Multimodal 3D Genome Pre-training

Minghao Yang[1], Pengteng Li[1], Yan Liang[2], Qianyi Cai[1], Zhihang Zheng[3], Shichen Zhang[3], Pengfei Zhang[1], Zhi-An Huang[4,*], and Hui Xiong[1,5,*]

[1]Thrust of Artificial Intelligence, The Hong Kong University of Science and Technology (Guangzhou), China
[2]School of Artificial Intelligence, South China Normal University, China
[3]Thrust of Bioscience and Biomedical Engineering, The Hong Kong University of Science and Technology (Guangzhou), China
[4]Department of Computer Science, City University of Hong Kong (Dongguan), China
[5]Department of Computer Science and Engineering, The Hong Kong University of Science and Technology Hong Kong SAR, China

Abstract

Deep learning techniques have driven significant progress in various analytical tasks within 3D genomics in computational biology. However, a holistic understanding of 3D genomics knowledge remains underexplored. Here, we propose ***MIX-HIC***, the first multimodal foundation model of 3D genome that integrates both Hi-C contact maps and epigenomic tracks, which obtains unified and comprehensive semantics. For accurate heterogeneous semantic fusion, we design the cross-modal interaction and mapping blocks for robust unified representation, yielding the accurate aggregation of 3D genome knowledge. Besides, we introduce the first large-scale dataset comprising over ***1 million*** pairwise samples of Hi-C contact maps and epigenomic tracks for high-quality pre-training, enabling the exploration of functional implications in 3D genomics. Extensive experiments show that MIX-HIC significantly surpasses existing state-of-the-art methods in diverse downstream tasks. This work provides a valuable resource for advancing 3D genomics research.

1 Introduction

The three-dimensional (3D) organization of chromosomes within the nucleus plays a pivotal role in gene regulation and cellular function [1, 2]. Key topological features of the 3D genome, such as chromatin loops that bring distant regulatory elements into close physical proximity with their target genes, are essential for cell-type-specific transcriptional regulation. High-resolution 3D chromatin interactions can be quantified through high-throughput chromosome conformation capture (Hi-C) technique [3]. Understanding the mechanisms of how the 3D genome influences gene expression can unveil pivotal insights into cellular functionality, developmental biology, and disease mechanisms [4].

Recently, computational models have emerged as a powerful tool to unravel the intricate associations between the 3D chromatin structure, epigenome, and transcriptome. Existing approaches predict various genomic features, including 3D chromatin contact maps [5, 6, 7], chromatin loops [8, 9, 10], and gene expression [11, 12], often leveraging DNA sequences, Hi-C contact maps and epigenomic tracks. Though successful, most of these methods are limited to a single specific task and struggle to integrate the diverse and heterogeneous information of the 3D genome, hindering a comprehensive understanding of its complex organization.

Recent progress in large-scale foundation models has demonstrated remarkable success in various fields of computational biology, such as molecular representations [13, 14], medical imaging [15, 16], proteomics [17, 18], and genomics [19, 20]. Inspired by these advancements, we aim to develop a

*Correspondence to Zhi-An Huang <huang.za@cityu-dg.edu.cn>, Hui Xiong <xionghui@ust.hk>.
The source code of MIX-HIC is available at `https://github.com/myang998/MIX-HIC`.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

multimodal foundation model to address the above-mentioned limitations of analyzing 3D genomic downstream tasks in isolation.

Modeling the multimodal foundation model of 3D genome introduces three key challenges. First, Hi-C contact maps and epigenomic tracks have inherently distinct characteristics, making integration difficult. Simply aligning features from the two modalities and projecting them into a unified latent space would primarily capture modal-invariant knowledge like gene regulatory mechanisms, which are governed by both chromatin spatial organization and epigenomic tracks. However, this approach tends to overlook modal-specific characteristics, such as precise chemical modifications and chromatin states revealed by epigenomic tracks, which are essential factors for fine-grained 3D genome analysis. This can lead to information loss and degrade downstream task performance (*Refer to Appendix B for theoretical analysis*). Second, the unified representation from heterogeneous 3D genomic and epigenomic data must exhibit robust generalization capabilities; otherwise, the model may struggle to adapt effectively to diverse downstream tasks *e.g.*, generation and regression. Third, uncovering implicit semantic relationships between 3D genomic and epigenomic data is crucial for addressing the data scarcity problem in 3D genomics. Especially, the high experimental costs of Hi-C sequencing in real-world applications limit data accessibility, leading to incomplete representations and degraded model performance. Pre-training a model to learn implicit multimodal structural connections offers significant benefits for downstream tasks, particularly when only single-modality data is available. This enables the model to leverage structural knowledge to compensate for missing modality semantics, thereby enriching the overall data representation.

Hence, we introduce MIX-HIC, the first multimodal foundation model for 3D genomics to extract fine-grained knowledge from 3D genome and epigenomic tracks, enabling efficient adaptation to diverse downstream tasks with superior performance. Specifically, MIX-HIC first incorporates two distinct encoders to capture the refined features from 3D genome contact maps and epigenomic tracks, leveraging cell type-specific information for accurate predictions in novel cell types. To address the challenge of integrating heterogeneous data, we propose a cross-modal interaction block to capture both modal-invariant and modal-specific representations, regularized by contrastive learning and orthogonal constraints, preserving both shared and distinctive information across modalities. Additionally, a cross-modal mapping block facilitates information exchange between modalities, ensuring robust representation even with single-modality input. To comprehensively capture 3D genome knowledge, we have curated a large-scale dataset that consists of 1,275,948 pair samples of Hi-C contact maps and epigenomic tracks for rapid adaptation to downstream tasks through task-specific decoders. Notably, this is the largest paired dataset for the 3D genome analysis to date. MIX-HIC is evaluated across diverse downstream tasks, demonstrating its effectiveness and robustness in comparison to other state-of-the-art methods. In summary, the main contributions are:

- We propose ***the first 3D genomic multimodal foundation model***, integrating Hi-C contact maps and epigenomic tracks to establish a new paradigm for 3D genome analysis.
- MIX-HIC features a novel architecture with two key components: (1) a cross-modal interaction block to capture both shared and unique biological patterns across modalities; and (2) a cross-modal mapping block to enable a reliable complement of missing modality features.
- For holistic representation learning in 3D genome analysis, we present ***the largest paired dataset of Hi-C and epigenomic tracks***, comprising over 1 million samples.
- Extensive experiments demonstrate that MIX-HIC achieves state-of-the-art performance on three critical downstream tasks across two cell lines.

2 Related Works

2.1 3D Genome-Related Tasks

This work evaluates the effectiveness of MIX-HIC on three downstream tasks, including Hi-C contact map prediction, chromatin loop detection, and CAGE-seq expression prediction.

Existing methods for Hi-C contact map prediction from DNA sequences show promise but lack cross-cell-type generalization [7, 21]. Models like EPCOT [11] and C.Origami [6] improve this by integrating DNA sequence with cell-type-specific epigenomic tracks. Epiphany [5] offers a more efficient solution using only the epigenomic tracks.

Chromatin loop detection methods are broadly divided into statistical and supervised learning methods. Statistical methods like HiCExplorer [22] and ChromoSight [9] rely on contact frequency distributions or expert-defined templates to identify loops. Supervised learning methods, including Peakachu [8], DLoopCaller [10] leverage labeled data and sophisticated architectures for loop detection. RefHiC [23] employs a coarse-to-fine training strategy that integrates multi-resolution Hi-C contact maps through contrastive learning mechanisms for model pre-training. However, RefHiC is limited by its reliance on small-scale data semantics, which hinders its generalizability to other downstream tasks.

CAGE-seq expression prediction typically uses multimodal inputs, including DNA sequences, epigenomic tracks, and Hi-C contact maps. Enformer [24] employs transformers for modeling DNA sequences, while EPCOT [11] integrates DNA sequences and epigenomic tracks with transformers or long short-term memory (LSTM) [25]. GraphReg [12] integrates epigenomic tracks and Hi-C contact maps using graph attention networks for expression prediction.

Despite their notable achievements, these methods often exhibit limitations in knowledge transfer across tasks and fail to fully capture the complex interaction patterns between multimodal data.

2.2 Foundation Models in Computational Biology

The emergence of foundation models has significantly advanced various fields of computational biology. For example, EvoRank [26] has demonstrated remarkable capabilities by harnessing extensive protein sequence datasets to derive latent representations through sequence alignment. VQDNA [19] employs large-scale DNA sequences to learn adaptive tokenization through vector-quantized codebooks. To capture more comprehensive views of data, multimodal foundation models such as ESM-IF [27] combine protein sequences and 3D structures to learn functional and structural representations, advancing protein function design and prediction. Similarly, UniCorn [28] integrates 2D and 3D molecular views through contrastive learning, effectively capturing complementary bimodal features. These foundation models have showcased strong representation learning capabilities of foundation models, inspiring the development of a universal model capable of effectively addressing various downstream tasks in the 3D genomic field.

3 Data Processing and Preparation

To distill the comprehensive semantics from the 3D genome, we collect and refine a large-scale dataset for pre-training MIX-HIC, using publicly available data from the hg38 assembly. The Hi-C contact maps are obtained from the 4DN Data Portal[1], while the epigenomic tracks (ATAC-seq and DNase-seq, which measure how 'open' or accessible DNA is for transcription), CAGE-seq expression data (which directly quantifies gene activity levels), and CTCF ChIA-PET [29] chromatin loops (which identify high-confidence interactions mediated by the key architectural protein CTCF) are downloaded from the ENCODE Portal [2]. Due to the high cost of deep sequencing for Hi-C experiments [30], publicly available datasets are typically limited to resolutions of 5 kb, 10 kb, or coarser. A 5 kb resolution is a fine-grained and effective choice for deep learning models [11, 12]. The MIX-HIC model processes 250,000 base pair (bp) genomic windows, a size selected to encompass key regulatory structures like chromatin loops, ensuring most functional units are fully contained within the inputs [5, 8]. This results in Hi-C contact maps being represented as 50×50 matrices at the 5 kb resolution. To reduce data variability, epigenomic tracks are averaged over every 100 bps, which generates 2,500-length sequences. Each bin in the Hi-C matrix corresponds to specific x and y coordinates, representing two distinct genomic segments. Their respective epigenomic sequences are concatenated into a 5,000-length representation for that bin. Thus, Hi-C contact maps and epigenomic tracks are treated as images and sequences, respectively. We focus on four cell lines for pre-training, including HepG2, HCT116, IMR90, and WTC111. To ensure data quality, we filter out windows with fewer than 10% non-zero Hi-C interactions or insufficient contact signals, which is a standard quality control step to remove uninformative data [8, 10]. Hi-C contact maps are inherently sparse, especially for

Table 1: Summary of pre-training data for MIX-HIC, including original and cleaned data counts.

Cell line	Original	Cleaned	Remain
HepG2	48, 415	38, 536	79.5%
HCT116	276, 384	189, 099	68.4%
IMR90	471, 557	316, 794	67.2%
WTC11	1, 032, 048	731, 519	70.9%
Total	**1, 828, 401**	**1, 275, 948**	69.8%

[1] `https://data.4dnucleome.org/`

[2] `https://www.encodeproject.org/`

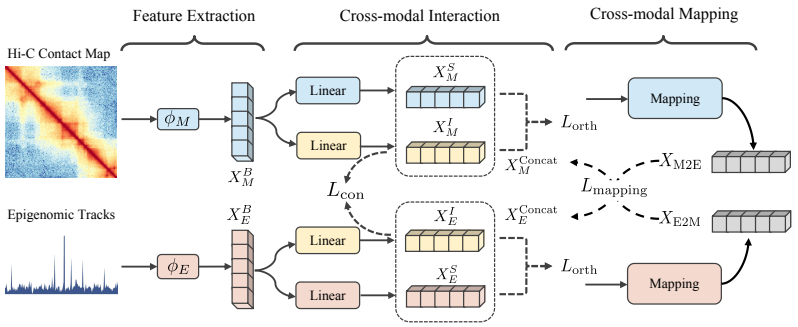

Figure 1: **Pre-training stage of MIX-HIC.** MIX-HIC employs a dual-encoder architecture to extract refined features from both Hi-C contact maps and epigenomic tracks. The modal-specific and modal-invariant representations are learned via contrastive learning and orthogonal constraints within the cross-modal interaction block. A cross-modal mapping block is developed to further regularize the bimodal representations and facilitate cross-modal complement. ϕ_M and ϕ_E denote the Hi-C contact map encoder and epigenomic track encoder, respectively.

long-range interactions distant from the diagonal. Including these extremely sparse, low-signal windows would introduce noise and degrade the model's training process. As summarized in Table 1, this filtering process removed approximately 30% of the raw windows. Crucially, a massive and high-quality dataset of over 1.2 million sample pairs was retained, which is more than sufficient for robust pre-training. *Further details on data processing, reference number, and downstream task data are provided in Appendix A.1.*

4 Methodology

4.1 Self-supervised Pre-training

Self-supervised learning [31, 32, 33] offers a powerful solution for learning unified and comprehensive representations from heterogeneous 3D genomic and epigenomic data. By leveraging large-scale pairwise data during pre-training, MIX-HIC effectively captures inherent biological patterns and relationships across these bimodal data, significantly enhancing its adaptability to diverse downstream tasks. As shown in Figure 1, the pre-training phase includes a feature extraction block, cross-modal interaction block, and cross-modal mapping block. The feature extraction block employs specialized encoders for epigenomic tracks and Hi-C contact maps, respectively, to generate refined representations. The learned representations are then fed into the cross-modal interaction block to learn both modal-specific and modal-invariant features. This learning process is regularized through a combination of contrastive learning loss and orthogonal loss. Finally, a cross-modal mapping block further explores latent connections and complementary information between the two modalities.

Feature extraction block. The feature extraction block consists of Hi-C and epigenomic feature encoders. Given the Transformer's capability to model long-range dependencies and its flexibility in capturing both spatial interactions (Hi-C contact maps) and sequential relationships (epigenomic tracks), we utilize a Transformer-based architecture for robust feature extraction and multimodal integration. Similar to Vision Transformer (ViT) [34], the single-channel Hi-C contact map $X_M \in \mathbb{R}^{1\times H\times W}$ is first transformed into a sequence of flattened patches $X_M^p \in \mathbb{R}^{N\times D}$. Here, H and W are the height and width of the Hi-C contact map, both equal to 50. The dimension $D = 1 \times P \times P$ corresponds to the initial size of each patch, where the patch size P is set to 2. The total number of these patches $N = H \times W/P^2$ is the resulting number of patches, which becomes the input sequence length. In the Hi-C feature encoder, a feedforward network projects these patches into an embedding $X_M^0 \in \mathbb{R}^{N\times C}$, where C denotes the predefined feature dimension. This embedding is then refined through three cascaded encoder layers, where each layer consists of T Transformer blocks followed by a downsampling layer, yielding progressively refined embeddings $X_M^i \in \mathbb{R}^{\alpha_i\times C_i}$, where the sequence length α_i is defined as $\alpha_i = \frac{N}{4^i}$ and the feature dimension C_i grows exponentially as $C_i = 2^i C$, with the encoder layer $i \in \{1, 2, 3\}$. Finally, a bottleneck layer equipped with T Transformer blocks is applied to produce the Hi-C contact map embedding $X_M^B \in \mathbb{R}^{\alpha_3\times C_3}$.

The epigenomic encoder processes input sequences $X_E \in \mathbb{R}^{L_1\times O}$, where $L_1 = 5,000$ is the initial sequence length, and O represents the two epigenomic tracks: ATAC-seq and DNase-seq.

Convolutional layers with max-pooling operations extract an initial embedding $X_E^0 \in \mathbb{R}^{L_2 \times C}$, where $L_2 = 100$ corresponds to two genomic segments along both x and y axes in the Hi-C contact map. Note that processing epigenomic tracks at high resolution (*e.g.*, 100bp) is common practice to avoid over-smoothing [5, 11]. Similarly, the embedding X_E^0 is processed through three encoder layers, where each layer consists of T Transformer blocks followed by a downsampling layer, resulting in $X_E^i \in \mathbb{R}^{\beta_i \times C_i}$, where the sequence length $\beta_i = \frac{L_2}{2^i}$ for $i \in \{1, 2, 3\}$. Then, the refined epigenomic representation $X_E^{\mathrm{B}} \in \mathbb{R}^{\beta_3 \times C_3}$ is also derived from a final bottleneck layer.

Cross-modal interaction block. Conventional multimodal learning architectures [35] often utilize contrastive learning to project features from different modalities into a shared embedding space. However, these modalities inherently contain both homogeneous and heterogeneous information. Direct aligning all features risks losing essential modal-specific characteristics, potentially diminishing downstream task performance [36, 37]. This is further analyzed in Theorem 1 as follows.

Theorem 1. *Let ϕ_1 and ϕ_2 be feature encoders for two modalities z_1 and z_2, respectively. If the encoded features $\mathbf{F}^1 = \phi_1(z_1)$ and $\mathbf{F}^2 = \phi_2(z_2)$ are perfectly aligned such that $\mathbf{F}^1 = \mathbf{F}^2$, we have:*

$$\inf_h \mathbb{E}_q[\mathcal{L}_{\text{CE}}(h(\mathbf{F}^1, \mathbf{F}^2), t)] - \inf_{h'} \mathbb{E}_q[\mathcal{L}_{\text{CE}}(h'(z_1, z_2), t)] \geq \Gamma_q. \tag{1}$$

Remarks. The information gap $\Gamma_q := \max\{U(z_1;t), U(z_2;t)\} - \min\{U(z_1;t), U(z_2;t)\}$ quantifies the effectiveness of the modalities in predicting target variable t, where $U(z_j;t)$ represents the mutual information. Here, $\mathbb{E}$ represents the expectation, q denotes the joint distribution of (z_1, z_2, t), $\mathcal{L}_{\text{CE}}$ is the cross-entropy loss, and h and h' are prediction functions for features and raw data, respectively. Theorem 1 demonstrates that perfect alignment results in prediction errors that are suboptimal by at least Γ_q compared to using raw modalities directly. This information gap widens when information content is imbalanced across modalities. The complete proof is provided in Appendix B.

To address this, our cross-modal interaction block captures both modal-specific and modal-invariant representations, enabling a more comprehensive understanding of the data. Specifically, we employ four independent dense networks to process the Hi-C contact map representation X_M^B and the epigenomic representation X_E^B. This generates the condensed modal-invariant representations $X_M^I \in \mathbb{R}^{\alpha_3 \times C_2}$ and $X_E^I \in \mathbb{R}^{\beta_3 \times C_2}$, as well as the modal-specific representations $X_M^S \in \mathbb{R}^{\alpha_3 \times C_2}$ and $X_E^S \in \mathbb{R}^{\beta_3 \times C_2}$. The mean pooling operation is then applied along the sequence length dimension, producing $\hat{X_M^I}$, $\hat{X_E^I}$, $\hat{X_M^S}$, and $\hat{X_E^S}$, each with a feature dimension of C_2. To ensure that the modal-invariant features capture shared knowledge across modalities, we incorporate a contrastive learning loss to regularize these representations. We propose a unified contrastive loss function $\mathcal{L}_{\text{pair}}$ [31] to compute the similarity between two modalities $\mathcal{A}$ and $\mathcal{B}$ as follows:

$$\mathcal{L}_{\text{pair}}(\mathcal{A}, \mathcal{B}) = -\frac{1}{J} \sum_{j=1}^{J} \log \frac{\exp\langle \mathcal{A}_j, \mathcal{B}_j \rangle / \tau}{\sum_{r=1}^{J} \exp\langle \mathcal{A}_j, \mathcal{B}_r \rangle / \tau}, \tag{2}$$

where τ denotes the temperature, commonly set to 0.07 [38], J refers to the batch size,$\langle \cdot, \cdot \rangle$ is the dot product operation, and $\mathcal{A}_j$ and $\mathcal{B}_j$ represent the embedding of the j-th sample in the mini-batch for modality $\mathcal{A}$ and $\mathcal{B}$, respectively. Finally, the overall contrastive loss is computed as follows:

$$\mathcal{L}_{\text{con}} = \frac{1}{2}(\mathcal{L}_{\text{pair}}(\hat{X_E^I}, \hat{X_M^I}) + \mathcal{L}_{\text{pair}}(\hat{X_M^I}, \hat{X_E^I})). \tag{3}$$

Moreover, we introduce an orthogonal constraint to maximize the dissimilarity between modal-specific and modal-invariant features. This ensures that the modal-specific features capture complementary information distinct from the shared knowledge represented by the modal-invariant features. The orthogonal loss $\mathcal{L}_{\text{orth}}$ is calculated by minimizing the inner product between these features:

$$\mathcal{L}_{\text{orth}} = \frac{1}{2}\left(\langle \hat{X_M^S}, \hat{X_M^I} \rangle + \langle \hat{X_E^S}, \hat{X_E^I} \rangle\right). \tag{4}$$

Cross-modal mapping block. Data scarcity, often due to high experimental costs, can lead to incomplete datasets with missing modalities. Integrating predicted features of a missing modality with existing modalities can enhance prediction performance [37, 39]. Our cross-modal mapping block aims to capture the implicit semantic relationships and facilitate knowledge transfer between modalities to address this issue.

Some downstream tasks require the MIX-HIC to preserve the sequence length of the input features. However, the differing lengths of Hi-C contact maps and epigenomic tracks pose a challenge for effective modality transfer. Therefore, we apply 1D adaptive pooling to the concatenated representations of Hi-C, $X_M^{\text{Concat}} = [X_M^I : X_M^S] \in \mathbb{R}^{\alpha_3 \times C_3}$ and epigenomic tracks, $X_E^{\text{Concat}} = [X_E^I : X_E^S] \in \mathbb{R}^{\beta_3 \times C_3}$ (where $[\cdot : \cdot]$ denotes concatenation), to align their lengths. This yields the complementary representations $X_{\text{M2E}} \in \mathbb{R}^{\beta_3 \times C_3}$ and $X_{\text{E2M}} \in \mathbb{R}^{\alpha_3 \times C_3}$ as follows:

$$X_{\text{M2E}} = \mathcal{F}_{\text{M2E}}(\mathcal{G}_{\text{M2E}}(X_M^{\text{Concat}})), X_{\text{E2M}} = \mathcal{F}_{\text{E2M}}(\mathcal{G}_{\text{E2M}}(X_E^{\text{Concat}})), \tag{5}$$

where $\mathcal{G}_{\text{M2E}}$ and $\mathcal{G}_{\text{E2M}}$ are 1D adaptive pooling operations. $\mathcal{F}_{\text{M2E}}$ and $\mathcal{F}_{\text{E2M}}$ denote dense layers with the same output dimensions matching their inputs. To ensure these mapped embeddings capture the relevant information from the target modality, we use a $\mathcal{L}_{\text{mapping}}$ loss for regularization:

$$\mathcal{L}_{\text{mapping}} = \frac{1}{2}(\left\|X_{\text{M2E}} - X_E^{\text{Concat}}\right\|_2^2 + \left\|X_{\text{E2M}} - X_M^{\text{Concat}}\right\|_2^2). \tag{6}$$

Overall, the final loss for self-supervised pre-training is computed as follows:

$$\mathcal{L}_{\text{pretrain}} = \mathcal{L}_{\text{con}} + \mathcal{L}_{\text{orth}} + \mathcal{L}_{\text{mapping}}. \tag{7}$$

4.2 Task-specific Fine-tuning

As depicted in Figure 2, MIX-HIC is a versatile framework capable of processing various kinds of inputs. Ideally, MIX-HIC takes both the Hi-C contact map and the epigenomic feature profile as inputs (represented as MIX-HIC-Bimodal), extracting their concatenated features X_M^{Concat} and X_E^{Concat} similar to the feature extraction block used in the self-supervised pre-training stage. However, in real-world scenarios, a certain modality may be absent for various unforeseen reasons.

Leveraging the powerful representation ability of pre-training to capture implicit connections between bimodal data, MIX-HIC incorporates a cross-modal mapping block to complement the features of the missing modality using the information from the available modality. For example, when the Hi-C contact map is missing, MIX-HIC can infer the missing modality features X_{E2M} from the concatenated epigenomic embedding X_E^{Concat} using Eq. 5. For the Hi-C contact map prediction task, even if only the epigenomic tracks are available, the corresponding contact map features X_{E2M} can still be utilized for prediction (denoted as MIX-HIC-Infer).

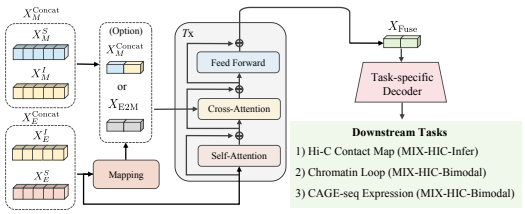

Figure 2: **Fine-tuning stage of MIX-HIC.** Epigenomic features are captured from the pre-trained encoder, while Hi-C contact map features are obtained either directly from the pre-trained encoder or through feature mapping based on the epigenomic features. MIX-HIC incorporates a modality fusion block to integrate the bimodal representations, followed by a task-specific decoder for final predictions of downstream tasks.

Modality-fusion block. We employ T stacked contact map-grounded fusion blocks to learn the interaction patterns between the bimodal representations. Each contact map-grounded fusion block consists of a self-attention layer, a cross-attention layer, and a feedforward network. The extracted epigenomic embeddings X_E^{Concat} are first input into the self-attention layer, generating query embeddings of the cross-attention layer. The contact map embeddings, either X_M^{Concat} or X_{E2M}, serve as the key and value embeddings, which are then processed through the cross-attention layer. Finally, a feedforward neural network is applied to produce the fusion embeddings $X_{\text{Fuse}} \in \mathbb{R}^{\beta_3 \times C_3}$.

Task-specific prediction block. Three types of decoders are involved in the task-specific prediction block. The chromatin loop detection task is a binary classification problem. Therefore, the decoder for this task includes a mean pooling operation along the sequence length axis, followed by a feedforward network, to classify a given sample as either a chromatin loop or a non-loop.

The CAGE-seq expression task is formulated as a regression problem, aiming to predict 100 values corresponding to the expression levels of two genomic segments (each of length 50) along x-axis and y-axis of the Hi-C contact map. The decoder consists of three transformer blocks, with each followed by an upsampling layer. At each stage, the encoder-derived features X_E^i from the i-th encoder layer are concatenated with the corresponding i-th layer of decoder outputs X_D^i from the

preceding stage via skip connections, facilitating feature integration across network depths. Finally, the decoder output features $X_{\text{out}} \in \mathbb{R}^{L_2 \times 2C}$ are processed through two feedforward networks to generate regression predictions.

To predict the 50×50 Hi-C contact maps at 5kb resolution, the decoder output features X_{out} are obtained similar to the CAGE-seq expression task. These features are then split into two feature segments $X^1_{\text{out}} \in \mathbb{R}^{50\times 2C}$ and $X^2_{\text{out}} \in \mathbb{R}^{50\times 2C}$ along the length axis. X^1_{out} and X^2_{out} are unsqueezed into $\mathbb{R}^{50\times 1\times 2C}$ and $\mathbb{R}^{1\times 50\times 2C}$, respectively, after which element-wise addition and multiplication operations are applied to generate feature maps of size $\mathbb{R}^{50\times 50\times 2C}$. Finally, two layers of feedforward networks are utilized to predict the Hi-C contact maps.

Training loss. The chromatin loop detection task employs the binary cross entropy (BCE) loss function, while the Hi-C contact map prediction and CAGE-seq expression prediction tasks use the mean squared error (MSE) loss function. It should be noted that we normalize CAGE-seq data using RPGC [40] and Hi-C data using KR [41] normalization to correct for sequencing depth and systematic biases, respectively. Subsequently, a log transformation is applied to both datasets. This is critical for stabilizing variance and compressing the highly skewed distribution of raw counts, ultimately producing continuous values representing normalized interaction frequencies. Consequently, the MSE loss becomes an appropriate and robust choice for these continuous values. The MSE and BSE loss functions can be formulated as follows:

$$\mathcal{L}_{\text{MSE}} = \frac{1}{J}\sum_{j=1}^{J}(y_j - \hat{y}_j)^2, \tag{8}$$

$$\mathcal{L}_{\text{BCE}} = -\frac{1}{J}\sum_{j=1}^{J}\left[y_j \log(\hat{y}_j) + (1-y_j)\log(1-\hat{y}_j)\right], \tag{9}$$

where y_i and $\hat{y}_i$ represent the true value and the predicted value, respectively.

5 Experiments

Implementation details, hyperparameter analysis, and analyses of the model's biological grounding and robustness for noise can be referred to Appendix.

5.1 Comparison Results on Downstream Tasks

To demonstrate the effectiveness of MIX-HIC, three downstream tasks are involved in this work, including Hi-C contact map prediction, chromatin loop detection, and CAGE-seq expression prediction. We construct three versions of MIX-HIC: (1) MIX-HIC-Bimodal, which leverages both Hi-C contact maps and epigenomic tracks through pre-training; (2) MIX-HIC-NonPre, a non-pretrained version using the same bimodal inputs; and (3) MIX-HIC-Infer, designed to handle missing Hi-C data by integrating epigenomic track embeddings with inferred Hi-C embeddings. *Detailed descriptions of the compared state-of-the-art methods are given in Appendix D.*

3D Chromatin Organization Prediction. In this task, the Hi-C contact maps serve as the target. MIX-HIC-Infer is compared with four state-of-the-art methods, including Epiphany [5], C.Origami [6], and two variants of EPCOT [11], (*i.e.* EPCOT-LSTM and EPCOT-Transformer). For a fair comparison, all methods receive the same input as MIX-HIC and are configured with their default parameter settings. The coefficient of determination (R^2) [42] is employed as the evaluation metric, since it effectively quantifies explained variance in analyses of long-tailed or sparse data, such as low-probability long-range interactions in Hi-C.

Table 2: Methods comparison for the Hi-C contact map prediction task on GM12878 and K562 cell lines using R^2. The results marked in **bold** and underlined denote the best and second-best performing methods, respectively.

Methods	GM12878	K562
Epiphany [5]	0.7970	0.6547
C.Origami [6]	0.7958	0.7055
EPCOT-Transformer [11]	0.5409	0.7648
EPCOT-LSTM [11]	<u>0.7993</u>	<u>0.7840</u>
MIX-HIC-Infer (Ours)	**0.8724**$_{+9.3\%}$	**0.8001**$_{+2.1\%}$

Table 2 shows the evaluation performance on GM12878 and K562 cell lines. Compared to other methods, MIX-HIC-Infer demonstrates superior performance, achieving the highest average R^2 values on both GM12878 and K562 cell lines. Specifically, it outperforms the runner-up method by approximately 9.3% and 2.1% in R^2 score on GM12878 and K562, respectively. We note that the EPCOT variants deliver competitive performance, but they suffer from fluctuations across the two datasets. Through extensive pre-training on large-scale pairwise datasets, MIX-HIC effectively

Table 3: Methods comparison for the supervised chromatin loop detection task on GM12878 and K562 cell lines.

Methods	GM12878				K562			
	Precision	Recall	F1	AUROC	Precision	Recall	F1	AUROC
Peakachu [8]	0.7763	0.8283	0.8015	0.8766	0.7895	0.7905	0.7900	0.8834
DLoopCaller [10]	0.8433	0.8075	0.8250	0.9046	0.8383	0.7526	0.7932	0.8924
MIX-HIC-Bimodal (Ours)	**0.8505**$_{+0.9\%}$	**0.8337**$_{+0.7\%}$	**0.8420**$_{+2.1\%}$	**0.9209**$_{+1.8\%}$	**0.8521**$_{+1.6\%}$	**0.8027**$_{+1.5\%}$	**0.8267**$_{+3.9\%}$	**0.9194**$_{+3.0\%}$

explores implicit semantic relationships between bimodal data, enabling robust compensation for missing modality semantics and achieving promising accuracy in Hi-C contact map prediction.

Chromatin Loop Detection. Two supervised machine learning-based methods DLoopCaller [10] and Peakachu [8] are used to compare with MIX-HIC-Bimodal. As illustrated in Table 3, MIX-HIC-Bimodal surpasses other machine learning-based methods across all classification metrics on two datasets. By integrating both epigenomic tracks and Hi-C contact maps, DLoopCaller slightly outperforms Peakachu, which relies solely on Hi-C contact maps. The enhanced performance of MIX-HIC can be attributed to the effective self-supervised pre-training, which facilitates the capture of both modal-invariant and modal-specific information from bimodal representations, in contrast to the simple concatenation employed by DLoopCaller.

We employ both statistical-based methods, namely ChromoSight [9] and HiCExplorer [22], as well as machine learning-based methods to annotate loops across entire chromosomes. The predicted loops are further compared with those loops validated by experimental data from ChIA-PET. *Method details for whole-chromosome chromatin loop annotation and corresponding results with the quantitative 'proportion' metric are available in Appendix A.2 and C.1, respectively*. Figure 3 compares the number of predicted loops versus the number of loops supported by ChIA-PET among various methods on the GM12878 dataset. Among all the methods, machine learning-based methods generally predict a higher number of loops that are validated by ChIA-PET on the GM12878 dataset, with MIX-HIC identifying the most ChIA-PET validated loops while maintaining the highest validated proportion. On the K562 dataset, which contains much less training data compared to GM12878, the efficacy of machine learning methods generally declines. Nevertheless, MIX-HIC consistently maintains the highest proportion of validated loops in comparison with other methods. *An example of whole-chromosome chromatin loop comparison is provided in Appendix C.2.*

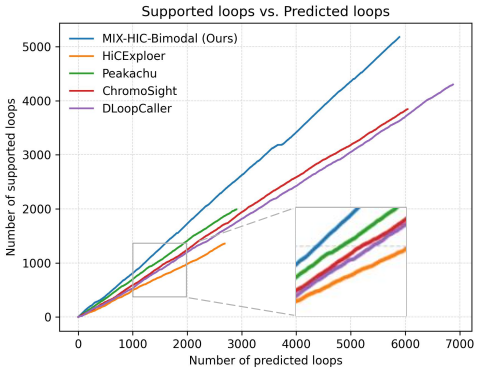

Figure 3: Comparison of the number of predicted loops with the number of corresponding ChIA-PET-supported loops across various deep learning methods on GM12878 cell line.

CAGE-seq Expression Prediction. MIX-HIC-Bimodal is pitted against four benchmark methods: two variants of GraphReg (EPI-CNN and EPI-Graph) [12], as well as two variants of EPCOT (EPCOT-LSTM and EPCOT-Transformer) [11]. EPI-CNN relies solely on epigenomic tracks to predict CAGE-seq expression, whereas EPI-Graph enhances this by incorporating Hi-C contact maps via graph attention networks for dual-modal modeling. Table 4 showcases the comparison results across two datasets. EPCOT achieves competitive performance, yet it suffers from a time-consuming problem due to the process of long-range DNA sequence. MIX-HIC demonstrates remarkable performance over other benchmark methods across all metrics on both datasets.

Table 4: Methods comparison for the CAGE-seq expression prediction task on GM12878 and K562 cell lines using R^2.

Methods	GM12878	K562
EPI-CNN [12]	0.7719	0.8033
EPI-Graph [12]	0.7965	0.8211
EPCOT-LSTM [11]	0.4723	0.8704
EPCOT-Transformer [11]	0.8578	0.8230
MIX-HIC-Bimodal (Ours)	**0.8833**$_{+3.0\%}$	**0.9077**$_{+4.3\%}$

Overall, MIX-HIC outperforms state-of-the-art methods across all downstream tasks, with the most significant improvement (9.3% R^2) in Hi-C contact map prediction. Large-scale pre-training enables

robust semantic representations for MIX-HIC, while other methods lack generalizability due to being narrowly optimized for specific tasks and often show inconsistent performance across datasets.

5.2 Few-shot Chromatin Loops Classification

We conduct a few-shot learning experiment on the chromatin loops classification task to evaluate the performance of MIX-HIC under limited training data scenarios. Specifically, we fine-tune MIX-HIC-Bimodal, MIX-HIC-NonPre, DLoopCaller, and Peakachu using four different ratios of the training data, ranging from 0.0001 to 0.1. As illustrated in Figure 4, MIX-HIC-Bimodal consistently outperforms other methods across all data ratios. Notably, the performance of most methods remains relatively robust even with minimal data, primarily stemming from the powerful and repetitive biological signatures of chromatin loops, which make the task tractable for most models. The performance variations among different architectures also highlight their inherent data efficiency. For instance, Peakachu, a random forest model, is inherently stable as it relies on engineered features that are less sensitive to data volume. In contrast, DLoopCaller's CNN architecture is more data-hungry and thus shows a steeper decline. Our MIX-HIC architecture, even without pre-training, proves more data-efficient: the self-attention mechanism is better suited for capturing the global and long-range dependencies in contact maps than local CNNs, and its bimodal input provides complementary information, enhancing robustness even in low-data regimes.

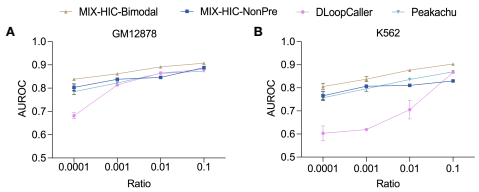

Figure 4: Few-shot chromatin loop classification performance across different training data ratios. Mean values and standard errors are calculated over five independent runs with varying random seeds.

With a training data ratio of 0.1, MIX-HIC-Bimodal achieves an AUROC of about 0.9 on two datasets, which is competitive with other state-of-the-art methods trained on full datasets. These findings highlight the robustness and efficiency of MIX-HIC in leveraging pre-trained knowledge to achieve superior performance even with limited labels.

5.3 Robust Performance Across Cell Types

The utilization of cell-type specific data, *i.e.*, Hi-C contact maps and epigenomic tracks, empowers MIX-HIC to achieve accurate predictions for novel cell types. We perform a cross-dataset evaluation with the GM12878 and K562 cell lines on the chromatin loop detection task to assess the generalization ability of MIX-HIC. In particular, all the models are trained on one cell line and evaluated on the other. The evaluation results of cross-cell type prediction are shown in Figure 5. We note a decrease in performance when all models are evaluated on external datasets compared to within-dataset testing (see Table 3). Nevertheless, MIX-HIC-Bimodal continues to outperform other methods, indicating its strong generalization ability in real-world scenarios.

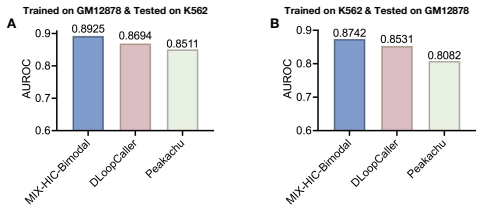

Figure 5: Methods comparison for cross-cell-type evaluation.

5.4 Ablation Study

To evaluate the effectiveness of key components in MIX-HIC, we perform ablation studies focusing on the proposed loss terms and the representation learning ability across multimodal data.

Three critical loss terms are employed during MIX-HIC pre-training (*i.e.*, $\mathcal{L}_{\text{con}}$, $\mathcal{L}_{\text{orth}}$, and $\mathcal{L}_{\text{mapping}}$). We examine the effects of each loss component. As shown in Table 5, the model equipped with $\mathcal{L}_{\text{con}}$ provides a strong baseline.

Table 5: AUROC results of ablation studies on loss terms for chromatin loop detection task. Symbols '✓' and '-' denote present and absent, respectively.

$\mathcal{L}_{\text{con}}$	$\mathcal{L}_{\text{orth}}$	$\mathcal{L}_{\text{mapping}}$	GM12878	K562
✓	-	-	0.9136	0.9099
✓	✓	-	0.9183	0.9156
✓	✓	✓	**0.9209**	**0.9194**

The orthogonal loss enhances discrimination between modal-invariant and modal-specific representations, contributing about 0.5% AUROC improvement compared to simply feature alignment, as demonstrated in Theorem 1. *Further details for the orthogonal constaint are provided in Appendix C.4.* Although the cross-modal mapping loss provides modest enhancement, it enables missing modality inference of MIX-HIC. These findings demonstrate that robust multimodal fusion is achieved by explicitly separating modal-invariant and modal-specific representations using appropriate loss terms.

In addition, we assess the representation learning ability of MIX-HIC across three aspects: the contribution of pre-training, the superiority of multimodal over unimodal representations, and the efficiency of modality completion. For the Hi-C contact map prediction task, the model is trained to predict the contact map using only 1D epigenomic tracks as input. The Hi-C contact map itself is the prediction target (output), making its use as an input feature methodologically invalid. Detailed comparison results appear in Table 6. The pre-trained bimodal MIX-HIC achieves superior performance compared to its non-pre-trained counterpart, confirming pre-training's benefit for unified representations. Moreover, while pre-trained bimodal MIX-HIC consistently outperforms single-modal variants, the non-pre-trained bimodal version underperforms compared to using Hi-C data alone for K562 chromatin loop detection due to heterogeneity between bimodal data, highlighting the importance of learning both modality-specific information and general patterns. Finally, when Hi-C data is unavailable, combining epigenomic embeddings with inferred Hi-C embeddings improves performance over using epigenomic data alone, validating the cross-modal mapping block's effectiveness for data scarcity scenarios.

Table 6: Ablation results of each modality, with values reported as R^2 (Hi-C contact map prediction), AUROC (chromatin loops dectection), and R^2 (CAGE-seq expression prediction). Symbol '$\circ$' represents inferred embeddings from the other modality.

Tasks	Epi.	Hi-C	Pre-trained	GM12878	K562
Hi-C contact map prediction	✓	-	-	0.8481	0.7709
	✓	$\circ$	✓	**0.8724**	**0.8001**
Chromatin loops dectection	✓	-	-	0.8236	0.8054
	✓	$\circ$	✓	0.8494	0.8226
	-	✓	-	0.9065	0.9072
	$\circ$	✓	✓	0.9135	0.9159
	✓	✓	-	0.9091	0.8859
	✓	✓	✓	**0.9209**	**0.9194**
CAGE-seq expression prediction	✓	-	-	0.8514	0.8710
	✓	$\circ$	✓	0.8684	0.8870
	✓	✓	-	0.8614	0.8755
	✓	✓	✓	**0.8833**	**0.9077**

6 Conclusion

In this work, we present MIX-HIC, a novel multimodal foundation model for diverse 3D genome downstream tasks by integrating Hi-C contact maps and epigenomic tracks. To facilitate a unified and comprehensive understanding of 3D genome organization, we construct the largest paired 3D genome dataset to date, comprising over one million high-quality samples. As the first multimodal foundation model in this domain, MIX-HIC incorporates two key innovations: (1) a cross-modal interaction block that jointly learns modal-invariant and modal-specific representations, effectively capturing both shared and unique biological patterns across modalities; and (2) a cross-modal mapping block that regularizes the multimodal feature space and enables robust imputation of missing modalities features, alleviating the practical challenges posed by the high costs of Hi-C contact map acquisition. Comprehensive experimental results demonstrate that MIX-HIC achieves state-of-the-art performance on three key downstream tasks across two cell lines.

7 Acknowledgments

This work was supported in part by the National Key R&D Program of China (Grant No.2023YFF0725001), in part by the National Natural Science Foundation of China (Grant Nos.92370204, 62572413), in part by the guangdong Basic and Applied Basic Research Foundation (Grant Nos.2023B1515120057, 2025A1515012944), and in part by the Education Bureau of Guangzhou.

References

[1] Yang Zhang, Lorenzo Boninsegna, Muyu Yang, Tom Misteli, Frank Alber, and Jian Ma. Computational methods for analysing multiscale 3D genome organization. *Nature Reviews Genetics*, 25(2):123–141, 2024.

[2] Ana Monteagudo-Sánchez, Daan Noordermeer, and Maxim VC Greenberg. The impact of DNA methylation on CTCF-mediated 3D genome organization. *Nature Structural & Molecular Biology*, 31(3):404–412, 2024.

[3] Erez Lieberman-Aiden, Nynke L Van Berkum, Louise Williams, Maxim Imakaev, Tobias Ragoczy, Agnes Telling, Ido Amit, Bryan R Lajoie, Peter J Sabo, Michael O Dorschner, et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, 326(5950):289–293, 2009.

[4] Jorge Ferrer and Nadya Dimitrova. Transcription regulation by long non-coding RNAs: mechanisms and disease relevance. *Nature Reviews Molecular Cell Biology*, 25(5):396–415, 2024.

[5] Rui Yang, Arnav Das, Vianne R Gao, Alireza Karbalayghareh, William S Noble, Jeffrey A Bilmes, and Christina S Leslie. Epiphany: predicting Hi-C contact maps from 1D epigenomic signals. *Genome Biology*, 24(1):134, 2023.

[6] Jimin Tan, Nina Shenker-Tauris, Javier Rodriguez-Hernaez, Eric Wang, Theodore Sakellaropoulos, Francesco Boccalatte, Palaniraja Thandapani, Jane Skok, Iannis Aifantis, David Fenyö, et al. Cell-type-specific prediction of 3D chromatin organization enables high-throughput in silico genetic screening. *Nature Biotechnology*, 41(8):1140–1150, 2023.

[7] Geoff Fudenberg, David R Kelley, and Katherine S Pollard. Predicting 3D genome folding from DNA sequence with Akita. *Nature Methods*, 17(11):1111–1117, 2020.

[8] Tarik J Salameh, Xiaotao Wang, Fan Song, Bo Zhang, Sage M Wright, Chachrit Khunsriraksakul, Yijun Ruan, and Feng Yue. A supervised learning framework for chromatin loop detection in genome-wide contact maps. *Nature Communications*, 11(1):3428, 2020.

[9] Cyril Matthey-Doret, Lyam Baudry, Axel Breuer, Rémi Montagne, Nadège Guiglielmoni, Vittore Scolari, Etienne Jean, Arnaud Campeas, Philippe Henri Chanut, Edgar Oriol, et al. Computer vision for pattern detection in chromosome contact maps. *Nature Communications*, 11(1):5795, 2020.

[10] Siguo Wang, Qinhu Zhang, Ying He, Zhen Cui, Zhenghao Guo, Kyungsook Han, and De-Shuang Huang. DLoopCaller: A deep learning approach for predicting genome-wide chromatin loops by integrating accessible chromatin landscapes. *PLoS Computational Biology*, 18(10):e1010572, 2022.

[11] Zhenhao Zhang, Fan Feng, Yiyang Qiu, and Jie Liu. A generalizable framework to comprehensively predict epigenome, chromatin organization, and transcriptome. *Nucleic Acids Research*, 51(12):5931–5947, 2023.

[12] Alireza Karbalayghareh, Merve Sahin, and Christina S Leslie. Chromatin interaction–aware gene regulatory modeling with graph attention networks. *Genome Research*, 32(5):930–944, 2022.

[13] Zhen Wang, Zhifeng Gao, Hang Zheng, Linfeng Zhang, Guolin Ke, et al. Exploring molecular pretraining model at scale. In *Advances in Neural Information Processing Systems*, 2024.

[14] Jinho Chang and Jong Chul Ye. Bidirectional generation of structure and properties through a single molecular foundation model. *Nature Communications*, 15(1):2323, 2024.

[15] Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. LLaVA-Med: Training a large language-and-vision assistant for biomedicine in one day. In *Advances in Neural Information Processing Systems*, 2023.

[16] Peng Xia, Ze Chen, Juanxi Tian, Yangrui Gong, Ruibo Hou, Yue Xu, Zhenbang Wu, Zhiyuan Fan, Yiyang Zhou, Kangyu Zhu, et al. Cares: A comprehensive benchmark of trustworthiness in medical vision language models. In *Advances in Neural Information Processing Systems*, 2024.

[17] Xingyi Cheng, Bo Chen, Pan Li, Jing Gong, Jie Tang, and Le Song. Training compute-optimal protein language models. In *Advances in Neural Information Processing Systems*, 2024.

[18] Mingchen Li, Yang Tan, Xinzhu Ma, Bozitao Zhong, Huiqun Yu, Ziyi Zhou, Wanli Ouyang, Bingxin Zhou, Liang Hong, and Pan Tan. Prosst: Protein language modeling with quantized structure and disentangled attention. In *Advances in Neural Information Processing Systems*, 2024.

[19] Siyuan Li, Zedong Wang, Zicheng Liu, Di Wu, Cheng Tan, Jiangbin Zheng, Yufei Huang, and Stan Z Li. VQDNA: Unleashing the power of vector quantization for multi-species genomic sequence modeling. In *International Conference on Machine Learning*, 2024.

[20] Yair Schiff, Chia-Hsiang Kao, Aaron Gokaslan, Tri Dao, Albert Gu, and Volodymyr Kuleshov. Caduceus: Bi-directional equivariant long-range DNA sequence modeling. In *International Conference on Machine Learning*, 2024.

[21] Ron Schwessinger, Matthew Gosden, Damien Downes, Richard C Brown, A Marieke Oudelaar, Jelena Telenius, Yee Whye Teh, Gerton Lunter, and Jim R Hughes. DeepC: predicting 3D genome folding using megabase-scale transfer learning. *Nature Methods*, 17(11):1118–1124, 2020.

[22] Joachim Wolff, Leily Rabbani, Ralf Gilsbach, Gautier Richard, Thomas Manke, Rolf Backofen, and Björn A Grüning. Galaxy HiCExplorer 3: a web server for reproducible Hi-C, capture Hi-C and single-cell Hi-C data analysis, quality control and visualization. *Nucleic Acids Research*, 48(W1):W177–W184, 2020.

[23] Yanlin Zhang and Mathieu Blanchette. Reference panel-guided super-resolution inference of Hi-C data. *Bioinformatics*, 39:i386–i393, 2023.

[24] Žiga Avsec, Vikram Agarwal, Daniel Visentin, Joseph R Ledsam, Agnieszka Grabska-Barwinska, Kyle R Taylor, Yannis Assael, John Jumper, Pushmeet Kohli, and David R Kelley. Effective gene expression prediction from sequence by integrating long-range interactions. *Nature Methods*, 18(10):1196–1203, 2021.

[25] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.

[26] Jiale Zhao, Wanru Zhuang, Jia Song, Yaqi Li, and Shuqi Lu. Pre-training protein bi-level representation through span mask strategy on 3D protein chains. In *International Conference on Machine Learning*, 2024.

[27] Chloe Hsu, Robert Verkuil, Jason Liu, Zeming Lin, Brian Hie, Tom Sercu, Adam Lerer, and Alexander Rives. Learning inverse folding from millions of predicted structures. In *International Conference on Machine Learning*, 2022.

[28] Shikun Feng, Yuyan Ni, Minghao Li, Yanwen Huang, Zhi-Ming Ma, Wei-Ying Ma, and Yanyan Lan. Unicorn: A unified contrastive learning approach for multi-view molecular representation learning. *International Conference on Machine Learning*, 2024.

[29] Guoliang Li, Melissa J Fullwood, Han Xu, Fabianus Hendriyan Mulawadi, Stoyan Velkov, Vinsensius Vega, Pramila Nuwantha Ariyaratne, Yusoff Bin Mohamed, Hong-Sain Ooi, Chandana Tennakoon, et al. ChIA-PET tool for comprehensive chromatin interaction analysis with paired-end tag sequencing. *Genome Biology*, 11:1–13, 2010.

[30] Lei Chang, Yang Xie, Brett Taylor, Zhaoning Wang, Jiachen Sun, Ethan J Armand, Shreya Mishra, Jie Xu, Melodi Tastemel, Audrey Lie, et al. Droplet Hi-C enables scalable, single-cell profiling of chromatin architecture in heterogeneous tissues. *Nature Biotechnology*, pages 1–14, 2024.

[31] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*, 2020.

[32] Xiao Wang, Daisuke Kihara, Jiebo Luo, and Guo-Jun Qi. EnAET: A self-trained framework for semi-supervised and supervised learning with ensemble transformations. *IEEE Transactions on Image Processing*, 30:1639–1647, 2020.

[33] Xiao Wang, Haoqi Fan, Yuandong Tian, Daisuke Kihara, and Xinlei Chen. On the importance of asymmetry for siamese representation learning. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pages 16570–16579, 2022.

[34] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

[35] Minghao Yang, Shichen Zhang, Zhihang Zheng, Pengfei Zhang, Yan Liang, and Shaojun Tang. Employing bimodal representations to predict DNA bendability within a self-supervised pre-trained framework. *Nucleic Acids Research*, 52(6):e33–e33, 2024.

[36] Hao Dong, Ismail Nejjar, Han Sun, Eleni Chatzi, and Olga Fink. SimMMDG: A simple and effective framework for multi-modal domain generalization. In *Advances in Neural Information Processing Systems*, 2023.

[37] Hu Wang, Yuanhong Chen, Congbo Ma, Jodie Avery, Louise Hull, and Gustavo Carneiro. Multi-modal learning with missing modality via shared-specific feature modelling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15878–15887, 2023.

[38] Taekyung Kim, Debasmit Das, Seokeon Choi, Minki Jeong, Seunghan Yang, Sungrack Yun, and Changick Kim. Neural transformation network to generate diverse views for contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[39] Donggeun Kim and Taesup Kim. Missing modality prediction for unpaired multimodal learning via joint embedding of unimodal models. In *European Conference on Computer Vision*, pages 171–187. Springer, 2025.

[40] Fidel Ramírez, Devon P Ryan, Björn Grüning, Vivek Bhardwaj, Fabian Kilpert, Andreas S Richter, Steffen Heyne, Friederike Dündar, and Thomas Manke. deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Research*, 44:W160, 2016.

[41] Arya Kaul, Sourya Bhattacharyya, and Ferhat Ay. Identifying statistically significant chromatin contacts from Hi-C data with FitHiC2. *Nature Protocols*, 15(3):991–1012, 2020.

[42] Nico JD Nagelkerke. A note on a general definition of the coefficient of determination. *Biometrika*, 78(3):691–692, 1991.

[43] Clifford A Meyer and X Shirley Liu. Identifying and mitigating bias in next-generation sequencing methods for chromatin biology. *Nature Reviews Genetics*, 15(11):709–721, 2014.

[44] Yanlin Zhang and Mathieu Blanchette. Reference panel guided topological structure annotation of hi-c data. *Nature Communications*, 13(1):7426, 2022.

[45] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[46] Normand J Beaudry and Renato Renner. An intuitive proof of the data processing inequality. *Quantum Information and Computation*, 12(5–6):432–441, 2012.

[47] Han Zhao, Chen Dan, Bryon Aragam, Tommi S Jaakkola, Geoffrey J Gordon, and Pradeep Ravikumar. Fundamental limits and tradeoffs in invariant representation learning. *Journal of Machine Learning Research*, 23(340):1–49, 2022.

[48] Lun Zhao, Shuangqi Wang, Zhilin Cao, Weizhi Ouyang, Qing Zhang, Liang Xie, Ruiqin Zheng, Minrong Guo, Meng Ma, Zhe Hu, et al. Chromatin loops associated with active genes and heterochromatin shape rice genome architecture for transcriptional regulation. *Nature Communications*, 10(1):3640, 2019.

[49] Merve Sahin, Wilfred Wong, Yingqian Zhan, Kinsey Van Deynze, Richard Koche, and Christina S Leslie. HiC-DC+ enables systematic 3D interaction calls and differential analysis for Hi-C and HiChIP. *Nature Communications*, 12(1):3366, 2021.

[50] Charles E Grant, Timothy L Bailey, and William Stafford Noble. Fimo: scanning for occurrences of a given motif. *Bioinformatics*, 27(7):1017–1018, 2011.

[51] Galip Gürkan Yardımcı, Hakan Ozadam, Michael EG Sauria, Oana Ursu, Koon-Kiu Yan, Tao Yang, Abhijit Chakraborty, Arya Kaul, Bryan R Lajoie, Fan Song, et al. Measuring the reproducibility and quality of Hi-C data. *Genome Biology*, 20(1):57, 2019.

[52] Yannick G Spill, David Castillo, Enrique Vidal, and Marc A Marti-Renom. Binless normalization of Hi-C data provides significant interaction and difference detection independent of resolution. *Nature Communications*, 10(1):1938, 2019.

[53] Eric Nguyen, Michael Poli, Marjan Faizi, Armin Thomas, Michael Wornow, Callum Birch-Sykes, Stefano Massaroli, Aman Patel, Clayton Rabideau, Yoshua Bengio, et al. HyenaDNA: Long-range genomic sequence modeling at single nucleotide resolution. In *Advances in Neural Information Processing Systems*, 2024.

A Implementation Details.

In this section, we present the implementation details of MIX-HIC, including data source and processing, whole-chromosome loop annotation, as well as experimental settings.

Table 7: Data sources and accession numbers from 4DN Data Portal and ENCODE Data Portal.

Cell line	Hi-C	DNase-seq	ATAC-seq	ChIA-PET	CAGE-seq
GM12878	4DNFI1UEG1HD	ENCSR000EMT	ENCSR095QNB	ENCSR184YZV	ENCSR000CKA
K562	4DNFITUOMFUQ	ENCSR000EOT	ENCSR956DNB	ENCSR597AKG	ENCSR000CJN
HepG2	4DNFICSTCJQZ	ENCSR000EJV	ENCSR042AWH	ENCSR411IVB	-
HCT116	4DNFIXTAS6EE	ENCSR000ENM	ENCSR872WGW	ENCSR278IZK	-
IMR90	4DNFIH7TH4MF	ENCSR477RTP	ENCSR200OML	ENCSR076TTY	-
WTC11	4DNFIVSCH2CH	ENCSR785ZUI	ENCSR541KFY	ENCSR353ASS	-

A.1 Data Source and Processing

Detailed reference numbers for all data sources are provided in Table 7. We download the BAM files for epigenomic tracks and convert them into bigWig files using deepTools [40] with RPGC normalization. Two 250 kb epigenomic regions, corresponding to the x-axis and y-axis of a Hi-C contact map, are first averaged over every 100 bp and transformed using $log(x + 1)$ to reduce data variability, and then the processed regions are concatenated to form the final epigenomic tracks input with a length of 5,000. The Hi-C contact maps are binned at 5 kb resolution and normalized using KR normalization, and then divided into 50×50 sub-matrices based on the loci of each sample.

As the first attempt to create the multimodal 3D genome foundation model, we prioritize epigenomic tracks with ATAC-seq and DNase-seq because they capture the genome's foundational regulatory information. While DNase-seq and ATAC-seq have some inherent redundancy, they also offer complementary insights due to their distinct enzymatic biases. The enzymes' different cutting biases cause each assay to detect unique accessible sites, which together yield a more complete accessibility landscape [43].

The training samples of three downstream tasks are generated using two widely adopted cell lines, GM12878 and K562, as described below. First, the Hi-C contact maps and epigenomic tracks within a 250 kb genomic region upstream and downstream of the gene transcription start site (TSS) are used to predict CAGE-seq expression, with TSS annotations obtained from previous work [12]. For this task, the number of training samples for GM12878 and K562 are 16,046 and 14,739, respectively. Second, the same pairwise samples of Hi-C contact maps and epigenomic tracks are used for the Hi-C contact map prediction task, with the same sample sizes as mentioned above. Third, the positive chromatin loops (1,344,270 for GM12878 and 137,558 for K562) are derived from the ChIA-PET data, while an equal number of negative loops are randomly sampled based on two criteria following prior research [10]: (1) matching the distance distribution of positive samples using the distances probability density function, and (2) selecting interactions with distances greater than the maximum distance observed in the positive samples to enhance the diversity of the negative samples. Following previous work [10, 6, 5], we partition the chromosomes into distinct training, validation, and test sets. Specifically, chromosomes 10 and 11 serve as the validation set, chromosomes 3, 13, and 17 as the test set, and the remaining chromosomes are used for model training across three downstream tasks.

A.2 Whole-chromosome Loops Annotation

The well-trained MIX-HIC model is capable of predicting all potential chromatin interactions across individual chromosomes. The detection of chromatin loops within each chromosome consists of two key steps: (1) scoring potential chromatin interactions using MIX-HIC and (2) aggregating loops through a clustering algorithm. First, candidate elements are identified by examining the diagonals of the raw Hi-C contact matrix, where observed interaction frequencies are statistically compared to expected values based on a Poisson distribution, retaining only elements with p-values below 0.01. Subsequently, the trained MIX-HIC model is employed to predict loop probability scores, with scores exceeding 0.5 referred to as candidate loops. These candidates are further grouped to identify significant loops using a density-based clustering algorithm following [44]. For each candidate, a

local density score and a minimum distance metric are computed. Candidates exhibiting low distance values are subsequently eliminated to minimize redundancy. Cluster centers are determined via a target-decoy search, ensuring that the false discovery rate remains below 5%. Within each cluster, the candidate demonstrating the highest local density is designated as the representative loop.

A.3 Experimental Settings

MIX-HIC is developed using Python and PyTorch, and executed on the Ubuntu platform with a Tesla A100 GPU. MIX-HIC processes input data consisting of 50×50 Hi-C contact maps and epigenomic tracks with a sequence length of 5,000 bps. The architecture is composed of feature encoders, task-specific decoders, and a modality fusion block. The feature encoders and task-specific decoders are structured with four layers, each comprising two transformer encoder blocks. Notably, the epigenomic feature encoder includes an additional preprocessing stage, integrating four convolutional layers followed by max-pooling operations before the transformer blocks, to effectively process and condense long sequences. Similarly, the modality fusion block is constructed with two transformer encoder blocks, ensuring efficient integration of features across modalities.

During the pre-training stage, MIX-HIC is configured with 500 epochs, a learning rate of 1e-5, and a batch size of 256. For the CAGE-seq expression prediction task, the predefined feature dimension C and learning rate are set to 256 and 1e-4, respectively. For the prediction of Hi-C contact maps and chromatin loops, these parameters are specified as 128 and 1e-5, respectively. The number of transformer blocks T in each encoder, contact map-grounded fusion block, and decoder is set to 2. Fine-tuning is conducted with a batch size of 64, utilizing the AdamW optimizer [45] with the momentum parameters β_1 and β_2 initialized to 0.9 and 0.999, respectively. The fine-tuning process is configured with a maximum of 200 epochs, and an early stopping strategy is employed with a patience parameter of 20 to prevent overfitting.

B Analysis of Information Gap Between Bimodal Representations.

Simple alignment of multimodal features can result in information loss, which may compromise the performance of downstream tasks. Under conditions of perfect feature alignment, the prediction error using aligned features is at least Γ_q greater than that using the raw inputs. Based on [36], we generalize the theorem to the case of two modalities as Theorem 1. This theorem implies that if one modality is more informative than the others (*i.e.*, there exists a significant information gap), perfect alignment may lead to a substantial increase in prediction error. The underlying reason is that perfect alignment constrains the aligned features to preserve only the predictive information shared across all modalities, which may result in the loss of modal-specific information that is potentially critical for achieving accurate predictions. The proof of this theorem is presented as follows:

Consider the joint mutual information $U(\mathbf{F}^1, \mathbf{F}^2; t)$. Applying the chain rule of mutual information, we obtain:

$$U(\mathbf{F}^1, \mathbf{F}^2; t) = U(\mathbf{F}^1; t) + U(\mathbf{F}^2; t|\mathbf{F}^1) \tag{10}$$

$$= U(\mathbf{F}^2; t) + U(\mathbf{F}^1; t|\mathbf{F}^2). \tag{11}$$

Under the condition of perfect alignment between the two features, $U(\mathbf{F}^2; t|\mathbf{F}^1) = U(\mathbf{F}^1; t|\mathbf{F}^2) = 0$, which implies:

$$U(\mathbf{F}^1, \mathbf{F}^2; t) = U(\mathbf{F}^1; t) = U(\mathbf{F}^2; t). \tag{12}$$

By applying the data processing inequality [46]:

$$U(\mathbf{F}^1; t) \leq U(z_1; t); U(\mathbf{F}^2; t) \leq U(z_2; t), \tag{13}$$

we derive $U(\mathbf{F}^1, \mathbf{F}^2; t)$ as follows:

$$\begin{aligned} U(\mathbf{F}^1, \mathbf{F}^2; t) &= \min\{U(F^1; t), U(F^2; t)\} \\ &\leq \min\{U(z_1; t), U(z_2; t)\} \\ &\leq \max\{U(z_1; t), U(z_2; t)\} \\ &\leq U(z_1, z_2; t). \end{aligned}$$

According to the variational form of conditional entropy $H(t|z_1, z_2) = \inf_h \mathbb{E}_q[\mathcal{L}_{\text{CE}}(h(z_1, z_2), t)]$ [47] and the definition of mutual information $U(X;Y) = H(Y) - H(X|Y)$, we can conclude the theorem as follows:

$$\begin{aligned}
&\inf_h \mathbb{E}_q[\mathcal{L}_{\text{CE}}(h(\mathbf{F}^1, \mathbf{F}^2), t)] - \inf_{h'} \mathbb{E}_q[\mathcal{L}_{\text{CE}}(h'(z_1, z_2), t)] \\
&\qquad = H(t|\mathbf{F}^1, \mathbf{F}^2) - H(t|z_1, z_2) \\
&\qquad = H(t) - U(\mathbf{F}^1, \mathbf{F}^2|t) - (H(t) - U(z_1, z_2|t)) \\
&\qquad = U(z_1, z_2; t) - U(\mathbf{F}^1, \mathbf{F}^2; t) \\
&\qquad \geq \Gamma_q.
\end{aligned}$$

Therefore, although perfect alignment of bimodal features achieves consistency, it risks losing critical modal-specific information, potentially leading to higher prediction errors, especially when the information gap between modalities is substantial, as seen in the case of Hi-C contact maps and epigenomic tracks. To solve this problem, we introduce cross-modal interaction and cross-modal mapping blocks to capture both modal-invariant and modal-specific features, enabling a more comprehensive representation of bimodal data. The effectiveness of this approach is validated through extensive ablation studies.

C Additional Experimental Results.

C.1 Proportion of Validated Loops versus Predicted Loops

We define the quantitative metric 'proportion' as the ratio of experimentally validated loops to computationally predicted loops [44]. Table 8 presents a comparative analysis of chromatin interaction loops identified through ChIA-PET experiments and those predicted by different approaches. MIX-HIC exceeds runner-up methods by 28% and 9% proportion on GM12878 and K562 datasets, respectively.

Table 8: Comparison of ChIA-PET-supported loops with those identified by various deep learning methods.

Methods	GM12878			K562		
	Validated Loops	Predicted Loops	Proportion	Validated Loops	Predicted Loops	Proportion
HiCExplorer	1358	2688	0.51	765	1450	0.53
ChromoSight	3845	6044	0.64	993	1628	0.61
Peakachu	1992	2904	0.69	795	1492	0.53
DLoopCaller	4301	6878	0.62	1048	2274	0.46
MIX-HIC-Bimodal (Ours)	5179	5893	0.88	1064	1588	0.67

C.2 Example of Whole-Chromosome Chromatin Loop Comparison

Figure 6 illustrates the performance evaluation of chromatin loops predicted by various methods (black points, lower left) against corresponding experimentally validated interactions (red points, upper right). Specifically, we use the same genome regions, chromosome 13: 20Mb-22Mb from the GM12878 dataset, and identical ground truth data for evaluating all compared methods. Our predictions (the first left panel) demonstrate significantly fewer false positive loops compared to experimentally validated ChIA-PET interactions, as evidenced by the correspondent counts between red and black points. Our method achieves a proportion of 0.57, compared to the runner-up's 0.32, highlighting MIX-HIC's superior ability to accurately identify chromatin loops while producing fewer false positives.

C.3 Hyperparameter Analysis

The number of Transformer T blocks in each layer of the encoder, contact map-grounded fusion block, and decoder, along with the predefined feature dimension C play a pivotal role in MIX-HIC. We evaluate the impact of this parameter by testing $C = \{64, 128, 256\}$ and $T = \{2, 4, 8\}$ as depicted in Figure 7 and Figure 8, respectively. The highest performance is achieved at $C = 128$ for HI-C contact maps and chromatin loops prediction tasks, while $C = 256$ performs best for CAGE-seq

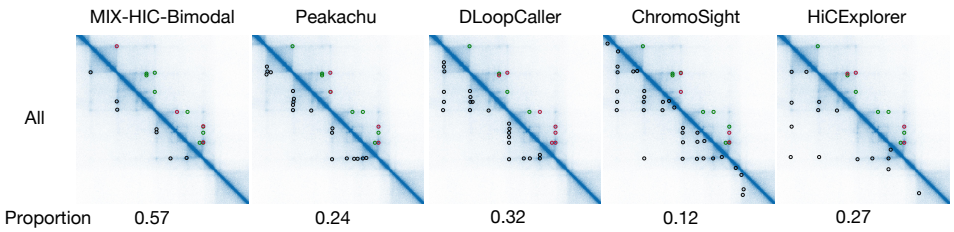

Figure 6: Comparison of predicted chromatin loops (black points, lower left) and corresponding experimentally validated ChIA-PET interactions (red points, upper right) in the chromosome 13 region (20Mb-22Mb) from the GM12878 dataset. Unmatched experimentally validated ChIA-PET interactions are represented by green points in the upper right.

expression prediction. Additionally, $T = 2$ demonstrates the most robust performance across all three downstream tasks. MIX-HIC exhibits strong resilience to parameter variations, demonstrating our model's robustness and parameter efficiency. The performance suggests our architecture reaches a "sweet spot" with a moderate parameter count and computational efficiency, sufficient to capture the essential biological patterns without incurring the high risk of overfitting during fine-tuning on smaller, task-specific datasets. As our pre-training corpus expands, we anticipate larger architectures will become beneficial. These experimental results substantiate the parameter configurations employed in MIX-HIC.

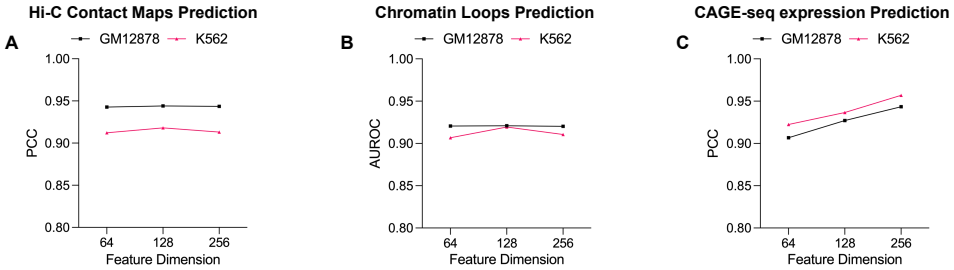

Figure 7: The performance across different predefined feature dimensions $C = \{64, 128, 256\}$ for three downstream tasks in GM12878 and K562 cell lines.

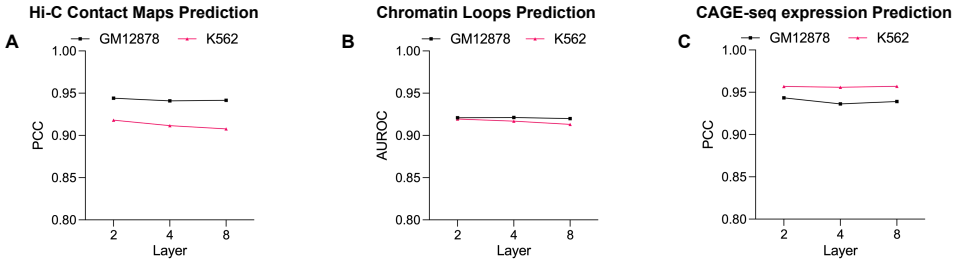

Figure 8: The performance across different Transformer layers $T = \{2, 4, 8\}$ for three downstream tasks in GM12878 and K562 cell lines.

C.4 Orthogonal Constraint Enhances Feature Diversity

Theorem 1 demonstrates that rigorous alignment can be detrimental to performance. The orthogonal constraint effectively mitigates this issue by promoting diversity between modality-invariant and modality-specific features, rather than enforcing strict non-overlap. To validate feature diversity, we

compute the inner product between modality-invariant $\hat{X_M^I}$ and modality-specific $\hat{X_M^S}$ features of Hi-C contact maps, as well as between their counterparts $\hat{X_E^I}$ and $\hat{X_E^S}$ in epigenomic tracks on the datasets of the CAGE-seq expression prediction task. As summarized in Table 9, the inner products under orthogonal constraints are orders of magnitude smaller than those without constraints (*e.g.*, $1\mathrm{e}{-5}$ versus 1.077 for Hi-C contact maps on GM12878 dataset). Our analysis of inner product results between these embeddings confirms that the constraint successfully generates near-orthogonal representations.

Table 9: Inner products between modality-invariant and modality-specific features.

Dataset	With Pre-training		Without Pre-training	
	Hi-C contact maps	Epigenomic tracks	Hi-C contact maps	Epigenomic tracks
GM12878	$0.003 \pm 3\mathrm{e}{-4}$	0.002 ± 0.001	1.419 ± 0.095	0.164 ± 0.059
K562	$1\mathrm{e}{-5} \pm 4\mathrm{e}{-4}$	0.002 ± 0.002	1.077 ± 0.074	0.144 ± 0.057

C.5 In Silico Perturbation Validates Biological Grounding of Chromatin Loop Detection

The presence and strength of certain epigenomic signals are highly associated with the formation of chromatin loops [48, 49]. For example, CTCF binding sites are typically located within regions of open chromatin, identifiable as peaks in assays such as DNase-seq or ATAC-seq.

To validate that MIX-HIC learns the fundamental principle linking epigenomic signals to 3D genome architecture, we design an in *silico* perturbation experiment. We hypothesize that the model's loop detection should be governed by the underlying epigenomic signals at the loop anchors. We focus our analysis on 118 high-confidence chromatin loops in the K562 cell line, each characterized by convergent CTCF motifs at their anchors (identified via FIMO [50]). We then systematically attenuate the input epigenomic tracks (ATAC-seq and DNase-seq) by down-sampling the signal intensity of peaks within these anchor regions at varying ratios. These perturbed epigenomic profiles are then fed into the MIX-HIC-InferMap model (trained on GM12878) to assess the impact on loop detection. The results are shown in Table 10.

Using the unaltered epigenomic data, MIX-HIC successfully recalls all 118 CTCF-mediated loops. As we progressively degrade the epigenomic signals at the loop anchors, the model's recall for these loops decreases. This demonstrates MIX-HIC's predictions are mechanistically grounded in biologically pertinent epigenomic features. This result demonstrates the model's biological interpretability, confirming the capture of a fundamental principle of genome organization.

Table 10: Impact of Epigenomic Signal Attenuation on MIX-HIC Loop Recall.

Varying ratio	0.0	0.5	0.7	0.8	0.9
MIX-HIC-InferMap	100% (118)	98% (116)	61% (72)	15% (18)	0% (0)

C.6 Robustness Analysis Under Noisy Conditions

The inherent noise and sparsity of Hi-C data [51, 52] present a critical challenge for developing robust and generalizable genomic models. To systematically assess the robustness of MIX-HIC, we conduct a controlled experiment simulating low-coverage and noisy scenarios. We corrupt Hi-C contact maps by perturbing different ratios of non-zero contacts with sparsity and Gaussian noise. Under these varying noise levels, MIX-HIC and the supervised baseline Peakachu are evaluated on the chromatin loop detection task using the K562 cell line dataset.

As shown in Table 11, the performance of Peakachu degrades sharply as noise increases, falling to near-random performance (0.5091 AUROC) when 70% of the contacts are disturbed. In contrast, MIX-HIC exhibits remarkable robustness, with its performance declining by less than 7% around all noise ratios.

We attribute this resilience to the power of our pre-training paradigm, which learns the fundamental principles of 3D genome organization from over 1 million samples. This experiment demonstrates that MIX-HIC develops a robust biological representation and is thus particularly suitable for analyzing noisy or low-coverage datasets. We will add this analysis to our manuscript to further strengthen the paper.

Table 11: AUROC comparison of MIX-HIC and Peakachu on chromatin loop detection under varying levels of simulated noise and sparsity.

Varying ratio	0.0	0.5	0.7	0.9
Peakachu	0.8833	0.7659	0.5091	-
MIX-HIC	0.9194	0.8899	0.8754	0.8486

D Baseline Methods and Assets

In this study, the effectiveness of MIX-HIC is evaluated on three downstream tasks, including Hi-C contact map prediction, chromatin loop detection, and CAGE-seq expression prediction. We employ four kinds of methods involving Epiphany [5], C.Origami [6], EPCOT-LSTM [11], and EPCOT-Transformer [11] for comparison on the Hi-C contact map prediction task. For chromatin loop detection task, two statistical-based methods (ChromoSight [9] and HiCExplorer [22]), as well as two supervised learning-based approaches (Peakachu [8] and DLoopCaller [10]) are utilized for evaluation. The CAGE-seq expression prediction involves EPI-CNN [12], EPI-Graph [12], EPCOT-LSTM [11], and EPCOT-Transformer [11] to compare with MIX-HIC. More details of these baselines are described below.

- Epiphany [5] predicts cell-type-specific Hi-C contact maps using bidirectional LSTMs to encode epigenomic tracks.
- C.Origami [6] introduces a multimodal framework that predicts chromatin organization from DNA sequence, CTCF binding signals, and epigenomic tracks, enabling the effective identification of regulatory elements.
- ChromoSight [9] proposes a computer vision-inspired algorithm for chromatin loop detection that employs expert-defined pattern templates, demonstrating computational efficiency across diverse species without requiring training data.
- HiCExplorer [22] identifies significant chromatin interactions by analyzing Hi-C contact matrices, employing binomial distribution modeling to distinguish true loops from background noise while controlling for distance-dependent contact probability.
- Peakachu [8] develops a supervised random forest classification framework that leverages chromatin interaction labels to predict chromatin loops from Hi-C contact maps, outperforming statistical enrichment methods in identifying short-range interactions.
- Xu et al. [10] propose DLoopCaller, a supervised deep learning method that predicts genome-wide chromatin loops by integrating epigenomic tracks with raw Hi-C contact maps.
- Karbalayghareh et al. [12] extracts local features from epigenomic tacks using convolutional neural networks (EPI-CNN) and incorporates Hi-C contact maps through graph attention networks (EPI-Graph) to predict CAGE-seq expression.
- EPCOT [11] introduces a pre-training and fine-tuning deep learning method that leverages Transformer (EPCOT-Transformer) or LSTM (EPCOT-LSTM) architectures to predict Hi-C contact maps and CAGE-seq expression profiles from epigenomic tracks and DNA sequences, achieving generalizable representations across diverse cell types.

The methods Epiphany, C.Origami, ChromoSight, Peakachu, and DLoopCaller are licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0), while EPI-CNN, EPI-Graph, EPCOT-Transformer, and EPCOT-LSTM are available under an Attribution-NonCommercial 4.0 International License (CC BY-NC 4.0). HiCExplorer is distributed under the GNU General Public License v3.0 (GPL-3.0). The licenses of other open-source tools utilized in this work are summarized in Table 12.

Table 12: License of softwares used in this study.

Software	License	URL
Juicer	MIT license	https://github.com/aidenlab/juicer
pyBigWig	MIT license	https://github.com/deeptools/pyBigWig
hicstraw	MIT license	https://github.com/aidenlab/straw
Huggingface	Apache-2.0	https://huggingface.co/
Scikit-Learn	BSD-3-Clause	https://scikit-learn.org/stable/
Numpy	BSD-3-Clause	https://numpy.org/
Pytorch	BSD-3-Clause	https://pytorch.org/
Matplotlib	Matplotlib License	https://matplotlib.org/

E Broader Impacts and Limitations

Broader Impacts. The three-dimensional chromatin architecture fundamentally governs both cellular differentiation and disease progression by mediating genomic interactions. This necessitates the development of Hi-C foundation models to systematically unravel the mechanistic basis of gene regulatory networks in both physiological and disease contexts. MIX-HIC fundamentally advances 3D genomics analysis by addressing two critical limitations of current approaches. First, they are typically designed for single tasks with limited cross-task knowledge transfer capability. Second, they predominantly rely on single-modality data (either Hi-C or epigenomic tracks alone) due to the scarcity of paired multimodal datasets, resulting in an incomplete understanding of chromatin organization.

To address these challenges, we develop MIX-HIC, the first multimodal foundation model for 3D genomics, with three key innovations: 1) We curate the largest paired dataset (over 1 million samples) of Hi-C and epigenomic tracks to overcome data scarcity. 2) Our novel cross-modal interaction and mapping blocks simultaneously capture both modality-invariant and modality-specific features, enabling a reliable complement of missing modality features. 3) As a foundation model, MIX-HIC enables versatile adaptation to diverse downstream tasks, facilitating knowledge transfer across different 3D genomics tasks where current approaches operate in isolation. In summary, MIX-HIC provides a universal computational platform for systematically deciphering the coordinated mechanisms between chromatin spatial organization and epigenetic regulation, while pioneering new avenues for multimodal data integration in disease mechanism research, precision medicine, and synthetic biology applications.

Limitations. Although MIX-HIC has exhibited promising performance, several areas warrant further improvement. Current methods in the 3D genome field for processing long-range DNA sequences are often time-intensive. In future work, MIX-HIC could further enhance its capabilities by integrating DNA sequence information through leveraging recent advancements like MambaDNA [20] and HyenaDNA [53] to facilitate feature extraction of genomic sequences. Additionally, while the current version of MIX-HIC serves as a bulk-seq level foundation model, developing a single-cell version through the integration of large-scale multimodal single-cell data would be valuable to effectively address the inherent sparsity and noise in single-cell analyses. Finally, the diversity of cell lines in our pre-training set is an area for future enhancement. Expanding the pre-training corpus by integrating the paired Hi-C and epigenomic datasets available from ENCODE will broaden MIX-HIC's applicability across a wider range of biological contexts. Overall, this study presents MIX-HIC, a versatile foundation model that integrates 3D genome structures with chromatin accessibility, providing an efficient framework for advancing genomic organization research and related fields.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The contributions of this work are summarized in the introduction section, while the score is clearly described in the abstract.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The limitations are included in Appendix E.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: The proofs of the theorems are detailed in Appendix B.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The implementation details and experimental settings are provided in Appendix A.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: All data used in this work are publicly available, as detailed in Appendix A. The code of MIX-HIC is available in our GitHub.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: All the training hyperparameters are provided in Appendix A.3. We also perform hyperparameter analysis for two critical hyperparameters in Appendix C.3.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We report error bars for the few-shot learning experiments in Figure 4, computed over five independent runs with varying random seeds.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)

- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provide the computer resources used in this work in Appendix A.3.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: This work follows the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We state the broader impacts of this work in Appendix E.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: This work poses no such risks and the model is free to share.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We cite all referenced sources appropriately, which are provided in Appendix D.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: This paper does not introduce new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: All datasets used in this work stem from humans are anonymized, and sourced from publicly available publications to ensure privacy compliance. The source publications address key considerations regarding human subjects research.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: All datasets used in this work stem from humans are anonymized, and sourced from publicly available publications to ensure privacy compliance. The source publications address key considerations regarding human subjects research.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.

- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: The LLM tools are only used for writing.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.