Distribution Agreement

In presenting this thesis as a partial fulfillment of the requirements for a degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis in whole or in part in all forms of media, now or hereafter now, including display on the World Wide Web. I understand that I may select some access restrictions as part of the online submission of this thesis. I retain all ownership rights to the copyright of the thesis. I also retain the right to use in future works (such as articles or books) all or part of this thesis.

Alex Belov April 8, 2025

Epiphany2: A Novel CNN-Transformer Method for Predicting 3D Chromatin Structure from Epigenetic Data

by

Alex Belov

Dr. Joyce Ho
Adviser

Computer Science

Dr. Joyce Ho

Adviser

Dr. Judy Gichoya

Committee Member

Dr. Bree Ettinger

Committee Member

2025

Epiphany2: A Novel CNN-Transformer Method for Predicting 3D Chromatin Structure from Epigenetic Data

By

Alex Belov

Dr. Joyce Ho
Adviser

An abstract of
a thesis submitted to the Faculty of Emory College of Arts and Sciences
of Emory University in partial fulfillment
of the requirements of the degree of
Bachelor of Science with Honors

Computer Science

2025

Abstract

Epiphany2: A Novel CNN-Transformer Method for Predicting 3D Chromatin Structure from Epigenetic Data
By Alex Belov

Understanding the three-dimensional (3D) organization of chromatin and its regulation by the epigenome is critical for unraveling the complexities of gene expression, cellular differentiation, and disease mechanisms. While current models leveraging 1D epigenomic data to predict 3D chromatin structure have shown promise, many suffer from key weaknesses, including a limited ability to capture cell-type-specific interactions and a lack of global contextual understanding of chromatin dynamics. This thesis presents a novel model architecture with state-of-the-art performance in predicting Hi-C contact maps. Our model is the first to use transformer layers to capture long-range dependencies for this task.

Epiphany2: A Novel CNN-Transformer Method for Predicting 3D Chromatin Structure from Epigenetic Data

By

Alex Belov

Dr. Joyce Ho
Adviser

A thesis submitted to the Faculty of Emory College of Arts and Sciences
of Emory University in partial fulfillment
of the requirements of the degree of
Bachelor of Science with Honors

Computer Science

2025

Acknowledgments

I would like to thank the Emory University Computer Science Department for its support and academic freedom over the last four years. I would also like to acknowledge the Biology department's expertise and exploratory nature that allowed me to gain a foundation in cell biology in epigenetics. I would also like to thank the Mathematics Department for helping build my statistical foundation and love for mathematical biology. I am grateful to have had Dr. Joyce Ho as my thesis advisor and for her guidance, patience, and support. I cannot thank Dr. Judy Gichoya enough for her continued mentorship and investment in my research career; without her, I would not be here today. I would also like to thank Dr. Bree Ettinger for her support in my statistics studies and for helping me find a passion in combining my three fields. Lastly, I would like to thank Dr. Christina Leslie, Dr. Rui Yang, and Wilfred Wong from Memorial Sloan Kettering Cancer Center for their continued support and mentorship on this project.

Table of Contents

Chapter 1 **8**
Introduction 8
Chapter 2 **11**
2.1 Background 11
2.2 Motivation 13
2.3 Challenges 14
Chapter 3: Existing Works **16**
3.1 Epigenome → Hi-C 17
3.2 Similar Biological Tasks 20
Chapter 4: Technical Designs **22**
Chapter 5: Experimental Settings **25**
5.1 Preprocessing 25
Chapter 6: Experiments & Results **27**
Chapter 7: Conclusion **30**
Chapter 8: Limitations & Future Work **31**
Bibliography **33**

Chapter 1

Introduction

The three-dimensional (3D) organization of chromatin within the nucleus plays a crucial role in regulating gene expression, replication timing, and cellular differentiation. Understanding how this 3D structure is influenced by epigenetic modifications, such as histone marks, is critical for deciphering the mechanisms that drive cellular function and phenotypic diversity. These epigenetic marks, which include modifications like methylation and acetylation, dictate how tightly or loosely DNA is packaged, affecting its accessibility to transcription machinery. As a result, different cell types, despite sharing the same DNA sequence, exhibit different 3D chromatin structures that contribute to their unique gene expression profiles.

Hi-C assays have emerged as one of the most powerful tools for studying the 3D folding of chromatin. These assays capture interactions between genomic regions and produce contact maps, which provide insights into how the genome is organized within the nucleus. However, the experimental generation of high-resolution Hi-C data is both time-consuming and costly. Additionally, interpreting this data requires sophisticated computational models due to the complexity of chromatin folding patterns.

Given the complexity and volume of chromatin interaction data, deep learning has become an essential tool for predicting 3D chromatin structure from more easily accessible 1D epigenomic

data. By leveraging information such as histone modifications, transcription factor binding, and chromatin accessibility data, machine learning models can generate predictions of chromatin contact maps, allowing for the exploration of chromatin organization without the need for expensive Hi-C experiments.

The primary motivation for predicting 3D chromatin structure from epigenomic data is:

1. To answer fundamental questions about how epigenetic modifications lead to 3D structural changes in chromatin. These predictions enable the study of chromatin remodeling processes that underlie gene regulation and cell differentiation.
2. To provide a proxy for studying enhancer-promoter interactions, which are essential long-range regulatory interactions that play a critical role in gene expression.
3. To generate synthetic Hi-C data for downstream analysis, making it possible to perform genomic studies in cases where high-resolution Hi-C data is not available due to cost or practical constraints.

Here, we aim to use transformers' success to find long-range dependencies in sequence data; our inputs are long sequences of epigenetic marks. In this thesis, we present Epiphany-2, a deep learning model that predicts 3D chromatin structure (Hi-C) from epigenetic marks. Our model is the first to use transformer layers to capture long-range dependencies between the epigenome and Hi-C. Epiphany-2 builds on Epiphany [5], a convolutional neural network (CNN) for the same task. Finally, we demonstrate that using transformer layers drastically improves our model's

ability to pick up chromatin loops and accurately predict topologically associated domains (TADs), presenting state-of-the-art performance.

Chapter 2

2.1 Background

Epigenomics is the study of the epigenetic changes in a cell that affect gene expression without altering the DNA sequence. These changes primarily involve chemical modifications of DNA or histone proteins around which DNA is wrapped. Histone marks are one of the most significant forms of epigenetic regulation. These chemical modifications—such as methylation and acetylation—on histones influence how tightly or loosely DNA is packaged in the chromatin.

Histone modifications vary between cell types, which leads to different regions of DNA being more accessible or restricted. This differential accessibility plays a key role in controlling gene expression, allowing different cells to take on specialized roles (i.e., different cell phenotypes) despite having the same underlying DNA sequence. For example, in muscle cells, certain genes that are required for muscle function are made more accessible, while the same genes may be tightly packed and inaccessible in neurons.

One way to analyze chromatin structure is through Hi-C, a genome-wide chromosome conformation capture assay that allows researchers to measure the 3D organization of chromatin. Hi-C produces a contact matrix that reflects how often two genomic loci interact in 3D space. The folding of chromatin influences which genes are active and which are silent, with tightly

packed regions (heterochromatin) typically being less transcriptionally active than loosely packed regions (euchromatin).

Through these 3D interactions, topologically associating domains (TADs) emerge, which are clusters of regions in the genome that interact more frequently with one another than with regions outside the TAD. Understanding how epigenetic marks impact the folding and structure of chromatin at this level is crucial for understanding gene regulation, development, and disease processes.

2.2 Motivation

A nuanced understanding of chromatin remodeling through epigenomic marks requires the analysis of vast datasets that are often too complex to interpret with heuristics or simple models. While we understand that epigenomic marks influence chromatin structure, we lack the tools to precisely predict or explain which marks lead to specific 3D structural changes. Machine learning provides a powerful tool to analyze these data and help identify which epigenomic features are most related to structural changes. With models trained on epigenomic data, we can conduct virtual perturbation studies—where we manipulate the influence of specific epigenomic marks and predict their impact on chromatin structure. Understanding 3D chromatin structure can serve as a proxy to predict long-range interactions, such as enhancer-promoter interactions. These interactions play a critical role in gene regulation, allowing regulatory elements far from the promoter to influence transcriptional activity. High-resolution Hi-C data is often expensive and impractical to collect for every sample. With machine learning models, we can generate synthetic Hi-C data based on epigenomic input, which can be used for downstream analyses such as identifying TAD boundaries, predicting interactions, and understanding regulatory changes.

2.3 Challenges

There are several key challenges in fully understanding how epigenetic modifications influence the 3D structure of chromatin. One major challenge is the complexity of the chromatin folding process, which is influenced by a combination of DNA sequence, epigenetic marks, and regulatory protein interactions. The sheer volume of potential combinations makes it difficult to generalize findings across different cell types and conditions.

For example, current models often use either DNA sequence or epigenomic data to predict 3D structure, but there is a growing need for cell-type-specific models that can incorporate multiple layers of information. Moreover, generating synthetic Hi-C data remains a challenge, as models must balance the accuracy of prediction with the biological realism of the synthetic data produced. Finally, another ongoing difficulty is transferring models across species or cell types without introducing biases or losing accuracy.

Summary of Challenges:

1. **Complexity of epigenetic regulation and chromatin folding:** Understanding how different histone marks and chromatin-associated proteins collectively impact 3D chromatin structure is a non-trivial problem that involves complex interactions.

2. **Need for cell-type-specific models:** Many models lack the capacity to fully integrate cell-type-specific information from both epigenomic marks and chromatin structure, limiting their generalizability.
3. **Synthetic Hi-C generation challenges:** Balancing the accuracy and realism of synthetic Hi-C data while ensuring it is usable for downstream biological applications is a key area of improvement for current models.

Chapter 3: Existing Works

There has been substantial work to model chromatin contact maps using epigenetic marks, DNA sequences, and other input features. These models have three significant differences: model architectures, prediction aggregation techniques, and input features. Because the genome is large, predicting the entire chromatin contact map is computationally infeasible. Instead, state-of-the-art methods predict sections of the Hi-C matrix and aggregate their predictions to output a final prediction. The three most successful prediction techniques we observed were predicting V-stripes, vertical (one-pixel-wide zigzags) stripes, and individual pixels of the Hi-C map.

3.1 Epigenome → Hi-C

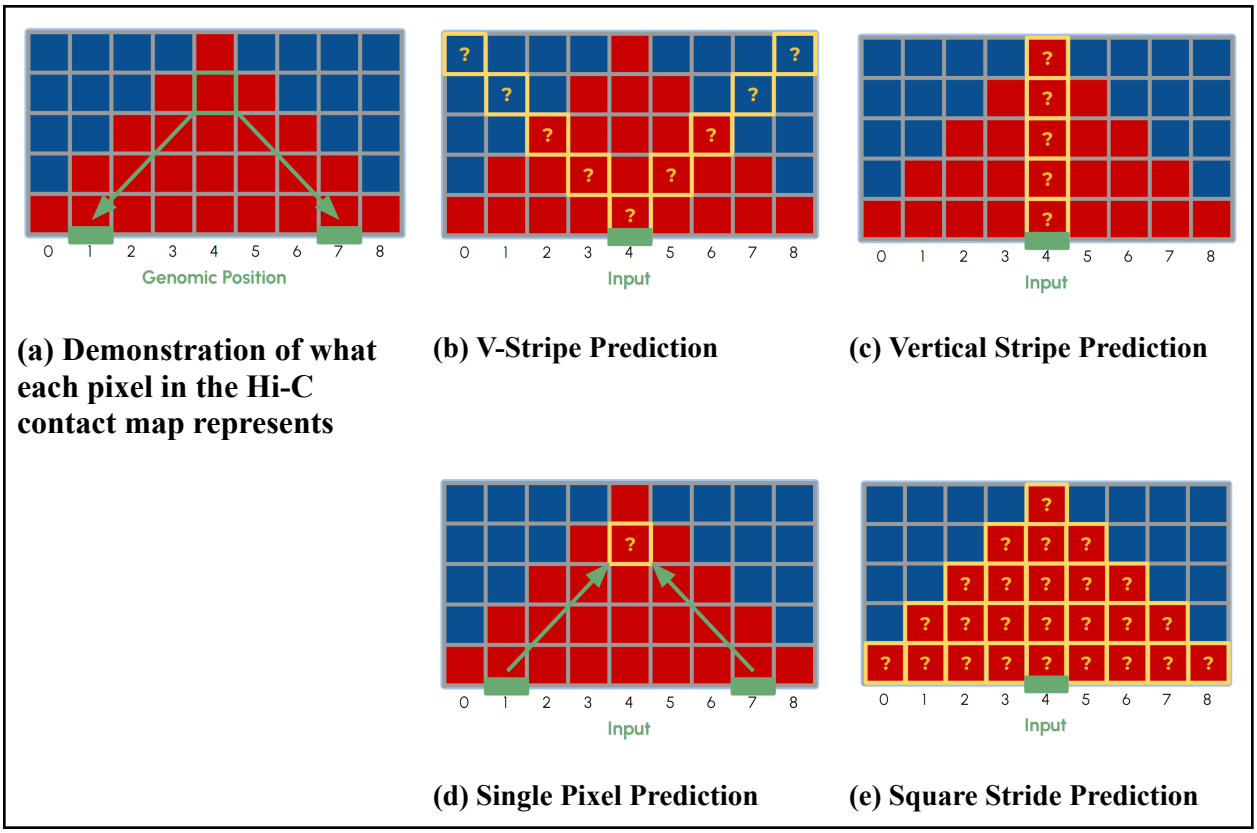
Chromafold [3], a model that uses bulk epigenetic signal data and single-cell ATAC-seq data, is the state-of-the-art model to predict Hi-C data. The reason for Chromafold's superior success, in terms of correlation of its predicted Hi-C map across distance away from the genome and insulation score, is difficult to pinpoint; its use of single-cell ATAC-seq data provides a unique variety of training examples which allow for greater model generalizability. However, its implementation of a V-stripe prediction is a strong choice compared to the vertical-stripe prediction of Epiphany. For a visual depiction of the V-stripe prediction method, see Figure #1b. Predicting a V-stripe from the genome represents using one epigenomic bin to predict the contact map values of each of its neighbors. However, predicting a vertical stripe (Figure #1c) poses a less biologically meaningful task. In the Hi-C matrix, each pixel h pixels above a genomic bin represents the contact value of the genomic positions h cells to the left and h cells to the right, which does not relate to the center genomic position, shown in Figure #1a. Therefore, vertical-stripe predictions predict contact values that are not necessarily representative of the epigenomic data given as their input. Epiphany uses a vertical stripe prediction method, but to improve biological relevance, in hopes of capturing more epigenetic signal correlated with 3D chromatin structure, we use the V-stripe prediction method in Epiphany2.

However, one major limitation of the V-stripe prediction is the idea of "one-sided predictions." By this, we mean that each pixel uses the epigenetic data from just one bin, instead of the two they represent. However, it is important to note that each pixel will be covered twice in our sliding window, once by the left arm and once by the right arm of the V-shape. However, each pixel value prediction only has local information from one bin. The solution used by

ChromaFold [3] is to represent a wide enough epigenetic window to cover all of the potential range of genomic positions represented in the V-shape. However, one problem with this solution is that it adds largely redundant input data between model runs.

Additionally, each pixel is predicted by two V-stripes, and these two "one-sided" predictions are simply averaged together. Zhang et al. [6], a predecessor to ChomaFold and Epiphany, takes in just the two epigenomic profiles bins and directly predicts the pixel representing the genomic interaction without including the signal from other bins (Figure #1d). Another important concept to consider is the variability in the exact locations of chromatin loops and topologically associated domains. Although this method likely decreases signal noise, it also risks omitting relevant signals outside its window. To represent the maximum window size for which we expect a meaningful epigenetic signal, we use a window size of 2.4 Mb. We expect ~1Mb to be the farthest distance for epigenomic interactions to our center point and include this distance to the left and right of our center with a 0.2 Mb buffer zone on each side.

Epiphany uses an adversarial loss term to create more realistic-looking, less blurry Hi-C maps, but this method has worsened the model's accuracy. To compare biologically meaningful results, all iterations of Epiphany (and Epiphany2) are trained without an adversarial loss term; however, we recognize this worsens Epiphany2's interpretability.



(a) Demonstration of what each pixel in the Hi-C contact map represents

(b) V-Stripe Prediction

(c) Vertical Stripe Prediction

(d) Single Pixel Prediction

(e) Square Stride Prediction

Figure 1: Illustrations of different prediction methods in Hi-C contact maps.

3.2 Similar Biological Tasks

Although we aim to understand only epigenetic inputs, DNA sequence is a well-studied input to predict chromatin contact maps. For a summary of existing methods, see Table #1. Because they use fundamentally different input data, the results of these sequence models, like DeepC [4] and Akita [2], are not directly comparable to those discussed above. However, we can learn from their model architecture and prediction aggregation methods. We identified several models that either use epigenetics and DNA sequence or just DNA sequence. Sliding window CNNs to predict the Hi-C contact map is standard for well-performing models. However, more nuanced implementation of transformers and attention mechanisms have been adopted in these tasks. C.Origami is a state-of-the-art model to predict Hi-C from both epigenetic marks and DNA sequence and uses CNN encoders, a self-attention transformer layer, and then a CNN decoder to leverage the strengths of both the CNN and the transformer [4]. However, C.Origami struggles to generalize to new cell types as it often focuses too heavily on its DNA input, which is identical across cell types of the same organism; we hypothesize that epigenetic information is sufficient to predict 3D chromatin structure and leads to more generalizable predictions. Surprisingly, transformers have not been widely adopted in predicting from just epigenetic data; we believe transformers' ability to achieve a nuanced understanding of global dependencies is a powerful tool that has been underutilized in the task we are interested in.

Article Name	Inputs	Architecture	Prediction Method
Epiphany [5]	Epigenome	CNN + Bi-LSTSM	Vertical Stripe
ChromaFold [3]	Epigenome	CNN	V-stripe
In silico prediction… [6]	Epigenome	Random Forest	Single Pixel
C.Origami [4]	Epigenome + DNA	CNN + Transformer	Square Stride
DeepC [4]	DNA	CNN	Vertical Stripe
Akita [2]	DNA	CNN	Square Stride

Table 1: Overview of Hi-C Contact Map Prediction Methods

Chapter 4: Technical Designs

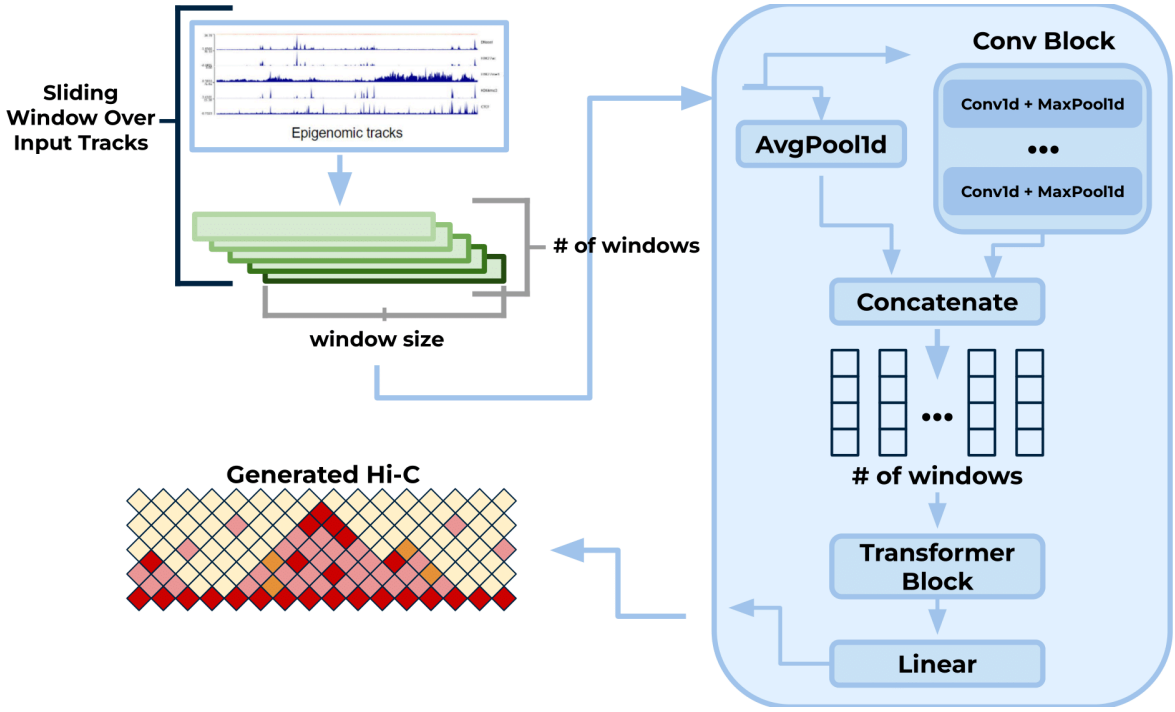


Figure 2: Outline of the Epiphany-2 Architecture to Predict Epigenomic Features with 3D Chromatin Structure (Hi-C Assay)

Epigenomic signal tracks are first presented to the model in a sliding window fashion, with a window size of 2.4 Mb and a step size of 10 kb. The generator first uses convolution modules to extract features from the processed input data, followed by a transformer block to capture the long-range dependencies between bins. After a fully connected layer, the predicted contact map is generated. An MSE loss between the predicted map and the ground truth is calculated to train the generator to predict correct structures. A visual depiction of our model is in Figure 1. We used a V-stripe prediction scheme to match Chromafold and other state-of-the-art models

because this prediction scheme is more consistent with the biological meaning of the HiC map values.

Our model employs a hybrid architecture combining convolutional layers, a transformer encoder, and attention pooling, culminating in a classification head. The input data is first processed through a 1D average pooling layer that reduces the resolution by pooling over a kernel size proportional to the input dimensions. This operation is followed by a sequence of 1D convolutional layers designed to extract hierarchical features. Each convolutional layer uses a progressively increasing dilation rate to capture multi-scale contextual information. The output of these layers undergoes batch normalization and ReLU activation, with intermediate max-pooling steps to downsample the feature maps.

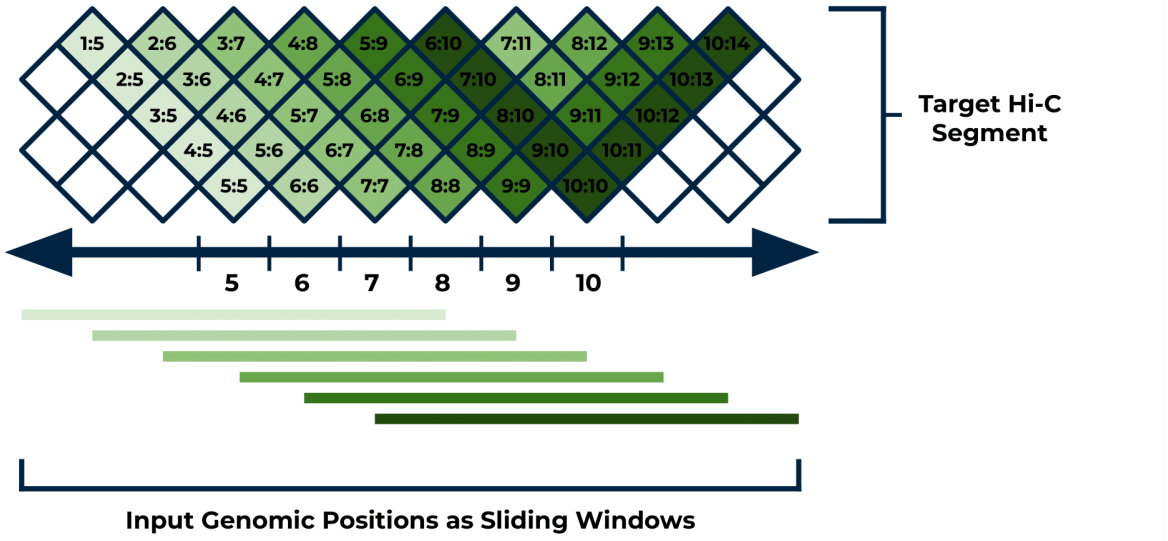


Figure 3: Illustration of the Sliding Window Prediction Scheme

The first window of input data (light green horizontal line, 2.4 Mb) is used to predict a vector on the Hi-C contact map that is diagonal (v-shape) to the center line (light green vector, covers 1 Mb from the diagonal to the left and right). Note that an extra .2 Mb of input is added to either side of each input window (a total length of 2.4 Mb instead of 2 Mb) to give the model additional context.

Next, the feature maps are concatenated with the pooled outputs along the channel dimension, forming a unified representation. This representation is projected into a 512-dimensional space using a linear transformation preceded by layer normalization. The resulting tensor is reshaped and fed into a transformer encoder. The transformer encoder consists of six layers of self-attention modules, each configured with eight attention heads and a feedforward network with 2048 hidden units. This architecture enables the model to capture long-range dependencies in the data.

After passing through the transformer, the features are aggregated using a multi-head attention pooling mechanism. This step selectively focuses on the most relevant features by applying learned attention weights. Finally, the aggregated features are processed through a fully connected classification head consisting of two linear layers. The first layer expands the feature space to 1024 dimensions, applies a ReLU activation, and then reduces the dimensionality to the desired output size of 200. The model integrates convolutional feature extraction, long-range contextualization through transformers, and attention-based feature aggregation to produce robust predictions

Chapter 5: Experimental Settings

5.1 Preprocessing

The epigenetic marks used in this study, including DNase I, CTCF, H3K4me3, H3K27ac, and H3K27me3, were sourced from the ENCODE data portal [7] for the hg38 genome assembly. DNase I is an assay similar to ATAC-seq that measures the openness of chromatin, which is essential for the binding of transcription factors and other regulatory proteins. CTCF is a zinc finger protein that is a regulator of chromatin looping, bringing together distant genomic elements for joint regulation. The other epigenomic marks, H3K4me3, H3K27ac, and H3K27me3, are common histone modifications that are known to regulate gene expression, however, their contextual meaning is difficult to study without computational methods. Our data was initially downloaded in BAM format, and replicate files were merged using the pysam Python module. The merged BAM files were then converted to BigWig format using the bamCoverage tool in deepTools, with a bin size of 10 bp and normalization by Reads Per Genomic Content (RPGC). The genome-wide coverage was segmented into 100-bp bins to prepare the data for model input, and bin-level signals for the five epigenomic tracks were extracted. All data used in this thesis can be found in Table #2.

Cell type	Dnase I	CTCF	H3K27ac	H3K27me3	H3K4me3	Hi-C
GM12878 - human B lymphocytes	ENCSR000EMT	ENCSR000DRZ	ENCSR000DRY	ENCSR000DRX	ENCSR000AKC	4DNFI1UEG1HD
H1-hESC - human embryonic stem cell	ENCSR000EMU	ENCSR000AMF	ENCSR000ANP	ENCSR216OGD	ENCSR019SQX	4DNFIQYQWPF5
K562 - human chronic myeloid leukemia	ENCSR000EOT	ENCSR000DWE	ENCSR000AKP	ENCSR000AKQ	ENCSR000DWD	4DNFITUQMFUQ

Table 2: Data Availability for Epigenomic Tracks and HiC Maps Used in Experiments

The Hi-C data, obtained from the 4DN data portal, were provided in the .hic format. The data was binned at resolutions of 10 kb for downstream analyses. To normalize our Hi-C, Z-scores were calculated with the HiC-DC+ package [8], which employs a negative binomial regression model to estimate expected interaction counts. This regression model adjusts for genomic distance, GC content, mappability, and effective bin sizes, enabling the computation of normalized ratios. These preprocessing steps ensured the Hi-C data was prepared at high resolution and corrected for confounding factors, making it suitable for integration with the epigenetic marks during model training.

Chapter 6: Experiments & Results

Each model was trained on chromosomes 1-22 while excluding chromosomes 3, 11, and 17 for testing; this train-test split is consistent with several other works using the hg38 genome assembly. Each training example was an epigenomic window-V-stripe pair, randomizing the order of training examples with stochastic gradient descent. Our hyperparameters chosen for all model versions were an initial learning rate of 1×10^{-4}, a learning rate decay factor of 0.5 applied every 16 epochs, a minimum learning rate of 1×10^{-6}, batch size of 1, and max epochs of 55. All models took approximately 50 hours to train on an NVIDIA RTX A6000. We used 12 convolutional layers and six transformer layers in our experiments.

To test our method's generalizability within cell types, we trained Epiphany2 on each cell type's training chromosome and evaluated it on its test chromosomes. We ran this experiment for GM12878, H1-hESC, and K562 separately and visualized their results in Table #3.

To test the generalizability of our method across cell types, we trained Epiphany2 on the GM12878 and H1-hESC cell types and evaluated this model on the K562 cell line. Unless stated otherwise, all head-to-head comparisons of models in this study are trained on the same data and from random parameters for a fair comparison with the same training time (55 epochs).

We visualized the outputs for our test chromosomes, including the first 4 kb of GM12878 chromosome 3 in Figure 4. There, we see several chromatin loops (small highly-associated regions) that can only be identified by our new model Epiphany-2 and were missed by Epiphany; this pattern was consistent across our test chromosomes. In Table 3, we report the correlations between the insulation scores of the predicted Hi-C map and the ground-truth Hi-C. Our

insulation score was calculated at a distance of 1Mb or 100 10Kb Hi-C bins, representing our entire HiC matrix. However, future experiments should consider analyzing all distances to determine differential performance for different distances away from the genome. We ran three intra-cell-type experiments, with models trained on one cell type and evaluated on that cell type's test chromosomes.

For all tests, we say Epiphany-2 outperforms Epiphany. However, there was not a significant difference between the intra-cell-type experiments for H1-hESC Chr11; perhaps this K562 (blood cancer) chromosome is difficult to predict or highly stochastic in its 3D organization, so both methods had only moderate performance in the intra-cell-type setting. We saw an average insulation score correlation increase of 0.34 in the GM12878 cell line, 0.14 in the H1-hESC line, and 0.16 in the K562 line.

We also ran an inter-cell-type experiment, training on the GM12878 and H1-hESC cell types and generalizing to the K562 cell type test set. This model generalized very well to this new cell type with an average insulation score correlation of 0.66, demonstrating our method can identify cell-type-specific and conserved epigenomic rules that regulate the 3D chromatin structure, even in cancerous cells. Using data from a similar cell type without cancer, GM12878 as the healthy lymphoblast line and K562 as the cancerous lymphoblast line seems to improve performance in unseen chromosomes, with an increase of 0.12 insulation score correlation points on average.

Trained On		Chr3	Chr11	Chr17
	GM12878			
GM12878	**Epiphany**	0.43	0.28	0.41
GM12878	**Epiphany-2**	0.78	0.53	0.84
	Δ	+0.35	+0.25	+0.43
	H1-hESC			
H1-hESC	**Epiphany**	0.63	0.42	0.68
H1-hESC	**Epiphany-2**	0.77	0.59	0.80
	Δ	+0.14	+0.17	+0.12
	K562			
K562	**Epiphany**	0.24	0.49	0.40
K562	**Epiphany-2**	0.56	0.50	0.56
	Δ	+0.32	+0.01	+0.14
H1-hESC+GM1278	**Epiphany-2**	0.71	0.57	0.70
	Δ	+0.15	+0.07	+0.14

Table 3: Insulation Score Correlations for Epiphany 1 and 2 on our Cell Lines.

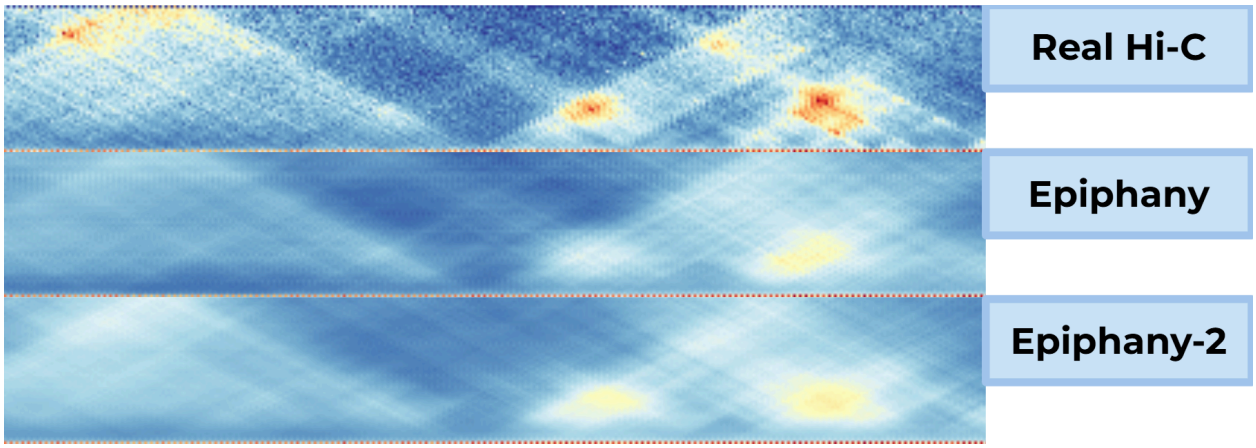


Figure 4: Example Output of Initial 4 kb of GM12878 Chromosome 3

Chapter 7: Conclusion

In this thesis, we introduced Epiphany2, a novel deep learning model that integrates convolutional neural networks with transformer layers to predict 3D chromatin structure from 1D epigenomic data. By incorporating a V-stripe prediction scheme alongside a transformer-based architecture, we address key limitations of prior models, such as Epiphany, which relied on biologically less relevant vertical stripe predictions and outdated mechanisms for modeling long-range dependencies.

Epiphany2 demonstrates state-of-the-art performance in predicting Hi-C contact maps, significantly improving the resolution of topologically associated domains (TADs) and chromatin loops. Our experiments show an increase in insulation score correlation between predicted and ground-truth Hi-C maps across multiple chromosomes and cell types. Our model generalizes well not only to unseen chromosomes within the same cell type but also across different cell types, including cancerous lines such as K562. This suggests Epiphany2 captures a robust, conserved understanding of the epigenomic rules governing 3D chromatin architecture across cell types.

This work positions transformer layers as a powerful and underutilized tool in epigenome-to-structure modeling. Their ability to capture global context enables our model to infer complex structural features that were previously difficult to predict with local architectures alone. The shift from vertical stripe to V-stripe prediction also enhances biological accuracy, ensuring that predictions more faithfully reflect the local epigenetic context of genomic interactions.

Ultimately, Epiphany2 brings us closer to a functional, computational proxy for experimental Hi-C assays. Its capacity to predict TAD boundaries and chromatin loops using only epigenomic input opens the door for cost-effective and scalable 3D genome inference—an essential step for understanding genome regulation, development, and disease.

Chapter 8: Limitations & Future Work

Several meaningful experiments fell out of the scope of this thesis. Primarily, we would like to investigate the performance of Epiphany2 at varying distances away from the genome; here, we evaluated the insulation scores (triangular sliding window average) out to 1Mb or our entire HiC contact map. However, evaluating the insulation score at closer distances may reveal a limited ability to capture certain short-range interactions and loops. We would also like to investigate which of our major contributions, transformer layers and the V-stripe prediction method, was more significant in improving our performance. We believe the transformer architecture was the major factor in Epiphany2's success, but further ablation studies should be performed. Moreover, biological ablation studies of zeroing out certain regions associated with chromatin loops may help us confirm our model has learned biologically accurate signals. Lastly, running experiments with different input tracks would reveal clues into what information is relevant to the 3D structure of chromatin. Finding a minimal subset of tracks that provide relatively high-quality 3D structures would help identify which marks are correlated with chromatin structure, and this line of experiments may find redundant or synergistic information between tracks.

Our current method is a sliding window across the genome with a relatively small window size. Each pixel is predicted twice, once by the left arm of the V and once by the right, and these predictions are averaged together. Inspired by the study of graph link prediction, implementing a graph-based method for aggregating Hi-C predictions will allow future models to overcome the major challenges with current methods, including "one-sided predictions," and improve its ability to capture long-range context [1]. Thinking about the epigenome as a graph where local epigenetic structure (nodes) are linked by edges (embedding distance on the 1D genome) leads to

adaptable and biologically meaningful predictions because 1) each prediction attends to its local context to infer local chromatin structure, and 2) message-passing algorithms can be used to learn complex long-range epigenetic relationships. For future studies, we recommend continued experimentation with our architectural choices and implementing a graph-based method for aggregating Hi-C pixel predictions, using our architecture as the encoder for each genomic position.

Bibliography

[1] Fan Feng. 2022. Connecting high-resolution 3D chromatin organization with epigenomic. Nature Communications 13, Article 2054 (2022). https://doi.org/10.1038/s41467-022-29695-6

[2] Geoff Fudenberg. 2020. Predicting 3D genome folding from DNA sequence with Akita. Nature Methods 17, Article 1111–1117 (2020). https://doi.org/10.1038/s41592-020-0958-x

[3] Vianne R. Gao. 2023. ChromaFold predicts the 3D contact map from single-cell chromatin accessibility. BioRxiv (2023). https://doi.org/10.1101/2023.07.27.550836

[4] Jimin Tan. 2023. Cell-type-specific prediction of 3D chromatin organization enables high-throughput in silico genetic screening. Nature Biotechnology 41, Article 1140–1150 (2023). https://doi.org/10.1038/s41587-022-01612-8

[5] Rui Yang. 2023. Epiphany: predicting Hi-C contact maps from 1D epigenomic signals. Genome Biology 24, Article 134 (2023). https://doi.org/10.1186/s13059-023-02934-9

[6] S. Zhang. 2019. In silico prediction of high-resolution Hi-C interaction matrices. Nature Communications 10, Article 5449 (2019). https://doi.org/10.1038/s41467- 019-13423-8

[7] Yunhai Luo. New developments on the Encyclopedia of DNA Elements (ENCODE) data portal. Nucleic Acids Res. 2020 Jan 8;48(D1): D882-D889. doi: 10.1093/nar/gkz1062

[8] Merve Sahin. HiC-DC+ enables systematic 3D interaction calls and differential analysis for Hi-C and HiChIP. Nat Commun 12, 3366 (2021). https://doi.org/10.1038/s41467-021-23749-x