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

Computer Science Conference for CSU Undergraduates

Meghana Indukuri
Dr. Carlos Rojas

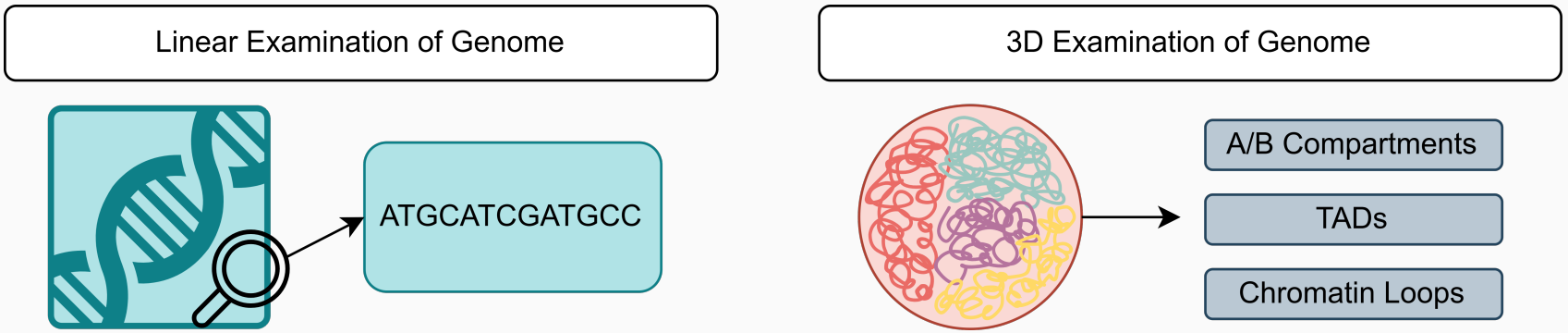
San Jose State University

Table of contents

Background

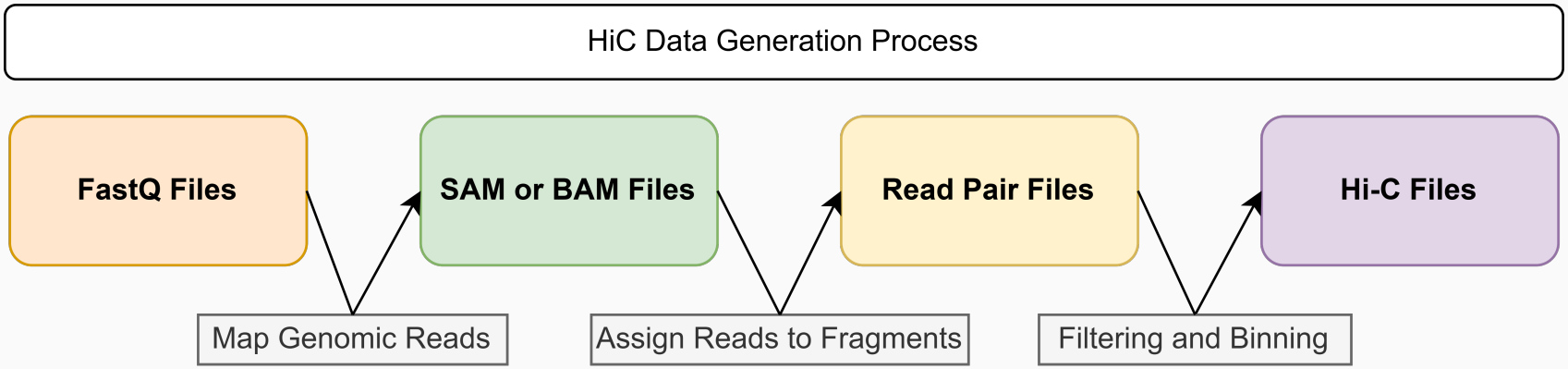
3D Examination of the Genome

The **genome** is often viewed through the linear lens; however, its 3D organization actually affects gene regulation and function.



3D Genomic Data: Hi-C

High-throughput Chromosome Capture **(Hi-C)** captures interactions frequencies between genomic regions and is generated through the following workflow [8]:



Hi-C Matrix

The output of the **Hi-C** workflow is a $N \times N$ matrix, where each cell displays the interaction frequencies between genomic bins.

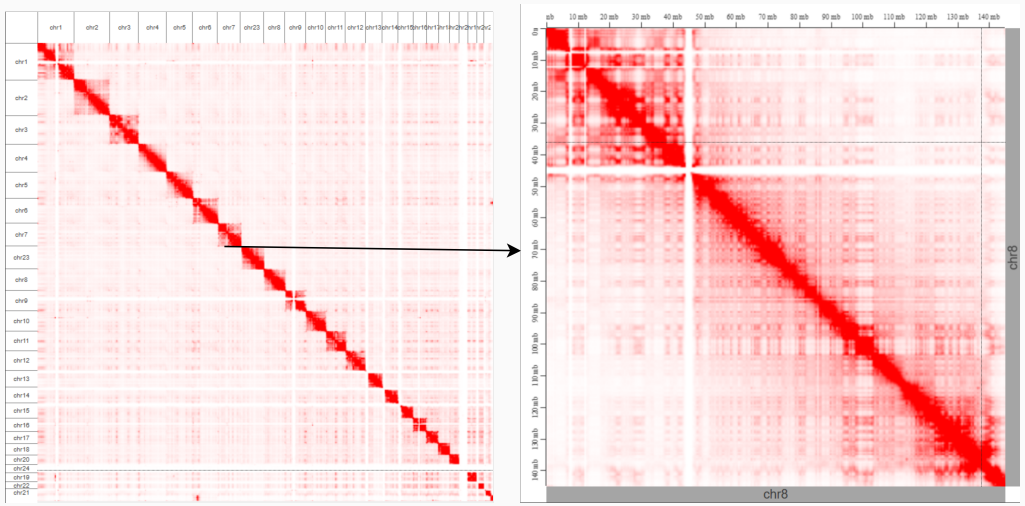
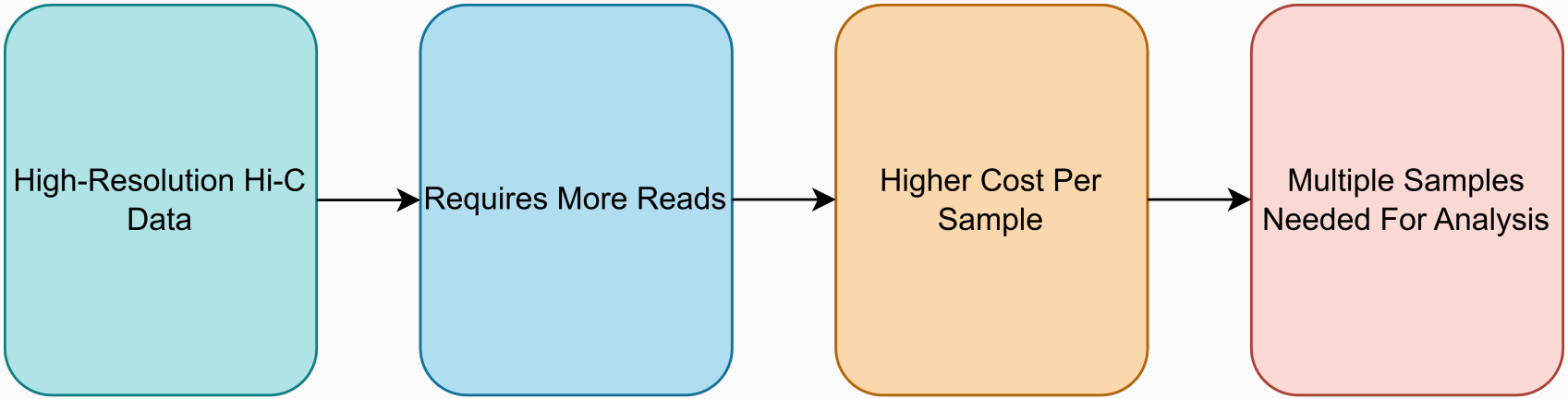


Figure 1: Hi-C Matrix of GM12878 cell line, generated using JuiceBox [5]

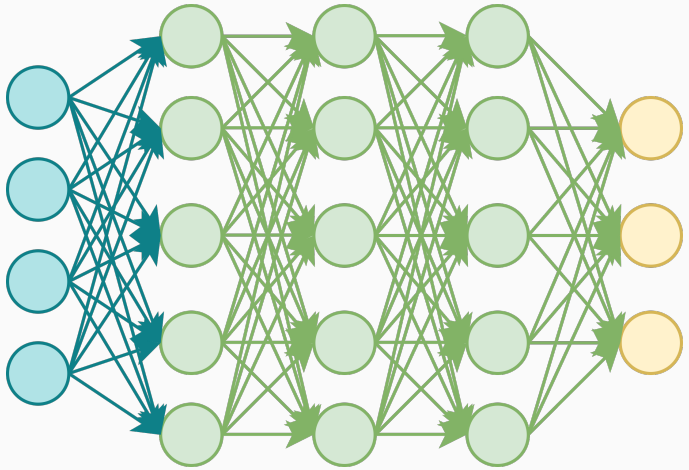
Hi-C Costs

Generating higher-resolution Hi-C matrices (e.g., Figure 1) is costly.



Deep Learning

Deep learning architectures such as CNNs, GANs, and autoencoders offer a cost-effective route for converting low-resolution data into high-resolution representations [2, 4, 1].



Hi-C super-resolution adopts the same principle, modeling each contact map as an image in which every interaction count serves as a pixel [16, 10, 7, 9, 3, 6].

Benchmarking Hi-C Predictions

Naive metrics

- **MSE**: element-wise squared error, averaged over all $N \times N$ entries.
- **Pearson** r: linear correlation.
- **Spearman** ρ: rank correlation.
- **SSIM**: structural similarity index. Originally developed for images [13].

Biologically informed metrics

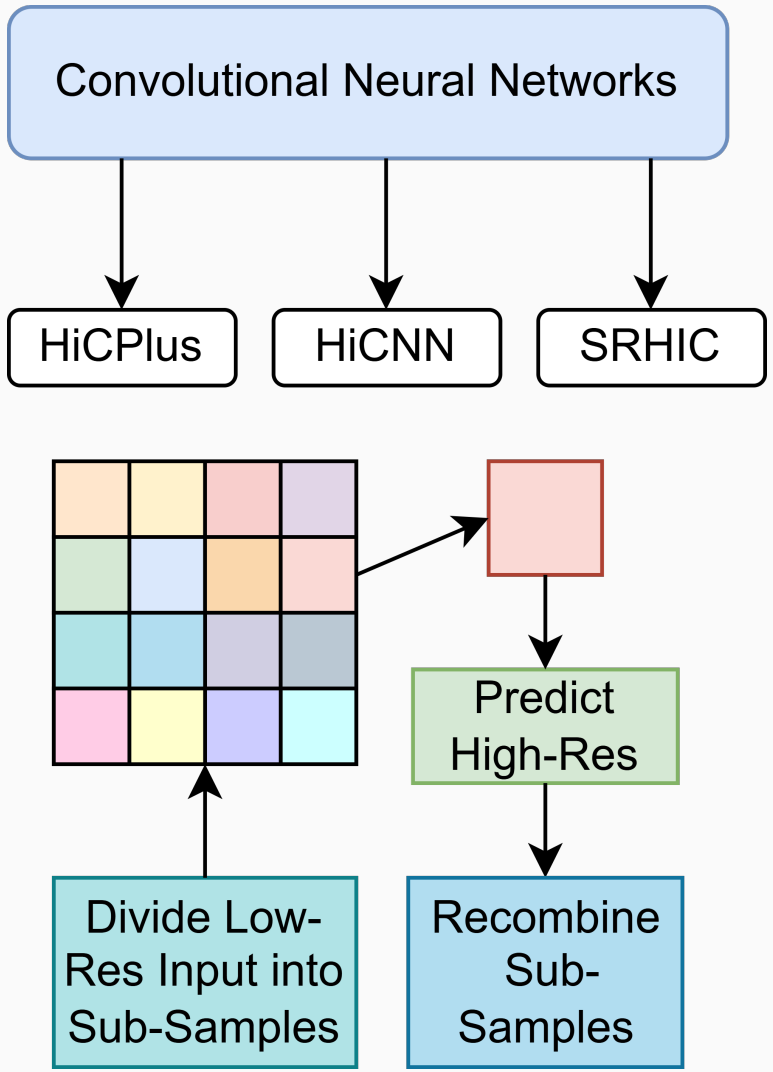
- **GenomeDisco**: graph-based [12].
- **HiCSpector**: eigenvalues and eigenvectors [14].
- **HiCRep**: uses Stratum-Adjusted Correlation Coefficient (SCC) [15].
- All three output a reproducibility score. Closer to $1 \Rightarrow$ strong similarity.

Deep Learning Architecture for Hi-C Super Resolution

Convolutional Neural Networks Overview

CNNs stack convolutional layers to map low-resolution features to their high-resolution counterparts [4, 2].

Their sensitivity to spatial locality makes them well suited for enhancing Hi-C data [2].



CNNs: HiCPlus

- **HiCPlus** (Zhang *et al.*, 2018): first CNN framework [16]
- Training set: 10 kb GM12878 data [11]; low-res version created by 1/16 down-sampling
- Performance assessed with Pearson and Spearman correlations
- Trained on one cell type, applicable to others

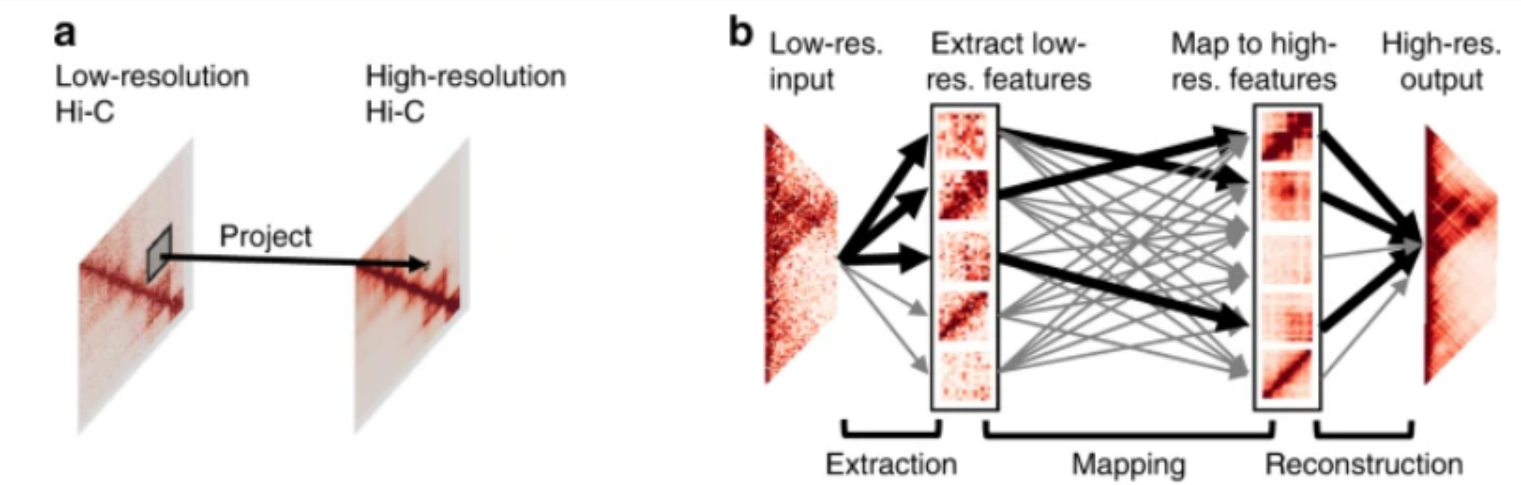


Figure 2: HiCPlus Diagram [16].

CNNs: HiCNN and SrHiC

- **HiCNN** (Liu & Wang, 2019): 54-layer CNN with residual paths [10]
- Faster training: ∼200 epochs vs 3500 for HiCPlus.
- Outperforms HiCPlus with higher Pearson *r* and lower MSE

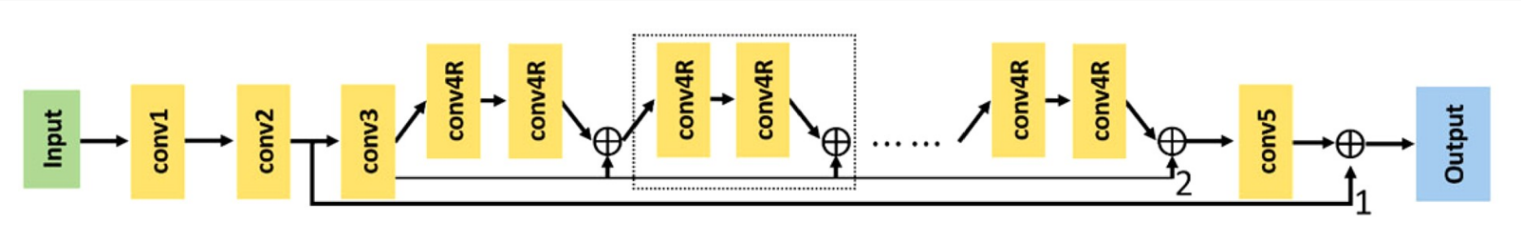
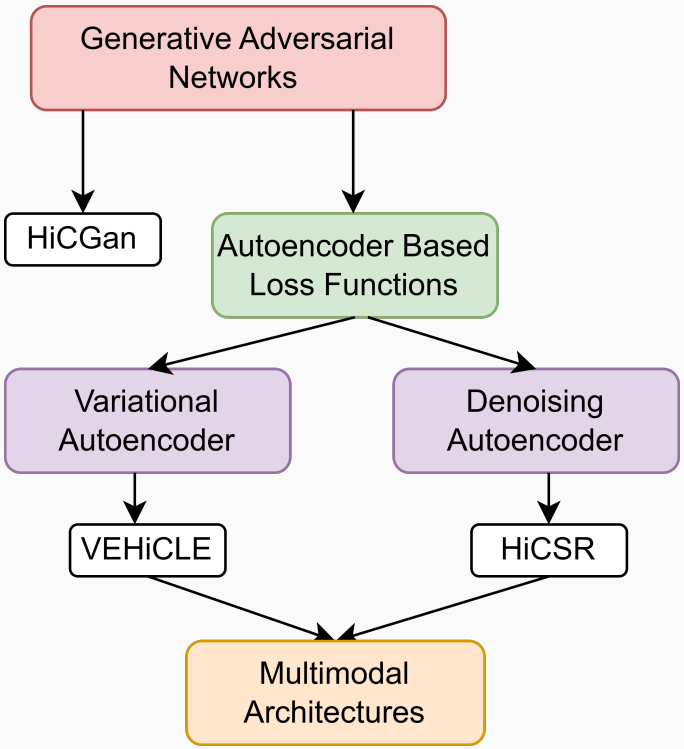


Figure 3: HiCNN Architecture [10].

- **SRHiC** (Li & Dai, 2020): lighter CNN architecture [7]
- Trains faster than both
- Higher Pearson correlation
- Shows greatest TAD reconstruction

General Adversarial Networks



- **hicGAN** (Liu *et al.*, 2019): first GAN framework [9]
- Training data: same as HiCPlus
- Uses adversarial loss rather than MSE
- Pearson and Spearman scores on par with HiCPlus

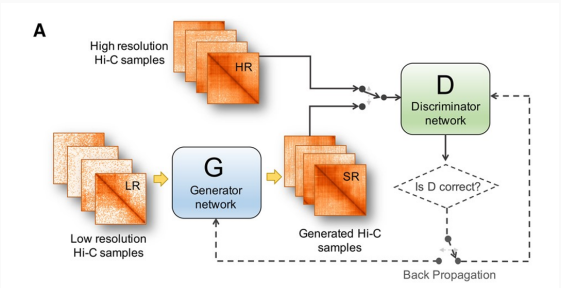
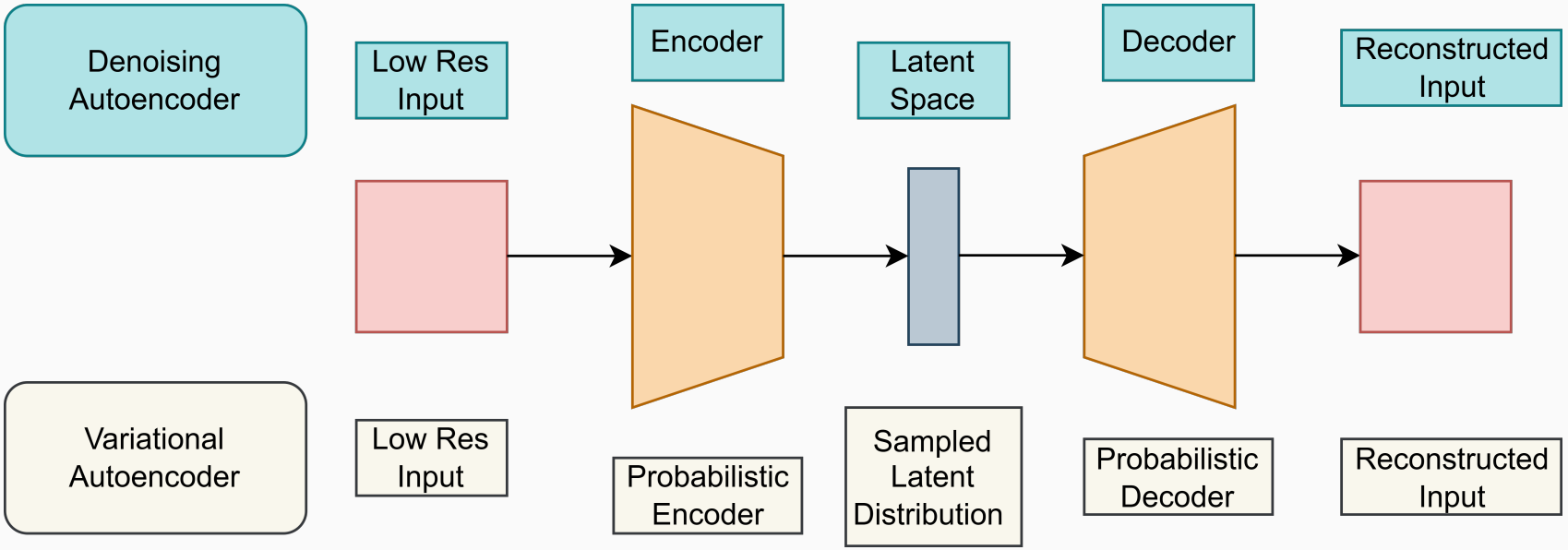


Figure 4: hicGAN Architecture [9].

Autoencoders Overview

AEs compress (encode) input data into a lower-dimensional latent space, which is then decoded to reconstruct or enhance the original data.



Multi-modal Architectures: HiCSR and VEHiCLE

HiCSR

- **HiCSR** (Dimmick *et al.*, 2020): GAN framework with a DAE for loss [3].
- Loss functions
  - Adversarial loss (generator vs discriminator)
  - Feature loss from DAE
  - Pixel loss (element-wise error)
- Tops HiC-specific metrics: GenomeDisco, HiC-Spector, HiCRep

VEHiCLE

- **VEHiCLE** (Highsmith *et al.*, 2021): combines a GAN with a VAE [6]
- Loss Functions
  - Adversarial loss
  - Variational loss (VAE loss)
  - Insulation score loss
  - Mean-squared error loss
- Outperforms HiCSR and HiCPlus on Pearson and Spearman correlation benchmarks

Multi-modal Architectures

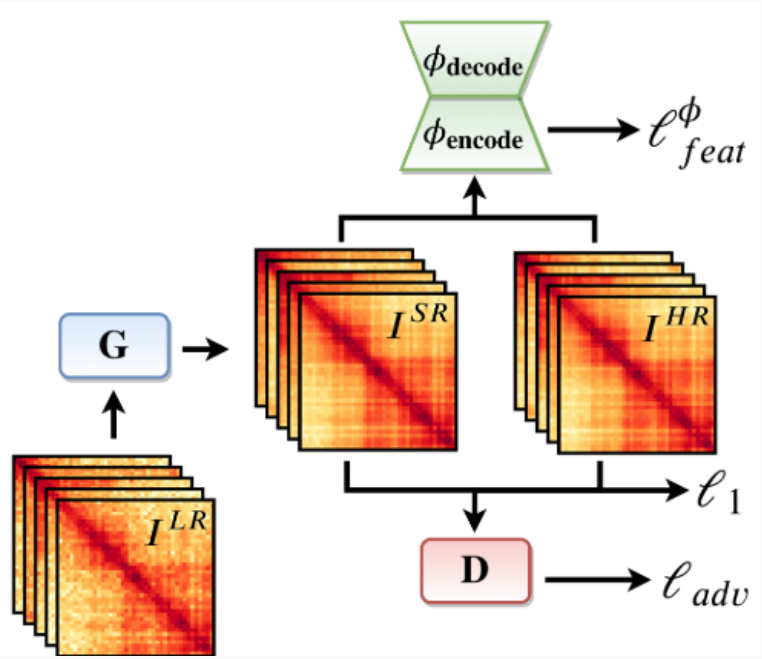


Figure 5: HiCSR Architecture [3].

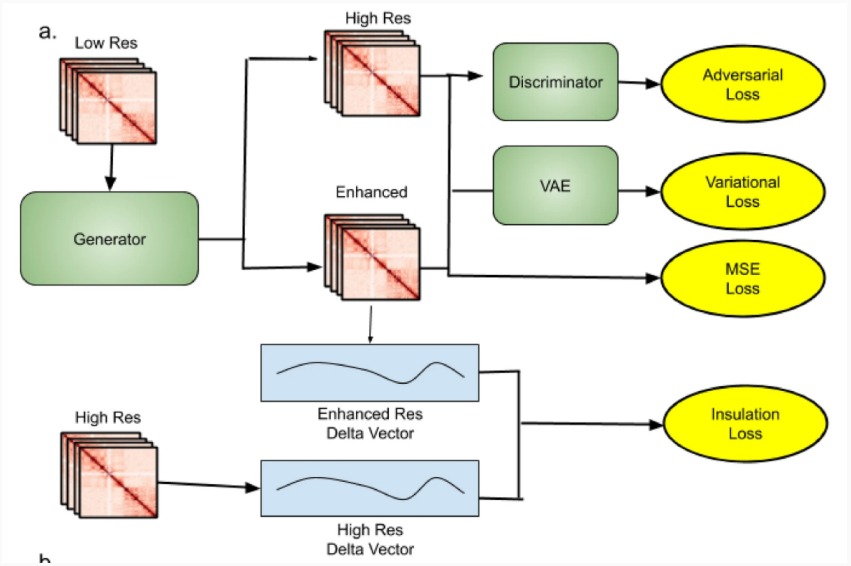


Figure 6: VeHiCLE Architecture [6].

Comparison of Architectures

Naive Benchmarking

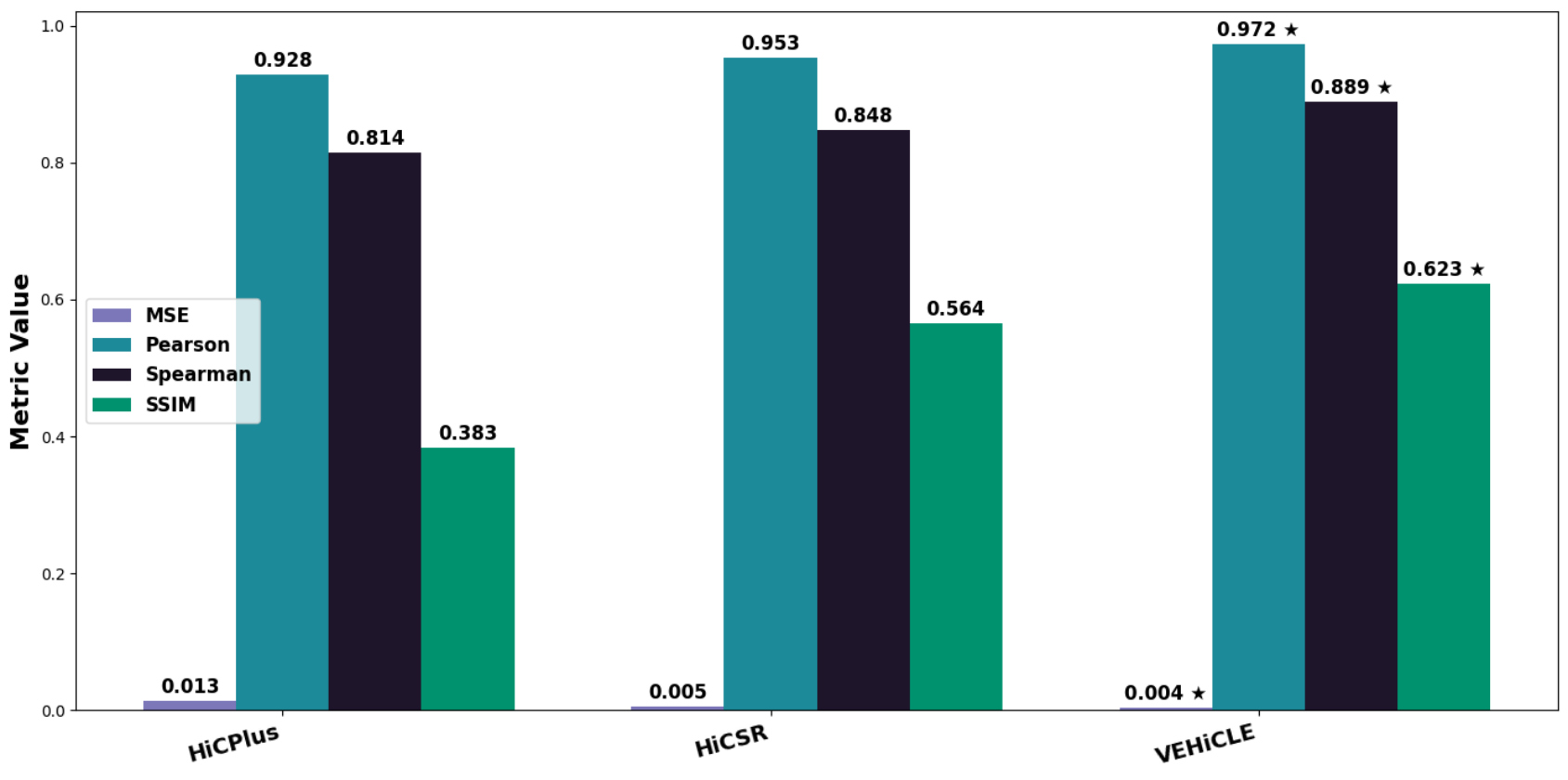


Figure 7: Grouped bar chart comparing the performance of 3 models for the GM12878 line (chromosomes 4, 14, 16, and 20). Averaged metrics are from VEHiCLE [6].

HiC-Focused Benchmarking

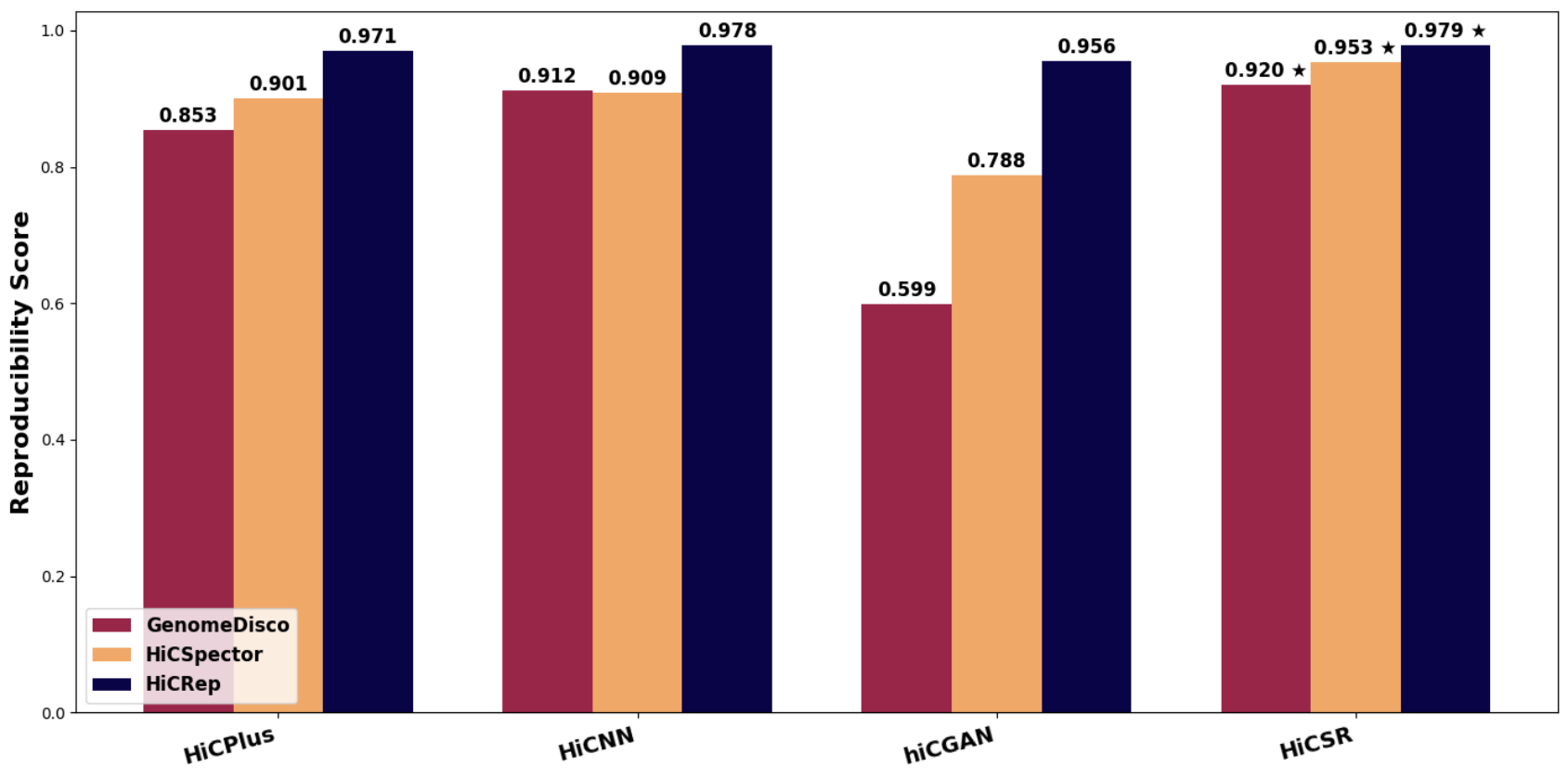


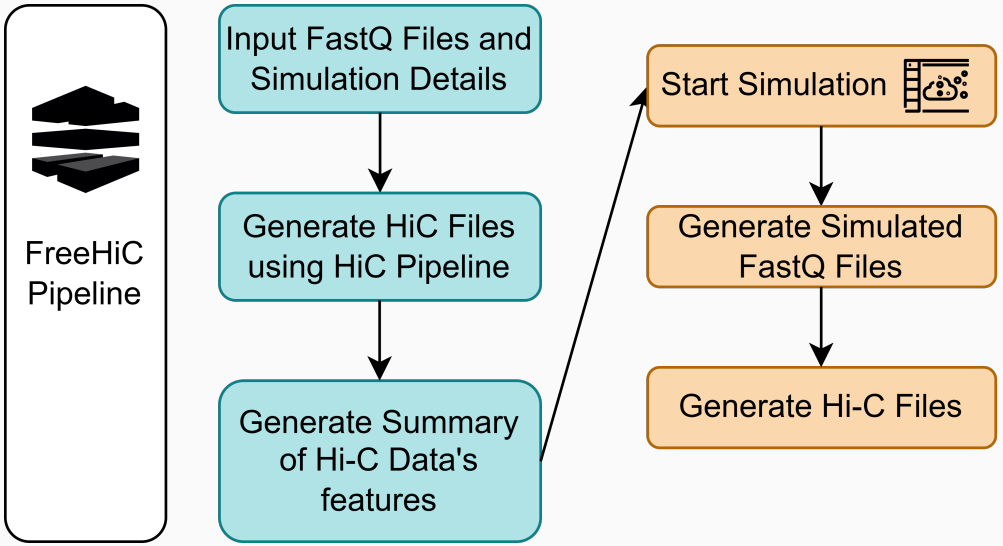
Figure 8: Grouped bar chart comparing the performance of 4 models for the GM12878 cell line (chromosomes 19-22). Averaged metrics are from HiCSR [3].

Future Work

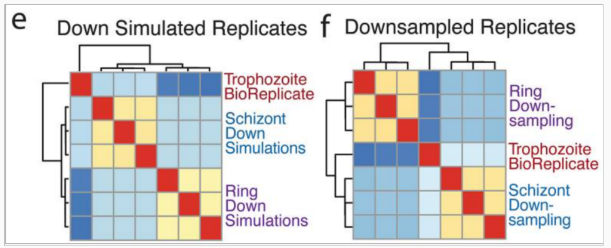
Down Simulation

Downsimulation of Hi-C matrices refers to simulating data at a lower sequencing depth, while preserving important biological features.

FreeHiC is a tool that enables effective simulation of Hi-C data [17].



Downsimulated vs Downsampled Hi-C Data



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

Other Models

- How effective are **autoencoders** when used *alone* for Hi-C resolution enhancement?
- Can **diffusion and transformer models** outperform current approaches in this domain?

Questions?

References i

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

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

M. C. Dimmick, L. J. Lee, and B. J. Frey.
Hicsr: a hi-c super-resolution framework for producing highly realistic contact maps.
bioRxiv, 2020.

References ii

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

N. C. Durand, J. T. Robinson, M. S. Shamim, I. Machol, J. P. Mesirov, E. S. Lander, and E. L. Aiden.
Juicebox provides a visualization system for hi-c contact maps with unlimited zoom.
Cell Systems, 3(1):99–101, 2016.

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

References iii

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

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

References iv

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

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

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

References v


O. Ursu, N. Boley, M. Taranova, Y. X. R. Wang, G. G. Yardimci, W. S. Noble, and A. Kundaje.
Genomedisco: a concordance score for chromosome conformation capture experiments using random walks on contact map graphs.
Bioinformatics, 34(16):2701–2707, 2018.


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


K.-K. Yan, G. G. Yardimci, C. Yan, W. S. Noble, and M. Gerstein.
Hic-spector: a matrix library for spectral and reproducibility analysis of hi-c contact maps.
Bioinformatics, 33(14):2199–2201, 2017.

References vi

T. Yang, F. Zhang, G. G. Yardımcı, F. Song, R. C. Hardison, W. S. Noble, F. Yue, and Q. Li.
Hicrep: assessing the reproducibility of hi-c data using a stratum-adjusted correlation coefficient.
Genome Research, 27(11):1939–1949, 2017.

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

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