# STRUCTURAL MODELING OF THE 3D GENOME USING MACHINE LEARNING

---

A Dissertation

presented to the

Faculty of the Graduate School

at the University of Missouri-Columbia

---

In Partial Fulfillment of the Requirements for the Degree
Doctor of Philosophy

---

by

Max Highsmith

Professor Jianlin Cheng, Dissertation Supervisor

December 2020

The undersigned, appointed by the Dean of the Graduate School, have examined the dissertation entitled

**STRUCTURAL MODELING OF THE 3D GENOME USING MACHINE LEARNING**

presented by Max Highsmith, a candidate for the degree of Doctor of Philosophy and hereby certifying that, in their opinion, it is worthy of acceptance.

---

Professor Jianlin Cheng

---

Professor Toni Kazic

---

Professor Yi Shang

---

Professor Rocio Rivera

---

Professor Chi-Ren Shyu

# Dedication

I dedicate this dissertation to my parents and my four brothers.

# Acknowledgements

I would like to first thank my advisor Dr. Jianlin Cheng for his constant guidance and support through my graduate school journey. I would also like to thank Dr. Yi Shang, Dr. Toni Kazic, Dr. Chi-Ren Shyu and Dr. Rocio Rivera for their roles as my graduate committee. I would like to thank Dr. Oluwatosin Oluwadare for his guidance as a senior lab member in the BDM lab and Frimpong Bodau for his interest in continuing the BDM labs progress in this research direction upon my graduation.

# Table of Contents

# List of Tables

# List of Figures

**Figure 2.3. VEHiCLE Latent Generative Model** (a) Diagram of synthetic Hi-C generation tool, a user tunable zero-centered feature vector is transformed via PCA reverse transform to latent space and then passed through our tuned decoder network. (b) The 0 vector corresponds to a purely linear contact map. (c) Increasing value of PC5 results in generation of TADs. (d) adjusting the value of PC2 shifts the position of TADs. (e) Adjusting PC11 creates stripes within TADS. (f) adjusting PC14 develops loops within TADs.

**Figure 2.4 Visual Comparison of VEHiCLE Contact Matrix Enhancement.** (a) Visual comparison of enhancement matrices. (b) Absolute difference matrices between target high resolution data and enhancement. All displayed matrices are derived from the GM12878 cell line. Architectures of previous models utilize original window size.

**Figure 2.5 VEHiCLE Enhanced 3DModel Generation**. (a) 3D reconstruction of Chro 20 0.6MB-3.1MB. (b) TM -score comparison of High Resolution structures to (red) Low resolution structures and (green) VEHiCLE enhanced structures. VEHiCLE enhanced scores are significantly higher (wilcoxon rank sum p value < 1e-20) (c) Average TM-Score comparison of ingroup structures generated by same contact matrix (red) low res, (yellow) high res, (green) VEHiCLE enhanced. VEHiCLE enhanced scores are significantly better than low-resolution scores (wilcoxon rank sum p value < 1e-20) Structures are all generated from GM12878 cell line using the test chromosome set: 4, 14,16,20.

**Figure 3.1 TAPIOCA Dataset Overview.** (a) Generation process of transitional gamma labels. (b) Hi-C Contact matrix (c) Generation process of Directionality Index labels. (d) Generation process of Insulation score labels. (e) vectors used as labels in the training process. (f) Distribution of numerical values for each metric.

**Figure 3.2 TAPIOCA Architecture.** (a) Overview of architecture (b) Details of multi-head attention block.

**Figure 4.2 Simulation of 4DMax Structures** Diagram of Outputs. (a) Outline of the different stages of the iPSC dataset. (b) Contact map of chromosome 13 by time, AB compartment vector shown above map. (c) 4DMax prediction of structural conformation of chromosome 13 at time. (d) Reconstructed contact map using simulated Hi-C of 4DMax structure, number below indicate spearman correlation between above reconstructed contact map and real contact maps at each time point.

**Figure 4.3: 4DMax Predictions of Hi-C Contact Maps** Example contact map comparison. (a) Contact maps of iPSC on day 2 chromosome 2 from Real Hi-C, 4DMax reconstruction and 4DMax day 2 agnostic interpolation model. (b) SPC of iPSC contact maps relative to Real Hi-C for each chromosome on day 2. (c) Contact maps of cardiomyocyte data on day 2 chromosome 7 from Real Hi-C, 4DMax reconstruction and 4DMax day 2 agnostic interpolation model. (d) SPC of cardiomyocyte contact maps to Real Hi-C for each chromosome on day 2.

**Figure 4.4: 4DMax AB Compartment Analysis.** Analysis of AB compartment features of 4DMax generated contact maps. (a) Pearson correlation matrices of chromosome 14 day 2 using Real Hi-C and synthetic contact maps obtained from the 4DMax model. (b) AB compartment vectors from chromosome 14 (red) real Hi-C data (blue) synthetic contact maps obtained from 4DMax model. (c) Trajectory curve of two largest principal components (red) real Hi-C (Blue) Reconstructed Hi-C. (d) Scatter plot of 100kb binned AB compartment vectors where x value is bins Real Data PC1 value and y value is interpolated Contact maps PC1 value.

**Figure 4.5: 4DMax Topologically Associated Domain Analysis** HiCtool identified topologically associated domains. (a) Select images of TAD boundaries on (black) Real Hi-C replicate 1, (blue) Real Hi-C replicate 2, (green) 4DMax Reconstructed Map, (orange) 4DMax Interpolated Hi-C Map and 4DMax Recon Map at a different time point. PO metric quantifies the percent of TAD boundaries found within 0.5Mb of a boundary identified in Hi-C rep1. (b) PO of Interpolated and Reconstructed 4DMax TAD positions for both replicates across all chromosomes.

**Figure 4.6: 4DMax Computational Evaluation** Evaluation of runtimes and computational stability. (a) Scatter plot of chromosome bin lengths and time to completion using 400 epoch (purple) 500kb resolution chromosomes and (blue) 50 kb resolution chromosomes. (b) 3D plot of predicted cardiomyocyte chromosome 10 on day 5 with varying granularity values. Spearman correlation Mean Squared distance compares (blue) granularity 15 structure to higher granularity structures (red). (c) 3D plots comparing (purple) 50 kb resolution chromosome 1 to (green) 500kb resolution iPSC chromosome 1 on each day in time series.

**Figure 5.1 GSDB HomePage** (top) home page of GSDB with data statistics (bottom) selection menu for viewing Hi-C data and 3D structures.

**Figure 5.2 Viewing window in GDSB.** (red) heat map of selected genome. (blue) color and (green) correlation settings for visualization of heatmap. (yellow) identifies 3D structure and options for dataset, resolution, chromosome, and algorithm to use.

**Figure 5.3 SpaceWalk Visualization** (a) A visualization of chromosome 1 within GSDB embedded pymol. Clicking the highlighted “View in Spacewalk” button will redirect the user to (b) the space walk visualization tool, which permits a more fluid and customizable visualization of all GSDB structures.

**Figure 5.4 GSDB PCA and Clustering** (a) Principal component analysis based on similarity of structures generated by different algorithms. (b) Hierarchical clustering based on structural similarity.

# Abstract

This dissertation, submitted as a partial requirement for completion of the Doctorate of Philosophy, outlines the research performed by Max Highsmith in the BDM Lab. This work includes a functional expansion of a three-dimensional genome conformation database, the development of a novel, deep-learning based strategy for the enhancement of Hi-C data, The development of deep learning approach for domain identification using epigenetic features, and the development of a novel computational tool for 4D modeling of chromosome dynamics.

# 1. Introduction

Moore's law is an empirical observation that the number of transistors in a dense integrated circuit has doubled every 2 years since the 1970s (November, 2018). This monumental observation illustrates the rapidness at which the power of computing has evolved over the course of just 50 years. In this time frame the transformative impact which computing has had on human society is difficult to overstate. With this observation in mind, the fact that the dropping price of comprehensive genomic sequencing has outpaced Moore's law since the human genome project's completion in 2003 becomes all the more astonishing(November, 2018). Through the development of next generation sequencing technologies the sheer amount of genomic data in existence has exploded, and is still growing. The potential impact of this data's availability on domains such as agriculture, medicine and biotech has yet to be fully realized. The next major challenge in propagating the metamorphic impact of computational genomics is the development of effective strategies for disseminating, visualizing, and integrating genomic data in a manner which permits effective reconciliation with computational assays and biological reality. In keeping with such philosophy, this dissertation explores each of these strategies with focus on the domain of epigenomic modeling.

## 1.1 Contributions

This dissertation makes the following contributions:

1. This dissertation presents a novel, deep learning based approach for the enhancement of Hi-C Data. The approach integrated domain specific knowledge

with deep learning algorithms protected from computer vision enhancement literature. The whole of this work was done independently with the exception of inception of problem area and manuscript review Dr. Jianlin Cheng

2. This dissertation presents a novel, deep learning based approach for the identification of topologically associated domains using epigenetic profiles of histone modification. The whole of this work was done independently with the exception of review and editing of the manuscript by Dr. Jianlin Cheng.
3. This dissertation presents a novel machine learning based approach for the modeling of 4D chromosome dynamics over a time dependent process. The whole of this work was done independently with the exception of review and editing of the manuscript by Dr. Cheng
4. This dissertation contributes to the development of the genome database, a centralized web browser for visualizing predictive 3D structures of chromosomes. The GSDB was initially created by Dr. Oluwatosin Oluwadare who gathered Hi-C datasets, ran them through 11 structure generating tools, hosted a mysql database and built a web interface. Throughout my PhD Thesis I implemented a new structure generating tool (Shrec3d)(Li et al., 2018), updated all structure formats to extend visualization capabilities, redesigned website features, developed a tutorial for users and added unsupervised analysis for comparison of method results. I also worked with Dr. Douglas Turner to integrate his spacewalk (Oluwadare et al., 2020) tool with the GSDB. This GSDB manuscript was edited and reviewed by Dr. Jianlin Cheng and reviewed by Dr. Erez Lieberman-Aiden.

## 1.2 Background Biological Assays

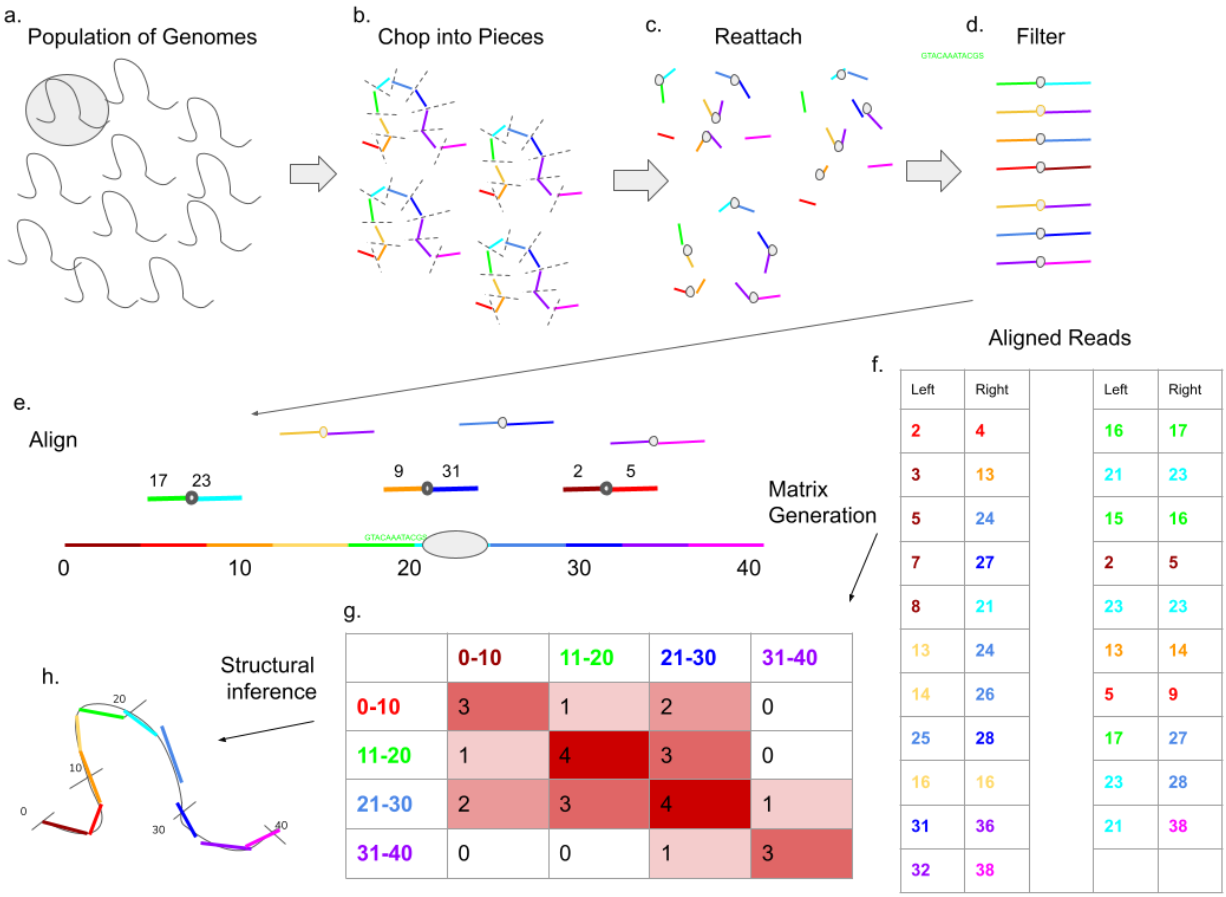
### 1.2.1 Chip-Seq

Chromatin Immunoprecipitation sequencing (ChIP-Seq) is a biological assay used to inspect chromatin interactions with a specified protein(Mifsud, 2018b)(Mifsud, 2018a, 2018b). ChIP-Seq is often applied to histone modifications where peaks are indicative of key epigenetic features such as enhancer and promoter placement. Chip-Seq can also be used to identify the binding sites of selected transcription factors or gene markers. Measurements from Chip-Seq are collections of individual reads which can be aligned to a reference genome to obtain 1 dimensional data. In this dissertation use Chip-Seq data in our TAPIOCA Network (Chapter 4) to assist in prediction of Topologically Associated Domains.

### 1.2.2 Hi-C

The central assay used in this dissertation is Hi-C(Lieberman-Aiden et al., 2009). Hi-C is a biological assay used to inspect chromatin-chromatin interactions. Hi-C improves upon previous chromosome conformation capture assays like 3C (Han et al., 2018) by providing information on the number of contacts between different regions of chromatin on a global scale (Figure 1.1). In Hi-C experiments chromosomes are cross linked with formaldehyde (Figure 1.1a) then cut into fragments with a restriction enzyme (Figures 1.1b). The fragments are ligated to other fragments in a manner which favors spatial proximity (Figure 1.1c). This results in a series of chimeric fragments (Figure 1.1 d) where each side of an individual fragment can be aligned to a separate portion of the

genome (Figure 1.1 e), indicating an interaction between those two loci. This collection of paired reads (Figure 1.1 f) can then be used to construct two-dimensional contact matrices (Figure 1.1 g) by selecting a binding resolution *R* and partitioning the genome into bins of size *R*. Each read pair corresponds to a contact whose row and columns are the bins to which the alignments belong. These contact matrices can be used to inspect various aspects of the chromosome (Figure 1.1h) A potential Hi-C analysis pipeline is displayed in (Figure 1.2)




| Left | Right | Left | Right |
|---|---|---|---|
| 2 | 4 | 16 | 17 |
| 3 | 13 | 21 | 23 |
| 5 | 24 | 15 | 16 |
| 7 | 27 | 2 | 5 |
| 8 | 21 | 23 | 23 |
| 13 | 24 | 13 | 14 |
| 14 | 26 | 5 | 9 |
| 25 | 28 | 17 | 27 |
| 16 | 16 | 23 | 28 |
| 31 | 36 | 21 | 38 |
| 32 | 38 | | |

| | 0-10 | 11-20 | 21-30 | 31-40 |
|---|---|---|---|---|
| 0-10 | 3 | 1 | 2 | 0 |
| 11-20 | 1 | 4 | 3 | 0 |
| 21-30 | 2 | 3 | 4 | 1 |
| 31-40 | 0 | 0 | 1 | 3 |

**Figure 1.1 Hi-C Assay Overview** (a) Population of Cells, spatially proximal regions are cross linked. (b) restriction enzyme cuts the genome into pieces (c) spatially close pieces are ligated. (d) chimeric read pairs are filtered for aberrant ligation junctions. (e) pairs are aligned to

reference the genome to create (f) table of read pairs. (g) Binning resolution is selected to create a contact matrix, from which downstream analysis is performed.

### 1.2.3 Upstream Analysis

Because of its two dimensional nature Hi-C data requires high numbers of reads to provide meaningful information. Consequently the alignment process is often heavily parallelized (Figure 1.2). Experimental bias such as PCR duplicates and random ligations can negatively impact quality of downstream analysis necessitating the inclusion of filtering steps and normalization of matrices, typically through a method called Iterative correction and eigenvector decomposition (ICE)(Imakaev et al., 2012). Even after filtering duplicates and random ligations, experimental bias such as restriction enzyme cutting sites, GC bias and low read coverage can impact the feasible scope of downstream analysis. These challenges as well as our working solutions are explored in Chapter 3 through our Hi-C contact map enhancing algorithm VEHiCLE.

### 1.2.4. Downstream Analysis

Downstream analysis begins after the creation of Hi-C contact matrices (Figure 1.2). To create a matrix requires the selection of a binning size. The selected bin size is constrained by read count and typically ranges from 1Mb-1kb. Lower bin size means higher resolution. Certain features such as A/B compartments can only be viewed from low resolution contact matrices due to issues of computational tractability while other features such as TADs can only be viewed from higher resolution contact matrices because the structural motif itself only spans small regions of the genome and would be entirely encompassed in a single low resolution bin. Furthermore resolution selection is necessarily constrained by readcount (Figure 1.3). In the scope of this dissertation we

observe 3 primary forms of downstream analysis: A/B compartments, Topologically Associated Domains, and 3D structural Models.

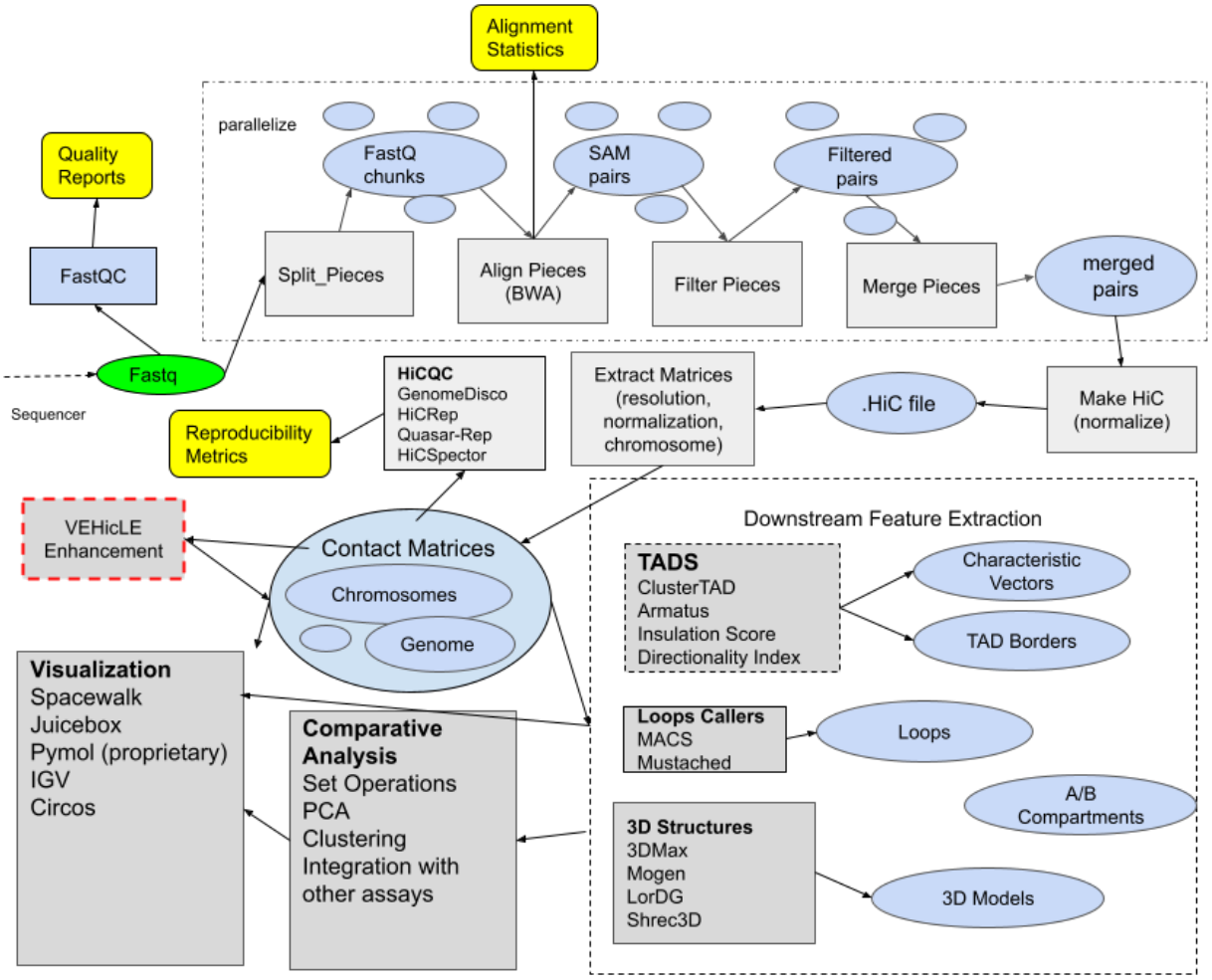



**Figure 1.2 Standard Hi-C Analysis Pipelines.** Pipeline showing possible HiC analysis from upstream parallelized alignment and filtering, downstream feature extraction and visualization.

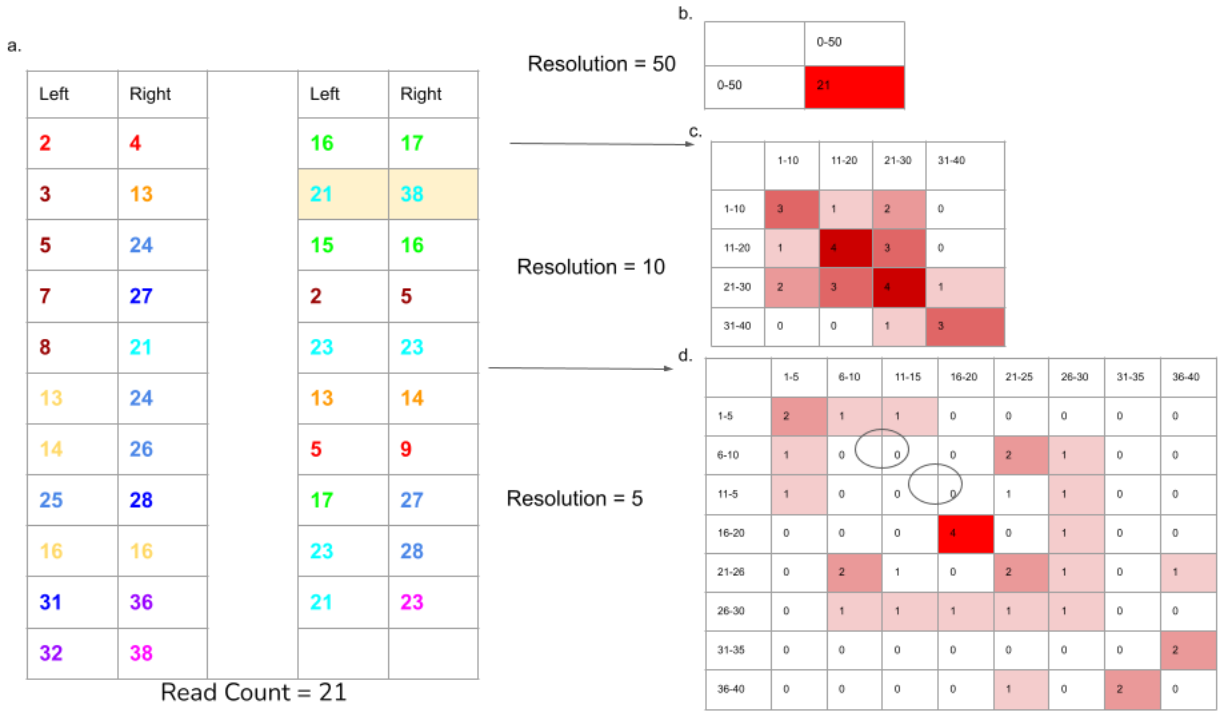



**Figure 1.3 Matrix resolution selection** Schematic showing read count (a) constraints bin resolution. Very low resolution contact matrices (b) skip meaningful features. Higher resolution matrices (c) allow more in depth inspection, but resolution higher than supported by read count (d) leads to gaps in information.

### 1.2.3.1 Hi-C QC Score

Four prolific Hi-C qualifying scores are frequently used in the literature to evaluate similarity between contact maps. These include Hi-CSpector(Yan et al., 2017), QuASAR(Yardımcı et al., 2019), HiCRep (T. Yang et al., 2017), and GenomeDisco(Ursu et al., 2018) (Figure 1.2). These metrics were originally developed to evaluate reproducibility of biological replicates, but can also be used as similarity metrics between contact maps. Hi-CSpector transformers contact maps to laplacian matrices and performs matrix decomposition. QuASAR defines a metric called interaction

correlation matrix which is weighted by interaction enrichment. HiCRep stratifies a smooth Hi-C contact matrix by distance before measuring the weighted similarity at each stratum. GenomeDISCO utilizes random walks on a network build using contact data information to smooth contact matrices prior to computing similarity. Each of these methods has correlation coefficient ranges and can be used to evaluate the similarity of two contact maps such as VEHiCLE enhanced matrices (Chapter 3) or interpolated matrices (Chapter 5).

#### 1.2.3.1 A/B Compartments

A/B compartments are two genomic compartments spanning an entire chromosome whose component bins preferentially interact with one another. A/B compartments are computed by first calculating an expectation matrix for contact maps where expected values are the mean contact frequency between bins of a given distance(Figure 1.4b). Then an observed over expected matrix (O/E)(Figure 1.4c) is computed by dividing contact matrices by this expectation. Some bins in the O/E will have values less than 1 while others will have values greater than one, forming the start of a banding pattern. To solidify this banding pattern the O/E matrix rows are treated as vectors and used to generate a Pearson correlation matrix(Figure 1.4d). The resultant matrix has sharp banding patterns corresponding to compartment assignment. Using principal component analysis (PCA) (Figure 1.4e) this compartment assignment, matching the banding pattern, can typically be extracted from PC1. (Figure 1.4f)

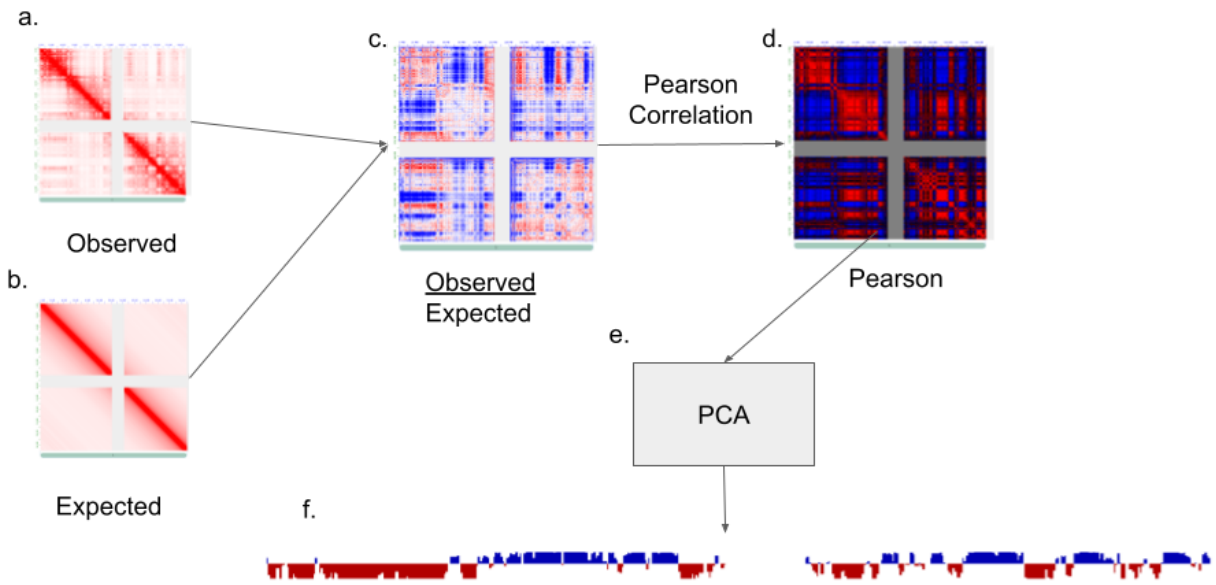



**Figure 1.4 AB Compartment Computation** (a) observed contact matrices and (b) expected contact matrices are used to build O/E (c). Rows are treated as vectors to obtain a Pearson correlation matrix, from which (e) PCA is applied. The sign of PC1 is used to define A/B compartment assignment (f).

#### 1.2.3.2 TADs

Topologically Associated Domains are regions of the genome which have higher levels of interconnectivity than interaction with adjacent regions. They can span multiple levels of resolution 10k-100kb and A wide ecosystem of tools has developed for the inspection of TADs. Oftentimes the computational methods used to identify TADs do not completely concur as the definition of a TAD is so loosely defined making TAD identification an active area of research(Zufferey et al., 2018). TAD identification tools used throughout this dissertation include: Insulation Score(Crane et al., 2015), Directionality Index(Dixon et al., 2012) and Armatus(Filippova et al., 2013) based Transitional Gamma(Rozenwald et al., 2020). Because the underlying mathematics of

these TAD identifying tools plays an important roles in other sections of this manuscript such as VEHiCLE's training mechanism (Chapter 3) and effectiveness of TAPIOCA at TAD categorization (Chapter 4) we discuss implementation at those stages in the manuscript rather than in this background section.

#### 1.2.3.2 3D Models

Hi-C contact maps can be used to construct three-dimensional (3D) chromosomal models. When working with the GSDB(Oluwadare et al., 2020) (Chapter 6) We compare a wide variety of tools for 3D Modeling and in other chapters of the manuscript we frequently use the 3DMax(Oluwadare et al., 2018) algorithm to build 3D models. We favor 3DMax because of its speed and ease of use. We exclude a full description of the underlying math of the 3DMax algorithm in this chapter, but discuss the algorithm's implementation when expanding the algorithm to predict time-series Hi-C data using 4DMax (Chapter 5).

## 1.3 Metric Formulas

Many of the metrics used for comparison of contact maps are reused throughout different problems throughout this dissertation. For simplicity we list their definitions here.

***Mean Squared Error***:

$$L_{mse}(x, y) = \sum_{i=1}(x_i - y_i)^2 \quad \text{(Eq 1.1)}$$

**Mean Absolute Error:**

$$L_{mae}(x,y) = \sum_{i=1} \frac{|y_i - x_i|}{n} \quad \text{(Eq 1.2)}$$

**Coefficient of Determination:**

$$1 - \frac{\sum_i (y_i - x_i}{\sum_i (y_i - \bar{y})} \quad \text{(Eq 1.3)}$$

**Pearson Correlation Coefficient:**

$$L_{pcc}(x,y) = \frac{\sum_{i=1}(x_i - x)(y_i - y)}{\sqrt{\sum_{i=1}(x_i - x)^2}\sqrt{\sum_{i=1}(y_i - y)^2}} \quad \text{(Eq 1.4)}$$

**Spearman Correlation Coefficient:**

Spearman Correlation is similar to pearson correlation differing in that it utilizes rank variables so as to evaluate monotonic relationship between the matrices without imposing a linearity condition that may not exist in nature.

$$L_{spc}(x,y) = \frac{\sum_{i=1}(rx_i - rx)(ry_i - ry)}{\sqrt{\sum_{i=1}(rx_i - rx)^2}\sqrt{\sum_{i=1}(ry_i - ry)^2}} \quad \text{(Eq 1.5)}$$

**Signal-To-Noise Ratio**

Signal-To-Noise Ratio uses a ratio of the clean signal to the difference between clean and noisy signals to represent how much signal is actually getting through. The higher the value of SNR the better the quality of the data.

$$L_{snr}(x,y) = \frac{\sum_{i,j} y_{i,j}}{\sqrt{\sum_{i,j}(x_{i,j} - y_{i,j})^2}} \quad \text{(Eq 1.6)}$$

**Structural Similarity Index**: SSI is calculated by sliding windows between images and averaging values. The constants $C_1$ and $C_2$ are used to stabilize the metric while the means, variances and covariances are computed via a Gaussian filter. We use the

implementation of SSI developed by Hong et al (Hong et al., 2020) keeping their default values for the size of sub-windows and variance value of the gaussian filter at 11 and 3 respectively.

$$L_{ssi}(x,y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \quad \text{(Eq 1.7)}$$

**Template Modeling Score:** TM-score(Yang Zhang & Skolnick, 2005) is a measure of similarity between two structures. It is the defacto standard measure for comparison of 3D chromosome structures(Trieu & Cheng, 2017)

$$TM - Score = max(\frac{1}{L_{target}} \sum_i \frac{1}{1 + (\frac{di}{d_0 L_{target}})^2}) \quad \text{(Eq 1.8)}$$

## 1.4 Outline

The content of each chapter in this dissertation is described as follows:

**Chapter 1** Introduction provides an overview of the biological assays used, common downstream analysis performed with these assays, and key metrics used for evaluation throughout the paper. The contents of this chapter are pulled primarily from the tutorial and publication:

**Highsmith, M.**, Cheng, J. *An Introduction to Computational Approaches for 3D Genomic Modeling*. ACM Conference on Bioinformatics and Computational Biology (ACM-BCB), 2021.

Oluwadare, O., **Highsmith, M.** & Cheng, J. An Overview of Methods for Reconstructing 3-D Chromosome and Genome Structures from Hi-C Data. *Biol Proced Online* 21, 7 (2019). https://doi.org/10.1186/s12575-019-0094-0

**Chapter 2** Describes VEHiCLE, a deep learning architecture incorporating a generative adversarial network, variational autoencoder, and biologically inspired loss functions

used to enhance the resolution of Hi-C contact maps. This work is pull primarily from the publication:

*VEHiCLE: a Variationally Encoded Hi-C Loss Enhancement algorithm for improving and generating Hi-C Data* ***Max Highsmith****, Jianlin Cheng Scientific Reports* *https://www.nature.com/articles/s41598-021-88115-9*

**Chapter 3** Describes TAPIOCA, a deep learning architecture which uses self-attention to infer the position of Topologically Associated Domains using epigenetic features. This work is pull primarily from the manuscript:

*TAPIOCA: Topological Attention and Predictive Inference of Chromatin arrangement Using Epigenetic Features* ***Max Highsmith****, Jianlin Cheng* *https://www.biorxiv.org/content/10.1101/2021.05.16.444378v1.full.pdf*

**Chapter 4** describes 4DMax, a machine learning based approach to modeling dynamic chromosomal structure changes over the course of a time-dependent genomic process. This work is pull primarily from the manuscript:

*Prediction of the 4D Chromosome Structure from Time-Series Hi-C Data* ***Max Highsmith****, Jianlin Cheng https://www.biorxiv.org/content/10.1101/2020.11.10.377002v1*

***Chapter 5*** *describes expansions applied* to GDSB, a publicly available repository that contains 3D structures of various Hi-C datasets. This work is pulled primarily from the publication:

GSDB: a database of 3D chromosome and genome structures reconstructed from Hi-C data *Oluwatosin Oluwadare;* ***Max Highsmith****; Douglass Turner; Erez Lieberman-Aiden; Jianlin Cheng, BMC Molecular and Cell Biology* *https://bmcmolcellbiol.biomedcentral.com/articles/10.1186/s12860-020-00304-y*

# 2. VEHiCLE: a **V**ariationally **E**ncoded **H**i-**C** **L**oss **E**nhancement algorithm for improving and generating Hi-C data

## 2.1 Abstract

Chromatin conformation plays an important role in a variety of genomic processes. Hi-C is one of the most popular assays for inspecting chromatin conformation. However, the utility of Hi-C contact maps is bottlenecked by resolution. Here we present VEHiCLE, a deep learning algorithm for resolution enhancement of Hi-C contact data. VEHiCLE utilises a variational autoencoder and adversarial training strategy equipped with four loss functions (adversarial loss, variational loss, chromosome topology-inspired insulation loss, and mean square error loss) to enhance contact maps, making them more viable for downstream analysis. VEHiCLE expands previous efforts at Hi-C super resolution by providing novel insight into the biologically meaningful and human interpretable feature extraction. Using a deep variational autoencoder, VEHiCLE provides a user tunable, full generative model for generating synthetic Hi-C data while also providing state-of-the-art results in enhancement of Hi-C data across multiple metrics.

## 2.2 Introduction

Hi-C data, an extension of chromosome conformation capture assay (3C) is a biological assay which can be used to inspect the three-dimensional (3D) architecture of a genome(Rao et al., 2014)(Lieberman-Aiden et al., 2009). Hi-C data can be used for downstream analysis of structural features of chromosomes such as AB compartment, Topological Associated Domains (TADs), loops, and 3D chromosome and genome models. Changes in chromosomal conformation have been empirically demonstrated to impact a variety of genomic processes

including gene methylation and gene expression(Dekker, 2008; P. Fraser & Bickmore, 2007; Miele & Dekker, 2008; Misteli, 2007).

When analysing Hi-C data, reads are usually converted into contact matrices, where each cell entry corresponds to the quantity of contacts between the two regions indexed by row and column. The size of an individual region in this contact matrix is referred to as the resolution or bin size(Oluwadare et al., 2019). The smaller the bin size, the higher the resolution. The resolution of a contact matrix is usually selected based on the quantity of read pairs in an individual Hi-C experiment, with a higher quantity of read pairs permitting a higher resolution. Certain genomic features, such as TADs, can only be meaningfully identified using high resolution contact matrices, however if a matrix resolution is selected with insufficient read coverage the matrices can be overly sparse. One method to address this issue is to run additional Hi-C experiments, however because of experimental costs this is not always a feasible solution.

To solve this problem previous groups have utilized methods from the field of Image super-resolution to improve Hi-C contact matrix resolution. The first of these networks was HiCPlus(Yan Zhang et al., 2018), a simple neural network optimized using mean squared error (mse). HiCPlus was then improved upon by HiCNN(T. Liu & Wang, 2019a, 2019b) by adjusting network architecture. Next hicGAN(Q. Liu et al., 2019) was proposed, introducing the use of Generative Adversarial Networks (GAN), which generated high resolution contact maps conditioned on low resolution input. The network DeepHiC(Hong et al., 2020) maintained the GAN loss function while extending it to also include a perceptual loss function derived from VGG-16 trained on image data. The model HiCSR(Dimmick et al., n.d.) continued the advancement by introducing the use of a deep autoencoder as a feature extraction mechanism.

Our network, the **V**ariationally **E**ncoded **Hi**-**C** **L**oss **E**nhancer (VEHiCLE) (Highsmith & Cheng, 2021), extends the approach of conditional generative adversarial networks by using an integrated training approach inspired by literature in the domains of deep learning and

genomics. First, VEHiCLE incorporates a variational autoencoder which extracts biologically meaningful features from Hi-C data. Second, VEHiCLE's decoder network is engineered to provide an easy to use generative model for Hi-C data generation which smoothly maps user tunable, low dimensional vectors to Hi-C contact maps independent of any low sampled input. Third, VEHiCLE incorporates a biologically explicit loss function based on Topologically Associated Domain identification to ensure accurate downstream genomic analysis.

VEHiCLE obtains state of the art results in the task of Hi-C superresolution across a variety of metrics pulled from the domains of Image analysis and Hi-C quality/reproducibility. VEHiCLE enhanced data show successful retrieval of important downstream structures such as TAD identification and 3DModel generation while also providing novel human interpretability of its enhancement process.

## 2.3 Approach

### 2.3.1 Description of VEHiCLE network training

Vehicle is trained as an adversarial network conditioned on low resolution input. The network is trained using a composite loss function made up of 4 sub loss functions: Adversarial loss, Variational loss, mean square error (MSE) loss, and Insulation loss. An overview of the training mechanism is displayed in Figure 2.1a. The intellectual motivation for each of these loss functions is outlined below.

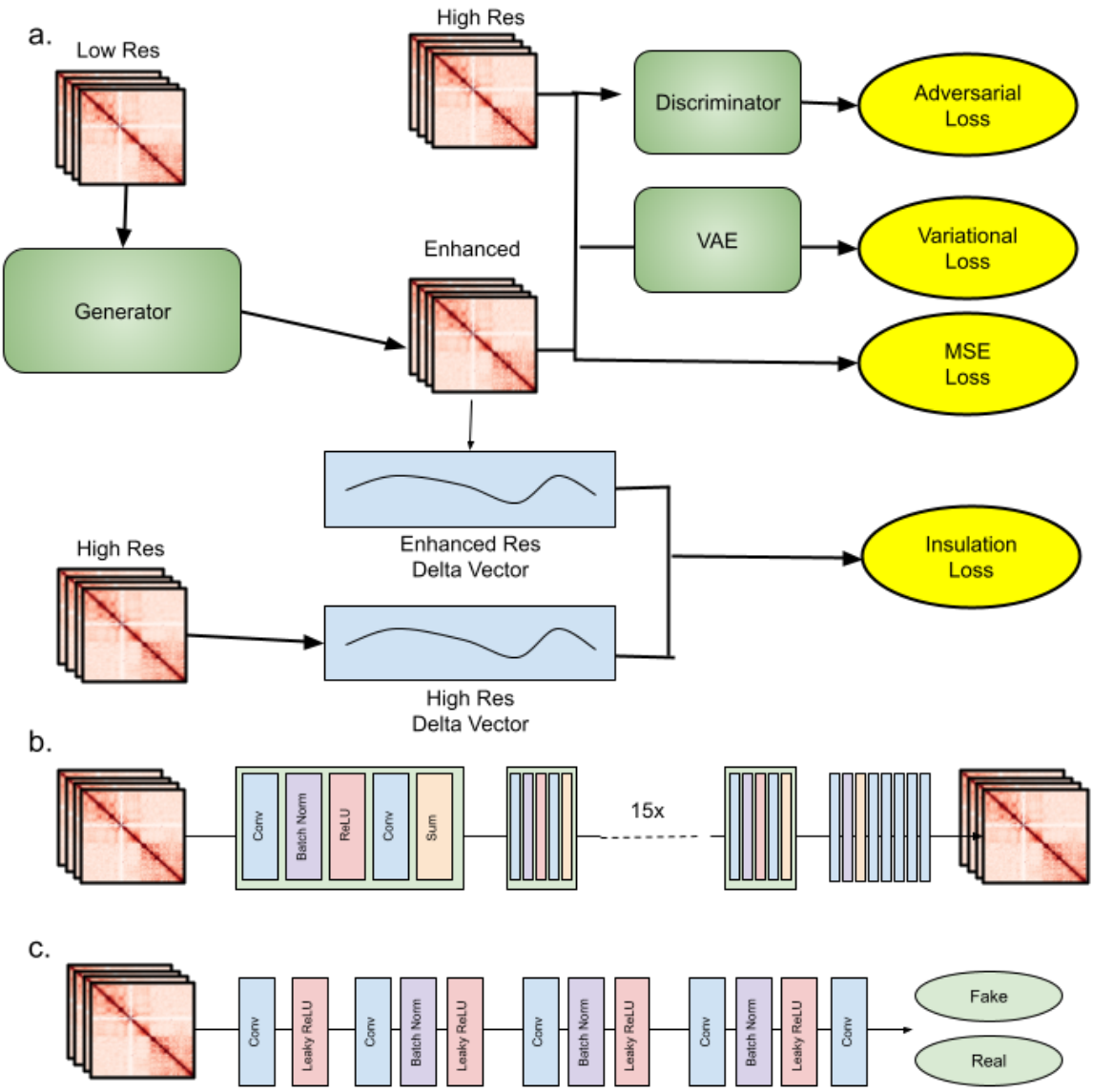



**Figure 2.1. VEHiCLE Architecture** (a) overview of training strategy (b) generator architecture (c) discriminator architecture.

## 2.3.2 Adversarial loss function

Generative adversarial networks (GANs) are a popular deep learning based framework for generative modeling which has gained traction in a wide variety of tasks including image superresolution. GANs were first introduced to the field of Hi-C super resolution through hicGAN, and later improved upon in DeepHiC and HiCSR. A GAN uses two key networks: a

generator $G$ (Figure 2.1b) and a discriminator $D$ (Figure 2.1c). The generator takes samples from an input distribution and generates enhanced matrices. The Discriminator is trained on a collection of inputs including real high resolution Hi-C samples as well as enhanced resolution Hi-C samples, and attempts to determine whether individual samples are real or enhanced. The two networks are trained in a game where the generator is rewarded for successfully tricking the discriminator and the discriminator tries to minimize classification mistakes.

$$min_G max_D V(D,G) = \mathbb{E}_{x\sim p_{data}(x)}[log(D(x))] + \mathbb{E}_{z\sim p_z(z)}[log(1-D(G(z)))] \quad \text{(Eq 2.1)}$$

The generator loss function is defined as:

$$L_{adb} = \sum_{n=1}^{N} -logD(G(X_{Low})) \quad . \quad \text{(Eq 2. 2)}$$

### 2.3.3 Variational Loss

Autoencoders are deep learning systems which map inputs from sample space to a condensed latent space via an encoder and then reconstruct images in sample space from the latent space using a decoder. The use of autoencoders for the task of Hi-C data super resolution was originally proposed in our preprint(Highsmith et al., n.d.) for the task of denoising Hi-C data. They were then suggested by Dimmick et al(Dimmick et al., n.d.) as tools for training super resolution networks by using the features extracted by passing Hi-C data through a trained autoencoder as a loss function. In this manuscript we expand upon this strategy, but replace their network with a different flavor of network called the variational autoencoder(Kingma & Welling, 2019).

Similar to vanilla autoencoders, variational autoencoders (VAE) aim to condense data into lower dimensional space, however they have the advantage of providing smooth feature representation which can permit the construction of powerful generative models. To obtain these advantages VAE relly upon a statistical method called variational inference(Blei et al., 2017). This method frames the tasks of encoding and decoding as an ancestral sampling problem with

two steps: First, a latent variable $z$ is sampled from a prior distribution $P_\theta(z)$. Second, the observed variables $x$ are drawn from a likelihood distribution $P_\theta(x|z)$.

To encode the observed variable $x$ requires the computation of the posterior distribution $p_\theta(z|x)$. However because this is computationally intractable, instead one approximates the posterior by choosing a parametric family of recognition models $q_\Phi(z|x)$ and selects parameters that minimize the divergence between the recognition model and the true underlying distribution via a probabilistic dissimilarity metric called KL-divergence,

$$D_{kl}(q_\theta(z|x)||p_\phi(z|x)) = \sum_{\mathbb{Z}} q(z) log(\frac{q(z)}{p(z)}) . \quad \text{(Eq 2.3)}$$

By performing some algebra outlined in Kingma and welling (Kingma & Welling, 2019) variational autoencoders are trained using the following loss function

$$L(\theta, \phi, x) = -D_{kl}(q_\phi(z|x)||p_\theta(z)) + \int_z q_\phi(z|x) log(p_\theta(x|z)) . \quad \text{(Eq 2.4)}$$

The integral term on the far right of the loss function ensures that the reconstruction outputs of our networks are highly similar to their original inputs, while the KL divergence term causes the latent space distribution of values to closely resemble a vector of gaussian random variables . This imposition of gaussian similarity on the latent space results in advantages in the quality of extracted features and the procurement of a generative model.

To create the variational loss function we first train our variational autoencoder using high resolution contact matrices as both inputs and labels. In each experiment our VAE network is trained using the same chromosomes as the overall VEHiCLE network. The variational autoencoder maps vectors from data space into condensed latent space, which we interpret as a lower dimensional feature vector (Figure 2ab). Because the variational autoencoder training strategy imposes a Gaussian distribution of the latent space variables and because our decoder maps latent vectors back into data space in a relatively smooth manner we expect highly similar Hi-C contact matrices to contain similar latent space profiles.

We compute variational loss by passing both the enhanced Hi-C contact matrix and target high resolution Hi-C contact matrices through the packpropagatable encoder component of our variational autoencoder network, extracting latent dimensional representations. We then compute the mean differences between their latent feature vectors,

$$L_{vae} = \frac{1}{n}\sum_{i=1}^{n} |f_{encode}(X_{enhanced}) - f_{encode}(X_{target})| \quad \text{(Eq 2.5)}$$

Where $f_{encode}$ is the encoding function, defined by our trained encoder network.

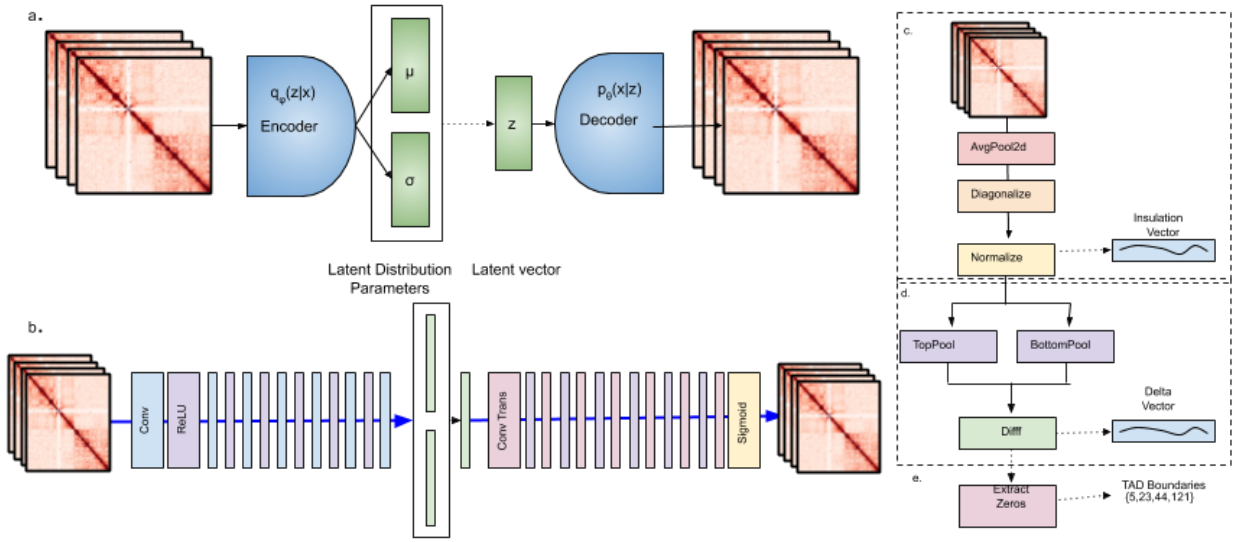



**Figure 2.2. VEHiCLE Variational Autoencoder Network.** (a) Overview of Variational Autoencoder Approach. (b) VEHiCLE architecture. Tad loss evaluated using a feedforward implementation of Insulation loss computing (c) Insulation Vector (d) Delta Vector and (e) Identification of TAD Boundaries.

## 2.3.4 Insulation Score Loss

Most of the previously proposed loss functions for developing Hi-C enhancement networks draw upon loss functions prolific in the fields of computer vision(Dimmick et al., n.d.; Hong et al., 2020; T. Liu & Wang, 2019b). While there are certainly advantages to these strategies, they derive from assumed similarities between the tasks of image superresolution

and Hi-C superresolution. However the tasks are not synonymous. Hi-C contact matrices contain important information used for downstream feature analysis such as loop calling, TAD identification and 3D model construction. Consequently images which are highly visually similar but which are blurry, shift positions of structural features, or contain noise might result in significant differences in downstream analysis. With this fact in consideration we used domain knowledge of computational genomics to devise an insulation loss function, which directly trains networks to correctly identify downstream features, specifically TAD placement.

One well-established strategy for the identification of TADs is the use of insulation scores (Crane et al., 2015). Insulation scores of a matrix are calculated by sliding a 20bin (200kb x200kb) window down the diagonal of a matrix and summing the signal across each bin, resulting in an insulation vector (Figure 2.2c). This insulation vector is normalized by taking the log2 ratio of each bin's insulation score and the mean of all insulation scores on the chromosome. From the insulation vector a delta vector is computed by observing the change in signal strength 100kb downstream and upstream of each bin on the insulation vector (Figure 2.2d). This delta vector is treated as a pseudo-derivative, and identifies insulation valleys in the regions where the delta vector crosses the x-axis from negative values to positive values, indicating a relative minimum in insulation. TAD Boundaries are assigned to each insulation valley whose difference in strength between the nearest left local max and right local min was >0.1 (Figure 2.2e).

The insulation TAD calling procedure can be encoded into a single, back propagatable network up until extraction of the delta vector (Figure 2.2cd). We define insulation loss,

$$L_{ins} = \frac{1}{n}\sum_{i=1}^{n} |D_{vec}(X_{enhanced}) - D_{vec}(X_{target})| \quad \text{(Eq 2.6)}$$

Where $D_{vec}$ is a backpropagable network which maps a contact matrix to a delta insulation vector.

### 2.3.5 Bin-Wise Mean Squared Error Loss.

Bin-wise mean square error loss is a thoroughly tested loss function used in previous Hi-C enhancement literature(Crane et al., 2015; Hong et al., 2020). It contributes to maintaining visual similarity between enhanced and target Hi-C contact matrices.

$$L_{mse} = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} |X_{enhanced} - X_{target}| \quad \text{(Eq 2.7)}$$

### 2.3.6 Composite Training Function

To capitalize on the advantages of all four loss functions we incorporate them into our comprehensive training process. First the variational network is trained on the train and validation datasets. Then the trained encoder is used for $L_{vae}$ along with the three other training losses to train the generator network, yielding our overall loss function

$$L_{tot} = \lambda_{adv} L_{adv} + \lambda_{mse} L_{mse} + \lambda_{vae} L_{vae} + \lambda_{ins} L_{ins} \quad \text{(Eq 2.8)}$$

Where $\lambda_x$ are hyperparameters used to determine loss contribution. We $\lambda_{adv} = 0.0025$, $\lambda_{mse} = 1$, $\lambda_{vae} = .01$, and $\lambda_{ins} = 1$.

## 2.4 Results

### 2.4.1 Latent space representations permit generation of synthetic Hi-C Data.

The KL divergence term in the loss function of our variational autoencoder imposes constraints on the latent dimension, pushing our estimate for the prior $q(z|x)$ towards a vector distribution of Gaussian random variables. Because all latent vector variables fall within

Gaussians centered around 0, most vectors near the center of these Gaussians can be successfully decoded into Hi-C space, resulting in a generative model for Hi-C data. We first perform principal component analysis (PCA) on our training set's learned, latent dimensional features. We then create a function mapping PCA values to the latent dimensional space. We then use our trained decoder network to transform the values in latent dimensional space into Hi-C space (Figure 2.3a). The result is a function mapping a profile of PCA values to a 2.57Mb x 2.57Mb block of Hi-C data. We hook this function into an interactive matplotlib widget, permitting manual visualization of changes to generated Hi-C data as input variables are adjusted. In our widget we set a NUM_SLIDERS=15 parameter to permit the manual tuning of PCA vector components. The widget passes a vector to our mapping function with user selected values in all manually adjusted components and dataset averages for all PCA's that are not manually selected or are above the NUM_SLIDERS component index threshold. The selection of 15 is arbitrary and can be manually increased by users interested in viewing the impact of adjusting higher PC values on the generated Matrix structure.

The zero vector results in a vanilla Hi-C map with interaction frequency between two regions following the inverse of genomic distance (Figure 2.3b). The biological interpretation of some adjustable features remains elusive, with changes to vector component values resulting merely in changes of diagonal signal strength or sporadic repositioning of contact regions. However, we observe that many of the tunable feature vector components correspond directly with biologically meaningful features in Hi-C space such as: formation of TADs, increasing TAD size (Figure 2.3c), increasing TAD frequency, shifting TAD position (Figure 2.3cd), formation of genomic stripes (Figure 2.3e)(Trenkmann, 2019) (Vian et al., 2018) and formation of chromatin loops (Roayaei Ardakany et al., 2020; Salameh et al., 2020) (Figure 2.3f).

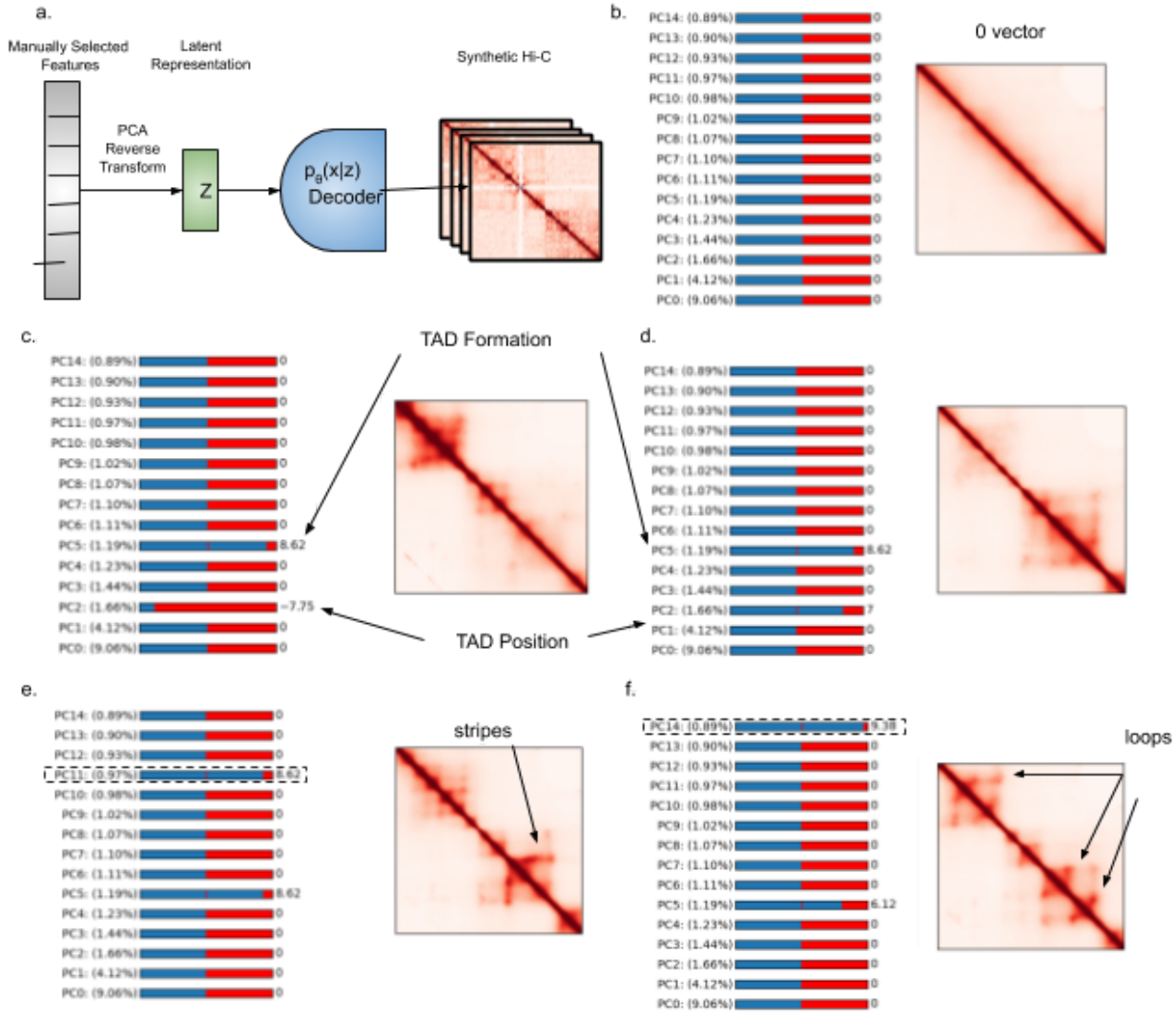



**Figure 2.3. VEHiCLE Latent Generative Model** (a) Diagram of synthetic Hi-C generation tool, a user tunable zero-centered feature vector is transformed via PCA reverse transform to latent space and then passed through our tuned decoder network. (b) The 0 vector corresponds to a purely linear contact map. (c) Increasing value of PC5 results in generation of TADs. (d) adjusting the value of PC2 shifts the position of TADs. (e) Adjusting PC11 creates stripes within TADS. (f) adjusting PC14 develops loops within TADs.

### 2.4.2 Low Resolution Hi-C contact matrices enhanced by VEHiCLE appear visually competitive with other Enhancement algorithms.

We generate visual heatmaps of Hi-C contact maps of the GM12878 dataset using VEHiCLE as well as three other previously developed algorithms: HiCSR, DeepHiC and HiCPlus. We observe high visual similarity between reconstructions by VEHiCLE and other enhancement algorithms (Figure 2.4a). We also subtracted high resolution contact maps from reconstructions by each tool to observe a visual difference matrix (Figure 2.4b). Visually VEHiCLE appears competitive with existing algorithms.

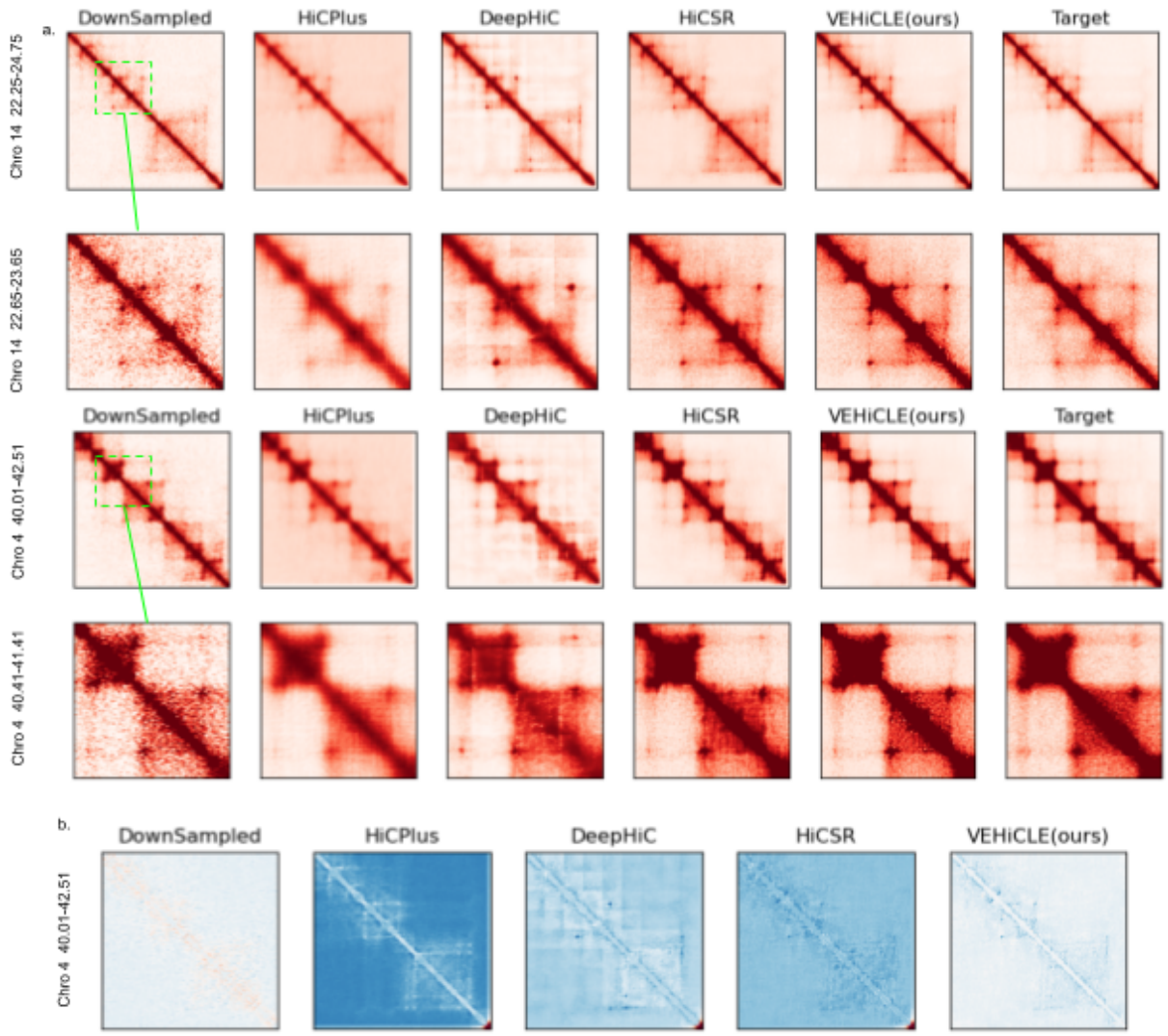



**Figure 2.4 Visual Comparison of VEHiCLE Contact Matrix Enhancement.** (a) Visual comparison of enhancement matrices. (b) Absolute difference matrices between target high resolution data and enhancement. All displayed matrices are derived from the GM12878 cell line. Architectures of previous models utilize original window size.

## 2.4.3 Notes on Evaluation Metrics

One of the major differences between the VEHiCLE algorithm and previous Hi-C enhancement tools is that our architecture is trained to enhance 2.69Mb x 2.69Mb regions along

diagonals of contact maps rather than splitting contact maps into 0.4Mb x 0.4Mb pieces, enhancing in a piecemeal fashion, and then reassembling (See Methods). This contribution permits the inclusion of more comprehensive information like TAD structure into training samples. However, it is possible to expand older architectures to full 2.69Mb x 2.69 Mb sizes rather than the condensed 0.4Mb x 0.4Mb window that appears in previous papers. In some cases this expansion of window size degrades older architecture performance, while in others it leads to enhancement. Thus, when comparing VEHiCLE to previous tools we include both original architectures without adjusting window size as well as alternative architectures trained using expanded window sizes.

### 2.4.4 Low Resolution Hi-C contact matrices enhanced by VEHiCLE achieve strong similarity to high resolution contact matrices using multiple metrics.

Using models trained and tested on the GM12878 cell line dataset We evaluated the effectiveness of VEHiCLE in predicting high resolution contacts using 5 common metrics: Pearson Correlation Coefficient (PCC), Spearman Correlation Coefficient( SPC), Mean Squared Error (MSE), Signal-to-noise ratio (SNR) and Structure Similarity Index (SSI) (see methods). We compared VEHiCLE reconstructions to the lower resolution data as well as other super resolution methods (HiCPlus, DeepHic and HiCSR.) VEHiCLE enhanced contact matrices consistently showed improvement relative to low resolution data along all 5 metrics (table 2.1). VEHiCLE frequently out-performed other Hi-C super resolution methods beating all older models with 0.4Mb window size along every test chromosome in every vision metric (table 2.1). VEHiCLE performs both the original and expanded window HiCPlus model in every vision metric across every chromosome (Table 2.1). VEHiCLE remained competitive with 2.69Mb window sized DeepHiC and HiCSR models scoring highest in PCC in 3 of the 4 test

chromosomes and scoring in the top 2 for 80% of the metric-chromosome combinations, a higher consistency of top-2 performance than any of the previous models (Table 3.1).

| **Chro 4** | Downsampled | HiCPlus 40 | DeepHiC 40 | HiCSR 40 | HiCPlus 269 | DeepHiC 269 | HiCSR 269 | VEHiCLE |
|---|---|---|---|---|---|---|---|---|
| PCC | 0.7592 | 0.9103 | 0.9212 | 0.9285 | 0.9467 | **0.9524** | 0.9463 | **0.9524** |
| SPC | 0.6259 | 0.805 | 0.7715 | 0.8292 | 0.8646 | **0.8837** | 0.8719 | **0.8739** |
| SSIM | 0.2336 | 0.3284 | 0.3784 | 0.4346 | 0.3785 | **0.4305** | **0.4526** | 0.3978 |
| MSE | 0.0468 | 0.0163 | 0.0162 | 0.0114 | 0.0091 | 0.0083 | **0.0097** | **0.0098** |
| SNR | 306.65 | 514.24 | 516.847 | 619.41 | 700.549 | 733.5042 | **673.73** | **670.9001** |
| | | | | | | | | |
| **Chro 14** | Downsampled | HiCPlus 40 | DeepHiC 40 | HiCSR 40 | HiCPlus 269 | DeepHiC 269 | HiCSR 269 | VEHiCLE |
| PCC | 0.8682 | 0.9374 | 0.9481 | 0.9583 | 0.9716 | 0.975 | **0.9753** | **0.9764** |
| SPC | 0.6692 | 0.8159 | 0.7188 | 0.85 | 0.88 | **0.9031** | 0.888 | **0.8892** |
| SSIM | 0.3524 | 0.4022 | 0.5481 | 0.5877 | 0.644 | **0.6588** | **0.67** | 0.6439 |
| MSE | 0.0145 | 0.0117 | 0.0052 | 0.0041 | 0.0027 | **0.0024** | **0.024** | **0.0024** |
| SNR | 341.5712 | 380.041 | 554.9034 | 627.752 | 786 | **830.2759** | 847.28 | **834.5026** |
| | | | | | | | | |
| **Chro 16** | Downsampled | HiCPlus 40 | DeepHiC 40 | HiCSR 40 | HiCPlus 269 | DeepHiC 269 | HiCSR 269 | VEHiCLE |
| PCC | 0.8798 | 0.9327 | 0.9479 | 0.9602 | 0.9694 | **0.9771** | 0.8771 | **0.9769** |
| SPC | 0.6684 | 0.8097 | 0.6949 | 0.8496 | 0.8887 | **0.9027** | 0.8884 | **0.8896** |
| SSIM | 0.3901 | 0.3935 | 0.5618 | 0.5924 | 0.6913 | **0.7058** | **0.7095** | 0.6948 |
| MSE | 0.0118 | 0.0124 | 0.0047 | 0.0036 | 0.0027 | **0.0021** | **0.0021** | **0.0022** |
| SNR | 332.8447 | 318.105 | 517.1062 | 597.73 | 703.324 | **808.4** | **810.28** | 797.5162 |
| | | | | | | | | |
| **Chro 20** | Downsampled | HiCPlus 40 | DeepHiC 40 | HiCSR 40 | HiCPlus 269 | DeepHiC 269 | HiCSR 269 | VEHiCLE |
| PCC | 0.9075 | 0.9303 | 0.9507 | 0.9656 | 0.9692 | 0.9825 | **0.983** | **0.983** |
| SPC | 0.6866 | 0.827 | 0.6857 | 0.8631 | 0.9094 | **0.9184** | 0.9033 | **0.905** |
| SSIM | 0.4373 | 0.4082 | 0.6022 | 0.6432 | 0.7522 | **0.7559** | **0.7619** | **0.7559** |
| MSE | 0.0076 | 0.0124 | 0.0038 | 0.0027 | 0.0023 | **0.0014** | **0.0013** | **0.0014** |
| SNR | 364.83 | 282.43 | 510.35 | 608.04 | 662.656 | 850.7672 | **886.20** | **868.3073** |

**Table 2.1 VEHiCLE Comparative Vision Metrics** Comparison of vision metrics across different

super-resolution algorithms. Networks are trained using the training set chromosomes of the GM12878 cell line and evaluated on the test chromosome set of the GM12878 cell line. Top 2 scores for each metric are bolded.

### 2.4.5 Downsampled Hi-C contact matrices enhanced by VEHiCLE display significant improvement using Hi-C specific metrics.

Using models trained on the GM12878 cell line dataset we next evaluated VEHiCLE reconstructions using 3 Hi-C specific metrics: GenomeDISCO, HiCRep and QuASAR-Rep (see methods). VEHiCLE enhanced metrics remain competitive with other methods (Table 2.2). Furthermore, even in instances where VEHiCLE is outperformed by another algorithm we consistently observe increased performance relative to original low resolution matrices. These results indicate biological consistency with VEHiCLE enhanced matrices.

| **Chr 4** | Downsampled | HiCPlus 40 | DeepHiC 40 | HiCSR 40 | HiCPlus 269 | DeepHiC 269 | VEHiCLE |
|---|---|---|---|---|---|---|---|
| GenomeDISCO | 0.941 | 0.972 | 0.945 | **0.98** | 0.972 | **0.98** | 0.972 |
| HiCRep | 0.967 | 0.974 | 0.972 | **0.989** | 0.972 | **0.99** | 0.972 |
| QuASAR-Rep | 0.924 | 0.995 | 0.993 | **0.995** | **0.995** | 0.589 | **0.995** |
| | | | | | | | |
| **Chr 14** | Downsampled | HiCPlus 40 | DeepHiC 40 | HiCSR 40 | HiCPlus 269 | DeepHiC 269 | VEHiCLE |
| GenomeDISCO | 0.942 | 0.933 | 0.907 | **0.979** | 0.975 | **0.977** | 0.972 |
| HiCRep | 0.982 | 0.969 | 0.97 | **0.991** | 0.987 | **0.992** | **0.991** |
| QuASAR-Rep | 0.944 | 0.995 | 0.993 | **0.996** | **0.996** | **0.996** | **0.996** |
| | | | | | | | |
| **Chr 16** | Downsampled | HiCPlus 40 | DeepHiC 40 | HiCSR 40 | HiCPlus 269 | DeepHiC 269 | VEHiCLE |
| GenomeDISCO | 0.927 | 0.904 | 0.88 | 0.972 | **0.967** | **0.972** | 0.969 |
| HiCRep | 0.974 | 0.948 | 0.96 | 0.987 | 0.978 | **0.988** | **0.987** |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| QuASAR-Rep | 0.941 | 0.992 | **0.99** | 0.994 | 0.994 | **0.995** | **0.995** |
| | | | | | | | |
| **Chr 20** | Downsampled | HiCPlus 40 | DeepHiC 40 | HiCSR 40 | HiCPlus 269 | DeepHiC 269 | VEHiCLE |
| GenomeDISCO | 0.934 | 0.895 | 0.864 | **0.974** | 0.968 | **0.973** | 0.948 |
| HiCRep | 0.981 | 0.949 | 0.959 | **0.988** | 0.984 | **0.989** | 0.979 |
| QuASAR-Rep | 0.955 | 0.994 | **0.99** | **0.996** | **0.996** | **0.996** | **0.996** |

**Table 2.2. VEHiCLE Comparative Hi-C Metrics** Comparison of Hi-C Superresolution algorithms using Hi-C reproducibility Metrics. Networks are trained using the training set chromosomes of the GM12878 cell line and evaluated on the test chromosome set of the GM12878 cell line. Top 2 scores for each metric are bolded.

*Our version of the HiCSR model with an expanded window size of 269 repeatedly failed to converge using these tools, thus we include only the author's original model for comparison.

## 2.4.6 VEHiCLE enhanced contact matrices effectively retrieve downstream features such as TADS

We identified TADs using the prolific insulation score method (Crane et al., 2015). This method assigns an insulation score vector by sliding a window across the diagonal of the contact matrix, constructing an insulation difference vector, and using the zeros of the insulation difference vector to discover TAD boundaries. We used models trained on the GM12878 cell line and evaluated insulation on test chromosomes for the HMEC, K562 and IMR90 cell lines as well as the GM12878 cell line. We expand the test set to evaluate the effectiveness of our network at predicting downstream biological features like TADs when the model is trained on different cell lines which may have different TAD profiles.

We compare the insulation difference vector of each matrix-enhancement algorithm to the insulation difference vector of our high resolution contact matrix using the L2 norm dissimilarity metric. In many cases VEHiCLE enhanced insulation difference vectors have higher similarity to target matrices relative to other matrix enhancing algorithms. (Table 2.3).

Furthermore, even in instances where VEHiCLE is outperformed by another algorithm we consistently observe higher similarity between the target high resolution matrices and VEHiCLE enhanced matrices relative to low resolution matrices. (Table 2.3).

| Norm of Insulation Score Difference Vectors | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **Chr 4** | Downsampled | HiCPlus 40 | DeepHiC 40 | HiCSR 40 | HiCPlus 269 | DeepHiC 269 | HiCSR 269 | VEHiCLE |
| GM18278 | 7.966 | 6.64 | 7.52 | 4.389 | 7.8217 | **4.3782** | **4.7323** | 4.763 |
| K562 | 10.942 | 9.1133 | 9.957 | 7.605 | 10.688 | 7.976 | **7.11** | **7.305** |
| IMR90 | 9.8344 | 8.5681 | 9.2244 | 5.9457 | 9.78 | **5.736** | 6.091 | **5.5729** |
| HMEC | 16.143 | 13.132 | 15.267 | 10.17 | 15.8212 | 11.6367 | **10.1420** | **11.2512** |
| | | | | | | | | |
| **Chr 14** | Downsampled | HiCPlus 40 | DeepHiC 40 | HiCSR 40 | HiCPlus 269 | DeepHiC 269 | HiCSR 269 | VEHiCLE |
| GM18278 | 2.68 | 2.619 | 3.774 | 2.898 | 2.97 | **2.305** | 2.473 | **2.3414** |
| K562 | 6.225 | 5.927 | 6.329 | 5.548 | 6.28 | 5.1104 | **4.8282** | **4.868** |
| IMR90 | 4.3609 | 4.6005 | 5.1827 | 3.871 | 4.838 | 3.309 | **3.284** | **3.244** |
| HMEC | 9.214 | 8.34 | 9.549 | 7.448 | 9.2113 | **6.9471** | **6.5814** | 7.0423 |
| | | | | | | | | |
| **Chr 16** | Downsampled | HiCPlus 40 | DeepHiC 40 | HiCSR 40 | HiCPlus 269 | DeepHiC 269 | HiCSR 269 | VEHiCLE |
| GM18278 | 4.162 | 3.467 | 3.769 | 3.099 | 4.3619 | 2.6623 | **2.3862** | **2.4376** |
| K562 | 6.653 | 5.903 | 6.485 | 5.14 | 6.817 | 4.6 | **4.465** | **4.572** |
| IMR90 | 5.806 | 5.0148 | 5.459 | 4.169 | 6.134 | 3.556 | **3.117** | **3.376** |
| HMEC | 8.957 | 8.353 | 8.799 | 7.527 | 9.1517 | **6.4103** | **6.068** | 6.4423 |
| | | | | | | | | |
| **Chr 20** | Downsampled | HiCPlus 40 | DeepHiC 40 | HiCSR 40 | HiCPlus 269 | DeepHiC 269 | HiCSR 269 | VEHiCLE |
| GM18278 | 2.077 | 2.419 | 2.587 | 2.624 | 2.5274 | **1.8383** | **1.807** | 1.922 |
| K562 | 5.316 | 4.835 | 5.021 | 4.307 | 5.4488 | 4.267 | **3.811** | **3.908** |
| IMR90 | 2.888 | 3.522 | 3.444 | 3.083 | 3.5723 | 2.3699 | **2.2602** | **2.3169** |
| HMEC | 6.383 | 6.662 | 6.579 | 5.805 | 6.562 | **4.7832** | **4.701** | 4.8159 |

**Table 2.3. VEHiCLE Comparison of TAD Insulation.** L2 norm of TAD Insulation difference vectors against target insulation vectors. Networks are trained using the training set chromosomes of the GM12878 cell line and evaluated on the test chromosome sets of the K562, IMR90, HMEC and GM12878 cell line. The top 2 scores for each metric are bolded.

### 2.4.7 3D chromatin model construction

We tested the effectiveness of reconstructed data in building 3D structure models using the structural modeling tool 3DMax. We extracted constraints from the low resolution, high resolution and VEHiCLE-enhanced 2.57Mbx2.57Mb regions of our test dataset chromosomes of the GM12878 dataset. From each constraint grouping we generated 3 models. We observed significantly higher visual similarity between VEHiCLE-enhanced and high-resolution matrices relative to low-resolution matrices (Figure 2.5a). We then used the TM-score metric to quantify structural similarity of models (Yang Zhang & Skolnick, 2005). We observed higher TM-scores between high resolution and VEHiCLE-enhanced matrices than between high resolution and low resolution models (Figure 2.5b). We also observed higher TM-score similarities between models generated by the same VEHiCLE-Enhanced matrices relative to models generated by the same low resolution matrices, indicating VEHiCLE enhanced models are more consistent (Figure 2.5c).

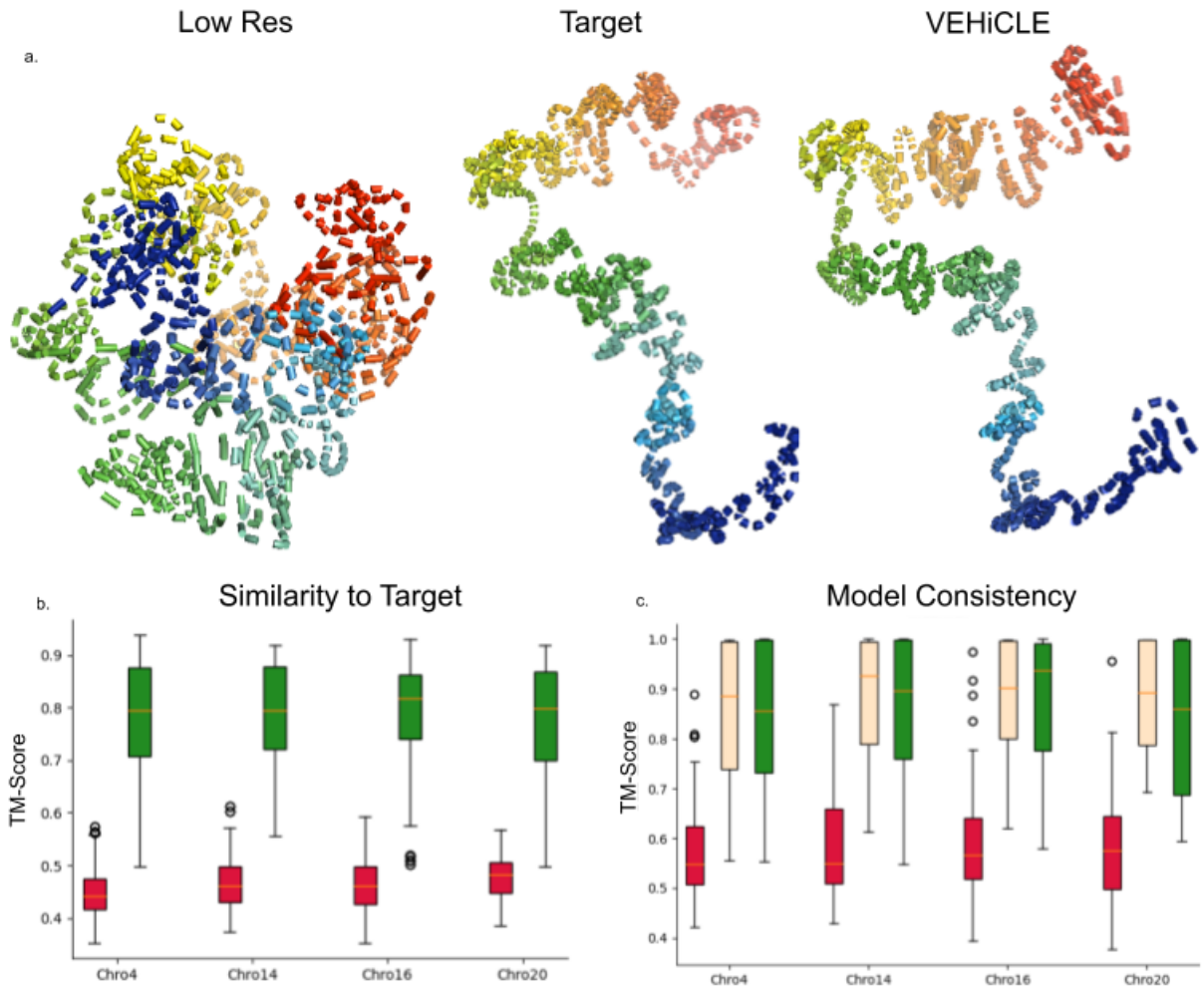



**Figure 2.5 VEHiCLE Enhanced 3DModel Generation**. (a) 3D reconstruction of Chro 20 0.6MB-3.1MB. (b) TM -score comparison of High Resolution structures to (red) Low resolution structures and (green) VEHiCLE enhanced structures. VEHiCLE enhanced scores are significantly higher (wilcoxon rank sum p value < 1e-20) (c) Average TM-Score comparison of ingroup structures generated by same contact matrix (red) low res, (yellow) high res, (green) VEHiCLE enhanced. VEHiCLE enhanced scores are significantly better than low-resolution scores (wilcoxon rank sum p value < 1e-20) Structures are all generated from GM12878 cell line using the test chromosome set: 4, 14,16,20.

## 2.5 Discussion

One of the most common challenges in Deep Learning projects is the opaque nature of a neural network's inner functioning. Consequently our ability to extract latent features and map them to biologically relevant structures provides a significant advance in increasing interpretability of Hi-C matrices. Our GUI tool can be used to generate Hi-C data through user tunable parameters with biologically relevant downstream structures such as TAD strength, TAD positioning, stripes and loops. Further inspection of these features have potential to enhance analysis of key characteristics of chromatin organization

Our introduction of the Insulation loss sets a new precedent of utilizing biological knowledge in the training of Hi-C networks. This may open the door for future improvement of Hi-C data enhancement by utilizing other forms of domain knowledge to increase usability of deep learning enhanced matrices. Future loss functions could incorporate algorithms for identification of other important downstream features such as loops or stripes.

In addition to the increased interpretability and inclusion of domain knowledge, VEHiCLE obtains resolution enhancement results competitive with the state-of-the art, often beating top algorithms on a variety of metrics, all while preserving the ability to convey meaningful structures such as TAD's and 3D structure in downstream analysis.

VEHiCLE's capacity to increase accuracy of insulation scores shows promise of utility for experimental biologists interested in chromosome architecture at specific genomic locations. By enhancing experimentally obtained Hi-C data a biologist could observe the frequency with which a list of genes or cis regulatory elements are found near TAD boundaries. Such analysis could provide further insight into the role of structural organization in a genomic process. Additionally, VEHiCLE enhanced matrices could be used to generate more accurate 3D models when building visualizations of genomic structure. These visualizations may provide insight into the underlying machinery of a genomic process of interest.

## 2.6 Methods

### 2.6.1 Dataset Assembly

Like many of the previous Hi-C super resolution networks we train VEHiCLE on high and low resolution Hi-C data for the GM12878 cell line(Lieberman-Aiden et al., 2009). While previous work often split chromosomes into training, validation and testing sets in a sequential manner we were concerned that differences in the 3D conformation of large vs small chromosomes(T. Liu & Wang, 2019b; Yan Zhang et al., 2018) may contain implicit bias in contact map features that could confound training. Consequently we assembled training, validation and test sets in a non sequential manner using chromosomes 1,3,5,6,7,9,11,12,13,15,17,18,19,21 as our training set, chromosome 2,8,10,22 as our validation set and chromosomes 4,14,16,20 as our test set.

Previous work on Hi-C super resolution consistently used network input window sizes of 0.4Mb x 0.4Mb at 10kb resolution, requiring networks to split chromosome contact maps into 40x40bin matrices(Dimmick et al., n.d.; T. Liu & Wang, 2019b; Yan Zhang et al., 2018). While this strategy has seen relative success, a major disadvantage is that certain important features of Hi-C such as TADs can span ranges larger than 0.4Mb, meaning that it is impossible for previous networks to explicitly encode important information about TAD organization. Furthermore this informational bottleneck of constraining window sizes to 40x40 bins is not incumbent upon the employed super-resolution networks as work in the field of computer vision has demonstrated the effectiveness of GAN and VAE networks on significantly larger images. With these considerations in mind we instead built our network to accept 2.69 Mb x 2.69Mb images, a range which is large enough to fully encompass the average TAD of length 1MB(Dali & Blanchette, 2017). Observing 2.69Mb x 2.69Mb regions of Hi-C contact maps at range 10kb results in submatrix images of 269x269 bin size. Because of the expanded window size we trained our network exclusively on diagonally centered submatrices, split by sliding a 269x269

window down the diagonal of each chromosome's Hi-C contact map. We move the window with a stride of 50 bins at a time, ensuring sufficient overlap between samples for our dataset to include all contacts between regions within 2Mb of each other. This results in a total of 3309 training, 1051 validation, and 798 testing matrices.

Because the convolutional arithmetic of our GAN architecture results in a decrease in output matrices by 12 bins, our output matrices are of dimension 257x257. Our variational loss is based on reconstruction of matrices output by our GAN, thus when training our variational autoencoder we use the inner 257 x 257 bins of each 269x269 sample in our dataset.

All Models were trained using the GM12878 cell line. When evaluating vision metrics, Hi-C qc metrics and 3D model comparison we use the test chromosomes from the GM12878 cell line. For our insulation score analysis we extend our test set to include the K562, IMR90 and HMEC cell lines so as to verify the effectiveness of our network at retrieving information when trained on a different cell line. Both low resolution and high resolution contact maps are normalized using the Knight-Ruiz algorithm, a standard normalization method in the Hi-C literature.

### 2.6.2 Variational Autoencoder architecture

The VAE component of VEHiCLE utilizes two neural networks for the encoding and decoding components, where the encoder is trained for the parameters of $q_\theta$ and the decoder is trained to optimize the parameters of $p_\theta$. The VEHiCLE encoder network contains 7 convolutional layers with kernel counts: 32,64,128,256,256,512,512. Each convolutional layer is separated by leaky ReLU and batch normalization. The decoder network has 7 layers of convolution transpose with the kernel counts 512, 512, 256, 256, 128, 64, 32, also separated by leaky ReLU and batch norm functions. The decoder network is appended by a Sigmoid activation function placing outputs in the range of [0,1].

### 2.6.3 Generative Adversarial Network Architecture

We use the discriminator and generator architecture defined in HiCSR, with the exception of our generator's output function, which is changed from tanh, to a sigmoid so that outputs are mapped to [0,1]. The generator architecture contains 15 residual blocks separated by skip connections, each containing 64 convolutional filters. The fully convolutional discriminator is a fully convolutional network with ReLU activation. Both the generator and discriminator are trained with batch normalization

### 2.6.4 Other Networks

We used the pytorch versions of HiCPlus, DeepHiC and HiCSR provided at https://github.com/wangjuan001/hicplus, https://github.com/omegahh/DeepHiC and https://github.com/PSI-Lab/HiCSR. We first tested networks using their literature provided weights, however we obtained very poor performance because these networks were trained on alternative training sets with key characteristic differences from ours. First, their training sets had bin value ranges of [-1,1], however our training datas range was [0,1] because negative values confound the probabilistically motivated VAE component. Second, the input size of contact maps for previous networks was 40x40, while our network aims to incorporate surrounding genomic information and utilizes a larger window input size of 269 x 269. To provide more accurate comparison we trained networks on our own GM12878 Dataset. Because our networks accept a large scale input matrix 269x269, but other networks were built to accept 40x40 pieces, we trained other networks by splitting each 269x269 into 36 non-overlapping pieces. Evaluation of Hi-C metrics was performed by feeding split pieces through networks as necessary, then reassembling pieces and comparing full chromosome contact maps.

### 2.6.5 Standard Evaluation Metrics

We utilize 5 reproducibility metrics pulled from image-super resolution literature: Pearson Correlation Coefficient (PCC), Spearman Correlation Coefficient( SPC), Mean Squared Error (MSE), Signal-to-noise ratio (SNR) and Structure Similarity Index (SSI) Implementation of these metrics is available in Chapter 2.

### 2.6.6 Hi-C Reproducibility Metrics

We consider 3 Hi-C specific reproducibility metrics: GenomeDISCO, HiCRep, and QuASAR-Rep. We use the 3DChromatin_ReplicateQC(Yardımcı et al., 2019) implementations of the metrics. When expanding previous models to a 269x269 window size the HiCSR model repeatedly failed to converge using these metrics, thus we only include the original 40x40 window version of HiCSR in our evaluation of Hi-C Reproducibility metrics.

### 2.6.7 Topologically Associated Domain Identification

Topologically associated domains were identified using Insulation score as identified in Crane et al(Crane et al., 2015). We mimicked their procedure entirely with the exception that our initial insulation score window size was condensed to 20 bins instead of 50 because this demonstrated greater visual accuracy in TAD positioning.

### 2.6.8 Three Dimensional Model Reconstruction

To generate models we utilize 3DMax(Crane et al., 2015; Oluwadare et al., 2018) with out-of-the-box parameters of 0.6 conversion factor, 1 learning rate, and 10000 max iteration. We create 3 models per input contact matrices. We generate models for every 5th 269Mbx269Mb input matrix from our training dataset, because this skipping

distance ensures coverage of each chromosome while minimizing model generation time. Similarity between structures was measured using TM-score(Yang Zhang & Skolnick, 2005).

### 2.6.9 Motivation for 269x269 window size

The decision to expand our window size to 2.69x2.69Mb is multifaceted. Philosophically the decision to expand beyond the previous standard of 0.4Mbx0.4Mb was to permit the inclusion of a wider range of genomic information in our deep learning methods.

(Len of insulation vector) = (Len of Hi-C Axis) - (insulation window) - (2*delta window-1) Thus, using a 20kb insulation window and 10kb Delta window with the previously applied 40x40 window would result in an insulation vector of length (40-20-19)=1, which is only a scalar and would contain insufficient information for meaningful feature extraction.

The decision to use 269 as opposed to a different, large number is due to our variational autoencoder. While passing through the variational autoencoder the dimension of an input matrix is compressed with each incremental layer. It was essential that at each step the output dimension remained a whole number and that when the latent representation is decoded back into contact matrix space the reconstructed matrix be of the same dimension as its input. 257 was the smallest number which both spanned 2Mb (a range that would encompases nearly all TADs) and resulted in the same dimensional input and output at each layer of our variational autoencoder. We account for the 12 bin decrease in size that occurs by passing through our GAN, resulting in a 269x269 matrix.

### 2.6.10 Data availability

All Hi-C data were downloaded from the Gene Expression Omnibus (GEO) GSE63525. For the Hi Resolution Matrices of GM12878, IMR90, K562 and HMEC we used GSE63525_GM12878_insitu_primary+replicate_combined_30.hic,

GSE63525_IMR90_combined_30.hic, GSE63525_K562_combined_30.hic and GSE63525_HMEC_combined_30.hic respectively. For low resolution matrices we used GSM1551550_HIC001_30.hic, GSM1551602_HIC053_30.hic, GSE63525_K562_combined_30.hic, and GSM1551610_HIC061_30.hic respectively.

# 3. TAPIOCA: **T**opological **A**ttention and **P**redictive **I**nference **o**f **C**hromatin **A**rrangement Using Epigenetic Features

## 3.1 Abstract

Chromatin conformation is an important characteristic of the genome which has been repeatedly demonstrated to play vital roles in many biological processes. Chromatin can be characterized by the presence or absence of structural motifs called topologically associated domains. The de facto strategy for determination of topologically associated domains within a cell line is the use of Hi-C sequencing data. However Hi-C sequencing data can be expensive or otherwise unavailable. Various epigenetic features have been hypothesized to contribute to the determination of chromatin conformation. Here we present TAPIOCA, a self-attention based deep learning transformer algorithm for the prediction of chromatin topology which circumvents the need for labeled Hi-C data and makes effective predictions of chromatin conformation organization using only epigenetic features. TAPIOCA outperforms prior art in established metrics of TAD prediction, while generalizing across cell lines beyond those used in training.

## 3.2 Introduction

Hi-C data can, in addition to many other applications, be used to identify regions of the genome with preferentially self-interacting regions termed topologically associated domains (TADs). Substantial scientific attention has been directed at the development of tools for identifying TAD's using Hi-C data (Zufferey et al., 2018).

It has been demonstrated that epigenetic features such as repressive histone modifications have preferential association with inter-TAD boundaries, indicating the potential contribution of epigenetic modifications in the construction of TADs. TAD prediction using epigenetic features was first formulated as a classification problem using logistic regression to predict boundaries(Ulianov et al., 2016). The approach was later expanded to include lasso regression and gradient boosting(Ramírez et al., 2018). The task was then reformulated as a linear regression problem in which epigenetic features were used to predict transitional gamma, a continuous metric created by the authors based on TAD identification tool Armatus. Recently the first application of neural networks, specifically LSTM obtained state-of-the art predictions(Rozenwald et al., 2020).

This paper provides two important contributions relative to prior research in this area. First we build upon the use of machine learning methods by applying self-attention through a variant of the state-of-the-art Transformer model which we call TAPIOCA (**T**opological **A**ttention and **P**redictive **I**nference of **C**hromatin **A**rrangement). Second, we extend the metrics for TAD characterization beyond the previously used transitional gamma to incorporate more prolific metrics for TAD characterization such as Insulation Score(Crane et al., 2015) and Directionality Index(Dixon et al., 2012). Through these extensions and comparative analysis to the results of previously suggested models, we strengthen the case for dependence between epigenetic profile and TAD formation.

## 3.3 Results

### 3.3.1 Overview of Dataset Features and Labels

Previous work on prediction of topological organization in Drosophila based on epigenetic features has used a metric called Transitional gamma. Transitional gamma is computed by performing TAD calling using the armatus tool with gamma values 1-10 and assigning the transitional gamma of a loci to be the first gamma value at which armatus identifies a TAD boundary (Figure 3.1a). In our experiments we use the transitional gamma values assigned to the feature dataset by (Ramírez et al., 2018).

In addition to using the standard transitional gamma, we use Hi-C data (Figure 1b) to extract two additional labels for TAD characterization, which are prolific in the Hi-C literature: Directionality Index (Figure 3.1c) and Insulation Score (Figure 3.1d).

We compute directionality index using a procedure based on Dixon et al (Dixon et al., 2012). Directionality Index is motivated by the observation that downstream portions of a domain are highly biased towards interactions with upstream bins. Directionality Index is computed using Equation 1, where A is a quantity of reads mapped from the observed bin to R bins downstream; B is a quantity of reads mapped from an observed bin to R bins upstream, and E is the mean of B and A. The result is a 1 dimensional vector with values corresponding to each genomic loci within R bins of the chromosome border. In the original Directionality Index literature the directionality vector is then passed to a hidden markov model, however, for ease of comparison we treat the directionality vector as our labels. Our directionality uses a radius of 10.

$$DI = \left(\frac{B-A}{\sqrt{B-A^2}}\right)\left(\frac{(A-E)^2}{E} + \frac{(B-E)^2}{E}\right) \quad \text{(Eq 3.1)}$$

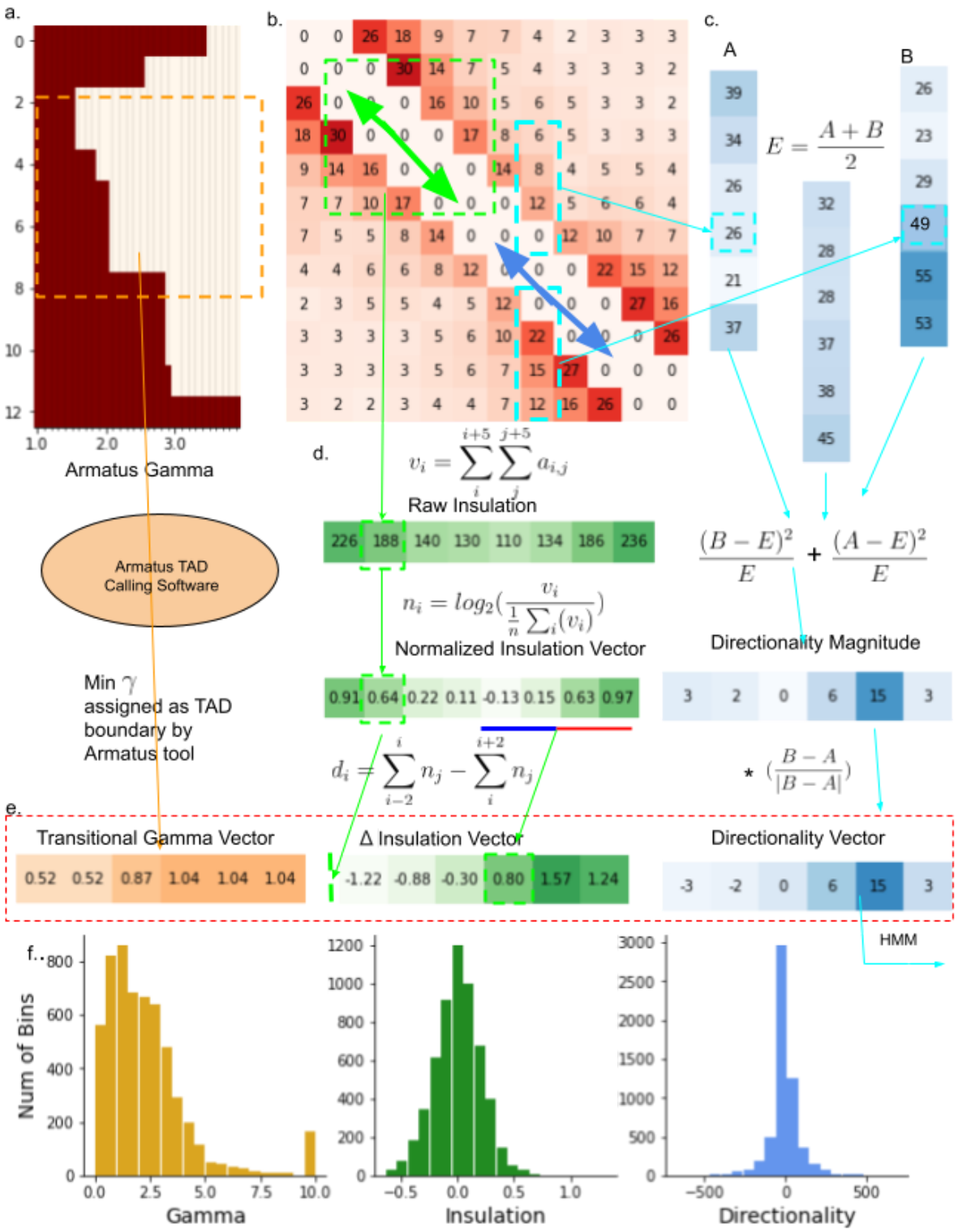



**Figure 3.1 TAPIOCA Dataset Overview.** (a) Generation process of transitional gamma labels. (b) Hi-C Contact matrix (c) Generation process of Directionality Index labels. (d) Generation

process of Insulation score labels. (e) vectors used as labels in the training process. (f) Distribution of numerical values for each metric.

We compute the insulation score using a procedure based on (Crane et al., 2015). Insulation Score is motivated by the intuition that regions which have drastic changes in quantity of interactions with their neighbors are likely boundaries for TADS. Insulation Score (Figure 3.1d) is calculated by sliding a window of radius R along the diagonal of a contact matrix and computing the sum of signals across each bin. This vector is then normalized and a Difference Vector is computed by observing changes in the summed value of L bins before and after a loci of interest. The result is a 1 dimensional vector with values corresponding to each genomic loci within R+L bins of the chromosome border. In the original Insulation Score literature the regions of the Difference vector where values switch sign are marked as potential TAD boundaries. In our experiments we use the full vector as the label. We use R=3 and L=10.

The extracted vectors are ultimately used as predictive labels in the formulation of a supervised regression problem (Figure 3.1e). We note that there are stark differences in the range and distribution shape of these three classifications of TADs (Figure 3.1f). For example Gamma remains bi-modal while insulation and directionality resemble normal distributions, with directionality showing extremely high central concentration. These differences in distribution may contribute to the varying effectiveness of our explored network on TAD prediction.

### 3.3.2 Overview of TAPIOCA Network

The TAPIOCA network is inspired by the transformer architecture as an approach to the problem of Seq2Seq language translation(Chen et al., n.d.; Duan & Zhao, 2020).

To convert the transformer network from the task of language translation to TAD prediction, we treat epigenetic features as though they are word embeddings and append a final linear prediction layer converting transformer outputs to numerical values for TAD labels (Figure 3.2a). The core processing component of the TAPIOCA architecture is made up of a series of attention layers each containing Scaled Dot-Product Attention with 7 heads followed by a linear neural network layer (Figure 3.2b). We treat the number of attention layers as a hyper parameter. We provide a composite view of the epigenetic inputs, labels, and network predictions (Figure 3.3).

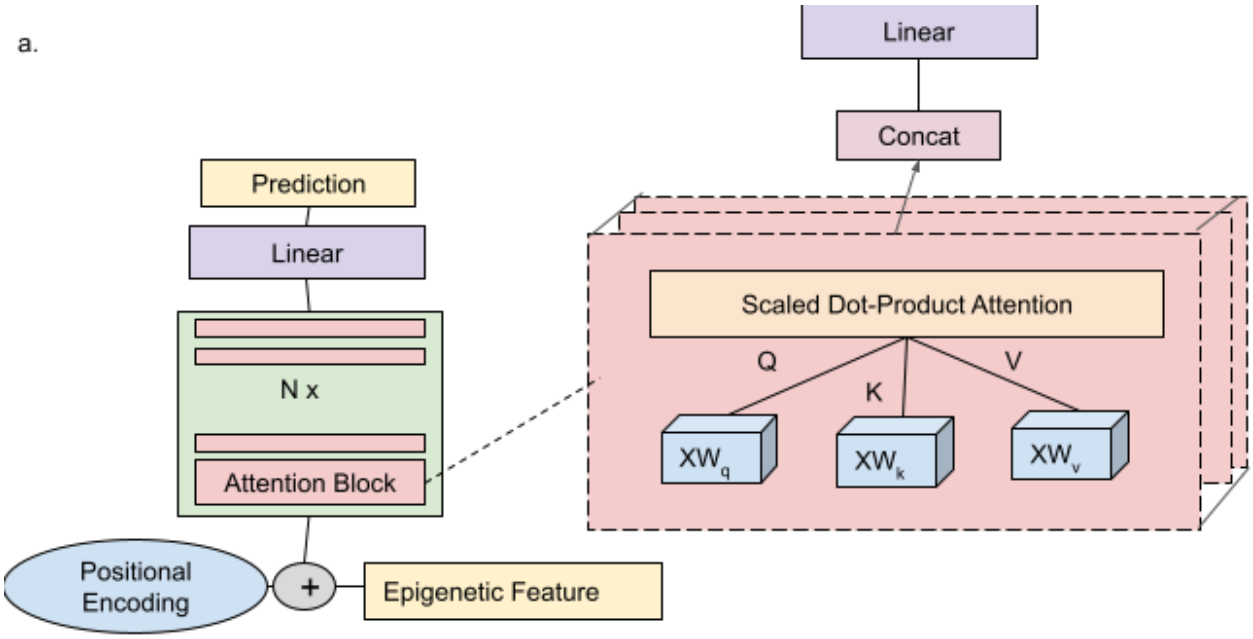



**Figure 3.2 TAPIOCA Architecture.** (a) Overview of architecture (b) Details of multi-head attention block.

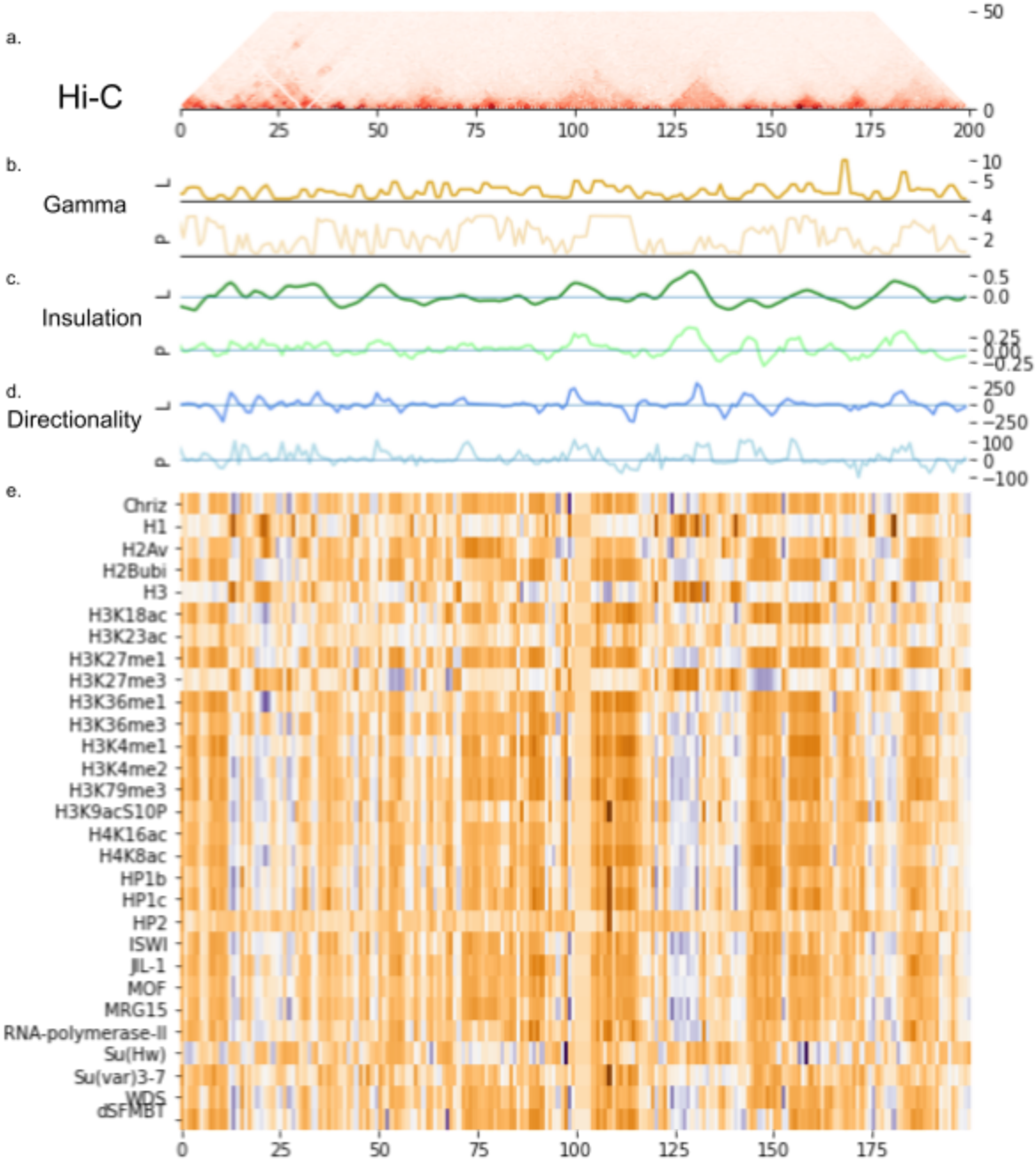



**Figure 3.3 Visualization of TAPIOCA Predictive Process.** (a) Hi-C track data (b) Label and predicted values for transitional gamma, (c) insulation vector, and (d) directionality index. (e) epigenetic track data values.

### 3.3.3 Benchmark of TAPIOCA Network Relative to Prior Art

We compared TAPIOCA-network's performance on the task of TAD prediction to the performance of previously used models such as linear regression(Ulianov et al., 2016), ridge and lasso regression (Ramírez et al., 2018) and Bi-directional Long Short-Term Memory (BILSTM) (Rozenwald et al., 2020) We observe visual similarity between the predictions of our network and Hi-C derived labels across all three metrics (Figure 3.4). TAPIOCA-network outperforms all previous approaches on the transitional gamma dataset (Table 3.1). In insulation vector experiments TAPIOCA-network outperforms all linear regression variants while remaining competitive with BILSTM (Table 3.2). TAPIOCA-network was the only network capable of effectively predicting Directionality index, even after extensive hyperparameter tuning of other networks (Figure 4c, Table 3.3).(Ramírez et al., 2018))

| metric | mse | mae | r2 | pcc | spc |
|---|---|---|---|---|---|
| linear | 5.296 | 2.063 | -3.290 | 0.530 | 0.548 |
| lasso | 3.993 | 1.695 | -2.235 | 0.521 | 0.544 |
| ridge | 1.941 | 1.123 | -0.573 | 0.522 | 0.542 |
| bilstm | 1.410 | 0.901 | -0.143 | 0.494 | 0.507 |
| transformer | **1.267** | **0.855** | **-0.026** | **0.506** | **0.553** |

**Table 3.1 TAPIOCA Gamma Metrics** Performance metrics of varying models using transitional gamma labels on s2 cell lines.

| metric | mse | mae | r2 | pcc | spc |
|---|---|---|---|---|---|
| linear | nan | nan | nan | nan | nan |
| lasso | 0.070 | 0.206 | -0.002 | 0.058 | 0.074 |
| ridge | 0.070 | 0.205 | 0.001 | 0.038 | 0.040 |
| bilstm | **0.058** | **0.190** | **0.162** | **0.420** | **0.394** |

| | | | | | |
|---|---|---|---|---|---|
| transformer | 0.065 | 0.200 | 0.062 | 0.325 | 0.333 |

**Table 3.2 TAPIOCA Insulation Metrics** Performance metrics of varying models using insulation vector labels on s2 cell lines.

| metric | mse | mae | r2 | pcc | spc |
|---|---|---|---|---|---|
| linear | 8499.176 | 57.258 | -0.019 | -0.039 | -0.024 |
| lasso | 8349.549 | 55.661 | -0.002 | -0.007 | 0.003 |
| ridge | 8337.046 | 55.499 | 0.000 | 0.019 | 0.013 |
| bilstm | 10696.080 | 78.295 | -0.283 | -0.055 | -0.020 |
| transformer | **6534.423** | **55.287** | 0.216 | **0.509** | **0.515** |

**Table 3.3 TAPIOCA Directionality Metrics** Performance metrics of varying models using directionality index labels on s2 cell lines.

### 3.3.4 TAPIOCA Network Remains Effective Across Cell Lines

One of the most important characteristics of any machine learning algorithm is its ability to generalize. To ensure that our network's predictive ability is not constrained to cell lines for which Hi-C data is already available, we test the effectiveness of TAPIOCA at predicting TAD organization on cell lines which differ from the network training dataset. We observe that in most instances the network's performance remains high even when training and test cell lines differ (Figure 3.5), in certain instances performing marginally better using different cell lines. Gamma obtains high values for R2 across all 5 metrics regardless of train test combination. The insulation vector and directionality index metric also obtain comparable results across training test cell line combinations in most instances.

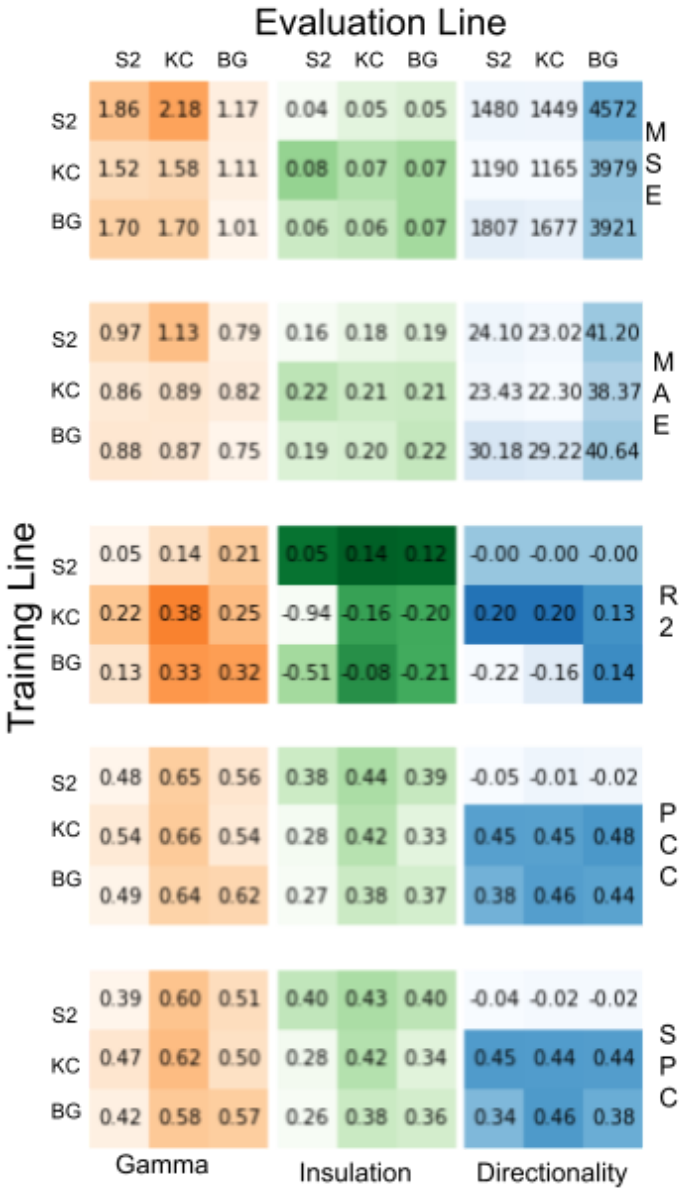



**Figure 3.5 TAPIOCA Performance Across Cell Lines.** Rows indicate training set cell lines, columns indicate testing set cell lines using (orange) transitional gamma, (green) insulation vector, and (blue) directionality index labels. Super rows show metric of evaluation: mean squared error, mean average error, r2, pearson correlation and spearman correlation.

### 3.3.5 Key Epigenetic Features in TAD prediction using TAPIOCA Network Resembles the Key Features Observed in Prior Art

We ran experiments excluding each epigenetic feature from training in both our TAPIOCA network and the previous state of the art BiLSTM network. We observe high consistency in the evaluated performance of the TAPIOCA network across the metrics: mean average error, mean squared error, pearson correlation and spearman correlation (Figure 3.6a). We observe similar degradation of performance across both networks when excluding features (Figure 3.6b). Removing certain epigenetic features such as Chriz and Su(HW) showed sharp relative drops in performance on both networks, however, other previously identified features such as H3K27me3 and H3K27ac (Rozenwald et al., 2020) showed low relative degradation in performance of the TAPIOCA network.

### 3.3.6 Epigenetic Features have different priority in predictive ability based on TAD label selection

We ran experiments excluding each epigenetic feature across the three TAD identifying metrics (Figure 3.6c). In each case networks were trained using the hyperparameters which obtained best results in full features experiments. In many instances the directionality networks failed to converge to meaningful results. This may be due to the directionality datasets demonstrated difficulty and requirement for extreme hyper parameter tuning. While the insulation networks and gamma networks both converged in most feature exclusion experiments, the prioritization of values for epigenetic features showed little correlation (Figure 3.6c). While removing features such as dSFMBT, WDS and CHRIZ all showed pronounced decreases in performance

for gamma prediction, the performance was not noticeably lower for these features in insulation score prediction relative to the removal of other epigenetic features such as H3K27Me3 and H3K27Ac .

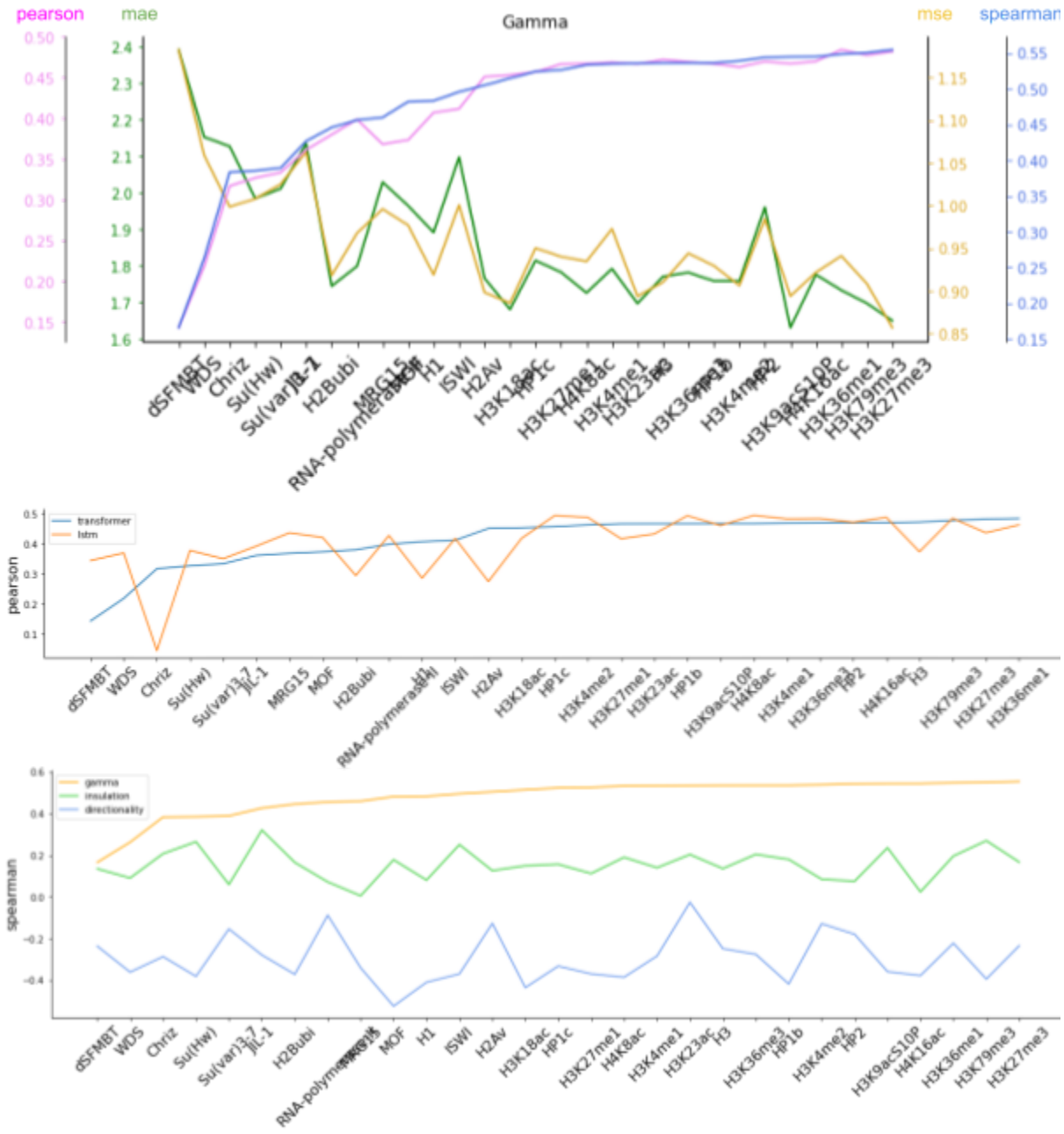



**Figure 3.6 TAPIOCA Feature Removal.** (a) performance of TAPIOCA network when excluding single epigenetic features on (pink) pearson correlation, (green) mean average error, (yellow)

mean squared error, and (blue) spearman correlation. (b) Pearson correlation of (blue) TAPIOCA network and (orange) bidirectional Long Short-Term memory network when excluding single epigenetic features . (c) Spearman correlation of TAPIOCA network predictions of (orange) transitional gamma (green) insulation score and (blue) directionality index when excluding single epigenetic features.

## 3.4 Discussion

We observe state of the art performance by TAPIOCA network on the well established metric of transitional gamma, indicating that the TAPIOCA approach should be considered when predicting TADs using epigenetic features. Furthermore TAPIOCA's high performance on Insulation score and its unique success in Directionality index demonstrate the power of the transformer approach to modeling the complex predictive relationship of epigenetic profile and chromatin topology.

TAPIOCA's ability to generalize across multiple cell lines indicates potential for real utility in saving the cost of Hi-C experiments, as this demonstrates that TAPIOCA can be used even without available Hi-C contact matrices from which to obtain labels. Future work could include examination of TAPIOCA's effectiveness across other model organisms beyond Drosophila. Such experiments may provide insight to similarity of the underlying biological mechanism of TAD formation in different organisms.

In our experiments where we removed individual features, we observed a different set of epigenetic features whose absence maximally degraded model performance when using the TAPIOCA network than when using the previously described BILSTM. This seems to indicate that those decreases in performance may be unelucidated consequences of the selected machine learning models, rather than

true biological relationships, raising uncertainty to the role of histone modifications H3K27Me3 and H3K27Ac in TAD formation as suggested in previous literature(Rozenwald et al., 2020). However, the epigenetic features where degradation of model performance was consistent between BILSTM and TAPIOCA, provide increased reason for hypothesizing an underlying relationship between TAD formation and presence of features such as Chriz.

We observed differing impacts on removing epigenetic features across different TAD characterization metrics. This disparity has multiple potential explanations and must be considered in conjunction with a few important observations. First, the explicit characterization of what makes a TAD is still an open area of discussion as multiple tools exist for TAD characterization and different tools often do not give fully concordant depictions(Zufferey et al., 2018). Second, the low performance of directionality index experiments in all epigenetic removal experiments is likely indicative of a failure of the model to capture the underlying data distribution, rather than anything grounded in biological reality. This claim is made in consideration with the demonstrated high hyperparameter sensitivity of the directionality dataset and the inability of BILSTM or regression variants to make successful predictions.

With these considerations, we do still observe differences in the contribution of removing single epigenetic features to successful prediction of insulation score and transitional gamma. One potential explanatory hypothesis for this disparity may be that different epigenetic features contribute to different scales or motifs of chromatin organization, some of which are more easily captured by armatus than insulation scores. Further work aiming to investigate this hypothesis may benefit from expanding

analysis even further to include some of the many other TAD characterization metrics outlined in Zufferey et al(Zufferey et al., 2018).

## 3.5 Conclusion

In this manuscript we present TAPIOCA, a tool for predicting TADs using epigenetic data via a self-attention based deep learning architecture. By reformulating the task of TAD prediction as a sequence transduction problem and developing an architecture inspired by the novel transformer network from machine learning literature we obtain state-of-the art results in inferring TAD characterization from epigenetic data. In addition to these results we contribute to the research community by expanding multiple Drosophila cell line datasets to include metrics for insulation score and directionality index.

## 3.6 Methods

### 3.6.1 Data Availability

All data is based on cell lines from the Drosophila model organism. We use three cell lines: Schneider-2 (S2) and Kc167 from late embryos and DmBG3-c2 (BG3) from the central nervous system. Epigenetic profiles and transitional gamma labels for all cell lines were found at https://github.com/MichalRozenwald/Hi-ChIP-ML. Hi-C data used to construct Insulation and Gamma labels is available on the Gene Expression Omnibus at GSE69013.

### 3.6.2 Code Availability

Cleaned datasets for all three metrics are available, along with all of our experiments at https://github.com/Max-Highsmith/TAPIOCA.

### 3.6.3 Model Hyper Parameter Tuning

In all hyper parameter tuning experiments, we used a random search over sets of discrete values for each parameter. Hyper parameters were determined separately for each network with each TAD label. With each network we tested batch sizes (1,4,16,64), learning rates (1e5, 1e4, 1e3, 1e2, 1e1). With BILSTM and TAPIOCA we varied dropouts (0,0.1, 0.2,0.3,0.5,0.7). And layer number (1,2,3,4,5,6). With TAPIOCA we varied the number of hidden units (512, 1024, 2048) and with BILSTM we varied bias existence for (True, False). All Models were initially trained with 10 iterations of random search for hyper parameters. Because none of these initial 10 results for the directionality dataset converged when using regression variants and BILSTM, we expanded the random search size to 20 but still did not obtain convergence on any network except TAPIOCA.

### 3.6.4 Model Architecture and Training Details

The TAPIOCA model is inspired by the transformer architecture (S.-W. Yang et al., 2020) (D. Liu et al., 2019). The Transformer was originally proposed for the task of seq2seq sentence translation and while our task is similar, there are a few key differences which inspired adjustments to the TAPIOCA architecture.

First, when working with seq2seq sentence translation, the fundamental unit of a sequence is a categorical token, typically a word. The preliminary step in translation

tasks is the conversion of tokens into numerical vector representations via embedding. In the task of TAD prediction we begin with normalized epigenetic feature vectors for each 20kb region instead of tokens. This removes the need for inclusion of an embedding step because we already have vector representations.

Secondly, because multihead attention does not use recurrence or convolution it permits increased ability to identify relationships between spatially distant features. While this characteristic is clearly advantageous, it also necessitates manual inclusion of positional information into propagated vectors. In the original transformer architecture this task is performed by adding a positional encoding vector to embedded inputs. In the original transformer the positional encoding vector is more information dense for certain vector components. The assumption is that because the embedding layer is high dimensional (512) that the necessary information will be passed, and multiple components can permit meaningful feature integration of position and embedding. However, in the TAD identification task we eschew embedding completely, instead using epigenetic feature vectors. Thus additive positional encoding would have potential to overwrite or give implicit preference to components which already encode specific information. To prevent this we instead concatenate the positional econding vector to the epigenetic features.

Third, Unlike seq2seq translation there is no variation in sentence length of inputs and outputs. When training we use a sentence length of 11 bins, (220kb region). We use the mean squared error of our full predicted vector and label vectors as a loss function. When evaluating performance on test data we pass each sequence through with stride 1, keeping the middle vector bin.

# 4. Four-Dimensional Chromosome Structure Prediction A probabilistic algorithm for the prediction of four-dimensional genome structure using time-series Hi-C data.

## 4.1 Abstract

Chromatin conformation plays an important role in a variety of genomic processes. Hi-C Data is frequently used to analyse structural features of chromatin such as AB compartments, topologically associated domains, and 3D structural models. Recently the genomics community has displayed growing interest in chromatin dynamics over time. Here we present 4DMax, a novel method which uses time-series Hi-C data to predict dynamic chromosome conformation. Using both synthetic data and real time-series Hi-C data from processes such as induced pluripotent stem cell reprogramming and cardiomyocyte differentiation, we construct smooth four dimensional models of individual chromosomes. These predicted 4D models effectively interpolate chromatin position across time, permitting prediction of unknown Hi-C contact maps at intermittent time points. Our results demonstrate that 4DMax correctly recovers higher order features of chromatin such as AB compartments and topologically associated domains, even at time points where Hi-C data is not made available to the algorithm. Use of 4DMax may alleviate the cost of expensive Hi-C experiments by interpolating intermediary timepoints while also providing valuable visualization of dynamic chromatin changes.

## 4.2 Introduction

The three-dimensional (3D) conformation of the genome has been shown to play an important role in a variety of genomic processes such as gene regulation(Dekker, 2008), gene replication(Wasim et al., 2021) and gene methylation(Buitrago et al., 2021). Various techniques have been developed for the analysis of three dimensional genome conformation, one of the most prominent being Hi-C, an improvement of the chromosome conformation captured (3C) technology. Hi-C data can be used to examine a plethora of higher order structural features such as: AB compartments (Fortin & Hansen, 2015), topological associated domains (TADs)(Zufferey et al., 2018) and 3D structural models(Oluwadare et al., 2019).

As genomic sequencing has become cheaper, more researchers have begun to generate time-series Hi-C data(Bertero et al., 2019),(Stadhouders et al., 2018). In such datasets Hi-C contact maps are obtained at multiple points in a time dependent genetic process. Some of these biological processes include:cardiomyocyte differentiation (Bertero et al., 2019) and induced stem-cell pluripotency(Stadhouders et al., 2018). While a plethora of meaningful and interesting observations have already been extracted from these datasets, analysis has been primarily constrained to comparing and contrasting individual points in the time series rather than the comprehensive analysis of four dimensional (4D) chromatin conformation changes over multiple time points (i.e. three dimensional conformation plus the 4th dimension of time). The need for novel 4D analysis has been identified as a critical and emerging area of research.

To address this need we introduce 4DMax, a maximum-likelihood based algorithm for predicting the transformation of chromatin conformation over the 4th dimension (time). By using spatial restraints derived from Hi-C contact matrices we provide a tool which permits the generation of a predictive 4D video of chromatin conformational changes throughout the time

series. 4DMax can be used to interpolate higher order chromatin features at times where no data is available while also providing valuable visualizations of chromosomal processes.

To date, only one other published computational method for modeling 4D transitions of chromatin exists, TADdyn(Di Stefano et al., 2020). Our 4DMax algorithm differs from TADdyn in 3 key ways.

Firstly, we utilize gradient descent optimization of a spatial restraint based maximum-likelihood function whereas the TADDyn approach utilizes polymer modeling and steered molecular dynamics followed by monte carlo based simulated annealing.

Secondly, TADyn focuses on small ~2MB segments of the genome with emphasis on transcriptional dynamics while our algorithm provides models of entire chromosomes. Our broader scope permits meaningful analysis of higher level structures such as TADs and AB compartments across time.

Thirdly, we demonstrate that 4DMax can use generated models as an interpolation mechanism for predicting AB compartments and TADS at time points for which no Hi-C data has been gathered.

The value of 4DMax is demonstrated through the construction of 4D models using contact maps derived from a mean-field simulated chromosomal looping process as well as multiple real time series Hi-C datasets. By studying the interrelation between contact maps we are capable of identifying meaningful characteristics of the genomic process unavailable from analysis of only individual timepoints. We successfully recover higher order conformational information such as AB compartments from the predicted 4D structure, even at time points where true Hi-C maps are intentionally excluded. Out of the box 4DMax can be easily inserted into any analytic pipeline focused on time-series Hi-C analysis.

## 4.3 Results

### 4.3.1 Overview of 4DMax approach.

In Figure 5.1 we outline the overall framework of 4DMax. First we gather intrachromosomal contact matrices from different time points in a genomic process. Next we convert contact matrices into spatial restraints using the relationship $D = IF^{\gamma}$ Where D is distance, IF is interaction frequency and $\gamma$ is a negative exponent. This inverse relationship with distance and interaction frequency is frequently used in 3D modeling literature(Oluwadare et al., 2019) . We then assign a parameter, granularity, to denote the number of temporal snapshots where the spatial position of our 4D model will be identified. Then, using a maximum likelihood approach from probability theory, we define a likelihood function which measures the agreement of our structure's position at each time point with temporally adjacent spatial restraints. We then initiate an unfolded structure and incrementally adjust its position to maximize our likelihood function using a gradient ascent algorithm. After training, a smooth 4d Model is created which can be visualized in movie format. From this 4DModel we extract synthetic Hi-C contact maps at time points of interest. We then use these extracted Hi-C contact maps for downstream Hi-C analysis such as AB compartment classification and topologically associated domain (TAD) identification.

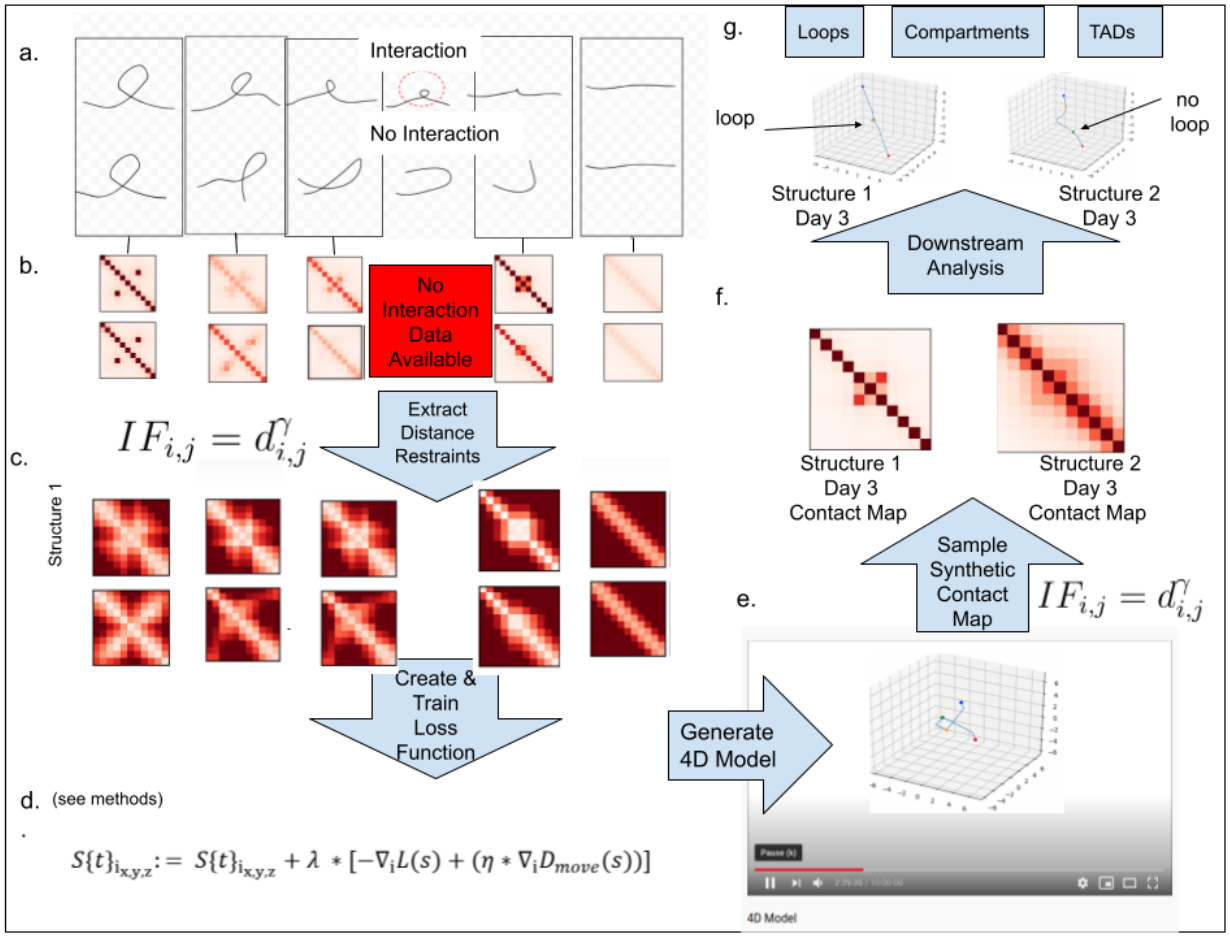



**Figure 4.1: Overview of 4DMax approach** Graphic elucidates the 4DMax workflow using a simplified synthetic dataset as illustration. (a) Drawings of two potential chromosomal trajectories from identical starting and ending conformations. A significant contact at the center exists in structure 1 but not structure 2. (b) Contact maps obtained through synthetic Hi-C experiments on each day in process. (c) Distance restraints derived from available contact maps. (d) Likelihood function for predicting 4D conformation. (e) Video of changing chromosome conformation. (f) Synthetic contact maps extracted at time of interest (g) Different 3D structural conformations on day 3.

### 4.3.2 4DMax correctly reconstructs models of synthetic time series Hi-C data.

We first created a simple, hypothetical chromosome and developed two theoretical structural progressions for the changing conformation of this chromosome. Both simulations are composed of 11 chromosomal bins and evolve over a 6 day process. Each 4D structure begins and ends identically, the initial chromatin state being in a looped formation and the final state being fully elongated. The two structures differ in their respective paths taken from their initial and final states. In structure 1 the loop unravels as if pulled on both ends while in structure 2 the loop swings open (Figure 4.1a). As a consequence of these differences in paths, on day 3, there is strong interaction between bins 4 and 6 on structure 1 but no such interaction exists on structure 2.

We first define contact maps for each of the 6 time points on both structures and use these contacts as inputs to 4DMax to generate novel 4D structures. We then simulate Hi-C experiments at the 6 time points using the generated structure and obtain contact maps with above .95 Pearson correlation (PCC) with corresponding input contact maps. Furthermore, visual inspection of the two generated videos accurately displays the unique behaviors of unraveling and swinging open previously described.

We test the effectiveness of 4DMax in capturing 4D movement and predicting 3D position at time points where contact map information is unavailable. We run four experiments for each synthetic structure excluding contact maps for days 1,2,3,4 respectively. The PCC values between original synthetic Hi-C maps and their corresponding interpolations remain high ranging from 0.82-0.99. Visually, we continue to observe the expected unraveling and swinging behaviors in each 4d video, even with excluded data (Supplementary Videos 1).

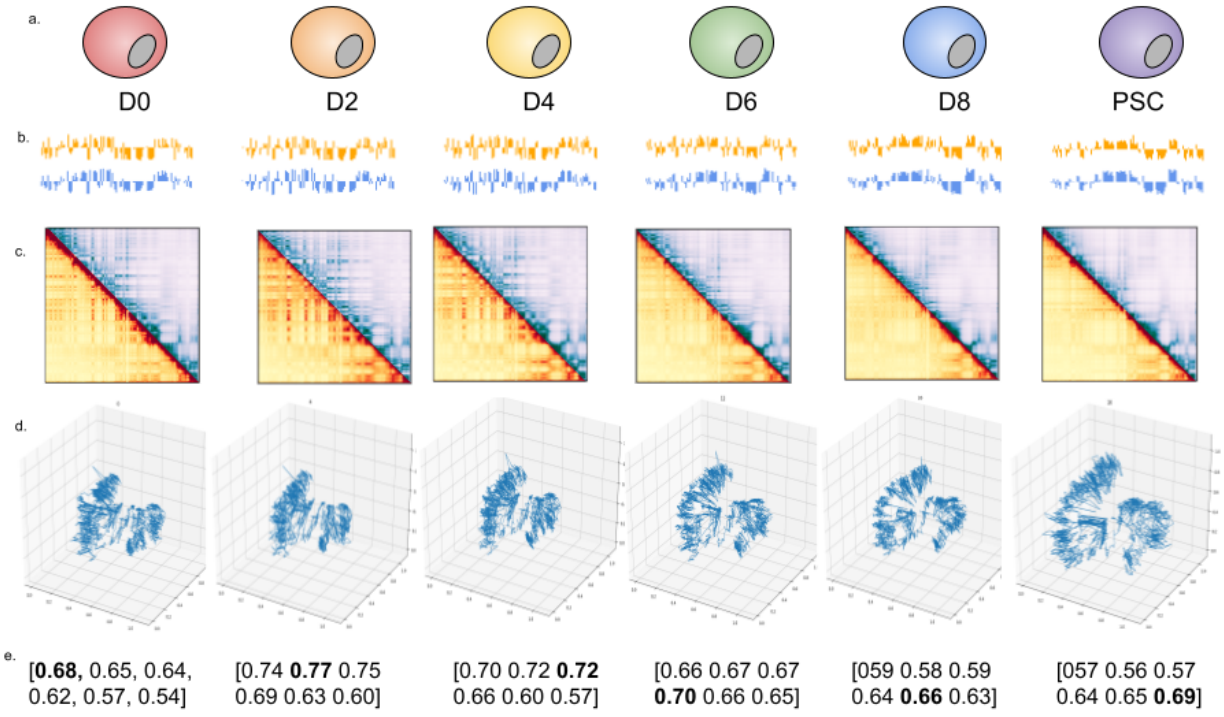



**Figure 4.2 Simulation of 4DMax Structures** Diagram of Outputs. (a) Outline of the different stages of the iPSC dataset. (b) Contact map of chromosome 13 by time, AB compartment vector shown above map. (c) 4DMax prediction of structural conformation of chromosome 13 at time. (d) Reconstructed contact map using simulated Hi-C of 4DMax structure, number below indicate spearman correlation between above reconstructed contact map and real contact maps at each time point.

### 4.3.3 4DMax predicts smooth 4D models of induced pluripotent stem cell differentiation in mice.

We apply 4DMax to a 10 day time series Hi-C dataset of induced stem cell pluripotency in mice . We use intrachromosomal Hi-C contact maps from day 0 (Beta), 2, 4, 6, 8 and 10 (PSC). We select a granularity of 21, ensuring that each time point for which real data is available occurs within the time interval partition. 4DMax successfully produces smoothly

changing structures for each chromosome (Figure 4.2c, Supplementary videos 2). We frequently observe a decrease in compression of 4D models as the induced pluripotency process progresses (Figure 4.2c, Supplementary videos 2). The 4DMax predictions for chromosomal position at the input times shows high similarity to 3D structures generated by previously built state of the art 3D modeling algorithms with average SPC=0.76 and PCC=0.75.

Using the 4DMax predictions we then simulated Hi-C experiments (Figure 4.2d) at each of the input time points to obtain synthetic Hi-C maps. We compare these synthetic maps to their corresponding real contact maps and observe high SPC values ranging (0.53-0.82). These values are consistently higher than the similarities seen between contact maps on Days 0 and 10 (0.46-0.68).

### 4.3.4 4DMax predicts smooth 4D models of cardiomyocyte differentiation in humans.

To verify the effectiveness of varied Hi-C datasets we also apply 4DMax to a 14 day time series Hi-C dataset of cardiomyocyte cell differentiation. The cardiomyocytes dataset contains Hi-C contact maps assayed at irregularly timed intervals on days: 0, 2, 5 and 14. We build 4DModels with a granularity of 15, ensuring that each time point for which real data is available occurs within the time interval partition., preserving the uneven timing of the contact maps. 4DMax again produces fluidly changing 4D models. (Supplementary Videos 3) We then simulate Hi-C experiments to obtain synthetic contact maps from the 4D model at the 4 input times and observe SPC values ranging from 0.54 to 0.92 between synthetic maps and their correspondingly timed real Hi-C data. We compare 4DMax reconstructed contact maps to real contact maps across all permutations of input times and observe SPC values are highest with corresponding times in 93.2% of the reconstructions, indicating the high correlation between real and reconstructed maps is significant relative to other Hi-C contact maps.

### 4.3.4 Interpolation of Time Series Hi-C Data using 4DMax generated models show high consistency with experimental Hi-C.

To evaluate the rigidity of 4DMax in its prediction of chromosomal position at timepoints between available contact maps we ran 4 experiments on each chromosome where we generated 4D models of the iPSC dataset while excluding Hi-C data for individual timepoints: D2, D4, D6, D8. We call these models the “iPSC Interp models”. The iPSC Interp models show high similarity to 4D models generated by the complete iPSC dataset (SPC>.99, in all chromosomes besides 1,4,5 PCC>.96), indicating the algorithm’s resilience to missing Hi-C data (Supplementary Videos 4). We then ran synthetic Hi-C experiments on the iPSC interp models at the time point for which their data was excluded to obtain interpolated contact maps. We compare these interpolated contact maps to corresponding real Hi-C contact maps and find high correlation with mean SPC=0.73 with values ranging from 0.62-0.80 (Figure 4.3a). In 24% of the experiments our interpolated contact maps show higher correlation to the real Hi-C contact maps than their biological replicate (Figure 4.3b). These results indicate that 4DMax is effective at predicting intermittent structures for time points where no Hi-C data is available.

We also perform interpolation experiments using the cardiomyocyte dataset where we exclude Hi-C input data on day 2. We refer to the resultant 4Dmodesl as the “Cardio Interp models”. The Cardio Interp models show high correlation to 4D models generated using the complete cardiomyocyte dataset (Supplementary Videos 5). We obtain synthetic Hi-C contact maps on day 2 from the Cardio Interp Model and compare these interpolation maps to the real day 2 Hi-C contact maps and find SPC values ranging from 0.57-0.87 (Figure 4.3c). In 6 of the chromosomes (28%) our interpolation shows higher correlation to the real Hi-C map than a

biological replicate (Figure 4.3d). These results indicate versatility in the time-series datasets for which our 4DMax algorithm can effectively interpolate Hi-C data.

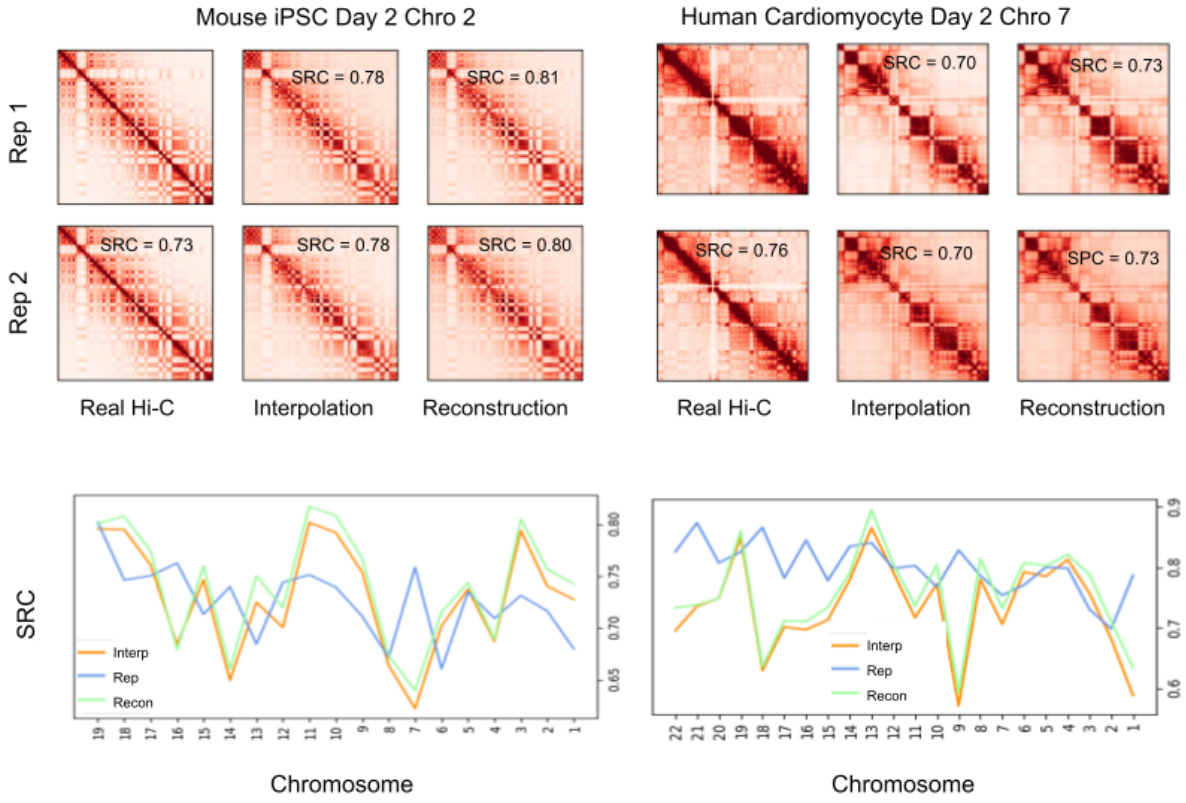



**Figure 4.3: 4DMax Predictions of Hi-C Contact Maps** Example contact map comparison. (a) Contact maps of iPSC on day 2 chromosome 2 from Real Hi-C, 4DMax reconstruction and 4DMax day 2 agnostic interpolation model. (b) SPC of iPSC contact maps relative to Real Hi-C for each chromosome on day 2. (c) Contact maps of cardiomyocyte data on day 2 chromosome 7 from Real Hi-C, 4DMax reconstruction and 4DMax day 2 agnostic interpolation model. (d) SPC of cardiomyocyte contact maps to Real Hi-C for each chromosome on day 2.

### 4.3.5 4DMax correctly preserves and predicts AB compartment assignment.

A primary value of Hi-C data is its utility in illuminating higher order structural features of chromatin (Lajoie et al., 2015)(Miura et al., 2018). One of the most prolific of these structural features are megabase scale subnuclear compartments called AB compartments. Regions of the genome are assigned to either compartment A or compartment B where the A compartment is associated with gene activity and euchromatin while the B compartment is associated with inactive heterochromatin. AB compartment assignment is known to change significantly during the iPSC process(Stadhouders et al., 2018). AB compartment assignment can be derived by principal component analysis (PCA) of Pearson correlation matrices derived from Hi-C contact maps (Methods). We first perform comparative AB compartment analysis on real Hi-C contact maps, and contact maps reconstructed from iPSC full models. We observe high visual similarity between Pearson correlation matrices of reconstructed and corresponding real Hi-C data across all chromosomes and timepoints (Figure 4.4a).

The iPSC dataset has previously been shown to undergo pronounced changes to compartmental organization as time progresses. Visually we observe high similarity between Reconstructed and Real AB compartment vectors at each point in the time series (Figure 4.4c). We quantify this progression by treating ab compartment vectors as input vectors to PCA to obtain trajectory curves for each chromosome (Figure 4.4b). The trajectories of real and reconstructed compartments match one another closely. These analyses indicate that the 4D models generated by 4DMax maintain the higher order information needed for AB compartment analysis.

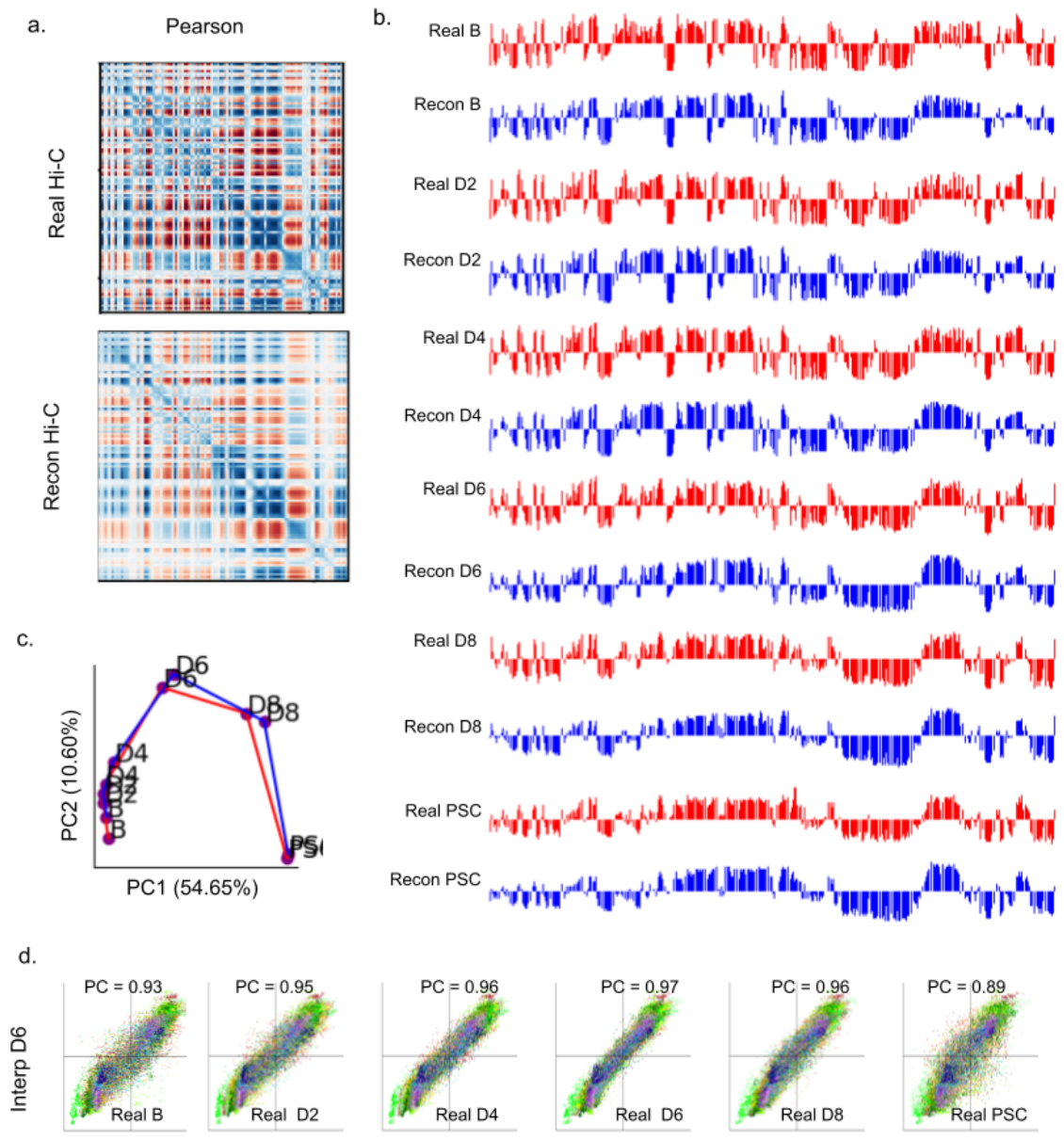



**Figure 4.4: 4DMax AB Compartment Analysis.** Analysis of AB compartment features of 4DMax generated contact maps. (a) Pearson correlation matrices of chromosome 14 day 2 using Real Hi-C and synthetic contact maps obtained from the 4DMax model. (b) AB compartment vectors from chromosome 14 (red) real Hi-C data (blue) synthetic contact maps obtained from 4DMax model. (c) Trajectory curve of two largest principal components (red) real

Hi-C (Blue) Reconstructed Hi-C. (d) Scatter plot of 100kb binned AB compartment vectors where x value is bins Real Data PC1 value and y value is interpolated Contact maps PC1 value.

We also compared the AB compartment profiles of our interpolated iPSC matrices to AB compartment profiles of real Hi-C contact maps (Figure 4.4d). In all 4 models we see PCC values greater than 0.95. Furthermore, when comparing interpolated AB compartment profiles to the AB compartment profiles of real Hi-C contact maps across all times in the iPSC process, we find the highest correlation at the interpolated timepoints (Figure 4.4d). For example, we built an interpolation model for 4D chromatin structure excluding contact information on day 6, instead only showing the algorithm contact information for days 0,2,4,8 and 10. 4DMax then made predictions for the chromosomal conformation on day 6. The output prediction for chromosomal conformations on day 6 were more similar to the real contact matrices on day 6 (0.97) than they were to any of the contact maps the algorithm was exposed to (D0= 0.93, D2= 0.95, D4=0.96, D10 = 0.88). This trend is consistent across all interpolation models and is crucial as it indicates that 4DMax is accurately predicting changes to AB compartment profiles, rather than simply obtaining high correlation due to maintained ab compartment profiles between adjacent timepoints.

### 4.3.6 4DMax correctly preserves and predicts TAD border positioning.

Another prolific use of Hi-C data is the identification of topologically associated domains (TADs). We used the Hi-C analysis tool HiCtool(Calandrelli et al., 2018) to identify TADs from contact maps in the iPSC dataset. We then use HiCtool to identify TADs with synthetic contact maps derived from 4DMax reconstruction and interpolation models (Figure 4.5ab). We observe high similarity in TAD profiles of reconstructed

synthetic maps and real Hi-C contact maps with a mean percent overlap of 85% and a peak of 99% on chromosome 9. We also observe high similarity in TAD profiles of interpolated synthetic maps and real Hi-C contact maps with a mean percent overlap of 84% and a peak of 97% on chromosome 11.

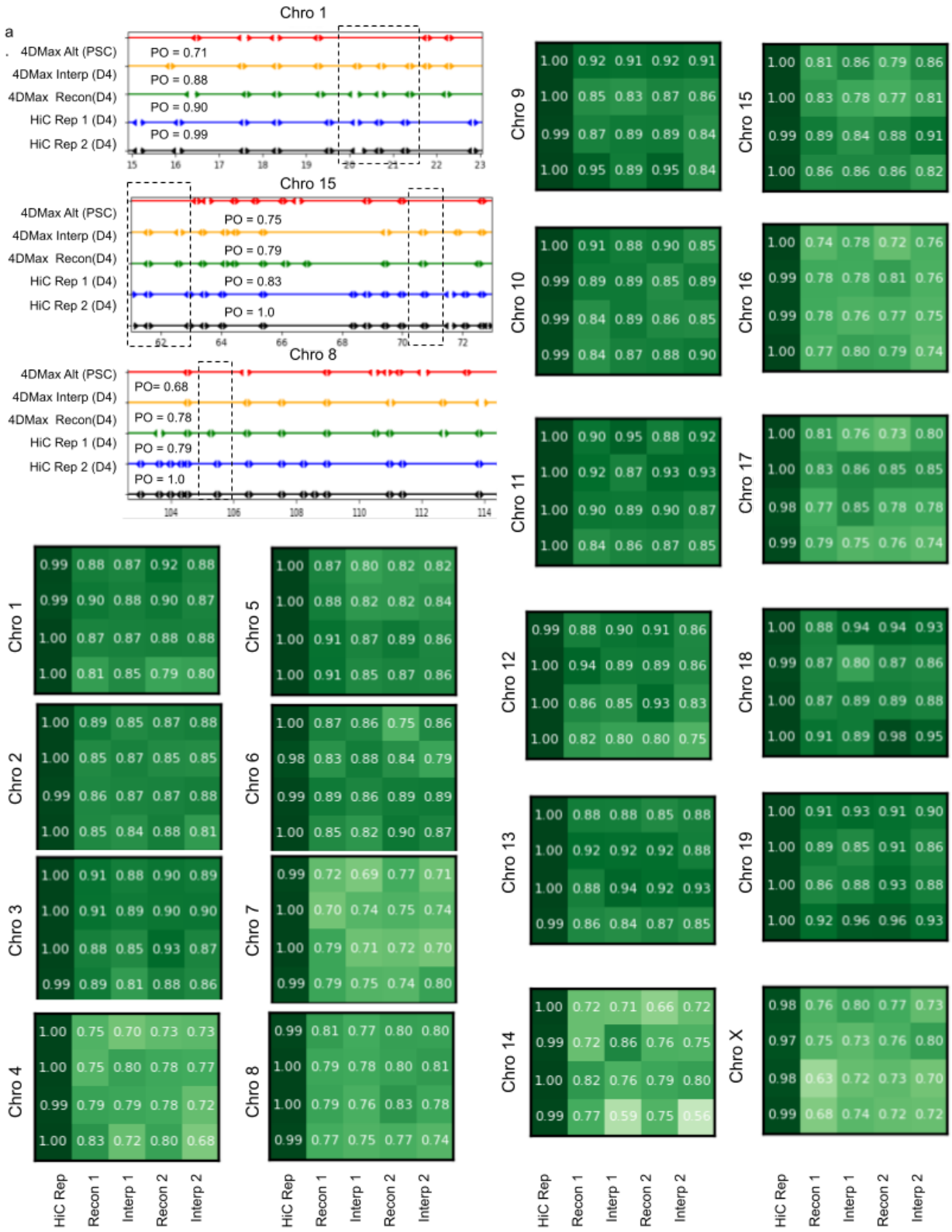

a
Chro 1
4DMax Alt (PSC)
4DMax Interp (D4)
4DMax Recon(D4)
HiC Rep 1 (D4)
HiC Rep 2 (D4)
PO = 0.71
PO = 0.88
PO = 0.90
PO = 0.99
15 16 17 18 19 20 21 22 23
Chro 15
PO = 0.75
PO = 0.79
PO = 0.83
PO = 1.0
62 64 66 68 70 72
Chro 8
PO= 0.68
PO = 0.78
PO = 0.79
PO = 1.0
104 106 108 110 112 114
Chro 1 0.99 0.88 0.87 0.92 0.88 / 0.99 0.90 0.88 0.90 0.87 / 1.00 0.87 0.87 0.88 0.88 / 1.00 0.81 0.85 0.79 0.80
Chro 5 1.00 0.87 0.80 0.82 0.82 / 1.00 0.88 0.82 0.82 0.84 / 1.00 0.91 0.87 0.89 0.86 / 1.00 0.91 0.85 0.87 0.86
Chro 2 1.00 0.89 0.85 0.87 0.88 / 1.00 0.85 0.87 0.85 0.85 / 0.99 0.86 0.87 0.87 0.88 / 1.00 0.85 0.84 0.88 0.81
Chro 6 1.00 0.87 0.86 0.75 0.86 / 0.98 0.83 0.88 0.84 0.79 / 0.99 0.89 0.86 0.89 0.89 / 1.00 0.85 0.82 0.90 0.87
Chro 3 1.00 0.91 0.88 0.90 0.89 / 1.00 0.91 0.89 0.90 0.90 / 1.00 0.88 0.85 0.93 0.87 / 0.99 0.89 0.81 0.88 0.86
Chro 7 0.99 0.72 0.69 0.77 0.71 / 1.00 0.70 0.74 0.75 0.74 / 1.00 0.79 0.71 0.72 0.70 / 0.99 0.79 0.75 0.74 0.80
Chro 4 1.00 0.75 0.70 0.73 0.73 / 1.00 0.75 0.80 0.78 0.77 / 0.99 0.79 0.79 0.78 0.72 / 1.00 0.83 0.72 0.80 0.68
Chro 8 0.99 0.81 0.77 0.80 0.80 / 1.00 0.79 0.78 0.80 0.81 / 1.00 0.79 0.76 0.83 0.78 / 0.99 0.77 0.75 0.77 0.74
HiC Rep Recon 1 Interp 1 Recon 2 Interp 2
Chro 9 1.00 0.92 0.91 0.92 0.91 / 1.00 0.85 0.83 0.87 0.86 / 0.99 0.87 0.89 0.89 0.84 / 1.00 0.95 0.89 0.95 0.84
Chro 15 1.00 0.81 0.86 0.79 0.86 / 1.00 0.83 0.78 0.77 0.81 / 0.99 0.89 0.84 0.88 0.91 / 1.00 0.86 0.86 0.86 0.82
Chro 10 1.00 0.91 0.88 0.90 0.85 / 0.99 0.89 0.89 0.85 0.89 / 0.99 0.84 0.89 0.86 0.85 / 0.99 0.84 0.87 0.88 0.90
Chro 16 1.00 0.74 0.78 0.72 0.76 / 0.99 0.78 0.78 0.81 0.76 / 0.99 0.78 0.76 0.77 0.75 / 1.00 0.77 0.80 0.79 0.74
Chro 11 1.00 0.90 0.95 0.88 0.92 / 1.00 0.92 0.87 0.93 0.93 / 1.00 0.90 0.89 0.90 0.87 / 1.00 0.84 0.86 0.87 0.85
Chro 17 1.00 0.81 0.76 0.73 0.80 / 1.00 0.83 0.86 0.85 0.85 / 0.98 0.77 0.85 0.78 0.78 / 0.99 0.79 0.75 0.76 0.74
Chro 12 0.99 0.88 0.90 0.91 0.86 / 1.00 0.94 0.89 0.89 0.86 / 1.00 0.86 0.85 0.93 0.83 / 1.00 0.82 0.80 0.80 0.75
Chro 18 1.00 0.88 0.94 0.94 0.93 / 0.99 0.87 0.80 0.87 0.86 / 1.00 0.87 0.89 0.89 0.88 / 1.00 0.91 0.89 0.98 0.95
Chro 13 1.00 0.88 0.88 0.85 0.88 / 1.00 0.92 0.92 0.92 0.88 / 1.00 0.88 0.94 0.92 0.93 / 0.99 0.86 0.84 0.87 0.85
Chro 19 1.00 0.91 0.93 0.91 0.90 / 1.00 0.89 0.85 0.91 0.86 / 1.00 0.86 0.88 0.93 0.88 / 1.00 0.92 0.96 0.96 0.93
Chro 14 1.00 0.72 0.71 0.66 0.72 / 0.99 0.72 0.86 0.76 0.75 / 1.00 0.82 0.76 0.79 0.80 / 0.99 0.77 0.59 0.75 0.56
Chro X 0.98 0.76 0.80 0.77 0.73 / 0.97 0.75 0.73 0.76 0.80 / 0.98 0.63 0.72 0.73 0.70 / 0.99 0.68 0.74 0.72 0.72
HiC Rep Recon 1 Interp 1 Recon 2 Interp 2

**Figure 4.5: 4DMax Topologically Associated Domain Analysis** HiCtool identified topologically associated domains. (a) Select images of TAD boundaries on (black) Real Hi-C replicate 1, (blue) Real Hi-C replicate 2, (green) 4DMax Reconstructed Map, (orange) 4DMax Interpolated Hi-C Map and 4DMax Recon Map at a different time point. PO metric quantifies the percent of TAD boundaries found within 0.5Mb of a boundary identified in Hi-C rep1. (b) PO of Interpolated and Reconstructed 4DMax TAD positions for both replicates across all chromosomes.

### 4.3.7 4DMax completes in tractable time for human and mouse chromosome construction.

4DModel generation time is determined by three parameters: training epochs, granularity and bin quantity. Run time scales linearly with the number of training epochs. Empirically we observe 400 epochs as sufficient to obtain organized and consistent conformational changes in 4D models for both datasets (Supplementary Videos 6). Granularity, defined as the number of tracked discrete time points in the interval, also impacts run time linearly. Bin quantity, defined as the number of discrete spatial points tracked per time point, is dependent on chromosomal length and resolution. We observe super linear growth of time as bin quantity increases (Figure 5.6a). Using a single GTX 1080 Ti graphics card 4DMax constructed 4DModels of 500kb resolution chromosomes in a matter of minutes and took under 1.5 hours to generate models from 50kb resolution chromosomes.

## 4.3.8 4DMax predictions remain stable against change in time point granularity.

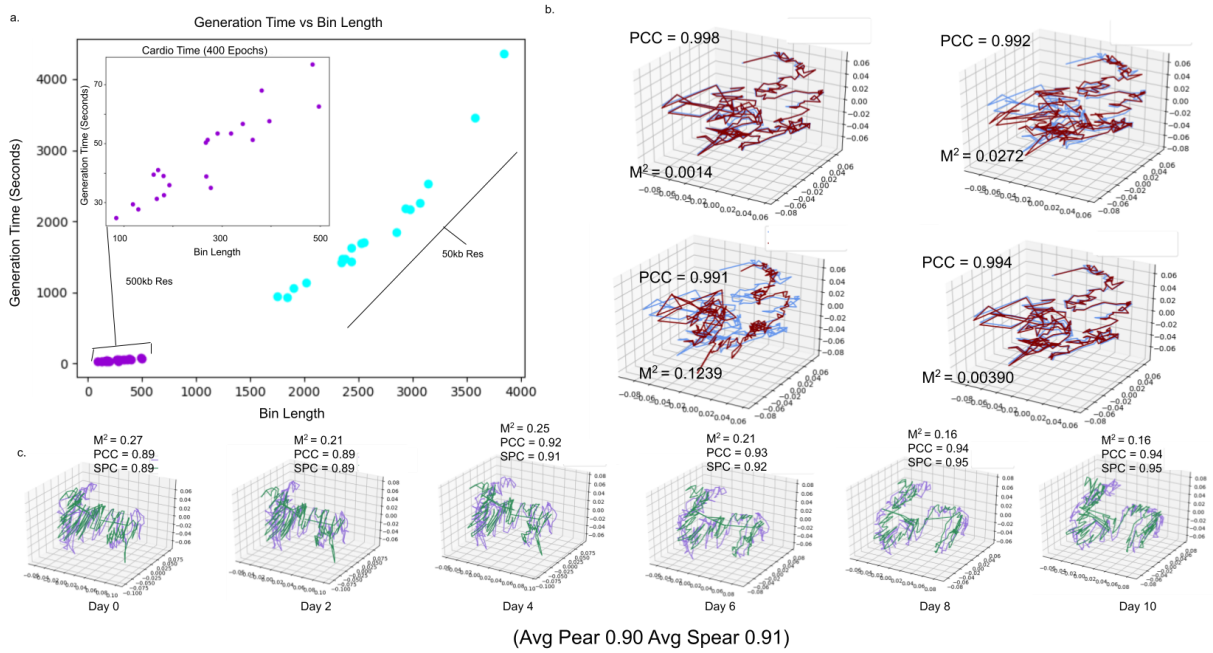



**Figure 4.6: 4DMax Computational Evaluation** Evaluation of runtimes and computational stability. (a) Scatter plot of chromosome bin lengths and time to completion using 400 epoch (purple) 500kb resolution chromosomes and (blue) 50 kb resolution chromosomes. (b) 3D plot of predicted cardiomyocyte chromosome 10 on day 5 with varying granularity values. Spearman correlation Mean Squared distance compares (blue) granularity 15 structure to higher granularity structures (red). (c) 3D plots comparing (purple) 50 kb resolution chromosome 1 to (green) 500kb resolution iPSC chromosome 1 on each day in time series.

We compared the 4D structures of the same chromosomes in the cardiomyocyte dataset with varying granularity of: 15, 29, 43, 57, 71. For each granularity comparison we used the average correlation between timepoints present in both structures (Figure 4.6b, Supplementary videos 7). We see minimal discrepancies between our maximal

and minimal granularity values with the average correlation (PCC=0.90; SPC=0.94) reaching as high as (PCC=0.94, SPC=0.99) on chromosome 9. These results indicate stability to changes in granularity.

### 4.3.9 4DMax predictions remain stable to variation in Hi-C contact matrix resolution.

We compared the 4D structures of the same chromosomes in the iPSC dataset at 500kb and 50kb resolution. The structures from 50kb resolution were reduced to 500kb by averaging the position of every 10 consecutive spatial points (Figure 4.6c). The average correlation between structures remains high (SPC=0.84, PCC=0.83) reaching (SPC=0.96, PCC=0.97 ) for chromosome 8. This indicates consistent 4DModel predictions across varying resolutions.

## 4.4 Discussion

Here we present 4DMax, a method used to examine time-dependent dynamics of chromatin during genomic processes. 4DMax is the second published tool to simulate structural changes to chromatin over time and is the first of its kind to provide comprehensive chromosome wide predictions of 4D dynamics. By converting contact maps at select times into spatial restraints, using these restraints to build a likelihood based objective function, and optimization with gradient ascent, 4DMax constructs smooth 4DModels.

We validate the effectiveness of 4DMax in predicting 4D conformations using both synthetic chromosomal conformations and real time-series Hi-C datasets from

mice and humans. From these visualizations we often observe pronounced changes to the positioning of chromosomes over time such as the progressive decompression of mice chromosomes as they become pluripotent. From our 4D models we visually observe the preservation of preferentially interacting regions such of TADs, providing valuable visual representations of how such TADS are actually positioned within a global chromosomal context.

In addition to the valuable visualizations, 4DMax accurately predicts chromosome position at timepoints where data is excluded from the 4DMax algorithm. The interpolated maps from 4DMax frequently show higher similarity to true contact maps at their corresponding time than to true contact maps at adjacent times presented to the model. This is particularly promising because it indicates the high similarity of retrieved biological features is not a product of low chromosomal structural change in temporal segments of the time series, but rather that novel inferences are being made to the actual position of the chromosome at times where no hi-c data is available. Given these findings 4DMax could be used by other labs as a preliminary substitute for expensive Hi-C experiments when examining a genomic process over time.

4DMax is easily integrated into any time series Hi-C pipeline. Our model stability experiments show computational stability to variation of parameters such as contact map resolution and granularity while maintaining a sufficiently short run time. The structures generated by 4DMax show high correlation to input contact matrices and the synthetic contact maps derived from predicted 4DMax structures frequently have high correlation with real contact maps, even at times where no contact map data is

presented to the model. 4DMax derived contact maps retain biologically relevant higher order features such as AB compartment and TAD placement.

Despite these promising results the time scale of all real Hi-C datasets tested is in the order of days, therefore it is possible that significant changes to chromosome conformation may occur at smaller intervals not captured by existing data. To address this concern in the future 4DMax will have to be applied to future time series Hi-C datasets with smaller time intervals and additional assays for validation of conformation such as Capture Hi-C and microscopy data.

## 4.5 Methods

### 4.5.1 Description of 4DMax algorithm

4DMax is intended for researchers interested in inspecting the changing structural conformation of the genome over the duration of a dynamic biological process. We assume that the biological process occurs over a time interval $I = (0, T)$, where $0$ is the start of the process and $T$ is the last point in the process. We represent a chromosomes movement over a time interval as a collection of $n$ points in 4D space. Let $S$ be a 4D chromosome structure. $S = S\{t\}t \in I$, where $S\{t\} = \{S_i\{t\}\}$and $S_i\{t\} \in R^3 x I$, where $S_i\{t\}$ denotes the $x$, $y$, $z$ coordinate of the $i^{th}$ bin of a chromosome at time *t*.

We view the chromosome structure $S$ as a structure in 4-dimensional space (3 spatial, 1 time), denoted $S \in (R^3 X I)$. We use $S\{t\}$ to denote the structure's spatial position at a given time.

### 4.5.2 Maximum likelihood

We use a likelihood function as a loss function to compute chromosome conformation from the contact maps. Let $H = \{H_\tau\}$ be a collection of Hi-C contact maps where $H_\tau$ is the Hi-C contact map at time $\tau$. The likelihood of a structural conformation S can be modeled as the product of the probability of the observed HiC contact maps $H$, conditioned on $S\{t\}$.

$$L(S) = \prod_t P(H|S\{t\}) \qquad \text{(Eq 4.1)}$$

$P(H|S\{t\})$ can be modeled as the product of individual distances in $H$ conditioned on S by assuming each constraint is independent. By assuming that each constraint $H_i \in H$ is conditionally independent of other constraints we rewrite the likelihood as

$$L(S) = \prod_t \prod_i P(H_i|S\{t\}) \qquad \text{(Eq 4.2)}$$

Because our Hi-C samples were taken during some point during the biological process being observed, we know $\mathbb{T} \in I$, however, if we select a high granularity of $I$ there are certain $t \in I$ such that $t \notin \mathbb{T}$. Thus we can separate our $L(S)$ by

$$L(S) = \prod_{t\in\mathbb{T}} \prod_i P(H_i|S\{t\}) * \prod_{t\notin\mathbb{T}} \prod_i P(H_i|S\{t\}) \qquad \text{(Eq 4.3)}$$

Because the logarithmic function is monotone we can take the logarithm of $L(S)$ without the argmax changing, yielding

$$L(S) = \sum_{t \in \mathbb{T}} \log(\prod_i P(H_i|S\{t\})) + \sum_{t \notin \mathbb{T}} log(\prod_i (P(H_i|S\{t\})) \quad \text{(Eq 4.4)}$$

**When** $t \in \mathbb{T}$ we assume that observed contact maps are drawn from a gaussian probability distribution

$$P(H_i|S\{t\}) = P(H_i\{t\}|S\{t\}) \sim \frac{1}{\sigma\sqrt{2\pi}} exp(\frac{-1}{2\sigma^2}(D_i\{t\} - H_i\{t\})^2) \quad \text{(Eq 4.5)}$$

where $D_i\{t\}$ is the actual euclidean distance between the pair of regions index by $i$, computed from $(x, y, z)$ coordinates $\sigma$ is the standard deviation of the gaussian distribution. By assumption of normal distribution we know

$$\sigma = \sqrt{\frac{\sum_i (H_i - D_i)^2}{n}} \quad \text{(Eq 4.6)}$$

Using algebra we can manipulate equation (5) to resemble a component of the first right hand summation term in equation (4) as shown in (7).

$$\prod_i P(H_i|S) = (\frac{1}{\sigma\sqrt{2\pi}})^n * exp(\frac{-1}{2\sigma^2} \sum_i^n (D_i\{t\} - H_i\{t\})^2) \quad \text{(Eq 4.7)}$$

Thus, by taking the logarithm of both sides of (7) we obtain

$$log(\prod_i P(H_i|S)) = -\frac{\sum_i (H_i\{t\} - D_i\{t\})^2}{2\sigma^2} - nlog(\sigma) \quad \text{(Eq 4.8)}$$

We can substitute equation (6) into equation (6) to remove all dependence on $\sigma$ and obtain

$$\log(\prod_i P(H_i|S)) = \frac{n}{2} - \log(\sqrt{\frac{\sum (H_i\{t\} - D_i\{t\})^2}{n}}) \quad \text{(Eq 4.9)}$$

**When** $t \notin \mathbb{T}$ we define

$$a_1(t) = \text{max } \tau \in S_\tau \text{ where } \tau < t$$

$$a_2(t) = \text{min } \tau \in S_\tau \text{ where } \tau > t$$

$$w_1(t) = \frac{t - a_2(t)}{|a_1(t) - a_2(t)|}, \;\; w_2(t) = 1 - w_1(t) \quad \text{(Eq 4.10)}$$

And assume

$$log(\prod_i P(H_i|S\{t\} = w_1(t)log(\prod_i P(a_1(t)|S\{t\})) + w_2(t)log(\prod_i P(a_2(t)|S\{t\}))) \quad \text{(Eq 4.11)}$$

.

Using equation (4.7) and (4.11) we obtain

$$log(\prod_i P(H_i|S\{t\}) =$$

$$w_1(t)[\frac{n}{2} - log(\sqrt{\frac{\sum(a_1(t)_i - D_i(t))^2}{n}})] + w_2(t)[\frac{n}{2} - log(\sqrt{\frac{\sum(a_2(t)_i - D_i(t))^2}{n}})]) \quad \text{(Eq 5.12)}$$

Thus by substituting equation (Eq 4.8) and (Eq 4.12) into the first and second terms of (Eq 4.4) we obtain a well defined likelihood function whose variables are either fixed values such as Hi-C contact constraints or functions of our structures coordinates

$$L(S) = \sum_{t \in \mathbb{T}} \frac{n}{2} - log(\sqrt{\frac{\sum_i (H_i\{t\} - D_i\{t\})^2}{n}}) + \quad \text{(Eq 4.13)}$$

$$\sum_{t \notin \mathbb{T}} w_1(t)[\frac{n}{2} - log(\sqrt{\frac{\sum(a_1(t)_i - D_i(t))^2}{n}})] + w_2(t)[\frac{n}{2} - log(\sqrt{\frac{\sum(a_2(t)_i - D_i(t))^2}{n}})])$$

Finally, to create a loss function we simply take the negation of our likelihood, so that our loss will maximize likelihood.

$$D_{likelihood}(S) = -L(S) \quad \text{(Eq 4.14)}$$

### 4.5.2 Distance Function:

Because the purpose of 4DMax is to represent structural changes in time as a continuous evolution, rather than provide individual snapshot images, it is important the motion between frames be smooth. To help ensure this we include a penalizing term, distance loss

$$D_{movement}(s) = \sum_t \delta(s(t))$$

$$= \sum_t \sum_i \delta(s_i\{t\}, s_i\{t+1\})$$

$$= \sum_t \sum_i \sqrt{\sum_c (s_{i,c}\{t\} - s_{i,c}\{t+1\})^2} \quad c \in \{x, y, z\}. \qquad \text{(Eq 4.15)}$$

This minimizes jumps between frames and results in more continuous structures.

### 4.5.3 Optimization

We optimize our structures coordinates by constructing a linear combination of our distance loss function and likelihood-loss function and incrementally adjusting via gradient descent, yielding

$$S\{t\}_{i_{x,y,z}} := S\{t\}_{i_{x,y,z}} \text{-} \lambda * [-\nabla_i L(S) + (\eta * \nabla_i D_{move}(S)))] \qquad \text{(Eq 4.16)}$$

Where $\eta$ is a weighing constant and $\lambda$ is the learning rate.

Unless stated otherwise we run experiments with $\eta = 1000$ and $\lambda = 0.0001$.

### 4.5.4 Interpolation of Contacts

We interpolate contacts by first running 4DMax excluding input Hi-C maps at time $t$ of interest. From the 4D model we extract the predicted 3D structure at time $t$. Using

this 3D mode we assume an inverse relationship between spatial distance and contact frequency as with map generation equation $IF_{i,j} = d_{i,j}^{\gamma}$ We use $\gamma$ values of -1.

### 4.5.5 AB Compartment Analysis

AB compartments are identified using the procedure outlined in ("Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome," 2010). We first obtain observed over expected (O/E) matrices for contact maps, where expected values are the mean contact frequency between bins of a given distance. From O/E matrices we treat rows as vectors and obtain Pearson correlation matrices. From the correlation matrices we perform principal component analysis (PCA). We assign compartments to each bin based on the sign of its corresponding row's PC1 value. Trajectories are obtained by performing PCA on AB compartment sign assignment vectors. Scatterplots are obtained by mapping pc1 values between two corresponding AB profiles as (x,y) coordinates.

### 4.5.6 TAD Identification

TADs were identified using the directionality index approach as implemented by HiCtool(Calandrelli et al., 2018). This procedure begins with the identification of a statistic called the directionality index on each genomic using equation 17.

$$DI = \frac{B-A}{|B-A|}\left(\frac{(A-E)^2}{E} + \frac{(B-E)^2}{E}\right) \quad E = \frac{A+B}{2} \qquad \text{(Eq 4.17)}$$

Where $A$ is the number of reads that map from a bin to 2Mb upstream and B is the number of reads that map from the same bin to 2Mb bins downstream. Using the directionality index a hidden markov model (HMM) is used to identify biased states via the viterbi algorithm. From the HMM emissions TAD coordinates are derived as consecutive downstream bias states . We compare TAD profiles of different contact maps using percent overlap (PO). We consider a TAD boundary from one profile overlapping if it occurs within .5Mb of a same direction TAD boundary on the compared profile. For more details on the directionality index computation and motivation see the methods original paper by Dixon et al (Dixon et al., 2012).

### 4.5.7 Statistical analysis

Disparity ($M^2$), Pearson correlation coefficient (PCC) and Spearman correlation coefficient (SPC) were used for evaluating similarity of contact matrices and distance vectors of 3D structures. Comparison between 4D structures were based on average correlation between 3D structures at each corresponding time point. Comparison is computed by first aligning the two structures around the origin and normalizing the matrix representation of their bins 3D position (A) such that $trace(AA^T) = 1$ using the scipy.spatial.procrustes method(Garreta & Moncecchi, 2013). After this normalization we compute disparity

$$M^2 = \sum_{i \in bins} (S_i^1\{t\} - S_i^2\{t\})^2 \qquad S_i\{t\} \tag{4.18}$$

Where $S_i^j\{t\}$ is the spatial coordinate of bin $i$ in structure $j$ at time $t$.

### 4.5.8 Data availability

All Hi-C data were downloaded from the Gene Expression Omnibus (GEO). Cardiomyocyte data was found at accession number GSE106690 (Bertero et al., 2019) and induced pluripotency data was found at the accession number GSE96611 (Stadhouders et al., 2018) .

### 4.5.9 Code Availability

4DMax was built using python.  4DMax as well as the software for all enclosed experiments is available at https://github.com/Max-Highsmith/4DMax.

# 5. Genome Structure Database (GSDB)

### 5.1 Maximum likelihoodAbstract

The proliferation of Hi-C technology has led to the implementation of a variety of approaches for predicting 3D structural conformations of chromosomes. These approaches can vary greatly in the guiding principles of their underlying methods and may lead to varied results. In this chapter, we describe the construction of the Genome Structure Database, a web portal for visualizing over 50,000 structures of chromosomes generated using 12 selected tools prolific in the literature. This database can be used to inspect hypothesised conformations across multiple cell lines as well as to compare structures predicted using existing tools to alternative structures proposed by newly developed algorithms. The database serves the research community by providing a centralized location for obtaining Hi-C contact maps, comparative visualizations and 3D models. GSDB can be accessed at http://sysbio.rnet.missouri.edu/3dgenome/GSDB.

### 5.2 Introduction

The 3D organization of genomes have been shown to play an important role in a variety of biological functions and processes such as gene regulation and gene expression. Consequently comprehensive understanding of gene dynamics requires increased inspection of the 3D conformation of the human genome. Multiple approaches to this task have developed including Fluorescent in-situ hybridization FISH(Cui et al., 2016) stimulated emission depletion (STED) (van der Wee et al., 2021) stochastic optical reconstruction microscopy (STORM)(Huang et al., 2010) and

photo-activated localization microscopy (PALM)(Betzig et al., 2006) . While these techniques have provided valuable insights, they are limited by their abilities to depict global, high resolution views of the genomes organization.

Because of these limits many research groups have moved to microbiology based 3C, technologies to examine the genomes structures, specifically Hi-C. Hi-C experiments can be used to create contact matrices, identifying frequency of interaction between different genomic bins along a partition of a chromosomal body. These contact maps are used by a variety of algorithms to create estimates for genomic structure. Broadly speaking such algorithms fall into one of three categories: distance-based, contact based, and probability based algorithms(Oluwadare et al., 2019). Distance based algorithms involve two steps: first the conversion of contact matrices into distance maps, and then an optimization function which creates a structure with similar predictive distances. Contact based algorithms treat observed contacts as constraints and use optimization functions to minimize the violation of such restraints. Probability algorithms formulate a probabilistic measure over the interaction frequency map and use probabilistic inference such as maximum a posteriori, monte carlo or maximum likelihood to choose from potential structures.

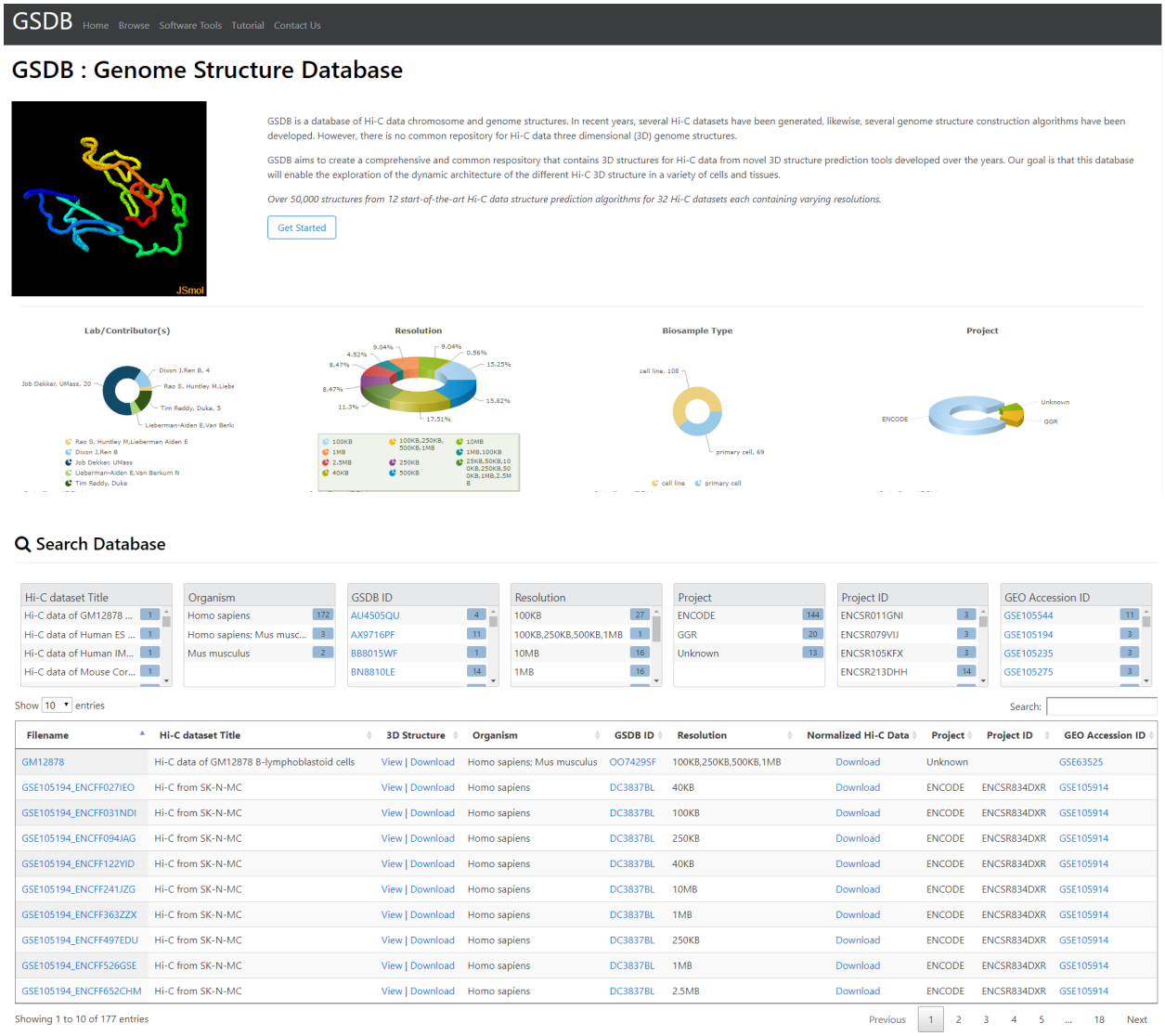



**Figure 5.1 GSDB HomePage** (top) home page of GSDB with data statistics (bottom) selection menu for viewing Hi-C data and 3D structures.

The GSDB pulls a total of 12 algorithms from each of these classes of reconstruction methods and applies them to 32 distinct Hi-C datasets at a variety of different resolutions ranging from 10kb to 2.5mb and generates a total of over 50,000 genomic structures (Figure 5.1). It provides multiple mechanisms for viewing these structures as well as a variety of comparative metrics, metadata and supplementary visualizations (Figure 5.2).

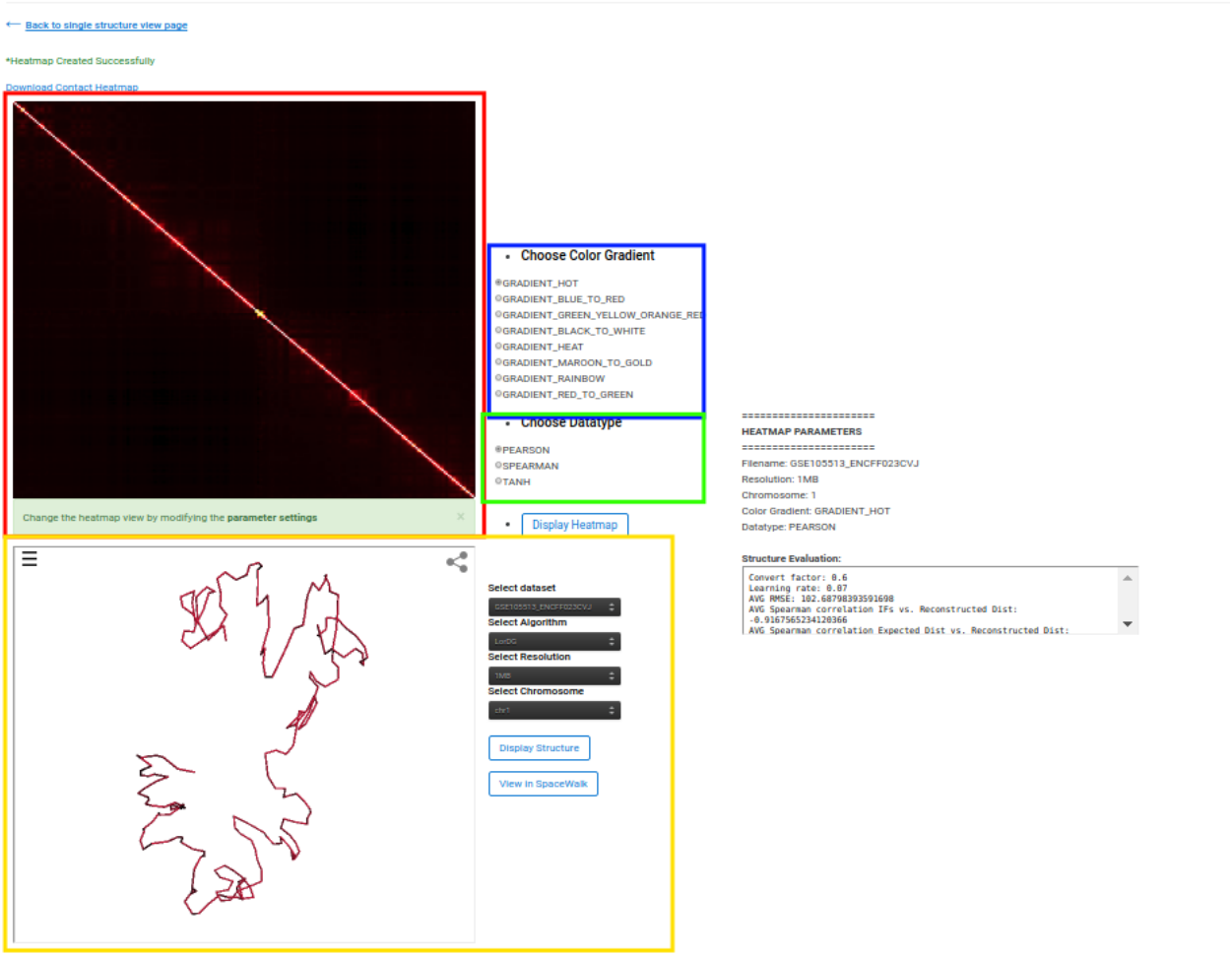



**Figure 5.2 Viewing window in GDSB.** (red) heat map of selected genome. (blue) color and (green) correlation settings for visualization of heatmap. (yellow) identifies 3D structure and options for dataset, resolution, chromosome, and algorithm to use.

### *5.2.1 Datasets*

Hi-C data is pulled from a variety of sources including the Gene Expression Omnibus (GEO) database, Accession numbers used include: GSE63525(Rao et al., 2014), GSE35156(Dixon et al., 2012) and GSE 18199(Lieberman-Aiden et al., 2009). Data is

also pulled from the Encode project(Consortium & The ENCODE Project Consortium, 2004).

### *5.2.2 Algorithms*

The GSDB uses twelve algorithms to construct models of the 3D genome. Broadly speaking these algorithms fall in the category of distance-based, contact-based, and probability-based algorithms. Our distance based algorithms include: LorDG(Trieu & Cheng, 2017), 3DMax(Oluwadare et al., 2018), Chromosome3d(Adhikari et al., 2016), HSA(Zou et al., 2016), ChromSDE(Z. Zhang et al., 2013), Shrec3D(J. Fraser et al., 2009) and InfoMod3Dgen(Wang et al., 2015). Our contact based algorithms include: MOGEN (Trieu & Cheng, 2016; Wang et al., 2015), GEM(Zhu et al., 2018) . Our Probability based algorithms are Pastis(Varoquaux et al., 2014) and SIMBA 3D(Rosenthal et al., 2019). As a brief overview, LorDG uses a nonlinear Lorentzian function as the objective function with the main objective of maximizing the satisfaction of realistic restraints rather than outliers. LorDG uses a gradient ascent algorithm to optimize the objective function. 3DMax used a maximum likelihood approach to infer the 3D structures of a chromosome from Hi-C data. A log-likelihood was defined over the objective function which was maximized through a stochastic gradient ascent algorithm with per-parameter learning rate. Chromosome3D uses distance geometry simulated annealing (*DGSA*) to construct chromosome 3D structure by translating the distance to positions of the points representing loci. HSA introduced an algorithm capable of taking multiple contact matrices as input to improve performance. HSA can generate the same structure irrespective of the restriction enzyme used in the Hi-C experiment.

ChromSDE (Chromosome Semidefinite Embedding) framed the 3D structure reconstruction problem as a semi-definite programming problem. Shrec3D formulated the 3D structure reconstruction problem *as* a graph problem and attempted to find the shortest-path distance between two nodes on the graph. The length of a link is determined as the inverse contact frequency between its end nodes. Each fragment is regarded as the nodes connected by a link. The represented 3D structure for a Hi-C data is one in which the distance between the nodes is the shortest. InfoMod3DGen converts the IF to a distance matrix and uses an expectation-maximization (EM) based algorithm to infer the 3D structure. In the contact-based category, we used MOGEN and GEM for the 3D structure reconstruction. MOGEN does not require the conversion of IF to distances and is suitable for large-scale genome structure modeling. GEM considers both Hi-C data and conformational energy derived from knowledge about biophysical models for 3D structure modeling. It used a manifold learning framework, which is aimed at extracting information embedded within a high-dimensional space, in this case the Hi-C data. Lastly, in the probability-based category, Pastis defined a probabilistic model of IF and casted the 3D inference problem as a maximum likelihood problem. It defined a Poisson model to fit contact data and used an optimization algorithm to solve it. SIMDA3D used a Bayesian approach to infer 3D structures of chromosomes from single cell Hi-C data.

## 5.3 Expansions

### *5.3.1 Expanded Viewing Capabilities*

In the previous version of GSDB, 3D models were displayed in Protein data bank (*.pdb) format. This format, intended for the display of protein molecules, has been the de facto standard for storing three-dimensional models of chromatin thus far. The major advantage of this format being that it permits visualization in popular protein modeling tools such as Chimera(Burkowski, 2014) and Pymol(Burkowski, 2014; Rigsby & Parker, 2016). However, the *.pdb format has a few key disadvantages. First, because the format is intended for protein files it contains entries with ambiguous meaning such as ATOM and HETATM records which do not make sense in the context of chromatin modeling. Secondly, the *.pdb format is missing useful metadata such as the identity of the chromosomes being represented and the genome index used for alignment. Thirdly the format is not readable by the newly developed integrative genomics viewer tool **spacewalk**, which provides genome specific visualization in browsers with the addition of visible annotations and .

To improve these issue the revised gsdb now provides data in two additional formats: *.g.pdb and *.sw. The *.g.pdb format is still visible in popular tools like Pymol and Chimera, while now including valuable metadata. The *.sw format is viewable within spacewalk (Figure 5.3).

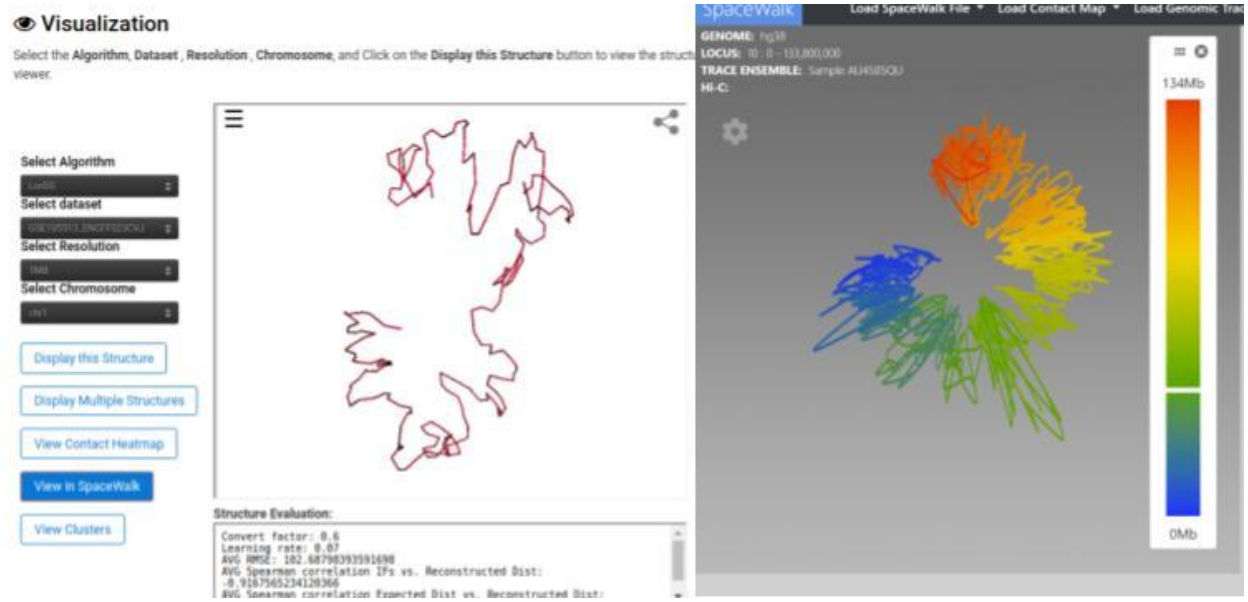



**Figure 5.3 SpaceWalk Visualization** (a) A visualization of chromosome 1 within GSDB embedded pymol. Clicking the highlighted “View in Spacewalk” button will redirect the user to (b) the space walk visualization tool, which permits a more fluid and customizable visualization of all GSDB structures.

Furthermore, to increase proliferation of the genome specific *.sw format, this paper's author worked in collaboration with the igv team to improve their spacewalk tool and integrate the spacewalk application into GSDB. As a consequence, now every structure stored within gsdb can be viewed in the spacewalk app without downloading any files, and in combination with user provided annotations all without leaving the browser.

### *5.3.2 Comparative analysis of Generated Structures*

To provide users with more quantifying information in selection of structure prediction tools we perform unsupervised data analysis on the generated structures. This includes principal component analysis and hierarchical clustering between scale normalized structures (Figure 5.4). The corresponding analysis can be visualized for each cell line and chromosome within the GSDB browser.

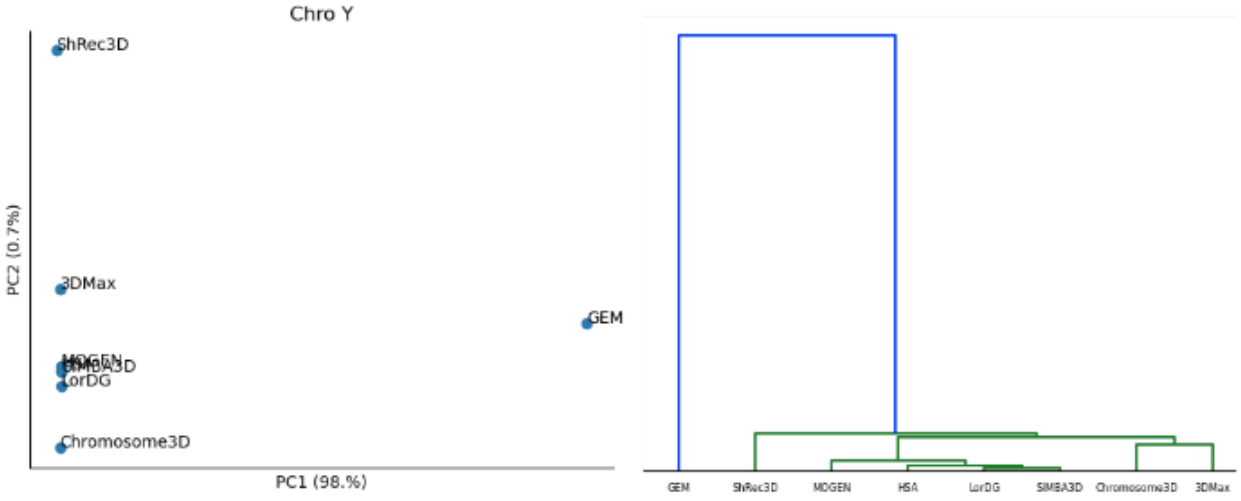



**Figure 5.4 GSDB PCA and Clustering** (a) Principal component analysis based on similarity of structures generated by different algorithms. (b) Hierarchical clustering based on structural similarity.

*5.3.3 Miscellaneous enhancement:*

We also provide a variety of miscellaneous viewing enhancements to the original GSDB. We expand the visualization utility of GSDB by introducing a variety of new pages such as a comparison page for viewing two structures simultaneously. We also provide a window for visualizing a customizable heatmap of contact matrices beside their 3D structures which permits observation of the impacts of different structural motifs such as TADs and the curvature of 3D models. Finally we provide a detailed tutorial accessible within the GDSB page to explain use of all GSDB features and tools to new users.

## 5.4 Discussion

The GSDB is the first repository for 3D structures generated from a wide range of Hi-C assays and Structure prediction algorithms. GSDB contains over 50,000 structures generated by 12 different modeling algorithms. Future development of GSDB will involve inclusion of more structure generating algorithms and expansion of the included datasets as more Hi-C data becomes publicly available.

# 6 Using tools

The tools VEHiCLE, TAPIOCA and 4DMax are all available for use at

https://github.com/Max-Highsmith/. Detailed instructions for use are provided in the

corresponding repositories README.md

## 6.1 Usage Instructions for VEHiCLE

### 6.1.1 Installation

VEHiCLE is built in python. The source code for VEHiCLE can be cloned from github at https://github.com/Max-Highsmith/VEHiCLE

Before using VEHiCLE the following python packages must be installed:
pytorch-lightning=1.0.3
pytorch=1.3.1
numpy=1.19.2
cupy=6.0
scipy=1.5.2
scikit-learn=0.23.2
Matplotlib=3.3.2

They can be manually installed or installed using the “vehicle_env.yml” file

### 6.1.2 Usage

To visualize our interactive tunable Hi-C contact matrix generating GUI run

“python Generative_GUI.py”

To enhance your own HiC data you will need to first set the following values in your configuration file

1. YOUR_CELL_LINE: <Name of User Cell Line>
2. LOW_RES_HIC: <location of Hi-C data”
3. CHRO : <chromosome number to be inspected>

After setting the required configuration variables run the command

"python Enhance_Your_Own_Data.py"

Once the above command is run a directory titled <Name of User Cell Line> will be created containing the following directories:

Constraints\ : <start coord> <end coord> <count>
Full_Enhanced\ : npy files containing enhanced matrices
Full_Mats\ :npy files containing unenhanced matrices
Full_Mats_Coords\ : coords mapping bins to genomic position
Splits\ : npy files containing pieces passed through the network.

To obtain the Insulation score identified TAD boundaries run

"python Insulation.py <enhanced_cell_line> <chromosome> <coordinate_file> <resolution> <tad filename>"

Example:

"Python insulation.py GM12878/Full_Enhanced/full_enh.npy 7 GM12878/Full_Mats_Coords/coords_chr7_res_10000.npy 10000 enh.txt"

## 6.2 Usage Instructions for TAPIOCA

### 6.2.1 Installation

TAPIOCA is built in python. The source code for TAPIOCA can be cloned from github at https://github.com/Max-Highsmith/TAPIOCA

Before using TAPIOCA the following python packages must be installed:
pytorch-lightning=1.1.8
pytorch=1.8
numpy=1.19.2
scipy=1.6.1
scikit-learn=0.23.2
matplotlib=3.3.2
jupyter=1.0.0

They can be manually installed or installed using the "tapioca_env.yml" file

### 6.2.2 Usage

Tapioca's primary value is demonstrating epigenetic features ability to predict chromosomal conformation using existing datasets. Because different types of epigenetic features may have different impacts on chromosome structure the pretrained weights from our experiment will only be applicable to novel data if all epigenetic features used in training are provided. However, to ensure reproducibility of the experiments outlined in the TAPIOCA paper each figure is generated through a jupyter notebook. These notebooks are named according to the figure or table which they generate.

To use TAPIOCA to make predictions on novel epigenetic data, follow the example jupyter notebook titles Transformer_Prediction.ipynb. Before running the notebook a few changes must be applied to fit your personal dataset. First you must adjust the input dataset, this can be done in two ways.

1. Creating a personal pytorch lightning data module described in https://www.pytorchlightning.ai/
2. Setting the parameters in Data/Drosophilla/FlyDataMod.py to point to the new dataset by editing the parameters:

    FEATURE_STRING: <path to csv of epigenetic features>

    (if training) LABEL_STRING: < path to csv of TAD labels>

After these strings are set to your new data location you must set the "ninp" parameter of the TransfromerModule to be the number of epigenetic features included in your training dataset. The default setting is 29.

## 6.3 Usage Instructions for 4DMax

### 6.3.1 Installation

4DMax is built in python. The source code for 4DMax can be cloned from github at https://github.com/Max-Highsmith/4DMax

Before using 4DMax the following python packages must be installed
pytorch-lightning=1.1.8
pytorch=1.8
numpy=1.19.2
scipy=1.5.2
matplotlib=3.3.2
jupyter=1.0.0
cupy=6.0.0
cudnn=7.6.5
torchvision=0.7.0

They can be manually installed or installed using the “4dmax_env.yml” file

### 6.3.2 Usage

1. Format your time series Hi-C data

*4DMax expects each Hi-C experiment in your time series to be represented as a 3 column tab separated text file where (pos1, pos2, val)*

2. Build dataset and hyper parameters configuration files

examples shown in:

1. Config/Datasets/example.json
2. Config/Hyper_Params/example.json
   *Hyper Params*
   - eta: weight of movement loss
   - alpha: contact map to distance constraint conversion ratio IF=d^alpha
   - lr: learning rate
   - epoch: number of epochs to train
3. *Data Set*
   - name: genomic Process name
   - step: granularity of 4D Model

- chro: chromosome number
- rep: biological replicate number
- taos: indx of hi-c experiments in time process
- datasets: hi-c experiment text files

3. Run 4DMax 'python 4dmax.py {input.dataset} {input.param}'

4 View Structures 'python Python_Scripts/create_gif.py {output.outfig} {input.npfile}

# References

Adhikari, B., Trieu, T., & Cheng, J. (2016). Chromosome3D: reconstructing three-dimensional chromosomal structures from Hi-C interaction frequency data using distance geometry simulated annealing. *BMC Genomics*, *17*(1), 886.

Bertero, A., Fields, P. A., Ramani, V., Bonora, G., Yardimci, G. G., Reinecke, H., Pabon, L., Noble, W. S., Shendure, J., & Murry, C. E. (2019). Dynamics of genome reorganization during human cardiogenesis reveal an RBM20-dependent splicing factory. *Nature Communications*, *10*(1), 1538.

Betzig, E., Patterson, G. H., Sougrat, R., Lindwasser, O. W., Olenych, S., Bonifacino, J. S., Davidson, M. W., Lippincott-Schwartz, J., & Hess, H. F. (2006). Imaging Intracellular Fluorescent Proteins at Nanometer Resolution. In *Science* (Vol. 313, Issue 5793, pp. 1642–1645). https://doi.org/10.1126/science.1127344

Blei, D. M., Kucukelbir, A., & McAuliffe, J. D. (2017). Variational Inference: A Review for Statisticians. In *Journal of the American Statistical Association* (Vol. 112, Issue 518, pp. 859–877). https://doi.org/10.1080/01621459.2017.1285773

Buitrago, D., Labrador, M., Arcon, J. P., Lema, R., Flores, O., Esteve-Codina, A., Blanc, J., Villegas, N., Bellido, D., Gut, M., Dans, P. D., Heath, S. C., Gut, I. G., Brun Heath, I., & Orozco, M. (2021). Impact of DNA methylation on 3D genome structure. *Nature Communications*, *12*(1), 3243.

Burkowski, F. J. (2014). *Computational and Visualization Techniques for Structural Bioinformatics Using Chimera*. CRC Press.

Calandrelli, R., Wu, Q., Guan, J., & Zhong, S. (2018). GITAR: An Open Source Tool for Analysis and Visualization of Hi-C Data. *Genomics, Proteomics & Bioinformatics*, *16*(5), 365–372.

Chen, C., Zha, Y., Zhu, D., Ning, K., & Cui, X. (n.d.). *Attention is all you need for*

*general-purpose protein structure embedding*. https://doi.org/10.1101/2021.01.31.428935

Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. (2010). In *Journal Club for Condensed Matter Physics*. https://doi.org/10.36471/jccm_february_2010_02

Consortium, T. E. P., & The ENCODE Project Consortium. (2004). The ENCODE (ENCyclopedia Of DNA Elements) Project. In *Science* (Vol. 306, Issue 5696, pp. 636–640). https://doi.org/10.1126/science.1105136

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

Cui, C., Shu, W., & Li, P. (2016). Fluorescence In situ Hybridization: Cell-Based Genetic Diagnostic and Research Applications. *Frontiers in Cell and Developmental Biology*, *4*, 89.

Dali, R., & Blanchette, M. (2017). A critical assessment of topologically associating domain prediction tools. *Nucleic Acids Research*, *45*(6), 2994–3005.

Dekker, J. (2008). Gene Regulation in the Third Dimension. In *Science* (Vol. 319, Issue 5871, pp. 1793–1794). https://doi.org/10.1126/science.1152850

Dimmick, M. C., Lee, L. J., & Frey, B. J. (n.d.). *HiCSR: a Hi-C super-resolution framework for producing highly realistic contact maps*. https://doi.org/10.1101/2020.02.24.961714

Di Stefano, M., Stadhouders, R., Farabella, I., Castillo, D., Serra, F., Graf, T., & Marti-Renom, M. A. (2020). Transcriptional activation during cell reprogramming correlates with the formation of 3D open chromatin hubs. *Nature Communications*, *11*(1), 2564.

Dixon, J. R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J. S., & Ren, B. (2012). Topological domains in mammalian genomes identified by analysis of chromatin interactions. In *Nature* (Vol. 485, Issue 7398, pp. 376–380). https://doi.org/10.1038/nature11082

Duan, S., & Zhao, H. (2020). Attention Is All You Need for Chinese Word Segmentation. In

*Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. https://doi.org/10.18653/v1/2020.emnlp-main.317

Filippova, D., Patro, R., Duggal, G., & Kingsford, C. (2013). Multiscale Identification of Topological Domains in Chromatin. In *Lecture Notes in Computer Science* (pp. 300–312). https://doi.org/10.1007/978-3-642-40453-5_23

Fortin, J.-P., & Hansen, K. D. (2015). Reconstructing A/B compartments as revealed by Hi-C using long-range correlations in epigenetic data. *Genome Biology*, *16*, 180.

Fraser, J., Rousseau, M., Shenker, S., Ferraiuolo, M. A., Hayashizaki, Y., Blanchette, M., & Dostie, J. (2009). Chromatin conformation signatures of cellular differentiation. In *Genome Biology* (Vol. 10, Issue 4, p. R37). https://doi.org/10.1186/gb-2009-10-4-r37

Fraser, P., & Bickmore, W. (2007). Nuclear organization of the genome and the potential for gene regulation. In *Nature* (Vol. 447, Issue 7143, pp. 413–417). https://doi.org/10.1038/nature05916

Garreta, R., & Moncecchi, G. (2013). *Learning scikit-learn: Machine Learning in Python*. Packt Publishing Ltd.

Han, J., Zhang, Z., & Wang, K. (2018). 3C and 3C-based techniques: the powerful tools for spatial genome organization deciphering. In *Molecular Cytogenetics* (Vol. 11, Issue 1). https://doi.org/10.1186/s13039-018-0368-2

Highsmith, M., & Cheng, J. (2021). VEHiCLE: a Variationally Encoded Hi-C Loss Enhancement algorithm for improving and generating Hi-C data. *Scientific Reports*, *11*(1), 8880.

Highsmith, M., Oluwadare, O., & Cheng, J. (n.d.). *Deep Learning For Denoising Hi-C Chromosomal Contact Data*. https://doi.org/10.1101/692558

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

Huang, B., Babcock, H., & Zhuang, X. (2010). Breaking the diffraction barrier: super-resolution

imaging of cells. *Cell*, *143*(7), 1047–1058.

Imakaev, M., Fudenberg, G., McCord, R. P., Naumova, N., Goloborodko, A., Lajoie, B. R., Dekker, J., & Mirny, L. A. (2012). Iterative correction of Hi-C data reveals hallmarks of chromosome organization. *Nature Methods*, *9*(10), 999–1003.

Kingma, D. P., & Welling, M. (2019). *An Introduction to Variational Autoencoders*. https://doi.org/10.1561/9781680836233

Lajoie, B. R., Dekker, J., & Kaplan, N. (2015). The Hitchhiker's guide to Hi-C analysis: practical guidelines. *Methods* , *72*, 65–75.

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

Li, J., Zhang, W., & Li, X. (2018). 3D Genome Reconstruction with ShRec3D and Hi-C Data. In *IEEE/ACM Transactions on Computational Biology and Bioinformatics* (Vol. 15, Issue 2, pp. 460–468). https://doi.org/10.1109/tcbb.2016.2535372

Liu, D., Wang, Y., & Kato, J. (2019). Supervised Spatial Transformer Networks for Attention Learning in Fine-grained Action Recognition. In *Proceedings of the 14th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*. https://doi.org/10.5220/0007257803110318

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

Liu, T., & Wang, Z. (2019a). HiCNN2: Enhancing the Resolution of Hi-C Data Using an Ensemble of Convolutional Neural Networks. *Genes*, *10*(11). https://doi.org/10.3390/genes10110862

Liu, T., & Wang, Z. (2019b). HiCNN: a very deep convolutional neural network to better enhance

the resolution of Hi-C data. *Bioinformatics* , *35*(21), 4222–4228.

Miele, A., & Dekker, J. (2008). Long-range chromosomal interactions and gene regulation. In *Molecular BioSystems* (Vol. 4, Issue 11, p. 1046). https://doi.org/10.1039/b803580f

Mifsud, B. (2018a). ChIP-seq-specific Quality Control. In *Practical Guide to ChIP-seq Data Analysis* (pp. 35–40). https://doi.org/10.1201/9780429487590-5

Mifsud, B. (2018b). Introduction to ChIP-seq. In *Practical Guide to ChIP-seq Data Analysis* (pp. 1–10). https://doi.org/10.1201/9780429487590-1

Misteli, T. (2007). Beyond the Sequence: Cellular Organization of Genome Function. In *Cell* (Vol. 128, Issue 4, pp. 787–800). https://doi.org/10.1016/j.cell.2007.01.028

Miura, H., Poonperm, R., Takahashi, S., & Hiratani, I. (2018). Practical Analysis of Hi-C Data: Generating A/B Compartment Profiles. *Methods in Molecular Biology* , *1861*, 221–245.

November, J. (2018). More than Moore's Mores: Computers, Genomics, and the Embrace of Innovation. In *Journal of the History of Biology* (Vol. 51, Issue 4, pp. 807–840). https://doi.org/10.1007/s10739-018-9539-6

Oluwadare, O., Highsmith, M., & Cheng, J. (2019). An Overview of Methods for Reconstructing 3-D Chromosome and Genome Structures from Hi-C Data. *Biological Procedures Online*, *21*, 7.

Oluwadare, O., Highsmith, M., Turner, D., Lieberman Aiden, E., & Cheng, J. (2020). GSDB: a database of 3D chromosome and genome structures reconstructed from Hi-C data. *BMC Molecular and Cell Biology*, *21*(1), 60.

Oluwadare, O., Zhang, Y., & Cheng, J. (2018). A maximum likelihood algorithm for reconstructing 3D structures of human chromosomes from chromosomal contact data. *BMC Genomics*, *19*(1), 161.

Ramírez, F., Bhardwaj, V., Arrigoni, L., Lam, K. C., Grüning, B. A., Villaveces, J., Habermann, B., Akhtar, A., & Manke, T. (2018). High-resolution TADs reveal DNA sequences underlying genome organization in flies. *Nature Communications*, *9*(1), 189.

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

Rigsby, R. E., & Parker, A. B. (2016). Using the PyMOL application to reinforce visual understanding of protein structure. In *Biochemistry and Molecular Biology Education* (Vol. 44, Issue 5, pp. 433–437). https://doi.org/10.1002/bmb.20966

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

Rosenthal, M., Bryner, D., Huffer, F., Evans, S., Srivastava, A., & Neretti, N. (2019). Bayesian Estimation of Three-Dimensional Chromosomal Structure from Single-Cell Hi-C Data. *Journal of Computational Biology: A Journal of Computational Molecular Cell Biology*, *26*(11), 1191–1202.

Rozenwald, M. B., Galitsyna, A. A., Sapunov, G. V., Khrameeva, E. E., & Gelfand, M. S. (2020). A machine learning framework for the prediction of chromatin folding in using epigenetic features. *PeerJ. Computer Science*, *6*, e307.

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

Stadhouders, R., Vidal, E., Serra, F., Di Stefano, B., Le Dily, F., Quilez, J., Gomez, A., Collombet, S., Berenguer, C., Cuartero, Y., Hecht, J., Filion, G. J., Beato, M., Marti-Renom, M. A., & Graf, T. (2018). Transcription factors orchestrate dynamic interplay between genome topology and gene regulation during cell reprogramming. *Nature Genetics*, *50*(2), 238–249.

Trenkmann, M. (2019). 3D genome shows its stripes in gene-rich regions. In *Nature Reviews*

*Genetics* (Vol. 20, Issue 4, pp. 192–193). https://doi.org/10.1038/s41576-019-0102-x

Trieu, T., & Cheng, J. (2016). MOGEN: a tool for reconstructing 3D models of genomes from chromosomal conformation capturing data. In *Bioinformatics* (Vol. 32, Issue 9, pp. 1286–1292). https://doi.org/10.1093/bioinformatics/btv754

Trieu, T., & Cheng, J. (2017). 3D genome structure modeling by Lorentzian objective function. *Nucleic Acids Research*, *45*(3), 1049–1058.

Ulianov, S. V., Khrameeva, E. E., Gavrilov, A. A., Flyamer, I. M., Kos, P., Mikhaleva, E. A., Penin, A. A., Logacheva, M. D., Imakaev, M. V., Chertovich, A., Gelfand, M. S., Shevelyov, Y. Y., & Razin, S. V. (2016). Active chromatin and transcription play a key role in chromosome partitioning into topologically associating domains. *Genome Research*, *26*(1), 70–84.

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

van der Wee, E. B., Fokkema, J., Kennedy, C. L., Del Pozo, M., de Winter, D. A. M., Speets, P. N. A., Gerritsen, H. C., & van Blaaderen, A. (2021). 3D test sample for the calibration and quality control of stimulated emission depletion (STED) and confocal microscopes. *Communications Biology*, *4*(1), 909.

Varoquaux, N., Ay, F., Noble, W. S., & Vert, J.-P. (2014). A statistical approach for inferring the 3D structure of the genome. In *Bioinformatics* (Vol. 30, Issue 12, pp. i26–i33). https://doi.org/10.1093/bioinformatics/btu268

Vian, L., Pękowska, A., Rao, S. S. P., Kieffer-Kwon, K.-R., Jung, S., Baranello, L., Huang, S.-C., El Khattabi, L., Dose, M., Pruett, N., Sanborn, A. L., Canela, A., Maman, Y., Oksanen, A., Resch, W., Li, X., Lee, B., Kovalchuk, A. L., Tang, Z., … Casellas, R. (2018). The Energetics and Physiological Impact of Cohesin Extrusion. *Cell*, *175*(1), 292–294.

Wang, S., Xu, J., & Zeng, J. (2015). Inferential modeling of 3D chromatin structure. *Nucleic Acids Research*, *43*(8), e54.

Wasim, A., Gupta, A., & Mondal, J. (2021). A Hi-C data-integrated model elucidates E. coli chromosome's multiscale organization at various replication stages. *Nucleic Acids Research*, *49*(6), 3077–3091.

Yang, S.-W., Liu, A. T., & Lee, H.-Y. (2020). Understanding Self-Attention of Self-Supervised Audio Transformers. In *Interspeech 2020*. https://doi.org/10.21437/interspeech.2020-2231

Yang, T., Zhang, F., Yardımcı, G. G., Song, F., Hardison, R. C., Noble, W. S., Yue, F., & Li, Q. (2017). HiCRep: assessing the reproducibility of Hi-C data using a stratum-adjusted correlation coefficient. In *Genome Research* (Vol. 27, Issue 11, pp. 1939–1949). https://doi.org/10.1101/gr.220640.117

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

Yardımcı, G. G., Ozadam, H., Sauria, M. E. G., Ursu, O., Yan, K.-K., Yang, T., Chakraborty, A., Kaul, A., Lajoie, B. R., Song, F., Zhan, Y., Ay, F., Gerstein, M., Kundaje, A., Li, Q., Taylor, J., Yue, F., Dekker, J., & Noble, W. S. (2019). Measuring the reproducibility and quality of Hi-C data. *Genome Biology*, *20*(1), 57.

Zhang, Y., An, L., Xu, J., Zhang, B., Zheng, W. J., Hu, M., Tang, J., & Yue, F. (2018). Enhancing Hi-C data resolution with deep convolutional neural network HiCPlus. *Nature Communications*, *9*(1), 750.

Zhang, Y., & Skolnick, J. (2005). TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Research*, *33*(7), 2302–2309.

Zhang, Z., Li, G., Toh, K.-C., & Sung, W.-K. (2013). Inference of Spatial Organizations of Chromosomes Using Semi-definite Embedding Approach and Hi-C Data. In *Lecture Notes in Computer Science* (pp. 317–332). https://doi.org/10.1007/978-3-642-37195-0_31

Zhu, G., Deng, W., Hu, H., Ma, R., Zhang, S., Yang, J., Peng, J., Kaplan, T., & Zeng, J. (2018). Reconstructing spatial organizations of chromosomes through manifold learning. *Nucleic Acids Research*, *46*(8), e50.

Zou, C., Zhang, Y., & Ouyang, Z. (2016). HSA: integrating multi-track Hi-C data for genome-scale reconstruction of 3D chromatin structure. In *Genome Biology* (Vol. 17, Issue 1). https://doi.org/10.1186/s13059-016-0896-1

Zufferey, M., Tavernari, D., Oricchio, E., & Ciriello, G. (2018). Comparison of computational methods for the identification of topologically associating domains. *Genome Biology*, *19*(1), 217.

# Vita

Max Highsmith was raised in St. Louis, Missouri. He obtained his Bachelor's degree in Mathematics from Truman State University, and Master's degree in Applied Mathematics from the University of Missouri, in 2015 and 2018 respectively. While in the Math Department his research was focused on mathematical modeling of large scale sales pipelines.

Prior to his Ph.D studies Max worked as a software engineer for multiple companies including Flight Safety International, MidwayUSA, and Conduce Inc.

Max started his Ph.D. studies in the Department of Electrical Engineering and Computer Science at the University of Missouri-Columbia in 2018. With research interests in genomics and deep learning, he has been published in reputable journals including: BMC Molecular and Cell Biology, BMC Bioinformatics, Biological Procedures Online, and Scientific Reports.