# Enhancing Hi-C contact matrices for loop detection with Capricorn, a multi-view diffusion model

Tangqi Fang*[1], Yifeng Liu*[1], Addie Woicik*[1], Minsi Lu[1], Anupama Jha[2], Xiao Wang[3], Gang Li[2,4], Borislav Hristov[2], Zixuan Liu[1], Hanwen Xu[1], William S. Noble†[1,2], and Sheng Wang†[1]

[1]Paul G. Allen School of Computer Science and Engineering, University of Washington, WA
[2]Department of Genome Sciences, University of Washington, WA
[3]Department of Computer Science, Purdue University, IN
[4]eScience Institute, University of Washington, USA

**Abstract**

High-resolution Hi-C contact matrices measure the detailed three-dimensional architecture of the genome, but high-resolution experimental Hi-C data are expensive to generate and relatively rare. Computational methods to enhance low-resolution contact matrices exist but are largely based on resolution enhancement methods for natural images and hence often employ models that do not distinguish between biologically meaningful contacts and background contacts. We present Capricorn, a tool for Hi-C resolution enhancement that incorporates high-order chromatin features as additional views of the input Hi-C contact matrix and leverages a diffusion probability model backbone to generate a high-resolution matrix. We show that Capricorn outperforms the state-of-the-art in a cross-cell-line setting, improving existing methods by 17.8% in mean-squared error and 22.9% in F1 score for loops called from the generated high-resolution data. We also show that Capricorn performs well in the cross-chromosome setting, again improving the downstream loop F1 score by 15.7% relative to existing methods. Capricorn's implementation and source code are freely available at https://github.com/CHNFTQ/Capricorn.

**Keywords:** Hi-C resolution enhancement, diffusion probability model, chromatin loop calling

*TF, YL, and AW contributed equally to this work. Co-first authors can prioritize their names when adding this reference to their resumes.

†Corresponding authors. Emails: william-noble@uw.edu; swang@cs.washington.edu

| Method | Publication date | Pixel-wise loss | Adversarial loss | TV loss | Perceptual loss | Includes biology |
|---|---|---|---|---|---|---|
| HiCPlus [44] | Feb. 2018 | ✓ | | | | |
| hicGAN [24] | July 2019 | | ✓ | | | |
| HiCNN2 [26] | Oct. 2019 | ✓ | | | | |
| HiCNN [25] | Nov. 2019 | ✓ | | | | |
| DeepHiC [19] | Feb. 2020 | ✓ | ✓ | ✓ | ✓ | |
| SRHiC [22] | Apr. 2020 | ✓ | | | | |
| HiCSR [9] | July 2020 | ✓ | ✓ | | ✓ | |
| VEHiCLE [17] | Apr. 2021 | ✓ | ✓ | | | ✓ |
| EnHiC [20] | July 2021 | ✓ | ✓ | | ✓ | |
| HiCARN-1 [16] | Apr. 2022 | ✓ | | ✓ | ✓ | |
| HiCARN-2 [16] | Apr. 2022 | ✓ | ✓ | ✓ | ✓ | |
| Capricorn | (ours) | ✓ | | | | ✓ |

Table 1: **A summary of existing Hi-C resolution enhancement methods.** We catalogue whether each method adopts common computer vision loss terms, with "TV" for total variation and where pixel-wise losses include mean squared error (MSE) and L1 loss. Notably, only VEHiCLE [17] and Capricorn incorporate biological features into the resolution enhancement model. Although VEHiCLE uses TAD identification in its loss, Capricorn is trained to enhance TAD and loop features along with the contact matrix. We limit the comparison to methods for Hi-C resolution enhancement for proximal contacts (interaction within 2 megabases) that do not require additional high-resolution input data, excluding distal-contact enhancement techniques like BoostHiC [5] and reference-based methods like RefHiC-SR [45].

# 1 Introduction

Chromosomes encode genetic and epigenetic cellular programs, leveraging a complex three-dimensional (3D) genome architecture in eukaryotic cells that is critical for many biological processes [3, 8], including modulating gene regulatory relationships, RNA splicing sites, and DNA repair mechanisms. These architectures within a cell can be measured with assays such as high-throughput chromosome conformation capture (Hi-C) [23], genome architecture mapping (GAM) [1], split-pool recognition of interactions by tag extension (SPRITE) [30], and HiChIP [27]. *High-resolution* Hi-C datasets employ smaller bin sizes, thereby grouping genomic loci to more precisely characterize genomic substructures, but consequently require more experimental reads in order to produce a suitably dense matrix. Importantly, doubling the resolution requires quadrupling the experiment read counts due to the pairwise nature of interactions [40].

In order to obtain denser contact matrices, computational *resolution enhancement* methods take low-resolution Hi-C matrices with fewer measured contacts and generate the corresponding higher-resolution matrices. Existing approaches adopt techniques from computer vision, such as convolutional networks [22, 25, 26, 44] and generative adversarial networks (GAN)[9, 16, 17, 19, 24] to produce detailed contact maps. The resulting data can then be used to perform analyses of genome folding, including A/B compartment classification [6, 12, 14, 15, 23], topologically associating domain (TAD) identification [7, 10, 11, 32, 41], and loop calling [33, 34, 36, 39, 43]. Although existing approaches produce lo mean-squared error (MSE) predictions and perform well according to metrics developed for analysis of natural images, we claim that these metrics do not accurately distinguish biologically significant contacts from background contacts in Hi-C data.

We hypothesize that resolution enhancement can produce contact matrices that can better capture these higher-order chromatin structures if we design a loss function that explicitly models structures like loops and TADs during resolution enhancement (Table 1). Such an approach can both provide additional supervision for the resolution enhancement task and help teach the model to distinguish and enhance particularly interesting genomic contacts. We, therefore, propose Capricorn, which incorporates additional biological views of the contact matrix to emphasize important chromatin interactions and leverages powerful computer vision diffusion models [37] for the model backbone (Figure 1). In particular, Capricorn learns a diffusion model that enhances a five-channel image, containing both the primary Hi-C matrix as well as representations of TADs, loops, and distance-normalized counts computed from the original low-resolution matrix. Capricorn thereby learns meaningful structural contacts as well as the general Hi-C matrix distribution.

We compare Capricorn to four existing Hi-C resolution enhancement approaches representing the state-of-the-art, and we find that Capricorn outperforms the others in terms of both MSE and its ability to detect loops from the predicted high-resolution contact matrix. We tested the models' generalizability across both cell line and chromosome and found that Capricorn's enhanced matrices had a 17.8% lower MSE and 22.9%

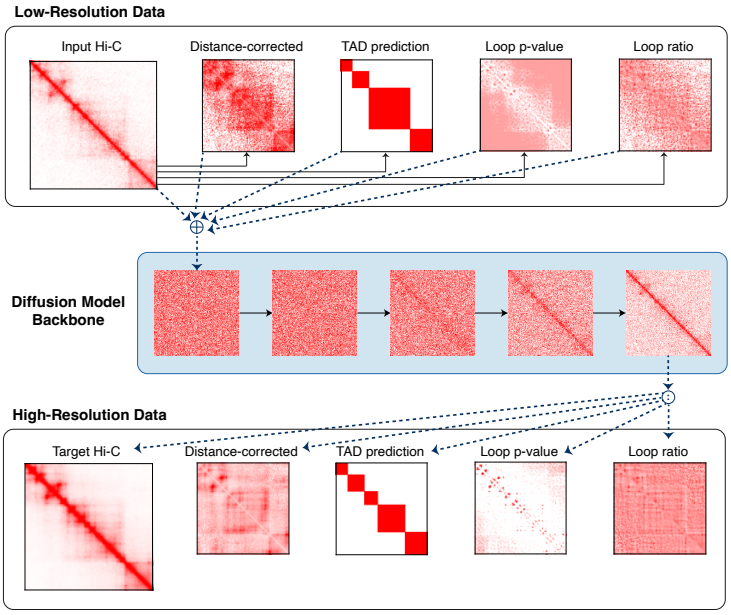


Figure 1: **Overview of Capricorn architecture.** Given the low- and high-resolution Hi-C contact matrices we compute high-order chromatin features that explicitly teach Capricorn to recognize meaningful 3D contacts. We then leverage a diffusion model backbone to iteratively de-noise a random sample, conditioned on the low-resolution data.

higher loop F1-score on average when transferring learned patterns across cell lines. We further found that Capricorn's key idea of incorporating higher-order chromatin features as additional input views is a broadly applicable technique that improved the comparison approaches as well, though Capricorn's diffusion model backbone still provides the best performance. Capricorn is a general-purpose approach for chromatin conformation capture contact matrix resolution enhancement, and additional high-order feature views can be easily incorporated into the framework for various downstream tasks of interest.

# 2 Method

## 2.1 Problem setting

For a set of experimental cell lines $\mathcal{C}$, we have interaction frequency matrices for chromosomes $\Xi$. In the supervised resolution enhancement task, we are given a dataset with pairs of low- and high-resolution contact matrices $\mathcal{D} = \{(\mathbf{X}_\xi^{(c)}, \mathbf{Y}_\xi^{(c)})\}_{c\in\mathcal{C},\xi\in\Xi}$. For chromosome $\xi$ containing $N_\xi$ total base pairs and a $\Delta = 10$ kb resolution, the contact matrix shapes are identical, with $\mathbf{X}_\xi^{(c)}, \mathbf{Y}_\xi^{(c)} \in \mathbb{N}^{N_\xi/\Delta\times N_\xi/\Delta}$. However, the high-resolution version $\mathbf{Y}_\xi^{(c)}$ has $\lambda$-fold more contacts measured in the matrix.

The task aim is then to obtain a model $f$ such that $\mathbf{Y}_\xi^{(c)} \approx f(\mathbf{X}_\xi^{(c)})$ for all $c \in \mathcal{C}$ and $\xi \in \Xi$. Furthermore, $f$ should generalize to both new cell lines and different chromosomes. As most Hi-C interactions occur between nearby intrachromosomal loci, we restrict the model to enhance intrachromosomal contacts within 2 Mb of each other as in previous work [16, 24].

Given the matrix $\mathbf{X}_\xi^{(c)} \in \mathbb{N}^{N_\xi/\Delta\times N_\xi/\Delta}$, each entry $\mathbf{X}_\xi^{(c)}[i,j]$ reports a $\Delta^2 = 10\text{kb}\times 10\text{kb}$ region of genomic interactions between genomic regions $[10i \text{ kb}, 10(i+1) \text{ kb})$ and $[10j \text{ kb}, 10(j+1) \text{ kb})$, where the bin size is defined as $\Delta = 10$ kb. We then tile the paired contact matrices into $40 \times 40$ non-overlapping submatrices $\mathbf{X}_\xi^{(c)}[i : i+40, j : j+40]$, each covering a $400^2$ kb region of contacts, consistent with many Hi-C resolution frameworks [9, 16, 19, 22, 25, 26, 24, 44]. For simplicity of notation, we drop the cell line and chromosome annotation when comparing a pair of low- and high-resolution tiled matrices, and we refer to a pair as $(\mathbf{X}, \mathbf{Y})$.

## 2.2 Chromatin structure resolution enhancement

The key idea behind Capricorn is that explicitly modeling high-order chromatin features such as TADs and loops will improve the method's ability to focus on meaningful contacts. We therefore train Capricorn to enhance the biological interpretation of the low-resolution contact matrix as well as the low-resolution contact matrix itself, thereby explicitly training the model to recognize important 3D chromatin structures from low-resolution data.

Toward this end, in addition to the resolution enhancement task with paired $(\mathbf{X}, \mathbf{Y})$, we further include additional views of chromatin features derived from $\mathbf{X}$ and $\mathbf{Y}$ for resolution enhancement.

1. Distance-corrected: $(\mathbf{X}^{(oe)}, \mathbf{Y}^{(oe)})$. The original contact matrices contain a strong bias based on inter-locus distance, with a higher number of measured contacts for locus pairs close to the diagonal. We therefore include a contact matrix corrected for expected contacts based on distance. Specifically, we compute the expected matrix by taking the average along each diagonal

$$\mathcal{E}(\mathbf{X})[i,j] = \frac{1}{n-|i-j|} \sum_{k=0}^{n-|i-j|} \mathbf{X}[k, k+|i-j|] \tag{1}$$

for all $i = 1, \ldots, n$ and $j = 1 \ldots, n$, where $n = N_\xi/\Delta$ for notational simplicity. We then normalize the observed matrix, clamp any large values, and finally normalize to the range $[0, 1]$ as

$$\mathbf{X}^{(oe)}[i,j] = \frac{\min\left(\mathbb{1}[\mathcal{E}(\mathbf{X})[i,j] > 0]\mathbf{X}[i,j]/\mathcal{E}(\mathbf{X})[i,j], 16\right)}{16} \tag{2}$$

where $\mathbb{1}[\cdot]$ is the indicator function, and compute $\mathbf{Y}^{(oe)}$ analogously.

2. TAD score: $(\mathbf{X}^{(tad)}, \mathbf{Y}^{(tad)})$. We compute the insulation score (IS) [7] for TAD detection, using sliding windows along the contact matrix diagonal to identify insulated regions with low scores and in-domain regions with high scores. We then smooth the IS scores by averaging them over a 21-bin domain and assign the within-TAD label 1 to contiguous regions with monotonically decreasing ISs centered around some locus $(i, j)$, where all ISs are still larger than the contact matrix average IS and the IS of $(i, j)$ is at least 0.1 larger than the IS at the region boundary.

3. Loop p-value and ratio: $(\mathbf{X}^{(loop-p)}, \mathbf{Y}^{(loop-p)}), (\mathbf{X}^{(loop-r)}, \mathbf{Y}^{(loop-r)})$. We follow Hi-C Computational Unbiased Peak Search (HiCCUPS) loop detection algorithm [33], which tests whether the measured contacts are significantly more frequent than would be expected. The method combines a $10 \times 10$ donut kernel centered at a specific locus with a horizontal, vertical, and lower-left quadrant kernel. For a given locus $(i, j)$, we use the distance-based expected matrix as defined in (1) to compute

$$\alpha_p = \kappa_p * \mathbf{X} \qquad\qquad \beta_p = \kappa_p * \mathcal{E}(\mathbf{X}) \tag{3}$$

for each kernel $\kappa_p = \{\kappa_1, \ldots, \kappa_4\}$. Then, we compute the *loop ratio* for the input matrix as

$$\mathbf{X}^{(loop-r)} = \min_{p \in \{1,\ldots,4\}} \frac{\beta_p \mathbf{X}}{\alpha_p \mathcal{E}(\mathbf{X})}. \tag{4}$$

We also follow the HiCCUPS algorithm and fit a Poisson distribution to the observed measurements with $\lambda$-chunking, where we set $\lambda = 2^{1/3}$ as in the default HiCCUPS algorithm [33]. In the $\lambda$-chunking protocol, we set $\mu_{i,j} = \max_c \lambda^c$ subject to $\frac{\alpha_p}{\beta_p}\mathcal{E}(\mathbf{X})[i,j] < \lambda^c$, grouping pixels into relative intensity ranges of $\{[0,1), [1, 2^{1/3}), [2^{1/3}, 2^{2/3}), \ldots\}$. Finally, we use the result of $\lambda$-chunking to compute the *loop p-value* as

$$\mathbf{X}^{(loop-p)}[i,j] = \max_{p \in \{1,\ldots,4\}} \mathtt{BH}_{\mu_{i,j}}\left(f(\mathbf{X}[i,j] \geq t | \mu = \mu_{i,j})\right), \tag{5}$$

where $f(\cdot \geq t|\mu)$ represents the survival function of the Poisson distribution with expected value $\mu$ and $\mathtt{BH}_{\mu_{i,j}}$ indicates the Benjamini-Hochberg false discovery rate correction [2, 4] over the p-values for all pixels with the same $\lambda$-chunked $\mu_{i,j}$. We compute $\mathbf{Y}^{(loop-r)}$ and $\mathbf{Y}^{(loop-p)}$ based on the ground-truth high-resolution matrix $\mathbf{Y}$ in an analogous manner.

Importantly, we compute each of the low-resolution chromatin features directly from the low-resolution experimental matrix $\mathbf{X}$, so the inputs to Capricorn have no knowledge of the high-resolution contact matrix $\mathbf{Y}$, which would prevent Capricorn's practical utility during inference. The original contact matrices and derived chromatin structures are then concatenated to form the full input and output

$$\begin{aligned}\tilde{\mathbf{X}}(\mathbf{X}) &= [\mathbf{X}, \mathbf{X}^{(oe)}, \mathbf{X}^{(tad)}, \mathbf{X}^{(loop-p)}, \mathbf{X}^{(loop-r)}] \in \mathbb{R}^{5\times N/\Delta\times N/\Delta} \\ \tilde{\mathbf{Y}}(\mathbf{Y}) &= [\mathbf{Y}, \mathbf{Y}^{(oe)}, \mathbf{Y}^{(tad)}, \mathbf{Y}^{(loop-p)}, \mathbf{Y}^{(loop-r}] \in \mathbb{R}^{5\times N/\Delta\times N/\Delta}. \end{aligned} \tag{6}$$

In order to prevent one experimental view from dominating model training, we first trained an exploratory version of Capricorn end-to-end and computed the validation set loss $\mathcal{L}'_{val}(\hat{\tilde{\mathbf{Y}}}|\tilde{\mathbf{X}})$. We compared the contribution of each generated view in $\hat{\tilde{\mathbf{Y}}}$ to the overall loss and then computed a weight vector $\omega \in \mathbb{R}+^5$ that would lead to each view contributing equally to the loss. When training the full version of Capricorn, we therefore weighted each channel by $\sqrt{\omega}$ when computing the input and output.

## 2.3 Image diffusion probability models

To carry out the resolution enhancement task given the low-resolution contact matrix and derived high-order chromatin features, we use a *diffusion probability model* backbone [18, 42], which has recently been shown to excel in image generation tasks [13, 21, 38] and is much more stable to train than GANs, which decompose the image generation process into a series of small steps. To the best of our knowledge, Capricorn is the first approach for Hi-C resolution enhancement that leverages diffusion probability models, and we therefore provide a brief overview here.

Diffusion models leverage a latent Markov chain framework that iteratively denoises an input to produce diverse and realistic outputs [18, 42]. The models are split into a *forward process* $q(\mathbf{y}_t|\mathbf{y}_{t-1})$ and *reverse process* $p(\mathbf{y}_{t-1}|\mathbf{y}_t, \mathbf{x})$, where $\mathbf{y}_0$ indicates the original, high-fidelity image, $\mathbf{y}_t$ are latent variables with the same dimension as $\mathbf{y}_0$, and $\mathbf{x}$ is an optional input term on which to condition the process. Specifically, the forward process computes progressively noisier latent representations of the true image as

$$q(\mathbf{y}_t|\mathbf{y}_{t-1}) \sim \mathcal{N}(\sqrt{\alpha_t}\mathbf{y}_{t-1}, \beta_t\mathbf{I}), \tag{7}$$

where $\beta_t$ is defined by a noise schedule, such as a cosine schedule [28, 37], $\alpha_t = 1 - \beta_t$, and the Markov chain repeats for $t = 1, \ldots, T$. The reverse diffusion process, optionally conditioned on input $\mathbf{c}$, can be then parameterized by a learnable $p_\theta(\cdot)$ as

$$p_\theta(\mathbf{y}_{t-1}|\mathbf{y}_t, \mathbf{x}, t) \sim \mathcal{N}\left(\frac{1}{\sqrt{\alpha_t}}\left(\mathbf{y}_t - \frac{\beta_t}{\sqrt{1 - \prod_{s=1}^{t}\alpha_s}\epsilon_\theta(\mathbf{y}_t, \mathbf{x}, t)}\right), \beta_t\mathbf{I}\right), \tag{8}$$

where $\epsilon(\cdot)$ is trained to predict noise from the previous step's $\mathbf{y}_t$ prediction and the conditional input $\mathbf{x}$. If the total noise $1 - \prod_{t=1}^{T}\alpha_t$ added by all steps in the Markov chain is sufficiently large, then the final noisy representation is well approximated by $\mathbf{y}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, enabling inference from a random Gaussian sample.

## 2.4 Low-resolution guided diffusion

We leverage conditional diffusion probability models for the resolution enhancement task. The combined contact matrix and its derived chromatin feature views can be interpreted as a five-channel image, lending the problem to image generation with low-resolution inputs used to guide the generation process. Capricorn is built around a backbone that approximates $\epsilon_\theta(\tilde{\mathbf{Y}}_{t-1}|\tilde{\mathbf{Y}}_t, \tilde{\mathbf{X}}, t)$ as part of the reverse process given in (8) (Figure 1).

Capricorn's multi-channel framework encodes additional biological understanding of the contact matrix data into the diffusion model resolution enhancement task, teaching the model to identify significant contacts while enabling the standard MSE loss formulation. Specifically, we optimize the objective

$$\mathcal{L}(\theta|\mathcal{D}) = \mathbb{E}_{(\mathbf{X},\mathbf{Y})\in\mathcal{D}}\Big[\mathbb{E}_{t\in\{1,\ldots,T-1\}}\Big[w_t\mathbb{E}\big[\|p_\theta(\tilde{\mathbf{Y}}_T|\tilde{\mathbf{Y}}_t, \tilde{\mathbf{X}}, t) - \tilde{\mathbf{Y}}\|^2\big]\Big]\Big], \tag{9}$$

where $w_t$ is a diffusion loss hyperparameter impacting the relative amount of sampled Gaussian noise [37], and the reverse diffusion process begins after sequentially introducing $t$ noise samples to the true matrix views $\tilde{\mathbf{Y}}$ for training.

### 2.5 Performance measures

We evaluate the results using both image-based and biologically-motivated metrics, focusing on the generated resolution-enhanced contact matrix. Specifically, given a model's predicted high-resolution contact matrix $\hat{\mathbf{Y}} = \tilde{\mathbf{Y}}_T[0]$, we compute the MSE as $\|\hat{\mathbf{Y}} - \mathbf{Y}\|_F^2$, for $\|\cdot\|_F$ the Frobenius norm.

We additionally use the enhanced high-resolution matrix to call loops with the HiCCUPS algorithm [33] to compute loop ratios and p-values as described in (4) and (5) respectively. For all loci $(i, j)$ with p-value $< 0.1$ and observed-over-expected ratio larger than 1.75 for the donut or lower-left-quadrant kernels, or larger than 1.5 for the horizontal or vertical kernels, we annotate $(i, j)$ as containing a loop. We then compute the loop $F1$ score $F1 = \frac{TP}{TP+\frac{1}{2}(FP+FN)}$ with a 5 pixel tolerance range, such that $TP$ are the true positive loops called from the predicted data that appear within $[i-5 : i+5, j-5 : j+5]$ in the loops called from the ground-truth data, $FP$ are the loops called from the predicted data that do not appear within the five-pixel range from the ground-truth data, and $FN$ the loops that are called from the ground-truth data but do not occur in the five-pixel tolerance range for predicted data.

### 2.6 Hi-C data preprocessing

We collected Hi-C data using the Gene Expression Omnibus (GEO) database for the GM12878 Epstein-Barr-virus-infected human lymphoblastoid cell line and K562 human chronic myelogenous leukemia lymphoblast cell lines from the Rao et al. (2014) [33] dataset (accession code GSE63525, https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE63525), restricting the contact matrix to read mapping quality $\geq 30$ and processed to 10 kilobase (kb) resolution. We adopt the contact matrix preprocessing techniques from HiCARN [16] and DeepHiC [19], including tiling the contact matrices into $40 \times 40$ submatrices, only retaining submatrices in the 2 megabase (Mb) region around the diagonal, clamping the high-resolution matrix to $[0, 255]$ and then normalizing to $[0, 1]$, and clamping the low-resolution matrix to $[0, 100]$ and then normalizing to $[0, 1]$.

In order to simulate low-resolution data, we downsampled the GM12878 and K562 cell line Hi-C matrices [33] to $\frac{1}{16}$ of the original read count. We treated these downsampled data as the low-resolution matrices and the original matrices as their full-resolution pairs. We conduct and report on two separate experiments to evaluate model generalizability: in the first, we train on the GM12878 data and test the model on K562; in the second, we train on K562 data and test on GM12878. In both experiments, we withhold chromosomes 4, 5, 11, and 14 from the training cell line as our validation set. The K562 data is excluded for chromosome 9 due to extreme sparsity at 10 kb resolution. After running the exploratory version of Capricorn to determine view weighting, as described in Section 2.2, we set view weights to $\sqrt{\omega} = [1, 0.073, 0.008, 0.038, 0.018]$ when training on GM12878 and $\sqrt{\omega} = [1, 0.036, 0.010, 0.049, 0.014]$ when training on K562.

### 2.7 Existing model implementations

We use the publicly available python HiCARN [16] repository from GitHub (https://github.com/OluwadareLab/HiCARN) for both their model and contact matrix preprocessing implementation, as well as source code implementations for HiCNN [25] and HiCSR [9].

We also use the conditional diffusion probability model Imagen [37] as the resolution enhancement backbone model, updating the model to condition on low-resolution contact matrices rather than text. We choose Imagen rather than other image diffusion models [29, 31, 35] due to its efficient U-Net architecture, which is faster and more memory efficient than other diffusion generators. We use the GitHub repository https://github.com/lucidrains/imagen-pytorch to access Imagen.

## 3 Results

### 3.1 Capricorn accurately enhances contact matrices and loop features

We find that Capricorn outperforms other methods both in terms of its ability to enhance chromatin loops from the low-resolution data and its accuracy in producing high-resolution data for both test cell lines in the cross-cell-line experimental setting. We first used the HiCCUPS [33] algorithm to call loops in

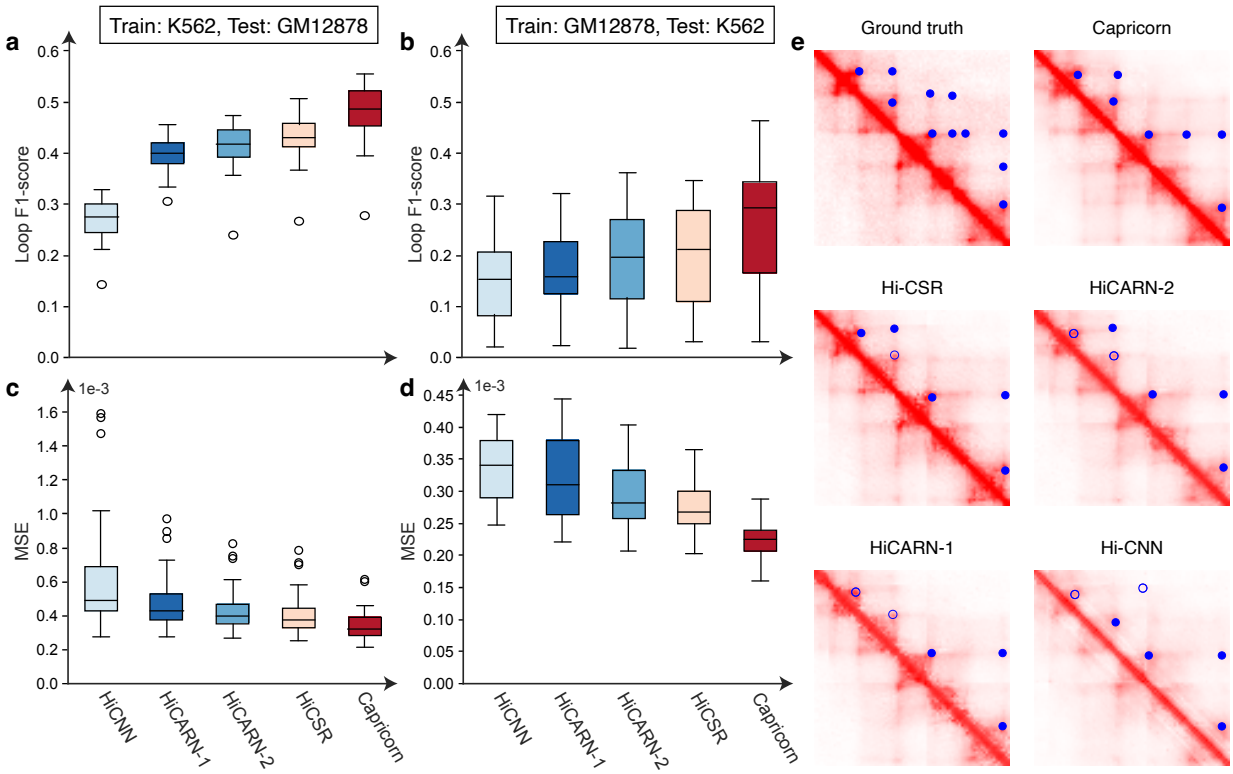


Figure 2: **Resolution enhancement model performance. a-d,** Boxplot comparison of F1-score for loop detection from the generated high-resolution matrices (**a,b**, higher is better) and generated matrix MSE (**c,d**, lower is better). The boxplots show the median, interquartile range (IQR), 1.5×IQR, and outliers by chromosome. **e**, Ground-truth high-resolution submatrix covering genomic loci from 47.3 Mb to 48.1 Mb on GM12878 chromosome 17, compared to the generated high-resolution submatrix output by each method. Blue circles indicated called loops in the ground-truth or predicted high-resolution data. Circles for loops that are also called in the ground truth data are filled in; loops that are called from the generated data but not the ground-truth data are empty circles.

the resolution-enhanced contact matrices, and we compared these to the loops HiCCUPS identified in the original full-resolution data. As shown in Figure 2a,b, Capricorn outperformed all other approaches in its ability to recognize and enhance loop chromatin structures when transferred to both the GM12878 and K562 cell lines, with an average loop $F_1$ score taken over chromosomes of 0.48 on GM12878 test data and 0.31 on K562 test data, relative to 0.43 and 0.20 for HiCSR, the best-performing comparison approach, and 0.30 and 0.19 if directly using the input low-resolution data scaled by the experimental downsampling rate. The performance difference for all methods between GM12878 and K562 test data can be explained by the difference in pairwise contacts in the original dataset: GM12878 has approximately five times more measured contacts than K562 [33]. These results demonstrate that Capricorn is able to successfully identify and enhance meaningful biological contacts, such as those involved in loop formation.

Furthermore, Capricorn achieved a lower prediction MSE than any of the comparison approaches, indicating that the overall resolution enhancement results are accurate. Figure 2c,d, shows that Capricorn has an average MSE taken over chromosomes of 3.5e−4 and 2.2e−4 for the GM12878 and K562 test cell lines, respectively, relative to 4.2e−4 and 2.8e−4 for HiCSR, again the best-performing comparison approach, and to 2.4e−3 and 1.1e−3 if directly using the input low-resolution data scaled by the experimental downsampling rate.

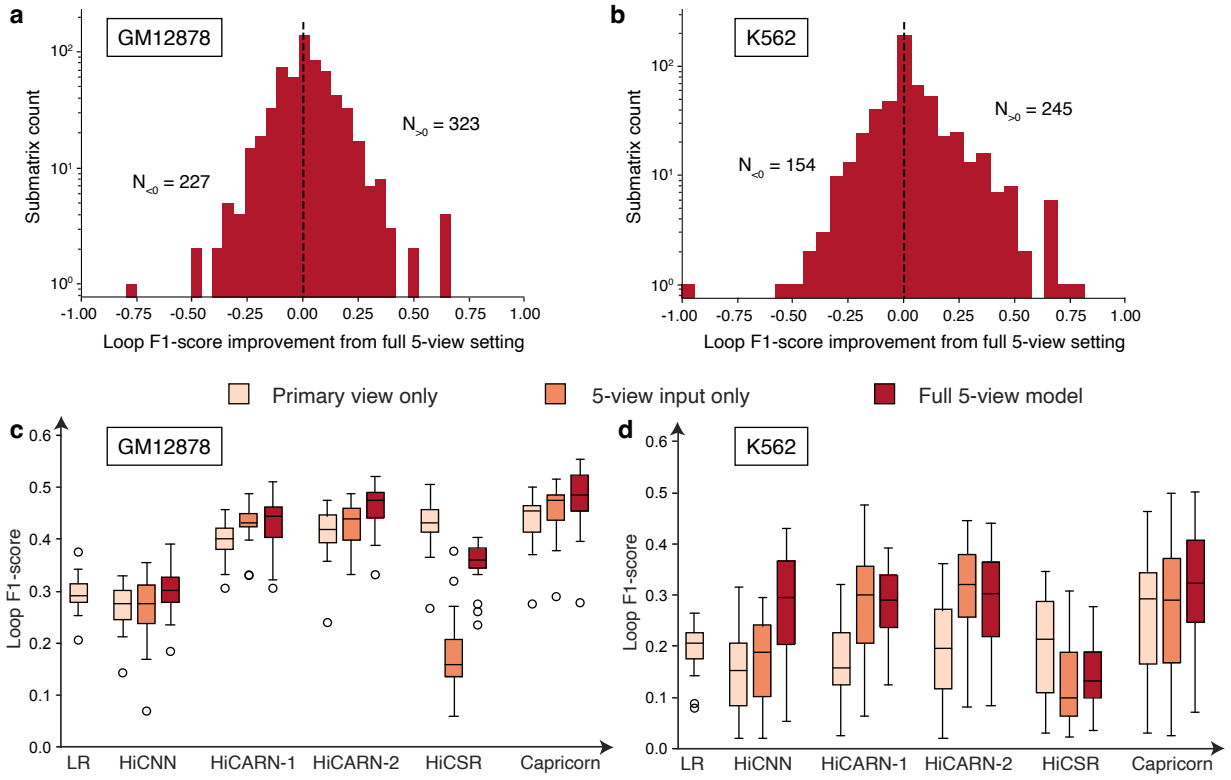


Figure 3: **Study of Capricorn's multi-view chromatin feature framework. a,b,** Histogram of the difference in loop F1-score from the predicted high-resolution contact matrix in the GM12878 (**a**) and K562 (**b**) test datasets using the full five-view Capricorn framework and an alternate version of Capricorn that only includes the primary Hi-C matrix view as the input and output for the diffusion model backbone. The counts $N_{>0}$ and $N_{<0}$ indicate the number of submatrices for which the five-view setting respectively outperform or underperform the primary-only setting. **c,d** Boxplot comparing the performance of resolution enhancement methods in the primary-view-only setting where enhancement is only performed with Hi-C matrices, the five-view-input setting where high-order chromatin features are included as input to the resolution model, and the full-five-view setting where high-order chromatin features are used both as input and output to train the model backbone. "LR" indicates the results taking low-resolution matrix and scaling all contacts by the downsampling factor 16. The boxplots show the median, IQR, 1.5× IQR, and outliers by predicted submatrix.

## 3.2 High-order chromatin features are critical to model improvement

We next investigated the impact of including high-order chromatin features as additional input views to train the resolution enhancement model to identify and enhance structurally meaningful contacts. We compared Capricorn's performance when trained to enhance chromatin features as well as the Hi-C matrix to its performance when performing resolution enhancement without any additional views. As shown in Figure 3a,b, we find that explicitly training the model to enhance meaningful biological contacts as captured in the high-order chromatin features significantly improves our performance to identify these features from the enhanced contact matrices (one-sided Wilcoxon signed-rank test p-value $< 6.4\text{e}{-5}$, $< 1.3\text{e}{-6}$ comparing loop F1-score over submatrices for test GM12878 and K562 respectively).

We then further tested whether the benefits of providing high-order chromatin features as additional views of the initial Hi-C matrix generalized to other resolution enhancement models. To this end, we updated the comparison models to accept multi-channel input matrices and compared the results of three model formulations. First, each *primary view* setting uses the model's default resolution enhancement pipeline

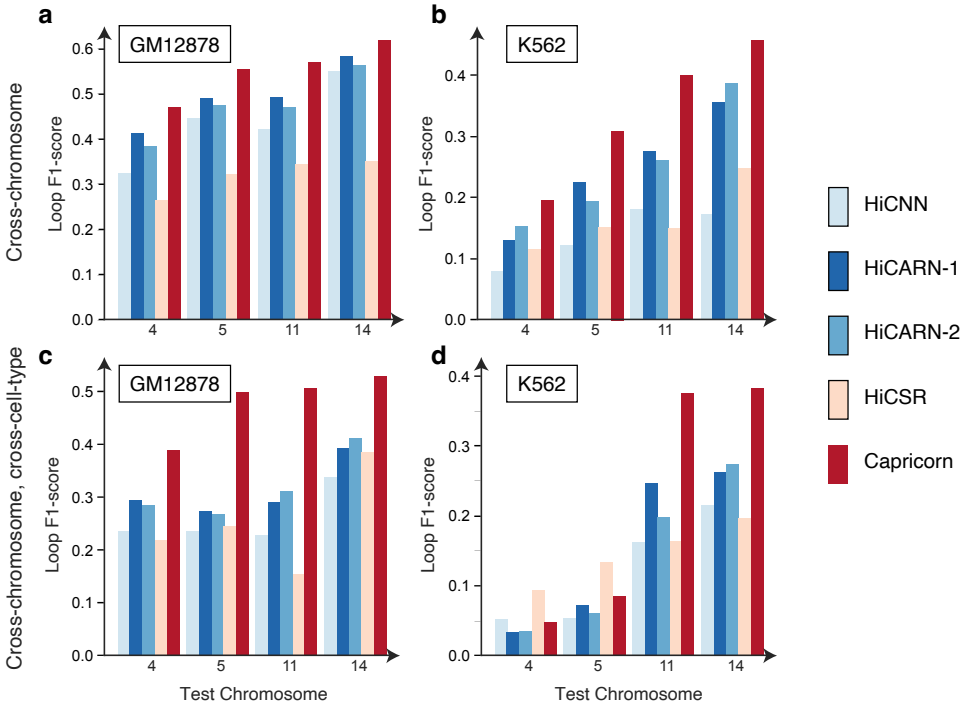


Figure 4: **Model performance comparison in cross-chromosome and cross-chromosome, cross-cell-type experimental settings. a-d**, Barplots showing loop F1-score by test chromosome in a cross-chromosome experiment (**a,b**) and cross-chromosome, cross-cell-type experiment (**c,d**). Plots are labeled with the cell line of the test data.

with the low- and high-resolution Hi-C matrices as input and output. Second, we constructed a *five-view input* setting that uses the additional chromatin feature views as input but is still trained to only predict the high-resolution contact matrix. Third, we evaluated the *full five-view* setting that takes the chromatin feature views as input and also trains the model to enhance the chromatin features in addition to the Hi-C matrices, as in Capricorn.

As shown in Figure 3c,d, several comparison methods benefit from including the high-order data, though Capricorn's diffusion model backbone remains state-of-the-art even when other models leverage the high-order chromatin features. This observation points to the broad applicability of Capricorn's key idea, and also suggests opportunities for including additional Hi-C-derived views based on expert domain knowledge for various downstream genome folding analyses.

## 3.3 Capricorn generalizes across chromosomes

To confirm that Capricorn's strong performance is not due to memorizing the training chromosomes with relatively small cell-line differences, we conducted a second experiment to rigorously test Capricorn in the cross-chromosome setting. Here, we withhold chromosomes 2, 6, 10, and 12 as a validation set and reserve chromosomes 4, 14, 16, and 20 as test data. We therefore carried out two experiments to examine cross-chromosome and intra-cell-line generalization as well as cross-chromosome and cross-cell-line generalization.

We find that Capricorn is able to transfer its learned resolution enhancement patterns to never-before-seen genomic loci in the cross-chromosome intra-cell-line setting better than other methods (one-sided Wilcoxon signed-rank test p-value $< 4.0\text{e}{-3}$). Across the four test chromosomes, Capricorn's generated high-resolution contact matrix had an average loop F1-score of 0.55 in the GM12878 experiment (Figure 4a) and 0.34 in the K562 experiment (Figure 4b), relative to the best-performing comparison approaches at 0.49 for HiCARN-1 and 0.25 for HiCARN-2. Similarly, Capricorn had the highest loop F1 scores in the cross-chromosome, cross-

cell-line experiments as shown in Figure 4c,d (one-sided Wilcoxon signed-rank test p-value $< 1.9e{-}2$), with an average loop F1 score of 0.48 on GM12878 and 0.22 on K562 test data, relative to the best comparison approaches respective loop F1-scores of 0.32 from HiCARN-2 and 0.15 from HiCARN-1. This result highlights Capricorn's ability to generalize the informative, high-order chromatin patterns it learns across genomic loci as well as cell line.

# 4 Discussion

In this work we present Capricorn as a tool for Hi-C resolution enhancement. Capricorn explicitly models the biology underlying the contact matrices by incorporating high-order chromatin features into the model formulation and loss function. Furthermore, we find that this key insight is widely applicable to other resolution enhancement approaches, improving performance in three out of four of the comparison approaches. However, the high-order chromatin feature views still perform best with Capricorn's conditional diffusion model backbone.

We demonstrate Capricorn's strong performance in cross-cell-line, cross-chromosome, and cross-chromosome-cross-cell-line settings with the GM12878 and K562 datasets. In all three measured settings, Capricorn is best able to generate high-resolution data containing accurate chromatin loops that can be called with standard tools [33]. This highlights Capricorn's generalizability as well as the value of including additional biological data views, differentiating Hi-C resolution enhancement from resolution enhancement in natural image applications. In the future, Capricorn's framework can be further broadened to include additional biological views designed specifically for downstream tasks of interest. We also imagine that future work will study the impact of different model backbones on the multi-view resolution enhancement problem.

# References


[1] Beagrie, R.A., Scialdone, A., Schueler, M., Kraemer, D.C.A., Chotalia, M., Xie, S.Q., Barbieri, M., de Santiago, I., Lavitas, L.M., Branco, M.R., Fraser, J., Dostie, J., Game, L., Dillon, N., Edwards, P.A.W., Nicodemi, M., Pombo, A.: Complex multi-enhancer contacts captured by genome architecture mapping. Nature **543**(7646), 519–524 (Mar 2017)

[2] Benjamini, Y., Hochberg, Y.: Controlling the false discovery rate: A practical and powerful approach to multiple testing. J. R. Stat. Soc. Series B Stat. Methodol. **57**(1), 289–300 (1995)

[3] Bonev, B., Cavalli, G.: Organization and function of the 3D genome. Nat. Rev. Genet. **17**(11), 661–678 (Oct 2016)

[4] Cai, T.T., Sun, W.: Simultaneous testing of grouped hypotheses: Finding needles in multiple haystacks. J. Am. Stat. Assoc. **104**(488), 1467–1481 (Dec 2009)

[5] Carron, L., Morlot, J.B., Matthys, V., Lesne, A., Mozziconacci, J.: Boost-HiC: computational enhancement of long-range contacts in chromosomal contact maps. Bioinformatics **35**(16), 2724–2729 (Aug 2019)

[6] Chakraborty, A., Wang, J.G., Ay, F.: dcHiC detects differential compartments across multiple Hi-C datasets. Nat. Commun. **13**(1), 6827 (Nov 2022)

[7] Crane, E., Bian, Q., McCord, R.P., Lajoie, B.R., Wheeler, B.S., Ralston, E.J., Uzawa, S., Dekker, J., Meyer, B.J.: Condensin-driven remodelling of X chromosome topology during dosage compensation. Nature **523**(7559), 240–244 (Jul 2015)

[8] Cremer, T., Cremer, C.: Chromosome territories, nuclear architecture and gene regulation in mammalian cells. Nat. Rev. Genet. **2**(4), 292–301 (Apr 2001)

[9] Dimmick, M.: HiCsr: A Hi-C Super-Resolution Framework for Producing Highly Realistic Contact Maps. Ph.D. thesis, University of Toronto (Canada) (2020)

[10] Dixon, J.R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J.S., Ren, B.: Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature **485**(7398), 376–380 (Apr 2012)

[11] Filippova, D., Patro, R., Duggal, G., Kingsford, C.: Identification of alternative topological domains in chromatin. Algorithms Mol. Biol. **9**, 14 (May 2014)

[12] Fortin, J.P., Hansen, K.D.: Reconstructing A/B compartments as revealed by Hi-C using long-range correlations in epigenetic data. Genome Biol. **16**(1), 180 (Aug 2015)

[13] Gao, S., Liu, X., Zeng, B., Xu, S., Li, Y., Luo, X., Liu, J., Zhen, X., Zhang, B.: Implicit diffusion models for continuous Super-Resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10021–10030 (Jun 2023)

[14] Haddad, N., Vaillant, C., Jost, D.: IC-Finder: inferring robustly the hierarchical organization of chromatin folding. Nucleic Acids Res. **45**(10), e81 (Jun 2017)

[15] Harris, H.L., Gu, H., Olshansky, M., Wang, A., Farabella, I., Eliaz, Y., Kalluchi, A., Krishna, A., Jacobs, M., Cauer, G., Pham, M., Rao, S.S.P., Dudchenko, O., Omer, A., Mohajeri, K., Kim, S., Nichols, M.H., Davis, E.S., Gkountaroulis, D., Udupa, D., Aiden, A.P., Corces, V.G., Phanstiel, D.H., Noble, W.S., Nir, G., Di Pierro, M., Seo, J.S., Talkowski, M.E., Aiden, E.L., Rowley, M.J.: Chromatin alternates between a and B compartments at kilobase scale for subgenic organization. Nat. Commun. **14**(1), 3303 (Jun 2023)

[16] Hicks, P., Oluwadare, O.: HiCARN: resolution enhancement of Hi-C data using cascading residual networks. Bioinformatics **38**(9), 2414–2421 (Apr 2022)

[17] Highsmith, M., Cheng, J.: VEHiCLE: a variationally encoded Hi-C loss enhancement algorithm for improving and generating Hi-C data. Sci. Rep. **11**(1), 8880 (Apr 2021)

[18] Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M.F., Lin, H. (eds.) Advances in Neural Information Processing Systems. vol. 33, pp. 6840–6851. Curran Associates, Inc. (2020)

[19] Hong, H., Jiang, S., Li, H., Du, G., Sun, Y., Tao, H., Quan, C., Zhao, C., Li, R., Li, W., Yin, X., Huang, Y., Li, C., Chen, H., Bo, X.: DeepHiC: A generative adversarial network for enhancing Hi-C data resolution. PLoS Comput. Biol. **16**(2), e1007287 (Feb 2020)

[20] Hu, Y., Ma, W.: EnHiC: learning fine-resolution Hi-C contact maps using a generative adversarial framework. Bioinformatics **37**(Suppl_1), i272–i279 (Jul 2021)

[21] Li, H., Yang, Y., Chang, M., Chen, S., Feng, H., Xu, Z., Li, Q., Chen, Y.: SRDiff: Single image super-resolution with diffusion probabilistic models. Neurocomputing **479**, 47–59 (Mar 2022)

[22] Li, Z., Dai, Z.: SRHiC: A deep learning model to enhance the resolution of Hi-C data. Front. Genet. **11**, 353 (Apr 2020)

[23] 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 (Oct 2009)

[24] Liu, Q., Lv, H., Jiang, R.: hicGAN infers super resolution Hi-C data with generative adversarial networks. Bioinformatics **35**(14), i99–i107 (Jul 2019)

[25] Liu, T., Wang, Z.: HiCNN: a very deep convolutional neural network to better enhance the resolution of Hi-C data. Bioinformatics **35**(21), 4222–4228 (Nov 2019)

[26] Liu, T., Wang, Z.: HiCNN2: Enhancing the resolution of Hi-C data using an ensemble of convolutional neural networks. Genes **10**(11) (Oct 2019)

[27] 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. Nat. Methods **13**(11), 919–922 (Nov 2016)

[28] Nichol, A., Dhariwal, P.: Improved denoising diffusion probabilistic models (Feb 2021)

[29] Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., Chen, M.: GLIDE: Towards photorealistic image generation and editing with Text-Guided diffusion models (Dec 2021)

[30] Quinodoz, S.A., Ollikainen, N., Tabak, B., Palla, A., Schmidt, J.M., Detmar, E., Lai, M.M., Shishkin, A.A., Bhat, P., Takei, Y., Trinh, V., Aznauryan, E., Russell, P., Cheng, C., Jovanovic, M., Chow, A., Cai, L., McDonel, P., Garber, M., Guttman, M.: Higher-Order inter-chromosomal hubs shape 3D genome organization in the nucleus. Cell **174**(3), 744–757.e24 (Jul 2018)

[31] Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical Text-Conditional image generation with CLIP latents (Apr 2022)

[32] 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. Nat. Commun. **9**(1), 189 (Jan 2018)

[33] Rao, S.S.P., Huntley, M.H., Durand, N.C., Stamenova, E.K., Bochkov, I.D., Robinson, J.T., Sanborn, A.L., Machol, I., Omer, A.D., Lander, E.S., Aiden, E.L.: A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell **159**(7), 1665–1680 (Dec 2014)

[34] Roayaei Ardakany, A., Gezer, H.T., Lonardi, S., Ay, F.: Mustache: multi-scale detection of chromatin loops from Hi-C and Micro-C maps using scale-space representation. Genome Biol. **21**(1), 256 (Sep 2020)

[35] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-Resolution image synthesis with latent diffusion models. arXiv [cs.CV] (2021)

[36] Rowley, M.J., Poulet, A., Nichols, M.H., Bixler, B.J., Sanborn, A.L., Brouhard, E.A., Hermetz, K., Linsenbaum, H., Csankovszki, G., Lieberman Aiden, E., Corces, V.G.: Analysis of Hi-C data using SIP effectively identifies loops in organisms from c. elegans to mammals. Genome Res. **30**(3), 447–458 (Mar 2020)

[37] Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E., Ghasemipour, S.K.S., Gontijo-Lopes, R., Ayan, B.K., Salimans, T., Ho, J., Fleet, D.J., Norouzi, M.: Photorealistic Text-to-Image diffusion models with deep language understanding. In: Oh, A.H., Agarwal, A., Belgrave, D., Cho, K. (eds.) Advances in Neural Information Processing Systems (2022)

[38] Saharia, C., Ho, J., Chan, W., Salimans, T., Fleet, D.J., Norouzi, M.: Image Super-Resolution via iterative refinement. IEEE Trans. Pattern Anal. Mach. Intell. **45**(4), 4713–4726 (Apr 2023)

[39] Salameh, T.J., Wang, X., Song, F., Zhang, B., Wright, S.M., Khunsriraksakul, C., Ruan, Y., Yue, F.: A supervised learning framework for chromatin loop detection in genome-wide contact maps. Nat. Commun. **11**(1), 3428 (Jul 2020)

[40] Schmitt, A.D., Hu, M., Ren, B.: Genome-wide mapping and analysis of chromosome architecture. Nat. Rev. Mol. Cell Biol. **17**(12), 743–755 (Dec 2016)

[41] Serra, F., Baù, D., Goodstadt, M., Castillo, D., Filion, G.J., Marti-Renom, M.A.: Automatic analysis and 3d-modelling of Hi-C data using TADbit reveals structural features of the fly chromatin colors. PLoS Comput. Biol. **13**(7), e1005665 (Jul 2017)

[42] Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., Ganguli, S.: Deep unsupervised learning using nonequilibrium thermodynamics. In: Bach, F., Blei, D. (eds.) Proceedings of the 32nd International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 37, pp. 2256–2265. PMLR, Lille, France (2015)

[43] Wolff, J., Backofen, R., Grüning, B.: Loop detection using Hi-C data with HiCExplorer. Gigascience **11** (Jul 2022)

[44] Zhang, Y., An, L., Xu, J., Zhang, B., Zheng, W.J., Hu, M., Tang, J., Yue, F.: Enhancing Hi-C data resolution with deep convolutional neural network HiCPlus. Nat. Commun. **9**(1), 750 (Feb 2018)

[45] Zhang, Y., Blanchette, M.: Reference panel-guided super-resolution inference of Hi-C data. Bioinformatics **39**(39 Suppl 1), i386–i393 (Jun 2023)