

Article

A Comprehensive Evaluation of Generalizability of Deep Learning-Based Hi-C Resolution Improvement Methods

Ghulam Murtaza ¹ , Atishay Jain ¹ , Madeline Hughes ¹, Justin Wagner ² and Ritambhara Singh ^{1,3,*}

¹ Department of Computer Science, Brown University, Providence, RI 02912, USA; ghulam_murtaza@brown.edu (G.M.); atishay_jain@brown.edu (A.J.); madeline_hughes@brown.edu (M.H.)
² Material Measurement Laboratory, National Institute of Standards and Technology, Gaithersburg, MD 20899, USA; justin.wagner@nist.gov
³ Center for Computational Molecular Biology, Brown University, Providence, RI 02912, USA
 * Correspondence: ritambhara@brown.edu

Abstract: Hi-C is a widely used technique to study the 3D organization of the genome. Due to its high sequencing cost, most of the generated datasets are of a coarse resolution, which makes it impractical to study finer chromatin features such as Topologically Associating Domains (TADs) and chromatin loops. Multiple deep learning-based methods have recently been proposed to increase the resolution of these datasets by imputing Hi-C reads (typically called upscaling). However, the existing works evaluate these methods on either synthetically downsampled datasets, or a small subset of experimentally generated sparse Hi-C datasets, making it hard to establish their generalizability in the real-world use case. We present our framework—Hi-CY—that compares existing Hi-C resolution upscaling methods on seven experimentally generated low-resolution Hi-C datasets belonging to various levels of read sparsities originating from three cell lines on a comprehensive set of evaluation metrics. Hi-CY also includes four downstream analysis tasks, such as TAD and chromatin loops recall, to provide a thorough report on the generalizability of these methods. We observe that existing deep learning methods fail to generalize to experimentally generated sparse Hi-C datasets, showing a performance reduction of up to 57%. As a potential solution, we find that retraining deep learning-based methods with experimentally generated Hi-C datasets improves performance by up to 31%. More importantly, Hi-CY shows that even with retraining, the existing deep learning-based methods struggle to recover biological features such as chromatin loops and TADs when provided with sparse Hi-C datasets. Our study, through the Hi-CY framework, highlights the need for rigorous evaluation in the future. We identify specific avenues for improvements in the current deep learning-based Hi-C upscaling methods, including but not limited to using experimentally generated datasets for training.

Keywords: chromosome conformation capture; Hi-C; resolution improvement; generalizability



Citation: Murtaza, G.; Jain, A.; Hughes, M.; Wagner, J.; Singh, R. A Comprehensive Evaluation of Generalizability of Deep Learning-Based Hi-C Resolution Improvement Methods. *Genes* **2024**, *15*, 54. <https://doi.org/10.3390/genes15010054>

Academic Editors: Zhiping Liu, Han Zhang and Junwei Han

Received: 4 December 2023

Revised: 24 December 2023

Accepted: 26 December 2023

Published: 29 December 2023



Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

The 3D organization of the genome plays a vital role in cell fate and disease onset. A high-throughput chromosome conformation capture experiment, or Hi-C, is a genome-wide sequencing technique that allows researchers to understand and study the 3D organization of the genome [1]. The sequencing results from Hi-C correspond to observed molecular contacts between two genomic loci. This contact information captures local and global interactions of the DNA molecule. In the past decade, analysis of Hi-C data facilitated the discovery of important genomic structural features, including but not limited to A/B compartments [1] that denote active and inactive genomic regions, topologically associated domains (TADs) [2] that represent highly interactive genomic regions, and enhancer-promoter interactions [3] that are involved in the regulation of genes. Therefore, Hi-C experiments are crucial in advancing our understanding of the spatial structure of the genome and its relationship with gene regulation machinery.

When studying the spatial structure of DNA, the quality of the downstream analysis is highly dependent on the resolution of its Hi-C contact map. For example, having precise locations of chromatin loops in a RAD-21 knock-out experiment [4] is crucial in understanding its impact on the chromatin structure. Data from a Hi-C experiment coalesce into a matrix (or contact map as shown in Figure 1A) in which rows and columns correspond to fixed-width windows (“bins”) tiled along the genomic axis, and values in the matrix are counts of read pairs that fall into the corresponding bins. The bin size typically ranges from 1 kilobase pair (kbp) to 1 megabase pair (mbp), where the choice of the bin size depends on the number of paired end reads from the experiment. Lower read count experiments result in sparser contact matrices that require large bin sizes, resulting in a “low-resolution” contact map. Consequently, the downstream analysis of the contact map cannot yield genomic features, such as enhancer–promoter interactions that typically occur in the 5 kbp to 10 kbp range [5]. Similarly, the output of a Hi-C experiment with a high number of read counts results in a “high-resolution” contact map with small bin sizes. A contact map of this resolution enables identifying fine-grained genomic features. However, due to the quadratic scaling of the sequencing cost, most tissue and cell line samples do not have any Hi-C data available. For example, on the ENCODE portal, there are only 92 samples that have a Hi-C experiment available across a collection of 450 cell lines and tissue samples. For cell and tissue samples that have a Hi-C experiment available, they have relatively low read counts (typically ≤ 100 Million reads) and, consequently, low-resolution (≥ 40 kbp bin size) contact maps [6]. Constructing Hi-C matrices with high resolution (≤ 10 kbp bin size) requires billions of Hi-C reads [5], which can be prohibitively expensive to obtain for many experiments. Thus, the absence of such matrices makes the comprehensive analysis of the spatial structure of DNA difficult. This limitation is even more apparent in single-cell variants of the Hi-C protocol [7], where the reads are even more sparse, and it is a complicated experimental challenge to acquire high-resolution contact matrices.

Recently, researchers developed several computational methods to upscale (synonymous with increasing the Hi-C resolution) Hi-C matrices by imputing Hi-C reads. Deep learning-based methods [6,8–11] have shown remarkable success in upscaling Hi-C matrices when trained and tested on simulated datasets. These datasets are typically constructed by uniformly removing a fixed fraction of reads from an experimentally generated high-resolution Hi-C contact map to simulate a low-resolution (downsampled) Hi-C contact map. This downsampling method tends to miss out on experimental artifacts, such as the diagonal effect [12], and produces a Hi-C contact map that has substantially different distribution from an experimentally generated sparse Hi-C contact map. Moreover, the existing methods report their performance on correlation-based metrics that ignore the genome’s multi-scale hierarchical organization. VeHiCle is an exception [13] that utilizes experimentally generated Hi-C datasets and state-of-the-art Hi-C-specific similarity metrics [14]. VeHiCle restricts its analysis to only four datasets with low-sparsity (high number of reads) Hi-C datasets. Therefore, it does not provide a comprehensive report on the utility of the method for its intended use case, where Hi-C matrices tend to be more sparse.

We investigate how existing deep learning-based methods generalize their upscaling performance to experimentally generated sparse Hi-C datasets. Given the rising popularity of single-cell chromatin capture protocols such as scHi-C [7] that are even sparser than their bulk sequencing counterparts, it is imperative to test how these methods perform for sparse Hi-C datasets for future methods development. To accomplish this, we developed a framework to pre-process, upscale, and evaluate methods on multiple Hi-C contact maps with varying levels of read counts originating from different cell lines. We present our framework Hi-CY in Figure 1. Hi-CY supports the investigation of Hi-C resolution upscaling efforts by providing access to a comprehensive set of Hi-C datasets and evaluation metrics. We contribute towards the recent efforts to standardize evaluation protocols and develop reproducible benchmarks, such as those for epigenome imputation [15], protein–drug interaction prediction [16], and Hi-C feature extraction [17]. These projects perform

a comprehensive evaluation of existing methods and provide robust and reproducible frameworks to make them accessible and to accelerate research.

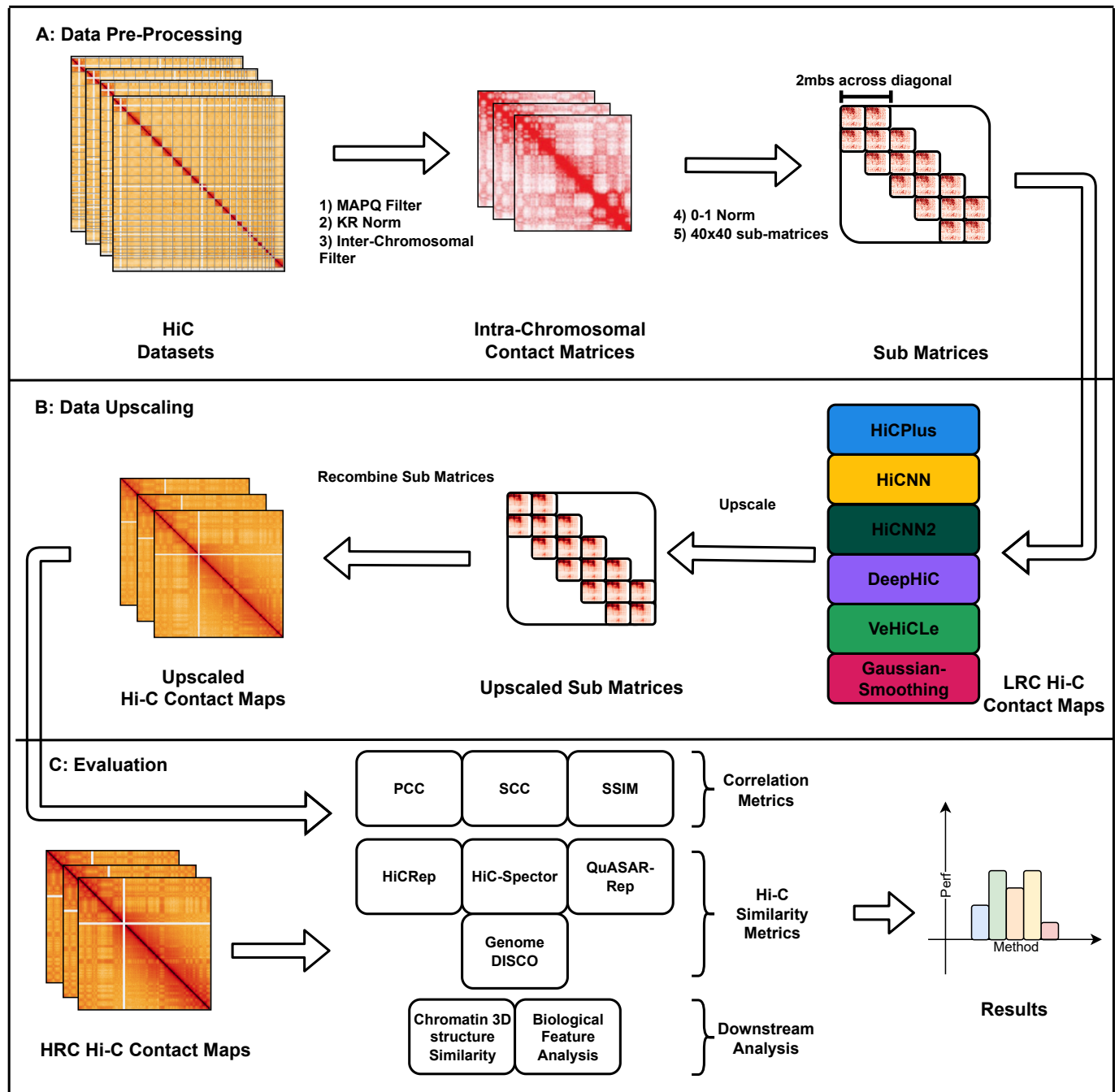


Figure 1. Overview of our benchmarking framework Hi-CY: **(A)** Data pre-processing pipeline. We (1) filtered the Hi-C matrices with the same MAPQ value (≥ 30), (2) normalized them with the same KR normalization algorithm to ensure a fair comparison, (3) removed inter-chromosomal contacts because of their extremely sparse nature, (4) performed a 0–1 normalization on the intra-chromosomal matrices to reduce the impact of extreme values; and (5) cropped appropriately sized sub-matrices to ensure that the input is in the correct format for each upscaling algorithm. **(B)** We upscaled the sub-matrices using a wide variety of deep learning-based upscaling models (and Gaussian Smoothing) and then recombined them to form upscaled intra-chromosomal Hi-C matrices. **(C)** We combined multiple Hi-C similarity metrics, correlation-based metrics and downstream analyses, like chromatin loops and TAD recovery analysis, to provide a comprehensive report that we can use to analyze the performance of each upscaling model.

Hi-CY includes seven sparse experimentally generated datasets from three cell lines (GM12878, K562, and IMR90) with sparsity ranging from $\frac{1}{9}$ to $\frac{1}{100}$ of the reads in comparison to the appropriate high-resolution Hi-C contact map. We also package six Hi-C upscaling methods (Gaussian Smoothing, HiCPlus, HiCNN, HiCNN2, DeepHiC, and VeHiCLe), which we evaluate on three correlation-based metrics; four Hi-C similarity metrics; and four downstream analysis tasks. Our metrics capture the bin-wise similarity (through correlation-based metrics), genome structural similarity, and biological information content to provide a comprehensive report on the utility of these generated Hi-C contact maps. Our results show:

- The existing deep learning-based methods struggle to generalize to sparse, experimentally generated Hi-C inputs. These methods show a reduction of up to 57% in performance when upscaling a Hi-C contact map with $\frac{1}{100}$ of the reads (compared to the high-resolution Hi-C contact map) in comparison to upscaling a Hi-C contact map with $\frac{1}{9}$ of the reads.
- Our results show that retraining existing deep learning-based models with experimentally generated Hi-C datasets improves performance by 31% on experimentally generated Hi-C contact maps.
- We further find that deep learning-based methods still struggle to recover biologically meaningful features in the upscaled Hi-C contact maps, even when retrained, particularly on chromatin loop and DNA hairpin recovery tasks.

2. Hi-Cy: A Comprehensive Evaluation Framework

We developed Hi-CY to facilitate the development and evaluation of Hi-C upscaling methods in a robust and reproducible setup. As shown in Figure 1, Hi-CY packages (1) a Hi-C pre-processing pipeline, (2) a Hi-C upscaling pipeline with five deep learning-based methods, and (3) a comprehensive evaluation pipeline under a unified framework. Our code is available publicly on GitHub at <https://github.com/rsinghlab/Hi-CY> (accessed on 25 December 2023).

2.1. Hi-C Pre-Processing

After reviewing existing deep learning-based upscaling methods [6,8–11,13], we use Hi-C experiments from GEO Accession GSE63525 for GM12878, IMR90, and K562 as our primary high resolution that we refer to as High Read Count (HRC; synonymous with high-resolution Hi-C contact map) datasets [5]. Similar to previous evaluation [6,8,10,11], we generate 12 downsampled datasets by uniformly downsampling primary GM12878, IMR90, and K562 datasets by factors of 16, 25, 50, and 100. We collect an additional seven experimentally generated Low Read Count (LRC; synonymous with low-resolution Hi-C contact map) Hi-C datasets to evaluate performance in real-world settings on sparse matrices. Five of these LRC datasets are for the GM12878 cell line and have sparsity ranging from $\frac{1}{9}$ to $\frac{1}{100}$ reads compared to the HRC dataset. The remaining two LRC datasets are for IMR90 and K562 cell lines with $\frac{1}{10}$ and $\frac{1}{14}$ reads, respectively. We also include a GM12878 HRC biological replicate cell line [5] in our analysis to calculate the “upper-bound” on the metric performance, which we show as a black dotted line wherever appropriate. We show the absolute read counts, sparsity, and the source experiment of all datasets in Table 1.

For these Hi-C contact maps, we pre-processed them by filtering reads using a MAPQ value of ≥ 30 and performed KR normalization to remove reads with low statistical confidence as well as accounting for the experimental artifacts. We sampled both LRC and HRC datasets at 10 Kbp resolution and only kept the intra-chromosomal contacts to generate twenty-two dense contact matrices. We performed 99.95th percentile normalization to scale all observed contacts between 0 and 1, which has been shown to improve the predictive capabilities of deep learning-based models [10]. Additionally, we cropped out sub-matrices (size depending on each model’s input parameters) across a 2 Mbp range of the diagonal, as this is shown to contain contacts with the highest biological information content [6]. Finally, we divided the data for 22 chromosomes into training (chr1–chr7 and chr12–chr18),

validation (chr8–chr11), and test (chr19–chr22) sets. We exclusively used GM12878 datasets to train our models as is the standard for all methods we compare.

Table 1. Summary of the datasets and their sources. All experiments used the Mbol enzyme and filtered fragments to be in the size range of 300–500 using SPRI beads. HRC refers to high-read-count, and LRC refers to low-read-count Hi-C contact matrices. Sparsity represents the fraction of reads in comparison to the relevant HRC Hi-C contact map.

Dataset	Absolute Read Counts	Sparsity	Source
GM12878-HRC-1	1,844,107,778	1	GEO (GSE63525)
GM12878-HRC-replicate	1,564,534,654	1	GEO (GSE63525)
GM12878-LRC-1	42,453,795	1/44	ENCODE (ENCSR382RFU)
GM12878-LRC-2	37,079,587	1/50	ENCODE (ENCSR382RFU)
GM12878-LRC-3	70,138,184	1/25	ENCODE (ENCSR968KAY)
GM12878-LRC-4	202,380,884	1/9	GEO (GSM1551575)
GM12878-LRC-5	18,696,952	1/100	GEO (GSM1551582)
IMR90-HRC-1	735,043,093	1	GEO (GSE63525)
IMR90-LRC-1	75,193,876	1/10	GEO (GSM1551606)
K562-HRC-1	641,402,880	1	GEO (GSE63525)
K562-LRC-1	44,882,605	1/14	GEO (GSM1551622)

2.2. Deep Learning-Based Hi-C Upscaling Models

We set up five state-of-the-art deep learning-based Hi-C upscaling methods divided into two broad categories—those that employ an adversarial loss to optimize weights and those that do not. HiCPlus [6] uses a three-layer convolutional network to optimize with Mean Squared Error (MSE) loss. HiCNN [8] extends the HiCPlus approach to a 54-layer architecture that uses residual connections between the CNN layers [18], improving the performance, training times, and stability. HiCNN2 [9] combines HiCNN and HiCPlus with a VDSR (Very Deep Super Resolution) model to generate an output that is a linear combination, with learned weighted contribution, of all the networks. HiCGAN [11], DeepHiC [10], and VeHiCle [13] employ Generative Adversarial Networks. GANs [19] jointly optimize two models—a generator that produces samples and a discriminator that tells fake samples apart from real ones—to learn to create increasingly more realistic outputs. HiCGAN uses an MSE loss and cross entropy (computed through discriminator output) to optimize the weights. DeepHiC extends HiCGAN to introduce a perceptual and total variation loss to generate Hi-C contact maps with sharper and more realistic features. Lastly, VeHiCle makes two modifications to the DeepHiC approach: (1) it replaces the perceptual loss with a domain-specific Hi-C loss using an unsupervised model trained to generate the Hi-C input, and (2) it adds insulation loss, forcing the model to learn the underlying biological structure (specifically topologically associated domains) and generate more informative Hi-C matrices. We picked these models because they capture a representative set of methods that capture the overall evolution of deep learning-based Hi-C resolution enhancement methods. To the best of our knowledge, none of the existing methods are explicitly designed to upscale experimentally generated LRC Hi-C contact matrices.

For our evaluations, we retrain HiCPlus, HiCNN, and HiCNN2 to ensure these models generate an output of value between [0, 1]. These three methods predict raw contact counts, making the generated Hi-C matrices incomparable with GAN-style methods and putting these models at a disadvantage against other methods as shown by DeepHiC [10]. Similar to DeepHiC, we train four different versions, each on $\frac{1}{16}$, $\frac{1}{25}$, $\frac{1}{50}$, and $\frac{1}{100}$ downsampled datasets to provide a range of Hi-C input sparsities [10]. When upscaling, we utilize the version with the smallest difference between the sparsity of input Hi-C data and the downsampled Hi-C data we used to train the method. We show the training curves in Supplementary Figures S1–S3. Note that we exclude HiCGAN [11] from our comparisons, as we could not set up the provided code base, as it depended on depreciated Python packages. For DeepHiC and VeHiCle, we used the provided weights. As a baseline

upsampling algorithm, we added Gaussian smoothing with a kernel size of $n = 17$ and 2D Gaussian distribution with $\sigma_x = \sigma_y = 7$. We found these parameters to provide best performance. We keep the architecture the same for all the retrained methods. We do not tune layer parameters to ensure that we compare the off-shelf versions of these methods to test their real-world applicability. Lastly, we use the same merging algorithms provided by these methods to combine the predicted samples into full chromosome contact maps.

2.3. Evaluation Metrics

Recent work has shown that correlation metrics, such as Pearson's Correlation Coefficient (PCC) and Spearman's Correlation Coefficient (SCC), fail to assign an accurate reproducibility score for Hi-C experiments due to a limitation in accounting for the underlying data distributions (e.g., distance effect in Hi-C matrices) [14]. Unfortunately, most of the current Hi-C upscaling studies have used PCC and SCC along with the Structural Similarity Index Metric (SSIM), which calculates similarity using the two images' luminance, contrast, and structure to evaluate their performance. Due to the limitations of these correlation-based metrics, we include the Hi-C similarity metrics [14]—GenomeDISCO, HiCRep, Hi-C-Spector, and QuASAR-Rep—in our evaluation pipeline. We show the results for GenomeDISCO as our main metric, which uses random graph walks for smoothing out the contact matrices at varying scales to compute the concordance scores between the two input maps. Similarity across these smoothed contacts corresponds to similarity at various genomic organizational scales [20]. We show the results for the rest of the metrics (and explain their mechanisms to compute similarity) in Supplementary Section S3. For all of the reported metrics, a higher score (closer to 1) is better.

2.4. Downstream Analyses for Biological Validation

For a biologically informed evaluation, we perform four additional downstream analyses to assess the utility of information recovered from the upscaled Hi-C matrices. While Hi-C similarity scores produced by GenomeDISCO, QuASAR-Rep, HiC-Rep, and HiC-Spector are developed to compare the higher-order structure of the chromatin that can be recovered through Hi-C contact maps, they do not compare biological features such as TADs and chromatin loops. Therefore, we evaluate the quality of 3D reconstruction, recovered chromatin loops, TADs, and DNA hairpins from upscaled Hi-C maps to provide a thorough analysis of the utility of the predicted Hi-C contact maps in their real-world downstream use case.

2.4.1. Quality of Chromatin Reconstruction from Upscaled Hi-C Maps

We use 3DMax [21] to generate a 3D model of chromatin from both the HRC and the upscaled Hi-C matrices. 3DMax uses iterative correction with eigen-decomposition to denoise and normalize the Hi-C contact map before constructing the 3D model. We use default 3DMax parameters and compare the 3D models using the Template Modeling score (TM-score), which is typically used in protein–structure comparison, to estimate the reconstruction accuracy. A score closer to 1 indicates a more similar 3D structure, which in our case corresponds to having similar genome organization and function. Typically, a score greater than 0.5 suggests that 3D models have similar underlying structures [22].

2.4.2. Quality of Recovered Chromatin Loops, TADs, and DNA Hairpins from the Upscaled Hi-C Maps

To analyze the structural features of the genome, we used Chromosight [23] on Hi-C contact matrices to call chromatin loops, TADs, and DNA hairpins. Recent works have shown these features to be crucial in understanding gene regulation, disease, and genome organization [3,5,24]. We recovered these features from the Hi-C matrices using Chromosight [23] with its default parameters and detection kernels. Given Chromosight's ability to recover an extensive set of biologically informative features [23], we consider the chromatin features retrieved from the HRC matrices to be a “ground truth” feature set

and compare the derived features from the upscaled matrices against them. We count the following: (1) True Positives (TP) that overlap (we consider an overlap if the position of both features are within n bins of each other; we tune the value of n for all features individually as summarized in Supplementary Figure S4) in both matrices; (2) False Positives (FP), features that are called on the imputed matrices but are not present in the HRC matrices; and (3) False Negatives (FN), features in HRC matrices that were absent in the imputed matrices. Then, we compute the F1 score using:

$$\text{F1 score} = \frac{2 \cdot TP}{2 \cdot TP + FP + FN}$$

3. Results

In this section, we detail the results of the following experimental steps: (1) reproduce previous results using our Hi-CY framework, (2) use Hi-CY to show that deep learning-based methods' performance does not generalize to experimentally generated LRC datasets, (3) characterize the potential issues leading to poor generalization, and (4) present potential strategies to improve the generalization of these methods on experimentally generated LRC datasets. For all presented results, we show region chr22:41-43Mbp to qualitatively visualize the feature recovery performance. We pick this region because it contains a high density of chromatin features, including TADs, chromatin loops, and DNA hairpins. The quantitative evaluations show averaged cross-chromosome scores on all four (chr19-chr22) test chromosomes for all metrics and downstream analyses.

3.1. Hi-Cy Framework Reproduces Similar Performance of Deep Learning-Based Models on Downsampled Datasets

All existing deep learning-based Hi-C resolution upscaling models [6,8–11] show that they can achieve correlation and reproducibility scores comparable to or better than a biological replicate when upscaling a downsampled Hi-C contact map. These methods show their performance on Hi-C contact maps with a downsampling ratio of up to $\frac{1}{100}$ reads compared to the output HRC contact map [10]. We reproduce performance trends using Hi-CY to establish the reproducibility and accuracy of our approach. We generate 12 Hi-C datasets for three cell lines—GM12878, IMR90, and K562—downsampled by factors of $\frac{1}{16}$, $\frac{1}{25}$, $\frac{1}{50}$ and $\frac{1}{100}$ of the reads in the corresponding HRC matrix. We evaluate the performance by comparing the generated Hi-C contact maps against the HRC contact maps. In Figure 2A, we visualize the predicted Hi-C contact maps when using an input GM12878 contact map downsampled by a factor of $\frac{1}{50}$ and compare them against the target HRC GM12878 dataset. All methods, excluding the baseline Gaussian Smoothing, can recover the finer features, including the sub-TAD structures highlighted by a blue dotted rectangle, with DeepHiC generating the most visually similar Hi-C matrices compared to the target HRC contact map.

We quantify the performance in Figure 2B,C by comparing SSIM and GenomeDISCO scores (on the y-axis) across all three cell lines and four different downsampling ratios (on the x-axis). Our results demonstrate that DeepHiC, HiCNN, HiCNN2, and HiCPlus perform better than or comparable to the biological replicate (shown as a dotted black line) on SSIM and GenomeDISCO metrics. Moreover, these methods generalize to both the downsampling ratios and cross-cell type inputs. We show the results on the rest of the metrics in Supplementary Figure S5, which further support the insights we gathered from the SSIM and GenomeDISCO metrics. In contrast, VeHiCLe offers consistent but lower performance than other methods on both metrics. We expected this result given that VeHiCLe was trained and evaluated on a experimentally generated LRC Hi-C dataset (GSE63525 HIC0001) [13]. These results indicate that Hi-CY can train and evaluate the existing methods in a unified pipeline and reproduce results from the previous studies.

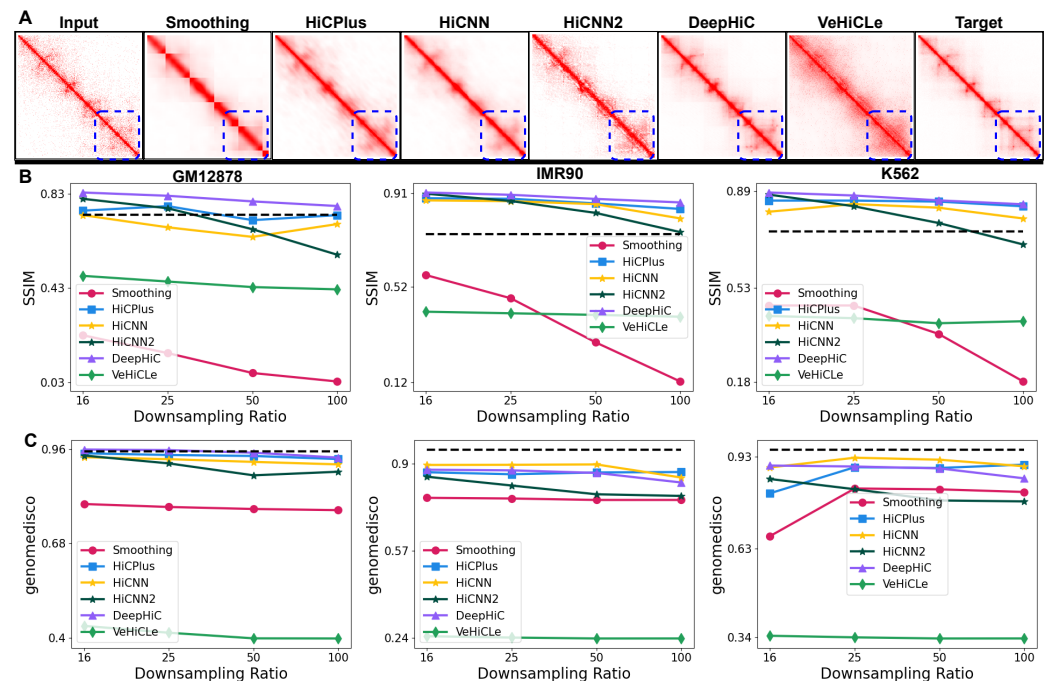


Figure 2. Hi-CY can reproduce the performance of deep learning-based methods on upscaling downsampled input Hi-C datasets. (A) Deep learning-based methods' output for chr22:41–43Mbp with input reads downsampled to 1/50 counts of the original HRC Hi-C matrix. (B,C) On the x-axis, we present the downsampling ratio in increasing order, and on the y-axis, we report SSIM (panel B) and GenomDISCO (panel C). As the downsampling ratio increases, HiCPlus, DeepHiC, HiCNN, and HiCNN2 show similar or comparable performance compared to the biological replicate shown as a dotted black line. VeHiCLe and Gaussian Smoothing baseline give a lower performance on these datasets.

3.2. Testing Generalization of Deep Learning-Based Models on Experimentally Generated Hi-C Datasets

We use the Hi-CY framework to evaluate the performance of existing models on experimentally generated LRC Hi-C datasets and contrast it with their performance on the downsampled Hi-C datasets we used in the previous experiment. We test generalizability on five datasets from the GM12878 cell line, with the read count sparsity ranging from $\frac{1}{9}$ to $\frac{1}{100}$ of the total reads of the GM12878 HRC Hi-C matrix. We further test one dataset each from K562 and IMR90 with a sparsity of $\frac{1}{14}$ and $\frac{1}{10}$, respectively, to evaluate the cross-cell-type generalization. First, we compare the output produced by the models when provided with the GM12878-LRC-4 dataset with $\frac{1}{50}$ of the reads of the GM12878 HRC Hi-C contact map. As shown in Figure 3A all methods, including VeHiCLe, struggle to recover chromatin features, particularly the sub-TAD features highlighted with a dotted blue rectangle, which they could recover when provided with a downsampled contact map with a similar number of reads. Next, as shown in Figure 3B, we observe that none of the methods can achieve scores similar to the biological replicate (shown with the dotted black line). Moreover, except VeHiCLe, all methods show a decrease in performance as the sparsity increases. For example, comparing the HiCNN results on the GM12878 datasets with $\frac{1}{9}$ to $\frac{1}{100}$ of the reads, SSIM goes from 0.77 to 0.45 while GenomDISCO goes from 0.94 to 0.4. Although VeHiCLe shows stable scores, they are lower than other methods for SSIM and GenomDISCO, suggesting that it may have overfit to the dataset that it was trained with (GSE63525 HIC0001).

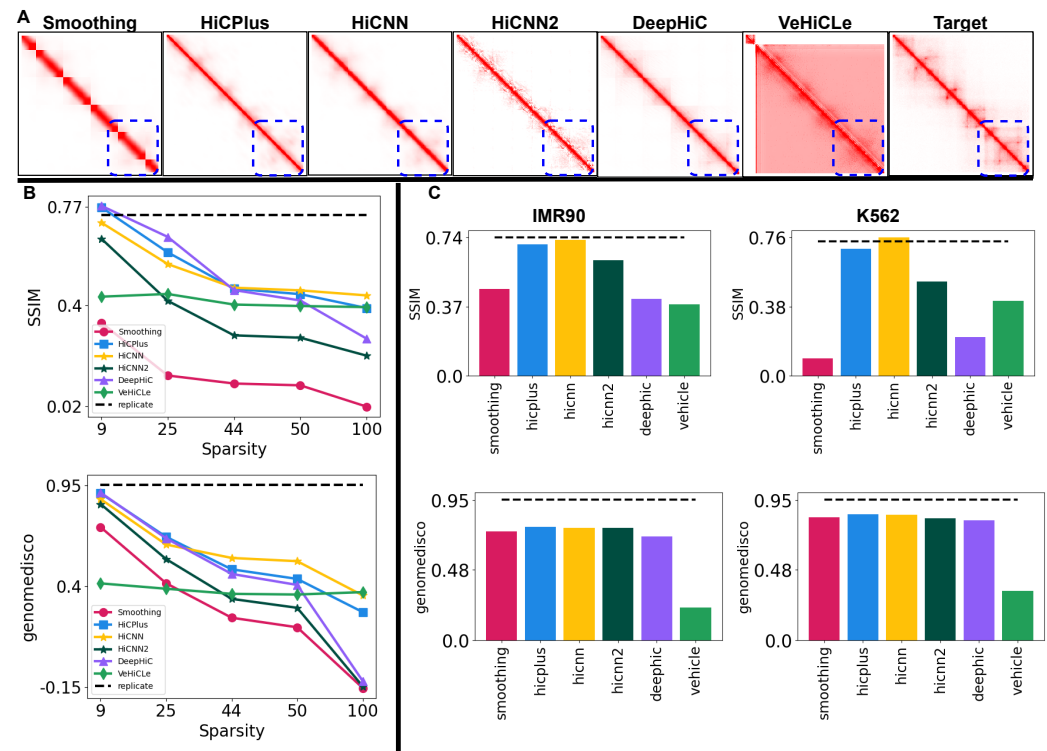


Figure 3. Performance comparison for upscaling experimentally generated LRC Hi-C datasets. **(A)** We visualize the outputs of our methods on the GM12878-LRC-2 dataset with a read count of $\frac{1}{50}$ of the total reads of GM12878 HRC Hi-C map for chr22:41–43 Mbp. We find that all methods struggle to recover the finer chromatin features in comparison to when provided with a downsampled Hi-C contact map. **(B)** We quantify the decrease in performance of the upscaling methods as we increase the sparsity (on the x-axis) of the GM12878 LRC datasets on the SSIM and GenomeDISCO metrics (on the y-axis). We observe that all methods show a substantial drop in performance as we increase sparsity, with HiCNN showing better robustness to the sparsity of data in comparison to other methods. **(C)** We show a drop in the performance of these methods on the IMR90 and K562 LRC datasets by comparing them against a biological replicate score shown with a dotted black line. We observe a substantial drop in performance in both cell lines on the GenomeDISCO metrics. Our results **(B,C)** show that all methods fail to generalize to sparse, experimentally generated Hi-C datasets.

Lastly, we test how these deep learning-based models generalize to experimentally generated cross-cell type sparse Hi-C contact maps. Figure 3C shows the SSIM and GenomeDISCO scores on the upscaled matrices when provided with the IMR90-LRC-1 and K562-LRC-1 datasets as inputs. Across the IMR90 and K562 cell lines, both HiCNN and HiCPlus achieve SSIM scores similar to the biological replicate. However, for the GenomeDISCO metric, all methods show a substantial reduction in performance across both cell lines. Our results show that all methods strongly favor downsampled datasets and perform significantly worse on experimentally generated sparse LRC Hi-C datasets. Our results for the rest of the metrics are presented in Supplementary Figure S9, adding further support for our findings. Overall, among the available Hi-C upscaling methods, HiCNN shows the best generalizability to its real-world out-of-the-box use cases with a competitive performance on the GM12878 datasets (particularly highly sparse $\frac{1}{100}$ dataset) and the best performance on IMR90 and K562 datasets. Furthermore, we believe that GAN-based models such as DeepHiC and VeHiCLe are not ideal candidates for retraining given the tendency of GANs to overfit to the training data [25–27], and their unstable training process [27,28] posing a technical challenge outside the scope of this work.

3.3. A Significant Distributional Shift between Downsampled Datasets and Experimentally Generated LRC Datasets Hurts the Generalizability of Deep Learning Models

Lack of generalizability in deep learning models is typically attributed to distributional differences [29]. We compare the distribution of reads in LRC and downsampled Hi-C datasets and the impact of that difference on the underlying chromatin structure. First, we compare the observed reads between an experimentally generated LRC Hi-C contact map and a downsampled contact map with similar reads. In Figure 4A, we plot the log (base 10) difference of region chr22:41-43Mbp between the experimentally generated GM12878 LRC dataset and downsampled datasets with $\frac{1}{50}$ of the reads of the HRC Hi-C contact map as a heat map. The heatmap highlights the region with higher log differences in yellow compared to regions with smaller differences in purple. We observe bins with a higher difference on the heatmap as we move further away from the diagonal. To establish this trend across all the chromosomes (test, train, and validation), we plot PCC between observed reads in experimentally generated or downsampled and HRC datasets at increasing genomic distances. HiCNN, HiCPlus, and DeepHiC [6,8–10] used this analysis to compare the similarity and distribution of Hi-C samples across each genomic distance. In this curve, a higher correlation on all genomic distances suggests a higher read count distribution similarity between the two samples. Figure 4B shows that across all genomic distances, the experimentally generated LRC Hi-C dataset has a lower correlation than the downsampled datasets, with the difference between those two increasing as the distance increases. We observe the same trend, shown in Supplementary Figure S6, across all the experimentally generated LRC datasets we gathered. These observations suggest that experimentally generated LRC Hi-C contact maps have a higher distributional difference with HRC Hi-C contact maps than with downsampled Hi-C contact maps with a similar number of reads. This difference arises because uniformly removing reads to downsample Hi-C contact matrices fails to account for the distance effect [12], suggesting that Hi-C reads are more likely to be observed on genomic bins closer to the diagonal.

We further investigate the impact of this distributional shift by comparing the content and biological information of downsampled datasets against experimentally generated LRC datasets. Figure 4C shows a steeper degradation in GenomeDISCO scores in LRC matrices compared to the downsampled matrices. Given that GenomeDISCO compares the underlying structural topology rather than bin-wise similarity, our results in Figure 4C show that there is a steeper loss in structural information in LRC Hi-C datasets compared to downsampled datasets. Lastly, we quantify the impact of this loss of structural information by comparing the chromatin loop F1 scores between the downsampled and experimentally generated LRC Hi-C contact maps. Unsurprisingly, Figure 4D shows that chromatin loop F1 scores decay more sharply for LRC datasets, while F1 scores approach 0 for both datasets. Our findings across all datasets, as shown in the Supplementary Figures S7 and S8 on other metrics and biological features, further support a more significant information loss in experimentally generated LRC datasets that scales with the sparseness of data. In the cross-cell-type setting, we find an interesting trend that suggests that IMR90-LRC-1 has more loss in structure information than K562-LRC-1, even though it has higher sparsity, suggesting the existence of cell-specific experimental artifacts. Our analysis identifies the distributional shift between the downsampled and LRC datasets as a crucial component that leads to the poor generalizability of deep learning-based models. We further show that upscaling LRC datasets is substantially more complicated due to a lack of structure in the LRC Hi-C matrices and their lower similarity to the target HRC Hi-C matrices.

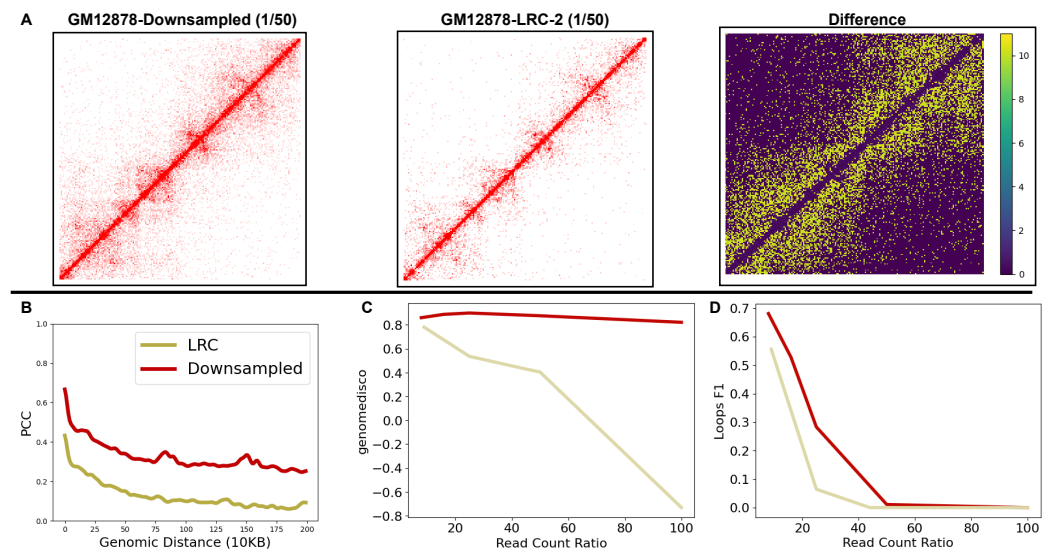


Figure 4. Comparing the distributions of downsampled Hi-C datasets with the experimentally generated LRC datasets. (A) Shows a pixel-wise comparison of $\frac{1}{50}$ experimentally generated LRC dataset with a similarly downsampled dataset for GM12878 chromosome 22 region 41–43 Mbp. As we move further from the diagonal, the difference amplifies (signified by yellow). (B) We compare Pearson's Correlation (on the y-axis) of the LRC dataset (beige) and the downsampled dataset (red) with $\frac{1}{50}$ sparsity against the HRC dataset for strata with increasing genomic distances (on the x-axis). The correlation is always smaller for experimentally generated LRC datasets across all genomic distances, and the difference increases as we move further away from the diagonal. (C) Compares GenomeDISCO score of experimentally generated LRC datasets against a similarly sparse downsampled dataset. Our results show that the GenomeDISCO score is lower for experimentally generated LRC datasets compared to an equally sparse downsampled dataset. (D) We compare the Biological Features recovery F1 score against the HRC feature set for LRC and Downsampled datasets on the GM12878 cell line. The F1 scores decay more steeply for experimentally generated LRC datasets.

3.4. Retraining Deep Learning Models with Experimentally Generated Datasets Improves Performance

Our results in the previous sections show that the distribution of experimentally generated LRC Hi-C data differs from the downsampled Hi-C datasets. Given this insight, we pick the best model—HiCNN—and retrain it with experimentally generated GM12878 LRC datasets. The core idea behind retraining is to expose the deep learning model to a more realistic data distribution during the training phase to improve generalizability in the testing phase. We retrain the HiCNN model on each LRC dataset individually (LRC-1— $\frac{1}{44}$ reads of HRC data; LRC-3— $\frac{1}{25}$ reads of HRC data; and LRC-4— $\frac{1}{9}$ reads of HRC data). We also train on an ensemble of aforementioned LRC datasets (Ensemble-LRC) as a means to expose the model to a range of realistic distribution of Hi-C datasets during the training phase. Figure 5A shows that even with retraining, HiCNN is still unable to recover the sub-TADs in the TAD cluster highlighted with a dotted blue square, even when retrained with an experimentally generated LRC dataset. Figure 5B shows the SSIM and GenomeDISCO scores of three versions of HiCNN trained with LRC-1, LRC-3, or LRC-4 GM12878 datasets and two versions of HiCNN trained either with an ensemble of LRC datasets or an ensemble of downsampled datasets. We evaluate the performance of these models on five GM12878 datasets with varying levels of read count sparsity. We find that the versions of HiCNN retrained with LRC datasets with high sparsity (LRC-1 and LRC-3 and Ensemble-LRC) improve the deep learning model performance on average by 11%, 10%, and 10% on SSIM and 18%, 31%, and 16% on GenomeDISCO in comparison to HiCNN trained with downsampled datasets. Conversely, we observe a reduction in

performance when retrained with a low sparsity (LRC-4), on average of 6% on SSIM and 8% on GenomeDISCO, compared to when trained with downsampled datasets. However, we observe the best performance improvements of 16%, 16%, 14% for LRC-1, LRC-3 and Ensemble-LRC, respectively, on SSIM. We also see improvements of 75%, 68%, 48% for LRC-1, LRC-3, LRC-4 and Ensemble-LRC, respectively, on the GenomeDISCO metric when upscaling GM12878-LRC-5, which is the sparsest dataset, having $\frac{1}{100}$ reads.

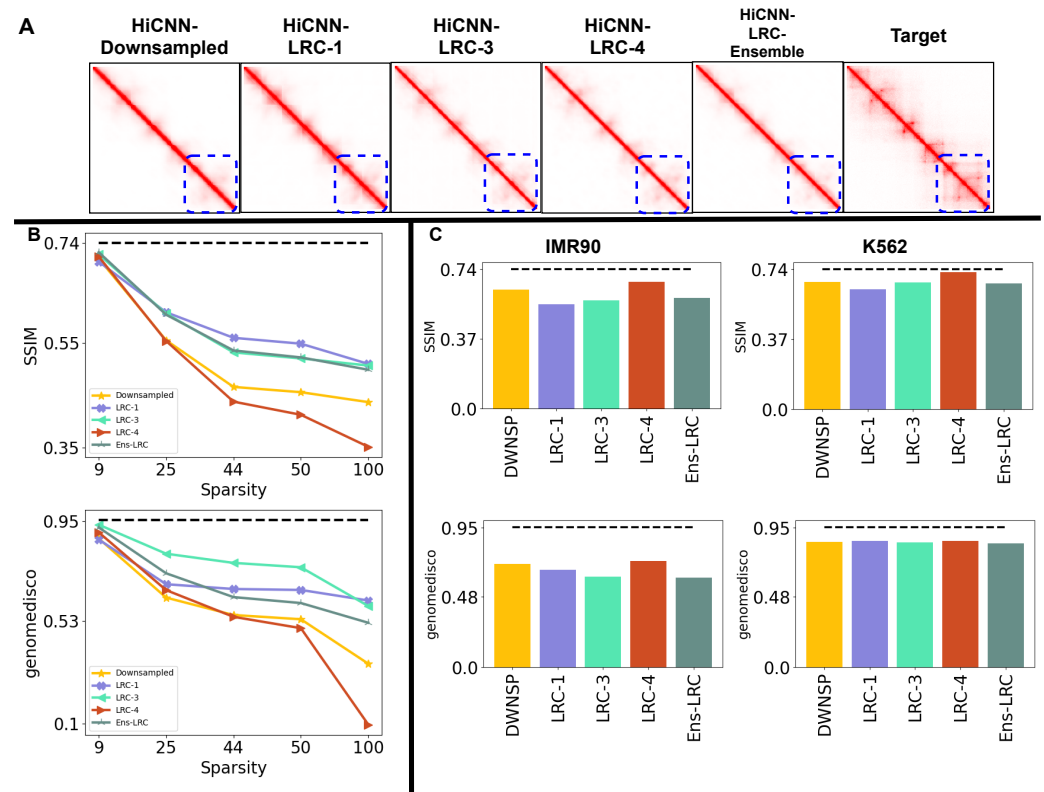


Figure 5. Retraining HiCNN with experimentally generated LRC datasets. **(A)** We visualize the outputs of our methods on GM12878-LRC-2 with $\frac{1}{50}$ on region chr22:41–43Mbp, and we show that all methods struggle to recover finer architectural features, such as TADs. **(B)** We quantify the decrease in performance of upscaling methods as we increase the sparsity (on the x-axis) of GM12878 LRC datasets on the SSIM and GenomeDISCO metrics (on the y-axis). We observe that retraining with LRC datasets or an ensemble of LRC datasets improves performance, with LRC-3 showing the most improvement in the GenomeDISCO metric, and retraining with LRC-1 providing the most improvement for the SSIM score. All versions struggle to achieve scores similar to the biological replicate, shown as the dotted black line. **(C)** We compare the performance of these methods for both the IMR90 and K562 LRC datasets. All methods perform similarly, with HiCNN trained with downsampled datasets performing better on the IMR90 dataset.

Next, we use the IMR90-LRC-1 and K562-LRC-1 datasets to compare the performance of HiCNN retrained with experimentally generated LRC datasets against HiCNN trained with downsampled datasets in the cross-cell-type setting. Figure 5C shows that when evaluated on the K562-LRC-1 dataset, the retrained versions of HiCNN performed similarly to the original HiCNN on both SSIM and GenomeDISCO. However, we observe reductions of 8%, 9%, and 7% when trained with LRC-1, LRC-3, and Ensemble-LRC, and conversely, increases of 7% when trained with LRC-4 on the SSIM scores, respectively. We observe similar GenomeDISCO scores across all variants of HiCNN when evaluated on the IMR90-LRC-1 dataset. Retrained HiCNN does not generalize to unseen cell types. We hypothesize that this happens because distributional differences across cell-type are independent of the read counts. Overall, we demonstrate that retraining HiCNN using LRC datasets

is a promising approach towards improving the generalizability of deep learning-based models because it exposes them to a more realistic Hi-C contact map distribution during the training phase.

3.5. Downstream Analysis to Quantify the Utility of Upsampled Hi-C Contact Maps

To determine whether the upscaled Hi-C matrices retain the necessary biological signals, we analyze the 3D structure of chromatin and the biological features. HiCNN, HiCPlus, and DeepHiC [6,8,10,11] present these results for upscaled outputs from synthetically downsampled Hi-C maps. We perform this downstream analysis on upscaled experimentally generated LRC datasets to provide insights into the performance for real-world applications.

3.5.1. Retrained Deep Learning Models Generate Highly Accurate 3D Structure of Chromatin

For 3D chromatin reconstruction analysis, we use the upscaled Hi-C maps produced by the methods from all seven experimentally generated LRC input matrices on GM12878, IMR90, and K562. For this, we incorporate 3DMax [21] into the Hi-CY framework, which generates the 3D structure for the 200×200 Hi-C sub-matrices for all of the sub-matrices in the test chromosomes. Next, we compare the spatial conformation of the chromatin using the TM-score, where a higher TM-score is proportional to a higher structural similarity of the chromatin. Our analysis follows the VeHiCLe paper [13], where the recovered 3D structure from HRC Hi-C contact map is compared with the 3D structure recovered from the upscaled Hi-C matrices to show that the upscaled matrices have the same structure. We also visualize 3D models of the region chr22:41-43Mbp for upscaled GM12878 in Figure 6A, which shows that the models, except for HiCNN2 and VeHiCLe, are similar in their overall 3D topology compared to the 3D model generated with the HRC Hi-C contact map.

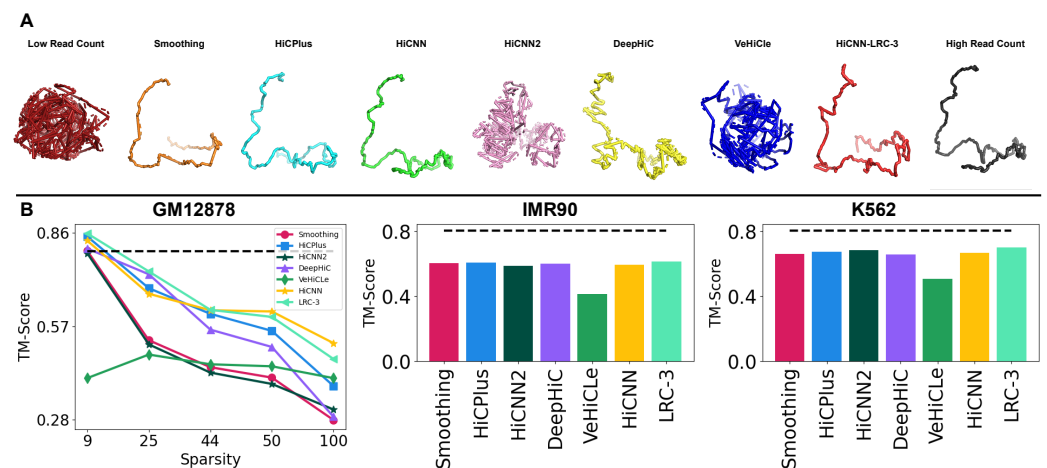


Figure 6. 3D reconstruction of Chromosome 22:41-43Mbp region **(A)** Most models, including HiCNN-LRC-3, produce highly similar 3D structures of the chromatin. **(B)** Our quantitative analysis on 3D reconstruction score comparison suggests that retraining improves performance in certain cases while performing similarly to HiCNN on average.

We summarize the TM-scores in Figure 6B. We observe that, on average, both the HiCNN-Baseline and HiCNN-LRC-3 show similar performance in the 3D reconstruction analysis with scores within 1% of each other. However, although the gains are marginal, we observe that HiCNN-LRC-3 outperforms all methods on both the K562 and IMR90 cell lines. Our results suggest that retraining improves 3D reconstruction performance or results in similar performance to the baseline version of HiCNN, highlighting the value of retraining for recovering biologically informative Hi-C matrices.

3.5.2. Deep Learning Models Have Similar Performance for Recovering Biological Features

We incorporate Chromosight [23] into the Hi-CY evaluation pipeline to recover and compare biological features, including chromatin loops, TADs, and DNA hairpins. To compare the utility of the upscaled Hi-C matrices in recovering biological features, we detect features on all seven experimentally generated LRC datasets upscaled from the complete set of models, including HiCNN retrained with the LRC-3 dataset. In Figure 7, we compare the F1 scores across all seven datasets for all three features we recover. For all models, we observe a trend similar to GenomeDISCO and SSIM, that F1 decreases as we increase the sparsity of the LRC dataset. Retraining with the LRC-3 dataset helps recover features far from the diagonal, such as chromatin loops and TADs, particularly in IMR90, K562, and lower-sparsity GM12878 cases. Regardless, the performance improvements, as observed earlier, are marginal. Surprisingly, HiCNN2 outperforms other methods in the loops and DNA hairpin case in feature recovery analysis on the GM12878 LRC datasets. This provides evidence of the need for multifaceted research that explores results across multiple scenarios and metrics to develop a holistic understanding of the costs as well as the benefits of using deep learning-based models. Currently, all existing methods struggle to recover biological features accurately, particularly in a cross-cell-type setting and on sparse Hi-C datasets, as we see a steep degradation in the F1 score in both scenarios. We observe similar scores or negligible improvements even when we retrain HiCNN with an experimentally generated LRC dataset. This suggests a need to revisit the architectural assumptions as well, including formulating Hi-C as a 2D image when the underlying measurement data correspond to a 3D structure.

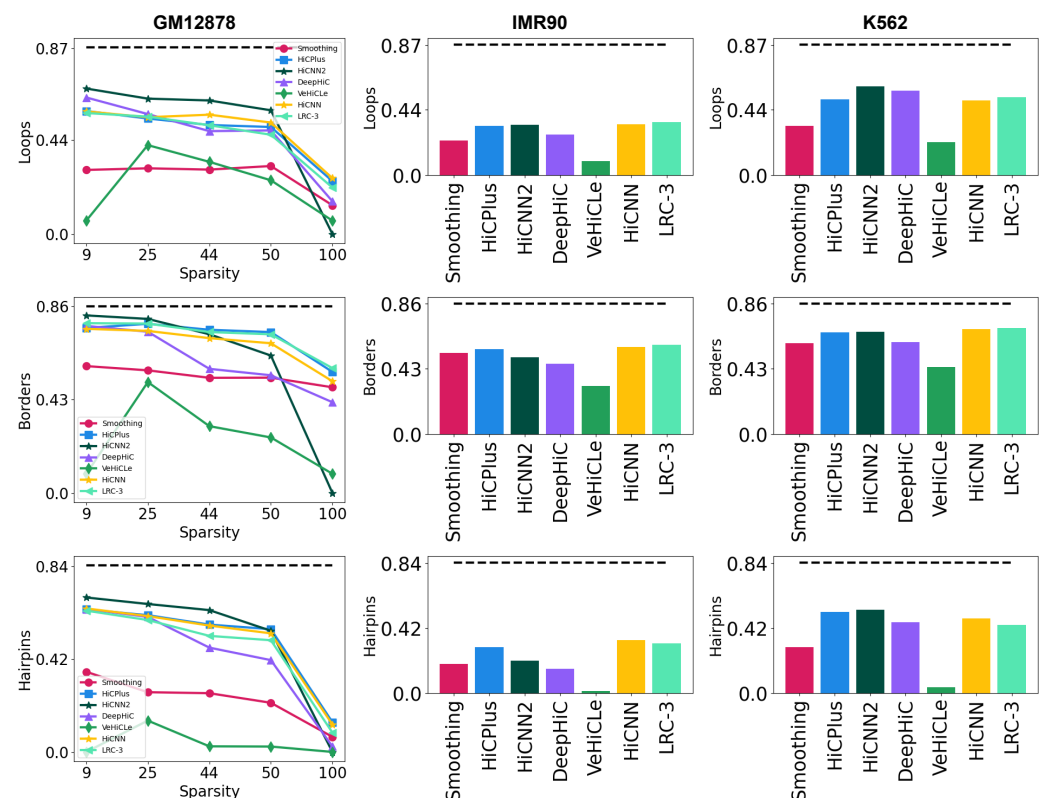


Figure 7. Biological feature comparison. Our quantitative analysis, by comparing the F1 scores, computed by comparing the recovered feature set with the HRC feature set, suggests that across chromatin loops, TADs, and DNA hairpins across all three cell lines, all of the methods, including retrained HiCNN, struggle to recover meaningful biological features in highly sparse settings. While retraining improves performance, the improvements are marginal, similar to the previous analysis.

4. Conclusions

We developed the Hi-CY framework that explores how deep learning-based Hi-C upscaling methods perform in their intended scenarios, which is to upscale experimentally generated sparse Hi-C datasets. Our results strongly suggest that existing deep learning-based methods do not generalize to their real-world use case. To provide potential solutions to improve generalizability, we explore retraining with dataset augmentations (such as adding Gaussian noise to the training datasets shown in Supplementary Figure S10), the ensembling of multiple downsampled datasets, and training with experimentally generated Hi-C datasets. Our analysis found that retraining with experimentally generated sparse datasets was the most promising approach to improving generalizability. Although we did not observe a significant improvement in the 3D structure generation or recovering chromatin loops, TADs, and DNA hairpin analysis tasks, we still recommend retraining with experimentally generated Hi-C datasets to improve their generalizability beyond the training datasets. Our Hi-CY framework provides a simple yet robust pipeline to streamline the training and evaluation of Hi-C upscaling methods with real-world Hi-C datasets on correlation, HiC similarity, and downstream analysis-based metrics.

Another trend we observe in our evaluation is that among the currently available techniques, HiCNN, which uses a non-adversarial model, outperforms the GAN-based models in generalizing to sparse experimentally generated Hi-C datasets. Interestingly, we observe that an even simpler method, HiCPlus (3-layer CNN), performs comparably to HiCNN (a 54-layer CNN) on most metrics across most datasets. In future work, we plan to investigate if this similar performance stems from formulating Hi-C data as an image and the upscaling as a super-resolution task. We plan to explore performance gains from using graph-based architectures to capture the underlying geometry of the DNA molecule more accurately. In conclusion, we present the Hi-CY framework, which provides a comprehensive evaluation toolkit coupled with real-world datasets that support the development of future models by investigating their performance on intended use cases.

Supplementary Materials: The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/genes15010054/s1>, Figure S1: This figure shows the loss curves for HiCPlus models on four version we trained with each downsampling ratio. We use the version of model (across all the epochs) that minimizes the validation loss. Figure S2: This figure shows the loss curves for HiCNN models on four version we trained with each downsampling ratio. We use the version of model (across all the epochs) that minimizes the validation loss. Figure S3: This figure shows the loss curves for HiCNN2 models on four version we trained with each downsampling ratio. We use the version of model (across all the epochs) that minimizes the validation loss. Figure S4: Overview of our Biological Feature Analysis Pipeline: A: We define True Positives as features that we find to be overlapping between the upscaled Hi-C matrix and the HRC matrix. Similarly, we define features that are present in the HRC matrix and not in the upscaled Hi-C matrix as False Negatives. Conversely, features we find in upscaled matrix and not in the HRC matrix as the false positives. We compute F1 scores using these definitions. B: We optimize the parameter “r” (relaxation parameter) for each feature, chromatin loops, TADs and DNA-hairpins. This relaxation parameter defines the overlapping radius between the position of features between the upsampled and HRC features. We tune this parameter by comparing features between the GM12878 HRC and GM12878 biological replicate cell lines and find the value of r that maximizes the F1 score while keeping the value of r small. We found a value of 3 for chromatin loops, 5 for TADs and 2 for DNA-hairpins. C: We summarize our biological feature analysis pipeline, we first extract biological features using Chromosight and then compute F1 scores. Figure S5: This figure shows the performance of deep-learning based Hi-C upscaling methods on computationally downsampled methods on four downsampling ratios (shown on x-axis), three cell-lines five metrics. All methods except VeHiCLe show performance very similar to the replicate shown as black dotted bold line. Figure S6: Our results on comparing the PCC (y-axis) across various genomic distances (x-axis) suggest that LRC datasets show a smaller similarity with HRC Hi-C dataset in both increasing levels of read sparsity and in cross-cell-type cases. Figure S7: Our results on comparing the PCC (y-axis) across various genomic distances (x-axis) suggest that LRC datasets show a smaller similarity with HRC Hi-C dataset in both increasing levels of read sparsity and in

cross-cell-type cases. Figure S8: We show the impact of the distributional difference in the structural information contained in downsampled and experimentally generated LRC Hi-C contact maps by comparing scores across both correlation and Hi-C similarity metrics. Figure S9: This figure shows performance on rest of the metrics, and our results highlight that there is a reduction in performance as the sparsity increases in GM12878 datasets. Figure S10: We show results of retraining across five additional metrics and also across additional retraining variants, such as downsampled datasets augmented with noise.

Author Contributions: R.S. conceived the idea that is presented in the paper. G.M. and A.J. re-implemented all the baseline methods in the latest PyTorch libraries. G.M. collected all the datasets, and developed standardized pre-processing scripts and evaluation scripts. G.M. and M.H. ran all the experiments. G.M. wrote the initial draft. R.S., J.W., A.J. and M.H. supported the writing phase by providing feedback on the draft. All authors have read and agreed to the published version of the manuscript.

Funding: Ghulam Murtaza's effort on this project was funded by the NIST PREP grant GR5245041. Ritambhara Singh's contribution to the work was supported by NIH award 1R35HG011939-01.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: All the scripts, programs and datasets used to conduct this analysis are available at <https://github.com/rsinghlab/Hi-CY> (accessed on 25 December 2023).

Acknowledgments: We thank Justin Zook from the National Institute of Standards and Technology (NIST) for providing incredibly valuable feedback on the draft. We also would like to thank Hyeyeon Hwang for sharing her valuable experience about Hi-C downstream analysis methods and having many helpful discussions with us on the draft. Lastly, we would also thank the Compbio Core at Brown for their support in setting up the resources required to conduct the experiments. Certain commercial equipment, instruments or materials are identified to adequately specify experimental conditions or reported results. Such identification does not imply recommendation or endorsement by NIST, nor does it imply that the equipment, instruments or materials identified are necessarily the best available for the purpose.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. 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.; et al. Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. *Science* **2009**, *326*, 289–293. [[CrossRef](#)] [[PubMed](#)]
2. 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* **2012**, *485*, 376–380. [[CrossRef](#)] [[PubMed](#)]
3. Rao, S.S.P.; Huntley, M.H.; Durand, N.C.; Stamenova, E.K.; Bochkov, I.D.; Robinson, J.T.; Sanborn, A.; Machol, I.; Omer, A.D.; Lander, E.S.; et al. A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. *Cell* **2014**, *159*, 1665–1680. [[CrossRef](#)] [[PubMed](#)]
4. Pugacheva, E.M.; Kubo, N.; Loukinov, D.; Tajmul, M.; Kang, S.; Kovalchuk, A.L.; Strunnikov, A.V.; Zentner, G.E.; Ren, B.; Lobanenkov, V.V. CTCF mediates chromatin looping via N-terminal domain-dependent cohesin retention. *Proc. Natl. Acad. Sci. USA* **2020**, *117*, 2020–2031. [[CrossRef](#)] [[PubMed](#)]
5. Jin, F.; Li, Y.; Dixon, J.R.; Selvaraj, S.; Ye, Z.; Lee, A.Y.; Yen, C.-A.; Schmitt, A.D.; Espinoza, C.A.; Ren, B.; et al. A high-resolution map of the three-dimensional chromatin interactome in human cells. *Nature* **2013**, *503*, 290–294. [[CrossRef](#)]
6. 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.* **2018**, *9*, 750. [[CrossRef](#)]
7. Nagano, T.; Lubling, Y.; Stevens, T.J.; Schoenfelder, S.; Yaffe, E.; Dean, W.; Laue, E.D.; Tanay, A.; Fraser, P. Single-Cell hi-C reveals cell-to-cell variability in chromosome structure. *Nature* **2013**, *502*, 59–64. [[CrossRef](#)]
8. Liu, T.; Wang, Z. HiCNN: A very deep convolutional neural network to better enhance the resolution of Hi-C data. *Bioinformatics* **2019**, *35*, 4222–4228. [[CrossRef](#)]
9. Liu, T.; Wang, Z. HiCNN2: Enhancing the Resolution of Hi-C Data Using an Ensemble of Convolutional Neural Networks. *Genes* **2019**, *10*, 862. [[CrossRef](#)]
10. Hong, H.; Jiang, S.; Li, H.; Du, G.; Sun, Y.; Tao, H.; Quan, C.; Zhao, C.; Li, R.; Li, W.; et al. DeepHiC: A generative adversarial network for enhancing HiC data resolution. *PLoS Comput. Biol.* **2020**, *16*, e1007287. [[CrossRef](#)]

11. Liu, Q.; Lv, H.; Jiang, R. hicGAN infers super resolution Hi-C data with generative adversarial networks. *Bioinformatics* **2019**, *35*, i99–i107. [\[CrossRef\]](#)
12. Cook, K.B.; Hristov, B.H.; Le Roch, K.G.; Vert, J.P.; Noble, W.S. Measuring significant changes in chromatin conformation with accost. *Nucleic Acids Res.* **2020**, *48*, 2303–2311. [\[CrossRef\]](#)
13. Highsmith, M.; Cheng, J. VeHiCle: A variationally encoded HiC loss enhancement algorithm for improving and generating hi-C data. *Sci. Rep.* **2021**, *11*, 8880. [\[CrossRef\]](#)
14. Yardimci, G.G.; Ozadam, H.; Sauria, M.E.; Ursu, O.; Yan, K.-K.; Yang, T.; Chakraborty, A.; Kaul, A.; Lajoie, B.R.; Song, F.; et al. Measuring the reproducibility and quality of hi-C data. *Genome Biol.* **2019**, *20*, 57. [\[CrossRef\]](#)
15. Toneyan, S.; Tang, Z.; Koo, P.K. Evaluating deep learning for predicting epigenomic profiles. *Nat. Mach. Intell.* **2022**, *4*, 1088–1100. [\[CrossRef\]](#)
16. Huang, K.; Fu, T.; Glass, L.M.; Zitnik, M.; Xiao, C.; Sun, J. DeepPurpose: A Deep Learning Library for drug–target interaction prediction. *Bioinformatics* **2020**, *36*, 5545–5547. [\[CrossRef\]](#)
17. Forcato, M.; Nicoletti, C.; Pal, K.; Livi, C.M.; Ferrari, F.; Bicciato, S. Comparison of computational methods for hi-C data analysis. *Nat. Methods* **2017**, *14*, 679–685. [\[CrossRef\]](#)
18. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. *arXiv* **2015**, arXiv:1512.03385.
19. Goodfellow, I.J.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative Adversarial Networks. *arXiv* **2014**, arXiv:1406.2661.
20. Ursu, O.; Boley, N.; Taranova, M.; Wang, Y.X.; Yardimci, G.G.; Stafford Noble, W.; Kundaje, A. GenomeDISCO: A concordance score for chromosome conformation capture experiments using random walks on contact map graphs. *Bioinformatics* **2018**, *34*, 2701–2707. [\[CrossRef\]](#)
21. Oluwadare, O.; Zhang, Y.; Cheng, J. A maximum likelihood algorithm for reconstructing 3D structures of human chromosomes from chromosomal contact data. *BMC Genom.* **2018**, *19*, 161. [\[CrossRef\]](#)
22. Kadauke, S.; Blobel, G.A. Chromatin loops in gene regulation. *Biochim. Biophys. Acta (BBA)-Gene Regul. Mech.* **2009**, *1789*, 17–25. [\[CrossRef\]](#)
23. Matthey-Doret, C.; Baudry, L.; Breuer, A.; Montagne, R.; Guiglielmoni, N.; Scolari, V.; Jean, E.; Campeas, A.; Chanut, P.H.; Oriol, E.; et al. Computer vision for pattern detection in chromosome contact maps. *Nat. Commun.* **2020**, *11*, 5795. [\[CrossRef\]](#)
24. Bikard, D.; Celine, L.; Baharoglu, Z.; Mazel, D. Folded DNA in action: Hairpin Formation and biological functions in prokaryotes. *Microbiol. Mol. Biol. Rev.* **2010**, *74*, 570–588. [\[CrossRef\]](#)
25. Thanh-Tung, H.; Tran, T.; Venkatesh, S. Improving generalization and stability of generative adversarial networks. *arXiv* **2019**, arXiv:1902.03984.
26. Li, K.; Malik, J. On the implicit assumptions of Gans. *arXiv* **2018**, arXiv:1811.12402.
27. Bau, D.; Zhu, J.-Y.; Wulff, J.; Peebles, W.; Strobel, H.; Zhou, B.; Torralba, A. Seeing what a gan cannot generate. *arXiv* **2019**, arXiv:1910.11626.
28. Zhang, Z.; Li, M.; Yu, J. On the convergence and mode collapse of gan. In Proceedings of the SIGGRAPH Asia 2018 Technical Briefs, Tokyo, Japan, 4–7 December 2018. [\[CrossRef\]](#)
29. Koh, P.W.; Sagawa, S.; Marklund, H.; Xie, S.M.; Zhang, M.; Balsubramani, A.; Hu, W.; Yasunaga, M.; Phillips, R.L.; Gao, I.; et al. WILDS: A Benchmark of in-the-Wild Distribution Shifts. *arXiv* **2021**, arXiv:2012.07421.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.