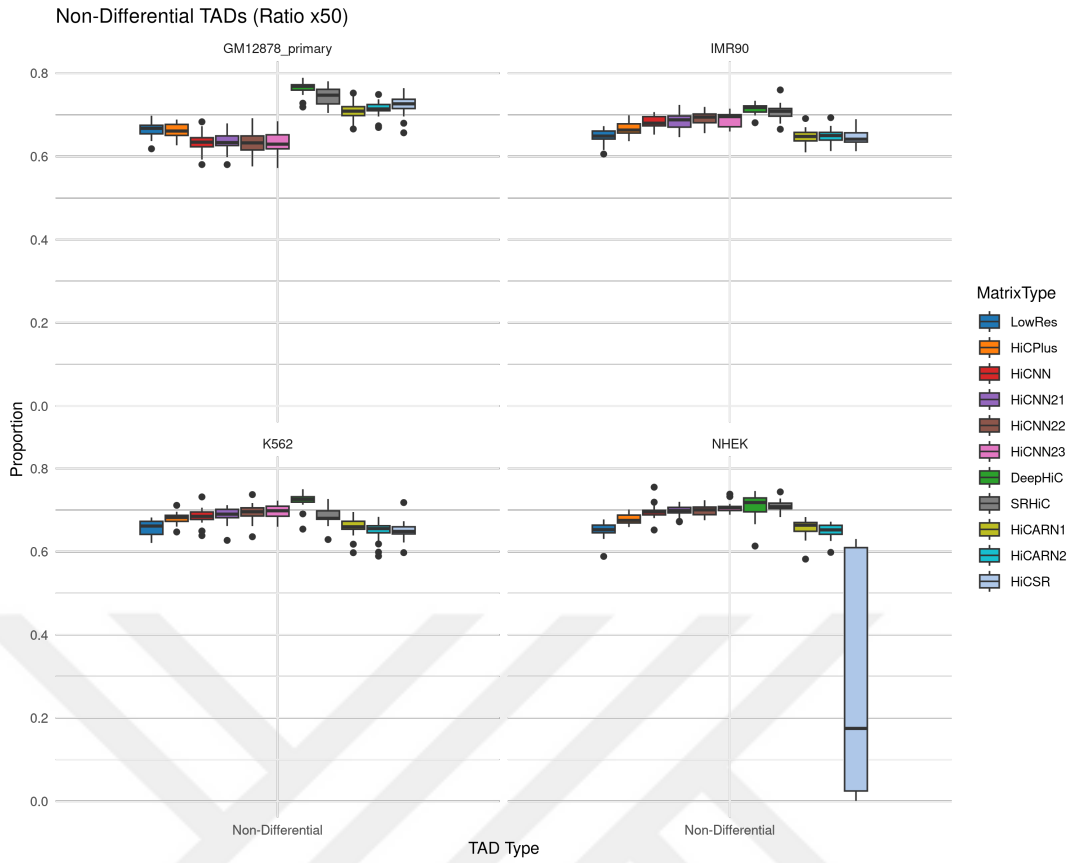


Figure 3.21: Non-differential TADs in all human cell lines at enhancement ratio of x50.

Table 3.4: Pair-wise comparison significance values for Non-differential TADs in GM12878 cell line of x50 ratio.

	LowRes	HiCPlus	HiCNN	HiCNN21	HiCNN22	HiCNN23	DeepHiC	SRHiC	HiCARN1	HiCARN2	HiCSR
LowRes	–	–	–	–	–	–	–	–	–	–	–
HiCPlus	ns	–	–	–	–	–	–	–	–	–	–
HiCNN	3.373e-05	3.596e-04	–	–	–	–	–	–	–	–	–
HiCNN21	4.964e-04	1.150e-03	ns	–	–	–	–	–	–	–	–
HiCNN22	2.680e-04	1.593e-03	ns	ns	–	–	–	–	–	–	–
HiCNN23	7.005e-05	2.871e-04	ns	ns	ns	–	–	–	–	–	–
DeepHiC	9.505e-13	1.438e-08	9.505e-13	9.505e-13	9.505e-13	9.505e-13	–	–	–	–	–
SRHiC	9.505e-13	1.438e-08	9.505e-13	9.505e-13	9.505e-13	9.505e-13	9.771e-04	–	–	–	–
HiCARN1	8.823e-09	1.202e-07	1.806e-11	2.852e-11	4.277e-11	1.141e-11	2.585e-10	2.021e-05	–	–	–
HiCARN2	8.697e-10	1.056e-07	6.654e-12	1.806e-11	1.806e-11	1.141e-11	4.829e-10	1.943e-04	ns	–	–
HiCSR	3.333e-09	1.202e-07	1.806e-11	1.806e-11	4.277e-11	1.806e-11	2.698e-08	5.413e-03	2.890e-02	ns	–

The inclusion of biologically informed loss functions in the current generation of models likely contributed to their superior performance in TAD-related tasks, particularly for non-differential TADs. However, the observed variability in differential TAD subgroup performance suggests room for further refinement. Analyzing these subgroups in greater detail may uncover model-specific trade-offs and guide the development of future enhancements.

For instance, differential TAD categories such as Complex or Strength Change may be particularly informative in understanding how models handle subtle interaction changes, while categories like Split or Merge could shed light on large-scale structural rearrangements. Incorporating these insights into training frameworks, such as by weighting loss functions toward biologically relevant subgroup characteristics, could significantly improve model robustness and generalizability. Additionally, these subgroup analyses may reveal underlying genomic features—such as chromatin accessibility or loop density—that influence model performance and could serve as additional inputs for biologically inspired designs.

Future work in this area could benefit from a systematic exploration of differential TAD categories alongside metrics like SSIM and PSNR. Understanding how models navigate these subgroups could provide a pathway for further improvements, particularly in handling sparse input data or reconstructing regions with complex chromatin organization. By aligning computational objectives with biologically meaningful features, next-generation models may achieve even greater accuracy and structural fidelity, offering new opportunities for high-resolution Hi-C analyses.

The inclusion of biologically informed loss functions in modern deep learning models has significantly enhanced their performance in tasks related to TAD preservation, particularly for non-differential TADs. These loss functions align computational objectives with biological features, enabling models to maintain the structural integrity of chromatin interaction matrices during resolution enhancement. However, the observed variability in performance across differential TAD subgroups—such as Complex, Strength Change, Split, and Merge—highlights opportunities for further refinement. Understanding how models handle these subgroups offers a pathway for optimizing their robustness, generalizability, and alignment with biologically meaningful features.

Differential TAD categories provide unique insights into model performance and reveal trade-offs in how various architectures prioritize structural fidelity. Subgroups such as Complex and Strength Change reflect subtle variations in chromatin interactions, requiring models to accurately capture fine-scale structural changes without introducing noise. Conversely, Split and Merge categories represent large-scale rearrangements, which challenge models to distinguish between true boundary changes and spurious artifacts. The variability in performance across these subgroups suggests that while current loss functions excel in general TAD preservation, they may lack the sensitivity to address subgroup-specific challenges. For instance, models like SRHiC, which prioritize large-scale structural fidelity, perform well in maintaining TAD boundaries but may introduce spurious splits or merges, highlighting the need for targeted optimization.

One promising direction for improvement lies in the development of biologically weighted loss functions that prioritize subgroup-specific characteristics. For example, weighting the loss function to emphasize Strength Change for subtle interaction dynamics or Split/Merge for large-scale structural rearrangements could guide models toward more precise reconstructions. Dynamic loss weighting could also help balance fine-scale reconstruction accuracy with structural fidelity, ensuring models perform consistently across diverse genomic contexts. Additionally, incorporating biologically relevant features, such as chromatin accessibility, loop density, or insulation scores, as inputs to the models could provide critical contextual information, enabling better differentiation between subtle and large-scale structural changes.

Systematic analyses of differential TAD subgroups, alongside metrics such as SSIM and PSNR, could uncover critical trade-offs in model design and training. For example, sparse input data often amplifies challenges in regions with subtle chromatin interactions or complex rearrangements. By examining subgroup-specific performance under different levels of sparsity, researchers could identify architectural features or loss functions that require refinement. Furthermore, analyzing genomic context dependence—such as how gene density, enhancer-promoter loops, or other chromatin features influence performance—could guide the development of next-generation models that are more robust across diverse chromosomal regions and cell types.

To address these challenges, future models could integrate biologically inspired ar-

chitectural features and optimization strategies. Attention mechanisms, such as transformer layers, could dynamically prioritize regions with high structural variability, reducing sensitivity to intrachromosomal heterogeneity. Hybrid architectures that combine GANs with variational autoencoders (VAEs) or other generative models could enhance robustness by balancing fine-scale reconstruction with global structural fidelity. Additionally, designing multi-objective loss functions that explicitly account for subgroup-specific challenges, such as penalties for shifted or missing TAD boundaries, could provide a more comprehensive framework for optimizing model performance.

Ultimately, aligning computational objectives with biologically meaningful features offers a powerful opportunity to advance resolution enhancement models. By incorporating differential TAD subgroup analyses and biologically informed loss functions into training and evaluation frameworks, future models could achieve unprecedented accuracy and structural fidelity. These improvements would enable more effective reconstructions of sparse datasets, improve insights into complex chromatin organization, and expand the applicability of Hi-C analyses to resource-constrained or biologically complex settings. The systematic integration of these features into next-generation models holds immense potential for advancing our understanding of genome architecture and supporting precision genomics research.

3.1.4. 16-Fold Enhancement of Matrices Across Organism: mESC

The performance of deep learning models for Hi-C resolution enhancement was assessed on mouse embryonic stem cells (mESCs) at an x16 enhancement ratio, revealing consistent trends across multiple metrics and reaffirming the generalizability of the top-performing models.

Among the tested models, HiCARN1, HiCARN2, and DeepHiC consistently outperformed their counterparts across all evaluation metrics, including Mean Squared Error (MSE), Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM), and Pearson correlation. HiCARN2 demonstrated particularly strong performance, achieving the lowest MSE (95,221.20), the highest PSNR (25.02), and near-optimal SSIM (0.8664) and Pearson correlation (0.8291). Similarly, HiCARN1 excelled with a comparable MSE of 95,299.47 and the highest SSIM score of 0.8665, reflecting its excep-

tional ability to preserve structural details. DeepHiC, while slightly trailing in individual metrics, consistently ranked among the top three models, showcasing its robust balance between noise reduction and structural fidelity.

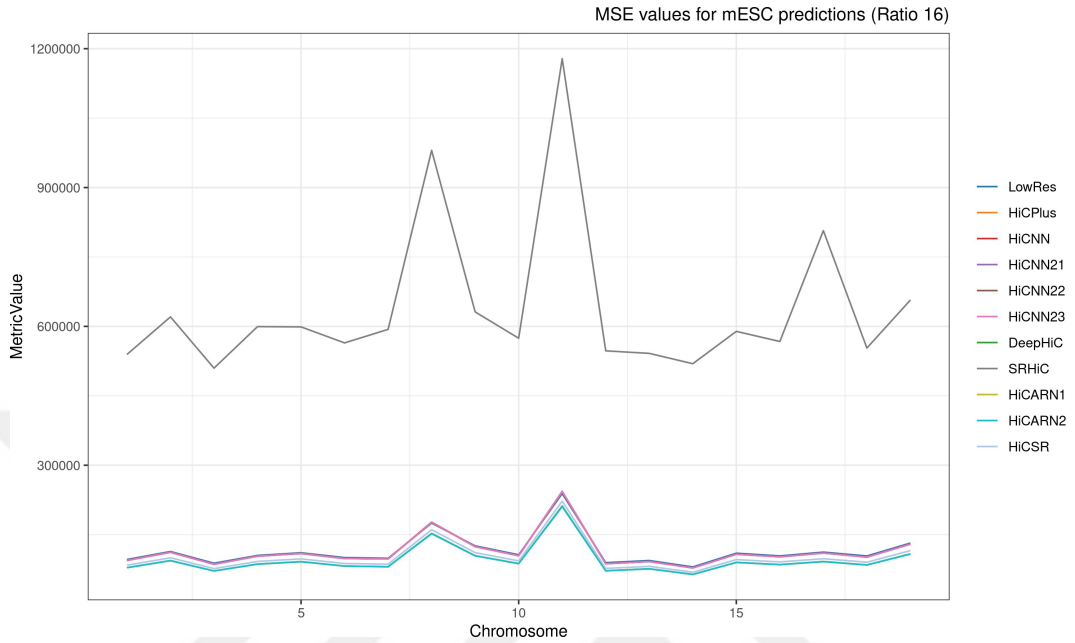
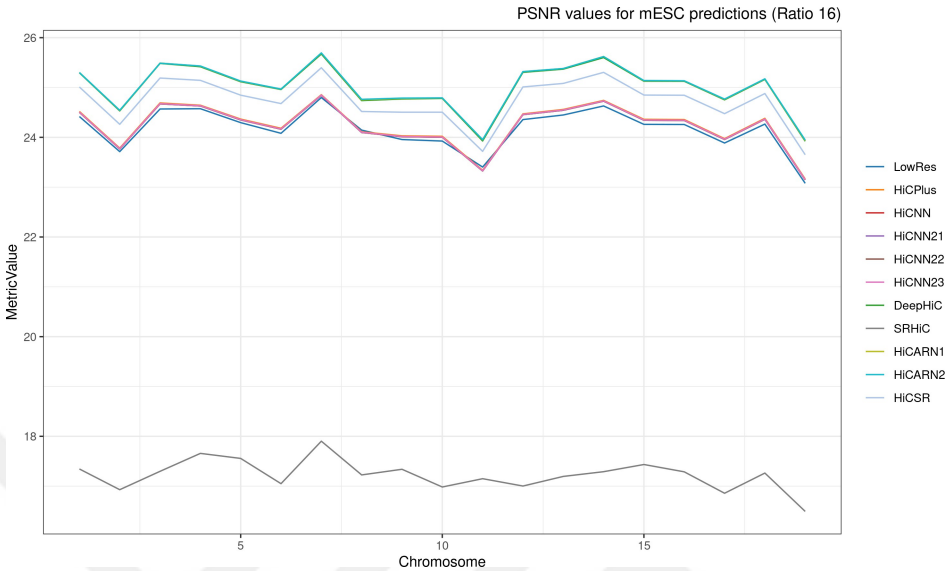


Figure 3.22: MSE values of each model in mouse embryonic stem cells (mESC). Enhancement ratio is x16.

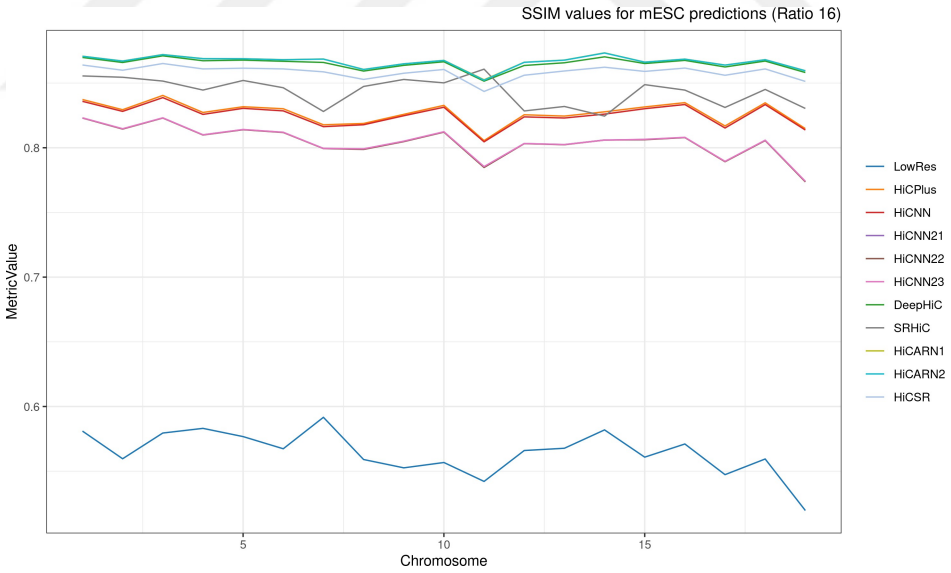
In contrast, HiCSR delivered moderate performance across metrics. With an MSE of 101,575.73 and a PSNR of 24.73, HiCSR outperformed the HiCNN family and HiCPlus but lagged behind the leading models. Its SSIM (0.8585) and Pearson correlation (0.8187) were similarly mid-range, demonstrating its reliability for tasks requiring moderate accuracy and structural preservation. However, its performance indicates room for improvement, particularly in error reduction and noise handling.

The HiCNN-based models and HiCPlus exhibited limited improvements over the low-resolution baseline. With MSE values ranging from 113,302.85 to 113,727.83 and PSNR scores of 24.22–24.24, these models struggled to reduce reconstruction errors and enhance signal quality. Their SSIM values (0.8047–0.8253) further highlighted their limited ability to recover fine structural details, while Pearson correlation scores (0.6112) underscored their inability to maintain strong similarity with high-resolution

data. These results suggest that HiCNN-based models and HiCPlus are less suitable for demanding reconstruction tasks, particularly for sparse input conditions such as mESC matrices.



(a)



(b)

Figure 3.23: (a) PSNR and (b) SSIM values of each model in mouse embryonic stem cells (mESC). Enhancement ratio is x16.

SRHiC presented an intriguing performance profile, with notable strengths in structural preservation but significant weaknesses in error reduction and noise handling. Despite achieving a moderate SSIM of 0.8436 and a Pearson correlation of 0.6435, SRHiC recorded the highest MSE (640,645.01) and the lowest PSNR (17.22). These results indicate that while SRHiC prioritizes large-scale structural preservation, it struggles with fine-scale reconstruction and noise reduction, making it less effective for high-resolution Hi-C analysis in mESC cells.

These results align with trends observed in human cell lines, reinforcing the superior generalizability of HiCARN1, HiCARN2, and DeepHiC across diverse genomic contexts. The ability of these models to excel in minimizing reconstruction errors, preserving structural fidelity, and maintaining strong correlations with high-resolution data highlights their potential for cross-species applications. Furthermore, the limitations observed in models like SRHiC and HiCSR underscore the importance of balancing large-scale structural preservation with fine-scale accuracy, particularly for genomic contexts with complex chromatin organization.

Future work could leverage insights from differential TAD subgroup analysis to further refine model performance for cross-organism applications. Incorporating biologically informed loss functions tailored to the structural characteristics of mESC chromatin may enhance the ability of these models to capture subtle interaction changes, ultimately improving their utility for comparative genomic studies.

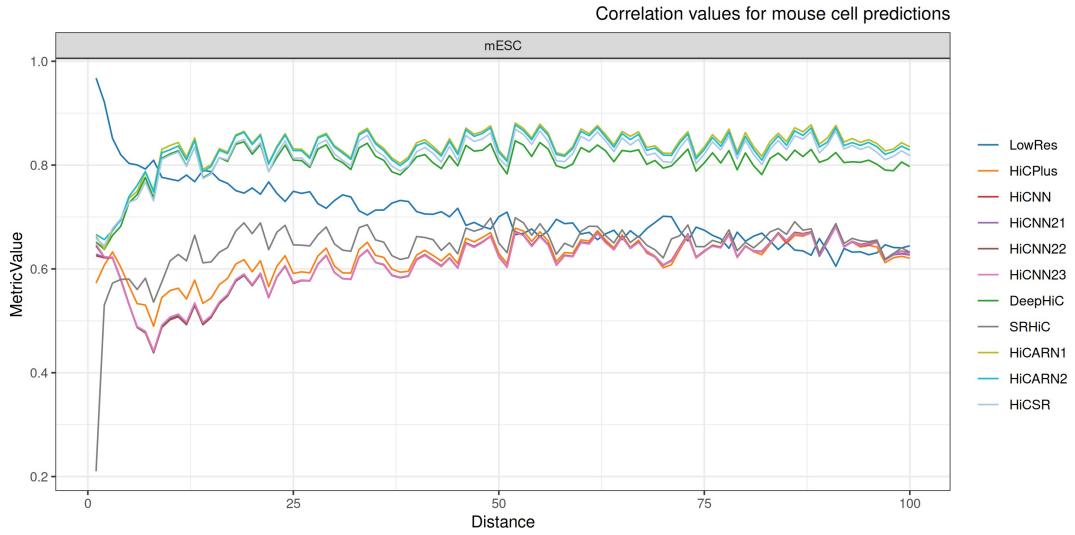
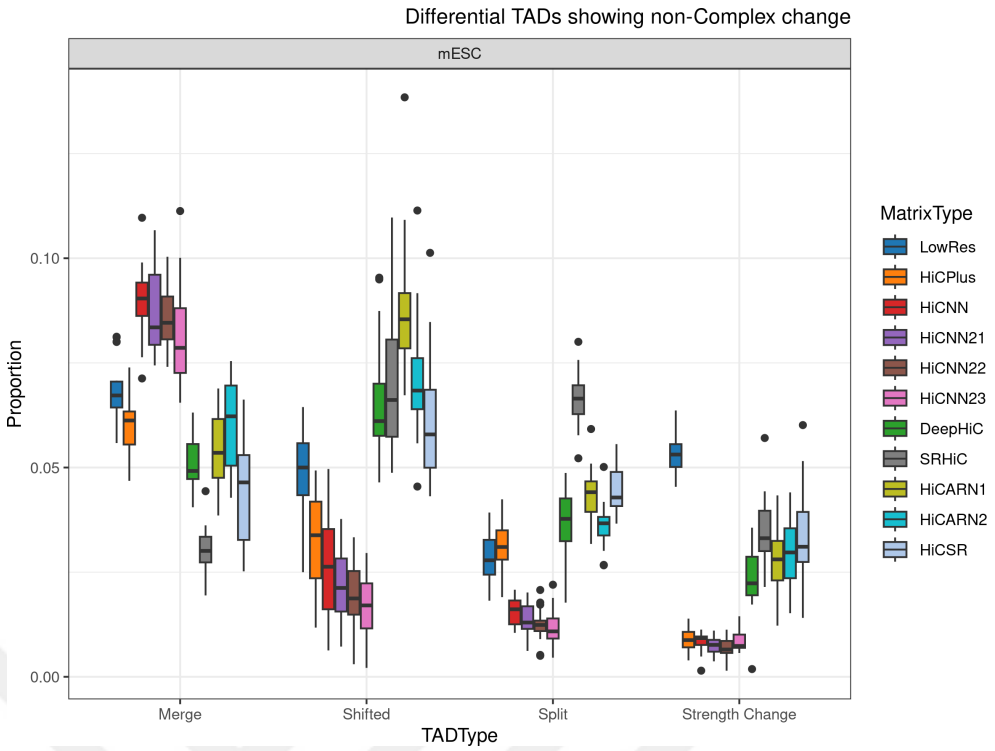
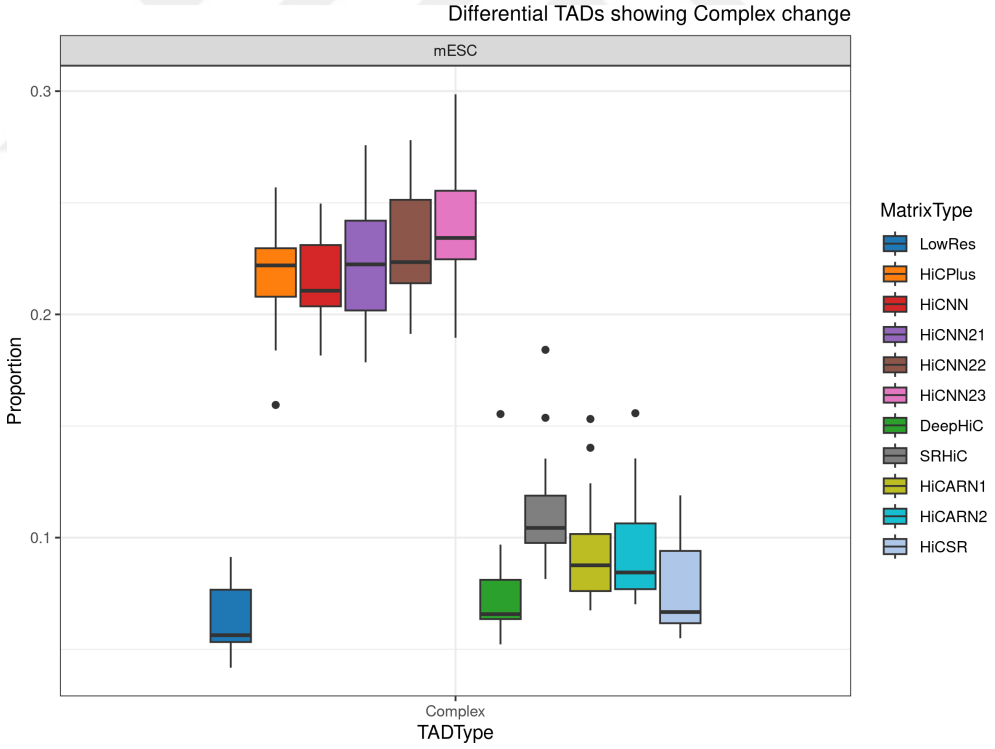


Figure 3.24: Pearson Correlation values of each model in mouse embryonic stem cells (mESC). Enhancement ratio is x16.



(a)



(b)

Figure 3.25: (a) Differential TADs (excluding Complex TADs) and (b) Complex TADs in mESCs at enhancement ratio of x16.

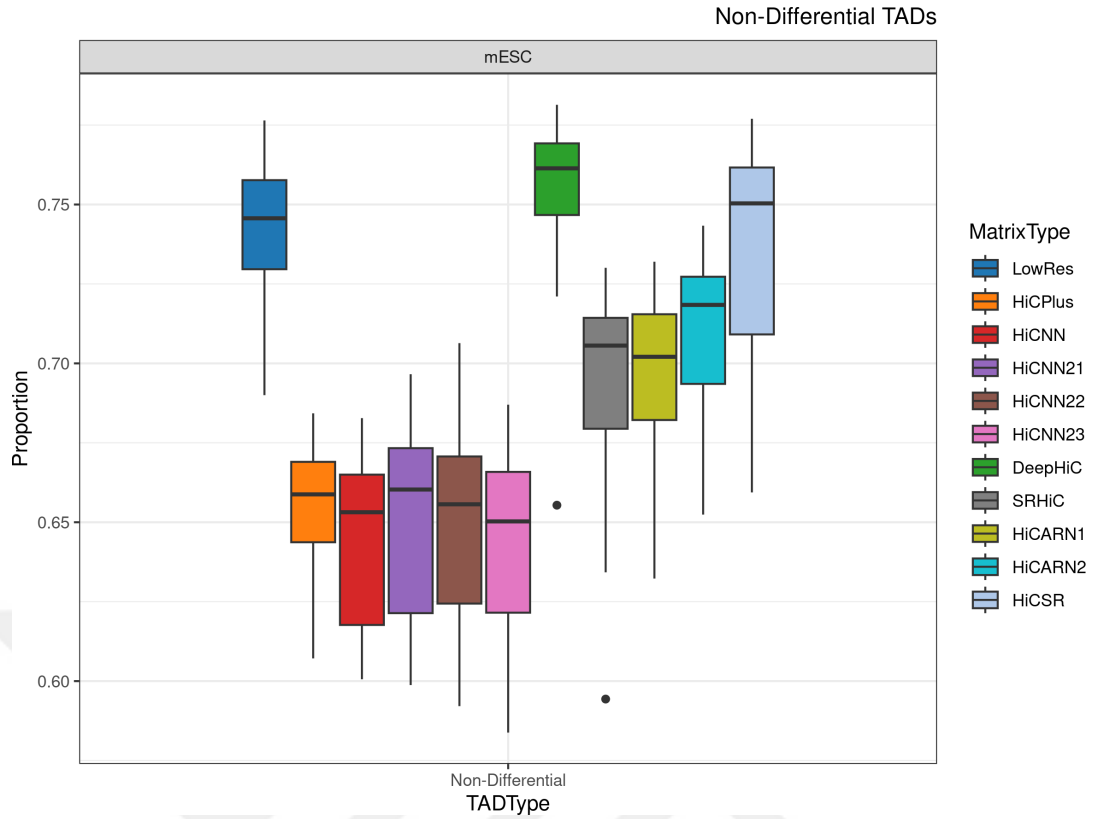


Figure 3.26: Non-differential TADs in mESCs at enhancement ratio of x16.

Table 3.5: Pair-wise comparison significance values for Non-differential TADs in mESCs of x16 ratio.

	LowRes	HiCPlus	HiCNN	HiCNN21	HiCNN22	HiCNN23	DeepHiC	SRHiC	HiCARN1	HiCARN2	HiCSR
LowRes	–	–	–	–	–	–	–	–	–	–	–
HiCPlus	5.658e-11	–	–	–	–	–	–	–	–	–	–
HiCNN	5.658e-11	ns	–	–	–	–	–	–	–	–	–
HiCNN21	1.132e-10	ns	ns	–	–	–	–	–	–	–	–
HiCNN22	2.263e-10	ns	ns	ns	–	–	–	–	–	–	–
HiCNN23	5.658e-11	ns	ns	ns	ns	–	–	–	–	–	–
DeepHiC	ns	1.539e-08	5.489e-09	7.865e-09	7.865e-09	3.791e-09	–	–	–	–	–
SRHiC	4.123e-07	6.159e-04	2.158e-04	3.694e-04	5.433e-04	3.238e-04	9.037e-08	–	–	–	–
HiCARN1	8.206e-07	6.751e-05	2.592e-05	6.751e-05	1.632e-04	6.751e-05	2.543e-07	ns	–	–	–
HiCARN2	ns	9.037e-08	2.875e-08	4.123e-07	5.209e-07	1.181e-07	3.550e-06	ns	1.090e-01	–	–
HiCSR	ns	5.489e-09	2.546e-09	1.103e-08	5.489e-09	2.546e-09	ns	3.238e-04	3.694e-04	3.063e-03	–

3.1.5. Conclusions on Benchmarking of Resolution Enhancement Tools

The analysis of deep learning models for Hi-C resolution enhancement has provided valuable insights into their strengths and limitations. Models such as DeepHiC, HiCARN1, and HiCARN2 demonstrated remarkable performance in reconstructing high-resolution matrices, preserving structural fidelity, and reducing noise. These models excelled in tasks requiring the identification of large-scale chromatin structures, such as TADs and significant chromatin interactions. However, the results also highlighted important caveats, particularly for applications requiring finer-scale analysis.

Enhancement methods inherently optimize for broader structural preservation, often prioritizing the reconstruction of large-scale features over granular accuracy. This bias, while beneficial for tasks like TAD identification or inter-domain interaction analysis, can inadvertently lead to false-positive interactions when the focus shifts to fine-scale resolution. For instance, in cases where Hi-C bins encompass multiple regulatory elements, such as enhancers or promoters, the enhancement process may introduce artificial interactions to align with larger structural patterns. This over-reconstruction could mislead downstream analyses, especially when the goal is to determine the mere existence of an interaction rather than delineating structural features.

Given these considerations, applying enhancement models to Hi-C matrices for cis-regulatory element mapping introduces the risk of biasing results. The task of identifying specific enhancer-promoter interactions demands an approach that prioritizes precision over structural context, as the resolution constraints of Hi-C data already pose significant challenges. Enhancing the matrix resolution in this scenario may amplify existing ambiguities rather than resolve them.

Thus, while the enhancement models evaluated in this work offer significant advantages for large-scale structural analysis, their application in fine-scale interaction mapping may not be appropriate. The next section of this thesis explores an alternative strategy to address these challenges, leveraging existing high-resolution Hi-C data without enhancement to focus on mapping enhancer-promoter interactions. This approach aims to minimize potential biases and provide a more reliable framework for analyzing the complex interplay of cis-regulatory elements within the chromatin landscape.

3.2. Promoter-Enhancer Networks

The promoter-enhancer network analysis identified distinct regulatory patterns among differentially expressed genes (DEGs) during cardiomyocyte differentiation. These patterns were categorized into High Positive, High Negative, Opposite High, Specific, and Synergistic/Competitive groups, each revealing unique transcriptional dynamics. The DEGs within these categories not only highlight diverse modes of regulatory interaction but also validate the methodology by aligning observed patterns with known biological functions.

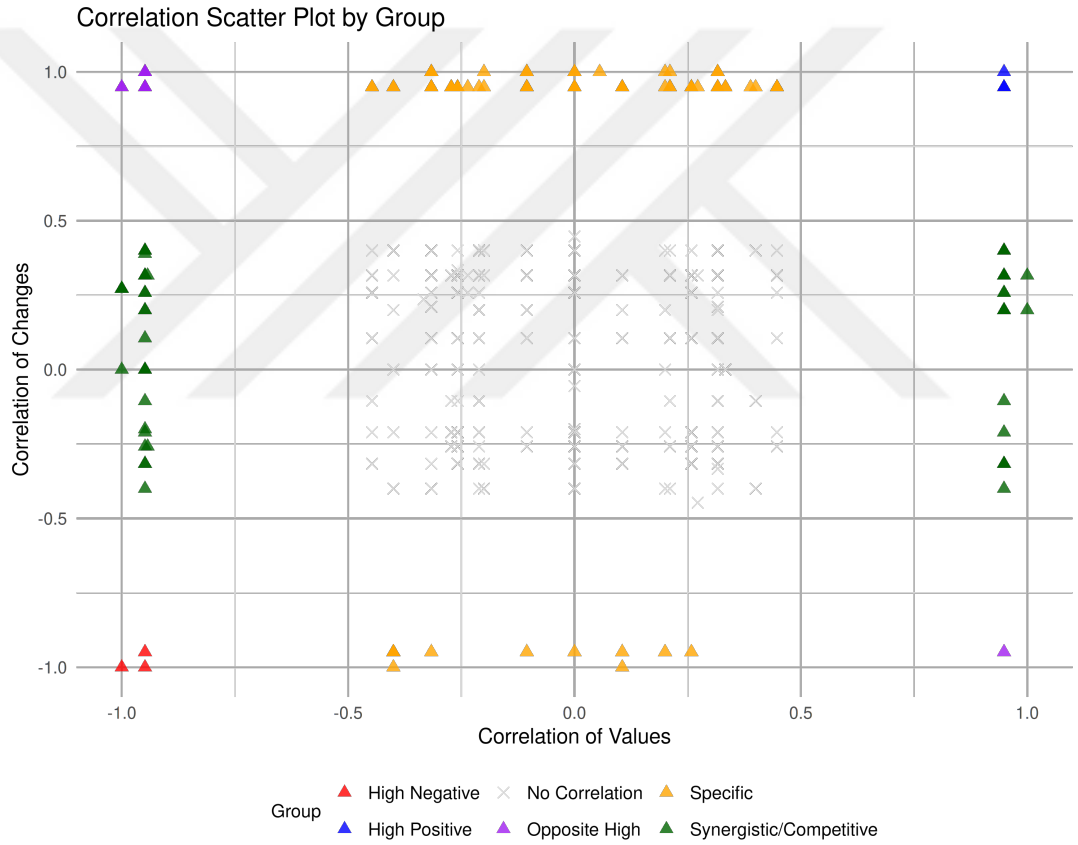


Figure 3.27: Enhancer Mode of Action Groups.

3.2.1. High Positive Group: Balanced Enhancer-Promoter Interactions

1. Correlation Pattern: The "High Positive" group consists of genes with strong positive correlations in both total network connectivity (*corS_edges*) and dynamic net-

work changes (*corS_changes*) relative to gene expression. This regulatory pattern suggests a harmonious relationship where both the static and dynamic aspects of enhancer-promoter interactions amplify gene expression.

2. Genes in the Group:

Table 3.6: Genes in the High Positive Group

Gene Name	*corS_edges*	*corS_changes*	Chromosome	Group
USP39	0.948	0.948	2	High Positive
HMGN4	0.948	0.948	6	High Positive
TRMT1	0.948	1.000	19	High Positive

The group characterized by high correlation in both network connectivity and expression levels was further divided into three subcategories: the high positive group, the high negative group, and the opposite high group. The high positive group includes genes where both correlation values are positive and exceed 0.9, indicating a strong direct relationship between the number of interactions and expression changes. This pattern suggests that enhancer-promoter interactions directly drive gene expression, consistent with models of balanced or additive enhancer activity. Genes in this group include USP39, HMGN4, and TRMT1, each showing a correlation of 0.948 or higher in both static connectivity and network changes. USP39, located on chromosome 2, is a spliceosome component essential for RNA splicing. HMGN4, on chromosome 6, is part of the high-mobility group (HMG) family involved in chromatin remodeling and transcription regulation. TRMT1, on chromosome 19, functions as a tRNA methyltransferase important for RNA stability. Despite these genes' significant regulatory roles, no specific functional enrichments were identified in canonical pathways, suggesting that these genes may be engaged in fundamental cellular processes that are not reflected in conventional pathway databases.

3.2.2. High Negative Group: Repressive Enhancer-Promoter Interactions

1. Correlation Pattern: The "High Negative" group features genes with strong negative correlations in both *corS_edges* and *corS_changes*, indicative of repressive enhancer-promoter dynamics. This pattern suggests a regulatory mechanism where increased connectivity suppresses gene expression.

2. Genes in the Group:

Table 3.7: Genes in the High Negative Group

Gene Name	*corS_edges*	*corS_changes*	Chromosome	Group
METTL16	-0.949	-1.000	17	High Negative
PPM1D	-1.000	-1.000	17	High Negative
UTP6	-0.949	-0.949	17	High Negative

The high negative group, where both correlation values are negative but still exceed 0.9, indicates an inverse relationship between connectivity and expression. This could point to mechanisms where an increased number of interactions represses gene expression, potentially reflecting regulatory feedback loops or repressive chromatin states. Genes in this group include METTL16, PPM1D, and UTP6, all located on chromosome 17. METTL16 is an RNA methyltransferase involved in RNA stability, splicing, and SAM homeostasis, while PPM1D acts as a phosphatase regulating stress response and the cell cycle. UTP6, a processome component, is critical for ribosome biogenesis. Enrichment analysis for this group highlighted 23S rRNA methyltransferase activity, linked to METTL16, and U6 snRNA methyltransferase activity, emphasizing the importance of RNA methylation in splicing regulation. These findings suggest that repressive enhancer-promoter interactions may modulate RNA processing and stability, contributing to tight regulatory control over gene expression.

3.2.3. Opposite High Group: Inverse Dynamics in Regulatory Networks

1. Correlation Pattern: The "Opposite High" group identifies genes with inverse correlations between *corS_edges* and *corS_changes*, revealing unique regulatory mechanisms where static connectivity opposes dynamic network changes.

2. Genes in the Group:

Table 3.8: Genes in the Opposite High Group

Gene Name	Chromosome	*corS_edges*	*corS_changes*	Group
DTNBP1	6	-0.949	0.949	Opposite High
HPS6	10	-0.949	0.949	Opposite High
PTK2	8	-1.000	0.949	Opposite High

Finally, the opposite high group, characterized by correlation values with absolute magnitudes greater than 0.9 but with opposite signs, highlights complex regulatory dynamics. Genes in this group include DTNBP1, HPS6, and PTK2. DTNBP1 and HPS6 are core components of the BLOC1-BLOC2 complex, essential for lysosomal biogenesis and trafficking. PTK2, a tyrosine kinase, plays a significant role in adhesion and signaling pathways. The enrichment results for this group revealed significant associations with the BLOC1-BLOC2 complex, underscoring its importance in lysosomal trafficking and cellular organization. This pattern suggests that static connectivity may inhibit dynamic changes, potentially reflecting a balance between structural and functional regulatory mechanisms.

3.2.4. Specific Group: Dynamic Enhancer Interactions

1. Correlation Pattern: The "Specific" group exhibits strong correlations with *corS_changes* but weak correlations with *corS_edges*, suggesting that dynamic enhancer-promoter interactions, rather than overall connectivity, drive gene expression.

2. Genes in the Group: There are 195 DEGs in this group. Following genes from the Specific group are examples included in the known TF list:

- **ARNT** (Figure3.28): A transcription factor involved in hypoxia signaling and

critical for adaptive responses under low oxygen conditions.

- **BCL11A** (Figure3.29): A zinc finger transcription factor essential for hematopoiesis and neural development.
- **FOSL2** (Figure3.30): An AP-1 transcription factor subunit critical for lineage-specific transcriptional activation.

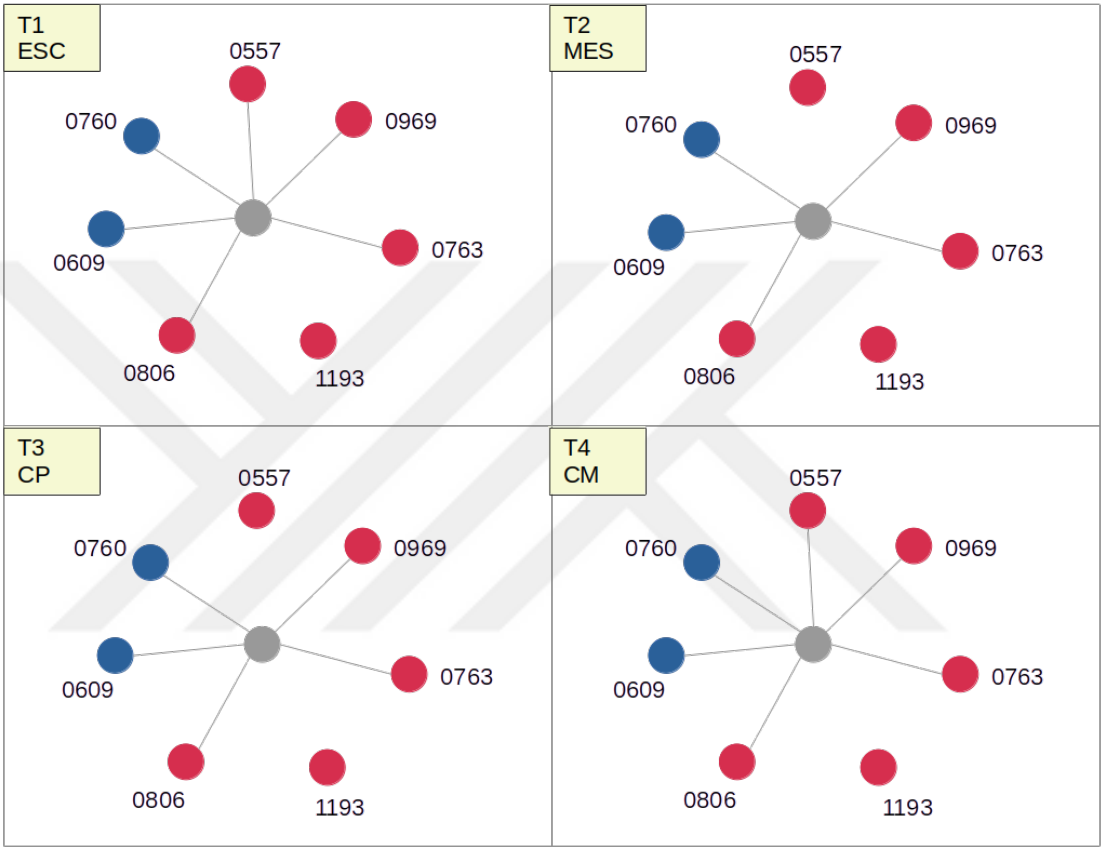


Figure 3.28: Change in Enhancer Network of ARNT (Promoter:GH01J150874) During Cardiomyocyte Differentiation. Elements shown in red are both promoters and enhancers whereas those shown in dark blue are enhancers. Element IDs are shortened in a way to follow common pattern: GH01J15.

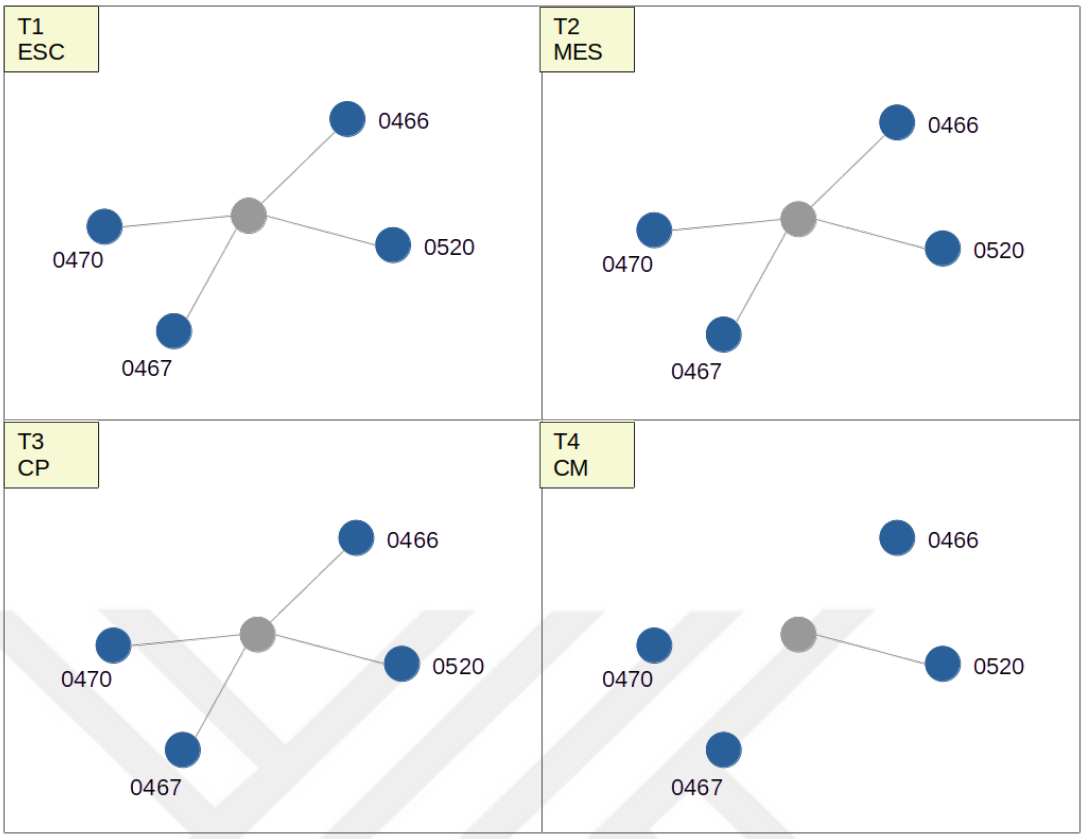


Figure 3.29: Change in Enhancer Network of BCL11A (Promoter:GH02J060543) During Cardiomyocyte Differentiation. Elements shown in dark blue are enhancers. Ele- ment IDs are shortened in a way to follow common pattern: GH02J06.

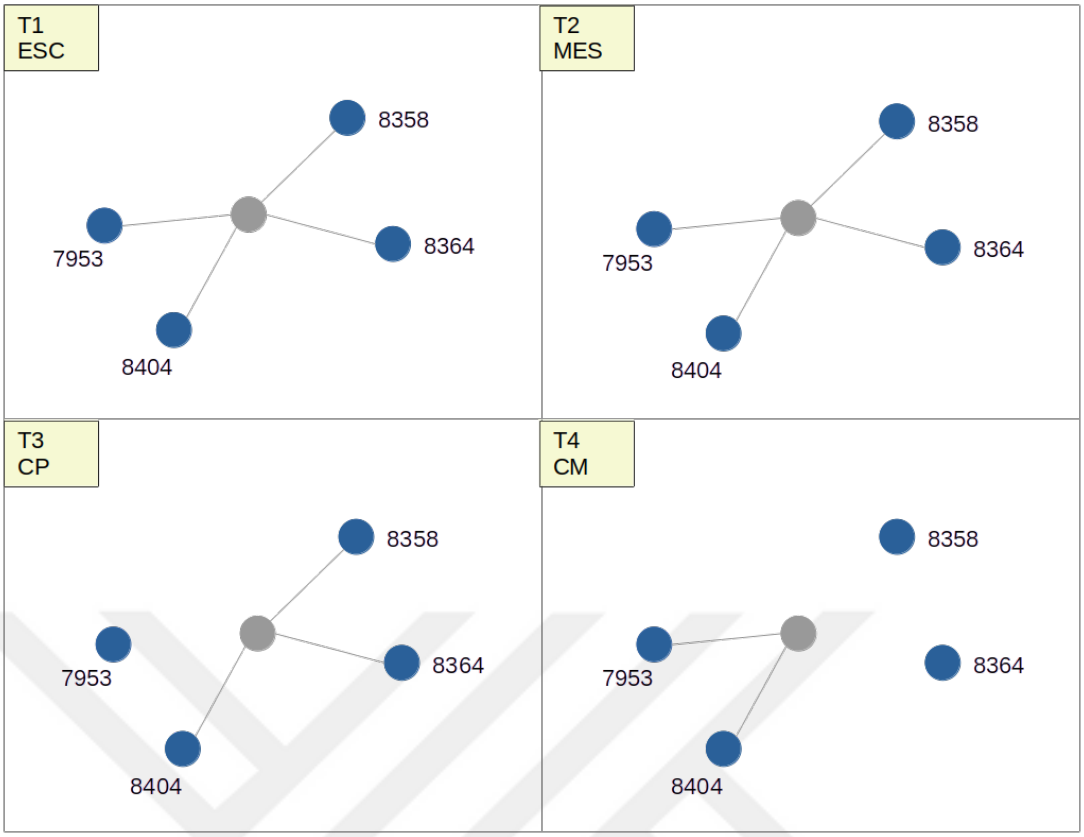


Figure 3.30: Change in Enhancer Network of FOSL2 (Promoter:GH02J028389) During Cardiomyocyte Differentiation. Elements shown in dark blue are enhancers. Ele- ment IDs are shortened in a way to follow common pattern: GH02J02.

One of the major findings in our study was the dominant enrichment of E2F transcription factor family motifs (Figure3.31), specifically E2F-1, E2F-2, E2F-3, and E2F-4. This result aligns with existing literature emphasizing the pivotal role of E2F transcription factors in regulating the G1/S transition of the cell cycle and their influence on the balance between proliferation and differentiation [52, 53]. E2F factors are known to dynamically interact with chromatin remodeling complexes, modulating enhancer activity in a context-specific manner. This is consistent with our observation of high network changes over time, suggesting that temporal and dynamic regulatory interactions are critical for cardiomyocyte differentiation.

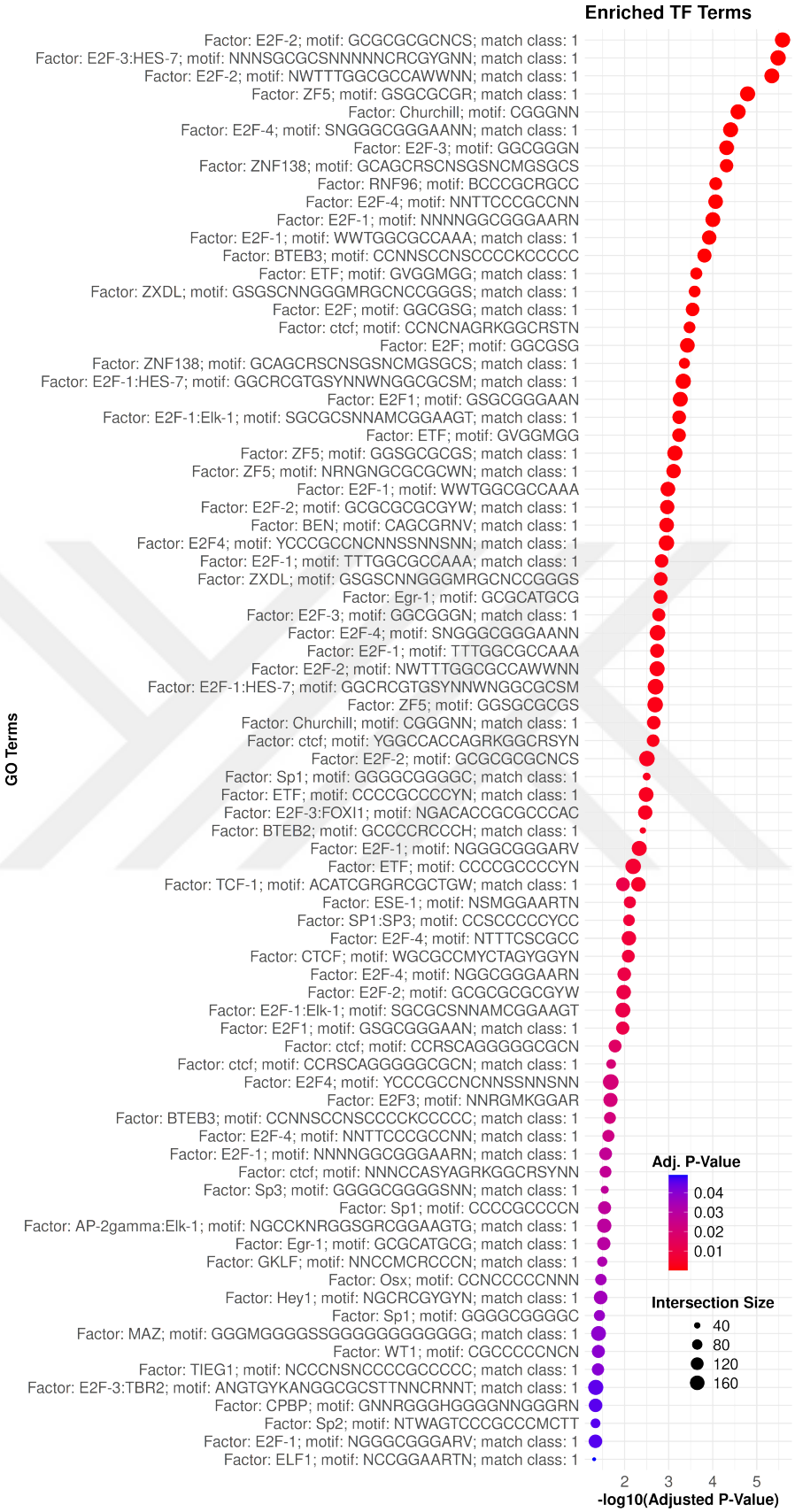


Figure 3.31: Enrichment of 195 Genes in the Specific Group Using TRANSFAC Database.

Furthermore, the enrichment analysis revealed CTCF motifs, supporting its established role as a chromatin insulator that facilitates the formation of regulatory loops [48]. CTCF's presence in the network underscores the importance of three-dimensional chromatin architecture in orchestrating stage-specific enhancer-promoter interactions. These findings validate the utility of incorporating Hi-C data into our methodology to refine the regulatory networks.

While the enrichment of E2F transcription factors was expected and logical given their regulatory role, it is notable that we did not detect the motifs of Nkx2.5, GATA4, Isl1, TBX5, or MEF2C, which are key transcription factors in cardiomyocyte differentiation [45, 46, 49].

The absence of Nkx2.5, Isl1, GATA4, TBX5, and MEF2C in the dynamic enhancer-promoter networks identified in this analysis can be attributed to their stage-specific activity, stable regulatory interactions, and indirect modes of action. Nkx2.5 and Isl1 undergo expression changes during cardiomyocyte differentiation, yet no high-confidence regulatory elements or interactions were associated with them. This suggests that their regulatory elements were either not well-annotated in the GeneHancer database or that they interact with promoters through alternative mechanisms such as distal enhancers, co-factors, or post-transcriptional modifications. GATA4, although detected, did not exhibit dynamic changes in its enhancer-promoter network, indicating that it may function through stable interactions rather than dynamic rewiring. Unlike early-stage regulators such as E2F, which exhibit high enhancer-promoter turnover, GATA4 might already be engaged in its regulatory role at the outset of differentiation, maintaining consistent expression patterns rather than initiating new regulatory events. TBX5 was present in the analysis but had only two associated elite enhancers, meaning its regulatory network was too sparse to provide meaningful insight. This suggests that its regulatory landscape might be incomplete in our dataset or that it interacts with enhancers in a non-canonical way that was not captured by current filtering criteria. MEF2C did not exhibit a strong correlation between enhancer-promoter network changes and expression, implying that its regulation is complex or indirect. Possible explanations include multi-layered regulation via signaling pathways, co-factor interactions, or non-linear relationships between enhancer connectivity and transcriptional output.

From a technical perspective, several factors may have limited the ability of our anal-

ysis to detect the regulatory contributions of these transcription factors. The resolution of our Hi-C data at 10kb may have been insufficient to capture fine-scale interactions, particularly for GATA4 and MEF2C, which might rely on short-range promoter-proximal interactions. Enhancing Hi-C resolution to 5kb or 1kb could improve detection. Similarly, if the binding sites of these transcription factors are within regions of low chromatin accessibility or compact chromatin states, they may not have been detected by ATAC-seq, potentially affecting MEF2C, which may rely on chromatin state changes rather than enhancer activation alone. Motif enrichment analysis might also have been a limiting factor since transcription factors that bind indirectly via co-factors or non-canonical motifs would not be captured. This could particularly impact Nkx2.5 and Isl1, which may require partner proteins for functional activity. Additionally, the differentiation stages sampled in our study may not have aligned with the peak activity windows of these transcription factors, potentially obscuring their regulatory roles. GATA4, for example, may only show network changes at later stages of differentiation beyond the scope of our data. Furthermore, the emphasis on dynamic enhancer interactions may have biased the analysis against transcription factors that regulate gene expression through stable or indirect mechanisms, as evidenced by the case of GATA4, whose interactions were present but did not show rewiring.

To better capture the roles of these transcription factors, several improvements to our approach can be considered. Enhancing Hi-C and ATAC-seq resolution could help detect fine-scale interactions. Incorporating ChIP-seq data specific to these transcription factors would provide direct evidence of their binding sites and confirm whether they participate in enhancer-promoter interactions. Expanding the temporal scope by including later differentiation stages could capture GATA4's delayed regulatory activity and provide insights into MEF2C's potential non-linear expression patterns. A complementary promoter-centric approach, rather than focusing solely on enhancer-driven regulation, could help detect stable transcription factor associations, which may be particularly relevant for GATA4 and TBX5. Adjusting criteria to include near-elite or distal enhancer interactions may help capture weakly annotated transcription factor networks. Investigating super-enhancer networks could also provide insight into whether Nkx2.5, GATA4, and Isl1 are preferentially associated with specialized enhancer clusters rather than dynamic regulatory regions. By integrating these approaches, we can

refine our understanding of how these transcription factors regulate cardiomyocyte differentiation and identify potential regulatory mechanisms that were not captured in the initial analysis.

The significant enrichment of biological processes such as regulation of RNA metabolic processes and DNA-templated transcription aligns with the necessity of robust transcriptional control during differentiation [47]. Additionally, the pathways related to reactive oxygen species (ROS) production and ethanol metabolism through the CYP2E1 pathway suggest that oxidative stress may play a regulatory role in modulating chromatin accessibility and enhancer activity, as highlighted in prior studies [54].

The identification of miRNAs such as hsa-let-7f-1-3p and hsa-miR-381-3p further validates the biological relevance of our results. These miRNAs are known to regulate differentiation and stress responses post-transcriptionally, complementing the enhancer-mediated transcriptional regulation observed in this study [55]. This highlights a multilayered regulatory mechanism where dynamic enhancer-promoter interactions are tightly coupled with miRNA-mediated fine-tuning of gene expression.

3.2.5. Synergistic/Competitive Group: Context-Dependent Regulation

1. Correlation Pattern: The "Synergistic/Competitive" group includes genes with one strong positive and one weak or moderately negative correlation between total network connectivity (*corS_edges*) and dynamic network changes (*corS_changes*) relative to gene expression. This pattern suggests a regulatory mechanism combining cooperative and competitive enhancer-promoter behaviors.

2. Genes in the Group:

The group characterized by high correlation in the number of edges but low correlation in network changes represents genes with relatively static network structures that strongly correlate with expression levels. This suggests that the overall connectivity of enhancer-promoter interactions, rather than dynamic changes, plays a critical role in regulating these genes. Genes in this group include ATF6, CASQ2, DISP1, GBP3, and KHDRBS1, all of which exhibited strong correlations between their static connec- tivity and gene expression. The enrichment analysis for this group revealed significant molecular and biological processes. The identification of "Inositol 1,4,5-trisphosphate-

Table 3.9: Genes in the Synergistic/Competitive Group

Gene Name	*corS_edges*	*corS_changes*	Chromosome	Group
ATF6	-0.949	0.316	1	Synergistic/Competitive
CASQ2	-0.943	-0.258	1	Synergistic/Competitive
DISP1	-0.949	0.316	1	Synergistic/Competitive
GBP3	-1.000	0.272	1	Synergistic/Competitive
KHDRBS1	0.949	0.316	1	Synergistic/Competitive

sensitive calcium-release channel activity" highlights the potential role of calcium signaling in modulating cellular functions, which is critical for cardiomyocyte physiology. The enrichment of protein-protein interaction-related functions suggests that these genes might play roles in maintaining structural and signaling integrity within cells. Pathway enrichment further identified adherens junctions as essential for cell-cell adhesion, which is critical during tissue formation and cardiomyocyte differentiation, and iron homeostasis, emphasizing roles in iron metabolism and regulation. The involvement in iron metabolism and regulation suggests a role in maintaining cellular redox balance, crucial for both differentiation and function of cardiomyocytes.

These results align with literature describing how stable regulatory networks contribute to maintaining essential cellular processes during differentiation. For instance, the enrichment in adherens junction pathways correlates with the requirement for strong intercellular connections in developing cardiomyocytes, while iron homeostasis is vital for energy production and mitigating oxidative stress [48]. The observed static network structures may represent robust, evolutionarily conserved mechanisms essential for cardiomyocyte development and function.

The findings from the enhancer-promoter network analysis in cardiac differentiation exemplify a bottom-up approach, where regulatory interactions are traced from enhancers and promoters to their associated transcription factors (TFs). This method highlights how specific regulatory elements orchestrate gene expression changes during cardiomyocyte differentiation. In contrast, the analysis of deregulated transcriptional protein complexes in aging T cells employs a top-down approach, beginning with TF complexes and examining their regulatory roles within the broader transcriptional

landscape. While these analyses were conducted independently, their complementary perspectives—one progressing from regulatory regions to TFs and the other from TF complexes to their regulatory effects—underscore the potential for future studies to bridge these distinct regulatory frameworks, revealing shared mechanisms in transcriptional regulation. In the next section, I implement the top-down analysis of deregulated protein complexes, providing insights into transcriptional regulation in aging T cells.

3.3. Deregulated Transcriptional Protein Complexes

3.3.1. Comparison of scRNA-seq and Bulk RNA-seq for Complex Dereg- ulation Analysis

The transcriptional protein complexes in classical and nonclassical monocytes were analyzed across four datasets: three scRNA-seq datasets (Yoshida22 [87], Ahern22 [88], Thomson23 [74]) and one bulk RNA-seq dataset (Williams17 [89]). This analysis aimed to identify differences in the number and composition of protein complexes between the two monocyte subsets, and to evaluate the capabilities of scRNA-seq versus bulk RNA-seq in capturing cell-type-specific transcriptional regulation.

Differences Between Classical and Nonclassical Monocytes

For all three scRNA-seq datasets, a significant decrease in the number of protein complexes was observed in nonclassical monocytes compared to classical monocytes. This trend was consistent across datasets:

- **Yoshida22:** Classical monocytes showed a significantly higher number of pre- dicted complexes compared to nonclassical monocytes in children aged newborn to 14 years ($p < 0.05$).
- **Ahern22:** A similar reduction in complex numbers was observed in nonclassical monocytes compared to classical monocytes in children aged 1–14 years ($p < 0.001$).
- **Thomson23:** The dataset for children aged 11–13 years also demonstrated fewer predicted complexes in nonclassical monocytes ($p < 0.001$).

In contrast, the bulk RNA-seq dataset (Williams17), which analyzed mixed populations

of classical and nonclassical monocytes, did not show a significant difference in the number of predicted complexes between the two subsets. This highlights a limitation of bulk RNA-seq, where the averaging effect across cell populations may obscure cell-type-specific transcriptional differences that are evident in single-cell analyses.

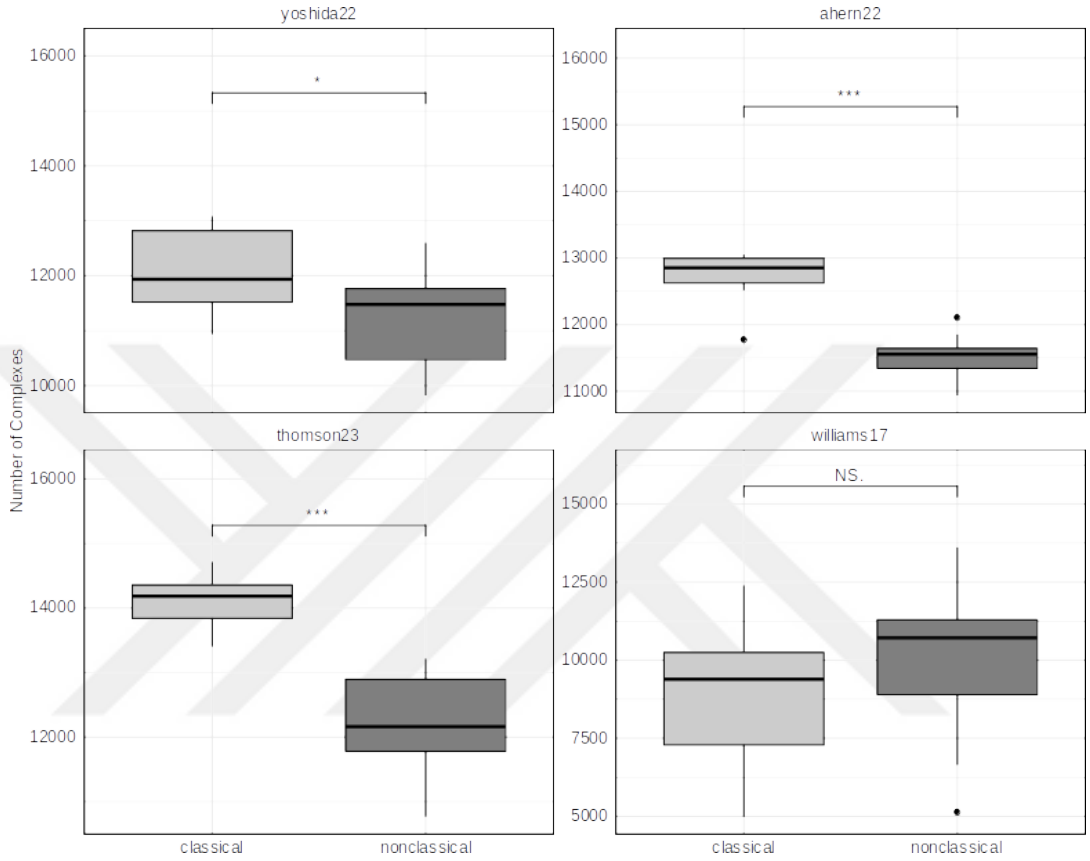


Figure 3.32: Number of Complexes Compared Between Cell Types in Each Study.

Variability in Complex Numbers Across Datasets

When comparing the number of predicted protein complexes across datasets for each cell type, consistent trends were observed in both classical and nonclassical monocytes:

- Among the scRNA-seq datasets:
  - **Thomson23** had the highest number of predicted protein complexes in both classical and nonclassical monocytes.
  - **Ahern22** and **Yoshida22** followed, with moderate numbers of complexes.
- The bulk RNA-seq dataset **Williams17** showed significantly fewer predicted pro-

tein complexes in both cell types compared to the scRNA-seq datasets.

Variation Across Datasets

An analysis of the variance in the number of predicted complexes across datasets revealed additional insights:

- The largest variation was observed in **Williams17** (bulk RNA-seq), reflecting inconsistencies in capturing transcriptional complexity at the population level.
- **Yoshida22** showed moderate variability, which may be attributed to the broader age range (newborn to 14 years) and other donor-specific factors.
- **Thomson23** and **Ahern22** demonstrated the smallest variation in complex numbers, likely due to narrower age ranges and consistent experimental protocols.

These findings underscore the enhanced resolution and consistency of scRNA-seq in profiling transcriptional protein complexes, particularly in distinguishing cell-type-specific differences. Additionally, the variability among datasets highlights the influence of technical and biological factors, such as sample preparation, sequencing depth, and donor heterogeneity, on the observed results.

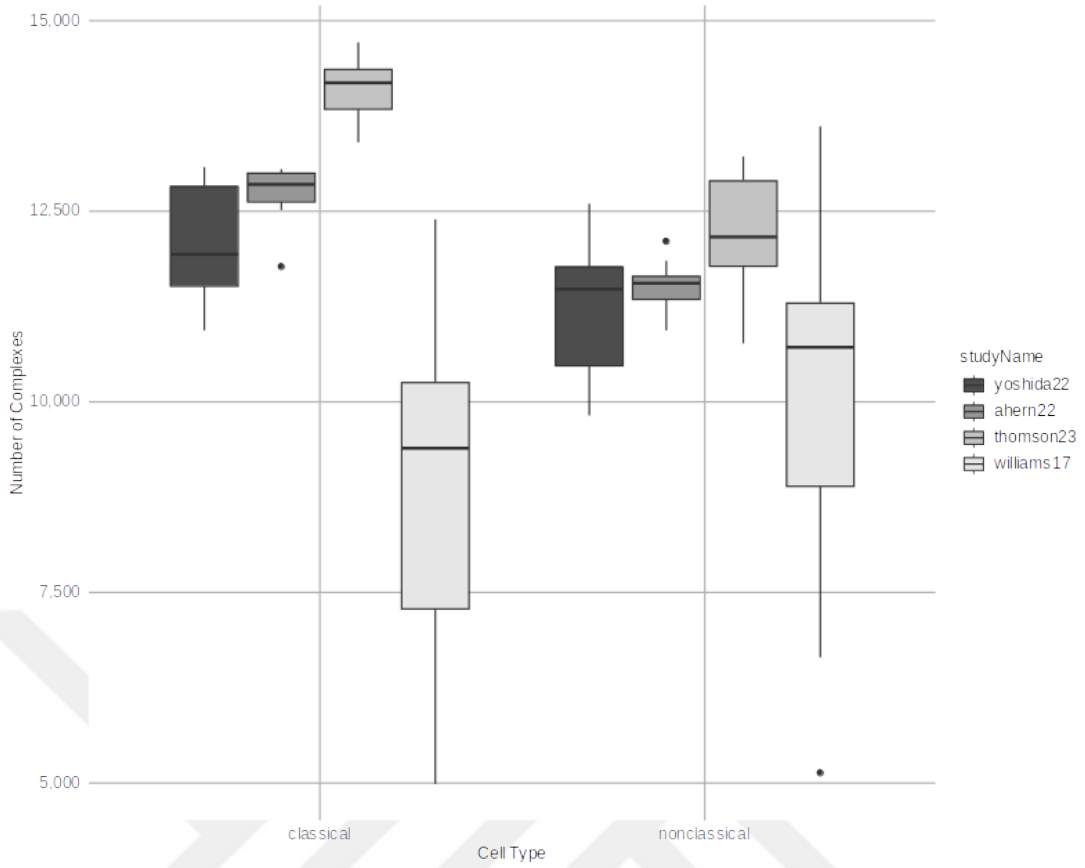


Figure 3.33: Number of Complexes Compared Between Studies for Each Cell Type.

3.3.2. Dataset 1: Lack of Significant Age-Associated Changes in PBMC Bulk RNA-seq

The transcriptional protein complexes from bulk RNA-seq of PBMCs (GSE262619) were analyzed across three age groups: Young (n=11, avg age 27.5 ± 2.54), Middle (n=10, avg age 60.4 ± 2.22), and Old (n=14, avg age 80.5 ± 4.57). Comparisons between these age groups (Old vs. Middle, Middle vs. Young, and Old vs. Young) did not reveal any significantly deregulated protein complexes. This lack of detectable differences highlights the limitations of bulk RNA-seq in capturing subtle, age-associated transcriptional changes, particularly within specific immune cell subpopulations.

3.3.3. Dataset 2: Impact of Aging and COVID-19 on Transcriptional Protein Complexes in Naive T Cells

The Age Factor on Protein Complexes Transcriptional protein complexes of naive thymus-derived CD4+ T (nCD4T) and naive thymus-derived CD8+ T (nCD8T) cells are predicted and compared between children and adults to investigate the age factor in both healthy and COVID conditions: healthy children (n=24 for both nCD4T and nCD8T cells) vs healthy adults (n=11 for nCD4T, n=7 for nCD8T), COVID children (n=14 for both nCD4T and nCD8T cells) vs COVID adults (n=28 for nCD4T, n=22 for nCD8T)

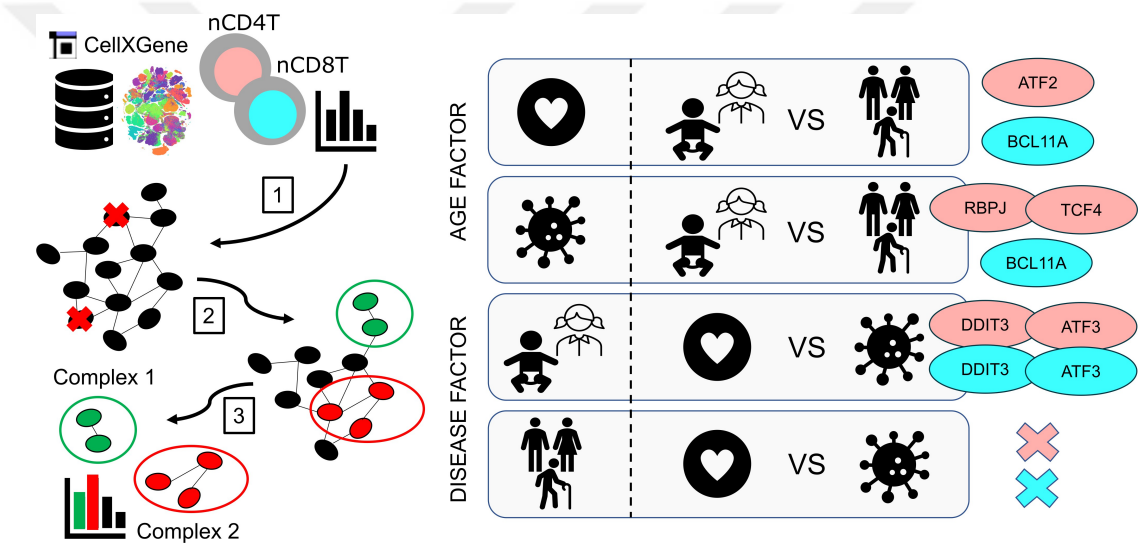


Figure 3.34: Graphical Abstract of the Analysis Made with Dataset 2.

Initially, the change in the protein complex repertoire due to age difference is investigated based on the number of unique proteins mapped to the protein-protein interaction network (PPIN) and domain-domain interaction network (DDIN), the number of unique interactions in these condition-specific networks and the number of unique predicted complexes (Figure 3.35). For nCD4T cells, the number of unique protein complexes predicted does not change significantly in healthy condition whereas it decreases significantly in COVID-infected adults compared to COVID-infected children (Figure 3.35). So, it can be implicated that age factor in healthy individuals at their state of quiescence

do not generate a significant change in the diversity of protein complexes; however, age factor manifests its effect in the activated or infected state with the lack or reduction in diversity of unique protein complex assemblies. For nCD8T cells, in both conditions age factor leads to significant reduction in the protein complex diversity, implicating that nCD8T cells might be more sensitive to aging than nCD4T cells with respect to regulatory complex repertoire. Their repertoire diversity is reduced in steady state by the increased age factor and that effect is emphasized more in the activated state.

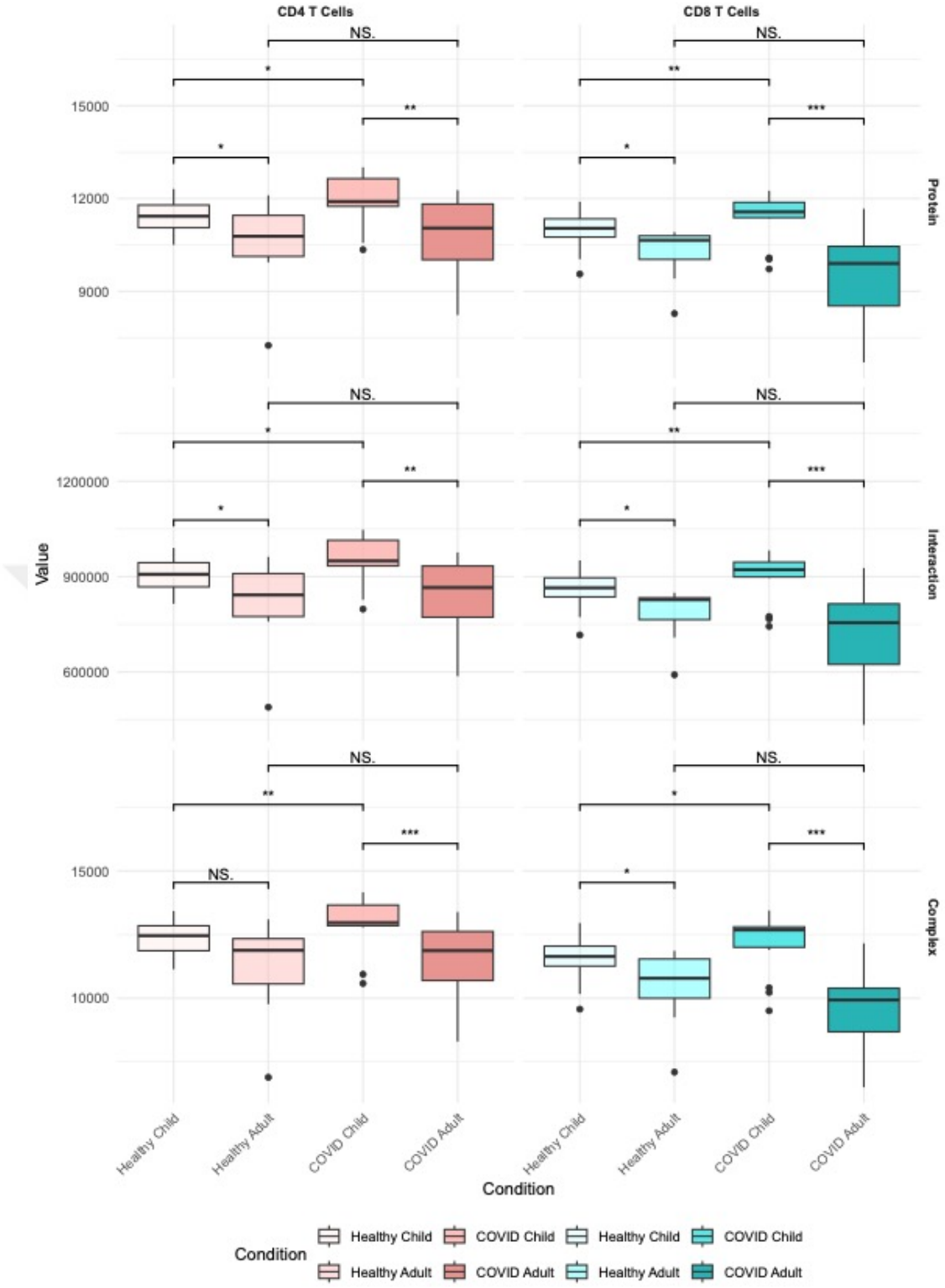


Figure 3.35: Number of mapped proteins, interactions and unique protein complexes predicted for nCD4T and nCD8T cells in 4 different conditions: Healthy Child, Healthy Adult, COVID Child, COVID Adult. Groups were compared by Wilcoxon rank sum exact test. NS: Not Significant (p-value > 0.05), *: Significant (p-value < 0.05), **: Significant (p-value < 0.01) ***: Significant (p-value < 0.001).

To test the effect of age factor on the transcriptional protein complexes, same composition complexes are quantified and quantified protein complexes that are common to at least 75% of the samples are compared. It has been observed that increased age generates a major deregulation in the amount of transcriptional protein complexes of both naïve T cells regardless of the disease status (Table 3.10). They show significant change in both healthy and COVID conditions: for nCD4T cells, 2742 protein complexes out of 7866 tested found to be deregulated in healthy adults compared to healthy children and 4214 protein complexes out of 8252 tested found to be deregulated in COVID adults compared to COVID children (Table 3.10). 1569 (29% of all) protein complexes were found to be deregulated in both comparisons, 1173 only in healthy comparison and 2645 only in COVID comparison (Figure 3.36). In both healthy and COVID comparisons, majority of the deregulated protein complexes are downregulated (2343/2742 and 4100/4214) in adults. To understand the similar patterns between deregulated complexes and differentially expressed genes, differentially expressed genes (DEGs) were determined. In the healthy comparison 174 up- and 1025 down-regulated genes, totaling 1199 were found which is larger than in the comparison with the disease where 394 up- and 571 down-regulated genes, totaling 965 were detected (Table 3.10). Similar trends are observed for nCD8T cells. 1634 protein complexes out of 7486 tested found to be deregulated in healthy adults compared to healthy children and 4996 protein complexes out of 7433 tested found to be deregulated in COVID adults compared to COVID children (Table 3.10). When shared deregulated complexes were investigated, 1071 (19% of all) protein complexes were found to be deregulated in both comparisons, 563 only in healthy comparison and 3925 only in COVID comparison (Figure 3.36). Decreased commonality in deregulation (19% of all complexes of nCD8T cells) compared to shared deregulation ratio of 29% in nCD4T might indicate larger sensitivity of nCD8T cells to aging. In both healthy and COVID comparisons, majority of the deregulated protein complexes are downregulated (1405/1634 and 4845/4996) in adults. A comparison of the number of deregulated protein complexes revealed a nearly threefold increase in the case of COVID disease. However, DEG analysis indicated a greater number of DEGs in the healthy comparison (83 up- and 571 down-regulated genes, totaling 654) than in the comparison with the disease (145 up- and 315 down-regulated genes, totaling 460). This suggests that the analysis of deregulated protein complexes

shows a different trend than that of differentially expressed genes (Table 3.10).

Table 3.10: Results of deregulated protein complex analyses, transcription factor enrichment, and differentially expressed gene analyses in different comparisons of naïve thymus-derived CD4+ (nCD4T) and CD8+ T (nCD8T) cells: AGE_HEALTHY represents the age factor tested in healthy individuals and AGE_COVID represents the age factor tested in COVID individuals. CHILD_DISEASE represents the infection factor tested in pediatric cases (COVID vs Healthy) and ADULT_DISEASE represents the infection factor tested in adults. Statistical test applied for complexes is Wilcoxon rank sum test (unpaired, non-parametric) and False Discovery Rate threshold is 0.05.

Cell Type	Comparison	Up-reg Comps	Down-reg Comps	Total Dereg Comps	TF in Up-reg Comps	TF in Down-reg Comps	Total TF in Comps	Up-reg DEGs	Down-reg DEGs	Total DEGs
nCD4T	age_healthy	399	2343	2742	52	25	77	174	1025	1199
nCD4T	age_covid	114	4100	4214	14	37	51	394	571	965
nCD4T	child_disease	2352	27	2379	49	13	62	30	391	421
nCD4T	adult_disease	0	0	0	28	53	81	117	107	224
nCD8T	age_healthy	229	1405	1634	52	15	67	83	571	654
nCD8T	age_covid	151	4845	4996	13	33	46	145	315	460
nCD8T	child_disease	1093	34	1127	46	17	63	11	289	300
nCD8T	adult_disease	0	0	0	9	49	58	44	38	82

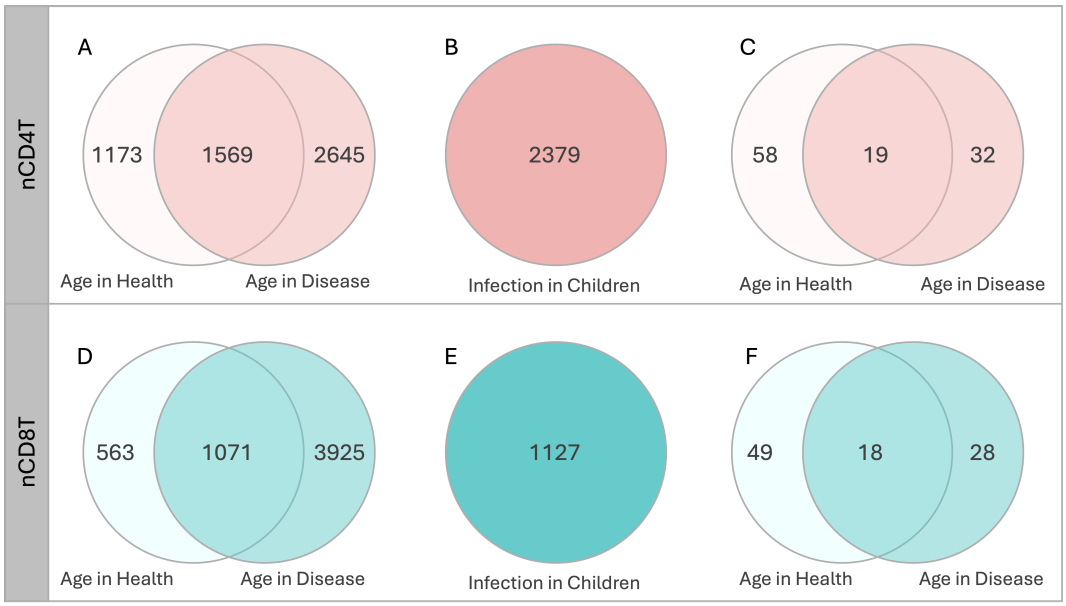


Figure 3.36: Intersection of deregulated protein complexes in age factor tests of nCD4T (A) and nCD8T (D) cells. Deregulated protein complexes in disease factor tests of nCD4T (B) and nCD8T (E) cells. Intersection of enriched transcription factors in deregulated protein complexes in age factor tests of nCD4T (C) and nCD8T (F) cells.

Top 25 deregulated transcriptional protein complexes are investigated further to find out most affected pathways and processes due to age difference. The highly enriched transcription factor in all deregulated protein complexes which is also the most frequent transcription factor (TF) found as a member of top 25 deregulated protein complexes is determined. In the nCD4T cells of healthy individuals, Cyclic AMP-dependent transcription factor ATF-2 (ATF2) emerged as the most frequently occurring TF within the top 25 deregulated protein complexes (10 out of 25, Fisher's Exact Test, p-value « 0.05), ranking 4th among 77 enriched TFs in the healthy comparison (Figure 3.37). In the COVID comparison, two transcription factors, Recombining binding protein suppressor of hairless (RBPJ, 15 out of 25, Fisher's Exact Test, p-value « 0.05) and Transcription factor 4 (TCF4, 5 out of 25, Fisher's Exact Test, p-value « 0.05), were notably present in the top 25 deregulated protein complexes. RBPJ ranked 2nd among the 51 enriched TFs in the COVID comparison, while TCF4 ranked 5th (Figure 3.38). Network analysis revealed no specific interaction or axis for ATF2-complexes, in contrast to the RBPJ-complexes, where significant interactions were observed between RBPJ

and MAML3, MAML2-MAML1, and MAML1-MAML3, forming a critical MAML2-MAML1-MAML3-RBPJ axis (Figure 3.39). Further gene enrichment analysis of the 25 unique proteins constituting the 10 ATF2-complexes in the healthy comparison highlighted significant Reactome and KEGG Pathway terms, which were predominantly grouped under "Mitochondrial Fatty Acid Beta-Oxidation." This suggests a role for ATF2-complexes in fatty acid oxidation metabolism associated with aging in nCD4T cells. Similarly, the 29 unique proteins comprising the 20 RBPJ- and TCF4-complexes in the healthy comparison were subjected to gene enrichment analysis, revealing enrichment terms such as "Signaling by NOTCH2/NOTCH3/NOTCH4", "NOTCH1/NOTCH2/- NOTCH3/NOTCH4 Intracellular Domain Regulates Transcription", "Pre-NOTCH Tran- scription and Translation", and "RUNX3 regulates NOTCH signaling" indicating a sig- nificant involvement of these complexes in NOTCH signaling pathways (Figure 3.39).

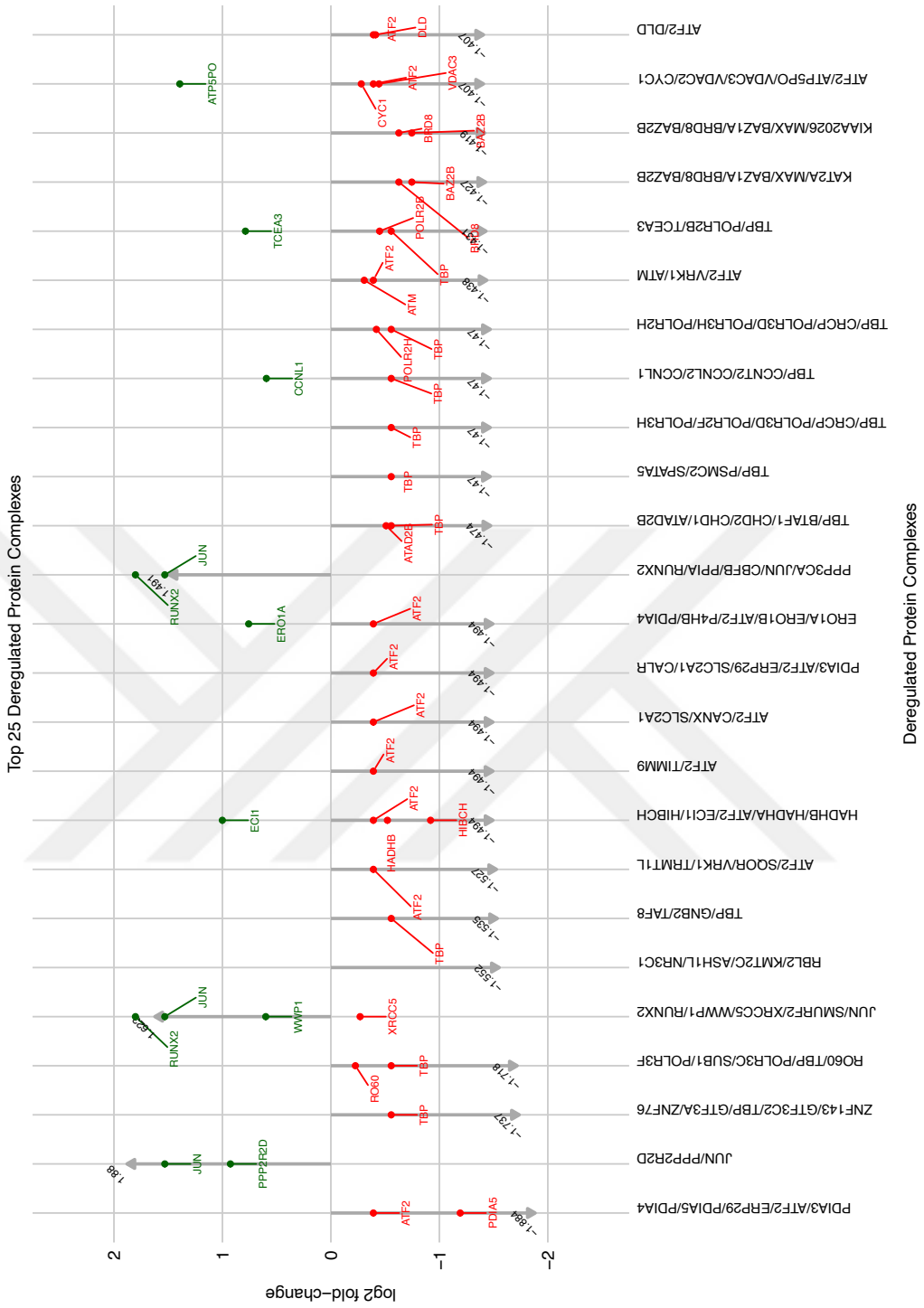


Figure 3.37: Top 25 deregulated regulatory protein complexes healthy children vs healthy adults comparison of nCD4T cells and their log2 fold change as a complex. Transcription factors in these complexes showing differential expression is represented as red or green dots and with corresponding gene name: Red dots representing the downregulation, green dots representing the upregulation. The position of the dot also shows the log2 fold change of the individual gene expression.

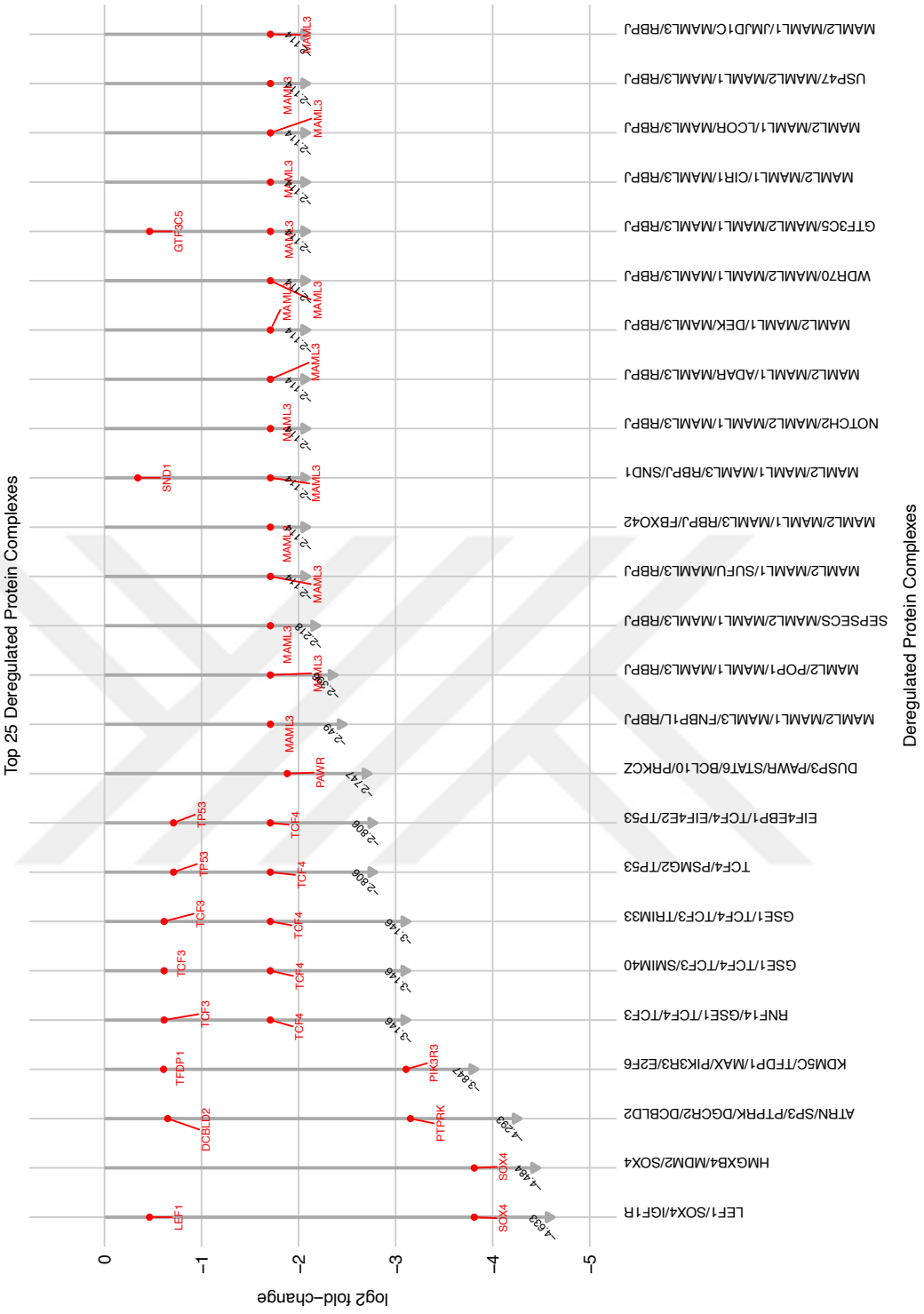


Figure 3.38: Top 25 deregulated regulatory protein complexes in COVID children vs COVID adults comparison of nCD4T cells and their log2 fold change as a complex. Transcription factors in these complexes showing differential expression is represented as red or green dots and with corresponding gene name: Red dots representing the downregulation, green dots representing the upregulation. The position of the dot also shows the log2 fold change of the individual gene expression.

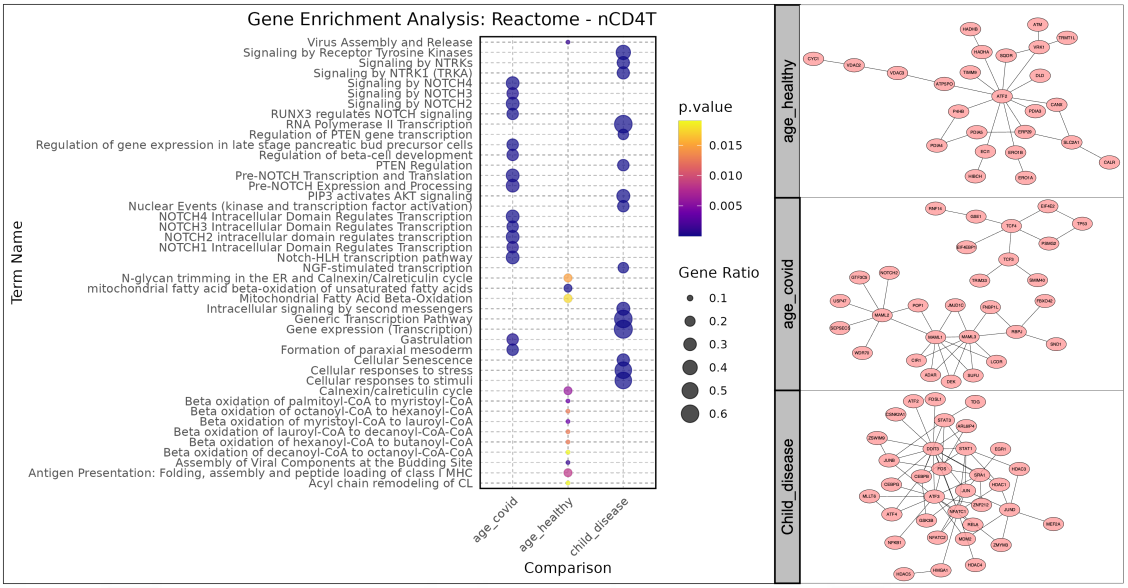


Figure 3.39: Network representation and gene enrichment analysis of unique proteins involved in various protein complexes across different conditions of nCD4T cells: 25 proteins from 10 ATF2-complexes in the healthy age comparison, 29 proteins from 20 RBPJ- and TCF4-complexes in the COVID age comparison, and 34 proteins from 25 DDIT3- and ATF3-complexes in the pediatric infection comparison, analyzed using Reactome pathway terms.

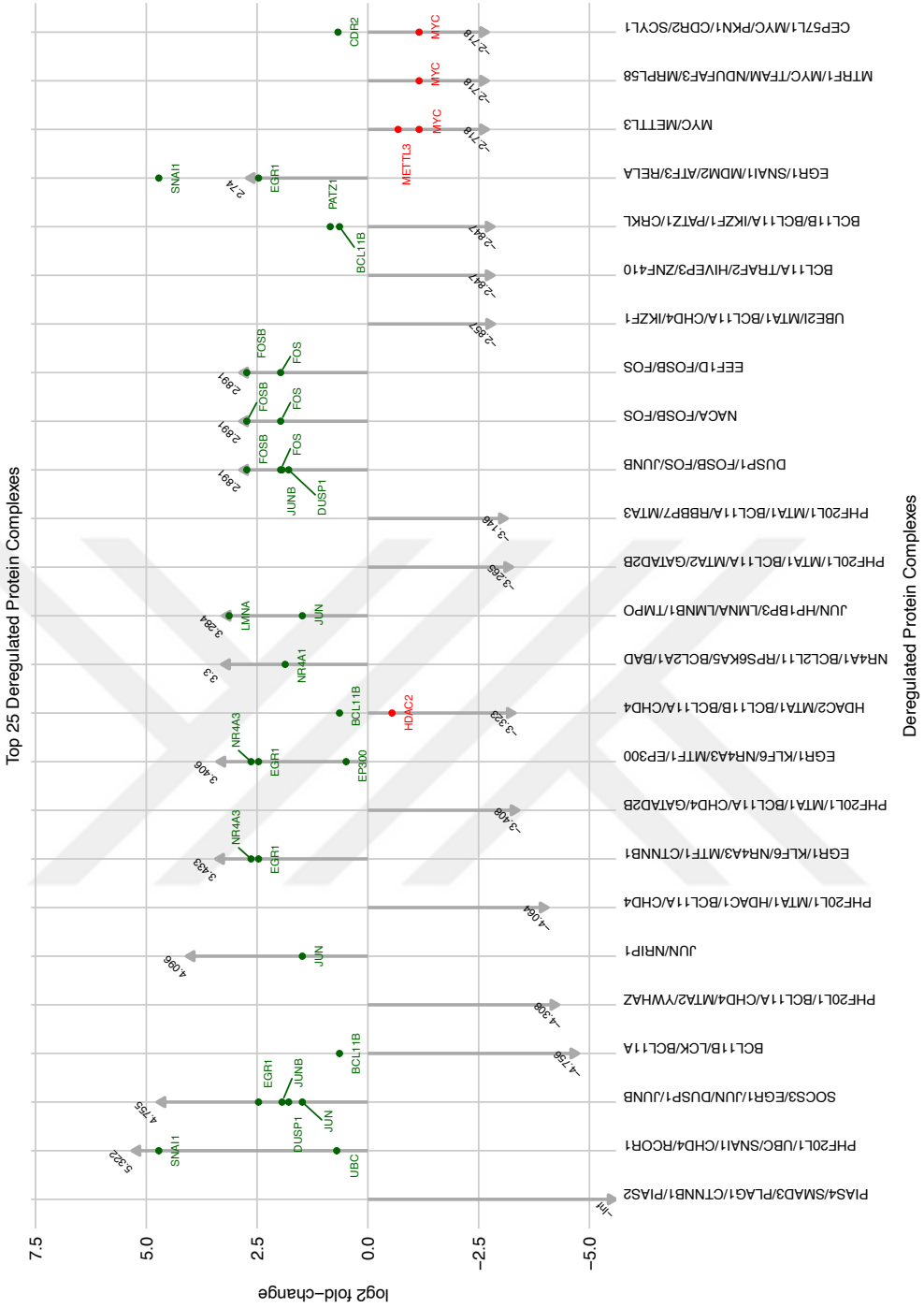


Figure 3.40: Top 25 deregulated regulatory protein complexes healthy children vs healthy adults comparison of nCD8T cells and their log2 fold change as a complex. Transcription factors in these complexes showing differential expression is represented as red or green dots and with corresponding gene name: Red dots representing the downregulation, green dots representing the upregulation. The position of the dot also shows the log2 fold change of the individual gene expression.

Interestingly, the TF that was found most frequently in the top 25 deregulated protein complexes of nCD8T cells in both healthy and COVID comparison was B-cell lymphoma/leukemia 11A (BCL11A, 10/25 in healthy and 13/25 in COVID, Fisher's Exact Test, p-value « 0.05), which ranked 3rd among the 67 enriched TFs of healthy comparison and 13th among the 46 enriched TFs of COVID comparison (Figures 3.40, 3.41). These 10 and 13 BCL11A-containing down-regulated protein complexes were further investigated for their role as a potential link between functional changes in nCD8T cells with regards to age difference (Figure 3.42). They are represented as networks and in both conditions the most frequent interactions occur between PHF20L1-MTA1, MTA1-BCL11A and BCL11A-CHD4. Together they indicate the key connections of BCL11A complexes and as a whole PHF20L1-MTA1-BCL11A-CHD4 the key axis in the network. 20 unique proteins that make up 10 BCL11A-complexes in healthy comparison and 27 unique proteins that make up 13 BCL11A-complexes in COVID comparison were further analyzed in gene enrichment analysis. Most significantly enriched Reactome and KEGG Pathway terms such as "Positive epigenetic regulation of rRNA expression", "HDACs deacetylate histones", "Epigenetic regulation of gene expression", "ERCC6 (CSB) and EHMT2 (G9a) positively regulate rRNA expression", "ATP-dependent chromatin remodeling" suggest a role in epigenetic modifications and chromatin remodeling for aging-associated BCL11A-complexes in nCD8T. Their enrichment in known complexes such as "SNF2h-cohesin-NuRD complex", "NuRD.1 (The Nucleosome Remodeling and Deacetylase) complex" also support this role (Figure 3.42). Second group of terms includes "Regulation of PTEN gene transcription", "PTEN regulation", "PIP3 activates AKT signaling", "Intracellular signaling by second messengers" implies another role for deregulated BCL11A-complexes in PTEN and AKT signaling pathways. Detecting the same enriched TF and most significantly deregulated protein complexes in both healthy and COVID comparisons of nCD8T cells might emphasize the dominating effect of age factor in this cell type.



Figure 3.41: Top 25 deregulated regulatory protein complexes in COVID children vs COVID adults comparison of nCD8T cells and their log2 fold change as a complex. Transcription factors in these complexes showing differential expression is represented as red or green dots and with corresponding gene name: Red dots representing the downregulation, green dots representing the upregulation. The position of the dot also shows the log2 fold change of the individual gene expression.

Transcription factors of deregulated protein complexes were further examined for their domains and functional families. 77 transcription factors were found to be enriched in deregulated protein complexes of nCD4T cells when testing age factor in healthy donors, and 51 were found to be enriched in deregulated protein complexes when testing age factor in COVID donors. 19 of these TFs were shared in both comparisons (Figure 3.36). Domain enrichment analysis in INTERPRO database resulted in significantly enriched domains of bZIP, AP-1, p53/RUNT-type, ZNF Homonal Receptor, Nuclear Hormonal Receptor. Transcription factors that were significantly enriched in nCD8T cell comparisons were similar in numbers: 67 in healthy condition, 46 in COVID, 18 in common. Domains enriched in these TFs were also similar such as bZIP, AP-1, ZNF Homonal Receptor, Nuclear Hormonal Receptor domains.

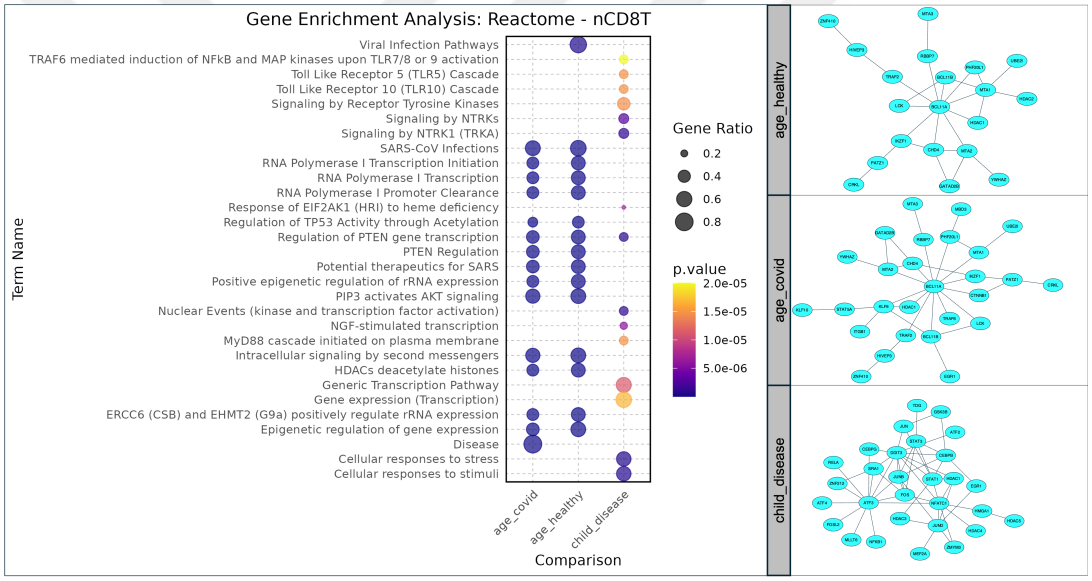


Figure 3.42: Network representation and gene enrichment analysis of unique proteins involved in various protein complexes across different conditions of nCD8T cells: 20 proteins from 10 BCL11A-complexes in the healthy age comparison, 27 proteins from 13 BCL11A-complexes in the COVID age comparison, and 29 proteins from 25 DDIT3- and ATF3-complexes in the pediatric infection comparison, analyzed using Reactome pathway terms.

The COVID-infection Factor on Protein Complexes

Transcriptional protein complexes of naive thymus-derived CD4+ T (nCD4T) and naive thymus-derived CD8+ T (nCD8T) cells are predicted and compared between healthy and COVID-infected donors to investigate the infection factor in both children and adults: healthy children (n=24 for both nCD4T and nCD8T cells) vs COVID children (n=14 for both nCD4T and nCD8T cells), healthy adults (n=11 for nCD4T, n=7 for

nCD8T) vs COVID adults (n=28 for nCD4T, n=22 for nCD8T).

The analysis of diversity and quantity of protein complexes revealed significant differences predominantly in pediatric cases. In both cell types, the increase in unique protein complexes was significant in COVID-infected children compared to their healthy counterparts, whereas in adults, the change in number of unique protein complexes was not significant (Figure 3.35).

Among the 8763 protein complexes tested, 2379 were identified as deregulated in COVID-infected children compared to their healthy counterparts in nCD4T. In contrast, none of the 6299 protein complexes tested in adults showed signs of deregulation between COVID-infected and healthy individuals (Table 3.10). Of the deregulated protein complexes in children, the majority (2352 out of 2379) were found to be upregulated. Additionally, the study of differentially expressed genes (DEGs) in pediatric cases revealed 30 upregulated and 391 downregulated DEGs, totaling 421. In adult cases, the DEG analysis showed 117 upregulated and 107 downregulated DEGs, summing up to 224 genes. Similarly in nCD8T cells, infection in pediatric cases generated an upregulation of 1093 protein complexes and a downregulation of 34 among 7851. Among the 6455 complexes tested, the change detected due to infection is not significant in adults. On the differential gene expression level, 300 DEGs were identified in pediatric cases and 82 DEGs were identified in adult cases. That shows deregulation in gene and complex level might differ both in their size and direction.

Detailed look in most deregulated protein complexes in pediatric cases show that in both nCD4T and nCD8T cells, most frequent TF members of complexes are DNA damage-inducible transcript 3 protein (DDIT3) and Cyclic AMP-dependent transcription factor ATF-3 (ATF3). In nCD4T cells, 7th of the enriched transcription factors, DDIT3, appear in 13 of the top 25 deregulated protein complexes (13/25, Fisher's Exact Test, p-value < 0.05) and 11th of the enriched transcription factors, ATF3, appear in 23 of the top 25 deregulated protein complexes (23/25, Fisher's Exact Test, p-value < 0.05) (Figure 3.43). Network analysis of the protein complexes revealed NFATC and JUNB as the most frequent interactors of ATF3 and ATF3, SRA1, STAT1 as the most frequent interactors of DDIT3; however, there was no particular axis emphasized in the network. Further gene enrichment analysis of the 34 unique proteins constituting the 25 DDIT3- and ATF3-complexes in the pediatric comparison were subjected to gene enrichment analysis, revealing enrichment terms such as "Cellular responses to stress/ stimuli", "Cellular Senescence", "RNA Polymerase II Transcription", "Disease",

"Nuclear Events (kinase and transcription factor activation)", "Signaling by NTRKs", "NGF-stimulated transcription" (Figure 3.39). In nCD8T cells, 1st of the enriched transcription factors, DDIT3, appear in 10 of the top 25 deregulated protein complexes (10/25, Fisher's Exact Test, p-value < 0.05) and 9th of the enriched transcription factors, ATF3, appear in 22 of the top 25 deregulated protein complexes (22/25, Fisher's Exact Test, p-value < 0.05) (Figure 3.44). Network and enrichment analyses of 29 unique proteins making up the 25 deregulated complexes of DDIT3 and ATF3 give similar results to the nCD4T analyses for frequent interactors and enriched pathway terms with additional terms such as "Response of EIF2AK1 (HRI) to heme deficiency", "Toll Like Receptor 10 (TLR10) Cascade", "Toll Like Receptor 5 (TLR5) Cascade", "MyD88 cascade initiated on plasma membrane","TRAF6 mediated induction of NFkB and MAP kinases upon TLR7/8 or 9 activation" (Figure 3.42).

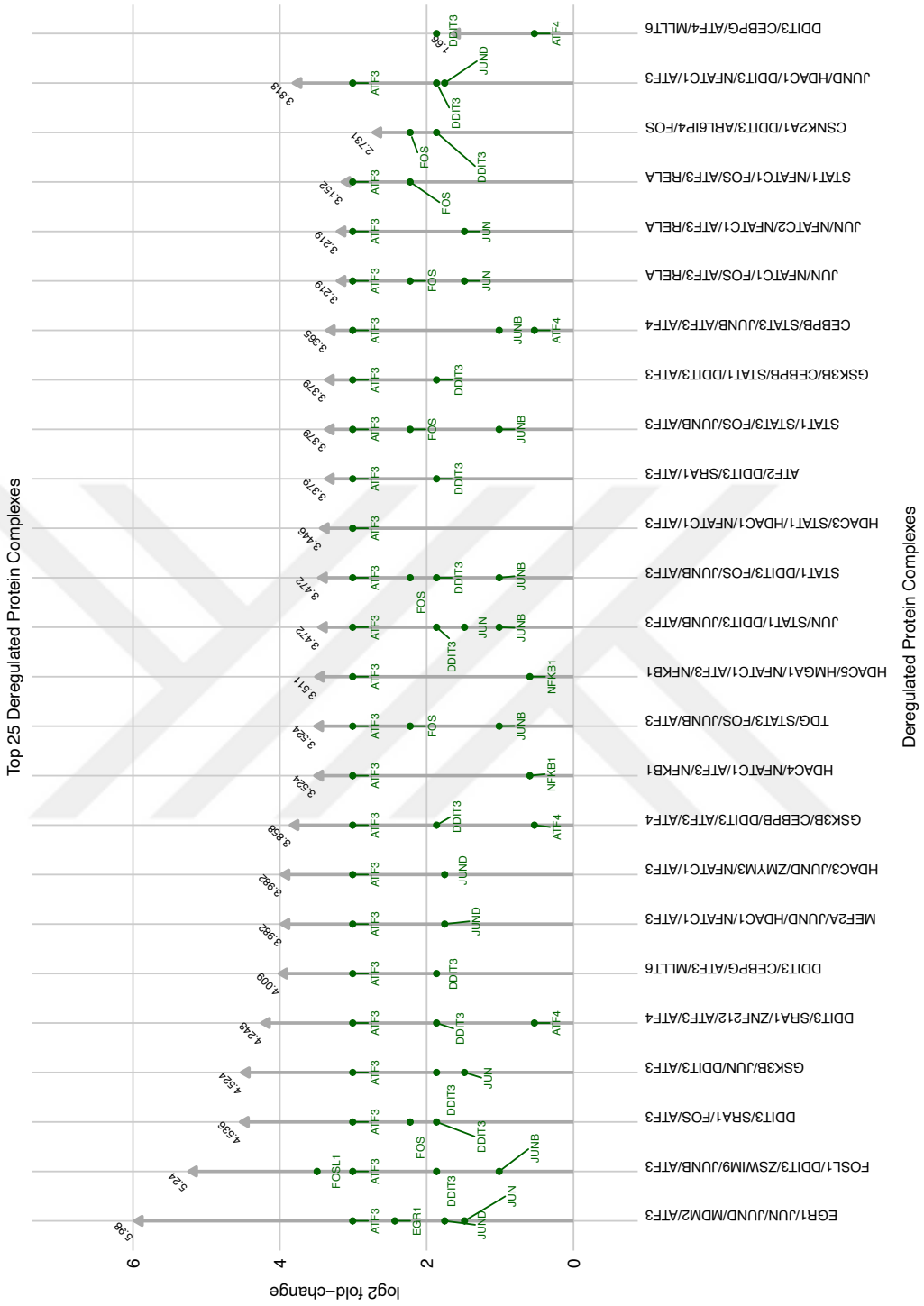


Figure 3.43: Top 25 deregulated regulatory protein complexes in Healthy children vs COVID children comparison of nCD4T cells and their log2 fold change as a complex. Transcription factors in these complexes showing differential expression is represented as red or green dots and official gene name: Red dots representing the downregulation, green dots representing the upregulation. The position of the dot also shows the log2 fold change of the individual gene expression.

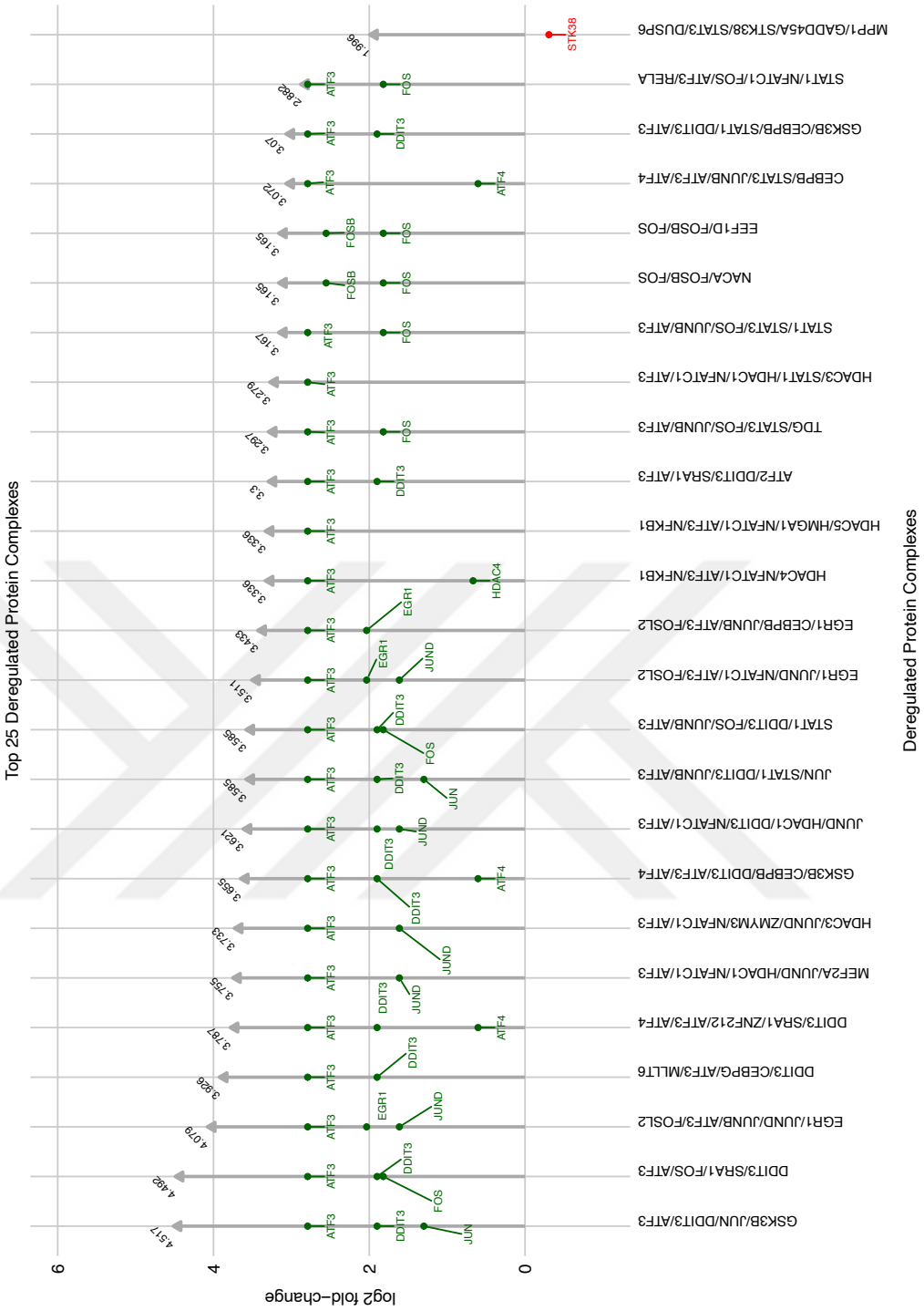


Figure 3.44: Top 25 deregulated regulatory protein complexes in Healthy children vs COVID children comparison of nCD8T cells and their log2 fold change as a complex. Transcription factors in these complexes showing differential expression is represented as red or green dots and official gene name: Red dots representing the downregulation, green dots representing the upregulation. The position of the dot also shows the log2 fold change of the individual gene expression.

Transcription factors of deregulated protein complexes were further examined for their domains and functional families. 62 transcription factors were found to be enriched in deregulated protein complexes of nCD4T cells when testing COVID-infection factor in pediatric donors, and 63 were found to be enriched in deregulated protein complexes of nCD8T cells. 19 of these TFs were shared in both cell type comparisons (Figure 3.36). Domain enrichment analysis in INTERPRO database resulted in significantly enriched domains of “SCAN”, “bZIP”, “E2F”, “JNK”, “Znf_C2H2_type”. Transcription factors that were significantly enriched in nCD8T cell comparisons were enriched in similar domains such as “SCAN”, “bZIP”, and “JNK”.

3.3.4. Dataset 3: Distinct Age-Related Patterns in nCD4T and nCD8T Cells

Aging-Associated Deregulation in Protein Complexes of Naive T Cells

This study aimed to investigate the dynamics of transcriptional protein complexes in naive CD4+ T (nCD4T) cells and naive CD8+ T (nCD8T) cells across three age groups: Child (CH; 11-13 years; $n = 16$), Young Adult (YA; 20-35 years; $n = 16$), and Older Adult (OA; 55-65 years; $n = 16$) (Figure 3.45[106]). By integrating expression data [74] with protein-protein interaction networks and domain-domain interaction networks, we modeled transcriptional protein complexes and quantified their abundance. These complexes were analyzed across age groups to determine aging-associated changes.

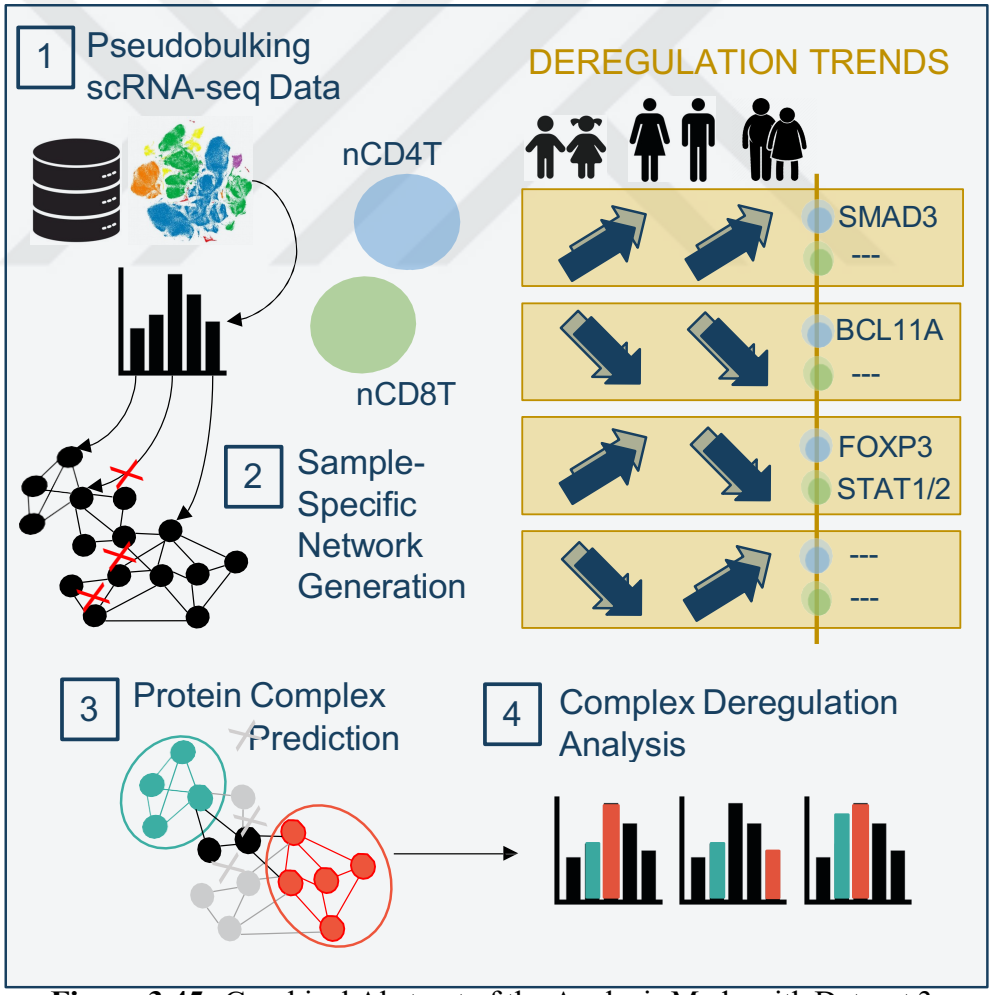


Figure 3.45: Graphical Abstract of the Analysis Made with Dataset 3.

Quantification of Transcriptional Protein Complexes Across Age Groups

The total number of unique protein complexes predicted in nCD4T cells remained stable across age groups (Figure 3.46A[106]), with no significant differences observed between the means of the groups (CH: 14200, YA: 14154, OA: 14188; Wilcoxon rank sum test, $p > 0.05$). Conversely, nCD8T cells exhibited a significant reduction in the number of unique complexes with increasing age (CH: 13418, YA: 12858, OA: 11839; Wilcoxon rank sum test, $p < 0.001$ for all comparisons). This reduction in nCD8T complexes was also reflected in the number of mapped proteins and total protein-protein interactions (Figures 3.46B-C[106]), highlighting a greater sensitivity of nCD8T cells to aging-associated deterioration compared to nCD4T cells.

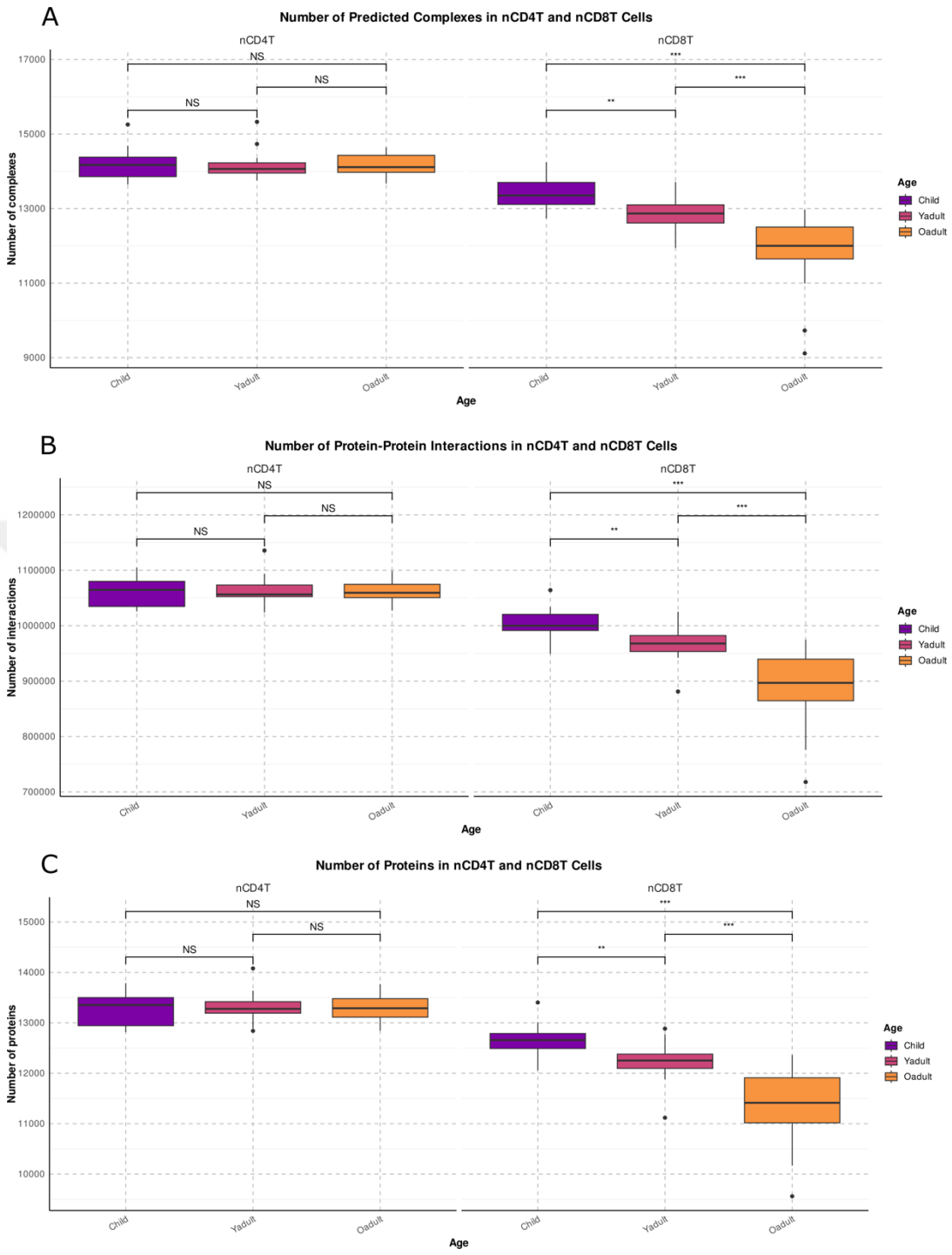


Figure 3.46: (A) Number of unique protein complexes predicted and compared for three age groups Child, Young Adult (Yadult), Older Adult (Oadult) in two cell types: naive CD4+ T cells (nCD4T) and naive CD8+ T cells (nCD8T). (B) Number of unique protein-protein interactions detected and compared for three age groups Child, Young Adult (Yadult), Older Adult (Oadult) in two cell types: naive CD4+ T cells (nCD4T) and naive CD8+ T cells (nCD8T). (C) Number of unique proteins mapped and com- pared for three age groups Child, Young Adult (Yadult), Older Adult (Oadult) in two cell types: naive CD4+ T cells (nCD4T) and naive CD8+ T cells (nCD8T). Wilcoxon rank sum exact test is used in all comparisons. NS: Not Significant p-value > 0.05, *: Significant p-value < 0.05, **: Significant p-value < 0.01 ***: Significant p-value < 0.001.

Deregulated Complexes and Transcription Factor Enrichment

The analysis of deregulated complexes revealed distinct patterns in nCD4T and nCD8T cells (Table 3.11). In nCD4T cells, the total number of deregulated protein complexes increased with age, with 1774 complexes deregulated between CH and YA, 378 between YA and OA, and 4179 between CH and OA. Most of these complexes were upregulated (e.g., 3917 of 4179 in CH vs OA), suggesting a reorganization of transcriptional regulatory networks during aging (Table 3.11, Figure 3.47A [106]).

Table 3.11: Results of deregulated protein complex analyses and transcription factor enrichment in various comparisons of naïve CD4+ (nCD4T) and CD8+ (nCD8T) cells. Comparisons include: CH vs YA (children vs young adults), YA vs OA (young adults vs old adults), and CH vs OA (children vs old adults). Statistical testing for complexes was performed using the Wilcoxon rank-sum test (unpaired, non-parametric), with a False Discovery Rate (FDR) threshold of 0.05.

Cell Type	Comparison	Total Dereg Comps	Up-reg Comps	TF in Up-reg Comps	Down-reg Comps	TF in Down-reg Comps
nCD4T	CH vs YA	1774	1695	62	79	69
nCD4T	YA vs OA	378	313	31	65	16
nCD4T	CH vs OA	4179	3917	61	262	58
nCD8T	CH vs YA	450	431	35	19	35
nCD8T	YA vs OA	109	16	51	93	48
nCD8T	CH vs OA	2379	1640	42	739	43

In nCD8T cells, the pattern differed. A smaller number of complexes were deregulated in each comparison (e.g., 450 in CH vs YA and 109 in YA vs OA). However, the largest change was observed between CH and OA, with 2379 complexes deregulated, including 1640 upregulated and 739 downregulated (Figure 3.47B [106]). Notably, transcription factors enriched in deregulated complexes varied between the two cell types. Enriched domains in nCD4T cells included the SMAD/FHA domain and the Basic-leucine zipper (bZIP) domain. In contrast, nCD8T cells exhibited enrichment in additional domains such as Rel homology, Ets, and High Mobility Group (HMG) box domains.

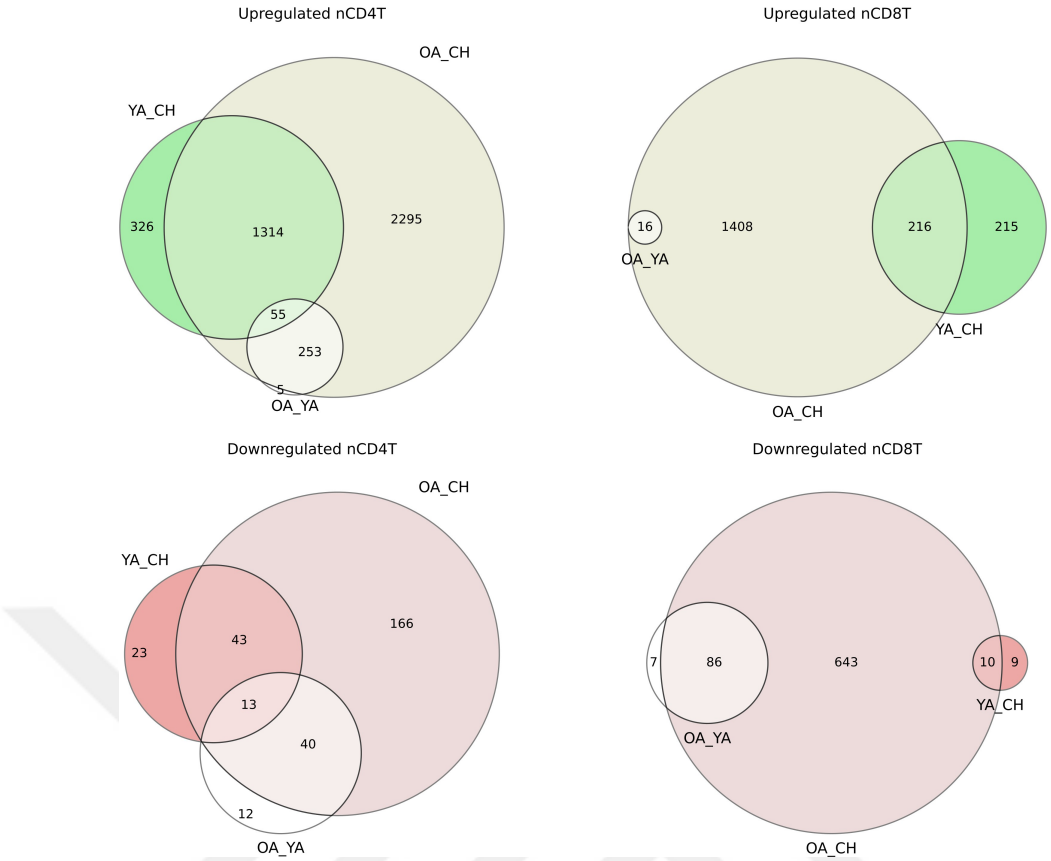


Figure 3.47: Number of up- and down-regulated protein complexes in three age group comparisons: Child vs Young Adult (CH_YA), Young Adult vs Older Adult (YA_OA), Child vs Older Adult (CH_OA) for two cell types: naive CD4+ T cells (nCD4T) and naive CD8+ T cells (nCD8T).

Trajectory Analysis of Protein Complexes in Aging

Trajectory analysis identified 70 protein complexes in nCD4T cells exhibiting specific trends across age groups: 55 showed continuous upregulation, 13 showed continuous downregulation, and 2 exhibited non-monotonic trends (Figure 3.48 [106]). In contrast, no such trends were observed in nCD8T cells, suggesting that aging-related regulatory mechanisms may differ fundamentally between these cell types. Principal Component Analysis (PCA) of these complexes separated the three age groups in nCD4T cells with 83% accuracy and 84% precision (Figure 3.49A [106]).

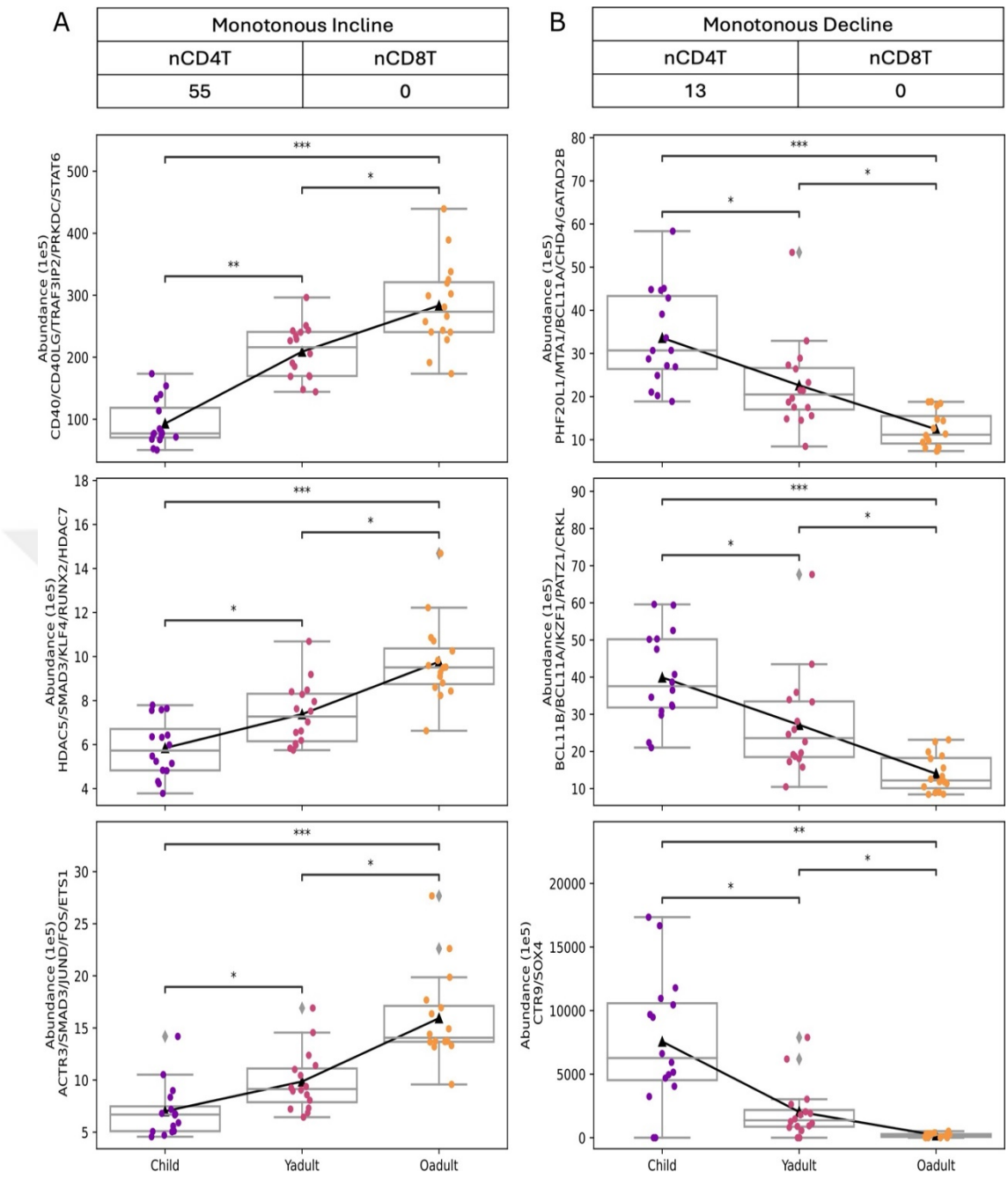


Figure 3.48: Representative deregulated protein complexes that show monotonous incline and decline across three age groups Child, Young Adult (Yadult), Older Adult (Oadult) and in two cell types: naive CD4+ T cells (nCD4T) and naive CD8+ T cells (nCD8T). Total number of complexes are shown at the top of each panel (A) Abundance values of 3 selected continuously up-regulated protein complexes: CD40 – CD40LG – TRAF3IP2 – PRKDC – STAT6; HDAC5 – SMAD3 – KLF4 – RUNX2 – HDAC7; ACTR3 – SMAD3 – JUND – FOS – ETS1. (B) Abundance values of 3 selected contin- uously down-regulated protein complexes: PHF20L1 – MTA1 – BCL11A – CHD4 – GATAD2B; BCL11B – BCL11A – KZF1 – PATZ1 – CRKL; CTR9 – SOX4. Wilcoxon rank sum exact test is used in all comparisons. NS: Not Significant p-value > 0.05, *: Significant p-value < 0.05, **: Significant p-value < 0.01 ***: Significant p-value < 0.001.

Representative upregulated complexes included those containing SMAD3 and FOS, while downregulated complexes were centered around BCL11A and CHD4 (Figure 3.48 [106]). SMAD3 complexes were particularly associated with pathways such as Th17 cell differentiation and IL-17 production, highlighting their relevance in age-related immune modulation.

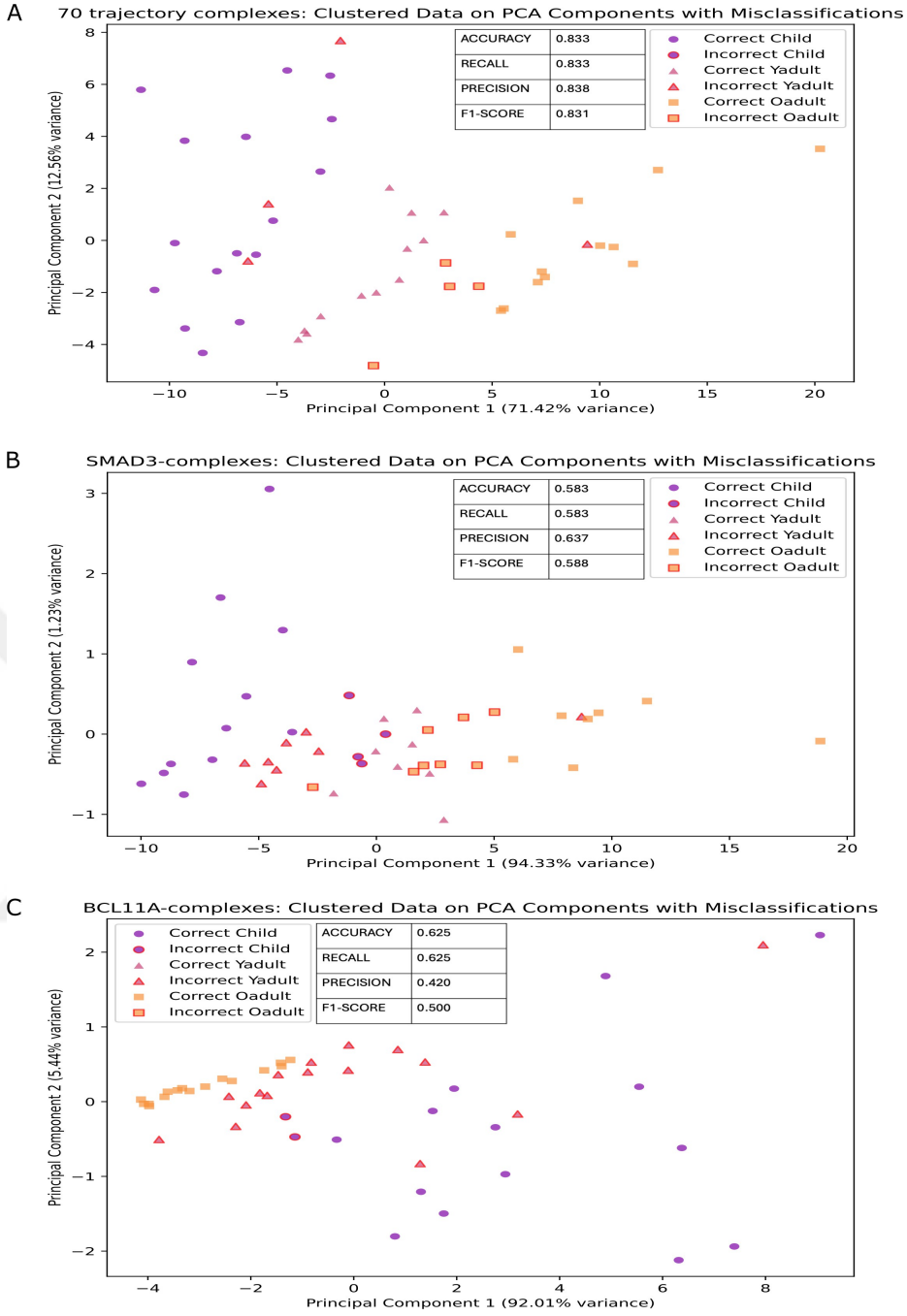


Figure 3.49: PCA of abundance data of (A) 70 deregulated trajectory protein complexes, (B) 39 SMAD3-containing up-regulated trajectory protein complexes and (C) 13 BCL11A-containing down-regulated trajectory protein complexes in three age groups Child, Young Adult (Yadult), Older Adult (Oadult) of naive CD4+ T cells repre- sented on the first two principal components out of 20 computed. Results of K-means clustering represented on two principal components and calculated evaluation metrics: Accuracy, Recall, Precision and F1-score.

Network Analysis and Validation

Network analysis revealed SMAD3 as a central hub in upregulated complexes, interacting most frequently with PIAS3, a protein implicated in T regulatory cell function. In contrast, BCL11A emerged as a key hub in downregulated complexes, with frequent interactions involving CHD4 and MTA1, proteins associated with chromatin remodeling and DNA repair (Figure 3.50 [106]).

Validation against external datasets confirmed the relevance of these complexes: 21 of the 70 trajectory complexes showed consistent deregulation in a separate child versus adult comparison dataset, including 11 involving BCL11A and 5 involving SMAD3.

In this study, we conducted a comprehensive deregulation analysis of transcriptional protein complexes in naive CD4+ and CD8+ T cells across three distinct age groups to explore the effects of aging and identify potential reversal pathways. Our findings revealed a marked contrast between the stability of the complexome repertoire in naive CD4+ T cells and the reduced diversity observed in naive CD8+ T cells with aging. While the mechanisms driving this decline in CD8+ T cells remain unclear, the reduction in complexome diversity likely contributes to impaired immune responses. These observations align with prior transcriptomic and epigenomic profiling studies [69, 72–74], which have reported similar age-related trends in transcription factor activity and chromatin organization.

Notably, our analysis identified transcription factors within deregulated complexes, such as those containing the bZIP domain, that corroborate previous studies[69, 72–74] on the impact of aging on gene regulation and epigenetic modulation. This reinforces the idea that aging-associated changes in transcription factor dynamics are integral to immune system decline. The observed trends in naive CD4+ T cells, such as monotonic increases or decreases and cyclic patterns in transcriptional protein complexes, suggest traceable aging pathways that may be amenable to intervention. However, the absence of such trends in naive CD8+ T cells indicates that their age-related changes might involve more intricate regulatory mechanisms, as supported by earlier studies on the epigenetic complexity of T cell subsets.

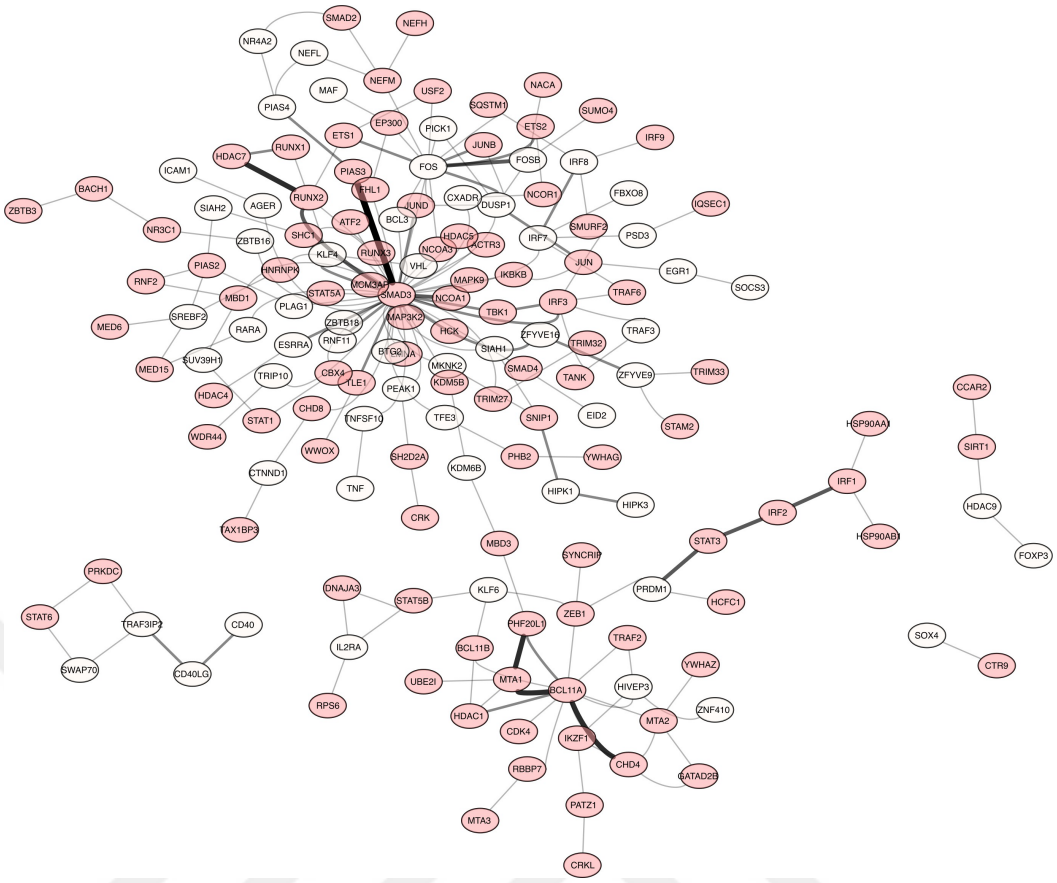


Figure 3.50: Network representation of 70 trajectory protein complexes of naive CD4+ T cells. MCD3 protein is indicated with a black arrowhead as the link between up- per SMAD3 cluster and lower BCL11A cluster. The width of the edges reflects the frequency of the interaction in complexes. SMAD3 is the highest degree node of the network with 50 interactions of which the most frequent (5 times) partner is E3 SUMO- protein ligase PIAS3 (PIAS3). FOS comes the second in degree with 14 and Protein FosB (FOSB) is the most frequent (3 times) interacting partner. BCL11A is the third with 12 and it has equally (4 times) interacting partners Chromodomain-helicase-DNA- binding protein 4 (CHD4) and Metastasis-associated protein MTA1 (MTA1). SMAD3 and FOS appear in up-regulated complexes and BCL11A appears in the continuously down-regulated complexes. Colorless nodes represent proteins which are not detected at protein level and pink nodes represent proteins which are detected at protein level in naïve CD4+ T cells.

Our investigation further identified SMAD3 and BCL11A as central transcription factors in trajectory protein complexes of aging naive CD4+ T cells. These findings resonate with previous studies emphasizing the importance of SMAD3 in TGF-β signaling and its role in Th17 differentiation and IL17 production. The interaction between SMAD3 and PIAS3, critical for Th17 and T regulatory cell function, mirrors epigenomic analyses that highlight the interplay between transcription factors and chromatin modifiers in regulating immune responses [35, 107]. Similarly, the phosphorylation status of SMAD3, influencing its cofactor function in T cell differentiation [108], underscores its pivotal role in immune aging. This finding complements earlier research showing increased IL-17 production in naive CD4+ T cells with aging, adding a protein complex-level perspective to the existing transcriptomic and epigenomic data.

BCL11A, another key regulator in aging naive CD4+ T cells, was found to interact with chromatin remodelers like CHD4 and MTA1, suggesting a role in age-related epigenetic alterations. Prior studies have demonstrated the importance of CHD4 in DNA damage repair and maintaining the genomic stability of T cells [109, 110], reinforc- ing its relevance to our findings. Although direct links between BCL11A and prema- ture aging are sparse[111, 112], our results are consistent with reports of its role in hematopoietic stem cell function and differentiation. Furthermore, the downregulation of BCL11A complexes in naive CD8+ T cells aligns with broader epigenomic evidence of distinct regulatory pathways in T cell subsets.

Intriguingly, the presence of MBD3 in both SMAD3- and BCL11A-associated complexes suggests a bridging role between these networks, hinting at potential crosstalk mechanisms that may contribute to age-associated immune changes. This observation complements findings from prior epigenomic studies that have highlighted the multifaceted roles of chromatin modifiers in coordinating transcriptional responses during aging.

Overall, our transcriptional protein complex analysis underscores the differential impact of aging on naive CD4+ and CD8+ T cells, revealing distinct molecular mechanisms that parallel and complement epigenomic studies. By providing a deeper understanding of the protein complex dynamics underlying immune senescence, this study highlights the potential of these complexes as biomarkers and therapeutic targets, offering valuable insights into the molecular basis of immune aging.

Identification of Regulatory Loops for the Significantly Deregulated Protein Complexes

The identification of deregulated protein complexes in aging T cells highlighted a potential link between transcription factor (TF) complex formation and regulatory chromatin interactions. While this analysis was conducted in a different biological system, it raised an important question: Could such complex-driven regulatory loops be directly examined in the same biological context? To explore this possibility, a focused analysis was performed, investigating whether genes targeted by a deregulated complex involving BCL11A, IKZF1, and CHD4 exhibited structured enhancer-promoter interactions. This minor analysis serves as a proof-of-concept, demonstrating that integrating TF complex composition with enhancer regulatory loops could be a viable strategy for deciphering transcriptional regulation.

By analyzing regulatory regions in chromosome 21 bound by these TFs and mapping their target genes, two distinct regulatory patterns emerged. The first (Group 1) represents a constrained, loop-based mode of regulation, while the second (Group 2) exhibits a more flexible, multi-enhancer-driven network. These findings align with broader insights from enhancer mode-of-action analysis, particularly the category where the temporal dynamics of enhancer interactions are more critical than enhancer quantity. Together, these results support the feasibility of extending the major analytical framework to investigate how TF complex deregulation impacts enhancer-driven transcriptional control across different biological contexts. The regulatory landscape of gene expression is shaped by complex interactions between transcription factors (TFs), regulatory elements, and chromatin architecture. In this study, two distinct modes of regulatory interactions were examined—one representing constrained, loop-based enhancer-promoter interactions (Group 1) and the other depicting a more flexible, dynamic regulatory framework (Group 2). Integrating these findings with enhancer mode-of-action analysis provides deeper insight into the balance between structural stability and temporal enhancer dynamics in transcriptional regulation.

The Group 1 analysis revealed 51 genes targeted by three regulatory regions, each bound by the TFs BCL11A, IKZF1, and CHD4 (Figure 3.51). Among these, OLIG1, OLIG2, IFNGR2, GART, and TMEM50B stood out as genes that are co-targeted by at least two regulatory regions, indicating potential regulatory reinforcement. The high

specificity and overlap between TF binding and enhancer-promoter interactions in this group suggest a constrained and functionally relevant transcriptional loop, where enhancer interactions are stable and possibly insulated from extensive cross-regulatory influences. These findings align with the ”balanced or additive” enhancer mode-of-action scenario, where both the total number of interactions and their temporal stability correlate strongly with gene expression. This suggests that enhancer engagement follows a structured pattern, maintaining stable regulatory control. Unlike scenarios where network rewiring is the key driver, here the cohesion and persistence of interactions appear to be strong determinants of expression outcomes. This interpretation aligns well with the enrichment of key transcription factor motifs in cardiomyocyte differentiation. If genes from Group 1 overlap with those identified in the enhancer mode-of-action analysis, it would suggest that their expression is not driven by frequent network reconfiguration but rather by the maintenance of stable enhancer-promoter interactions, ensuring precise control over differentiation programs.

In contrast, Group 2 presented a much more diffuse and heterogeneous regulatory landscape, comprising 159 genes and 1108 regulatory regions. Each gene was targeted by an average of 7 regulatory regions, with influence from 2.5 TFs, suggesting a more flexible enhancer-promoter interaction pattern rather than a tightly constrained loop. A subset of 114 genes was characterized by at least two regulatory regions and two TFs, indicating a balance between regulatory redundancy and enhancer cooperation.

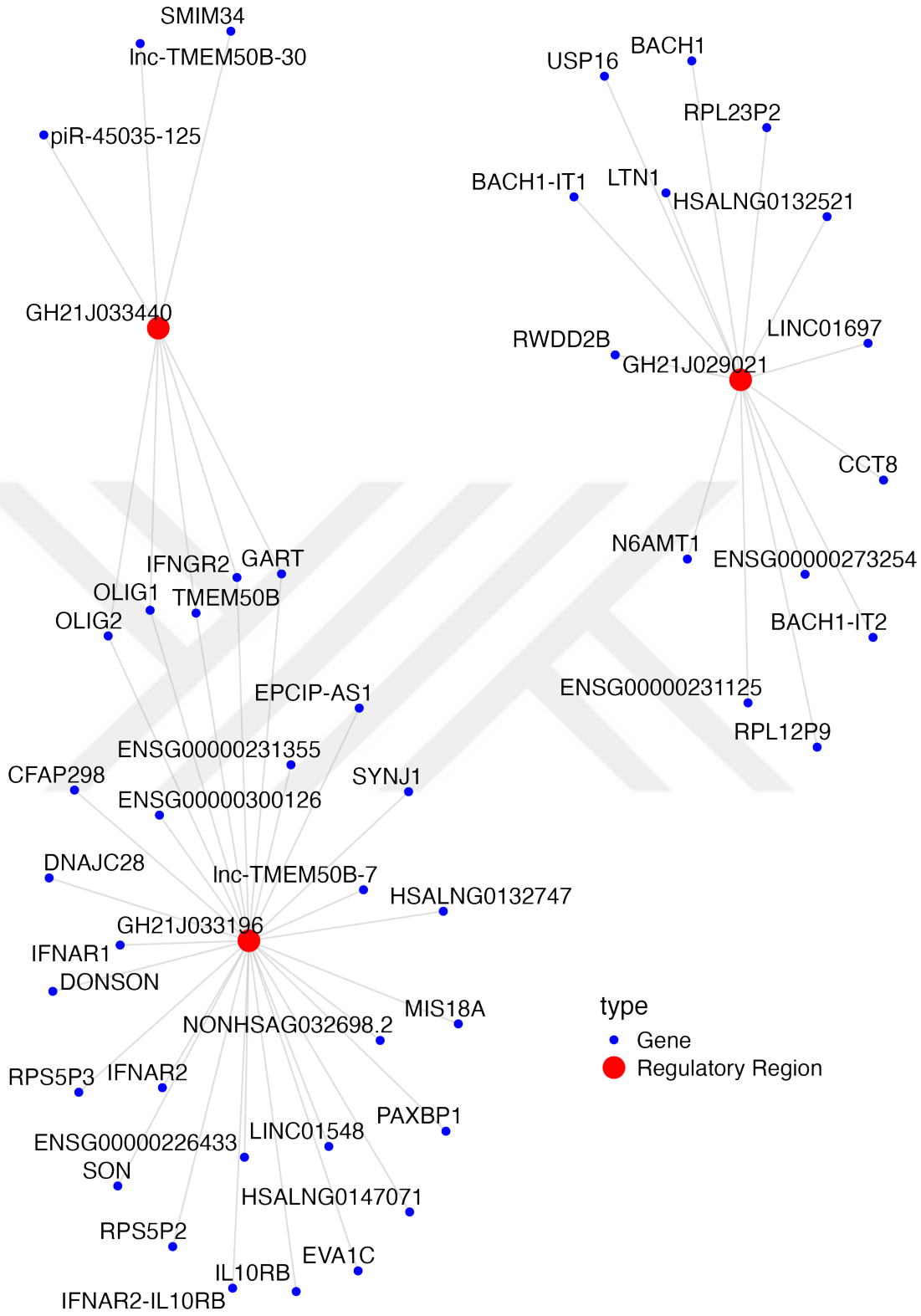


Figure 3.51: Regulatory Regions That Can be Bound by all Three TFs (BCL11A, IKZF1, CHD4) and their Target Genes.

Focusing on three example genes—PKNOX1, RUNX1, and IFNAR2—we found that each was targeted by multiple regulatory regions, but no single TF bound all of them simultaneously (Figure3.52). Instead, BCL11A, IKZF1, and CHD4 were distributed across different enhancer elements, implying a more dispersed regulatory influence. Rather than following a strict loop-based model, these genes appear to rely on context-dependent enhancer usage, where different enhancers are engaged dynamically un- der specific conditions. This behavior corresponds closely to the "dynamic enhancer-driven" scenario, where the changes in regulatory interactions (rather than their total number) correlate with gene expression levels. Unlike Group 1, where stability and persistence drive expression, Group 2 appears to function through regulatory rewiring, meaning that specific enhancers take on dominant roles at different time points. This suggests a fine-tuning mechanism, where different enhancer sets provide conditional activation depending on cellular states.

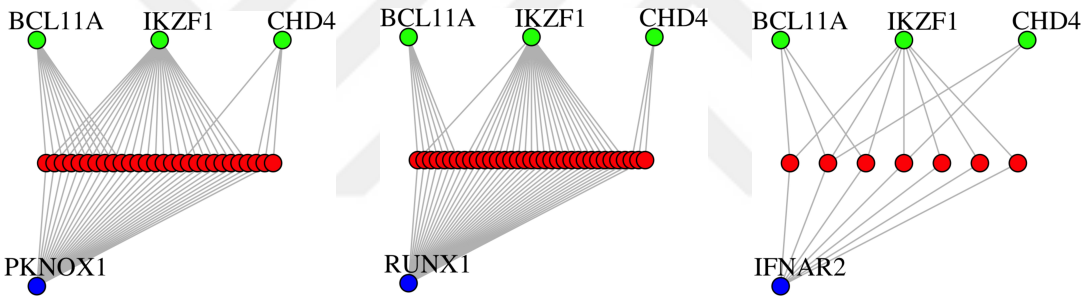


Figure 3.52: Network visualizations of PKNOX1, RUNX1, and IFNAR2.

While this flexibility may be advantageous for dynamic gene expression control, it also introduces more regulatory variability. The distinction between functional adaptability (transcriptional robustness) and stochastic noise remains an open question. However, if genes from Group 2 partially overlap with enhancer mode-of-action categories identified in prior analyses, it could provide deeper insights into how different regulatory mechanisms contribute to transcriptional control.

This study highlights the importance of integrating TF complex dynamics, enhancer-promoter interactions, and network-based enhancer classifications. While the origi-

nal protein complex analysis was conducted in aging T cells, the focused analysis of BCL11A, IKZF1, and CHD4 interactions in chromosome 21 regulatory regions provided a framework to explore their role in shaping transcriptional networks. If genes from Group 1 align with the stable enhancer-promoter category, it suggests that certain transcriptional loops may be reinforced through stable TF occupancy, providing a persistent regulatory scaffold. Conversely, the dynamic nature of Group 2 enhancer interactions implies that some genes rely more on enhancer switching and reconfiguration than on pre-existing structural loops. This distinction between stable and flexible enhancer usage could have profound implications for understanding gene regulation during differentiation and disease progression.

By dissecting enhancer-promoter interactions into structurally constrained vs. dynamically flexible networks, this study provides a useful framework for understanding transcriptional regulation. Future research integrating higher-resolution Hi-C data, TF binding kinetics, and temporal enhancer activity profiling will be crucial in refining these classifications. Additionally, validating these interactions in matched biological contexts (e.g., linking the T cell complex analysis to T cell-related enhancer studies) will offer new insights into how regulatory networks function across biological systems. Ultimately, these findings emphasize the importance of bridging 3D genome architecture, protein complex dynamics, and enhancer-driven regulatory networks to achieve a comprehensive understanding of gene expression control.

4. CONCLUSION

This thesis provides an integrative approach to understanding transcriptional regulation through enhancer-promoter interactions and transcriptional protein complexes, with a focus on enhancer multiplicity and its implications in diverse biological contexts. By addressing the limitations of current Hi-C data resolution, benchmarking methods for enhancing data fidelity, and leveraging high-resolution datasets, this study offers a nuanced understanding of chromatin architecture. The evaluation of Hi-C enhancement methods highlighted both their potential and limitations, particularly in defining inter-regional contacts, reinforcing the need for caution in their application.

The network-based framework used to analyze chromatin interactions represents a significant advance over traditional pairwise analyses. This approach enabled the identification of multi-way enhancer-promoter interactions and their correlation with gene expression changes, particularly in the context of cardiomyocyte differentiation. Despite limited data availability, this analysis successfully characterized the enhancer mode of action, emphasizing the utility of network-based methodologies in elucidating the regulatory dynamics of chromatin interactions.

Further, the investigation of transcriptional protein complexes in aging immune T cells complements the enhancer-promoter analyses. The results revealed key transcription factors enriched in altered protein complexes that align with findings from previous epigenomic and genomic accessibility studies. This convergence underscores the interplay between chromatin accessibility, regulatory complex dynamics, and gene expression, providing a comprehensive view of the regulatory mechanisms driving cellular aging and differentiation.

In this thesis, I present two distinct yet conceptually interconnected analyses: enhancer-promoter network dynamics in cardiac differentiation and deregulated transcriptional protein complexes in aging T cells. Both approaches independently illuminate critical regulatory mechanisms, with the former elucidating enhancer-driven gene expression changes and the latter revealing altered TF interactions within protein complexes during aging. While a direct link between these findings lies beyond the scope of this work, their intersection offers a compelling avenue for future research. Specifically, TFs iden-

tified within enhancer-regulated groups could serve as key components of deregulated complexes, suggesting a shared regulatory framework. Moreover, exploring enhancer enrichment for TFs implicated in transcriptional complexes may uncover novel mechanisms of transcriptional control. These promising results lay the groundwork for integrating these analyses to deepen our understanding of TF-driven regulation in both differentiation and aging processes, advancing the field of systems biology and transcriptional regulation.

This study not only highlights the distinct regulatory mechanisms of enhancer action but also bridges findings across different biological systems, demonstrating the versatility of the developed methodologies. By integrating omics datasets and network analysis, it contextualizes the dynamic interplay between 3D genome architecture and transcriptional regulation, advancing our understanding of pathological phenotypes and providing a foundation for future research in both cellular differentiation and immune aging.

Ultimately, this work underscores the critical role of integrative computational frameworks in uncovering the intricacies of transcriptional regulation, offering insights that are both parallel to and complementary with prior studies in epigenomics. These findings pave the way for further exploration of enhancer-promoter interactions and transcriptional complexes, contributing to the broader understanding of genome function and regulation in health and disease.

REFERENCES

[1] Finotti A., Fabbri E., Lampronti I., Gasparello J., Borgatti M., Gambari R., (2019), "Micrornas and long non-coding rnas in genetic diseases", Molecular Diagnosis & Therapy, 23, 155–171.

[2] Maston G. A., Evans S. K., Green M. R., (2006), "Transcriptional regulatory elements in the human genome", Annual Review of Genomics and Human Genetics, 7, 29–59.

[3] Poulos R. C., Sloane M. A., Hesson L. B., Wong J. W. H., (2015), "The search for cis-regulatory driver mutations in cancer genomes", Oncotarget, 6, 32509–32525.

[4] Mora A., Sandve G. K., Gabrielsen O. S., Eskeland R., (2015), "In the loop: promoter–enhancer interactions and bioinformatics", Briefings in Bioinformatics, bbv097.

[5] Snetkova V., Skok J. A., (2018), "Enhancer talk", Epigenomics, 10, 483–498

[6] Consortium E. P., (2012), "An integrated encyclopedia of dna elements in the human genome", Nature, 489, 57–74.

[7] Kundaje A., Meuleman W., Ernst J., Bilenky M., Yen A., Heravi-Moussavi A., Kheradpour P., Zhang Z., Wang J., Ziller M. J., Amin V., Whitaker J. W., Schultz M. D., Ward L. D., Sarkar A., Quon G., Sandstrom R. S., Eaton M. L., Wu Y.- C., Pfenning A. R., Wang X., Claussnitzer M., Liu Y., Coarfa C., Harris R. A., Shoresh N., Epstein C. B., Gjoneska E., Leung D., Xie W., Hawkins R. D., Lister R., Hong C., Gascard P., Mungall A. J., Moore R., Chuah E., Tam A., CanfieldT. K., Hansen R. S., Kaul R., Sabo P. J., Bansal M. S., Carles A., Dixon J. R., Farh K.-H., Feizi S., Karlic R., Kim A.-R., Kulkarni A., Li D., Lowdon R., Elliott G., Mercer T. R., Neph S. J., Onuchic V., Polak P., Rajagopal N., Ray P., Sallari R. C., Siebenthall K. T., Sinnott-Armstrong N. A., Stevens M., Thurman R. E., Wu J., Zhang B., Zhou X., Beaudet A. E., Boyer L. A., Jager P. L. D., Farnham P. J., Fisher S. J., Haussler D., Jones S. J. M., Li W., Marra M. A., McManus M. T., Sunyaev S., Thomson J. A., Tlsty T. D., Tsai L.-H., Wang W., Waterland R. A., Zhang M. Q., Chadwick L. H., Bernstein B. E., Costello J. F., Ecker J. R., Hirst M., Meissner A., Milosavljevic A., Ren B., Stamatoyannopoulos J. A., Wang T., Kellis M., (2015), "Integrative analysis of 111 reference human epigenomes", Nature, 518, 317–330.

[8] Halfon M. S., (2019), "Studying transcriptional enhancers: The founder fallacy, validation creep, and other biases", Trends in Genetics, 35, 93–103.

[9] Sanyal A., Lajoie B. R., Jain G., Dekker J., (2012), "The long-range interaction landscape of gene promoters", Nature, 489, 109–113.

[10] Hariprakash J. M., Ferrari F., (2019), "Computational biology solutions to identify enhancers-target gene pairs", Computational and Structural Biotechnology Journal, 17, 821–831.

[11] Pennacchio L. A., Bickmore W., Dean A., Nobrega M. A., Bejerano G., (2013), "Enhancers: five essential questions", Nature Reviews Genetics, 14, 288–295.

[12] Lieberman-Aiden E., van Berkum N. L., Williams L., Imakaev M., Ragoczy T., Telling A., Amit I., Lajoie B. R., Sabo P. J., Dorschner M. O., Sandstrom R., Bernstein B., Bender M. A., Groudine M., Gnirke A., Stamatoyannopoulos J., Mirny L. A., Lander E. S., Dekker J., (2009), "Comprehensive mapping of long-range interactions reveals folding principles of the human genome", Science, 326, 289– 293.

[13] Grob S., Cavalli G., (2018), "Technical Review: A Hitchhiker's Guide to Chromosome Conformation Capture", 233–246.

[14] Belaghzal H., Dekker J., Gibcus J. H., (2017), "Hi-c 2.0: An optimized hi-c pro- cedure for high-resolution genome-wide mapping of chromosome conformation", Methods, 123, 56–65.

[15] Lajoie B. R., Dekker J., Kaplan N., (2015), "The hitchhiker's guide to hi-c analy- sis: Practical guidelines", Methods, 72, 65–75.

[16] Gan M., Li W., Jiang R., (2019), "Encontact: predicting enhancer-enhancer con- tacts using sequence-based deep learning model", PeerJ, 7, e7657.

[17] Zhang Y., An L., Xu J., Zhang B., Zheng W. J., Hu M., Tang J., Yue F., (2018), "Enhancing hi-c data resolution with deep convolutional neural network hicplus", Nature Communications, 9, 750.

[18] Srihari S., Yong C. H., Patil A., Wong L., (2015), "Methods for protein complex prediction and their contributions towards understanding the organisation, func- tion and dynamics of complexes", FEBS Letters, 589, 2590–2602.

[19] Zahiri J., Emamjomeh A., Bagheri S., Ivazeh A., Mahdevar G., Tehrani H. S., Mirzaie M., Fakheri B. A., Mohammad-Noori M., (2020), "Protein complex prediction: A survey", Genomics, 112, 174–183.

[20] Nott A., Holtman I. R., Coufal N. G., Schlachetzki J. C. M., Yu M., Hu R., Han C. Z., Pena M., Xiao J., Wu Y., Keulen Z., Pasillas M. P., O'Connor C., Nickl C. K., Schafer S. T., Shen Z., Rissman R. A., Brewer J. B., Gosselin D., Gonda D. D., Levy M. L., Rosenfeld M. G., McVicker G., Gage F. H., Ren B., Glass C. K., (2019), "Brain cell type–specific enhancer–promoter interactome maps and disease - risk association", Science, 366, 1134–1139.

[21] Lin C. Y., Erkek S., Tong Y., Yin L., Federation A. J., Zapatka M., Haldipur P., Kawauchi D., Risch T., Warnatz H.-J., Worst B. C., Ju B., Orr B. A., Zeid R., Polaski D. R., Segura-Wang M., Waszak S. M., Jones D. T. W., Kool M., Hovestadt V., Buchhalter I., Sieber L., Johann P., Chavez L., Gröschel S., Ryzhova M., Kor- shunov A., Chen W., Chizhikov V. V., Millen K. J., Amstislavskiy V., Lehrach H., Yaspo M.-L., Eils R., Lichter P., Korbel J. O., Pfister S. M., Bradner J. E., North- cott P. A., (2016), "Active medulloblastoma enhancers reveal subgroup-specific cellular origins", Nature, 530, 57–62.

[22] Sato T., Yoo S., Kong R., Sinha A., Chandramani-Shivalingappa P., Patel A., Fridrikh M., Nagano O., Masuko T., Beasley M. B., Powell C. A., Zhu J., Watan- abe H., (2019), "Epigenomic profiling discovers trans-lineage sox2 partnershipsdriving tumor heterogeneity in lung squamous cell carcinoma", Cancer Research, 79, 6084–6100.

[23] Liu T., Wang Z., (2019*a*), "Hicnn: a very deep convolutional neural network to better enhance the resolution of hi-c data", Bioinformatics, 35, 4222–4228.

[24] Liu T., Wang Z., (2019*b*), "Hicnn2: Enhancing the resolution of hi-c data using an ensemble of convolutional neural networks", Genes, 10, 862.

[25] Yan K.-K., Yardımcı G. G., Yan C., Noble W. S., Gerstein M., (2017), "Hic-spector: a matrix library for spectral and reproducibility analysis of hi-c contact maps", Bioinformatics, 33, 2199–2201.

[26] Li Z., Dai Z., (2020), "Srhic: A deep learning model to enhance the resolution of hi-c data", Frontiers in Genetics, 11, 1–9.

[27] Liu Q., Lv H., Jiang R., (2019), "hicgan infers super resolution hi-c data with generative adversarial networks", Bioinformatics, 35, i99–i107.

[28] Hong H., Jiang S., Li H., Quan C., Zhao C., Li R., Li W., Du G., Yin X., Huang Y., Li C., Chen H., Bo X., (2019), "Deephic: A generative adversarial network for enhancing hi-c data resolution".

[29] Dimmick M. C., Lee L. J., Frey B. J., (2020), "Hicsr: a hi-c super-resolution framework for producing highly realistic contact maps".

[30] Hu Y., Ma W., (2021), "Enhic: learning fine-resolution hi-c contact maps using a generative adversarial framework", Bioinformatics, 37, i272–i279.

[31] Hicks P., Oluwadare O., (2022), "Hicarn: resolution enhancement of hi-c data using cascading residual networks", Bioinformatics, 38, 2414–2421.

[32] Wang B., Liu K., Li Y., Wang J., (2023), "Dfhic: a dilated full convolution model to enhance the resolution of hi-c data", Bioinformatics.

[33] Highsmith M., Cheng J., (2021), "Vehicle: a variationally encoded hi-c loss en- hancement algorithm for improving and generating hi-c data", Scientific Reports, 11, 8880.

[34] Cameron C. J., Dostie J., Blanchette M., (2018), "Estimating dna-dna interaction frequency from hi-c data at restriction-fragment resolution".

[35] Xu Z., Zhang G., Jin F., Chen M., Furey T. S., Sullivan P. F., Qin Z., Hu M., Li Y., (2016), "A hidden markov random field-based bayesian method for the detection of long-range chromosomal interactions in hi-c data", Bioinformatics, 32, 650– 656.

[36] Yang T., Zhang F., Yardımcı G. G., Song F., Hardison R. C., Noble W. S., Yue F., Li Q., (2017), "Hicrep: assessing the reproducibility of hi-c data using a stratum-adjusted correlation coefficient", Genome Research, 27, 1939–1949.

[37] Forcato M., Nicoletti C., Pal K., Livi C. M., Ferrari F., Bicciato S., (2017), "Comparison of computational methods for hi-c data analysis", Nature Methods, 14, 679–685.Lyu H., Liu E., Wu Z., (2020), "Comparison of normalization methods for hi-c data", BioTechniques, 68, 56–64.

[38] Yardımcı G. G., Ozadam H., Sauria M. E. G., Ursu O., Yan K.-K., Yang T., Chakraborty A., Kaul A., Lajoie B. R., Song F., Zhan Y., Ay F., Gerstein M., Kundaje A., Li Q., Taylor J., Yue F., Dekker J., Noble W. S., (2019), "Measuring the reproducibility and quality of hi-c data", Genome Biology, 20, 57.

[39] Paulsen J., Ali T. M. L., Nekrasov M., Delbarre E., Baudement M.-O., Kurscheid S., Tremethick D., Collas P., (2019), "Long-range interactions between topologi- cally associating domains shape the four-dimensional genome during differentia- tion", Nature Genetics, 51, 835–843.

[40] Song W., Sharan R., Ovcharenko I., (2019), "The first enhancer in an enhancer chain safeguards subsequent enhancer-promoter contacts from a distance", Genome Biology, 20, 197.

[41] Rao S. S., Huntley M. H., Durand N. C., Stamenova E. K., Bochkov I. D., Robin- son J. T., Sanborn A. L., Machol I., Omer A. D., Lander E. S., Aiden E. L., (2014), "A 3d map of the human genome at kilobase resolution reveals principles of chro- matin looping", Cell, 159, 1665–1680.

[42] Thiel D., Conrad N. D., Ntini E., Peschutter R. X., Siebert H., Marsico A., (2019), "Identifying lncrna-mediated regulatory modules via chia-pet network analysis", BMC Bioinformatics, 20, 292.

[43] Chen H., Xiao J., Shao T., Wang L., Bai J., Lin X., Ding N., Qu Y., Tian Y., Chen X., Liu H., Liu H., Xu J., Li X., (2019), "Landscape of enhancer-enhancer cooperative regulation during human cardiac commitment", Molecular Therapy - Nucleic Acids, 17, 840–851.

[44] Zheng B., Wang J., Tang L., Shi J., Zhu D., (2017), "mtorc1 and mtorc2 play different roles in regulating cardiomyocyte differentiation from embryonic stem cells", The International Journal of Developmental Biology, 61, 65–72.

[45] Chan S. S.-K., Li H.-J., Hsueh Y.-C., Lee D. S., Chen J.-H., Hwang S.-M., Chen C.-Y., Shih E., Hsieh P. C. H., (2010), "Fibroblast growth factor-10 promotes cardiomyocyte differentiation from embryonic and induced pluripotent stem cells", PLoS ONE, 5, e14414.

[46] Holtzinger A., Rosenfeld G. E., Evans T., (2010), "Gata4 directs development of cardiac-inducing endoderm from es cells", Developmental Biology, 337, 63–73.

[47] Lee J. H., Protze S. I., Laksman Z., Backx P. H., Keller G. M., (2017), "Human pluripotent stem cell-derived atrial and ventricular cardiomyocytes develop from distinct mesoderm populations", Cell Stem Cell, 21, 179–194.e4.

[48] Rysä J., Tenhunen O., Serpi R., Soini Y., Nemer M., Leskinen H., Ruskoaho H., (2010), "Gata-4 is an angiogenic survival factor of the infarcted heart", Circulation: Heart Failure, 3, 440–450.

[49] Nakajima K., Inagawa M., Uchida C., Okada K., Tane S., Kojima M., Kubota M., Noda M., Ogawa S., Shirato H., Sato M., Suzuki-Migishima R., Hino T., Satoh Y., Kitagawa M., Takeuchi T., (2011), "Coordinated regulation of differentiation and proliferation of embryonic cardiomyocytes by a jumonji (jarid2)-cyclin d1

pathway", Development, 138, 1771–1782.

[50] Burridge P. W., Matsa E., Shukla P., Lin Z. C., Churko J. M., Ebert A. D., Lan F., Diecke S., Huber B., Mordwinkin N. M., Plews J. R., Abilez O. J., Cui B., Gold J. D., Wu J. C., (2014), "Chemically defined generation of human cardiomy- ocytes", Nature Methods, 11, 855–860.

[51] Kim Y.-J., Tamadon A., Kim Y.-Y., Kang B.-C., Ku S.-Y., (2021), "Epigenetic reg- ulation of cardiomyocyte differentiation from embryonic and induced pluripotent stem cells", International Journal of Molecular Sciences, 22, 8599.

[52] Borden A., Kurian J., Nickoloff E., Yang Y., Troupes C. D., Ibetti J., Lucchese A. M., Gao E., Mohsin S., Koch W. J., Houser S. R., Kishore R., Khan M., (2019), "Transient introduction of mir-294 in the heart promotes cardiomyocyte cell cycle reentry after injury", Circulation Research, 125, 14–25.

[53] Vilches J., Pulido A., Hernández-Torres F., Franco D., Aránega A., (2014), "mir- 27b and mir-23b modulate cardiomyocyte differentiation from mouse embryonic stem cells", Journal of Cardiovascular Development and Disease, 1, 41–51.

[54] Zhen L.-X., Gu Y.-Y., Zhao Q., Zhu H.-F., Lü J.-H., Li S.-J., Xu Z., Li L., Yu Z.-
R., (2019), "Mir-301a promotes embryonic stem cell differentiation to cardiomy- ocytes", World Journal of Stem Cells, 11, 1130–1141.

[55] Giurgiu M., Reinhard J., Brauner B., Dunger-Kaltenbach I., Fobo G., Frishman G., Montrone C., Ruepp A., (2019), "Corum: the comprehensive resource of mam- malian protein complexes—2019", Nucleic Acids Research, 47, D559–D563.

[56] Meldal B. H., Forner-Martinez O., Costanzo M. C., Dana J., Demeter J., Du- mousseau M., Dwight S. S., Gaulton A., Licata L., Melidoni A. N., Ricard-Blum S., Roechert B., Skyzypek M. S., Tiwari M., Velankar S., Wong E. D., Hermjakob H., Orchard S., (2015), "The complex portal - an encyclopaedia of macromolec- ular complexes", Nucleic Acids Research, 43, D479–D484.

[57] Spirin V., Mirny L. A., (2003), "Protein complexes and functional modules in molecular networks", Proceedings of the National Academy of Sciences, 100, 12123–12128.

[58] Dongen S. V., (2008), "Graph clustering via a discrete uncoupling process", SIAM Journal on Matrix Analysis and Applications, 30, 121–141.

[59] Michalak W., Tsiamis V., Schwämmle V., Rogowska-Wrzesińska A., (2019), "Complexbrowser: A tool for identification and quantification of protein com- plexes in large-scale proteomics datasets", Molecular & Cellular Proteomics, 18, 2324–2334.

[60] Will T., Helms V., (2019), "Differential analysis of combinatorial protein com- plexes with complexchange", BMC Bioinformatics, 20, 1–14.

[61] Rizzetto S., Moyseos P., Baldacci B., Priami C., Csikász-Nagy A., (2018), "Context-dependent prediction of protein complexes by sicompre", npj Systems

Biology and Applications, 4, 37.

[62] Will T., Helms V., (2016), “Ppixpress: Construction of condition-specific protein interaction networks based on transcript expression”, Bioinformatics, 32, 571–578.

[63] Will T., Helms V., (2014), “Identifying transcription factor complexes and their roles”, Bioinformatics, 30, 415–421.

[64] Weiskopf D., Weinberger B., Grubeck-Loebenstein B., (2009), “The aging of the immune system”, Transplant International, 22, 1041–1050.

[65] Nidadavolu L. S., Walston J. D., (2021), “Underlying vulnerabilities to the cytokine storm and adverse covid-19 outcomes in the aging immune system”, The Journals of Gerontology: Series A, 76, e13–e18.

[66] Thomas R., Wang W., Su D.-M., (2020), “Contributions of age-related thymic involution to immunosenescence and inflammaging”, Immunity & Ageing, 17, 2.

[67] Tu W., Rao S., (2016), “Mechanisms underlying t cell immunosenescence: Aging and cytomegalovirus infection”, Frontiers in Microbiology.

[68] Hu B., Jadhav R. R., Gustafson C. E., Saux S. L., Ye Z., Li X., Tian L., Weyand C. M., Goronzy J. J., (2020), “Distinct age-related epigenetic signatures in cd4 and cd8 t cells”, Frontiers in Immunology.

[69] Li S., Wu B., Ling Y., Guo M., Qin B., Ren X., Wang C., Yang H., Chen L., Liao Y., Liu Y., Peng X., Xu C., Wang Z., Shen Y., Chen J., Liu L., Niu B., Zhu M., Liu L., Li F., Zhu T., Zhu Z., Zhou X., Lu H., (2021), “Epigenetic landscapes of single-cell chromatin accessibility and transcriptomic immune profiles of t cells in covid-19 patients”, Frontiers in Immunology.

[70] Page A. L., Dupuis G., Larbi A., Witkowski J. M., Fülöp T., (2018), “Signal transduction changes in cd4 + and cd8 + t cell subpopulations with aging”, Experimental Gerontology, 105, 128–139.

[71] Moskowitz D. M., Zhang D. W., Hu B., Saux S. L., Yanes R. E., Ye Z., Buenrostro J. D., Weyand C. M., Greenleaf W. J., Goronzy J. J., (2017), “Epigenomics of human cd8 t cell differentiation and aging”, Science Immunology.

[72] Zheng Y., Liu X., Le W., Xie L., Li H., Wen W., Wang S., Ma S., Huang Z., Ye J., Shi W., Ye Y., Liu Z., Song M., Zhang W., Han J.-D. J., Belmonte J. C. I., Xiao C., Qu J., Wang H., Liu G.-H., Su W., (2020), “A human circulating immune cell landscape in aging and covid-19”, Protein & Cell, 11, 740–770.

[73] Thomson Z., He Z., Swanson E., Henderson K., Phalen C., Zaim S. R., Pebworth M.-P., Okada L. Y., Heubeck A. T., Roll C. R., Hernandez V., Weiss M., Genge P. C., Reading J., Giles J. R., Manne S., Dougherty J., Jasen C. J., Greenplate A. R., Becker L. A., Graybuck L. T., Vasaikar S. V., Szeto G. L., Savage A. K., Speake C., Buckner J. H., jun Li X., Bumol T. F., Wherry E., Torgerson T. R., Vella L. A., Henrickson S. E., Skene P. J., Gustafson C. E., (2023), “Trimodal single- cell profiling reveals a novel pediatric cd8αα+ t cell subset and broad age-related molecular reprogramming across the t cell compartment”, Nature Immunology, 24, 1947–1959.

[74] Ucar D., Márquez E. J., Chung C.-H., Marches R., Rossi R. J., Uyar A., Wu T.- C., George J., Stitzel M. L., Palucka A. K., Kuchel G. A., Banchereau J., (2017), "The chromatin accessibility signature of human immune aging stems from cd8+ t cells", Journal of Experimental Medicine, 214, 3123–3144.

[75] Wang Z., Wang P., Li Y., Peng H., Zhu Y., Mohandas N., Liu J., (2021), "Interplay between cofactors and transcription factors in hematopoiesis and hematological malignancies", Signal Transduction and Targeted Therapy.

[76] Dong Y., Mo X., Hu Y., Qi X., Jiang F., Jiang Z., Tong S., (2020), "Epidemiology of covid-19 among children in china", Pediatrics.

[77] Cruz A. T., Zeichner S. L., (2020), "Covid-19 in children: Initial characterization of the pediatric disease", Pediatrics.

[78] Jovanovic M., Sekulic S., Jocic M., Jurisevic M., Gajovic N., Jovanovic M., Ar- senijevic N., Jovanovic M., Mijailovic M., Milosavljevic M., Jovanovic I., (2023), "Increased pro th1 and th17 transcriptional activity in patients with severe covid- 19", International Journal of Medical Sciences, 20, 530–541.

[79] Lin Y., Li Y., Chen H., Meng J., Li J., Chu J., Zheng R., Wang H., Pan P., Su J., Jiang J., Ye L., Liang H., An S., (2023), "Weighted gene co-expression network analysis revealed t cell differentiation associated with the age-related phenotypes in covid-19 patients", BMC Medical Genomics, 16, 59.

[80] Zufferey M., Tavernari D., Oricchio E., Ciriello G., (2018), "Comparison of com- putational methods for the identification of topologically associating domains", Genome Biology, 19, 1–18.

[81] Cresswell K. G., Dozmorov M. G., (2020), "Tadcompare: An r package for dif- ferential and temporal analysis of topologically associated domains", Frontiers in Genetics, 11, 1–16.

[82] Bertero A., Fields P. A., Ramani V., Bonora G., Yardimci G. G., Reinecke H., Pabon L., Noble W. S., Shendure J., Murry C. E., (2019), "Dynamics of genome reorganization during human cardiogenesis reveal an rbm20-dependent splicing factory", Nature Communications, 10, 1–19.

[83] Fishilevich S., Nudel R., Rappaport N., Hadar R., Plaschkes I., Stein T. I., Rosen N., Kohn A., Twik M., Safran M., Lancet D., Cohen D., (2017), "Genehancer: genome-wide integration of enhancers and target genes in genecards", Database : the journal of biological databases and curation, 2017, 1–17.

[84] Perez G., Barber G. P., Benet-Pages A., Casper J., Clawson H., Diekhans M., Fischer C., Gonzalez J. N., Hinrichs A. S., Lee C. M., Nassar L. R., Raney B. J., Speir M. L., van Baren M. J., Vaske C. J., Haussler D., Kent W. J., Haeussler M., (2025), "The ucsc genome browser database: 2025 update", Nucleic Acids Research, 53, D1243–D1249.

[85] Durand N. C., Robinson J. T., Shamim M. S., Machol I., Mesirov J. P., Lander E. S., Aiden E. L., (2016), "Juicebox provides a visualization system for hi-c contact maps with unlimited zoom", Cell Systems, 3, 99–101.

[86] Yoshida M., Worlock K. B., Huang N., Lindeboom R. G., Butler C. R., Kumasaka N., Conde C. D., Mamanova L., Bolt L., Richardson L., Polanski K., Madissoon E., Barnes J. L., Allen-Hyttinen J., Kilich E., Jones B. C., de Wilton A., Wilbrey- Clark A., Sungnak W., Pett J. P., Weller J., Prigmore E., Yung H., Mehta P., Saleh A., Saigal A., Chu V., Cohen J. M., Cane C., Iordanidou A., Shibuya S., Reuschl A. K., Herczeg I. T., Argento A. C., Wunderink R. G., Smith S. B., Poor T. A., Gao C. A., Dematte J. E., Budinger G. R., Donnelly H. K., Markov N. S., Lu Z., Reynolds G., Haniffa M., Bowyer G. S., Coates M., Clatworthy M. R., Calero-Nieto F. J., Göttgens B., O'Callaghan C., Sebire N. J., Jolly C., Coppi P. D., Smith C. M., Misharin A. V., Janes S. M., Teichmann S. A., Nikolić M. Z., Meyer K. B., (2021), "Local and systemic responses to sars-cov-2 infection in children and adults", Nature 2021 602:7896, 602, 321–327.

[87] Ahern D. J., Ai Z., Ainsworth M., Allan C., Allcock A., Angus B., Ansari M. A., Arancibia-Cárcamo C. V., Aschenbrenner D., Attar M., Baillie J. K., Barnes E., Bashford-Rogers R., Bashyal A., Beer S., Berridge G., Beveridge A., Bibi S., Bi- canic T., Blackwell L., Bowness P., Brent A., Brown A., Broxholme J., Buck D., Burnham K. L., Byrne H., Camara S., Ferreira I. C., Charles P., Chen W., Chen Y. L., Chong A., Clutterbuck E. A., Coles M., Conlon C. P., Cornall R., Cribbs A. P., Curion F., Davenport E. E., Davidson N., Davis S., Dendrou C. A., Dequaire J., Dib L., Docker J., Dold C., Dong T., Downes D., Drakesmith H., Dunachie S. J., Duncan D. A., Eijsbouts C., Esnouf R., Espinosa A., Etherington R., Fairfax B., Fairhead R., Fang H., Fassih S., Felle S., Mendoza M. F., Ferreira R., Fischer R., Foord T., Forrow A., Frater J., Fries A., Sanchez V. G., Garner L. C., Geeves C., Georgiou D., Godfrey L., Golubchik T., Vazquez M. G., Green A., Harper H., Harrington H. A., Heilig R., Hester S., Hill J., Hinds C., Hird C., Ho L. P., Hoekzema R., Hollis B., Hughes J., Hutton P., Jackson-Wood M. A., Jainarayanan A., James-Bott A., Jansen K., Jeffery K., Jones E., Jostins L., Kerr G., Kim D., Klenerman P., Knight J. C., Kumar V., Sharma P. K., Kurupati P., Kwok A., Lee A., Linder A., Lockett T., Lonie L., Lopopolo M., Lukoseviciute M., Luo J., Marinou S., Marsden B., Martinez J., Matthews P. C., Mazurczyk M., McGowan S., McKechnie S., Mead A., Mentzer A. J., Mi Y., Monaco C., Montadon R., Napolitani G., Nassiri I., Novak A., O'Brien D. P., O'Connor D., O'Donnell D., Ogg G., Overend L., Park I., Pavord I., Peng Y., Penkava F., Pinho M. P., Perez E., Pollard A. J., Powrie F., Psaila B., Quan T. P., Repapi E., Revale S., Silva-Reyes L., Richard J. B., Rich-Griffin C., Ritter T., Rollier C. S., Row- land M., Ruehle F., Salio M., Sansom S. N., Peres R. S., Delgado A. S., Sauka- Spengler T., Schwessinger R., Scozzafava G., Screaton G., Seigal A., Semple M. G., Sergeant M., Karali C. S., Sims D., Skelly D., Slawinski H., Sobrinodiaz A., Sousos N., Stafford L., Stockdale L., Strickland M., Sumray O., Sun B., Tay- lor C., Taylor S., Taylor A., Thongjuea S., Thraves H., Todd J. A., Tomic A., Tong O., Trebes A., Trzupek D., Tucci F. A., Turtle L., Udalova I., Uhlig H., van Grinsven E., Vendrell I., Verheul M., Voda A., Wang G., Wang L., Wang D., Watkinson P., Watson R., Weinberger M., Whalley J., Witty L., Wray K., Xue L., Yeung H. Y., Yin Z., Young R. K., Youngs J., Zhang P., Zurke Y. X., (2022), "A blood atlas of covid-19 defines hallmarks of disease severity and specificity", Cell, 185, 916–938.e58.

[88] Williams C. R., Baccarella A., Parrish J. Z., Kim C. C., (2017), "Empirical assess- ment of analysis workflows for differential expression analysis of human samples using rna-seq", BMC Bioinformatics, 18, 1–12.

[89] Zhang Q. C., Petrey D., Garzón J. I., Deng L., Honig B., (2013), "Preppi: a structure-informed database of protein–protein interactions", Nucleic Acids Research, 41, D828.

[90] Mosca R., Céol A., Stein A., Olivella R., Aloy P., (2014), "3did: a catalog of domain-based interactions of known three-dimensional structure", Nucleic Acids

Research, 42, D374.

[91] Kulakovskiy I. V., Vorontsov I. E., Yevshin I. S., Soboleva A. V., Kasianov A. S., Ashoor H., Ba-Alawi W., Bajic V. B., Medvedeva Y. A., Kolpakov F. A., Makeev V. J., (2016), "Hocomoco: expansion and enhancement of the collection of transcription factor binding sites models", Nucleic Acids Research, 44, D116–D125.

[92] Kolberg L., Raudvere U., Kuzmin I., Adler P., Vilo J., Peterson H., (2023), "g:profiler—interoperable web service for functional enrichment analysis and gene identifier mapping (2023 update)", Nucleic Acids Research, 51, W207– W212.

[93] Huang D. W., Sherman B. T., Lempicki R. A., (2009), "Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists", Nucleic Acids Research, 37, 1–13.

[94] Desiere F., (2006), "The peptideatlas project", Nucleic Acids Research, 34, D655– D658.

[95] Drew K., Wallingford J. B., Marcotte E. M., (2021), "hu.map 2.0: integration of over 15,000 proteomic experiments builds a global compendium of human multiprotein assemblies", Molecular Systems Biology.

[96] Tsitsiridis G., Steinkamp R., Giurgiu M., Brauner B., Fobo G., Frishman G., Mon- trone C., Ruepp A., (2023), "Corum: the comprehensive resource of mammalian protein complexes–2022", Nucleic Acids Research, 51, D539–D545.

[97] Gao S.-H., Cheng M.-M., Zhao K., Zhang X.-Y., Yang M.-H., Torr P., (2021), "Res2net: A new multi-scale backbone architecture", IEEE Transactions on Pat- tern Analysis and Machine Intelligence, 43, 652–662.

[98] Zhao Z., Hu J., Zeng Z., Yang X., Qian P., Veeravalli B., Guan C., (2022), Mmgl: Multi-scale multi-view global-local contrastive learning for semi-supervised car- diac image segmentation, in "2022 IEEE International Conference on Image Processing (ICIP)", IEEE, 401–405.

[99] Kumar P., Goel N., (2020), Image resolution enhancement using convolutional autoencoders, in "Proceedings of 7th International Electronic Conference on Sen- sors and Applications", MDPI, 8259.

[100] Liu X., Gao Z., Chen B. M., (2020), "Mlfcgan: Multilevel feature fusion-based conditional gan for underwater image color correction", IEEE Geoscience and Remote Sensing Letters, 17, 1488–1492.

[101] Deabes W., Abdel-Hakim A. E., Bouazza K. E., Althobaiti H., (2022), "Adversarial resolution enhancement for electrical capacitance tomography image reconstruction", Sensors, 22, 3142.

[102] Tang Y., Zhong J., Wang Z., Zeng J., (2024), Swinfusion: attention fusion super-resolution network based on swinir, in J. Zhang N. Sun, eds, "Third Interna- tional Conference on Electronic Information Engineering, Big Data, and Com- puter Technology (EIBDCT 2024)", SPIE, 82.

[103] Chen X., Zhang K., Abdoli N., Gilley P. W., Wang X., Liu H., Zheng B., Qiu Y., (2022), "Transformers improve breast cancer diagnosis from unregistered multi- view mammograms", Diagnostics, 12, 1549.

[104] Ren J., Qin J., Ma Q., Cao Y., (2024), Autoface: a novel method of automatic high-resolution face manipulation, in C. Qin H. Zhou, eds, "International Confer- ence on Image Processing and Artificial Intelligence (ICIPAl 2024)", SPIE, 13.

[105] Kökrek E., Pir P., (2024), "Distinct deregulation trends of transcriptional protein complexes in aging naive t cells", Journal of Leukocyte Biology.

[106] Lee J. S., Lee W.-W., Kim S. H., Kang Y., Lee N., Shin M. S., Kang S. W., Kang I., (2011), "Age-associated alteration in naive and memory th17 cell response in humans", Clinical Immunology, 140, 84–91.

[107] Yoon J.-H., Sudo K., Kuroda M., Kato M., Lee I.-K., Han J. S., Nakae S., Ima- mura T., Kim J., Ju J. H., Kim D.-K., Matsuzaki K., Weinstein M., Matsumoto I., Sumida T., Mamura M., (2015), "Phosphorylation status determines the oppos- ing functions of smad2/smad3 as stat3 cofactors in th17 differentiation", Nature Communications, 6, 7600.

[108] Jia M., Zou X., Yin S., Tian W., Zhao Y., Wang H., Xu G., Cai W., Shao Q., (2022), "Chd4 orchestrates the symphony of t and b lymphocytes development and a good mediator in preventing from autoimmune disease", Immunity, Inflam- mation and Disease.

[109] Pan M.-R., Hsieh H.-J., Dai H., Hung W.-C., Li K., Peng G., Lin S.-Y., (2012), "Chromodomain helicase dna-binding protein 4 (chd4) regulates homol- ogous recombination dna repair, and its deficiency sensitizes cells to poly(adp- ribose) polymerase (parp) inhibitor treatment", Journal of Biological Chemistry, 287, 6764–6772.

[110] Bateman A., Martin M.-J., Orchard S., Magrane M., Ahmad S., Alpi E., Bowler-Barnett E. H., Britto R., Bye-A-Jee H., Cukura A., Denny P., Dogan T., Ebenezer T., Fan J., Garmiri P., da Costa Gonzales L. J., Hatton-Ellis E., Hussein A., Ig- natchenko A., Insana G., Ishtiaq R., Joshi V., Jyothi D., Kandasaamy S., Lock A., Luciani A., Lugaric M., Luo J., Lussi Y., MacDougall A., Madeira F., Mahmoudy M., Mishra A., Moulang K., Nightingale A., Pundir S., Qi G., Raj S., Raposo P., Rice D. L., Saidi R., Santos R., Speretta E., Stephenson J., Totoo P., Turner E., Tyagi N., Vasudev P., Warner K., Watkins X., Zaru R., Zellner H., Bridge A. J., Aimo L., Argoud-Puy G., Auchincloss A. H., Axelsen K. B., Bansal P., Baratin D., Neto T. M. B., Blatter M.-C., Bolleman J. T., Boutet E., Breuza L., Gil B. C., Casals-Casas C., Echioukh K. C., Coudert E., Cuche B., de Castro E., Estreicher A., Famiglietti M. L., Feuermann M., Gasteiger E., Gaudet P., Gehant S., Gerrit- sen V., Gos A., Gruaz N., Hulo C., Hyka-Nouspikel N., Jungo F., Kerhornou A., Mercier P. L., Lieberherr D., Masson P., Morgat A., Muthukrishnan V., Paesano S., Pedruzzi I., Pilbout S., Pourcel L., Poux S., Pozzato M., Pruess M., Redaschi N., Rivoire C., Sigrist C. J. A., Sonesson K., Sundaram S., Wu C. H., Arighi C. N., Arminski L., Chen C., Chen Y., Huang H., Laiho K., McGarvey P., Natale D. A., Ross K., Vinayaka C. R., Wang Q., Wang Y., Zhang J., (2023), "Uniprot: the universal protein knowledgebase in 2023", Nucleic Acids Research, 51, D523– D531.

[111] Luc S., Huang J., McEldoon J. L., Somuncular E., Li D., Rhodes C., Mamoor S., Hou S., Xu J., Orkin S. H., (2016), “Bcl11a deficiency leads to hematopoietic stem cell defects with an aging-like phenotype”, Cell Reports, 16, 3181–3194.

BIOGRAPHY

Emel KÖKREK completed a B.Sc. in Chemical Engineering at Boğaziçi University in 2013, followed by an M.Sc. in the Erasmus Mundus BioHealth Computing Master Program at the University of Torino and Joseph Fourier University (2013–2015), where she focused on systems biology modeling. Currently, she is pursuing a Ph.D. in Bioinformatics and Systems Biology at Gebze Technical University, with a thesis titled "Investigation of Enhancer-Protein-Promoter Interactions Within Regulatory Loops." Since 2020, she has been a Research and Teaching Assistant at Kadir Has University, contributing to bioinformatics education and research. Her professional experience includes optimizing NGS pipelines, applying machine learning for genomic analyses, and collaborating on TUBITAK-funded projects. She has also worked as a bioinformatician at PHI Tech Bioinformatics R&D and as a researcher at İstanbul Medipol University's Regenerative Medicine Research Center.

PUBLICATIONS WITHIN the SCOPE of THESIS STUDY

Kökrek E., Pir P., (2024), “Distinct deregulation trends of transcriptional protein complexes in aging naive t cells”, Journal of Leukocyte Biology.

APPENDICES

A. Benchmarking Resolution Enhancement Models

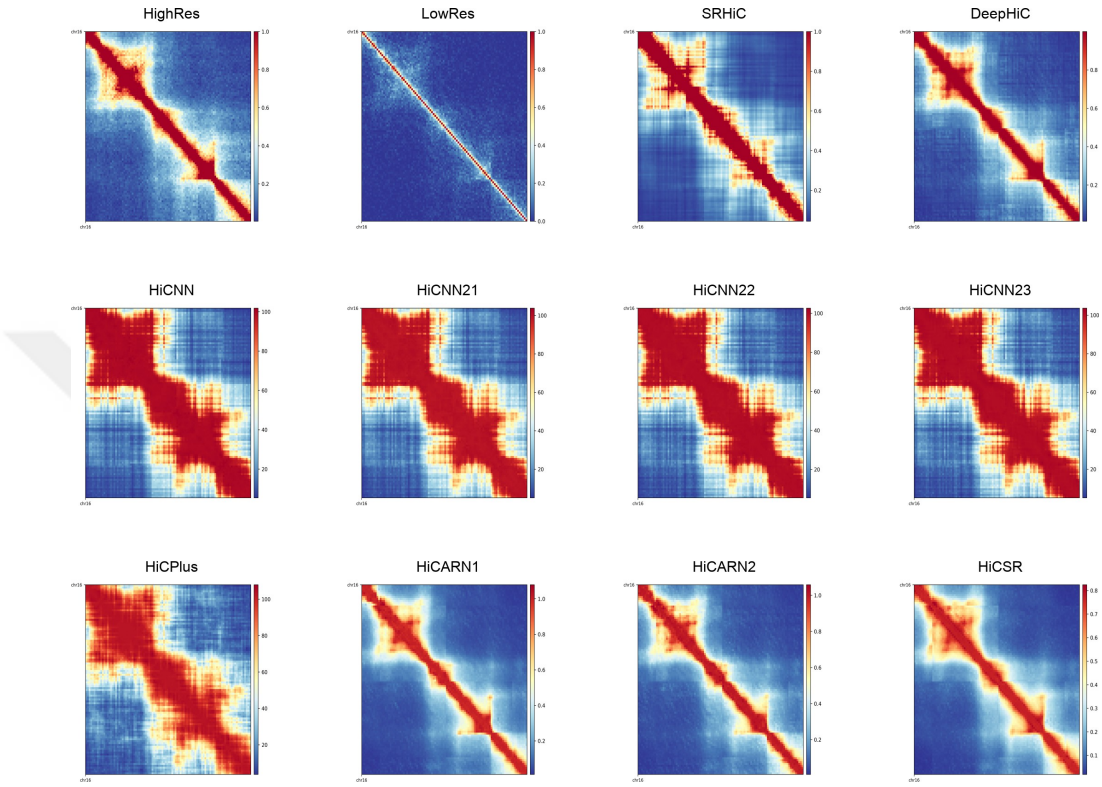


Figure A.1: Heatmap of a 1 Mb region of the GM12878 primary cell line, corresponding to chromosome 16: 3,150,000-4,160,000. The heatmap shows the original high- resolution matrix, the down-sampled low-resolution matrix, and the ×16 enhanced ma- trices generated by various models.

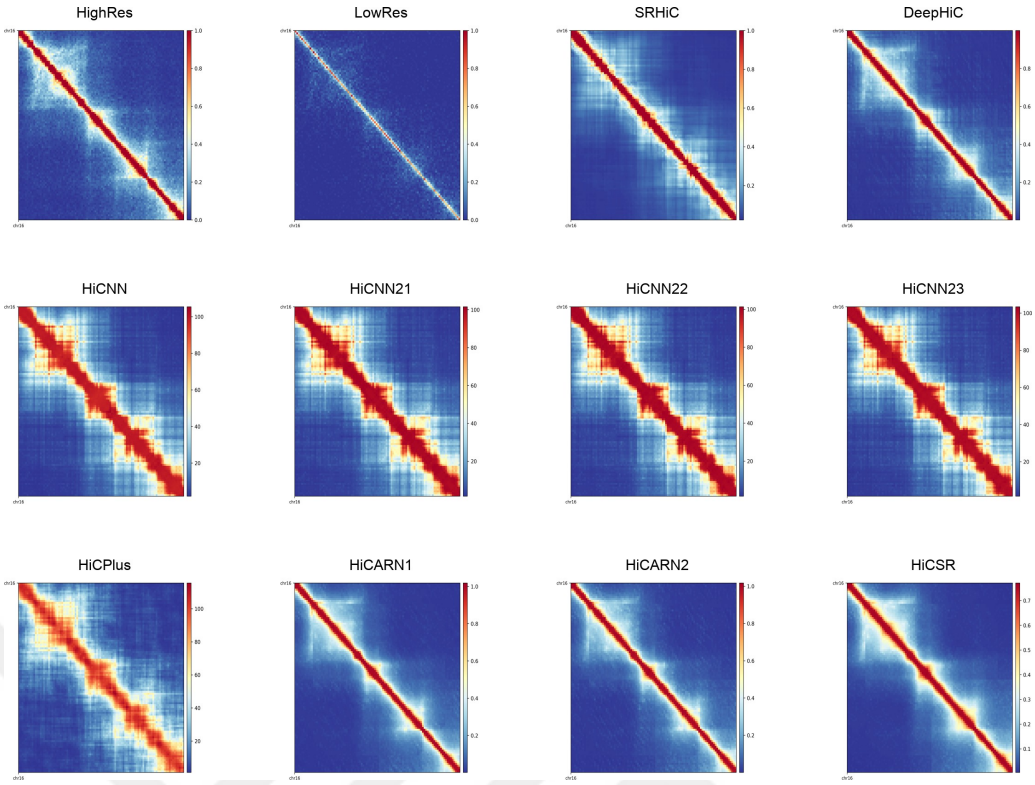


Figure A.2: Heatmap of a 1 Mb region of the IMR90 cell line, corresponding to chromosome 16: 3,150,000-4,160,000. The heatmap shows the original high-resolution matrix, the down-sampled low-resolution matrix, and the ×16 enhanced matrices gen- erated by various models.

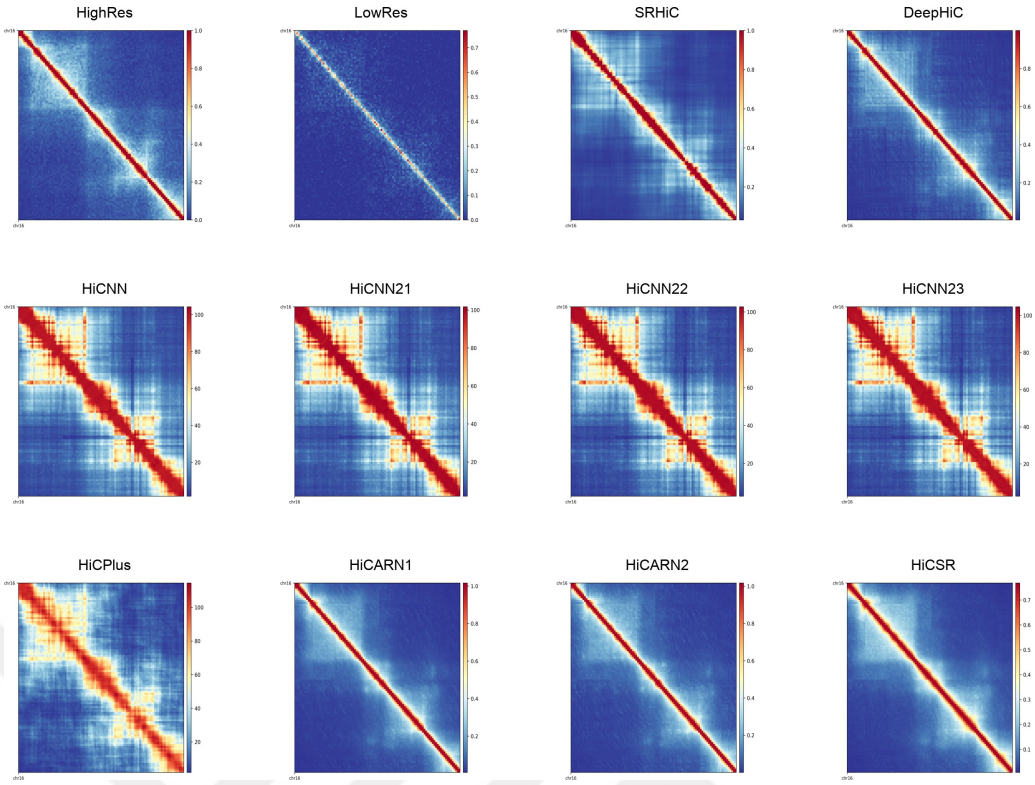


Figure A.3: Heatmap of a 1 Mb region of the K562 cell line, corresponding to chromosome 16: 3,150,000-4,160,000. The heatmap shows the original high-resolution matrix, the down-sampled low-resolution matrix, and the ×16 enhanced matrices gen- erated by various models.

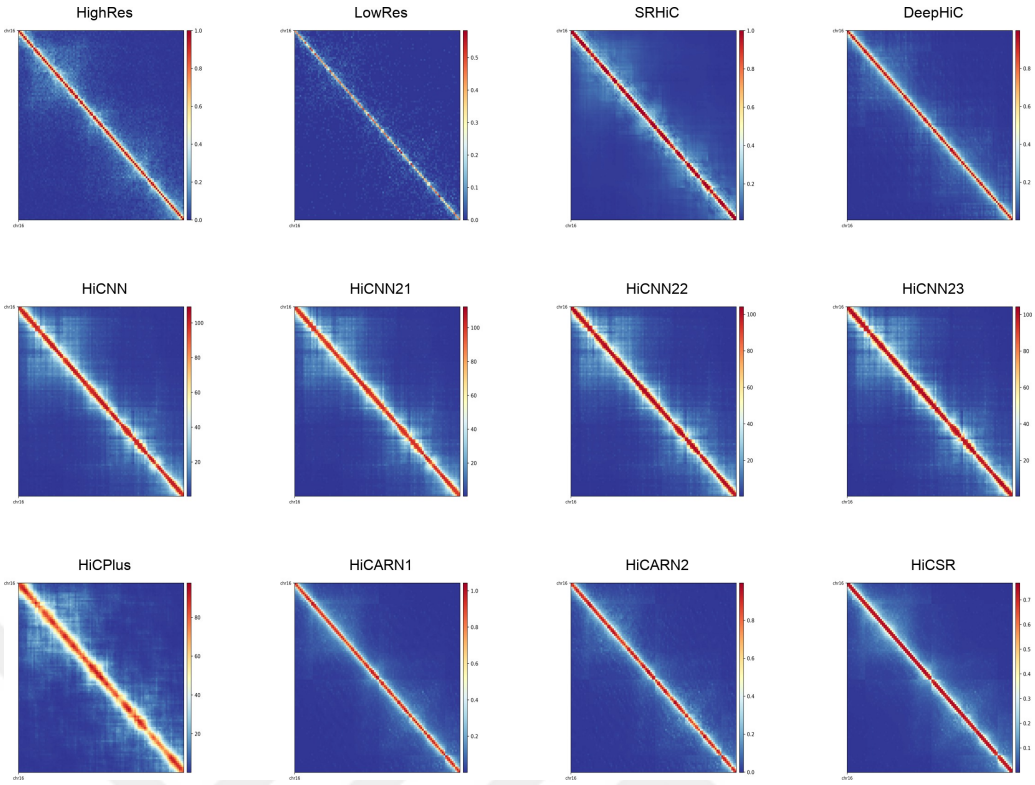


Figure A.4: Heatmap of a 1 Mb region of the NHEK cell line, corresponding to chromosome 16: 3,150,000-4,160,000. The heatmap shows the original high-resolution matrix, the down-sampled low-resolution matrix, and the ×16 enhanced matrices gen- erated by various models.

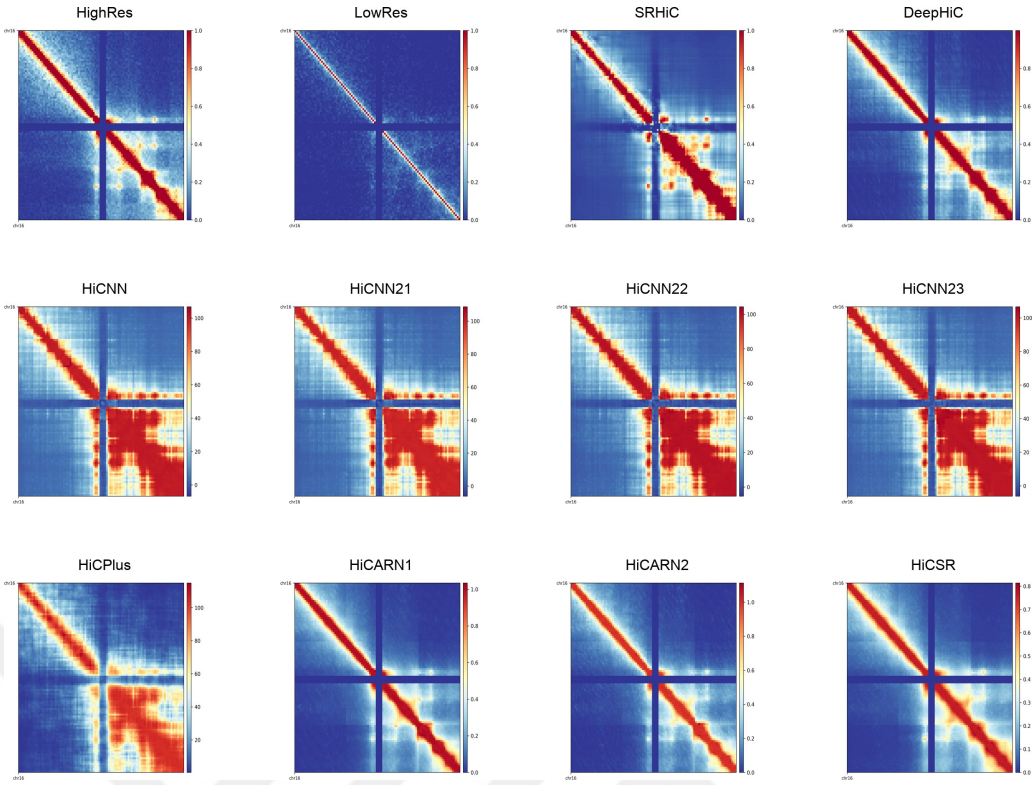


Figure A.5: Heatmap of a 1 Mb region of the GM12878 primary cell line, corresponding to chromosome 16: 8,150,000-9,160,000. The heatmap shows the original high- resolution matrix, the down-sampled low-resolution matrix, and the ×16 enhanced ma- trices generated by various models.

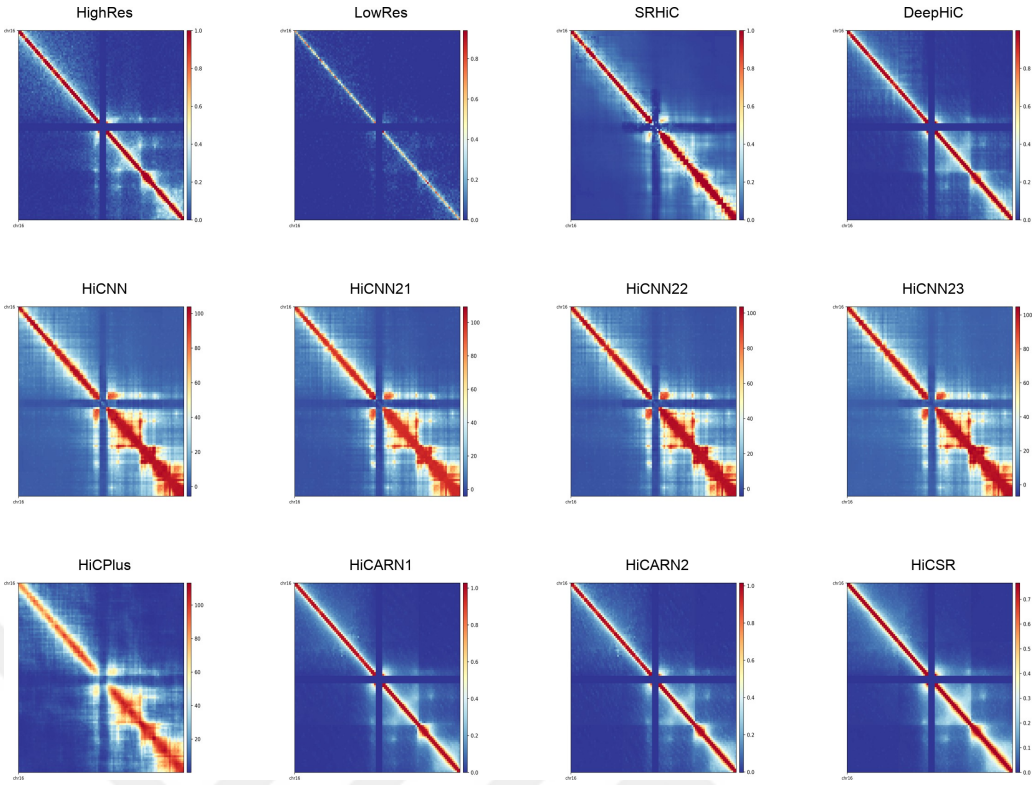


Figure A.6: Heatmap of a 1 Mb region of the IMR90 cell line, corresponding to chromosome 16: 8,150,000-9,160,000. The heatmap shows the original high-resolution matrix, the down-sampled low-resolution matrix, and the ×16 enhanced matrices gen- erated by various models.

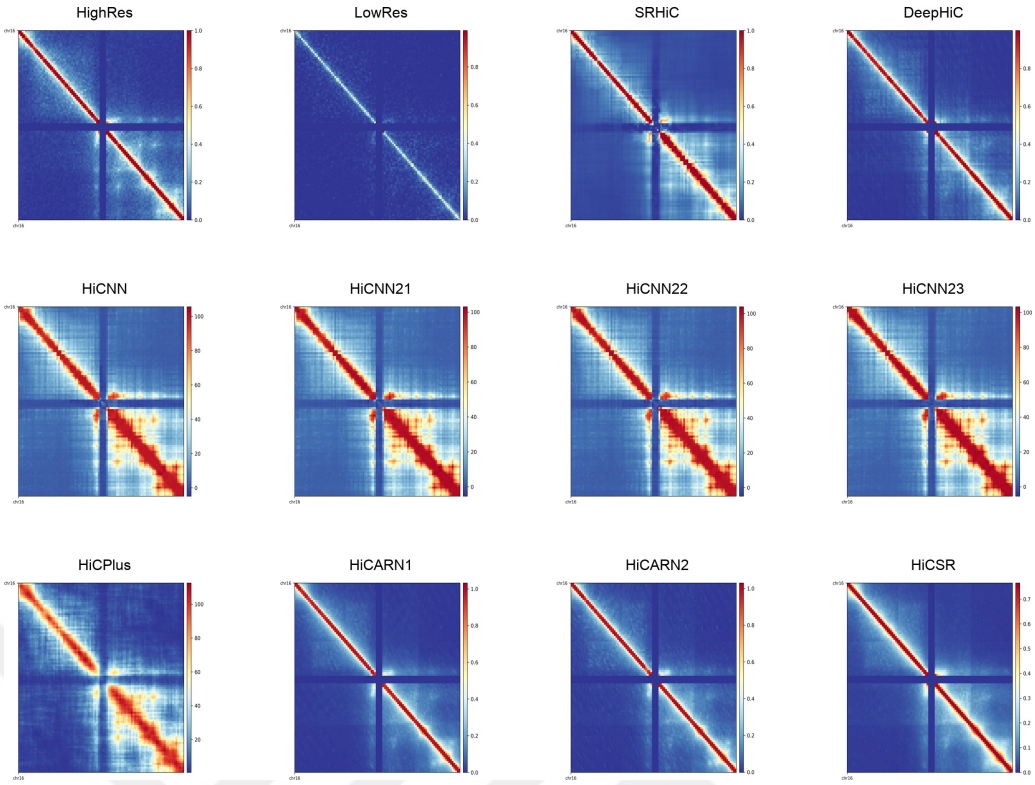


Figure A.7: Heatmap of a 1 Mb region of the K562 cell line, corresponding to chromosome 16: 8,150,000-9,160,000. The heatmap shows the original high-resolution matrix, the down-sampled low-resolution matrix, and the ×16 enhanced matrices gen- erated by various models.

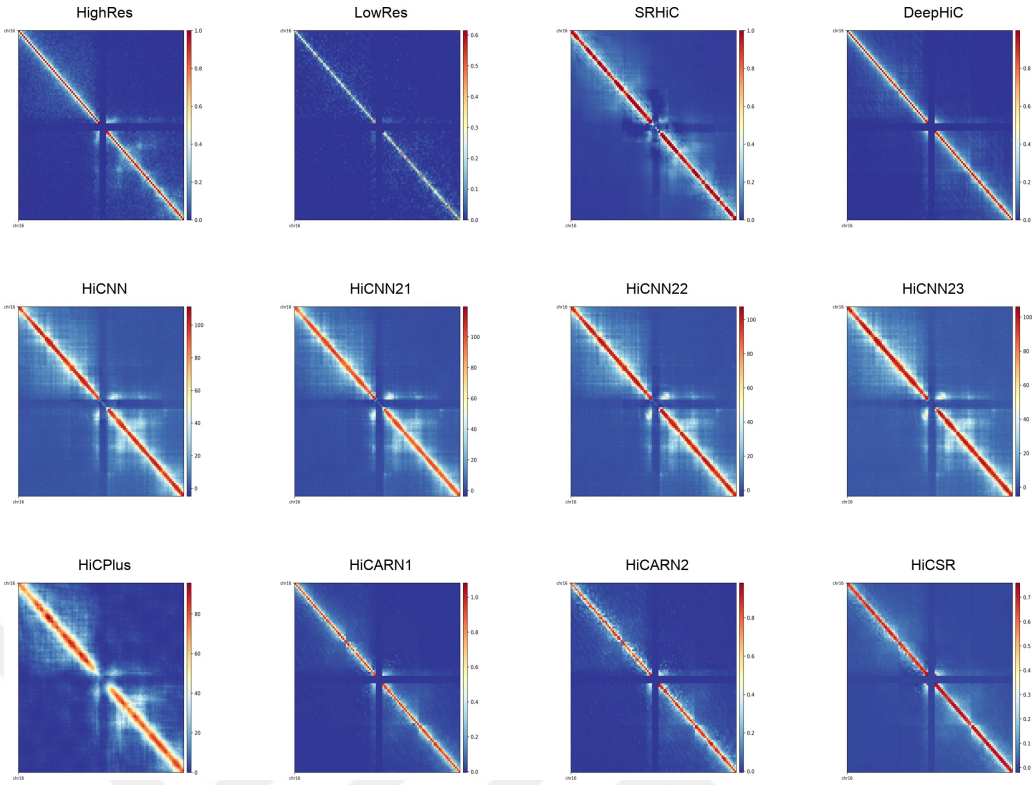


Figure A.8: Heatmap of a 1 Mb region of the NHEK cell line, corresponding to chromosome 16: 8,150,000-9,160,000. The heatmap shows the original high-resolution matrix, the down-sampled low-resolution matrix, and the ×16 enhanced matrices gen- erated by various models.

Table A.1: Pair-wise comparison significance values for Non-differential TADs in IMR90 cell line of x16 ratio.

	LowRes	HiCPlus	HiCNN	HiCNN21	HiCNN22	HiCNN23	DeepHiC	SRHiC	HiCARN1	HiCARN2	HiCSR
LowRes	–	–	–	–	–	–	–	–	–	–	–
HiCPlus	ns	–	–	–	–	–	–	–	–	–	–
HiCNN	ns	ns	–	–	–	–	–	–	–	–	–
HiCNN21	ns	ns	ns	–	–	–	–	–	–	–	–
HiCNN22	ns	ns	ns	ns	–	–	–	–	–	–	–
HiCNN23	ns	ns	ns	ns	ns	–	–	–	–	–	–
DeepHiC	4.285e-09	1.112e-08	2.819e-07	6.182e-08	9.183e-08	8.823e-09	–	–	–	–	–
SRHiC	1.561e-04	3.472e-02	2.549e-02	1.969e-02	2.243e-02	1.612e-02	6.806e-06	–	–	–	–
HiCARN1	7.881e-05	6.050e-04	1.688e-03	8.088e-04	1.289e-03	3.304e-04	6.817e-03	ns	–	–	–
HiCARN2	6.806e-06	4.964e-04	2.410e-04	2.977e-04	4.057e-04	2.977e-04	3.080e-03	ns	ns	–	–
HiCSR	4.016e-07	2.973e-05	4.490e-04	9.940e-05	1.248e-04	4.887e-05	1.506e-02	1.969e-02	ns	ns	–

Table A.2: Pair-wise comparison significance values for Non-differential TADs in IMR90 cell line of x25 ratio.

	LowRes	HiCPlus	HiCNN	HiCNN21	HiCNN22	HiCNN23	DeepHiC	SRHiC	HiCARN1	HiCARN2	HiCSR
LowRes	–	–	–	–	–	–	–	–	–	–	–
HiCPlus	7.941e-07	–	–	–	–	–	–	–	–	–	–
HiCNN	2.680e-04	ns	–	–	–	–	–	–	–	–	–
HiCNN21	4.490e-04	ns	ns	–	–	–	–	–	–	–	–
HiCNN22	9.505e-13	9.505e-13	9.505e-13	9.505e-13	–	–	–	–	–	–	–
HiCNN23	3.373e-05	ns	ns	ns	9.505e-13	–	–	–	–	–	–
DeepHiC	1.141e-11	1.518e-09	1.112e-08	2.819e-07	9.505e-13	3.368e-07	–	–	–	–	–
SRHiC	9.220e-11	1.035e-05	4.964e-04	2.604e-03	9.505e-13	4.625e-03	8.584e-04	–	–	–	–
HiCARN1	1.112e-08	ns	ns	ns	9.505e-13	ns	7.312e-07	2.604e-03	–	–	–
HiCARN2	2.819e-07	2.890e-02	ns	ns	9.505e-13	ns	2.580e-09	1.176e-03	ns	–	–
HiCSR	1.141e-11	5.109e-06	1.176e-03	9.181e-03	9.505e-13	4.625e-03	4.324e-05	ns	3.940e-03	1.140e-02	–

Table A.3: Pair-wise comparison significance values for Non-differential TADs in IMR90 cell line of x32 ratio.

	LowRes	HiCPlus	HiCNN	HiCNN21	HiCNN22	HiCNN23	DeepHiC	SRHiC	HiCARN1	HiCARN2	HiCSR
LowRes	–	–	–	–	–	–	–	–	–	–	–
HiCPlus	5.049e-08	–	–	–	–	–	–	–	–	–	–
HiCNN	3.811e-06	ns	–	–	–	–	–	–	–	–	–
HiCNN21	5.672e-07	ns	ns	–	–	–	–	–	–	–	–
HiCNN22	6.182e-08	ns	ns	ns	–	–	–	–	–	–	–
HiCNN23	3.336e-08	4.148e-02	ns	ns	ns	–	–	–	–	–	–
DeepHiC	1.901e-12	6.502e-10	5.476e-09	7.545e-08	1.349e-07	2.353e-07	–	–	–	–	–
SRHiC	1.901e-12	3.368e-07	3.004e-04	2.013e-03	4.270e-03	7.352e-03	1.742e-04	–	–	–	–
HiCARN1	1.061e-02	4.057e-04	3.346e-03	2.165e-04	8.855e-05	2.302e-05	1.806e-11	2.585e-10	–	–	–
HiCARN2	9.183e-08	ns	ns	ns	ns	3.074e-02	1.152e-09	5.672e-07	4.490e-04	–	–
HiCSR	1.959e-07	ns	ns	ns	ns	ns	1.518e-09	1.553e-05	1.176e-03	ns	–

Table A.4: Pair-wise comparison significance values for Non-differential TADs in IMR90 cell line of x50 ratio.

	LowRes	HiCPlus	HiCNN	HiCNN21	HiCNN22	HiCNN23	DeepHiC	SRHiC	HiCARN1	HiCARN2	HiCSR
LowRes	–	–	–	–	–	–	–	–	–	–	–
HiCPlus	2.604e-03	–	–	–	–	–	–	–	–	–	–
HiCNN	1.396e-08	3.940e-03	–	–	–	–	–	–	–	–	–
HiCNN21	6.182e-08	1.412e-03	ns	–	–	–	–	–	–	–	–
HiCNN22	1.518e-09	5.517e-05	ns	ns	–	–	–	–	–	–	–
HiCNN23	4.829e-10	3.821e-05	ns	ns	ns	–	–	–	–	–	–
DeepHiC	9.505e-13	2.852e-11	4.285e-09	3.368e-07	2.080e-06	7.941e-07	–	–	–	–	–
SRHiC	6.654e-12	2.580e-09	6.806e-06	1.248e-04	2.392e-03	1.176e-03	3.911e-02	–	–	–	–
HiCARN1	ns	1.544e-03	6.182e-08	2.353e-07	5.476e-09	6.969e-09	1.901e-12	1.141e-11	–	–	–
HiCARN2	ns	2.392e-03	1.114e-07	4.016e-07	1.112e-08	4.285e-09	1.901e-12	2.852e-11	ns	–	–
HiCSR	ns	2.165e-04	2.698e-08	1.746e-08	1.152e-09	3.545e-10	1.901e-12	3.802e-12	ns	ns	–

Table A.5: Pair-wise comparison significance values for Non-differential TADs in K562 cell line of x16 ratio.

	LowRes	HiCPlus	HiCNN	HiCNN21	HiCNN22	HiCNN23	DeepHiC	SRHiC	HiCARN1	HiCARN2	HiCSR
LowRes	–	–	–	–	–	–	–	–	–	–	–
HiCPlus	3.333e-09	–	–	–	–	–	–	–	–	–	–
HiCNN	1.073e-03	ns	–	–	–	–	–	–	–	–	–
HiCNN21	6.050e-04	ns	ns	–	–	–	–	–	–	–	–
HiCNN22	2.680e-04	ns	ns	ns	–	–	–	–	–	–	–
HiCNN23	3.663e-04	ns	ns	ns	ns	–	–	–	–	–	–
DeepHiC	9.505e-13	5.476e-09	3.336e-08	6.969e-09	4.111e-08	5.049e-08	–	–	–	–	–
SRHiC	8.697e-10	ns	ns	4.930e-02	2.715e-02	3.686e-02	7.941e-07	–	–	–	–
HiCARN1	2.819e-07	ns	ns	ns	ns	ns	5.672e-07	ns	–	–	–
HiCARN2	2.174e-08	ns	ns	ns	ns	ns	1.520e-06	ns	ns	–	–
HiCSR	2.174e-08	ns	ns	4.397e-02	3.074e-02	3.686e-02	8.855e-05	ns	ns	ns	–

Table A.6: Pair-wise comparison significance values for Non-differential TADs in K562 cell line of x25 ratio.

	LowRes	HiCPlus	HiCNN	HiCNN21	HiCNN22	HiCNN23	DeepHiC	SRHiC	HiCARN1	HiCARN2	HiCSR
LowRes	–	–	–	–	–	–	–	–	–	–	–
HiCPlus	3.333e-09	–	–	–	–	–	–	–	–	–	–
HiCNN	4.016e-07	ns	–	–	–	–	–	–	–	–	–
HiCNN21	4.016e-07	ns	ns	–	–	–	–	–	–	–	–
HiCNN22	9.505e-13	9.505e-13	9.505e-13	9.505e-13	–	–	–	–	–	–	–
HiCNN23	9.183e-08	ns	ns	ns	9.505e-13	–	–	–	–	–	–
DeepHiC	9.505e-13	1.901e-12	6.969e-09	6.502e-10	9.505e-13	8.697e-10	–	–	–	–	–
SRHiC	1.854e-10	1.114e-04	3.268e-02	9.181e-03	9.505e-13	3.074e-02	1.959e-07	–	–	–	–
HiCARN1	6.050e-04	3.304e-04	1.073e-03	7.348e-04	9.505e-13	4.964e-04	6.369e-11	2.353e-07	–	–	–
HiCARN2	6.806e-06	7.922e-03	6.317e-03	2.833e-03	9.505e-13	3.080e-03	9.505e-13	1.349e-07	ns	–	–
HiCSR	1.561e-04	1.612e-02	1.506e-02	2.103e-02	9.505e-13	7.352e-03	1.854e-10	6.806e-06	ns	ns	–

Table A.7: Pair-wise comparison significance values for Non-differential TADs in K562 cell line of x32 ratio.

	LowRes	HiCPlus	HiCNN	HiCNN21	HiCNN22	HiCNN23	DeepHiC	SRHiC	HiCARN1	HiCARN2	HiCSR
LowRes	–	–	–	–	–	–	–	–	–	–	–
HiCPlus	3.336e-08	–	–	–	–	–	–	–	–	–	–
HiCNN	1.349e-07	1.061e-02	–	–	–	–	–	–	–	–	–
HiCNN21	1.114e-07	2.243e-02	ns	–	–	–	–	–	–	–	–
HiCNN22	6.182e-08	1.969e-02	ns	ns	–	–	–	–	–	–	–
HiCNN23	4.111e-08	3.686e-02	ns	ns	ns	–	–	–	–	–	–
DeepHiC	9.505e-13	1.806e-11	6.969e-09	1.112e-08	1.112e-08	1.112e-08	–	–	–	–	–
SRHiC	4.111e-08	1.176e-03	ns	ns	ns	ns	3.368e-07	–	–	–	–
HiCARN1	ns	2.080e-06	4.778e-07	6.718e-07	3.368e-07	3.368e-07	9.505e-13	1.959e-07	–	–	–
HiCARN2	ns	3.373e-05	1.520e-06	2.825e-06	1.103e-06	7.941e-07	9.505e-13	4.778e-07	ns	–	–
HiCSR	ns	4.490e-04	4.887e-05	3.373e-05	3.821e-05	2.973e-05	2.852e-11	6.806e-06	ns	ns	–

Table A.8: Pair-wise comparison significance values for Non-differential TADs in K562 cell line of x50 ratio.

	LowRes	HiCPlus	HiCNN	HiCNN21	HiCNN22	HiCNN23	DeepHiC	SRHiC	HiCARN1	HiCARN2	HiCSR
LowRes	–	–	–	–	–	–	–	–	–	–	–
HiCPlus	7.837e-06	–	–	–	–	–	–	–	–	–	–
HiCNN	7.837e-06	ns	–	–	–	–	–	–	–	–	–
HiCNN21	1.103e-06	ns	ns	–	–	–	–	–	–	–	–
HiCNN22	4.778e-07	6.817e-03	ns	ns	–	–	–	–	–	–	–
HiCNN23	2.174e-08	2.392e-03	3.472e-02	ns	ns	–	–	–	–	–	–
DeepHiC	4.829e-10	8.823e-09	2.819e-07	4.111e-08	5.901e-06	1.780e-06	–	–	–	–	–
SRHiC	1.780e-06	ns	ns	ns	ns	2.243e-02	2.819e-07	–	–	–	–
HiCARN1	ns	4.490e-04	5.483e-04	1.553e-05	2.426e-06	4.778e-07	1.152e-09	1.742e-04	–	–	–
HiCARN2	ns	6.182e-08	2.825e-06	4.111e-08	2.698e-08	1.152e-09	3.545e-10	1.349e-07	ns	–	–
HiCSR	ns	1.959e-07	4.416e-06	4.016e-07	3.368e-07	2.174e-08	3.545e-10	4.016e-07	3.074e-02	ns	–

Table A.9: Pair-wise comparison significance values for Non-differential TADs in NHEK cell line of x16 ratio.

	LowRes	HiCPlus	HiCNN	HiCNN21	HiCNN22	HiCNN23	DeepHiC	SRHiC	HiCARN1	HiCARN2	HiCSR
LowRes	–	–	–	–	–	–	–	–	–	–	–
HiCPlus	9.505e-13	–	–	–	–	–	–	–	–	–	–
HiCNN	9.505e-13	1.780e-06	–	–	–	–	–	–	–	–	–
HiCNN21	9.505e-13	1.349e-07	ns	–	–	–	–	–	–	–	–
HiCNN22	9.505e-13	9.940e-05	ns	ns	–	–	–	–	–	–	–
HiCNN23	9.505e-13	1.553e-05	ns	ns	ns	–	–	–	–	–	–
DeepHiC	9.505e-13	6.050e-04	ns	ns	ns	ns	–	–	–	–	–
SRHiC	9.505e-13	7.922e-03	ns	ns	ns	ns	ns	–	–	–	–
HiCARN1	7.348e-04	6.502e-10	9.505e-13	9.505e-13	1.141e-11	1.901e-12	3.802e-12	1.984e-09	–	–	–
HiCARN2	4.285e-09	2.165e-04	6.654e-12	9.505e-13	6.969e-09	8.697e-10	8.823e-09	1.520e-06	5.850e-03	–	–
HiCSR	2.585e-10	ns	6.671e-04	4.887e-05	2.392e-03	9.771e-04	8.532e-03	ns	6.718e-07	1.015e-03	–

Table A.10: Pair-wise comparison significance values for Non-differential TADs in NHEK cell line of x25 ratio.

	LowRes	HiCPlus	HiCNN	HiCNN21	HiCNN22	HiCNN23	DeepHiC	SRHiC	HiCARN1	HiCARN2	HiCSR
LowRes	–	–	–	–	–	–	–	–	–	–	–
HiCPlus	8.823e-09	–	–	–	–	–	–	–	–	–	–
HiCNN	3.802e-12	5.049e-08	–	–	–	–	–	–	–	–	–
HiCNN21	1.854e-10	5.049e-08	ns	–	–	–	–	–	–	–	–
HiCNN22	9.505e-13	9.505e-13	9.505e-13	9.505e-13	–	–	–	–	–	–	–
HiCNN23	1.141e-11	2.426e-06	ns	ns	9.505e-13	–	–	–	–	–	–
DeepHiC	6.654e-12	3.304e-04	ns	3.911e-02	9.505e-13	2.115e-01	–	–	–	–	–
SRHiC	9.505e-13	2.618e-05	ns	ns	9.505e-13	ns	2.115e-01	–	–	–	–
HiCARN1	ns	1.959e-07	6.654e-12	1.854e-10	9.505e-13	4.277e-11	1.321e-10	6.654e-12	–	–	–
HiCARN2	ns	2.353e-07	2.852e-11	1.854e-10	9.505e-13	4.277e-11	9.220e-11	9.505e-13	7.537e-01	–	–
HiCSR	ns	2.080e-06	1.321e-10	2.585e-10	9.505e-13	6.502e-10	1.518e-09	1.854e-10	4.355e-01	3.335e-01	–

Table A.11: Pair-wise comparison significance values for Non-differential TADs in NHEK cell line of x32 ratio.

	LowRes	HiCPlus	HiCNN	HiCNN21	HiCNN22	HiCNN23	DeepHiC	SRHiC	HiCARN1	HiCARN2	HiCSR
LowRes	–	–	–	–	–	–	–	–	–	–	–
HiCPlus	1.553e-05	–	–	–	–	–	–	–	–	–	–
HiCNN	4.285e-09	2.021e-05	–	–	–	–	–	–	–	–	–
HiCNN21	1.806e-11	9.368e-07	ns	–	–	–	–	–	–	–	–
HiCNN22	6.654e-12	1.520e-06	ns	ns	–	–	–	–	–	–	–
HiCNN23	9.220e-11	4.416e-06	ns	ns	ns	–	–	–	–	–	–
DeepHiC	1.518e-09	9.940e-05	ns	ns	ns	ns	–	–	–	–	–
SRHiC	2.852e-11	2.353e-07	ns	ns	ns	ns	ns	–	–	–	–
HiCARN1	ns	5.483e-04	9.183e-08	4.829e-10	6.502e-10	4.285e-09	5.049e-08	2.585e-10	–	–	–
HiCARN2	ns	2.973e-05	4.285e-09	9.220e-11	6.654e-12	1.854e-10	1.518e-09	4.277e-11	4.635e-01	–	–
HiCSR	7.545e-08	6.654e-12	6.654e-12	9.505e-13	9.505e-13	9.505e-13	9.505e-13	9.505e-13	1.112e-08	4.111e-08	–

Table A.12: Pair-wise comparison significance values for Non-differential TADs in NHEK cell line of x50 ratio.

	LowRes	HiCPlus	HiCNN	HiCNN21	HiCNN22	HiCNN23	DeepHiC	SRHiC	HiCARN1	HiCARN2	HiCSR
LowRes	–	–	–	–	–	–	–	–	–	–	–
HiCPlus	2.117e-07	–	–	–	–	–	–	–	–	–	–
HiCNN	2.829e-09	2.006e-04	–	–	–	–	–	–	–	–	–
HiCNN21	5.804e-11	3.364e-05	ns	–	–	–	–	–	–	–	–
HiCNN22	2.902e-11	9.648e-06	ns	ns	–	–	–	–	–	–	–
HiCNN23	1.451e-11	7.371e-09	ns	ns	ns	–	–	–	–	–	–
DeepHiC	1.061e-07	1.192e-04	ns	ns	ns	ns	–	–	–	–	–
SRHiC	1.451e-11	1.328e-08	ns	ns	ns	ns	ns	–	–	–	–
HiCARN1	ns	6.004e-05	7.371e-09	2.017e-09	1.741e-10	1.451e-11	6.178e-07	2.902e-11	–	–	–
HiCARN2	ns	6.532e-08	3.176e-07	2.902e-11	1.451e-11	1.451e-11	8.341e-08	1.451e-11	1.918e-01	–	–
HiCSR	6.529e-10	1.451e-11	1.451e-11	1.451e-11	1.451e-11	1.451e-11	1.741e-10	1.451e-11	3.946e-09	2.017e-09	–