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

Abstract

Understanding how the three-dimensional (3D) organization of chromatin shapes gene regulation requires moving beyond pairwise contacts to capture higher-order interactions among multiple genomic loci. Although emerging experimental techniques such as Pore-C provide snapshots of such multi-way interactions, their limited coverage and inherent sparsity hinder systematic characterization and cross-cell-type generalization. Here we present PHOCI (Predictor of Higher-Order Chromatin Interactions), a computational framework for probabilistic modeling of multi-way chromatin interactions directly from Hi-C and epigenomic data. Across multiple cell lines, PHOCI successfully captures experimentally observed interactions from challenging negative configurations and generalizes to previously uncharacterized cell types. The framework further enables generation of realistic multi-way interaction datasets and identification of recurrent multi-locus regulatory modules through association rule mining. Notably, at the MYB locus in K562 cells, predicted enhancer combinations exhibit non-additive, synergistic effects on gene expression, as validated by CRISPR interference experiments, providing direct evidence that higher-order interactions encode regulatory logic beyond pairwise contacts. By transforming sparse pairwise measurements into a probabilistic representation of higher-order chromatin architecture, PHOCI bridges a key gap in genomic modeling and provides a scalable foundation for studying the multi-body regulatory principles underlying genome function.

Introduction

Chromatin is a dynamic, hierarchically organized polymer whose three-dimensional (3D) architecture is fundamental to gene regulation, genomic stability, and cellular identity[1–4]. While conventional models focus on linear annotations or pairwise contacts, emerging evidence suggests that multi-way interactions, i.e., simultaneous contacts among multiple genomic loci, function as the definitive units of transcriptional control[5–8]. Our recent theoretical model and related works further demonstrated that such multi-way interactions are pervasive throughout the genome and are intimately linked to major features of chromatin organization[9–11].

Despite their biological importance, systematically characterizing higher-order chromatin interactions remains a major challenge. Emerging technologies such as Tri-C[12], Pore-C[13,14], ChIA-Drop[15], and SPRITE[16] provide high-resolution snapshots of these assemblies, yet they are often

constrained by substantial experimental costs, technical complexity, and incomplete genomic coverage. Because multi-way interactions exist in a significantly higher-dimensional space than two-body contacts, experimental sampling is naturally sparse, making it difficult to systematically distinguish functional regulatory hubs from incidental spatial proximity. Moreover, current experimental datasets lack a systematic probabilistic framework, limiting their utility for quantitative modeling and cross-cell-type generalization.

In parallel, artificial intelligence (AI) has emerged as a powerful paradigm for modeling biological systems. The success of AlphaFold[17–19] highlights the ability of deep learning to extract complex patterns from large-scale data. At the same time, there is growing interest in building an AI-driven "virtual cell"[20,21], capable of predicting cellular behavior from molecular and genomic information. However, a substantial gap remains between these two levels: while molecular structure prediction has achieved remarkable accuracy, predictive modeling of cellular behavior remains limited. Bridging this gap requires accurate modeling of genome regulation, as the genome forms the central layer linking molecular interactions to cellular function. In this context, genomic prediction represents a critical intermediate step.

Existing computational approaches have begun to address aspects of this problem[22–30]. Models such as AlphaGenome[31] demonstrate that deep learning can infer chromatin features and contact maps from large-scale data, while widely used frameworks such as the Activity-by-Contact (ABC) model[32] provide practical tools for linking enhancers to target genes. However, these approaches are largely pairwise-centric and typically lose predictive power beyond short genomic distances (<100 kb) [31,32]. These limitations highlight a fundamental gap between current predictive models and the true complexity of genome organization. Chromatin regulation is not merely a collection of independent pairwise contacts, but a higher-order system composed of combinatorial, multi-locus assemblies. Capturing this complexity requires a transition from pairwise representations to probabilistic modeling of higher-order interactions, alongside computational strategies that can overcome the sparsity and biases of experimental data.

Here we address this challenge by introducing PHOCI (Predictor of Higher-Order Chromatin Interactions), a computational framework for predicting multi-way chromatin interactions directly from widely available Hi-C[33] and epigenomic data. We represent the genome as a cis-topological graph derived from Hi-C, in which nodes correspond to genomic bins and edges encode pairwise contacts. Within this framework, multi-way interactions are represented as hyperedges, and biased random walks are used to efficiently sample candidate multi-locus configurations, enabling scalable exploration of the combinatorial interaction space. We then develop a graph neural network based on GraphSAGE[34] to encode genomic loci by integrating chromatin topology with epigenetic features. Candidate interactions are evaluated through an embedding aggregation scheme to predict the probability of simultaneous interactions among multiple loci, providing a probabilistic description of higher-order chromatin organization. Notably, PHOCI does not require experimentally measured multi-way interactions as input; instead, it is trained using experimentally derived positives alongside carefully designed hard negative samples, enabling the model to learn discriminative features of biologically meaningful configurations. This design supports generation of realistic multi-way interaction datasets for previously uncharacterized cell types.

Furthermore, by coupling these predictions with association rule mining, we identify recurrent multi-locus enhancer–promoter modules, offering a transformative lens for dissecting the functional logic of the genome. Notably, these regulatory predictions are supported by experimental validation. For example, at the MYB locus in K562 cells, our model predicts specific multi-way interactions involving distinct combinations of distal enhancers. Using CRISPR interference (CRISPRi)[35], we demonstrate that simultaneous perturbation of selected enhancer pairs leads to non-additive, synergistic effects on MYB transcription, consistent with the predicted multi-way regulatory configurations. These results highlight the ability of our framework to uncover functionally relevant higher-order regulatory logic that is not accessible through pairwise models.

Together, our work establishes a scalable and interpretable framework for reconstructing the probabilistic, higher-order architecture of chromatin, bridging the gap between sparse experimental measurements and the combinatorial complexity of genome regulation. By enabling prediction and analysis of multi-way interactions across cell types, this approach provides a critical step toward comprehensive genomic modeling and advances the broader goal of predictive, AI-driven biology.

Overview of the PHOCI framework

To capture the combinatorial complexity of higher-order chromatin organization, we developed PHOCI (Predictor of Higher-Order Chromatin Interactions), an integrated computational framework that combines graph representation learning[34,36], stochastic sampling and association rule mining to reconstruct the probabilistic landscape of multi-way interactions (Fig. 1a). Starting from Hi-C–derived cis-topological graphs, in which genomic loci (5 kb bins) are represented as nodes and chromatin contacts as edges, we formulate multi-way interaction prediction as a graph-based learning problem and employ a constrained random walk–based sampling strategy to efficiently explore the combinatorial interaction space without requiring Pore-C data as input (Fig. 1b). Candidate multi-locus interaction sets generated in this manner are evaluated by a GraphSAGE-based[34] graph neural network that learns latent node embeddings from epigenomic features, including histone modifications and transcription factor occupancy, capturing both local regulatory states and their broader genomic context. These embeddings are processed through an embedding sampling and aggregation scheme to derive interaction-level descriptors, which are subsequently scored by a multilayer perceptron to prioritize high-confidence multi-way interactions. The model is trained using experimentally derived Pore-C interactions as positives together with structured negative sampling strategies that enforce discrimination across different levels of topological similarity. To further extract biologically interpretable regulatory patterns, we integrate an association rule mining framework centered on gene transcription start sites (TSS), applying the Apriori algorithm[37] to model-predicted interactions with distance-aware background controls to identify enriched multi-way interaction rules. Together, PHOCI provides a scalable and unified framework for the prediction and systematic characterization of higher-order chromatin organization. Further details are provided in the Methods section.

PHOCI distinguishes higher-order chromatin interactions from high-dimensional noise

Having established the PHOCI framework, we next evaluated its ability to identify authentic multi-locus assemblies within complex genomic environments. We benchmarked the model using three

increasingly stringent negative sampling strategies: Sized (SNS), Motif (MNS), and Clique Negative Sampling (CNS) (Fig. 1c–f) [38,39]. While SNS and MNS capture coarse structural constraints and genomic distance biases, CNS provides a stringent test by replacing a single node in an experimentally derived Pore-C cluster with a decoy node that preserves pairwise contacts with all neighbor nodes but disrupts higher-order coherence. Across validation, intra-cell line, and inter-cell line test sets, PHOCI achieved consistently high performance, with strong Area Under the ROC Curve (AUC) and Average Precision (AP) scores[40,41] (Fig. 2a–c; Supplementary Table 6). These results demonstrate robust discrimination between true multi-way interactions and structurally plausible but functionally inconsistent configurations across both GM12878 and K562 cell lines.

To further dissect the model's decision-making, we analyzed its response to CNS perturbations across interaction orders (three- to six-way interactions; Supplementary Fig. 2a–c). While predicted probabilities generally scaled with inverse genomic distance, consistent with the polymer nature of chromatin folding, distance alone was insufficient to explain PHOCI's predictions. We identified multiple "epigenetic override" cases in which the model rejected spatially closer configurations in favor of more distal loci with distinct epigenomic characteristics (Supplementary Fig. 2h–j). For example, in a three-way interaction (bins 780–782–659), replacing an active POLR2A-enriched locus with a closer H3K9me3-marked region substantially reduced the predicted probability. Similarly, in four- and five-way assemblies, substitutions that altered enhancer-associated signals (H3K27ac and ATAC-seq) or architectural features (CTCF binding) consistently led to marked decreases in predicted interaction probability. Notably, these prediction shifts could not be readily explained by any single epigenetic feature, nor by simple combinations of known chromatin marks. Instead, they suggest that PHOCI learns complex higher-order representations that integrate chromatin topology with epigenomic context. Together, these results indicate that PHOCI leverages both epigenetic information and distance-derived topological constraints to predict multi-way chromatin interactions. More broadly, they highlight the intricate and largely unresolved nature of the epigenetic encoding underlying higher-order genome organization, suggesting that predictive patterns governing multi-way chromatin interactions may extend beyond current biological intuition and mechanistic frameworks.

PHOCI generalizes across genomic contexts and captures multiscale chromatin organization

We next assessed the robustness of PHOCI across chromosomes, cell types, and interaction complexities. Model performance remained stable across all autosomes in both intra- and inter-cell line evaluations (Fig. 2d–f), with only a modest decrease observed on the X chromosome, likely reflecting its distinct epigenetic landscape and absence from the training set (Supplementary Fig. 3–5).

Analysis across interaction orders revealed distinct trends depending on the sampling strategy. For SNS and MNS, performance improved with increasing interaction order, as higher-dimensional configurations diverge more strongly from random expectations. In contrast, CNS becomes progressively more challenging at higher orders, as single-node substitutions represent smaller perturbations to the overall interaction set. Despite this, PHOCI maintained good predictive accuracy (Fig. 2g–i), demonstrating its ability to resolve subtle differences in high-dimensional interaction space. Moreover, predicted probabilities showed strong correlation with experimental

contact frequencies (Fig. 2j), supporting the quantitative fidelity of the model. Furthermore, a high consistency in prediction scores between the intra- and inter-cell line models highlights PHOCI's robust generalizability, underscoring its potential to effectively predict multi-way interactions in uncharacterized cell lines using only baseline genomic and epigenetic inputs (Fig. 2k).

To probe the internal representations learned by PHOCI, we analyzed latent embeddings derived from the GraphSAGE encoder. Dimensionality reduction using UMAP[42,43] revealed clear segregation corresponding to A/B chromatin compartments (Fig. 3a, e). K-means Clustering analysis[44] further demonstrated that these embeddings recapitulate hierarchical chromatin structures, with large clusters corresponding to topologically associating domains (TADs) and smaller clusters reflecting sub-TAD organization (Fig. 3d, h). These results indicate that PHOCI captures multiscale genome organization as an emergent property of its representation learning.

PHOCI captures the functional coupling between higher-order chromatin interactions and transcriptional regulation

We next investigated whether PHOCI's predictions reflect functional regulatory states. Across the genome, high-probability multi-way interactions were significantly enriched for active enhancer signatures and highly expressed genes[45,46] (Fig. 4a–d). To assess the model's sensitivity to cellular context, we compared standardized expression values and ChromHMM states[47,48] across cell lines. We observed that PHOCI consistently assigns high probabilities to structural configurations that align with functional activity; specifically, the model's predictions closely track the observed biological transitions, maintaining its predictive power even when nodes shift from enhancer-rich to repressive states.

At specific loci, PHOCI captures cell-type-specific regulatory contexts. For example, at the *CRYBB2* locus, which is transcriptionally repressed in K562 cells, PHOCI assigned high probability to a multi-way assembly enriched for Polycomb-associated chromatin states. In contrast, in GM12878 cells, where *CRYBB2* is active, the same configuration exhibited reduced probability, reflecting the underlying shift in epigenetic environment. Similarly, at the *FBXO33* locus, interaction probability tightly tracked the dynamic shifts in gene expression levels across cell types, with high-probability assemblies corresponding to enhancer-rich regulatory hubs, while mismatched scenarios—such as enhancer-rich states with low gene expression—consistently received low probabilities. Notably, these associations emerge despite the absence of gene expression data during model training, indicating that PHOCI implicitly captures functional regulatory logic from chromatin topology and epigenomic features alone (Fig. 4e–f).

PHOCI enables de novo discovery and experimental validation of higher-order regulatory logic

Leveraging its generalizability, we applied PHOCI to generate de novo multi-way interaction datasets in cell types lacking Pore-C data (A549, HepG2, and H1-hESC). Using biased random walk sampling followed by stringent probability filtering ($P > 0.88$), we identified stringently filtered interaction sets enriched in active regulatory regions (Supplementary Fig. 9), demonstrating the scalability of the framework for reconstructing higher-order chromatin architecture in previously uncharacterized systems.

We next sought to determine whether PHOCI can uncover functionally relevant regulatory modules. Focusing on the *MYB* proto-oncogene in K562 cells, a key regulator in leukemogenesis, we anchored our analysis at a genomic region proximal to the transcription start site and generated candidate multi-way interactions across the cis-topological graph. Applying Apriori-based association rule mining[37] to high-confidence predictions (Fig. 5a), we identified recurrent higher-order regulatory assemblies involving both proximal and distal regulatory elements.

PHOCI revealed two dominant multi-way modules: (*MYB* promoter, Locus B, Locus C) and (MYB promoter, Locus A, Locus C). While Locus B corresponds to a known proximal regulatory element, Loci A and C are more distal and appear as weak or isolated signals in conventional pairwise contact maps. Notably, comparison with the widely used Activity-by-Contact (ABC) model[32] showed that while ABC successfully identified the promoter–Locus B interaction, it failed to detect the distal Locus C or the upstream Locus A, highlighting the limitation of pairwise frameworks in capturing higher-order regulatory dependencies.

Importantly, these multi-way assemblies are not readily accessible through conventional analysis or expert inspection. The pairwise contacts underlying these modules are often weak or statistically insignificant in Hi-C maps, and therefore do not stand out as candidate regulatory interactions. Moreover, the regulatory logic governing these assemblies cannot be attributed to any single epigenetic feature; instead, it emerges from a complex integration of multiple epigenetic signals and sequence-distance–dependent topological constraints. PHOCI effectively captures this high-dimensional combinatorial effect, enabling the identification of regulatory modules that remain obscured under traditional pairwise or feature-centric analyses.

To directly test the functional relevance of these predicted assemblies, we performed CRISPR interference (CRISPRi)[35] experiments targeting Loci A, B, and C. Simultaneous perturbation of Loci B and C resulted in a pronounced reduction in *MYB* expression that significantly exceeded the additive effect of individual perturbations (Fig. 5e), indicating synergistic regulatory coupling. Similarly, co-perturbation of Loci A and C produced a strong synergistic repression of *MYB* transcription (Fig. 5c). In contrast, simultaneous perturbation of Loci A and B resulted in only additive effects (Fig. 5g), consistent with PHOCI's prediction that this combination does not form a functional higher-order module. These results were reproducible across multiple sgRNAs (Fig. 5d, f, h; Supplementary Data), providing strong experimental validation of the model's ability to distinguish functionally interdependent multi-way assemblies from independent pairwise interactions.

Together, these findings demonstrate that PHOCI not only predicts higher-order chromatin interactions with high accuracy, but also reveals the combinatorial regulatory logic underlying gene expression, enabling the identification of functional multi-locus hubs that are inaccessible to traditional pairwise analyses.

Discussion and conclusions

In this study, we introduced PHOCI, a computational framework designed to transcend the inherent

limitations of sparse, pairwise genomic measurements by reconstructing the probabilistic landscape of multi-way chromatin interactions. Deciphering long-range, higher-order regulatory modes of gene regulation remains a central challenge in genome biology: while enhancer–promoter interactions are often studied as pairwise events, growing evidence suggests that gene expression is frequently governed by cooperative assemblies involving multiple distal elements. However, such higher-order structures are difficult to resolve experimentally and are largely inaccessible to conventional computational approaches. By representing the genome as a cis-topological graph and integrating multi-modal epigenetic features, PHOCI provides a principled approach to navigating this combinatorial space, enabling the inference of multi-locus regulatory assemblies without requiring direct multi-way experimental input.

Importantly, PHOCI addresses a key gap left by existing modeling paradigms. Widely used methods such as the Activity-by-Contact (ABC) model[32] are fundamentally limited to pairwise interactions and therefore cannot capture the cooperative or non-additive effects that arise in multi-element regulatory hubs. Similarly, recent sequence-based deep learning models, such as AlphaGenome[31], excel at predicting local regulatory features from DNA sequence but do not explicitly model the topological coupling between multiple distal loci. As a result, both classes of approaches are inherently constrained in their ability to resolve long-range, higher-order regulatory logic. In contrast, PHOCI explicitly models multi-body interactions as first-class entities, allowing it to uncover regulatory modules that remain hidden when considering pairwise contacts or individual epigenetic features in isolation.

Our results support a shift toward viewing genome regulation as an emergent property of high-dimensional chromatin organization. The alignment of PHOCI's latent representations with A/B compartments and topologically associating domains, together with the experimental validation of synergistic multi-way interactions at the *MYB* locus, demonstrates that the model captures biologically meaningful structure across scales. In particular, the observation that combinatorial perturbations produce non-additive effects provides direct evidence that higher-order chromatin assemblies encode functional regulatory logic beyond the sum of their pairwise components. These findings suggest that deep learning frameworks can begin to approximate the "epigenetic grammar" governing multi-locus coordination, advancing toward a predictive, systems-level understanding of genome function.

Despite its predictive power, PHOCI also inherits limitations from both the underlying data and the modeling framework. We observed a "fragment-shifting" phenomenon, whereby predicted regulatory hubs are displaced by approximately 1–2 genomic bins (5–10 kb) from their precise functional locations. This likely reflects the biochemical constraints of proximity ligation assays, in which physically adjacent but functionally distinct fragments can be cross-linked and sequenced together. As PHOCI is trained on these data, it inevitably incorporates this spatial ambiguity. Future improvements may address this limitation by integrating higher-resolution sequence information or by developing deconvolution strategies that disentangle functional interactions from incidental spatial proximity.

More broadly, the regulatory landscape of the genome involves layers of complexity that are not yet

fully captured. While PHOCI incorporates a diverse set of epigenetic features, it does not explicitly model transcription factor binding dynamics, enhancer RNA activity, or the dynamic behavior of transcriptional condensates, which may play an important role in driving multi-way genomic interactions. In addition, the current implementation operates at a 5-kb resolution, which is fundamentally constrained by the resolution of Hi-C data and may limit sensitivity to fine-scale regulatory features. At the same time, PHOCI models interactions within 5-Mb cis-topological subgraphs, enabling the capture of long-range regulatory dependencies that extend beyond the range of many existing approaches; further extending this contextual window may provide an avenue for resolving even larger-scale regulatory architectures. A practical limitation is that the current framework requires both Hi-C data and matched epigenetic profiles as input, which may not be readily available for all cell types or conditions. As with many deep learning approaches, interpretability also remains an open challenge. Although our analyses provide insight into the learned representations, a fully quantitative understanding of how specific features contribute to individual multi-way predictions remains to be established.

Looking forward, PHOCI highlights the potential of artificial intelligence to address one of the most challenging problems in genome biology: the reconstruction of higher-order regulatory architectures from incomplete and noisy data. Future developments may extend this framework toward multi-scale, foundation-model-like architectures that integrate sequence, epigenomic, and transcriptomic information, as well as single-cell variability. Such models could enable the systematic exploration of long-range, higher-order cis-regulatory interactions and their dynamic reorganization during development and disease. While PHOCI does not yet fully resolve the complexity of multi-locus regulation, it provides a concrete step toward this goal, demonstrating that AI-driven approaches can uncover regulatory principles that are difficult to access through experimental or analytical methods alone.

Methods

1. Deep learning model

In order to discern and forecast potential multi-way interactions from a dataset generated through random walk sampling, we developed a deep learning model based on graph neural networks using PyTorch[49] and PyTorch Geometric[50]. This model was specifically designed to capture complex relationships between nodes, as illustrated in Figure 1(a).

Our model's network architecture incorporates the GraphSAGE Encoder module to perform graph convolution and encoding on input cis-topological subgraphs (consecutively along the DNA sequence for 1000 bins × 5kb, each bin is a node), obtaining hidden embeddings for each node. Specifically, the GraphSAGE Encoder consists of three GraphSAGE layers[34]. It processes features composed of 15 types of epigenetic modifications and transcription factor through a Linear layer to compute embeddings, which are then normalized. Sequentially, these embeddings are fed into the GraphSAGE layers, obtaining hidden embeddings for each node within the entire cis-topological subgraph. After that, the next step involves the Embedding Sampling and Aggregation (ESA) module for sampling and aggregating node embeddings with multi-way interactions. In this module, we extract the embeddings of a group of nodes to predict the existence of multi-way interactions. Subsequently, we perform dual aggregation on this set of hidden embeddings, aggregating maximum minus minimum values and averaging[39]. Next, by concatenating these aggregated results

and incorporating node count and sequence span distance, we construct final descriptors for these multi-way interactions. These descriptors are then inputted into a multi-layer perceptron to obtain the probability score of multi-way interactions occurring in this group of nodes. Further details of the specific structural configuration of the model can be found in Supplementary Table 1.

For effective model training, the utilization of appropriate positive and negative samples was imperative. Positive samples consisted of multi-way interaction data, obtained from high-throughput Pore-C experiments, between nodes on cis-topological subgraphs. In contrast, the selection of negative samples involved three distinct methods: Sized Negative Sampling (SNS), Motif Negative Sampling (MNS), and Clique Negative Sampling (CNS) [38,39]. Specifically, SNS involved the random selection of k nodes to fill a set; MNS merged duplicates with a randomly related multi-way interaction to fill a set of size k, while CNS replaced one node from a randomly selected positive sample with a random node multi-way interacted with all other nodes. Notably, distinguishing the negative samples generated by CNS from the positive samples proved to be the most challenging (Figure 1 (c-f)). To maintain balanced training data, we sampled an equal number of three types of negative samples as positive samples. For each epoch, we randomly selected 1/3 of each type and combined them to create a balanced set of negative samples used for training. This approach ensured effective model learning while maintaining data balance.

Once convergence in model training was achieved, comprehensive testing on the validation set was conducted for AUC (Area Under the Curve) and AP (Average Precision). AUC, a metric assessing classification model performance, measures the model's ability to rank samples, prioritizing positives over negatives. Its values range from 0 to 1, where higher values indicate a stronger ability to distinguish between positive and negative samples. AP evaluates the quality of ranking results in retrieval tasks[51]. Its values also range from 0 to 1, with higher values indicating better model performance in sorting tasks by accurately ordering positive class samples according to the true order[41]. By meticulously evaluating the performance of AUC and AP on the validation set for each epoch, we selected the model corresponding to the epoch with the highest scores. This step aimed to ensure that the model performance achieved optimal validation, thus providing a reliable foundation for further model application and deployment.

2. Generating multi-way interaction sets through random walk sampling

We employed a random walk sampling method based on the cis-topological graph formed by HiC to generate a random multi-way interaction dataset. Specifically, we randomly selected a node as a starting point and performed a random walk between its neighboring nodes along the edges of the graph. The length of the random walk path was chosen to match the distribution of multi-way interaction data lengths in the training set. We utilized the 'random walk' module[52] from the Python package PyTorch Geometric[50] to generate random sequences of nodes in the graph. To prevent looping back or generating duplicate nodes during the random walk, we performed deduplication on the nodes in the generated path, which will be subsequently used with trained deep learning models to filter out genuinely probable multi-way interaction sets (Figure 1 (b)).

3. Mining significant multi-way association rules related to specific gene TSS using the Apriori algorithm

We employed the Apriori algorithm[37] to mine significant multi-way association rules related to specific gene transcription start sites (TSS). First, we select the TSS region of the target gene from

the genomic dataset and determine its corresponding bin. This target bin serves as the starting point for generating a multi-way interaction dataset across the cis-topological graph, referred to as the signal dataset. Next, we apply a low support threshold to the signal dataset for an initial Apriori mining process to identify the proximal regions of the target gene in the sequence. Within these proximal regions, we randomly select new points as starting bins and generate new multi-way interaction datasets across the cis-topological graph, termed the background dataset. Subsequently, we calculate a normalized value based on the inverse of the distance between the closest bin in each multi-way interaction and the target bin. This value is used as the extraction probability, meaning interactions closer to the target bin are more likely to be selected. Then, using this probability, we extract an amount of data from the background dataset equal to the size of the signal dataset. Additionally, we randomly extract half as much data from the background dataset. Finally, we combine the signal dataset with the two sets of extracted background data. This step aims to use the background dataset for contrast, constructing a suitable itemset dataset for association rule mining. We then set a high confidence threshold and reapply the Apriori algorithm to the combined dataset, identifying frequently occurring multi-way association rules. This method allows us to identify significant multi-way interaction patterns related to specific gene TSS, providing crucial insights into gene expression regulation mechanisms and laying the foundation for further functional validation experiments.

4. Cell culture

K562 cells (human myelogenous leukemia cells, female, ATCC CCL-243) were cultured in RPMI-1640 medium (Thermo Fisher Scientific) supplemented with 10% FBS (VISTECH), GlutaMAX (Thermo Fisher Scientific), and Penicillin-Streptomycin (Thermo Fisher Scientific). Stable integration of EF1α-dCas9-KRAB into K562 cells was achieved via lentiviral transduction. HEK293T cells (human embryonic kidney epithelial cells, female; ATCC CRL-3216) were maintained in DMEM (Thermo Fisher Scientific) containing 10% FBS (VISTECH). All cells were cultured at 37°C in a humidified incubator with 5% CO_2.

5. sgRNA design

We used Benchling to design enhancer sgRNAs, selecting sgRNAs with on-target scores >50 and off-target scores < 50, and utilized CHOPCHOP to design sgRNAs for the TSS locus of the target gene, choosing sgRNAs with on-target scores > 60.

6. Plasmid construction

To clone a single sgRNA, the lentiGuide-Puro plasmid (Addgene #52963) was linearized with BsmBI (NEB) and gel-purified. A double-stranded DNA fragment containing the target sgRNA sequence was created by annealing oligonucleotides and then ligated into the linearized backbone using T4 DNA ligase (NEB), following the manufacturer's protocol.

For dual sgRNA expression vector construction, we first modified the lentiGuide-Puro backbone (Addgene #52963) by inserting an mU6-sgRNA scaffold. The vector was then linearized with BsmBI and gel-purified. A PCR-amplified fragment containing two sgRNA cassettes (sgRNA1-hU6 sgRNA scaffold-mU6 promoter-sgRNA2) was gel-purified and cloned into the linearized vector using homologous recombination (Vazyme).

7. Lentivirus production

HEK293T cells were transiently transfected with sgRNA constructs and packaging plasmids psPAX2 and PMD2.0G to produce lentivirus. Lentivirus was then collected 48 hours after transfection by filtering the supernatant through 0.45 μm filters.

8. Cell treatments

K562 cells stably expressing dCas9-KRAB were seeded at a density of 1×10^5 cells per well in 24-well plates and then infected with either single or paired sgRNA viruses targeting enhancers or TSS. One day later, the medium was replaced with fresh medium. After three days, 3 μg/ml puromycin was added for 4 days. On the seventh day, the cells were harvested for qRT-PCR analysis.

9. Quantitative real-time PCR (qRT-PCR)

The qRT-PCR experiments were performed to determine the fold change of the expression of MYB, ADAMTS14, and other tested genes. For each sample, total RNA was isolated using the FastPure Cell/Tissue Total RNA Isolation Kit V2 (Vazyme), followed by cDNA synthesis using the Promega cDNA Synthesis system. qRT-PCR was performed using the PrimePCR assay with the SYBR Green Master Mix (CWBIO) with primers ordered from Tsingke on a Bio-Rad CFX384 real-time system, according to the manufacturer's instructions.

10. Depiction of additive effect regions

The additive regional efficiency was calculated through the following procedure: First, qPCR data from three independent experiments measuring two individual inhibitory effects were averaged separately. The lower efficiency boundary was then determined by summing the lowest mean values across groups, while the upper boundary was established by summing the highest mean values. Finally, the valid range of additive regional efficiency was defined based on these two boundary values.

Data Sources

The Hi-C, ChIP-seq, and ATAC-seq data employed in this study were obtained from 4D Nucleome project[53] and ENCODE project[54–56]. High-throughput Pore-C data were obtained from Zhong et al.[14] and deposited in the NCBI GEO database[57,58]. Detailed accession codes for the data and descriptions of the data processing methods are provided in the Supplementary Information.

Code Availability

The code developed for this research is available on GitHub at https://github.com/yunyiwu/PHOCI under the GNU GENERAL PUBLIC LICENSE Version 3. Interested parties may access the codebase via the provided link.

Acknowledgements

This work was funded by grants from the National Natural Science Foundation of China (32571445 and 32521002 to K.H., 32370691 to C.H.), and the Guangdong Provincial Project (2024TQ08A012 to Y.C.), Yunnan Revitalization Talent Support Program Top Team (202505AT350003 and 202405AS350022 to C.H.); the Science and Technology Department of Yunnan Province (202305AH340007 to C.H.). We acknowledge computational support from the

Shenzhen Bay Laboratory High Performance Computing and Informatics Core.

Author contributions

K.H. conceived, guided and supervised the project; Y.W. developed and trained the AI model, performed computational predictions and analyses, with contributions from Z.Y., Y.W., L.L. and J.X.; X.J. designed and conducted CRISPRi experiments under the supervision of C.Y.; Y.W. and K.H. wrote the manuscript. All authors discussed the results and reviewed the manuscript.

Competing interests

The authors declare no competing interests.

References

1. Misteli, T. The Self-Organizing Genome: Principles of Genome Architecture and Function. *Cell* **183**, 28–45 (2020).
2. Bonev, B. & Cavalli, G. Organization and function of the 3D genome. *Nat. Rev. Genet.* **17**, 661–678 (2016).
3. Sreenivasan, V. K. A., Yumiceba, V. & Spielmann, M. Structural variants in the 3D genome as drivers of disease. *Nat. Rev. Genet.* **26**, 742–760 (2025).
4. Zheng, H. & Xie, W. The role of 3D genome organization in development and cell differentiation. *Nat. Rev. Mol. Cell Biol.* **20**, 535–550 (2019).
5. Allahyar, A. *et al.* Enhancer hubs and loop collisions identified from single-allele topologies. *Nat. Genet.* **50**, 1151–1160 (2018).
6. Tsai, A., Alves, M. R. & Crocker, J. Multi-enhancer transcriptional hubs confer phenotypic robustness. *eLife* **8**, e45325 (2019).
7. Pinglay, S. *et al.* Synthetic regulatory reconstitution reveals principles of mammalian *Hox* cluster regulation. *Science* **377**, eabk2820 (2022).
8. Sabari, B. R. *et al.* Coactivator condensation at super-enhancers links phase separation and gene control. *Science* **361**, eaar3958 (2018).
9. Huang, K. *et al.* Physical and data structure of 3D genome. *Sci. Adv.* **6**, eaay4055 (2020).
10. Carignano, M. A. *et al.* Local volume concentration, packing domains, and scaling properties of chromatin. *eLife* **13**, RP97604 (2024).
11. Li, W. S. *et al.* Mature chromatin packing domains persist after RAD21 depletion in 3D. *Sci. Adv.* **11**, eadp0855 (2025).
12. Oudelaar, A. M. *et al.* Single-allele chromatin interactions identify regulatory hubs in dynamic compartmentalized domains. *Nat. Genet.* **50**, 1744–1751 (2018).
13. Deshpande, A. S. *et al.* Identifying synergistic high-order 3D chromatin conformations from genome-scale nanopore concatemer sequencing. *Nat. Biotechnol.* **40**, 1488–1499 (2022).
14. Zhong, J.-Y. *et al.* High-throughput Pore-C reveals the single-allele topology and cell type-specificity of 3D genome folding. *Nat. Commun.* **14**, 1250 (2023).
15. Zheng, M. *et al.* Multiplex chromatin interactions with single-molecule precision. *Nature* **566**, 558–562 (2019).
16. Quinodoz, S. A. *et al.* Higher-Order Inter-chromosomal Hubs Shape 3D Genome Organization in the Nucleus. *Cell* **174**, 744-757.e24 (2018).
17. Senior, A. W. *et al.* Improved protein structure prediction using potentials from deep learning. *Nature* **577**, 706–710 (2020).
18. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).

19. Abramson, J. *et al.* Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* **630**, 493–500 (2024).
20. Bunne, C. *et al.* How to build the virtual cell with artificial intelligence: Priorities and opportunities. *Cell* **187**, 7045–7063 (2024).
21. Cui, H. *et al.* scGPT: toward building a foundation model for single-cell multi-omics using generative AI. *Nat. Methods* **21**, 1470–1480 (2024).
22. Cheng, W. *et al.* DNALONGBENCH: a benchmark suite for long-range DNA prediction tasks. *Nat. Commun.* **16**, 10108 (2025).
23. Fan, Y. *et al.* GFETM: Genome foundation-based embedded topic model for scATAC-seq modeling. *Cell Syst.* **17**, 101563 (2026).
24. He, Y. *et al.* Diffusion-enhanced characterization of 3D chromatin structure reveals its linkage to gene regulatory networks and the interactome. *Genome Res.* **33**, 1354–1368 (2023).
25. Wang, T. *et al.* CellNavi predicts genes directing cellular transitions by learning a gene graph-enhanced cell state manifold. *Nat. Cell Biol.* **27**, 1863–1874 (2025).
26. Zhang, Z. *et al.* Developing a general AI model for integrating diverse genomic modalities and comprehensive genomic knowledge. *Nucleic Acids Res.* **53**, gkaf1269 (2025).
27. Wang, X., Luan, Y. & Yue, F. EagleC: A deep-learning framework for detecting a full range of structural variations from bulk and single-cell contact maps. *Sci. Adv.* **8**, eabn9215 (2022).
28. Nguyen, E. *et al.* Sequence modeling and design from molecular to genome scale with Evo. *Science* **386**, eado9336 (2024).
29. Brixi, G. *et al.* Genome modelling and design across all domains of life with Evo 2. *Nature* **652**, 1349–1361 (2026).
30. Theodoris, C. V. *et al.* Transfer learning enables predictions in network biology. *Nature* **618**, 616–624 (2023).
31. Avsec, Ž. *et al.* Advancing regulatory variant effect prediction with AlphaGenome. *Nature* **649**, 1206–1218 (2026).
32. Fulco, C. P. *et al.* Activity-by-contact model of enhancer–promoter regulation from thousands of CRISPR perturbations. *Nat. Genet.* **51**, 1664–1669 (2019).
33. Lieberman-Aiden, E. *et al.* Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. *Science* **326**, 289–293 (2009).
34. Hamilton, W., Ying, Z. & Leskovec, J. Inductive Representation Learning on Large Graphs. in *Advances in Neural Information Processing Systems* (eds Guyon, I. et al.) vol. 30 (Curran Associates, Inc., 2017).
35. Larson, M. H. *et al.* CRISPR interference (CRISPRi) for sequence-specific control of gene expression. *Nat. Protoc.* **8**, 2180–2196 (2013).
36. Kipf, T. N. & Welling, M. Semi-Supervised Classification with Graph Convolutional Networks. in *International Conference on Learning Representations* (2017).
37. Toivonen, H. Apriori Algorithm. in *Encyclopedia of Machine Learning* (eds Sammut, C. & Webb, G. I.) 39–40 (Springer US, Boston, MA, 2010).
38. Patil, P., Sharma, G. & Murty, M. N. Negative Sampling for Hyperlink Prediction in Networks. in *Advances in Knowledge Discovery and Data Mining* (eds Lauw, H. W. et al.) 607–619 (Springer International Publishing, Cham, 2020).
39. Hwang, H., Lee, S., Park, C. & Shin, K. AHP: Learning to Negative Sample for Hyperedge Prediction. in *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval* 2237–2242 (Association for Computing Machinery, New York, NY, USA, 2022).
40. Ling, C. X., Huang, J. & Zhang, H. AUC: a statistically consistent and more discriminating measure than

accuracy. in *Proceedings of the 18th International Joint Conference on Artificial Intelligence* 519–524 (Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2003).

41. Zhang, E. & Zhang, Y. Average Precision. in *Encyclopedia of Database Systems* (eds LIU, L. & ÖZSU, M. T.) 192–193 (Springer US, Boston, MA, 2009).
42. McInnes, L., Healy, J. & Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. Preprint at https://doi.org/10.48550/arXiv.1802.03426 (2018).
43. Healy, J. & McInnes, L. Uniform manifold approximation and projection. *Nat. Rev. Methods Primer* **4**, 82 (2024).
44. Jin, X. & Han, J. K-Means Clustering. in *Encyclopedia of Machine Learning* (eds Sammut, C. & Webb, G. I.) 563–564 (Springer US, Boston, MA, 2010).
45. 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).
46. Diamant, I., Clarke, D. J. B., Evangelista, J. E., Lingam, N. & Ma'ayan, A. Harmonizome 3.0: integrated knowledge about genes and proteins from diverse multi-omics resources. *Nucleic Acids Res.* **53**, D1016–D1028 (2025).
47. Ernst, J. & Kellis, M. ChromHMM: automating chromatin-state discovery and characterization. *Nat. Methods* **9**, 215–216 (2012).
48. Ernst, J. & Kellis, M. Chromatin-state discovery and genome annotation with ChromHMM. *Nat. Protoc.* **12**, 2478–2492 (2017).
49. Paszke, A. *et al.* PyTorch: An Imperative Style, High-Performance Deep Learning Library. in *Advances in Neural Information Processing Systems* (eds Wallach, H. et al.) vol. 32 (Curran Associates, Inc., 2019).
50. Fey, M. & Lenssen, J. E. Fast Graph Representation Learning with PyTorch Geometric. in *ICLR 2019 Workshop on Representation Learning on Graphs and Manifolds* (2019).
51. Bradley, A. P. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognit.* **30**, 1145–1159 (1997).
52. Grover, A. & Leskovec, J. node2vec: Scalable Feature Learning for Networks. in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 855–864 (Association for Computing Machinery, New York, NY, USA, 2016).
53. Reiff, S. B. *et al.* The 4D Nucleome Data Portal as a resource for searching and visualizing curated nucleomics data. *Nat. Commun.* **13**, 2365 (2022).
54. Dunham, I. *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
55. Hitz, B. C. *et al.* The ENCODE Uniform Analysis Pipelines. Preprint at https://doi.org/10.1101/2023.04.04.535623 (2023).
56. Luo, Y. *et al.* New developments on the Encyclopedia of DNA Elements (ENCODE) data portal. *Nucleic Acids Res.* **48**, D882–D889 (2020).
57. Edgar, R., Domrachev, M. & Lash, A. E. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.* **30**, 207–210 (2002).
58. Barrett, T. *et al.* NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Res.* **41**, D991–D995 (2013).

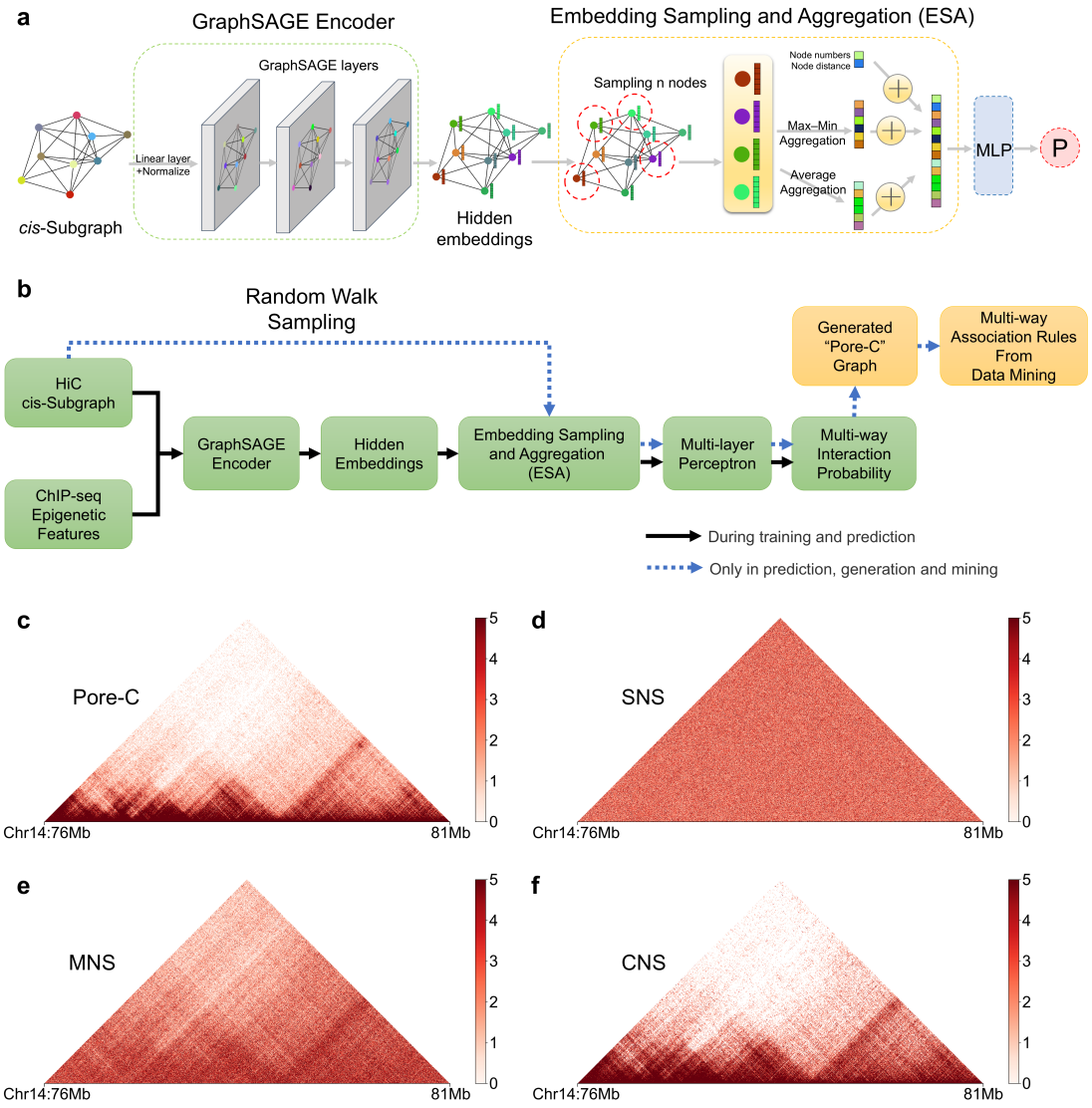

Figure 1. Architecture of PHOCI framework and comparison of three negative sampling methods in GM12878 data with Pore-C Pairwise contact maps. (a) Deep learning model base on GraphSAGE Encoder module and Embedding Sampling and Aggregation module. (b) Training, prediction, generation and mining flowchart. (c) Pore-C Pairwise contact maps. (d) SNS Pairwise contact maps. (e) MNS Pairwise contact maps. (f) CNS Pairwise contact maps.

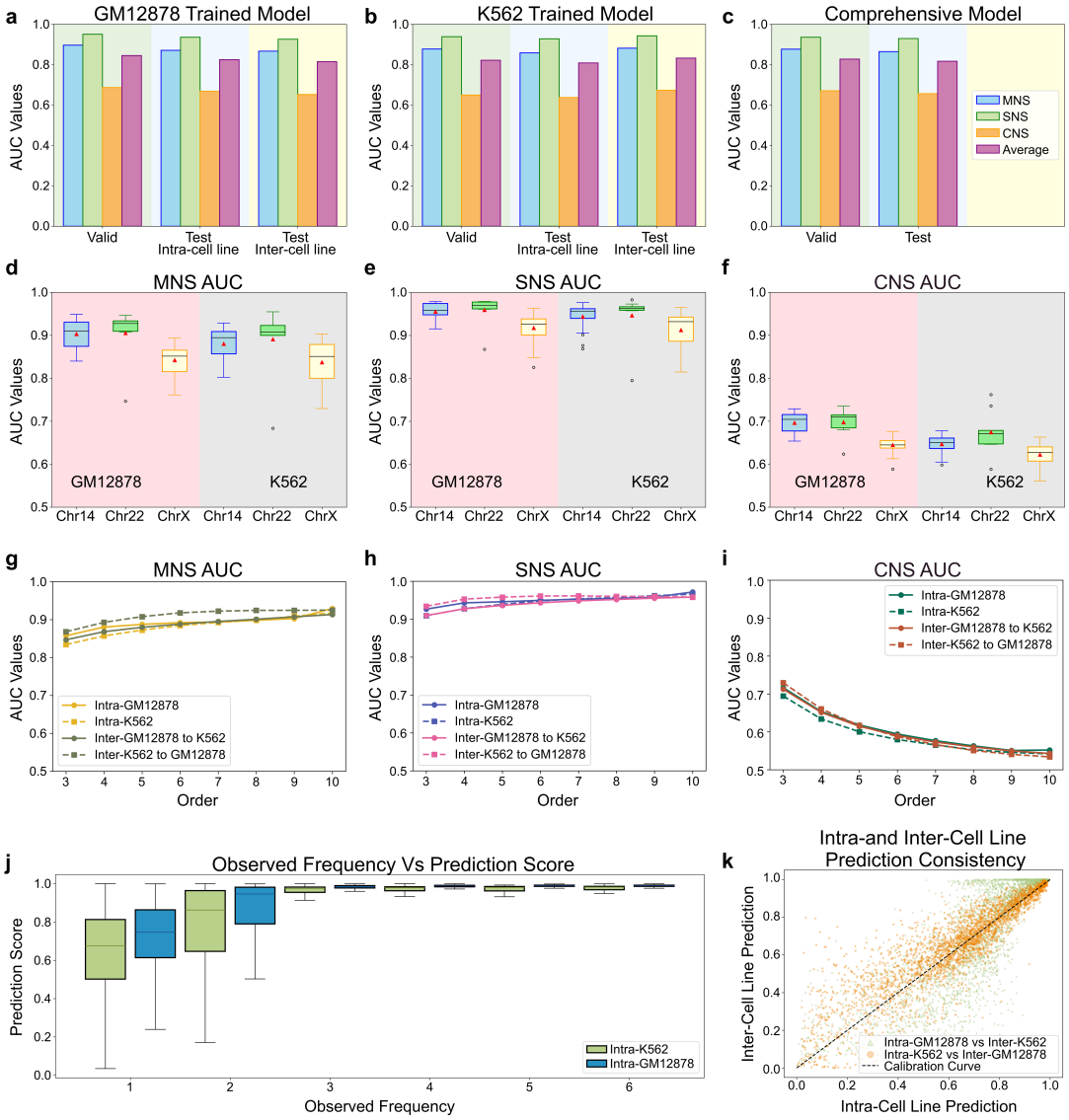

Figure 2. Performance analysis of multi-way interaction prediction models in intra-cell and inter-cell lines. (a-c) AUC scores on validation and test sets (including intra- and inter-cell line) for models trained on different cell lines; (d-f) The AUC of different chromosomes in intra-cell line testing (box plot) (Chr14, chr22, chrX); (g-i) The AUC scores under different orders; (j) Relationship between multi-way interaction occurrence frequency and prediction scores of the intra-cell line model (each box containing >2000 samples); (k) Consistency analysis of prediction scores between intra-cell line and inter-cell line models (randomly selected approximately 2900 multi-way interaction samples).

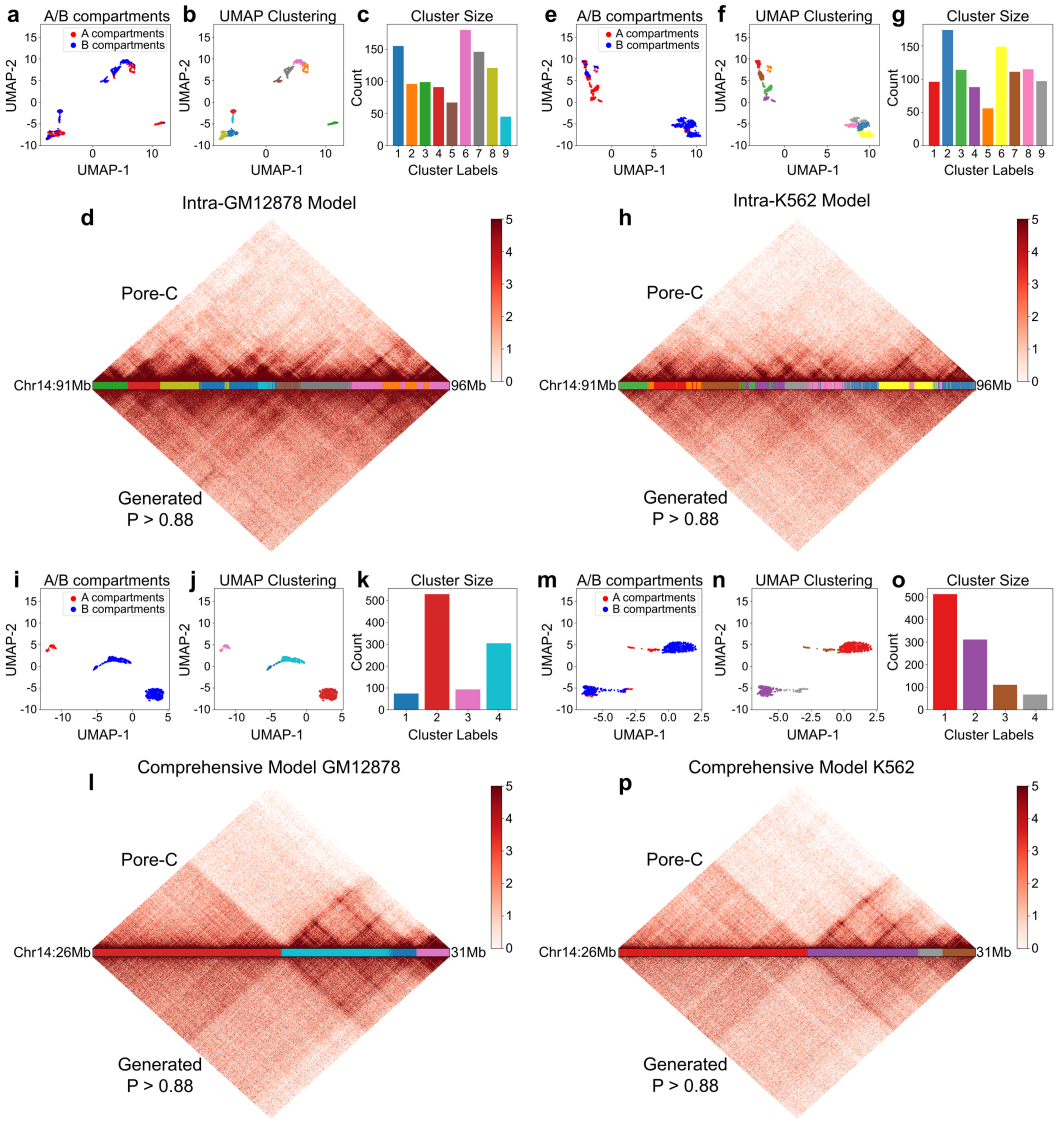

Figure 3. Analysis of hidden embeddings from intra-cell line models and comprehensive model of GM12878 and K562 cell lines. (a, e, i, m) Distribution of A/B compartments. (b, f, j, n) Distribution of clusters after k-means clustering. (c, g, k, o) Node count statistics for each corresponding cluster. (d, h, l, p) Correspondence between cluster positions and TADs of Pore-C pairwise contact maps and the corresponding generated versions.

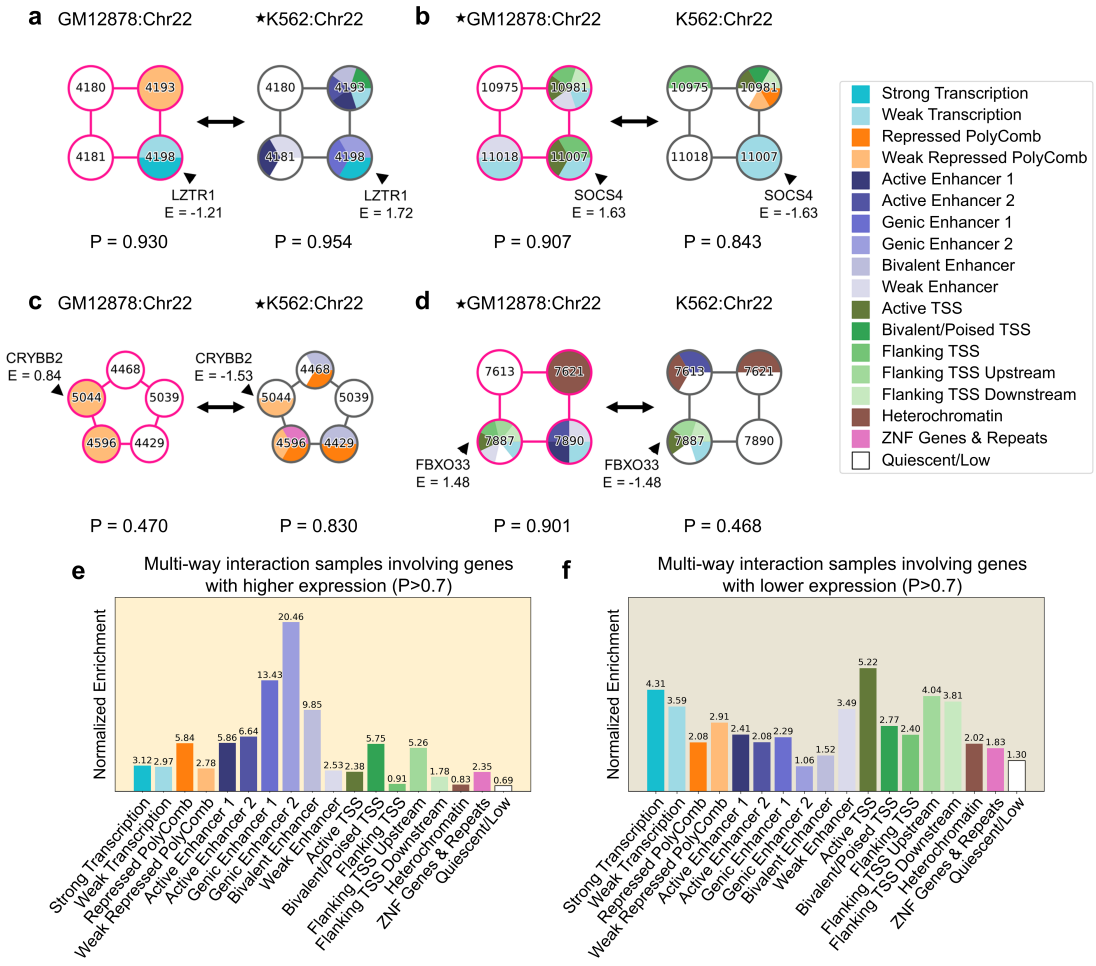

Figure 4. Examples of predictive system scores and analysis of gene expression and chromatin states. (a-b) Gene analysis and predictive system score examples. Asterisks indicate experimental measurement data. Reasonable examples with high scores provided by the comprehensive model. (c-d) Gene analysis and the predictive system's examples of high and low scores. Asterisks indicate experimental measurement data. Reasonable examples with high scores and unreasonable examples with low scores provided by the comprehensive model. (e) Normalized enrichment of 18 chromatin states of higher expression samples with different prediction scores (Chr22). (f) Normalized enrichment of 18 chromatin states of lower expression sample groups with different prediction scores (Chr22).

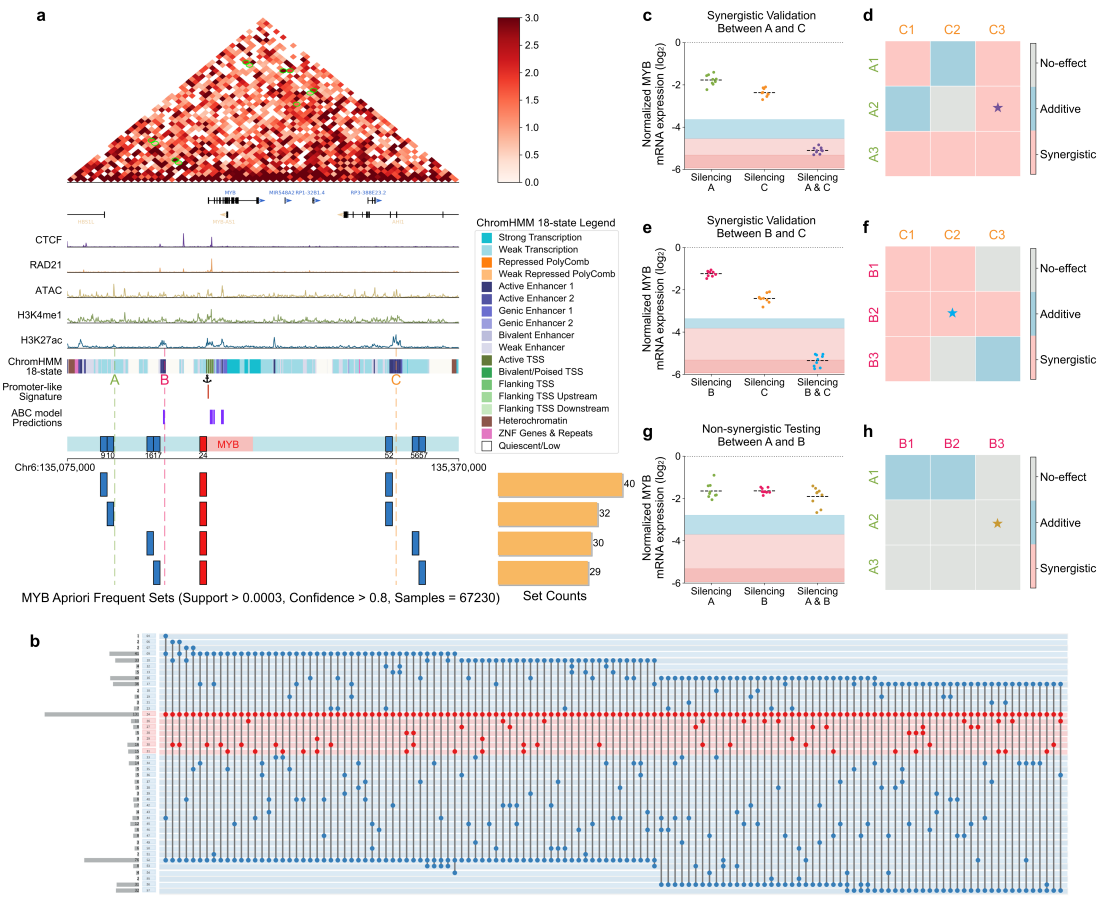

Figure 5. Multi-way association rules of the MYB gene and corresponding experimental validation.

(a) Multi-way association rules derived from the Apriori algorithm, incorporating chromatin features such as CTCF, RAD21, ATAC, H3K4me1, and H3K27ac signals. The Hi-C contact map at the top illustrates interaction frequencies, with yellow circles marking interactions between associated loci. ChromHMM 18-state annotations highlight various chromatin states, including active enhancers and repressed regions. The bar graph at the bottom displays the frequency distribution of the identified association rules based on support and confidence scores. (b) All involved loci (131 in total) from the multi-way interactions corresponding to the association rules identified in panel (a). (c) Synergistic experimental validation results between loci A and C. The pink regions indicate synergistic area, with dark pink marking the area for direct repression of the MYB transcription start site (TSS), while blue area represent additive effects. (d) Replicate experimental results for different sgRNAs targeting loci A and C. (e) Synergistic experimental validation results between loci B and C. The pink regions indicate synergistic area, with dark pink marking the area for direct repression of the MYB transcription start site (TSS), while blue area represent additive effects. (f) Replicate experimental results for different sgRNAs targeting loci B and C. (g) Non-synergistic experimental results between loci A and B. The pink regions indicate synergistic area, with dark pink marking the area for direct repression of the MYB transcription start site (TSS), while blue area represent additive effects. (h) Replicate experimental results for different sgRNAs targeting loci A and B.