A Comprehensive Literature Review on Enhancing Hi-C Resolution with Deep Learning

Meghana Indukuri[1], Carlos Rojas[2]

[1] San José State University, San Jose, United States

Abstract

Hi-C data is an effective tool for analyzing the 3D genome. However, the expenses associated with generating high-resolution Hi-C data are a significant obstacle. This literature review presents a comprehensive overview of the deep learning techniques that have emerged in response to this challenge. Specifically, we examine three major machine learning paradigms in the domain of Hi-C super resolution: Convolutional Neural Networks, Generative Adversarial Networks, and Autoencoders. We also explore down-simulation tools and benchmarking methodologies for Hi-C super resolution machine learning models, all while discussing potential future research directions.

1. Introduction

The advent of High-throughput Chromosome Capture (Hi-C) in 2009 [1] has widely altered the landscape for mapping and analyzing the three-dimensional (3D) genome. By capturing the interaction frequencies between genomic loci, Hi-C has revealed the complex 3D structures of the genome and their effects on gene expression and function [2]. Notably, this technique has unveiled the two genomic compartments, A and B, and further revealed their hierarchical organization into topologically associating domains (TADs) and chromatin loops [2, 3].

Hi-C data is generated through a similar workflow to that of sequencing a genome which includes mapping genomic reads, assigning reads to fragments, filtering, and binning. Binning divides the genome into equal-sized bins, with each bin representing a specific genomic region [3]. The output of this procedure is represented by an $N \times N$ matrix, where each cell displays the interaction frequencies between genomic bins. The bin size determines the resolution of the Hi-C data (e.g, a 10 kb bin size refers to 10 kb resolution). Brighter red areas in the Hi-C matrix indicate higher levels of genomic interactions, which is directly correlated to a closer 3D distance between the genomic regions, as shown in Figure 1.

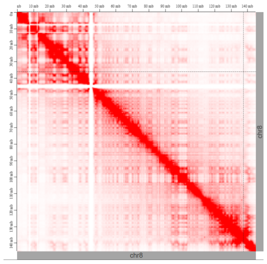

Figure 1. Hi-C Matrix of chromosome 8 from GM12878 cell line, generated using JuiceBox [4]

To derive an effective depiction of interactions between genomic loci the genome must be extensively sequenced. For example, sequencing approximately 100 million genomic reads would result in only a 40 kb resolution for Hi-C data. As the amount of sequencing increases, the size of the bin becomes smaller, resulting in higher resolution matrices [3]. As the matrix resolution increases, more information about the genome's inner 3D organizational structures can be derived. Researchers can make biological discoveries such as showing that TAD disruptions are linked to various diseases [5]. However, sequencing a genome to generate high-resolution Hi-C matrices is expensive. For 100 million reads, it costs approximately $1150 per sample [6]. These costs increase because experiments will require greater resolution and multiple samples to include the control and experiment.

Machine learning (ML) and recently deep learning (DL) have generated state-of-the-art results in a wide-variety of fields within bioinformatics. DL provides a cost-effective approach to generate high-resolution images from low-resolution inputs, with inspiration from methods in Computer Vision such as Convolutional Neural Networks (CNNs), Generative Adversarial Networks (GANs), and Autoencoders (AEs) [7–9]. These models leverage neural networks to enhance their capabilities beyond human limitations [7]. Deep learning models utilize numerous layers of neural networks to extract and map features from a large input dataset. Many Hi-C super-resolution methods utilize variations of these DL techniques by treating Hi-C matrices as images, with each interaction frequency analogous to a pixel [10–15].

This review surveys existing super-resolution DL frameworks and their application to Hi-C data, exploring both foundational works and current advancements in the Hi-C domain. We address the common benchmarking and quality assessment techniques for DL-generated high-resolution Hi-C matrices. We also examine the concept of down simulation and how this methodology may be utilized for the Hi-C super-resolution problem. A comprehensive understanding of existing DL models in this field could improve their real-world applicability and provide a baseline for future DL research in Hi-C data resolution.

2. Benchmarking and Quality Assessments

To effectively assess deep learning approaches for Hi-C super-resolution, it is first essential to understand the common benchmarking metrics used for comparing predicted Hi-C matrices with real Hi-C data. As with many super-resolution problems, determining the quality of these deep learning approaches is required to verify if a model performs well both during training and afterward.

Currently, several rudimentary methods are utilized for benchmarking and quality assessments. The **Mean Squared Error (MSE)** is defined as the average of the squared differences between each element of the real high-resolution matrix and its corresponding predicted value over all entries $N \times N$. A lower MSE indicates a closer match between the predicted and actual matrices. In addition, **Pearson's Correlation Coefficient** is used to check the linear similarity between two Hi-C matrices. The closer the value is to 1, the more similar the matrices are. Similarly, **Spearman's Rank Correlation Coefficient** can be utilized for Hi-C data; however, it does not demand linear similarity and can handle non-linear relationships. Furthermore, the **Structural Similarity Index (SSIM)** checks how similar two Hi-C maps are in terms of interaction frequency values and overall structure. The higher the index value, the more similar the matrices are. This methodology was originally developed to assess image quality, and is now applicable to Hi-C matrices [16].

Beyond these rudimentary techniques, there also exist Hi-C-focused benchmarking methods, which take into account the biological aspects of Hi-C data. For example, **GenomeDisco** [17] works by converting Hi-C matrices into graphs. Once they are in graphs, it compares the structure and similarity between the matrices, providing a reproducibility score. This score, when high, indicates that the two matrices are similar. Additionally, **HiCSpector** [18] works by converting Hi-C matrices into eigenvectors and eigenvalues. By doing so, the tool can produce the spatial relationships and organization of the data that may be lost in more naive methods (i.e. Pearson's correlation). HiCSpector produces a reproducibility score that indicates whether or not two Hi-C matrices are similar. Finally, **HiCRep** [19] also produces a reproducibility score that helps to understand the spatial locality of Hi-C data. The authors of this tool developed the Stratum-Adjusted Correlation Coefficient to compare two Hi-C matrices while accounting for their biological characteristics.

3. Deep Learning Architecture for Hi-C Resolution

To fully understand the application of DL models for enhancing Hi-C data resolution, we examine three different DL paradigms: Convolutional Neural Networks (CNNs), Generative Adversarial Networks (GANs), and Autoencoders (AEs). This section briefly explores each model's application in general computer vision and describes the corresponding works in Hi-C super-resolution.

3.1. Convolutional Neural Networks

CNNs have emerged as a prominent approach for solving a variety of computer vision problems, such as increasing the resolution of images [8]. Specifically, this architecture utilizes multiple convolutional layers to analyze an image, mapping low-resolution features to their high-resolution counterparts [7,8]. CNNs also take advantage of an image's spatial locality and resolution via convolutional filters, making them an important tool for understanding the inherent 3D genomic organization in Hi-C matrices [7].

In 2018 Zhang et al., developed HiCPlus a foundational and novel approach to Hi-C super-resolution using CNNs [10]. To train this model, 10 kb resolution Hi-C data was first acquired. A corresponding low-resolution Hi-C matrix was generated by randomly down sampling the original 10 kb Hi-C sequencing data by a factor of 1/16. During model training, the CNN effectively learns to map features in the low-resolution Hi-C matrix to those in the original high-resolution data generated using the GM12878 cell line [20]. To conduct enhancement of low-resolution data, the model splits the matrix into smaller windows and enhances each portion separately. The model then recombines the windows into a singular high-resolution matrix. This framework was evaluated using Pearson and Spearman correlation values, which indicated that the model's predicted Hi-C matrix from low-resolution data closely resembled the actual high-resolution Hi-C matrix. Once the model was trained on a specific cell line, it was also able to predict other cell types. Furthermore, the author's experimentation with CNN hyperparameters revealed that a three-layer

CNN with a Rectified Linear Unit (ReLU) activation function provides the best results for capturing interaction frequencies.

Building upon HiCPlus, Liu and Wang proposed HiCNN, which utilized a deeper CNN to enhance accuracy [11]. While HiCPlus utilized a shallow network with just three layers [10], HiCNN used 54 layers with global and local residual learning. These two residual learning approaches allow HiCNN to produce better results with regard to Hi-C matrix resolution. This model utilizes the same training format as HiCPlus, except for the fact that the original data is down sampled using a factor of 1/14. The authors also utilized methods to speed up the training process, allowing for HiCNN to finish training in 200 epochs. In comparison, HiCPlus takes 3000+ epochs to complete the entire training process. Furthermore, the model also outperforms HiCPlus in terms of the Pearson correlation value and mean squared values, when comparing the predicted and original high-resolution Hi-C matrices.

Li and Dai developed SRHiC [12] to further optimized the CNN Hi-C super-resolution architecture to reduce the computational cost of predecessor models. The architecture of this model consists of multiple convolutional layers (fewer than 54), which effectively predicts a high-resolution Hi-C matrix from a low-resolution input matrix. Without the speedup techniques utilized by the authors of HiCNN, both HiCPlus and SRHiC have shorter training times. In terms of prediction performance, SRHiC outperforms both HiCPlus and HiCNN on the same testing metrics. The Pearson correlation coefficient between SRHiC's predicted high-resolution matrix and the original high-resolution data was higher than that of the predecessor models, HiCPlus and HiCNN. SRHiC also computed the TAD overlap between the predicted matrix and the original high-resolution Hi-C data. For this metric, SRHiC also did better than HiCPlus and HiCNN.

3.2. Generative Adversarial Networks

GANs have emerged as another alternative to CNNs for Hi-C super-resolution. GANs consist of two core parts: a discriminator and a generator. The generator creates fake data, while the discriminator acts as a classifier, distinguishing the falsified data from real ones [9]. During the training process, the generator and discriminator improve in their respective roles. The overarching goal of this procedure is to successfully fool the discriminator such that it cannot tell apart the generated data from real images. Once training is complete, the generator can be used to generate high-resolution data from low-resolution inputs. Often, GANs utilize CNN's in their generator and discriminator partitions to improve performance.

In 2019, following the advent of HiCPlus [10], Liu et al. developed hicGAN [13] to overcome the shortcomings of previous CNN-based works (i.e., overly smooth predicted matrices, a lack of fine-grain details, and limited window sizes).

As the name suggests, hicGAN utilizes the GAN architecture to predict Hi-C data. Similarly to HiCPlus, low-resolution data are created by downsampling real high-resolution Hi-C sequencing reads by a factor of 1/16. The training process for this model begins by feeding a low-resolution Hi-C sample into the generator, which outputs a predicted high-resolution Hi-C matrix. The original high-resolution Hi-C matrix is then fed into the discriminator alongside the high-resolution Hi-C matrix generated by the generator. This framework allows both the discriminator and generator to improve, such that hicGAN can generate high-resolution Hi-C matrices with enhanced accuracy. Unlike CNN models, such as HiCPlus, hicGAN does not minimize mean squared error, which helps the model avoid too much smoothing in the predicted matrices. HicGAN also performs comparably with HiCPlus in regards to Pearson and Spearman correlation coefficients. This model can effectively predict high-resolution Hi-C matrices given low-resolution inputs even if the data is from a cell line on which hicGAN was not trained.

Following hicGAN, numerous GAN-based approaches to Hi-C super-resolution emerged. HiCSR [14] by Dimmick et al., where the authors employed a GAN alongside novel loss functions that had not previously been used in the Hi-C resolution enhancement domain. This framework utilizes a GAN to generate high-resolution data from low-resolution Hi-C matrices while also incorporating a Denoising Autoencoder (DAE) as a loss function to enhance the performance of both the generator and the discriminator. Much like previous models, HiCSR works with a high-resolution Hi-C matrix that is artificially downsampled to a low-resolution Hi-C matrix. The downsampling is performed by removing 93.7% of the original reads randomly. During training, the model optimizes three distinct loss functions:

1. *Adversarial loss* – Used and defined by the generator to create an image that can fool the discriminator.
2. *Feature loss* – Computed using a Denoising Autoencoder by comparing the generated high-resolution Hi-C matrix with the real high-resolution data.
3. *Pixel loss* – Derived by directly comparing the pixel values between the generated and real high-resolution matrices.

Compared to previous models, HiCSR is more effective at reconstructing unique 3D structures such as TADs that are present in real high-resolution Hi-C data. Furthermore, in terms of Pearson correlation and Spearman rank values, HiCSR performs comparably with HiC-GAN, HiCPlus, and HiCNN. However, HiCSR outperforms all other models when it is evaluated using Hi-C specific comparison metrics: GenomeDISCO [17] , HiC-Spector [18], and HiCRep [19]. This is highlighted in Figure 2 below.

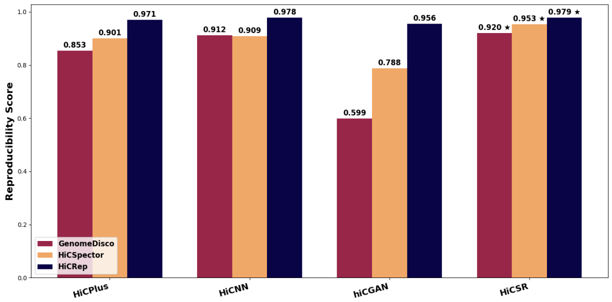

Figure 2. Grouped bar chart comparing the performance of HiCPlus, HiCNN, HiCGAN, and HiCSR across three Hi-C focused metrics for the GM12878 cell line (chromosomes 19-22). This diagram was generated using averaged metrics from HiCSR's paper [14].

Inspired by HiCSR, there also exists VEHiCLE [15], which utilizes a GAN with a Variational Autoencoder (VAE) instead of a DAE to predict high-resolution Hi-C matrices from low-resolution inputs. In VEHiCLE, the architecture uses 4 different loss functions to improve the enhancement of Hi-C data.

1. *Adversarial loss* – Used and defined by the generator (similar to HiCSR).
2. *Variational loss* – Computed by using a Variational Autoencoder.
3. *Insulation Score Loss* – Defined by a computing a delta vector to detect TADs.
4. *Mean Squared Error Loss* - Computed by squaring the differences between bins in real and generated matrices.

VEHiCLE measures well against previous models, such as HiCSR and HiCPlus, in terms of common metrics such as Pearson and Spearman coefficients. This can be seen in Figure 3 below.

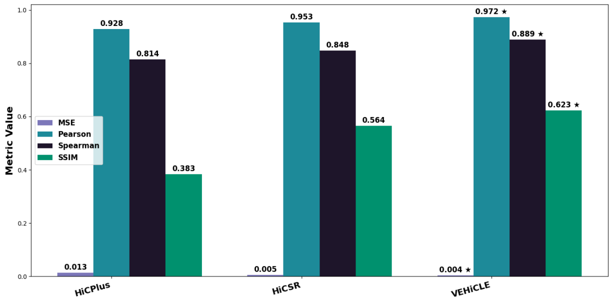

Figure 3. Grouped bar chart comparing the performance of HiCPlus, HiCSR, and VEHiCLE across three common metrics for the GM12878 line (chromosomes 4, 14, 16, and 20). This diagram was generated using averaged metrics from VEHiCLE's paper [15].

3.3. Autoencoders

Autoencoders are another effective model that is utilized for enhancing data resolution [21]. Autoencoders compress (encode) the inputted data into a smaller latent space and then decoding it to reconstruct the original data. Two well-known expansions of Autoencoders are Variational Autoencoders (VAEs) and Denoising Autoencoders (DAEs). In VAEs, the encoder utilizes a Gaussian distribution to introduce variability into the latent space. A fully trained VAE can generate multiple distinct outputs from a single input [15, 21]. On the other hand, DAEs work by corrupting the input and then attempting to reconstruct the original data from that corrupted input. DAEs can be useful for learning important features of the input data.

As mentioned previously, HiCSR utilizes a DAE to compute feature loss [14]. To train the DAE, Gaussian-distributed noise is added to the high-resolution Hi-C data. This noisy data is then input into the DAE, which learns to denoise and regenerate the original high-resolution Hi-C data. In doing so, the DAE captures the features of the high-resolution data, which helps determine the feature loss for GAN training.

VEHiCLE [15] utilizes a VAE to compute variational loss. To train the VAE, the high-resolution Hi-C matrix is fed into the encoder, which encodes the data into a latent space. Then, the decoder decodes it to recreate the original matrix. The variational loss is computed through the difference between the original and the reconstructed matrix. Furthermore, since VAEs introduce variability into the latent space, VEHiCLE can be utilized to generate synthetic high-resolution Hi-C data.

Looking at these multimodal architectures that utilize both GANs and Autoencoders, a significant question arises: How effective would using only Autoencoders be for the Hi-C super-resolution problem, and could this approach reduce computational costs?

4. Future Work

A common theme in Hi-C super-resolution models is the down sampling of existing high-resolution Hi-C data to create low-resolution samples for training. Although random down sampling of original Hi-C sequence reads to reduce sequencing depth is effective, we aim to explore down simulation, as well. Down simulation of Hi-C matrices refers to simulating data at a lower sequencing depth, while preserving important biological features.

A key tool that enables effective down simulation, and simulation in general, is FreeHiC [22]. This tool operates by taking an existing Hi-C dataset and algorithmically extracting real patterns from it. By doing so, FreeHiC does not assume a random structure for the data it generates; instead, it produces simulated Hi-C data that is closely connected with the original input.

The input for FreeHiC is real DNA from Hi-C experiments in FASTQ format. FASTQ data contains sequencing reads, which FreeHiC processes to generate synthetic read pairs. These simulated read pairs closely mimic real-world DNA interactions, including patterns. Down simulation occurs when we modify the parameters of FreeHiC to generate fewer sequencing reads. By doing so, we effectively reduce the sequencing depth. As such, the following questions arise for future research:

Can down simulation replace down sampling in current DL architectures to improve training efficacy? By using down simulation, can we also learn how down sampling affects these DL frameworks for Hi-C super resolution?

Future work in this domain should also explore the efficacy of autoencoders alone and other novel deep learning techniques such as diffusion models and transformers.

5. Conclusion

In this review, our objective was to generate a comprehensive understanding of deep learning techniques in the Hi-C super-resolution domain. Specifically, we examined how deep learning methods have become effective tools for enhancing Hi-C data. We explored three main DL paradigms: CNNs, GANs, and autoencoders. In addition, we discussed various benchmarking tools for Hi-C data while also providing directions for future research.

References

[1] Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, Amit I, Lajoie BR, Sabo PJ, Dorschner MO, Sandstrom R, Bernstein B, Bender MA, Groudine M, Gnirke A, Stamatoyannopoulos J, Mirny LA, Lander ES, Dekker J. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science Oct. 2009;326(5950):289–293.

[2] Rowley M, Corces V. Organizational principles of 3d genome architecture. Nature Reviews Genetics 2018; 19:789–800.

[3] Lajoie BR, Dekker J, Kaplan N. The hitchhiker's guide to hi-c analysis: Practical guidelines. Methods 2015;72:65–75.

[4] Durand NC, Robinson JT, Shamim MS, Machol I, Mesirov JP, Lander ES, Aiden EL. Juicebox provides a visualization system for hi-c contact maps with unlimited zoom. Cell Systems 2016;3(1):99–101.

[5] Lupianez DG, Kraft K, Heinrich V, Krawitz P, Brancati F, Klopocki E, Horn D, Kayserili H, Opitz JM, Laxova R, Santos-Simarro F, Gilbert-Dussardier B, Wittler L, Borschiwer M, Haas SA, Osterwalder M, Franke M, Timmermann B, Hecht J, Spielmann M, Visel A, Mundlos S. Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. Cell 2015;161(5):1012–1025.

[6] G-CORE USD. Hi-c data portal.

[7] Alzubaidi L, Zhang J, Humaidi A, et al. Review of deep learning: concepts, cnn architectures, challenges, applications, future directions. Journal of Big Data 2021;8:53.

[8] Dong C, Loy CC, He K, Tang X. Image super-resolution using deep convolutional networks. IEEE Transactions on Pattern Analysis and Machine Intelligence 2016;38(2):295–307.

[9] Aggarwal A, Mittal M, Battineni G. Generative adversarial network: An overview of theory and applications. International Journal of Information Management Data Insights 2021;1(1):100004.

[10] Zhang Y, An L, Xu J, et al. Enhancing hi-c data resolution with deep convolutional neural network hicplus. Nature Communications 2018;9:750.

[11] Liu T, Wang Z. Hicnn: a very deep convolutional neural network to better enhance the resolution of hi-c data. Bioinformatics 2019;35(21):4222–4228.

[12] Li Z, Dai Z. Srhic: A deep learning model to enhance the resolution of hi-c data. Frontiers in Genetics 2020;11:353.

[13] Liu Q, Lv H, Jiang R. hicgan infers super resolution hi-c data with generative adversarial networks. Bioinformatics 2019;35(14):i99–i107.

[14] Dimmick MC, Lee LJ, Frey BJ. Hicsr: a hi-c super-resolution framework for producing highly realistic contact maps. bioRxiv 2020;.

[15] Highsmith M, Cheng J. Vehicle: a variationally encoded hi-c loss enhancement algorithm for improving and generating hi-c data. Scientific Reports 2021;11(1):8880.

[16] Wang Z, Bovik A, Sheikh H, Simoncelli E. Image quality assessment: from error visibility to structural similarity. IEEE Transactions on Image Processing 2004;13(4):600–612.

[17] Ursu O, Boley N, Taranova M, Wang YXR, Yardimci GG, Noble WS, Kundaje A. Genomedisco: a concordance score for chromosome conformation capture experiments using random walks on contact map graphs. Bioinformatics 2018; 34(16):2701–2707.

[18] Yan KK, Yardimci GG, Yan C, Noble WS, Gerstein M. Hic-spector: a matrix library for spectral and reproducibility analysis of hi-c contact maps. Bioinformatics 2017; 33(14):2199–2201.

[19] Yang T, Zhang F, Yardımcı GG, Song F, Hardison RC, Noble WS, Yue F, Li Q. Hicrep: assessing the reproducibility of hi-c data using a stratum-adjusted correlation coefficient. Genome Research 2017;27(11):1939–1949.

[20] Rao SSP, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, Sanborn AL, Machol I, Omer AD, Lander ES, Aiden EL. A 3d map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell Dec. 2014;159(7):1665–1680.

[21] Bennouna K, Chougrad H, Khamlichi Y, Boushaki AE, Ali SEHB. Variational autoencoders versus denoising autoencoders for recommendations. In WITS 2020. Springer Singapore, 2022; 179–188.

[22] Zheng Y, Keleş S. Freehi-c simulates high-fidelity hi-c data for benchmarking and data augmentation. Nature Methods 2020;17(1):37–40.