Supplementary Information

Predicting Higher-Order Chromatin Interactions with PHOCI

Yunyi Wu[1], Xing Jiang[2], Zhi Yang[1,3], Yanqing Wang[1,3], Lipeng Li[1], Jinsheng Xu[4], Pan Deng[5], Chunhui Hou[4,*], Chen Yu[2,*], Kai Huang[1,3,6,*]

[1] Institute of Systems and Physical Biology, Shenzhen Bay Laboratory, Shenzhen 518132, China

[2] Institute of Cancer Research, Shenzhen Bay Laboratory, Shenzhen 518132, China

[3] Shenzhen Medical Academy of Research and Translation (SMART), Shenzhen 518107, China

[4] China State Key Laboratory of Genetic Resources and Evolution, Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming 650223, China

[5] Zhongguancun Academy, Beijing 100094, China

[6] Lead contact

* Correspondence: houchunhui@mail.kiz.ac.cn, yu@szbl.ac.cn, huangkai@szbl.ac.cn

The PDF file includes:

Supplementary Text

Supplementary Tables. 1 to 6

Supplementary Figures. 1 to 11

Model Configuration

During model training, we employed the Adam optimizer and simultaneously trained the GraphSAGE Encoder, ESA, and MLP. The learning rate was set to 0.0001, considering the extensive nature of multi-way interaction data, while we adjusted the batch size to 512 or 1024. To ensure stable training, we implemented gradient clipping with a threshold set at 0.01 and utilized the binary cross-entropy loss function (BCELoss).

Supplementary Table 1 Configuration of the deep learning model

Module	Layer	Layer number	Layer dimension configuration
GraphSAGE Encoder	Linear+Normalize	1	Input: 15 Output: 400
	GraphSAGE layers	3	Input:400 Hidden output 1: 400 Output: 400
ESA	Sampler	1	Input: n Output: 400*n
	Min – Max aggregator	1	Input: 400*n Output: 400
	Average aggregator	1	Input: 400*n Output: 400
	All aggregator (concat)	1	Input: 400+400+2 Output: 802
MLP	Linear	7	Input: 802 Hidden output 1: 400 Hidden output 2: 256 Hidden output 3: 128 Hidden output 4: 32 Hidden output 5: 8 Output: 1

Data Sources, Collection, and Preprocessing

In this section, we provide a detailed account of our methodology about data, including the source, acquisition and pre-processing of the data. In addition, we explain the calculation of chromatin state normalized enrichment. The datasets used in our study include HiC data showing chromatin pairwise interactions, ChIP-seq data associated with epigenetics, ATAC-seq data and high-throughput Pore-C data reflecting multi-way interactions. Specifically, we used data from ChromHMM-18 states, gene expression data, and A/B compartment data for results analysis.

HiC data sources and preprocessing

Specifically, the HiC data were obtained from the 4D Nucleome project[1] and ENCODE project[2–4], involving experimental datasets generated through Hi-C technology. We utilized experimental datasets from the 4DNESQWI9K2F dataset (GM12878), the 4DNES54GS5KI dataset (K562), the ENCFF689CUX dataset (A549), the 4DNFICSTCJQZ (HepG2) dataset, and the 4DNFID162B9J dataset (H1-hESC). To ensure the applicability of the model to other cell lines and

Hi-C data, we chose a resolution of 5kb per bin as the highest resolution. Within the experimental datasets, we processed mcool and hic file formats. The mcool files were processed using the Python package cooler[5] to read the mcool files, and the KR balance method was applied to globally smooth the Hi-C maps. Similarly, hic files were processed using the Python package hicstraw[6] to read the hic files, and the KR balance method was used to globally smooth the Hi-C maps. Finally, we segmented the entire Hi-C data based on chromosomes to generate cis-topological graphs for each chromosome. In these maps, each bin represented a node, while interactions between two bins were represented as edges. This processing method aimed to capture the spatial structure and interaction relationships of chromatin, providing the basis for subsequent analysis and modelling.

ChIP-seq and ATAC-seq data sources and preprocessing

Regarding the ChIP-seq data associated with epigenetics, we used bigWig format data from the ENCODE project[2–4] and 4D Nucleome project[1]. A total of 15 epigenetic features were employed, including H3K4me3, H3K27ac, H3K27me3, H3K4me1, H3K36me3, H3K9me3, H3K9ac, H3K4me2, H4K20me1, H2AFZ, H3K79me2, CTCF, POLR2A, RAD21, and ATAC (Supplementary Table 1). We used datasets at a resolution of 5kb. Using the Python package pyBigWig[7,8], we read the bigWig format files and obtained the mean peak signal value for each bin. Finally, we obtained the values of these 15 epigenetic features for each bin along the chromosome and considered them as features for each node in the cis-topological graph.

Supplementary Table 2 Source of epigenetic features from ChIP-seq and ATAC-seq

Epigenetic Features	GM12878	K562	A549	HepG2	H1-hESC
H3K4me3	ENCFF003DXG	ENCFF525ZRM	ENCFF242FAU	ENCFF500VAH	ENCFF493QWY
H3K27ac	ENCFF340JIF	ENCFF381NDD	ENCFF070DKP	ENCFF022TZG	ENCFF314KQD
H3K27me3	ENCFF039JOT	ENCFF928NWQ	ENCFF702IOJ	ENCFF437XHN	ENCFF345VHG
H3K4me1	ENCFF564KBE	ENCFF761XBZ	ENCFF160YWB	ENCFF576YVM	ENCFF088MXE
H3K36me3	ENCFF380LZI	ENCFF440XMD	ENCFF473XIC	ENCFF488DNL	ENCFF488THD
H3K9me3	ENCFF683HCZ	ENCFF812HRW	ENCFF142SPT	ENCFF754ROM	ENCFF183MHJ
H3K9ac	ENCFF599TRR	ENCFF937QUK	ENCFF808VAQ	ENCFF053ROV	ENCFF084JKQ
H3K4me2	ENCFF627OKN	ENCFF959YJV	ENCFF479HXK	ENCFF057BKO	ENCFF860NVB
H4K20me1	ENCFF479XIQ	ENCFF605FAF	ENCFF417UUX	ENCFF330AIV	ENCFF156JZY
H2AFZ	ENCFF601YET	ENCFF494WCA	ENCFF177CPK	ENCFF253PND	ENCFF296IBP
H3K79me2	ENCFF931USZ	ENCFF544AVW	ENCFF375NRQ	ENCFF655XBP	ENCFF401PZS
CTCF	ENCFF485CGE	ENCFF675GVW	ENCFF109XKO	ENCFF301SGJ	ENCFF648BTZ
POLR2A	ENCFF200WHZ	ENCFF124WLE	ENCFF774RVE	ENCFF761IJZ	ENCFF933YTR
RAD21	ENCFF571ZJJ	ENCFF652NKM	ENCFF498DXU	ENCFF242MRW	ENCFF002NBT
ATAC	ENCFF603BJO	ENCFF754EAC	ENCFF872SDF	ENCFF664EJT	4DNFICPNO4M5

High-throughput Pore-C data sources and preprocessing

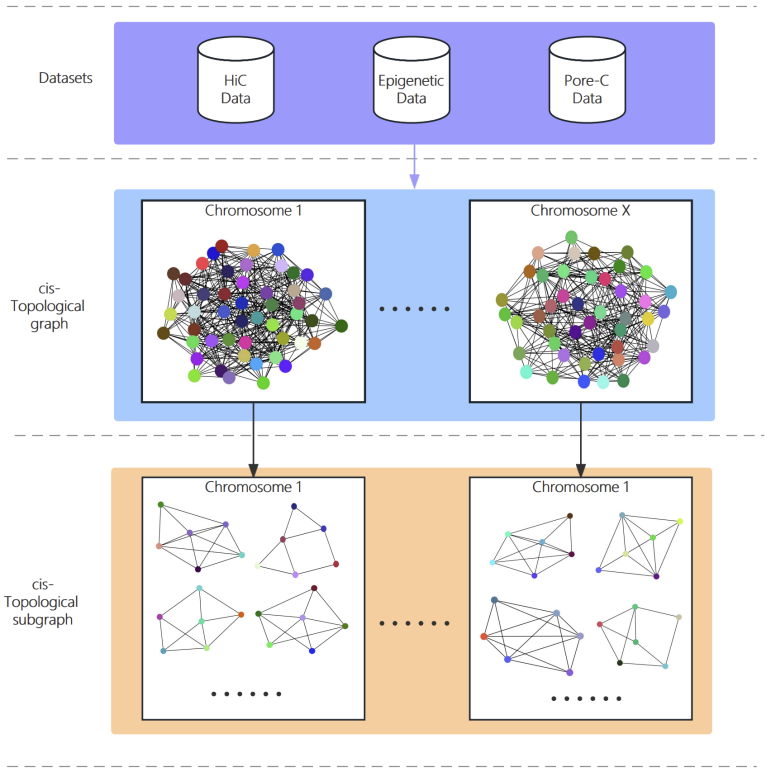
For the high-throughput Pore-C data concerning multi-way interactions, we utilized the dataset on the GM12878 and K562 cell lines produced by Zhong et al.[9] This dataset is stored within the GSE202539 dataset in the NCBI GEO database[10,11] After obtaining the Pore-C read alignment data, we performed a read fragment filtering process to ensure accurate mapping and to minimize

systematic noise due to mapping errors. We retained read fragments with a mapping quality (MapQual) ≥ 10, and determined the chromosomal bin of each fragment based on its midpoint. This approach allowed us to view the multi-way interactions between different bins in the cis-topological graph through each read.

From cis-topological graphs to cis-topological subgraphs

By amalgamating these diverse data sources, we successfully constructed cis-topological graphs of each chromosomes (Supplementary Figure 1). In this representation, each node denotes a 5kb genomic segment (bin) characterized by a vector comprising 15 epigenetic features. The edges within the cis-topological graph indicate pairwise interactions from HiC experiment, and we additionally recorded multi-way interactions between nodes. Due to computer memory limitations, we segmented the topological graph of the entire chromosome into contiguous segments, forming cis-topological subgraphs, each comprising 1000 adjacent bins on genomic sequence.

For the GM12878 cell line, we gained a total of 566 cis-topological subgraphs encompassing 566,000 nodes, 118,090,406 edges, and 71,077,284 multi-way interactions. On average, each cis-topological subgraph contained 1000 nodes, 208,640 edges, and 125,578 multi-way interactions. Similarly, for the K562 cell line, we produced 563 cis-topological subgraphs, with 563,000 nodes, 77,828,060 edges, and 54,309,928 multi-way interactions. The average statistic for each cis-topological subgraph included 1000 nodes, 138,238 edges and 96,465 multi-way interactions.



Supplementary Figure 1 Preprocessing flow for data to construct cis-topological graphs to cis-topological subgraphs (Conceptual Example)

Splitting of training, validation, testing sets

Subsequently, we partitioned the dataset into training, validation and testing sets. To prevent

data leakage within the same chromosome, our dataset partitioning was based on chromosomal integrity rather than random assignment. Additionally, we employed data from another cell line as a inter-cell-line test set. Combining data of these two cell lines, we trained a comprehensive model. The specific partitioning methodologies are shown in Supplementary Table 2. After partitioning the training, validation and testing sets, we utilized this data for subsequent training of deep learning models and random walk sampling.

Supplementary Table 3 Dataset partitioning of training, validation and testing sets

	GM12878 Trained Model	K562 Trained Model	Comprehensive Model (GM12878 + K562)
Training	chr1, chr2, chr3, chr4, chr5, chr6, chr7, chr8, chr10, chr11, chr12, chr16, chr17, chr18, chr20, chr21	chr1, chr2, chr3, chr4, chr5, chr6, chr7, chr8, chr10, chr11, chr12, chr16, chr17, chr18, chr20, chr21	chr1, chr2, chr3, chr4, chr5, chr6, chr7, chr8, chr10, chr11, chr12, chr16, chr17, chr18, chr20, chr21
	Total cis-subgraphs: 469 Total nodes: 469000 Total edges: 98521602 Total multi-way interactions: 60013619	Total cis-subgraphs: 466 Total nodes: 466000 Total edges: 67213404 Total multi-way interactions: 46997811	Total cis-subgraphs: 935 Total nodes: 935000 Total edges: 165735006 Total multi-way interactions: 107011430
Validation	chr13, chr15	chr13, chr15	chr13, chr15
	Total cis-subgraphs: 37 Total nodes: 37000 Total edges: 7565970 Total multi-way interactions: 4647794	Total cis-subgraphs: 37 Total nodes: 37000 Total edges: 4417140 Total multi-way interactions: 3064658	Total cis-subgraphs: 74 Total nodes: 74000 Total edges: 11983110 Total multi-way interactions: 7712452
Testing (Intra-cell line)	chr14, chr22, chrX	chr14, chr22, chrX	chr14, chr22, chrX
	Total cis-subgraphs: 60 Total nodes: 60000 Total edges: 12002834 Total multi-way interactions: 6415871	Total cis-subgraphs: 60 Total nodes: 60000 Total edges: 6197516 Total multi-way interactions: 4247459	Total cis-subgraphs: 120 Total nodes: 120000 Total edges: 18200350 Total multi-way interactions: 10663330
Testing (Inter-cell line)	K562 cell line data	GM12878 cell line data	\

Data sources and preprocessing of ChromHMM-18 states labels

ChromHMM[12,13] is a series of models based on a multivariate hidden Markov model, enabling effective analysis of complex patterns of chromatin marks and providing powerful capabilities for analyzing chromatin states in the genome. We utilized data from ChromHMM-18 states model from the ENCODE project[2–4], which includes 18 chromatin states. The datasets we used include ENCFF671FDK (GM12878), ENCFF319VXX (K562), ENCFF969EVA (A549), ENCFF123MSN (HepG2), and ENCFF789ZOX (H1-hESC). After converting the files from bed format to bigBed format, we used the Python package pyBigWig[7,8] for file reading. We divided the genome sequence

into bins of 5kb resolution, summarized the types of chromatin states contained in each bin, and removed duplicates. After that, we obtained labeled data for the chromatin state types in each bin.

Data sources and preprocessing of A/B compartment labels

Megabase[14,15] is a neural network model based on maximum entropy, designed to analyze local epigenomic data (e.g., ChIP-Seq data for histone modifications) to predict the A/B compartment classification of chromosome structure, including compartments and subcompartments. We utilized Megabase to generate A/B compartment labels at a 50kb resolution for GM12878 and K562 cell lines. Due to the model's minimum resolution being 50kb, we subsequently divided the 50kb bins into 10 bins of 5kb each to obtain A/B compartment labels for each 5kb bin.

Data sources and preprocessing of gene expression labels

In the cell lines of GM12878, K562, A549, and HepG2, we employed the standard gene expression levels relative to other cell lines and tissues from the Harmonizome[16] database to annotate the relative gene expression levels. However, due to the absence of corresponding gene expression data for H1-hESC cell lines in Harmonizome, we resorted to RNA-seq data (ENCFF816ERP) from the ENCODE project [2–4] and extracted their TPM values for ranking.

Chromatin states normalized enrichment of multi-way interaction group samples

To determine the proportions of chromatin states within multi-way interactions in different prediction probability ranges, we first counted the occurrences of each chromatin state within a given set of multi-way interactions (State count). Subsequently, we divided this count by the total number of points in the set (Total count) to derive the proportions of different chromatin states (State genome ratio). To further standardize our analysis, we divided these proportions by the corresponding occurrence rates of the chromatin states in the genome, resulting in the calculation of normalized enrichment. The same calculations were performed for each set of probability ranges.

$$Normalized\ Enrichment = \frac{State\ count}{Total\ count \times State\ genome\ ratio}$$

Visualization of Chromatin Multi-way Interactions

We used PAOHvis[17], a tool based on the Parallel Aggregated Ordered Hypergraph (PAOH) technique, to visualize chromatin multi-way interactions in Figure 6 (b) and (j). In PAOHvis, genomic loci are represented as parallel horizontal bars, and multi-way interactions are shown as vertical lines (hyperedges) connecting them.

Experimental Materials

Supplementary Table 4 sgRNA targets for *MYB* gene perturbation

Experiment	sgRNA Id	sgRNA Target	sgRNA Sequence	Related Bin
CRISPRi	sgA-1	A	TTGTGTTTTAAGTAGAGACG	9
CRISPRi	sgA-2	A	TGTAATCCCAGCACTTTGGG	9
CRISPRi	sgA-3	A	CATCAGCCTCCCAAAGTGCT	9
CRISPRi	sgB-1	B	GTGAAGGAAACAGAGACATG	17
CRISPRi	sgB-2	B	CATACCTTCCAAATACAGAT	17
CRISPRi	sgB-3	B	AGAAGAAGAGAGGAAGGAGA	17
CRISPRi	sgC-1	C	CAGGTAACCAGATAGGACTG	52
CRISPRi	sgC-2	C	AATTTAGCCCTCATTTATGG	52
CRISPRi	sgC-3	C	TTCAGCTGCTCTTCTGCAAG	52
CRISPRi	sgTSS-1	TSS	CCTGAGAAACTTCGCCCCAG	24
CRISPRi	sgTSS-2	TSS	CGCCATGGCCCGAAGACCC	24
CRISPRi	sgTSS-3	TSS	AAACTTCGCCCCAGCGGTG	24
CRISPRi	sg-NTC	NTC	GAACGACTAGTTAGGCGTGTA	NTC

Supplementary Table 5 qPCR primer sequences for target gene analysis

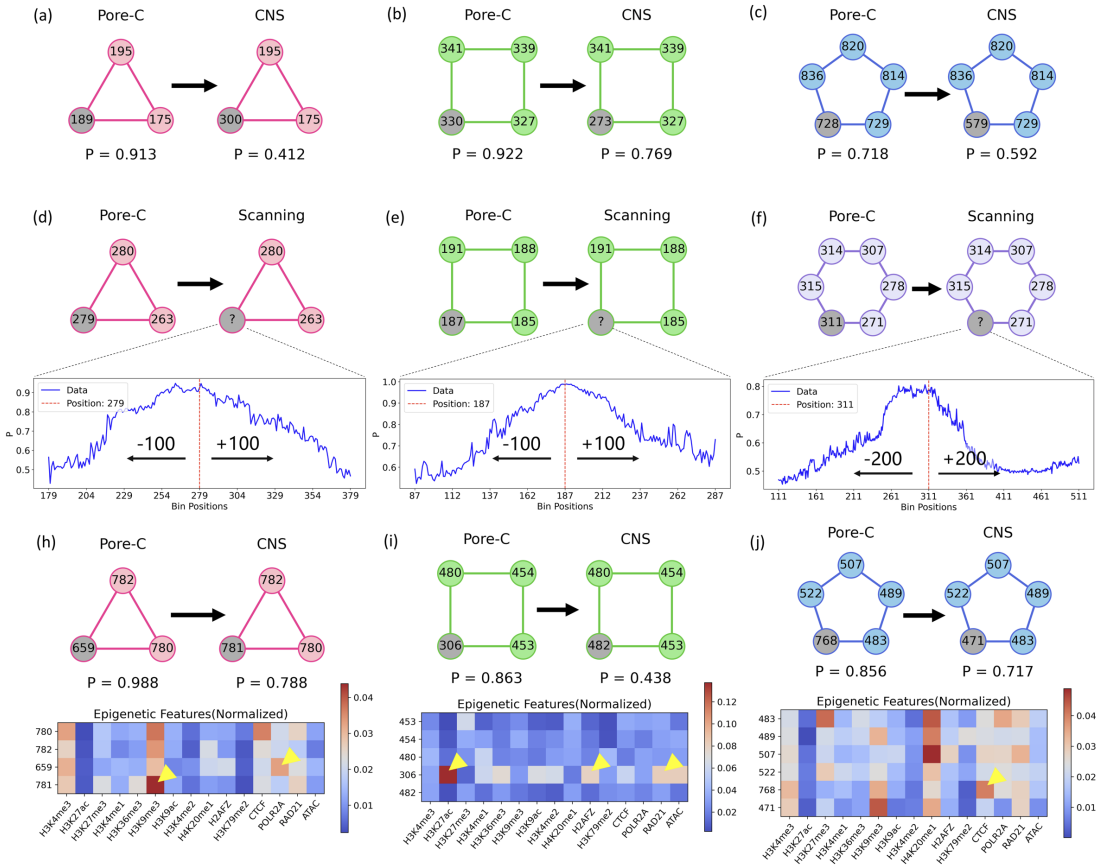
Gene Symbol	Primer Name	Sequence (5'→3')	Amplicon Size (bp)
GAPDH (reference gene)	GAPDH-F	GTCTCCTCTGACTTCAACAGCG	131
GAPDH (reference gene)	GAPDH-R	ACCACCCTGTTGCTGTAGCCAA	
MYB	MYB-F	AAGGTCGAACAGGAAGGTTATC	81
MYB	MYB-R	ACTGTTCTTCTGGAAGCTTGT	

Supplementary Figures and Tables of Results

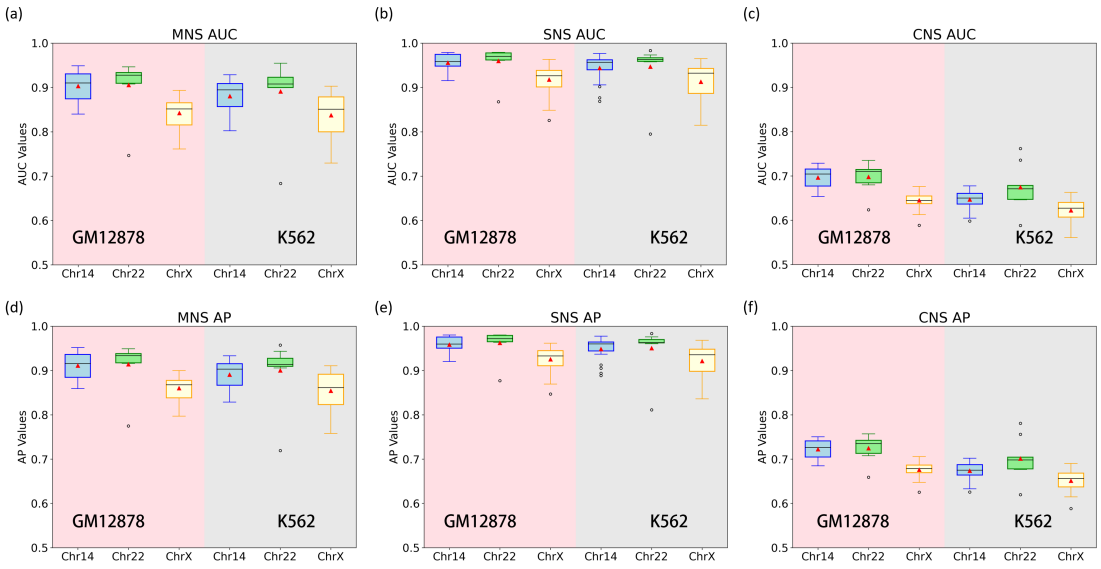
Supplementary Table 6 Performance of the Models

	GM12878 Trained Model							
	AUC				AP			
	MNS	SNS	CNS	Average	MNS	SNS	CNS	Average
Valid	0.897	0.951	0.687	0.845	0.905	0.953	0.713	0.857
Test (Intra-cell line)	0.871	0.936	0.669	0.825	0.884	0.941	0.698	0.841
Test (Inter-cell line)	0.867	0.926	0.653	0.815	0.879	0.931	0.680	0.830
	K562 Trained Model							
	AUC				AP			
	MNS	SNS	CNS	Average	MNS	SNS	CNS	Average

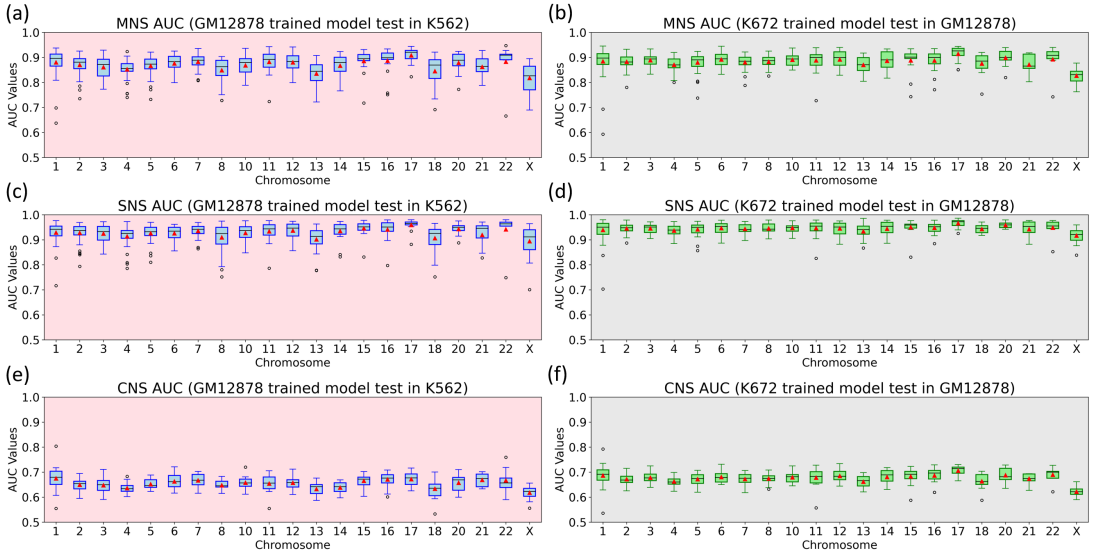
Valid	0.878	0.939	0.650	0.822	0.888	0.944	0.677	0.836
Test (Intra-cell line)	0.859	0.928	0.638	0.809	0.873	0.934	0.666	0.824
Test (Inter-cell line)	0.882	0.943	0.674	0.833	0.891	0.948	0.698	0.845
	Comprehensive Model							
	AUC				AP			
	MNS	SNS	CNS	Average	MNS	SNS	CNS	Average
Valid	0.877	0.936	0.671	0.828	0.887	0.940	0.697	0.841
Test	0.865	0.930	0.657	0.817	0.878	0.935	0.684	0.832



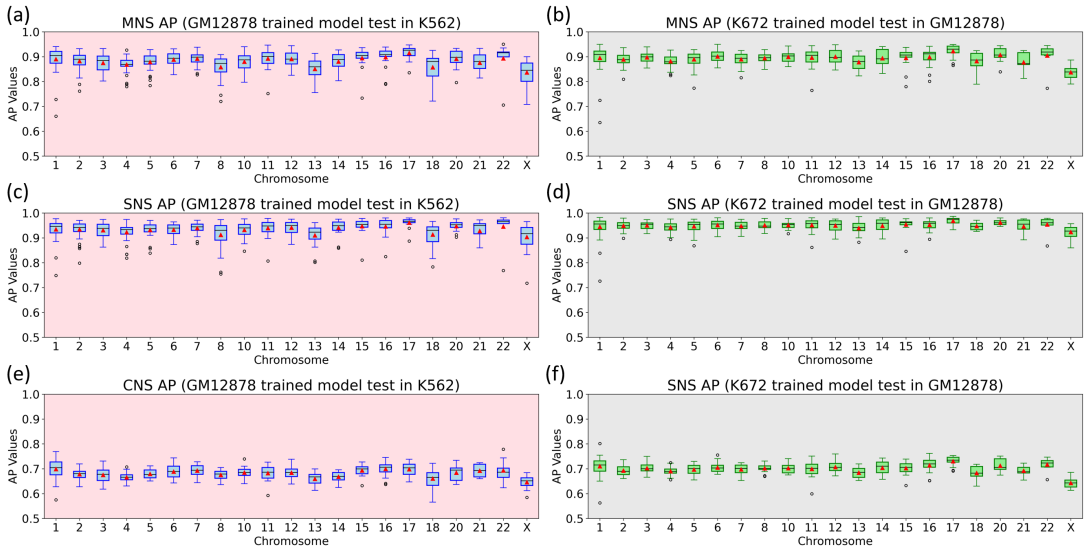
Supplementary Figure 2 Schematic illustration depicting the effect of distance range and feature distribution on model prediction (Test set Intra-GM12878 Chr14:76Mb-81Mb). (a-c) CNS transformation to father nodes. (d-f) Scanning analysis along upstream and downstream. (h-j) Analysis of epigenetic features as a contributing factor.



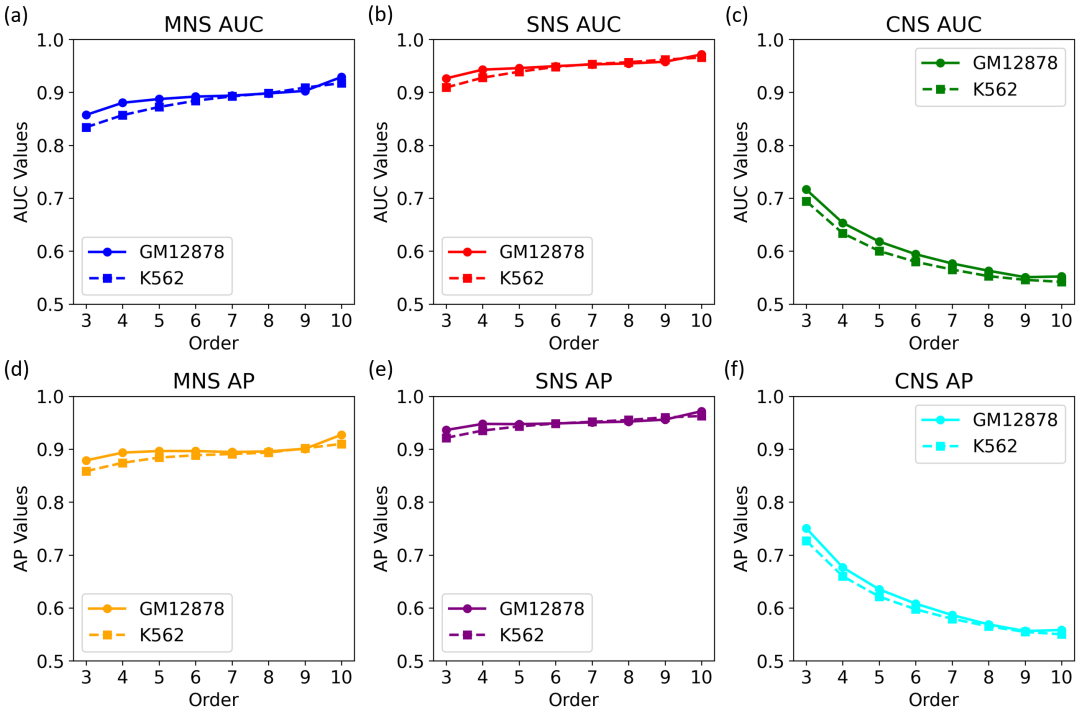
Supplementary Figure 3 Analysis of AUC and AP in intra-cell line testing. (a-c) Box plots of AUC scores for chromosomes in test sets (Chr14, Chr22, and ChrX). (d-f) Box plots of AP scores for chromosomes in test sets (Chr14, Chr22, and ChrX).



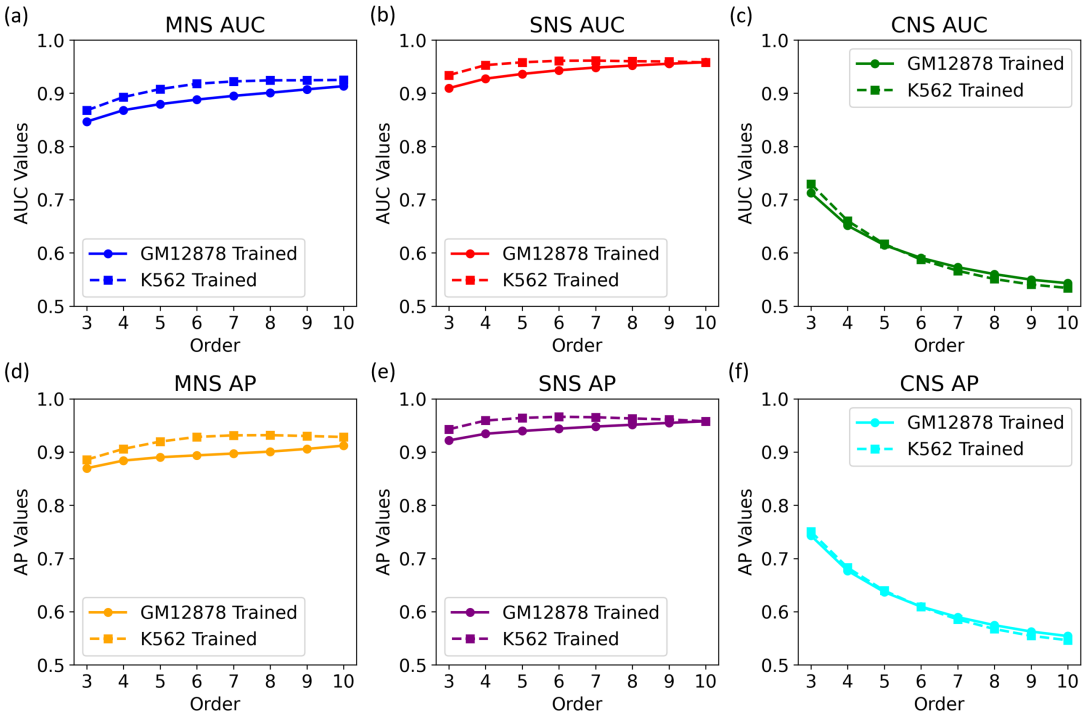
Supplementary Figure 4 Analysis of AUC in inter-cell line testing for each chromosome.



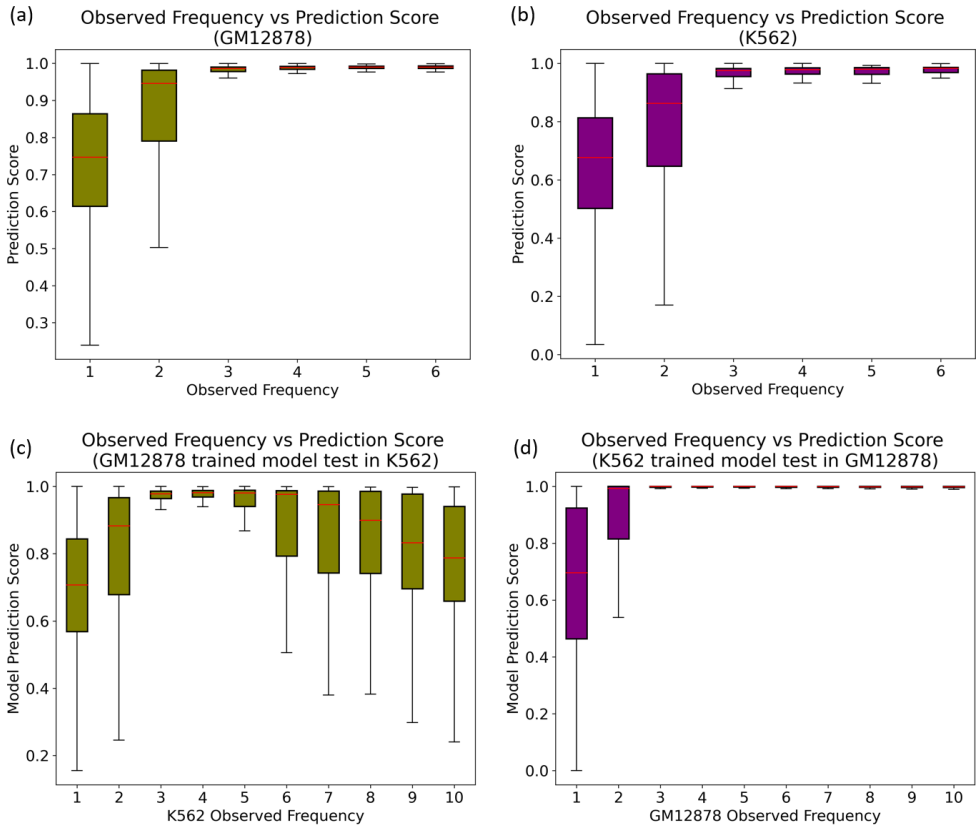
Supplementary Figure 5 Analysis of AP in inter-cell line testing for each chromosome.



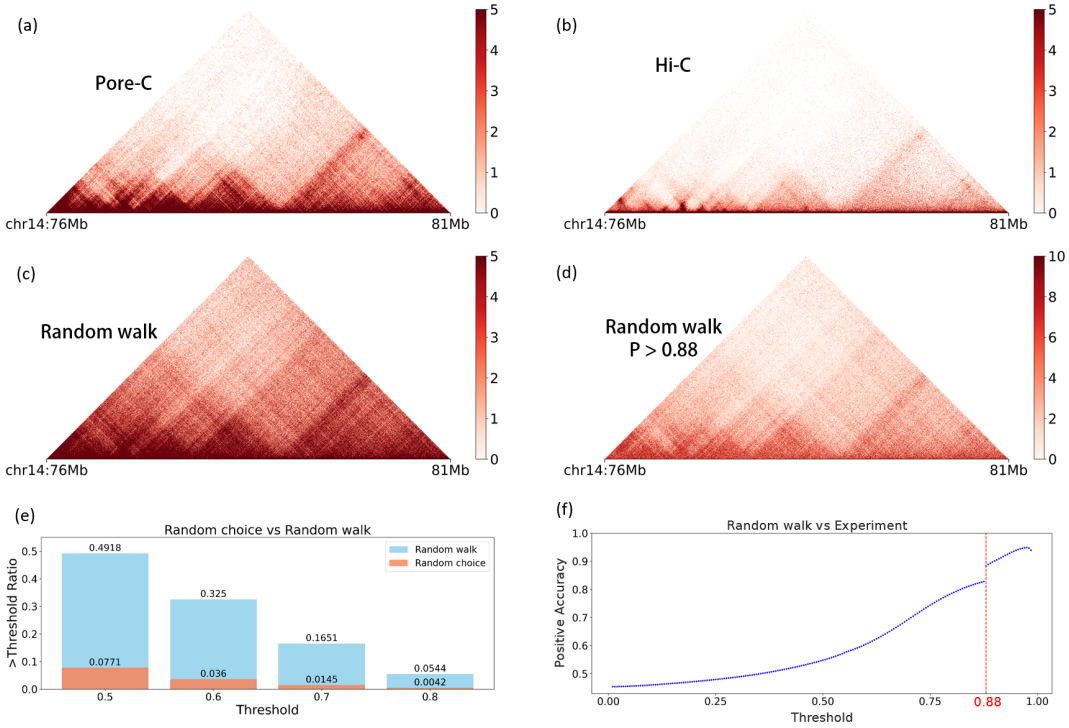
Supplementary Figure 6 Analysis of AUC and AP in intra-cell line testing for different order groups.



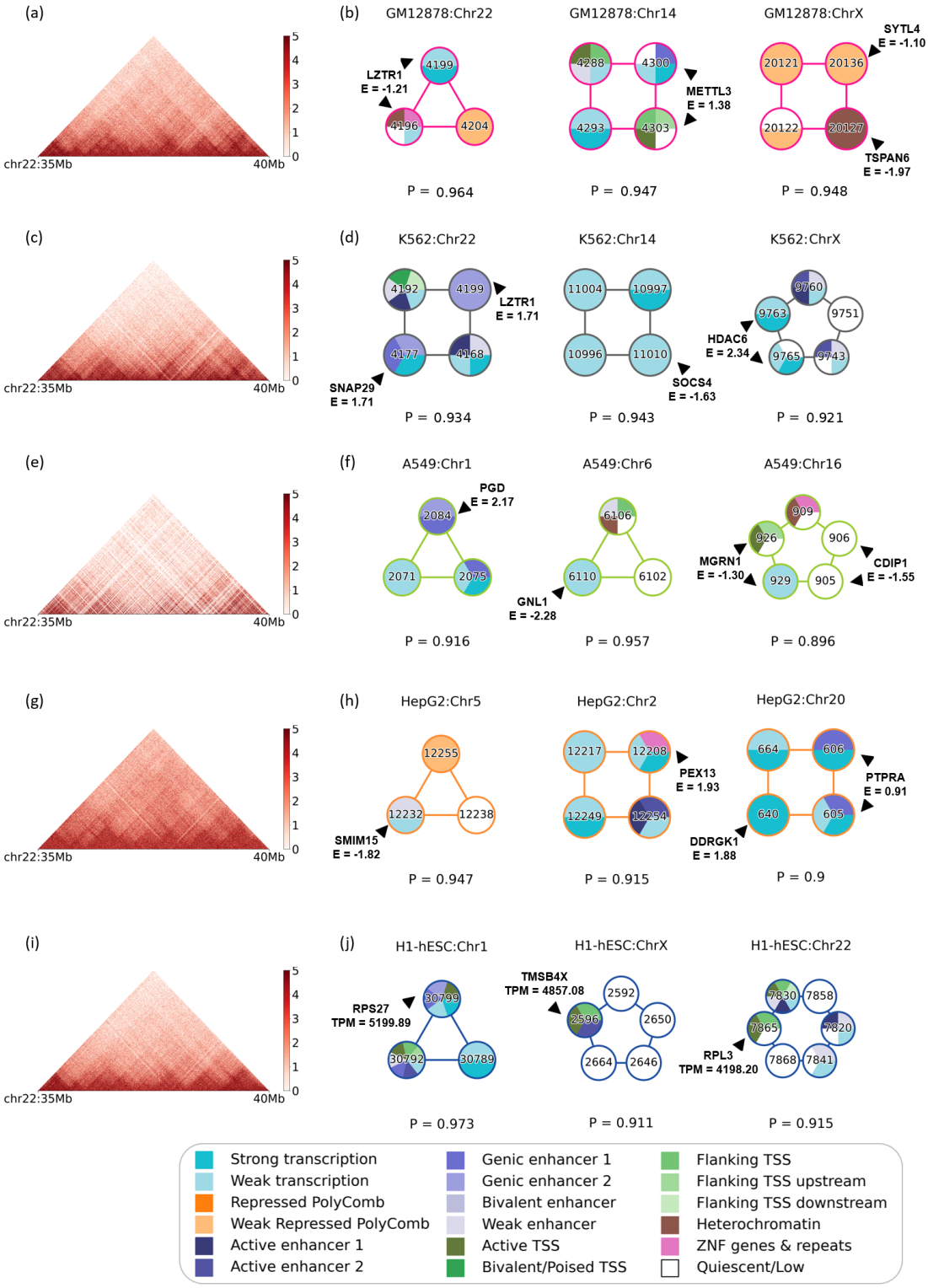
Supplementary Figure 7 Analysis of AUC and AP in inter-cell line testing for different order groups.



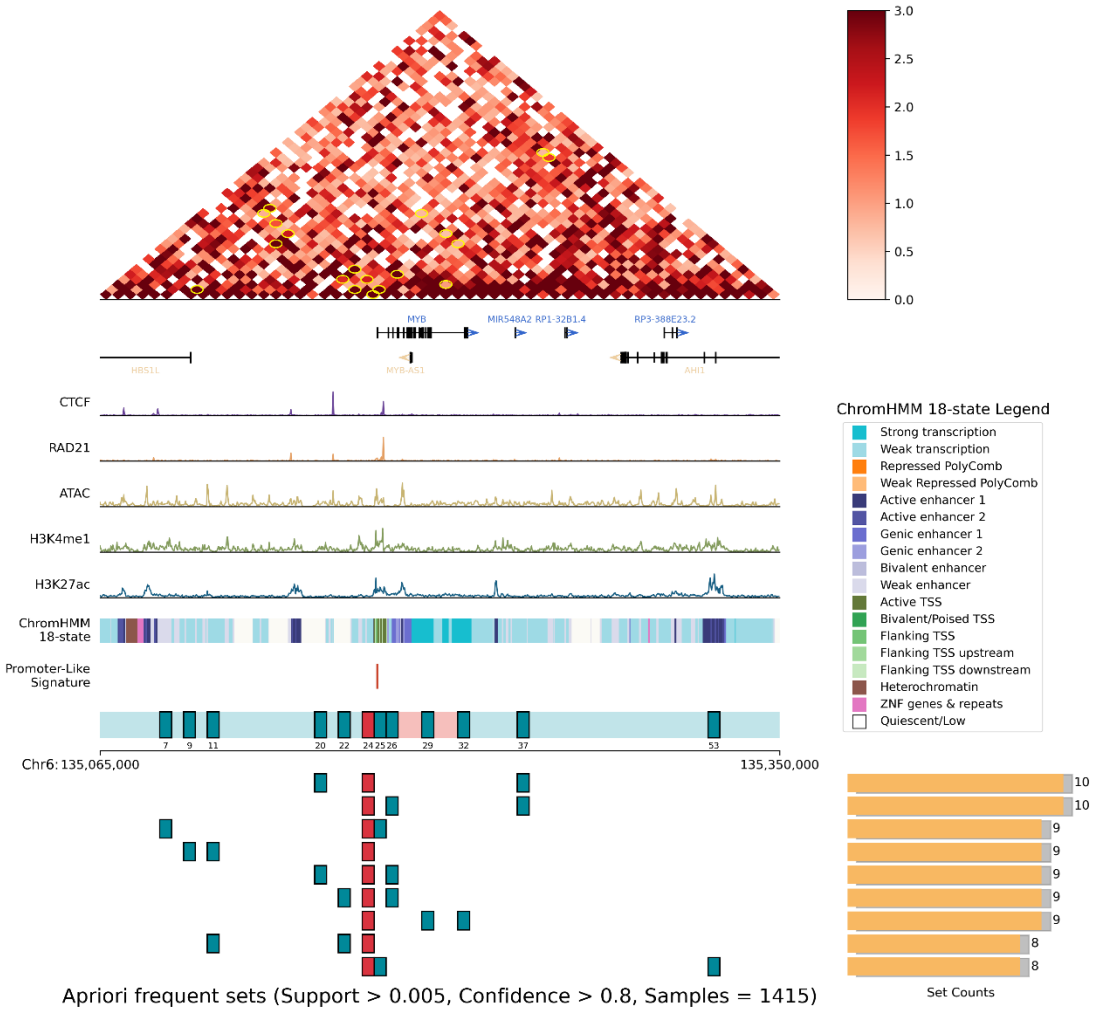
Supplementary Figure 8 Comparison of multi-way interactions with different observed frequencies against prediction scores. (a-b) Intra cell lines. (c-d) Inter cell lines.



Supplementary Figure 9 Comparison of sampling methods and positive accuracy analysis. (a) Experimentally observed Pore-C pairwise contact map (GM12878). (b) Experimental Hi-C pairwise contact map (GM12878). (c) Pairwise contact map of multi-way interactions generated by random walk (2000000 samples, GM12878). (d) Pairwise contact map of multi-way interactions generated by random walk (2000000 samples, GM12878). (e) Comparing the ratio of multi-way interaction prediction scores exceeding different thresholds generated from random choice against those from random walk on the HiC cis-topological graph. (f) Positive accuracy of comprehensive model (GM12878 and K562, Test set).



Supplementary Figure 10 Example of generated multi-way interaction pairwise contact map. Example of gene expression and chromatin state in multi-way interaction samples generated and filtered by the deep learning model.



Supplementary Figure 11 Apriori based multi-way association rules from experimental Pore-C data of *MYB* gene.

References

1. Reiff, S. B. *et al.* The 4D Nucleome Data Portal as a resource for searching and visualizing curated nucleomics data. *Nature Communications* **13**, 2365 (2022).
2. Dunham, I. *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
3. Benjamin C. Hitz *et al.* The ENCODE Uniform Analysis Pipelines. *bioRxiv* 2023.04.04.535623 (2023) doi:10.1101/2023.04.04.535623.
4. Luo, Y. *et al.* New developments on the Encyclopedia of DNA Elements (ENCODE) data portal. *Nucleic Acids Research* **48**, D882–D889 (2020).
5. Abdennur, N. & Mirny, L. A. Cooler: scalable storage for Hi-C data and other genomically labeled arrays. *Bioinformatics* **36**, 311–316 (2020).
6. Durand, N. C. *et al.* Juicebox Provides a Visualization System for Hi-C Contact Maps with Unlimited Zoom. *Cell Systems* **3**, 99–101 (2016).
7. Ramírez, F. *et al.* deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Research* **44**, W160–W165 (2016).

8. Xiao, M., Zhuang, Z. & Pan, W. Local Epigenomic Data are more Informative than Local Genome Sequence Data in Predicting Enhancer-Promoter Interactions Using Neural Networks. *Genes* **11**, (2020).
9. Zhong, J.-Y. *et al.* High-throughput Pore-C reveals the single-allele topology and cell type-specificity of 3D genome folding. *Nature Communications* **14**, 1250 (2023).
10. Edgar, R., Domrachev, M. & Lash, A. E. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Research* **30**, 207–210 (2002).
11. Barrett, T. *et al.* NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Research* **41**, D991–D995 (2013).
12. Ernst, J. & Kellis, M. ChromHMM: automating chromatin-state discovery and characterization. *Nature Methods* **9**, 215–216 (2012).
13. Ernst, J. & Kellis, M. Chromatin-state discovery and genome annotation with ChromHMM. *Nature Protocols* **12**, 2478–2492 (2017).
14. Dodero-Rojas, E. *et al.* PyMEGABASE: Predicting Cell-Type-Specific Structural Annotations of Chromosomes Using the Epigenome. *Journal of Molecular Biology* **435**, 168180 (2023).
15. Contessoto, V. G. *et al.* The Nucleome Data Bank: web-based resources to simulate and analyze the three-dimensional genome. *Nucleic Acids Research* **49**, D172–D182 (2021).
16. Rouillard, A. D. *et al.* The harmonizome: a collection of processed datasets gathered to serve and mine knowledge about genes and proteins. *Database* **2016**, baw100 (2016).
17. P. Valdivia, P. Buono, C. Plaisant, N. Dufournaud, & J. -D. Fekete. Analyzing Dynamic Hypergraphs with Parallel Aggregated Ordered Hypergraph Visualization. *IEEE Transactions on Visualization and Computer Graphics* **27**, 1–13 (2021).