Epiphany: predicting Hi-C contact maps from 1D epigenomic signals

Rui Yang[1†], Arnav Das[2†], Vianne R. Gao[1], Alireza Karbalayghareh[1], William S. Noble[2], Jeffrey A. Bilmes[2*], Christina S. Leslie[1*]

[1]Memorial Sloan Kettering Cancer Center, New York, NY, USA

[2]University of Washington, Seattle, WA, USA

[†]Equal contribution, [*]Correspondence: cleslie@cbio.mskcc.org, bilmes@uw.edu

December 2, 2021

Abstract

Recent deep learning models that predict the Hi-C contact map from DNA sequence achieve promising accuracy but cannot generalize to new cell types and indeed do not capture cell-type-specific differences among training cell types. We propose Epiphany, a neural network to predict cell-type-specific Hi-C contact maps from five epigenomic tracks that are already available in hundreds of cell types and tissues: DNase I hypersensitive sites and ChIP-seq for CTCF, H3K27ac, H3K27me3, and H3K4me3. Epiphany uses 1D convolutional layers to learn local representations from the input tracks, a bidirectional long short-term memory (Bi-LSTM) layers to capture long term dependencies along the epigenome, as well as a generative adversarial network (GAN) architecture to encourage contact map realism. To improve the usability of predicted contact matrices, we trained and evaluated models using multiple normalization and matrix balancing techniques including KR, ICE, and HiC-DC+ Z-score and observed-over-expected count ratio. Epiphany is trained with a combination of MSE and adversarial (i.a., a GAN) loss to enhance its ability to produce realistic Hi-C contact maps for downstream analysis. Epiphany shows robust performance and generalization to held-out chromosomes within and across cell types and species, and its predicted contact matrices yield accurate TAD and significant interaction calls. At inference time, Epiphany can be used to study the contribution of specific epigenomic peaks to 3D architecture and to predict the structural changes caused by perturbations of epigenomic signals.

Introduction

In vertebrate genomes, the three-dimensional (3D) hierarchical folding of chromatin in the nucleus plays a critical role in the regulation of gene expression, replication timing, and cellular differentiation [1, 2]. This 3D chromatin architecture has been elucidated through genome-wide chromosome conformation capture (3C) assays such as Hi-C, Micro-C, HiChIP, and ChIA-PET [3, 4, 5, 6] followed by next generation sequencing, yielding a contact matrix representation of pairwise chromatin interactions. Early Hi-C analyses revealed an organization of $\sim$1Mb self-interacting topologically associating domains (TADs) that may insulate within-TAD genes from enhancers outside of TAD boundaries [7]. High-resolution 3C-based studies have mapped regulatory interactions, often falling within TADs, that connect regulatory elements to target gene promoters [8, 9].

Over the past decade, large consortium projects as well as individual labs have extensively used 1D epigenomic assays to map regulatory elements and chromatin states across numerous human and mouse cell types. These include methods to identify chromatin accessible regions (DNase I hypersensitive site mapping, ATAC-seq) as well as transcription factor occupancy and histone modifications (ChIP-seq, CUT&RUN). While at least some of these 1D assays have become routine, mapping 3D interactions with Hi-C remains relatively difficult and prohibitively costly, and high-resolution contact maps (5 kb resolution, 2 billion read pairs) are still only available for a small number of cell types. This raises the question of whether it is possible to train a model to accurately predict the Hi-C contact matrix from more easily obtained 1D epigenomic data in a cell-type-specific fashion. Such a model could ultimately be used to predict how perturbations in the 1D epigenome—including deletion of TAD boundaries or inactivation of distal regulatory elements—would impact 3D organization.

Initial machine learning methods to predict Hi-C interactions from 1D epigenomic data or DNA sequence took a pairwise approach, treating each interacting or non-interacting pair of genomic bins as an independent training example [10, 11, 12]. For example, HiC-Reg [10] used a random forest regression model to predict the Hi-C contact signal from epigenomic features of the pair of anchoring genomic intervals. Two more recent models, DeepC [13] and Akita [14], respectively predict 'stripes' or submatrices of the Hi-C contact matrix from DNA sequence, capturing the non-independence of interaction bins. Neither method uses epigenomic data as an input signal. DeepC [13] presented a transfer learning framework by pre-training a model to predict epigenomic marks from DNA sequence in order to learn useful local sequence representations, then fine-tuning the model to predict the Hi-C contact map. Akita [14] designed a deep convolutional neural network to predict the Hi-C contact maps of multiple cell types from DNA sequence. These prior studies represent a significant advance in predicting 3D genomic structure, and the DeepC and Akita models demonstrated some success in predicting the impact of sequence perturbations like structural genetic variants on local chromatin folding. However, there are also clear limitations to these approaches. Models that start with DNA sequence need considerable computational resources to extract and propagate useful information from base-pair resolution to megabase scale. More importantly, by learning mappings from only DNA sequence to Hi-C contact map data in the training cell types—and therefore lacking any cell-type-specific feature inputs—the resulting models cannot generalize to new cell types that are not seen in training. In fact, it has also been observed that sequence-based models capture very limited cell-type-specific information about 3D genomic architecture even across the training data and instead predict similar structures in every cell type [14].

Here we propose a novel neural network model called Epiphany to predict the cell-type-specific Hi-C contact map from five commonly generated epigenomic tracks that are already available for a wide number of cell types and tissues: DNase I hypersensitive sites and CTCF, H3K27ac, H3K27me3, and H3K4me3 ChIP-seq. Epiphany uses 1D convolutional layers to learn local representations from the input tracks as well as bidirectional long short term memory (Bi-LSTM) layers to capture long term dependencies along the epigenome and a generative adversarial network (GAN) architecture to encourage realism. One goal of our study is to predict contact maps that are usable for downstream computational analyses such as TAD and interaction calls. To this end, we assessed model performance using multiple normalization and matrix balancing techniques including Knight-Ruiz (KR) [15], iterative correction (ICE) [16], and HiC-DC+ [17] Z-score and observed-over-expected count ratio. Epiphany is trained with a combination of mean-squared error (MSE) and adversarial loss to enhance its ability to produce realistic Hi-C contact maps for downstream analysis. The adversarial loss is calculated using a simultaneously trained GAN-style discriminator network, which distinguishes real contact maps from predicted ones, and helps the model to improve its prediction quality. Epiphany shows robust performance and generalization abilities to held-out chromosomes within and across cell types and species, and its predicted contact matrices yield accurate TAD and significant interaction calls. At inference time, Epiphany can be used to study the contribution of specific epigenomic signals to 3D architecture and to predict the structural changes caused by perturbations of epigenomic signals.

Results

Epiphany: A CNN-LSTM trained with an adversarial loss accurately predicts Hi-C contact maps

Epiphany uses epigenomic signals (DNaseI, CTCF, H3K27ac, H3K27me3, H3K4me3) to predict normalized Hi-C contact maps. Epigenomic signals are extracted at 100bp resolution from normalized .bigWig files without applying a peak calling step. Hi-C contact maps were initially binned at 10 kb resolution and normalized using the HiC-DC+ package [17] to produce Z-scores and observed-over-expected count ratios, Juicer Tools [18] for KR normalization, and HiCExplorer [19] for ICE normalization. The normalization approaches provided by HiC-DC+ are estimated from a negative binomial regression that is estimated directly from count data and adjusts for genomic distance and other covariates.

Epiphany consists of two parts: a generator to extract information and make predictions, and a discriminator to introduce adversarial loss into the training process (**Fig. 1A** and **Methods**). In the generator, we first used a series of convolution modules to featurize epigenomic information in a sliding window fashion. For one output vector, which covers a distance of 1Mb orthogonal to the diagonal, we used a window size of 1.4 Mb centered at the corresponding region as input (**Fig. 1B**). Then a Bi-LSTM layer was employed to capture the dependencies between output vectors, so that a total of 3.4 Mb input were processed in one pass for prediction of 200 output vectors. At the end, a fully connected layer was used to integrate signals and make the final prediction. We also introduced an adversarial loss and a discriminator, which consists of several convolution modules that are applied during training and pushes the generator to produce realistic samples (**Fig. 1C**).

Given the sequential nature of Hi-C contact maps, interactions on consecutive output vectors are unlikely to be independent from one another. We found that Bi-LSTM layers introduce strong dependencies between the output vectors, which allows Epiphany to leverage structures that span multiple genomic positions in Hi-C maps (such as edges of TADs). Furthermore, Bi-LSTM layers overcome the limitation of convolutional neural networks (CNNs) by enabling each output vector to make use of important signals beyond the input window. This is conducive to studying the contribution of distal regulatory elements towards 3D genome structures and reduces the sensitivity of model performance to the choice of window size.

Past approaches that predict the 3D genome structure from 1D inputs use pixel-wise MSE to quantify the similarity between predicted and ground truth Hi-C maps. However, pixel-wise losses for images have been shown by the computer vision community to be overly sensitive to noise [20] and to yield blurry results when used as objectives for image synthesis [21, 22]. In the context of predicting Hi-C maps, MSE loss can over-penalize poor performance on featureless, noisy regions while failing to penalize underestimation of significant interactions. These issues can be mitigated with an adversarial loss, which enables the model to generate highly realistic samples while circumventing the need to explicitly define similarity metrics for complex modalities of data. Thus, Epiphany is trained using a convex combination of MSE loss and adversarial loss. A parameter λ was introduced to balance the proportion of MSE loss and adversarial loss, and the loss function was defined as

$$\min_{\theta^{\mathcal{G}}} \max_{\theta^{\mathcal{D}}} (1-\lambda)\mathcal{L}_{adv}(\theta^{\mathcal{G}}, \theta^{\mathcal{D}}) + \lambda\mathcal{L}_{MSE}(\theta^{\mathcal{G}}) \tag{1}$$

where $\lambda\mathcal{L}_{adv}(\theta^{\mathcal{G}}, \theta^{\mathcal{D}})$ is the adversarial loss, and $\mathcal{L}_{MSE}(\theta^{\mathcal{G}})$ is the MSE between the predicted contact map and ground truth. Intuitively, the MSE loss ensures that the Hi-C maps predicted by Epiphany are aligned with their corresponding epigenomic tracks, while the adversarial loss ensures that the predictions are realistic. We find that using this customized training objective yields Hi-C maps that can be directly processed by commonly used downstream analysis tools.

Epiphany accurately predicts the Hi-C contact map

We first benchmarked the model at 10 kb resolution to compare between two loss functions: MSE only and the convex combination of MSE and adversarial loss. Both losses use the observed-over-expected count ratio normalization based on HiC-DC+. Models were trained on data from the GM12878 ENCODE cell line, with chr3, 11, and 17 as completely held-out chromosomes. Epiphany demonstrates good performance for both the Pearson and Spearman correlation metrics using the observed-over-expected count ratio (**Table 1**), while MSE produced higher correlations than the convex combination of MSE and adversarial loss. However, we observed that the high correlations from MSE trained models were associated with blurriness in the predicted contact maps (**Fig. 2B**), whereas the correlations produced by the combined loss models may have been slightly diminished due to small deviations in the sharper predictions. We also found that downstream algorithms such as TAD or significant loop callers would not function properly on such blurry maps. Therefore, we reasoned that correlation may not be an appropriate evaluation metric and decided to use the combined loss (MSE+adversarial loss) for downstream analysis.

We then tested the robustness of Epiphany with various normalization methods, including KR normalization, ICE normalization, and Z-scores from HiC-DC+. All models were set up with the same training approach as before, where chr3, 11, and 17 were used as held-out chromosomes and models were trained with the combined loss. Epiphany shows robust performance in all normalization methods (**Fig. 2A**). ICE normalization obtained the highest correlations, with an average Pearson correlation of 0.7028 and Spearman correlation of 0.5303 on completely held-out chromosomes (**Table 1**).

Normalization Method	λ	Pearson (all)	Pearson (train)	Pearson (test)	Spearman (all)	Spearman (train)	Spearman (test)
Obs/Exp	0.95	0.7408	0.7687	0.5636	0.6899	0.7191	0.5048
Obs/Exp	1 (MSE only)	0.7833	0.8045	0.6494	0.7381	0.7605	0.5963
Z-score	0.95	0.6881	0.7222	0.4722	0.6695	0.7034	0.4544
KR	0.35	0.7289	0.7510	0.5889	0.5909	0.6135	0.4477
ICE	0.35	0.8108	0.8288	0.7028	0.6631	0.6852	0.5303

Table 1: Mean Pearson and Spearman correlation for different normalization methods

To explore the capacity of Epiphany to capture key structures in genome architecture, we next evaluated the ability of Epiphany predictions to recover TAD boundaries. For all normalization methods and their predictions, we called TAD boundaries using TopDom [23] with window sizes ranging from 10 to 50 (corresponding to 100 kb to 500 kb regions). Because TAD calls depend on the normalization method, we first used KR normalization as the gold standard and compared TAD insulation scores computed on ground truth data on the test chromosomes using different normalization methods. Note that we chose to compare insulation scores rather than TAD boundaries, since the latter relies on finding local extrema in the insulation score signal and therefore can be unstable. Among all these methods, ICE had the highest consistency with KR, followed by Z-scores calculated from HiC-DC+. The observed-over-expected count ratios had the least consistency and showed large variation over the three test chromosomes (**Fig. 2C**, left). We then compared the insulation score calculated from the Epiphany-predicted contact maps trained with different normalization methods vs. the corresponding ground truth on the test chromosomes. ICE showed robust predictions on all test chromosomes, whereas HiC-DC+ observed-over-expected count ratio normalization displayed strong mean performance but had larger variance, especially for larger window sizes. HiC-DC+ Z-scores and KR normalization showed lower consistency between predicted vs. ground truth insulation scores (**Fig. 2C**, right). From a visual comparison of ground truth and predicted contact maps with different normalization approaches, we could see Epiphany consistently predicts accurate TAD structures (**Fig. 2D**). Overall, this analysis suggests that, for accurate prediction of TAD structure, Epiphany trained on ICE normalized contact maps gave the best performance, with HiC-DC+ observed-over-expected count ratio as runner-up.

One advantage of HiC-DC+ normalization is that it readily allows the comparison of significant interactions between predicted contact maps and ground truth. HiC-DC+ [17] fits a negative binomial regression using genomic distance, GC content, mappability and effective bin size based on restriction enzyme sites to estimate the expected read count for each interaction bin, which allows an assessment of significance of the observed count. For convenience, we defined the significant interactions as ground truth Z-scores greater than 2. Significant interactions were called with various thresholds from test chromosomes on predicted contact maps using Z-scores and observed-over-expected count ratio, yielding the ROC curves for each test chromosome (**Fig. 2E**). The average AUC is 0.7639 for the two models, suggesting solid performance at a difficult task.

Epiphany shows robust performance at finer resolution

Due to good overall performance and the ability to directly identify significant interactions, we chose observed-over-expected count ratios rather than Z-score from HiC-DC+ for further analysis. We again trained Epiphany to predict interactions within 1Mb from the diagonal at 5 kb resolution. Epiphany showed robust performance at 5 kb resolution, with an average Pearson correlation of 0.5625 and Spearman correlation of 0.5270. In addition to the distance-dependent correlations, we also used both MSE loss and insulation scores calculated from HiCExplorer [19] to evaluate model performance. Since Epiphany jointly predicts multiple interaction vectors, the model can predict a submatrix of the contact map that covers a 2Mb distance along the diagonal (400 vectors for 5 kb resolution) and up to 1Mb from the diagonal. We calculated the average MSE loss between the predicted submatrix vs. ground truth as well as Pearson correlation between insulation scores calculated from the corresponding submatrices. Results for all 2Mb submatrices from the three held-out chromosomes (chr3, 11, 17, **Fig. 3B**) show that Epiphany displays consistent prediction performance across held-out chromosomes with diverse length and gene densities. In particular, 84.4% (173 out of 203) of submatrices have insulation correlation higher than 0.50. Epiphany showed robust performance in most regions along the genome but sometimes produced inaccurate predictions at regions without clear signals or in low mappability regions. (**Fig. 3A**).

We also compared Epiphany with Akita [14] on common test regions, restricting our evaluation to regions that were held out by Akita and overlap with our test chromosomes. We binned the Hi-C contact map at 5 kb resolution and followed the normalization approaches suggested in the Akita study (Methods). Epiphany was re-trained using the training chromosomes as before (all chromosomes except for chr3, 11, 17) and evaluated on the 42 test regions from Akita's held-out set falling in our test chromosomes. Akita's predictions at 2048 bp resolution were average-pooled to 4096 bp in order to obtain relatively consistent resolution. For each test region, we calculated the Pearson correlation between predicted contact matrices and ground truth for both Akita and Epiphany (**Fig. 3C**). We also visually compared the predictions of Akita and Epiphany with ground truth on the held-out examples (**Fig. 3D**). Quantitatively and qualitatively, both models showed similar performance.

Epiphany predicts cell-type specific 3D structure

Since Epiphany uses epigenomic marks as input, it can potentially generalize to predict cell-type-specific 3D structures in a new cell type. We first compared Epiphany's cell-type-specific predictions with those of Akita, where five different cell types were simultaneously predicted in a multi-task framework. We selected H1ESC and GM12878 from these five cell types for the comparison. Akita's cell-type-specific predictions were directly obtained from its multi-task output. Epiphany was trained on GM12878 and evaluated on H1ESC test chromosomes (chr3, 11, 17) at inference time. We checked the visual comparison of Akita and Epiphany cell-type-specific predictions relative to their respective ground truths and also calculated the absolute difference for ground truth and predictions between the two cell types (**Fig. 4A**). The results suggest that Epiphany, which was trained only on GM12878 data, can generalize to a new cell type and accurately predict the differential structure between cell types based on cell-type-specific 1D epigenomic data. By contrast, the DNA-sequence-based Akita model, although trained on Hi-C/Micro-C data in these and other cell types, largely predicts the same 3D structure in GM12878 and H1ESC.

We next explored the ability of Epiphany to identify the contribution of cell-type-specific epigenomic input features to differential 3D structures using feature attribution. In recent years, feature attribution has become a powerful tool to study the contribution of input features to prediction of a specific output. For each interaction bin in the predicted contact map, we first calculated the saliency score [24], which is a gradient-based attribution on input values. We then calculated the SHAP value [25] with baseline signals equal to zero, which highlights the contribution of epigenomic peaks to a specific output. We compared a region (chr17:70,500,000-73,500,000) with differential interactions between GM12878 and K562 (**Fig. 4B**). Epigenomic signals between chr17:72,000,000-72,500,000 in GM12878 contributed to the prediction of the highlighted interaction, while the absence of signals in K562 input led to the correct prediction of a weak interaction.

Ablation analyses suggests redundancies between 1D inputs

In the previous cell-type-specific analysis, distal H3K4me3 peaks gained importance in the K562 prediction when there were no signals at the anchors of the investigated interaction (**Fig. 4B**). We wondered whether features from different epigenomic tracks could compensate for each other in predicting interactions and more generally whether there exist redundancies between the input tracks.

We performed a feature ablation experiment to address these questions. Instead of including all five epigenomic tracks as input, we re-trained the model with one or several of the tracks completely masked as zero. We reasoned that re-training the model rather than masking a specific input region at test time could better serve our goal. For example, using a model trained on all five input tracks, if we simply masked one important peak from

DNaseI track during test time, we expected that the model would inevitably fail to predict the corresponding interactions. However, if we re-trained the model with the entire DNaseI track masked, we expected the model to identify alternative signals from other tracks during training and potentially retain the ability to predict these interactions.

Indeed, this idea matched our observations from the ablation analysis. We re-trained Epiphany with a) an additional SMC3 ChIP-seq track, b) CTCF track masked as zero, c) DNaseI track masked as zero, d) only CTCF and H3K27ac tracks as training inputs, and compared their predictions with the results using all input (**Fig. 5A**). We found that by removing DNaseI, the model achieved similar performance as using all input tracks. Models with CTCF masked or using only two tracks (CTCF+H3K27ac) showed weaker performance.

As we have seen in previous example (**Fig. 4B**), DNaseI and H3K27ac contributed to the differential predictions between GM12878 and K562 at the region chr17:70,670,000-73,880,000. We therefore compared the prediction for this region using a model trained with all input tracks, without DNaseI, or with CTCF+H3K27ac only (**Fig. 5B**). Epiphany was still able to accurately predict interactions in this region after ablating DNaseI; feature attribution indicated that in place of the DNaseI signal (**Fig. 5B**, grey box), the model gave higher importance to H3K27me3 peaks (purple box) in order to predict the interaction. However, after ablating all signals except for CTCF and H3K27ac, the model failed to find alternative predictive signals and missed the boundary.

Bi-LSTM layers capture the contribution of distal elements

Given the sequential nature of Hi-C contact maps, interactions on consecutive output vectors are unlikely to be independent from one another. We tested whether the Bi-LSTM layers in Epiphany indeed captured the dependencies between the output vectors better than regular convolutional layers. We made predictions using Epiphany with Bi-LSTM layers and compared with a modified Epiphany where the Bi-LSTM layers were replaced by convolutional layers. We also calculated the saliency score and SHAP values for the bin chr17:57,140,000-57,750,000 (**Fig. 5C**). The interaction at chr17:57,140,000-57,750,000 was better predicted by Epiphany with Bi-LSTM layers, and feature attribution showed that one distal peak at around chr17:58,350,000-58,400,000 contributed to the prediction. Compared with regular convolutional layers, Bi-LSTM layers introduce stronger dependencies between the output vectors and overcome the limitation of CNNs by enabling each output vector to make use of important signals beyond the input window.

Epiphany predicts perturbations in 3D architecture

Since Epiphany models the contribution of epigenomic signals to 3D structures, we explored whether Epiphany could predict 3D structural changes caused by perturbations to the epigenome. In particular, we considered examples where structural variations eliminate important epigenomic features. Despang *et al.* [26] studied the TAD fusion caused by deletion of CTCF sites in vivo in the mouse embryonic limb bud at the *Sox9-Kcnj2* locus. They used the promoter capture Hi-C data in the E12.5 mouse limb bud to show the structural changes after deleting major CTCF sites (mm9, GSE78109, GSE125294). In WT TAD structures, *Kcnj2* and *Sox9* are separated into two TADs. After deleting four consecutive CTCF sites within a 15 kb boundary region (C1 site mm9 chr11:111,384,818–111,385,832, C2-C4 site chr11:111,393,908-111,399,229), the TAD boundaries disappeared and the two TADs fused together. When all CTCF binding sites between *Kcnj2* and *Sox9* were deleted, they observed a more complete TAD fusion (**Fig. 6A**). These experiments revealed a TAD fusion caused by the deletion of major CTCF sites at the boundaries and within the TAD.

We then tested Epiphany's ability to predict these structural changes after we perturbed the CTCF input track. Epiphany was trained on data from the human cell line GM12878 and used to make cross-species prediction in the mouse embryo. Epigenomic tracks were downloaded from ENCODE [27] and BioSamples [28] for mouse limb tissue aligned to the mm10 assembly. In the WT prediction, Epiphany predicted a strong boundary separating *Kcnj2* and *Sox9* into two TADs. Upon masking the C1-4 CTCF peaks (mm10 chr11:111,520,000-111,540,000) at the boundary and further masking all CTCF sites, Epiphany predicted behavior consistent with the ground truth experiments, where the two TADs gradually merged together (**Fig. 6B**, top). We further explored the relationship between the CTCF peaks and TAD formation using feature attribution methods (**Fig. 6B**, bottom). The SHAP values are calculated and averaged for the bins in the vertical highlighted dashed lines. We can see that CTCF peaks at the boundary contribute to TAD separation in unperturbed prediction, while with the masked CTCF track, the feature attribution scores focus on more distal regions at the boundaries of the fused TAD.

We also evaluated whether Epiphany could predict structural changes caused by genomic deletions. Yang *et al.* [29] observed upregulation of the *FLT3* gene in acute lymphoblastic leukemia (ALL) patients with a 13q12.2 deletion and attributed this gene expression change to chromatin structural reorganization and enhancer hijacking. *FLT3* was found to be controlled by three regulatory elements in the 13q12.2 region: DS1 (chr13:28100363-28100863), the promoter of *FLT3*; DS2 (chr13:28,135,863-28,140,863); and DS3 (chr13:28,268,863-28,269,363) in the intron of *PAN3*. In normal hematopoietic cells, *FLT3* is primarily controlled by the interaction of DS1 and DS2 due to the separation of two TADs (**Fig. 6C**, top, highlighted with dashed lines), where DS2 overlaps with the TAD boundary, and DS3 is located in a nearby TAD. In patients with the 13q12.2 deletion where DS2 was lost, they observed a fusion of the two nearby TADs and a strengthened long-range interaction between DS1 and DS3.

We simulated this deletion by excising the DS2 region from all epigenomic input tracks and predicted the resulting contact matrix with Epiphany. That is, epigenomic signals in all five input tracks for the corresponding region were deleted, and the up- and downstream tracks were concatenated together. Epiphany predicted TAD structures consistent with reported observations. Before 13q12.2 deletion, Epiphany predicted a small TAD separation between *FLT3* and *PAN3* genes at (chr13:27,600,000-28,600,000) region, consistent with the ground truth in GM12878 (**Fig. 6C**, top and middle). After the deletion, Epiphany shows the fusion of the two TADs and increased interactions between the *FLT3* and *PAN3* gene regions (**Fig. 6C**, bottom).

Discussion

In this study, we developed Epiphany, a neural network model to predict the cell-type specific Hi-C contact map for entire chromosomes up to a fixed genomic distance using commonly generated epigenomic tracks that are already available for diverse cell types and tissues. We showed that Epiphany accurately predicts cell-type-specific 3D genome architecture and shows robust performance for Hi-C different normalization procedures and at different resolutions. Epiphany was able to accurately predict cross-chromosome, cross-cell type and even cross-species 3D genomic structures. From feature ablation and attribution experiments, we showed that Epiphany could be used to interpret the contribution of specific epigenomic signals to local 3D structures. Through in silico perturbations of epigenomic tracks followed by contact map prediction with Epiphany, we were able to accurately predict the cell-type-specific impact of epigenetic alterations and structural variants on TAD organization in previously studied loci.

Although we used five specific epigenomic tracks (DNase I, CTCF, H3K27ac, H3K27me3, H3K4me3) and Hi-C data in this study, we believe that Epiphany could be used as a more general framework to link cell-type-specific epigenomic signals to 3D genomic structures. In the future, we plan to explore different combinations of the epigenomic input tracks to assess their biological and statistical relevance for prediction of 3D structure.

In addition to using the epigenomic information, we also tried to incorporate DNA information into the model. Previous models have used a one-hot encoding of long genomic DNA sequences ($\tilde{1}$Mb), incurring significant computational costs [14, 13]. We therefore tried an alternative strategy of extracting DNA representations from a pre-trained DNABERT model [30], a new method that adapts the state-of-the-art natural language processing model BERT [31] to the setting of genomic DNA. During the pre-training phase, DNA sequences were first truncated to 510 bp length sequences as the 'sentences' and further divided into k-mers as the 'words' of the vocabulary. The model learns the basic syntax and grammar of DNA sequences by self-supervised training to predict randomly masked k-mers within each sentence. After pre-training, each 510 bp sequence was represented by a 768-length numerical vector. However, since Epiphany covers a 3.4 Mb region as input during training, it was still extremely computationally intensive to directly incorporate the pre-trained representations from DNABERT. We therefore excluded the DNABERT component in order to keep the model relatively light-weighted and concise, although we do not rule out its utility in the future.

Beyond these computational issues, a more conceptual modeling challenge is retaining the ability to generalize to new cell types while also incorporating DNA sequence information. In principle, training on genomic sequence may learn DNA sequence features that are specific to the training cell types and do not generalize to other cell types. Epiphany learns a general model for predicting the Hi-C contact map in a cell type of interest from cell-type-specific 1D epigenomic data, giving state-of-the-art prediction accuracy while allowing generalization across cell types and across species.

Methods

Data sources and pre-processing

Training and test sets. We used three human cell lines (GM12878, H1ESc, K562) and one mouse cell line (mES) for training and testing the model. All human data (Hi-C, ChIP-seq) were processed using the hg38 assembly and mouse data with mm10. For all experiments, chromosome 3, 11, 17 were used as completely held-out data for testing.

Epigenomic data. All input epigenomic tracks including DNaseI, CTCF, H3K4me3, H3K27ac, H3K27me3 for genome assembly hg38 were downloaded from the ENCODE data portal [27]. Data were downloaded as bam files, and the replicates for each epigenomic track were merged using the pysam (https://github.com/pysam-developers/pysam) python module. We then converted merged bam files into bigWig files with deepTools [32] bamCoverage (binSize 10, RPGC normalization, other parameters as default). Genome-wide coverage bigWig tracks were later binned into 100-bp bins, and bin-level signals for the 5 epigenomic tracks were extracted as input data for the model.

Hi-C data. High quality and deeply sequenced Hi-C data as .hic format for all human and mouse cell lines were downloaded from 4DN data portal [1]. Data were binned at 5 kb and 10 kb resolution and normalized using multiple approaches. KR normalization was calculated by Juicer tools [18] and ICE normalization by the HiCExplorer package [19]. Observed-over-expected count ratio and Z-score normalizations were calculated by HiC-DC+ [17]. ICE normalization for 5 kb resolution was calculated using Cooler [33], and all additional matrix balancing steps followed the Akita pipeline [14]. For the observed-over-expected count ratios from HiC-DC+, raw counts for interaction bins are modeled using negative binomial regression to estimate a background model, giving an expected count value based on the genomic distance and other covariates associated with the anchor bins (GC content, mappability, effective size due to restriction enzyme site distribution). The observed-over-expected count ratio is then calculated using observed raw counts divided by the expected counts from the HiC-DC+ model.

Biological validation data. Capture Hi-C and corresponding CTCF tracks from Despang *et al.* [26] were downloaded from (GSE78109, GSE125294). Data were visualized using Coolbox [34].

Model and training

CNN layers. The input epigenomic tracks were divided into overlapping windows, with a window length of $m = 14,000$ bins (1.4Mb) and a stride of 1,000 bins (100 kb). We refer to the windowed inputs as $X = \{x_1, ..., x_n\}$, where $x_i \in \mathbb{R}^{c\times m}$ corresponds to window i, n is the total number of windows, and c is the number of epigenomic tracks. A series of four convolution modules were used to featurize each window into a vector of dimension $d = 900$ (after flattening), where each convolution module consists of a convolutional layer with ReLU activation, max pooling, and dropout. We define $Z = \{z_1, ..., z_n\}$ as the flattened output of the final convolution module where $z_i \in \mathbb{R}^d$ is the representation for window x_i.

Bi-LSTM layers. The Bi-LSTM layers receive sequence $Z = \{z_1, ..., z_n\}$ as an input and generate a new sequence $\tilde{Z} = \{\tilde{z}_1, ..., \tilde{z}_n\}$, where $\tilde{z}_i \in \mathbb{R}^{2d}$. To produce the final output, every element of $\tilde{Z}$ is passed through a fully connected layer yielding the output sequence $\hat{Y} = \{\hat{y}_1, ..., \hat{y}_n\}$. Each $\hat{y}_i \in \mathbb{R}^{d'}$ is a vector of dimension $d' = 100$ (or $d' = 200$ if predicting 5 kb resolution Hi-C) and corresponds to a *zig-zag* stripe in a Hi-C matrix, similar to DeepC (shown in **Fig. 1**). Epiphany uses multiple Bi-LSTM layers, with skip connections between successive layers.

Adversarial loss. Generative adversarial networks (GAN) consist of two networks, a generator $\mathcal{G}$ with parameters $\theta^{\mathcal{G}}$ and a discriminator $\mathcal{D}$ with parameters $\theta^{\mathcal{D}}$, that are adversarially trained in a zero-sum game [22, 35]. During training, the generator learns to fool the discriminator by synthesizing realistic samples from a given input, while the discriminator learns to distinguish real samples from synthetic samples. To train Epiphany, we employed a convex combination of pixel-wise MSE and adversarial loss. Given a dataset D and a trade-off parameter λ, Epiphany solves the following optimization problem during training:

$$\min_{\theta^{\mathcal{G}}} \max_{\theta^{\mathcal{D}}} \lambda \mathcal{L}_{adv}(\theta^{\mathcal{G}}, \theta^{\mathcal{D}}) + (1-\lambda)\mathcal{L}_{MSE}(\theta^{\mathcal{G}}) \tag{2}$$

$$\mathcal{L}_{adv}(\theta^{\mathcal{G}}, \theta^{\mathcal{D}}) = \mathbb{E}_{(X,Y)\sim D}\left[\log(\mathcal{D}(Y)) + \log(1 - \mathcal{D}(\mathcal{G}(X)))\right] \tag{3}$$

$$\mathcal{L}_{MSE}(\theta^{\mathcal{G}}) = \mathbb{E}_{(X,Y)\sim D}\left[\sum_{i\in[n]}\sum_{j\in[d']}(Y_{ij} - [\mathcal{G}(X)]_{ij})^2\right], \tag{4}$$

where X corresponds to epigenomic tracks and Y the corresponding Hi-C matrix.

In our framework, $\mathcal{G}$ is the CNN-LSTM architecture described in the previous sections while $\mathcal{D}$ is a simple four layer 2D CNN. Note that in practice, many tricks and heuristics are used in order to speed up convergence when training GANs, as described below.

Training. In Algorithm 1, we show the specific procedure used to approximately solve the optimization problem described above. Note that rather than setting $\mathcal{L}_{\mathcal{D}}$ to $-\mathcal{L}_{\mathcal{G}}$, we employ the target flipping heuristic outlined in [22] (Section 3.2.3) for faster convergence. The parameter updates (lines 5 and 8) are computed via the Adam optimizer [36]. We determine when to conclude training based on when $\mathcal{L}_G$ ceases to decrease.

Algorithm 1 Epiphany Training

```
Require: $\mathcal{G}, \mathcal{D}, \lambda$
 1: while not converged do
 2:     for $(X, Y) \sim D$ do                         ▷ X is the set of tracks, Y is the Hi-C
 3:         $\mathcal{L}_{MSE} \leftarrow MSE(\mathcal{G}(X), Y)$                  ▷ Update generator
 4:         $\mathcal{L}_{adv} \leftarrow -\log(\mathcal{D}(\mathcal{G}(X)))$
 5:         $\mathcal{L}_G \leftarrow \lambda\mathcal{L}_{adv} + (1-\lambda)\mathcal{L}_{mse}$
 6:         Update $\mathcal{G}$ using $\mathcal{L}_G$
 7:
 8:         $\mathcal{L}_D \leftarrow -\log(\mathcal{D}(Y)) - \log(1 - \mathcal{D}(\mathcal{G}(\mathcal{X})))$      ▷ Update discriminator
 9:         Update $\mathcal{D}$ using $\mathcal{L}_D$
10:     end for
11: end while
```

Performance evaluation and application

Model performance. We evaluated the model performance using Pearson and Spearman correlation of the predicted contact map vs. ground truth, computed as a function of genomic distance from the diagonal. Predicted contact maps were saved as .hic files for downstream analysis. We visualized Hi-C matrices and epigenomic tracks using CoolBox [34]. The insulation score was calculated using the TAD-separation score from HiCExplorer [37]. Then a correlation of these scores between ground truth vs. predicted contact maps was calculated. For each 2Mb length submatrix (200 bin matrix), we calculated MSE loss and insulation score correlation between the predicted and true maps.

TAD boundaries and significant interactions. We identified TAD structures and significant interactions in the predicted contact maps vs. ground truth. TAD structures were identified using TopDom [23], with various window sizes of 10, 20, 30, 40, 50. Since the binary TAD boundaries would be less robust towards hyper-parameter selection and small value perturbations on the contact maps, we used insulation score for the comparisons. In this evaluation, we first ran TopDom on all ground truth contact maps with different normalization methods, and used KR normalization as the gold standard, to compare the agreement between these normalization approaches (e.g. ICE vs. KR, Z-score vs. KR). We then compared TopDom results called from predicted contact map vs. the ground truth with their corresponding normalization approach (e.g. predicted Z-score vs. ground truth Z-score), to evaluate Epiphany's ability to predict key structures. These experiments were all run on test chromosomes 3, 11, 17.

HiC-DC+ [17] used interaction bin counts to fit a negative binomial regression with genomic distance, GC content, mappability and effective bin size based on restriction enzyme sites, providing an estimated expected read count for each interaction bins. Z-scores and observed-over-expected count ratios are then computed to evaluate the significance of the observed counts. We defined significant interactions as ground truth Z-score greater than or equal to 2. For test chromosomes 3, 11, 17 with Z-score and observed-over-expected count ratio normalization, we called significant interactions with various cut-off thresholds ranging from 0.5 to 3.5 and plotted the ROC curve.

Comparison with Akita. We followed provided tutorials and extracted the pre-trained Akita model from (Akita repository). Hi-C contact maps were first balanced using ICE normalization, followed by additional steps including adaptive coarse-grain, distance adjustment, rescaling and 2D Gaussian filter suggested by Akita. Test matrices were extracted from Akita held-out test regions that overlapped with Epiphany's test chromosomes (42 regions in total). For calculating the Pearson correlation between the predicted contact map vs. ground truth, we average-pooled Akita matrices from 2048 bp into 4096 bp, in order to keep relative consistency with our 5 kb resolution. For extracting cell-type-specific predictions, we extracted the multi-task output from Akita for H1ESc and GM12878.

Prediction of cell-type-specific structurea. In these experiments, Epiphany was trained on the training chromosomes in GM12878, and tested on test chromosomes chr 3, 11, 17 on H1ESc and K562. Therefore, the predictions were cross-chromosome and cross-cell-type. Feature attributions were calculated using Captum [38], with saliency score to show the gradient attribution on input regions and SHAP values to calculate the contribution of specific epigenomic peaks for predicting 3D structure. The baseline was set to zero when calculating SHAP values.

Feature ablation models. Feature ablation experiments were performed by re-training the model with one or several input epigenomic tracks completely masked as zero. We tested three ablation models: CTCF masked; DNaseI masked; only CTCF and H3K27ac not masked. In addition, we also re-trained Epiphany with an additional SMC3 ChIP-seq track to include cohesin occupancy information. Whole chromosome predictions were generated with trained models and compared to ground truth using Pearson and Spearman correlations as a function of genomic distance. Feature attributions were calculated as described above.

Biological application on mouse data. Epigenomic tracks for mouse limb bud tissue using genome assembly mm10 were downloaded from the ENCODE portal. In the CTCF deletion experiments, CTCF peaks were masked with the average value for the entire CTCF track (masked with the background). Epiphany was trained on the human cell line GM12878 and tested on mouse limb bud data (E11.5 for DNaseI and CTCF tracks, and E12.5 for H3K27ac, H3K27me3 and H3K4me3). Data were visualized using CoolBox [34].

Data Availability

Datasets in this study are all publicly available.

Hi-C data

Hi-C data are available from 4DN data portal. All human data are with hg38 assembly, and mouse with mm10.

GM12878	H1ESc	K562	mES
4DNFI1UEG1HD	4DNFIQYQWPF5	4DNFITUOMFUQ	4DNFI8KBXYNL

Table 2: Hi-C data source

Epigenomic data

Epigenomic tracks for human are available from ENCODE. Cohesin for GM12878 is available at ENCSR000DZP.

Cell type	Dnase I	CTCF	H3K27ac	H3K27me3	H3K4me3
GM12878	ENCSR000EMT	ENCSR000DRZ	ENCSR000DRY	ENCSR000DRX	ENCSR000AKC
H1ESc	ENCSR000EMU	ENCSR000AMF	ENCSR000ANP	ENCSR216OGD	ENCSR019SQX
K562	ENCSR000EOT	ENCSR000DWE	ENCSR000AKP	ENCSR000AKQ	ENCSR000DWD

Table 3: Epigenomic data source

Experimental validation data

Capture Hi-C data and CTCF tracks for mES E12.5 with mm9 assembly for biological validation from Despang *et al.* [26] are publicly available at GSE78109, GSE125294.

Epigenomic tracks for mouse limb validation are available at ENCODE [27] and BioSamples [28]. All data are aligned with mm10.

Track	Cell type	Days	Accession Number	Source
DNaseI	Mouse limb buds	E11.5	ENCSR661HDP	ENCODE
CTCF	Mouse limb buds	E11.5	SAMD00019977	BioSamples
H3K27ac	Mouse limb buds	E12.5	ENCSR737QWV	ENCODE
H3K27me3	Mouse limb buds	E12.5	ENCSR229LTY	ENCODE
H3K4me3	Mouse limb buds	E12.5	ENCSR938MUD	ENCODE

Table 4: Hi-C data source

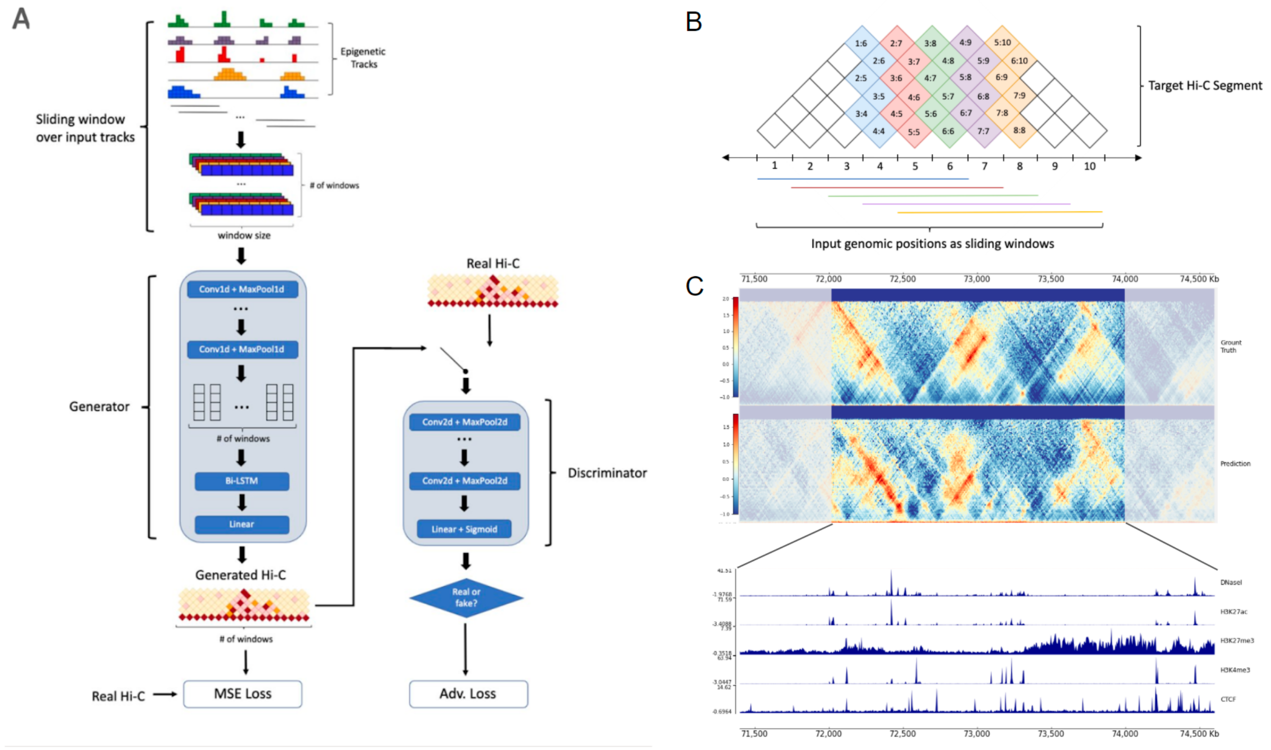

Figure 1: Epiphany employs long short-term memory and adversarial loss to predict the Hi-C contact map. (A) Architecture of Epiphany. Epigenomic signal track are first presented to the model in a sliding window fashion, with window size of 1.4 Mb and step size of 10 kb. During training, we take a total length of 3.4 Mb of the input (200 windows) in one pass. In the generator, the processed input data are first featurized by convolution modules, followed by a Bi-LSTM layer to capture the dependencies between nearby bins. After a fully connected layer, the predicted contact map is generated. An MSE loss between the predicted map and the ground truth is calculated in order to train the generator to predict correct structures. To mitigate the overly-smoothed predictions by the pixel-wise losses, we further introduced a discriminator and adversarial loss. The discriminator consists of several convolution modules, and an adversarial loss was calculated to enable the model to generate highly realistic samples. We trained Epiphany with a combined loss of these two components. **(B)** An illustration of prediction scheme. The first window of input data (blue horizontal line, 1.4 Mb) is used to predict a vector on the Hi-C contact map that is orthogonal to the diagonal (blue bin vector, covers 1Mb from the diagonal). During training, a 3.4 Mb length of input are processed using sliding windows (200 windows) in one pass, and 200 consecutive vectors are being predicted. **(C)** An example region of input epigenomic tracks (bottom), target Hi-C map (top row), and predicted Hi-C map (second row).

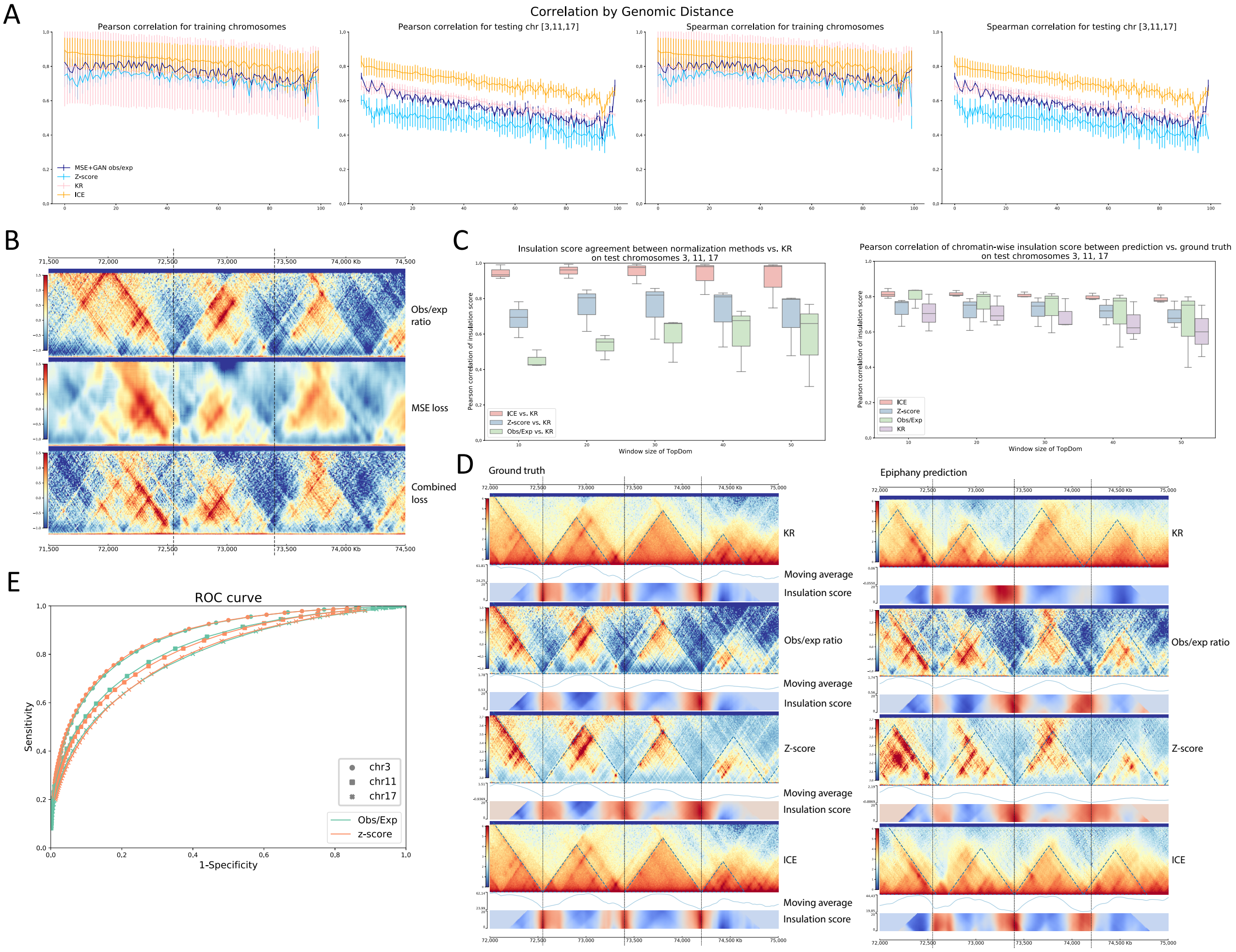

Figure 2: Epiphany-predicted contact maps identify TADs and significant interactions. (A) Epiphany performance using correlation by genomic distance for different normalization approaches. From left to right: Pearson correlation on training chromosomes, on testing chromosomes (chr3, 11, 17), and Spearman correlation on training and on testing chromosomes. Dark blue shows the performance of HiC-DC+ observed-over-expected count ratio, light blue shows HiC-DC+ Z-score, pink shows KR normalization, and yellow shows ICE normalization. **(B)** A visual comparison of the ground truth contact map (chr17:70,670,000-73,880,000, top row), blurry prediction made by MSE trained model (middle row), and more realistic prediction by combined loss (bottom row). **(C)** Top: Agreement of insulation score between different normalization methods vs. KR normalization on test chromosomes. Insulation scores were calculated using TopDom with different window sizes (X-axis) on ground truth contact maps with different normalization methods. KR normalization was used as the gold standard, and a Pearson correlation (Y-axis) was calculated to measure the agreement between each normalization method vs. KR (red: ICE vs. KR, blue: HiC-DC+ Z-score vs. KR, green: HiC-DC+ obs/exp vs. KR). Bottom: Pearson correlation of insulation score between predicted contact map vs. corresponding ground truth of the same normalization (red: ICE, blue: HiC-DC+ Z-score, green: HiC-DC+ obs/exp, purple: KR). **(D)** Left: Ground truth contact maps of different normalization methods (from top to bottom: ICE, HiC-DC+ obs/exp, HiC-DC+ Z-score, KR). Blue dashed lines denotes the TAD calls with window size of 50 on each contact map, and black dashed lines are the TAD boundaries called from KR normalized contact map. Right: Predicted contact maps of different normalization. **(E)** ROC curve of significant interactions between prediction contact maps vs. ground truth for the three test chromosomes for HiC-DC+ obs/exp ratios (green) and for Z-scores (orange).

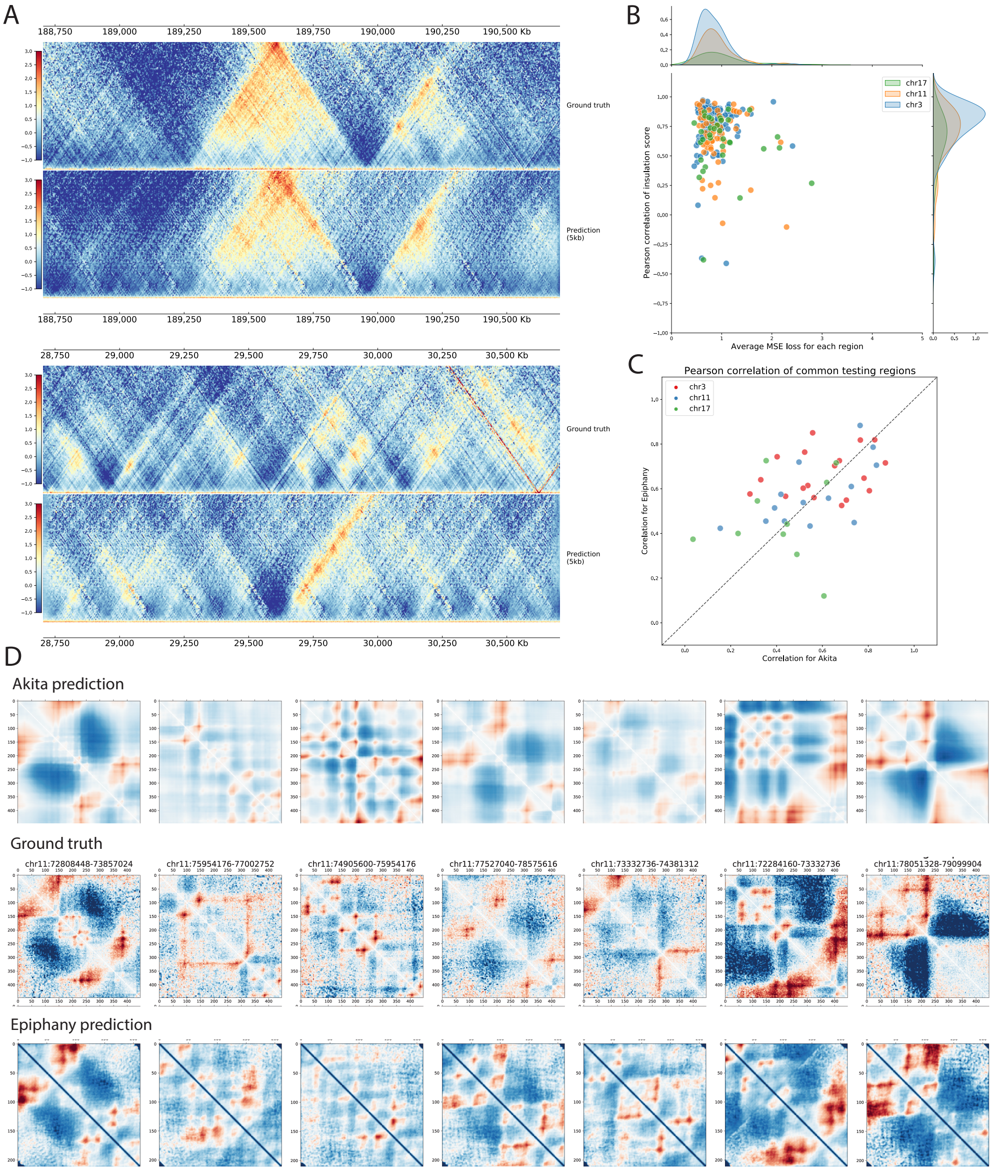

Figure 3: Epiphany achieves state-of-the-art performance at fine resolution. (A) Epiphany performance evaluation at 5 kb resolution. Top: one of the best predicted submatrices (chr3:188,610,000-190,610,000) with ground truth matrix on the top, and predicted matrix on the bottom. Bottom: one of the problematic matrices (chr17:28,705,000-30,705,000) predicted by Epiphany. **(B)** Evaluation of predicted submatrices. X-axis denotes the average MSE loss between predicted matrix and ground truth, and Y-axis shows the Pearson correlation of insulation score of the 2 Mb region. Dots are colored by chromosomes, and density plots for dot distribution are added on the side. **(C)** Model performance comparison between Epiphany and Akita on 42 common regions between Akita held-out test regions and our test chromosomes (chr3, 11, 17). X-axis shows the Pearson correlation of Akita prediction vs. ground truth, and Y-axis shows the correlation of Epiphany. Epiphany was re-trained using data with the same normalization steps of Akita at 5 kb resolution, and Akita predictions were average-pooled into 4096 bp resolution for better comparison. Dots are colored by chromosomes. **(D)** Visual comparison of Akita prediction (2048 bp resolution, top row), ground truth matrices (2048 bp resolution, middle row), and Epiphany prediction (5 kb, bottom row).

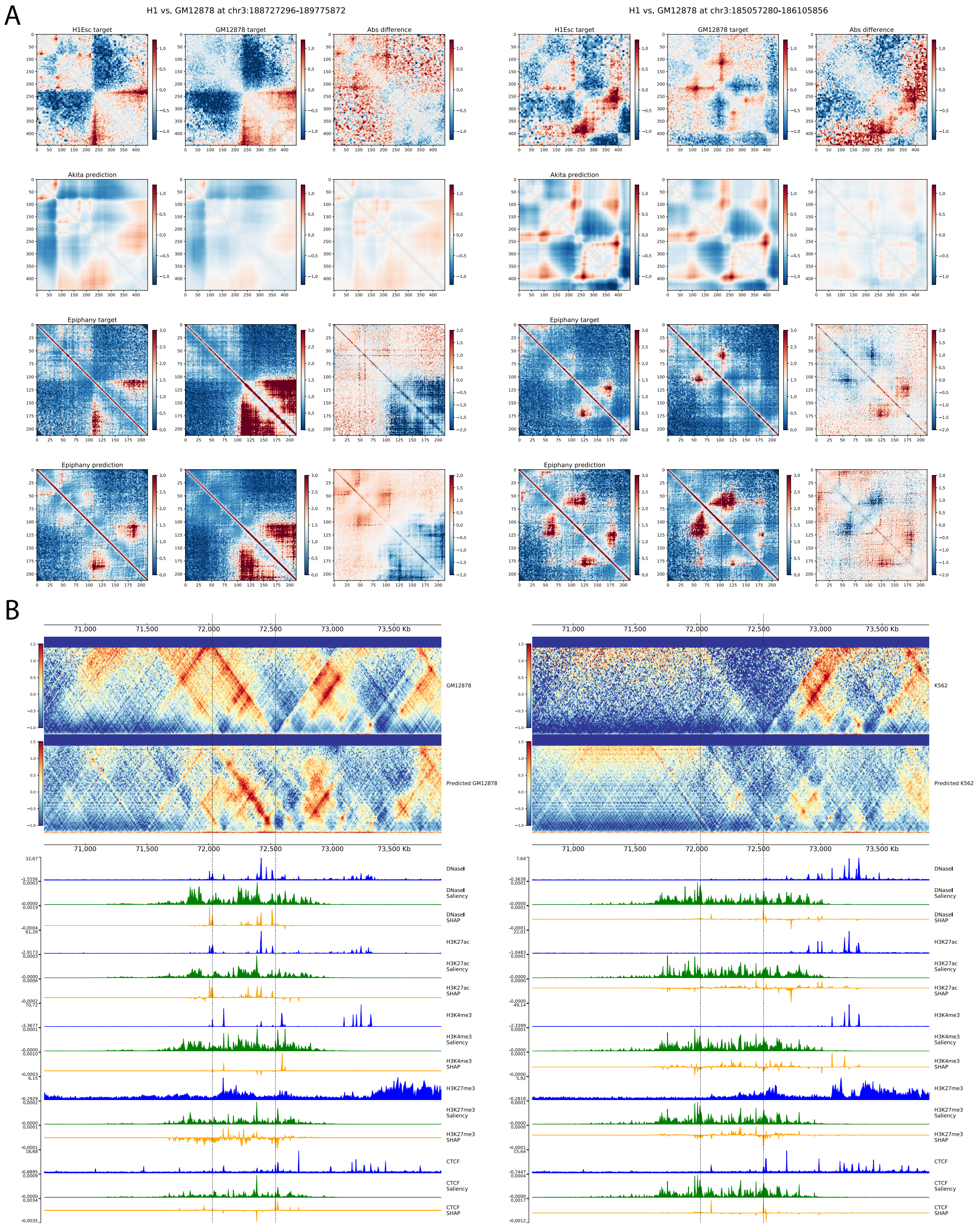

Figure 4: Epiphany accurately predicts cell-type-specific 3D structures. (A) Two examples (chr3:188,727,296-189,775,872) and (chr3:185,057,280-186,105,856) of cell-type specific predictions in H1ESc and GM12878. Two regions are selected from the overlapped region of Akita held-out test set and Epiphany's test chromosomes. Columns from left to right: contact map in H1ESc, same region in GM12878, and the absolute difference between the two cell types (H1ESc-GM12878). Rows from top to bottom: Ground truth matrices with Akita normalization, Akita prediction, ground truth with HiC-DC+ observed-over-expected count ratio, Epiphany prediction of observed-over-expected count ratio. Akita predictions were obtained from the multi-task output, and Epiphany predictions were generated with model trained on GM12878. **(B)** Cell type specific prediction at a differential region between GM12878 and K562. On the left is the ground truth matrix (top) and predicted matrix (middle), followed by epigenomic input tracks (blue), Saliency score (green), and SHAP values (yellow) for feature attributions. On the right is the predictions for K562. Epiphany was trained in GM12878 in training chromosomes, and predicted both cell types for test chromosomes.

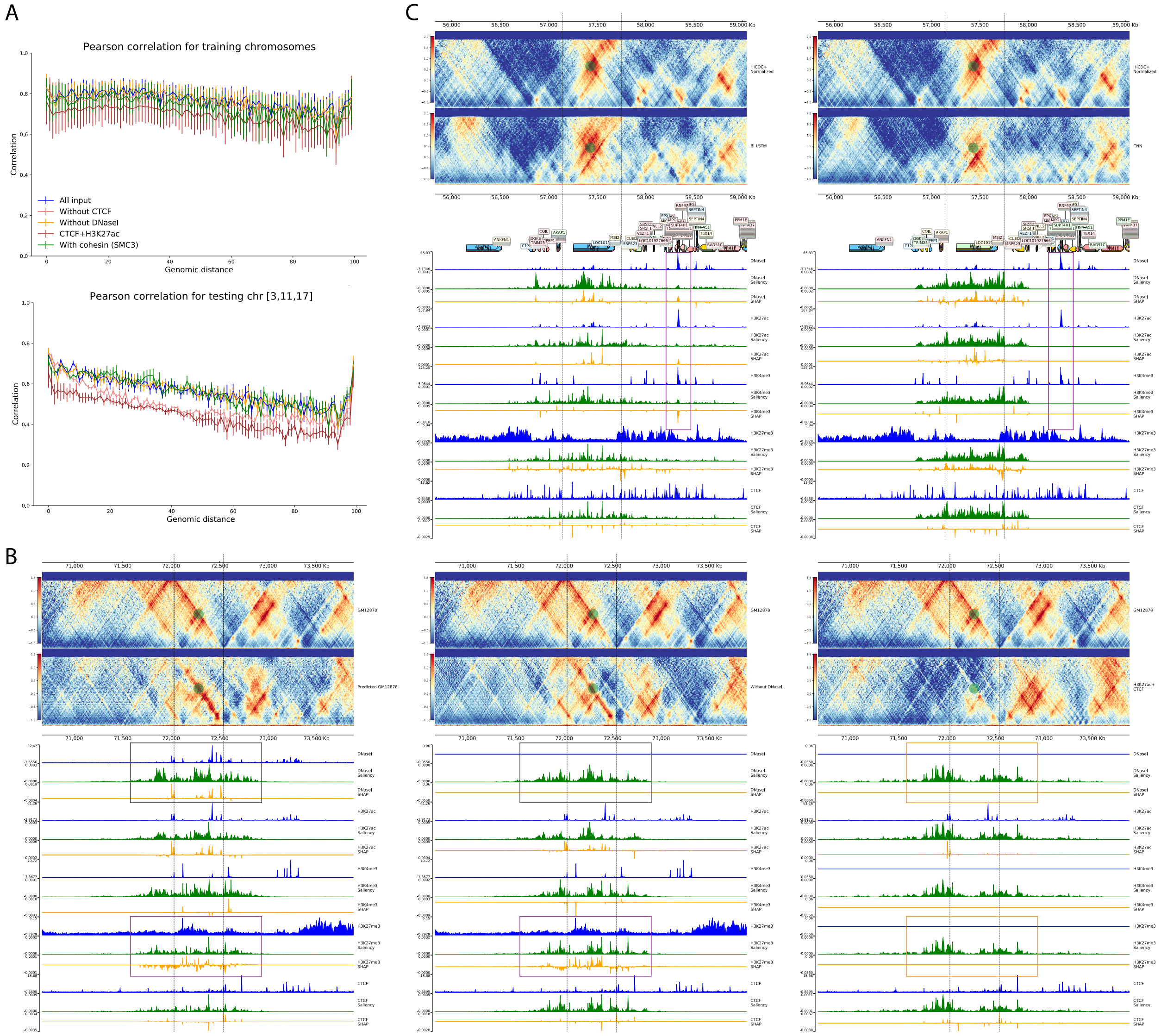

Figure 5: Feature ablation and attribution identify the contribution of epigenomic marks to 3D structure. (A) Correlation by distance for feature ablation experiments. Top: Pearson correlation for training chromosomes. Bottom: Pearson correlation for test chromosomes. Blue track for model performance using all 5 epigenomic input tracks; green for training with an additional track SMC3; pink for model trained without CTCF; yellow for without DNaseI, and dark red for model trained with only CTCF and H3K27ac tracks. **(B)** Feature attribution for bin (chr17:72,030,000-72,540,000) with full model (left), model without DNaseI (middle), model with only CTCF and H3K27ac (right). **(C)** Feature attribution for bin (chr17:57,140,000-57,750,000) with Epiphany with Bi-LSTM layer (left) vs. modified Epiphany with convolution layer (right).

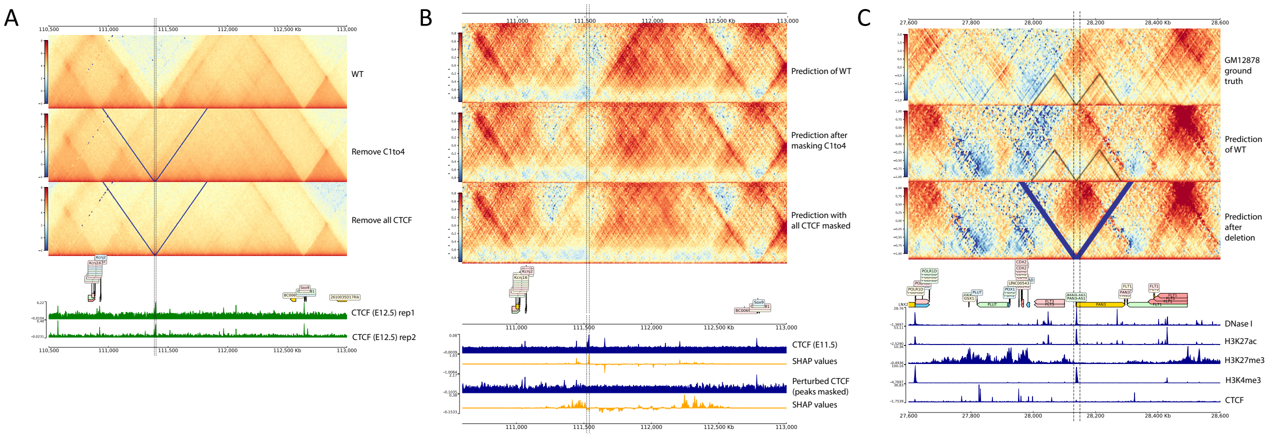

Figure 6: Epiphany predicts TAD boundary changes due to epigenomic perturbations. (A) Mouse ES E12.5 Capture Hi-C data from Despang *et al.* [26] for WT (top), 4 CTCF sites depletion (middle), and all CTCF depletion between gene *Kcnj2* and *Sox9* (bottom). Data are publicly available at (GSE78109, GSE125294). Four CTCF sites depleted in the middle figure were at region (C1 site(mm9 chr11:111,384,818–111,385,832), C2-C4 site (chr11:111,393,908-111,399,229), marked with black dashed lines). Data are mapped relative to mm9. **(B)** Epiphany cross-species prediction of structural changes caused by CTCF perturbation. Epiphany was trained using human cell line GM12878, and predicted using mouse limb bud epigenomic data mapped relative to the mm10 assembly. The panel shows Epiphany prediction of WT mES Hi-C map with HiC-DC+ obs/exp ratio normalization (top row), the prediction of TAD fusion after masking CTCF sites at (mm10 chr11:111,520,000-111,540,000) (middle row), and the prediction of further TAD fusion after masking all CTCF peaks between *Kcnj2* and *Sox9* genes (bottom row). Epigenomic tracks at the bottom are showing feature attribution (SHAP value) for highlighted vertical vector in the Hi-C contact map. The upper two tracks are the original CTCF track and corresponding SHAP values. The lower two tracks are CTCF tracks with peaks between *Kcnj2* and *Sox9* masked to the background and the corresponding SHAP values. **(C)** Human GM12878 Hi-C contact map at 5 kb resolution at chr13:27,600,000-28,600,000. Ground truth contact map (top), predicted contact map with unperturbed input (middle), predicted contact map with 20 kb deletion at region chr13:28,130,000-28,150,000 (bottom, deleted region highlighted in dashed line).

Supplementary Information

Model Architecture Details

In the following tables, we describe in detail Epiphany's architecture. In each table, n is the total number of Hi-C stripes we seek to predict.

Operation	Number of Filters	Filter Size	Stride	Activation	Output Shape
Input	-	-	-	-	n x 5 x 14000
Convolution	70	5 x 17	1	ReLU	n x 70 x 13984
Max Pool	1	1 x 4	1	-	n x 70 x 3496
Dropout (p = .1)	-	-	-	-	n x 70 x 3496
Convolution	90	70 x 7	1	ReLU	n x 90 x 3490
Max Pool	1	1 x 4	1	-	n x 90 x 872
Dropout (p = .1)	-	-	-	-	n x 90 x 872
Convolution	70	90 x 5	1	ReLU	n x 70 x 868
Max Pool	1	1 x 4	1	-	n x 70 x 217
Dropout (p = .1)	-	-	-	-	n x 70 x 217
Convolution	20	70 x 5	1	ReLU	n x 20 x 213
Adaptive Max Pool	1	-	1	-	n x 20 x 45
Dropout (p = .1)	-	-	-	-	n x 20 x 45
Flatten	-	-	-	-	n x 900

Table 5: Parameterization for 1D CNN for 5kb and 10kb Hi-C prediction

Operation	Hidden Layer Size	Activation	Output Shape
Bi-LSTM	1200	ReLU	n x 2400
Bi-LSTM	1200	ReLU	n x 2400
Bi-LSTM	2400	ReLU	n x 2400
Dense	-	ReLU	n x 900
Dense	-	None	n x 100

Table 6: Parameterization for Bi-LSTM for 10kb Hi-C prediction

Operation	Hidden Layer Size	Activation	Output Shape
Bi-LSTM	2400	ReLU	n x 4800
Bi-LSTM	2400	ReLU	n x 4800
Dense	-	ReLU	n x 1200
Dense	-	None	n x 200

Table 7: Parameterization for Bi-LSTM for 5kb Hi-C prediction

References

[1] Dekker, J., Belmont, A.S., Guttman, M., Leshyk, V.O., Lis, J.T., Lomvardas, S., Mirny, L.A., O'shea, C.C., Park, P.J., Ren, B., *et al.*: The 4D nucleome project. Nature **549**(7671), 219–226 (2017)

[2] Zheng, H., Xie, W.: The role of 3D genome organization in development and cell differentiation. Nature Reviews Molecular Cell Biology **20**(9), 535–550 (2019)

[3] Lieberman-Aiden, E., Van Berkum, N.L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B.R., Sabo, P.J., Dorschner, M.O., Sandstrom, R., Bernstein, B., Bender, M.A., Groudine, M., Gnirke, A., Stamatoyannopoulos, J., Mirny, L.A., Lander, E.S., Dekker, J.: Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science **326**(5950), 289–293 (2009)

[4] Hsieh, T.-H.S., Weiner, A., Lajoie, B., Dekker, J., Friedman, N., Rando, O.J.: Mapping nucleosome resolution chromosome folding in yeast by Micro-C. Cell **162**(1), 108–119 (2015)

[5] Mumbach, M.R., Rubin, A.J., Flynn, R.A., Dai, C., Khavari, P.A., Greenleaf, W.J., Chang, H.Y.: HiChIP: efficient and sensitive analysis of protein-directed genome architecture. Nature Methods **13**(11), 919–922 (2016)

[6] Fullwood, M.J., Liu, M.H., Pan, Y.F., Liu, J., Xu, H., Mohamed, Y.B., Orlov, Y.L., Velkov, S., Ho, A., Mei, P.H., Chew, E.G., Huang, P.Y., Welboren, W.J., Han, Y., Ooi, H.S., Ariyaratne, P.N., Vega, V.B., Luo, Y., Tan, P.Y., Choy, P.Y., Wansa, K.D., Zhao, B., Lim, K.S., Leow, S.C., Yow, J.S., Joseph, R., Li, H., Desai, K., Thomsen, J., Lee, Y., Karuturi, R., Herve, T., Bourque, G., Stunnenberg, H.G., Ruan, X., Cacheux-Rataboul, V., Sung, W.K., Liu, E.T., Wei, C.L., Cheung, E., Ruan, Y.: An oestrogen-receptor-α-bound human chromatin interactome. Nature **462**(7269), 58–64 (2009)

[7] Krijger, P.H.L., De Laat, W.: Regulation of disease-associated gene expression in the 3d genome. Nature Reviews Molecular Cell Biology **17**(12), 771 (2016)

[8] Shen, Y., Yue, F., McCleary, D.F., Ye, Z., Edsall, L., Kuan, S., Wagner, U., Dixon, J., Lee, L., Lobanenkov, V.V., *et al.*: A map of the cis-regulatory sequences in the mouse genome. Nature **488**(7409), 116–120 (2012)

[9] Javierre, B., Burren, O., Wilder, S., Kreuzhuber, R., Hill, S., Sewitz, S., Cairns, J., Wingett, S., Varnai, C., Thiecke, M., Burden, F., Farrow, S., Cutler, A., Rehnström, K., Downes, K., Grassi, L., Kostadima, M., Freire-Pritchett, P., Wang, F., The BLUEPRINT Consortium, Stunnenberg, H., Todd, J., Zerbino, D., Stegle, O., Ouwehand, W., Frontini, M., Wallace, C., Spivakov, M., Fraser, P.: Lineage-specific genome architecture links enhancers and non-coding disease variants to target gene promoters. Cell **167**(5), 1369–138419 (2016). doi:10.1016/j.cell.2016.09.037

[10] Zhang, S., Chasman, D., Knaack, S., Roy, S.: In silico prediction of high-resolution Hi-C interaction matrices. Nature Communications **10**(1), 1–18 (2019)

[11] Zhang, S., Chasman, D., Knaack, S., Roy, S.: In silico prediction of high-resolution hi-c interaction matrices. Nature Communications **10**(1), 1–18 (2019)

[12] Trieu, T., Martinez-Fundichely, A., Khurana, E.: DeepMILO: a deep learning approach to predict the impact of non-coding sequence variants on 3D chromatin structure. Genome Biology **21**(1), 1–11 (2020)

[13] Schwessinger, R., Gosden, M., Downes, D., Brown, R.C., Oudelaar, A.M., Telenius, J., Teh, Y.W., Lunter, G., Hughes, J.R.: DeepC: predicting 3D genome folding using megabase-scale transfer learning. Nature Methods **17**(11), 1118–1124 (2020)

[14] Fudenberg, G., Kelley, D.R., Pollard, K.S.: Predicting 3D genome folding from dna sequence with akita. Nature Methods **17**(11), 1111–1117 (2020)

[15] Knight, P.A., Ruiz, D.: A fast algorithm for matrix balancing. IMA Journal of Numerical Analysis **33**(3), 1029–1047 (2012). doi:10.1093/imanum/drs019. https://academic.oup.com/imajna/article-pdf/33/3/1029/1876772/drs019.pdf

[16] Imakaev, M., Fudenberg, G., McCord, R.P., Naumova, N., Goloborodko, A., Lajoie, B.R., Dekker, J., Mirny, L.A.: Iterative correction of Hi-C data reveals hallmarks of chromosome organization. Nature Methods **9**(10), 999–1003 (2012). doi:10.1038/nmeth.2148. Accessed 2021-11-23

[17] Sahin, M., Wong, W., Zhan, Y., Van Deynze, K., Koche, R., Leslie, C.S.: HiC-DC+: systematic 3D interaction calls and differential analysis for Hi-C and HiChIP. bioRxiv (2020)

[18] Durand, N.C., Shamim, M.S., Machol, I., Rao, S.S., Huntley, M.H., Lander, E.S., Aiden, E.L.: Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. Cell Systems **3**(1), 95–98 (2016)

[19] Ramírez, F., Bhardwaj, V., Arrigoni, L., Lam, K.C., Grüning, B.A., Villaveces, J., Habermann, B., Akhtar, A., Manke, T.: High-resolution TADs reveal DNA sequences underlying genome organization in flies. Nature Communications **9**(1), 189 (2018). doi:10.1038/s41467-017-02525-w. Accessed 2021-09-17

[20] Zhao, H., Gallo, O., Frosio, I., Kautz, J.: Loss functions for image restoration with neural networks. IEEE Transactions on Computational Imaging **3**(1), 47–57 (2017). doi:10.1109/TCI.2016.2644865

[21] Xu, X., Sun, D., Pan, J., Zhang, Y., Pfister, H., Yang, M.-H.: Learning to super-resolve blurry face and text images. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 251–260 (2017)

[22] Goodfellow, I.: NIPS 2016 tutorial: Generative adversarial networks. arXiv preprint arXiv:1701.00160 (2016)

[23] Shin, H., Shi, Y., Dai, C., Tjong, H., Gong, K., Alber, F., Zhou, X.J.: TopDom: an efficient and deterministic method for identifying topological domains in genomes. Nucleic Acids Research **44**(7), 70 (2016). doi:10.1093/nar/gkv1505. Accessed 2021-09-16

[24] Simonyan, K., Vedaldi, A., Zisserman, A.: Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. arXiv:1312.6034 [cs] (2014). arXiv: 1312.6034. Accessed 2021-11-25

[25] Lundberg, S., Lee, S.-I.: A Unified Approach to Interpreting Model Predictions. arXiv:1705.07874 [cs, stat] (2017). arXiv: 1705.07874. Accessed 2021-11-25

[26] Despang, A., Schöpflin, R., Franke, M., Ali, S., Jerković, I., Paliou, C., Chan, W.-L., Timmermann, B., Wittler, L., Vingron, M., Mundlos, S., Ibrahim, D.M.: Functional dissection of the Sox9–Kcnj2 locus identifies nonessential and instructive roles of TAD architecture. Nature Genetics **51**(8), 1263–1271 (2019). doi:10.1038/s41588-019-0466-z. Accessed 2021-10-31

[27] Davis, C.A., Hitz, B.C., Sloan, C.A., Chan, E.T., Davidson, J.M., Gabdank, I., Hilton, J.A., Jain, K., Baymuradov, U.K., Narayanan, A.K., *et al.*: The Encyclopedia of DNA elements (encode): data portal update. Nucleic Acids Research **46**(D1), 794–801 (2018)

[28] Courtot, M., Cherubin, L., Faulconbridge, A., Vaughan, D., Green, M., Richardson, D., Harrison, P., Whetzel, P.L., Parkinson, H., Burdett, T.: BioSamples database: an updated sample metadata hub. Nucleic Acids Research **47**(D1), 1172–1178 (2018). doi:10.1093/nar/gky1061. https://academic.oup.com/nar/article-pdf/47/D1/D1172/27437327/gky1061.pdf

[29] Yang, M., Safavi, S., Woodward, E.L., Duployez, N., Olsson-Arvidsson, L., Ungerbäck, J., Sigvardsson, M., Zaliova, M., Zuna, J., Fioretos, T., Johansson, B., Nord, K.H., Paulsson, K.: 13q12.2 deletions in acute lymphoblastic leukemia lead to upregulation of FLT3 through enhancer hijacking. Blood **136**(8), 946–956 (2020). doi:10.1182/blood.2019004684. Accessed 2021-11-14

[30] Ji, Y., Zhou, Z., Liu, H., Davuluri, R.V.: DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome. Bioinformatics **37**(15), 2112–2120 (2021). doi:10.1093/bioinformatics/btab083. https://academic.oup.com/bioinformatics/article-pdf/37/15/2112/39622303/btab083.pdf

[31] Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805 [cs] (2019). arXiv: 1810.04805. Accessed 2021-11-10

[32] Ramírez, F., Ryan, D.P., Grüning, B., Bhardwaj, V., Kilpert, F., Richter, A.S., Heyne, S., Dündar, F., Manke, T.: deepTools2: a next generation web server for deep-sequencing data analysis. Nucleic Acids Research **44**(W1), 160–165 (2016). doi:10.1093/nar/gkw257. Accessed 2021-10-28

[33] Abdennur, N., Mirny, L.A.: Cooler: scalable storage for Hi-C data and other genomically labeled arrays. Bioinformatics (2019). doi:10.1093/bioinformatics/btz540

[34] Xu, W., Zhong, Q., Lin, D., Li, G., Cao, G.: CoolBox: A flexible toolkit for visual analysis of genomics data. bioRxiv (2021)

[35] Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A.C., Bengio, Y.: Generative adversarial nets. In: Advances in Neural Information Processing Systems (2014)

[36] Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: International Conference on Learning Representations (ICLR) (2015)

[37] Wolff, J., Rabbani, L., Gilsbach, R., Richard, G., Manke, T., Backofen, R., Grüning, B.A.: 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), 177–184 (2020)

[38] Kokhlikyan, N., Miglani, V., Martin, M., Wang, E., Alsallakh, B., Reynolds, J., Melnikov, A., Kliushkina, N., Araya, C., Yan, S., et al.: Captum: A unified and generic model interpretability library for pytorch. arXiv preprint arXiv:2009.07896 (2020)