Epiphany: Predicting the Hi-C Contact Map from 1D Epigenomic Data

Arnav Das [*1] **Rui Yang** [*2] **Vianne Gao** [2] **Alireza Karbalaghareh** [2] **William Noble** [1] **Jeff A. Bilmes** [1] **Christina Leslie** [2]

Abstract

We propose Epiphany, a light-weight neural network to predict the Hi-C contact map from five commonly generated epigenomic tracks: DNase I hypersensitive sites and CTCF, H3K27ac, H3K27me3, and H3K4me3 ChIP-seq. Epiphany uses 1D convolutional layers to learn local representations from the input tracks as well as bidirectional Long Short Term Memory (Bi-LSTM) layers to capture long term dependencies along the epigenome. To improve the usability of predicted contact matrices, we perform statistically principled preprocessing of Hi-C data using HiC-DC+ (1) and train Epiphany using an adversarial loss, enhancing its ability to produce realistic Hi-C contact maps for downstream analysis. We show that Epiphany generalizes to held-out chromosomes within and across cell types, and that Epiphany's predicted contact matrices yield accurate TAD and significant interaction calls.

1. Introduction

In mammalian genomes, the three-dimensional (3D) hierarchical folding of chromatin in the nucleus plays a critical role in the regulation of gene expression (2; 3). The 3D architecture of chromatin has been elucidated through genome-wide chromosome conformation capture (3C) assays such as Hi-C, Micro-C, HiChIP, and ChIA-PET (4; 5; 6; 7) followed by next generation sequencing, yielding a contact matrix representation of pairwise chromatin interactions.

Over the past decade, large consortium projects as well as individual labs have extensively used 1D epigenomic assays to map regulatory elements and chromatin states across numerous human and mouse cell types. While at least some of these 1D assays have become routine, mapping 3D interactions with Hi-C remains relatively difficult and prohibitively costly, and high-resolution contact maps (5Kb resolution, 2 billion read pairs) are still only available for a small number of cell types. This raises the question of whether it is possible to train a model to accurately predict the Hi-C contract matrix from more easily obtained 1D epigenomic data in a cell-type specific fashion.

Recent deep learning models have made advances in predicting 3D genomic structure. For example, DeepC (8) presented a transfer learning framework by learning useful DNA representation from epigenetic marks, then fine-tuning the model to predict the Hi-C contact map. Akita (9) designed a convolutional neural network to predict the Hi-C contact maps of multiple cell types from DNA sequence. However, neither method uses epigenomic data as an input signal, and the resulting models capture very limited cell-type specific information about 3D genomic architecture. We therefore propose Epiphany, a light-weight neural network to predict the cell-type specific Hi-C contact map from commonly generated epigenomic tracks.

2. Methods

2.1. Dataset

Hi-C data. Deeply sequenced Hi-C data for GM12878 from the 4DN data portal was used to train the model. The data set was processed using the hg38 genome assembly and binned at 10Kb resolution to generate chromosome contact maps. Normalization was conducted using the HiC-DC+ R package using the observed over expected (Obs/Exp) ratio from a negative binomial regression that accounts for genomic distance and other covariates. During training, chromosome 3, 11, 17 were held-out as test chromosomes.

Epigenetics data. Five input epigenomic tracks including DNaseI, CTCF, H3K4me3, H3K27ac, H3K27me3 for genome assembly hg38 were downloaded from the ENCODE data portal. Data were downloaded as bam files, and the replicates for each track were merged using the pysam python module. We then converted merged bam files into bigWig files with deepTools bamCoverage (binSize 10, RPGC normalization, other parameters as default). Genome-wide coverage bigWig tracks were later binned into 100bp

[*]Equal contribution [1]University of Washington, Seattle, WA, US [2]Computational Systems Biology Program, Memorial Sloan Kettering Cancer Center, New York City, NY, US. Correspondence to: Christina Leslie <cleslie@cbio.mskcc.org>, Jeff A. Bilmes <bilmes@uw.edu>.

The 2021 ICML Workshop on Computational Biology.

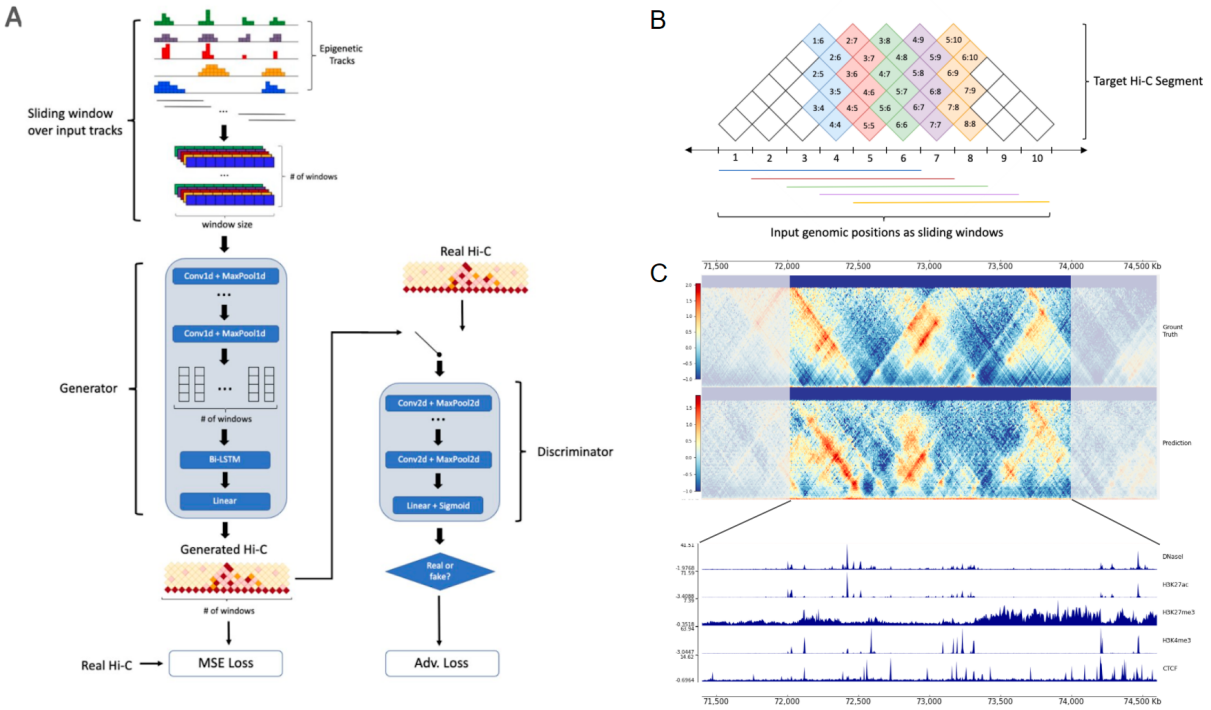

Figure 1. Model architecture and prediction scheme. **(A)** Model Architecture. **(B)** Prediction scheme of Hi-C contact map. **(C)** Ground truth target, model prediction on the Hi-C contact map (2Mb region), and input epigenomic tracks (3.2Mb receptive field).

bins, and bin-level signals for the 5 epigenomic tracks were extracted as input data for the model.

2.2. Model Architecture

Epiphany consists of two parts: a generator to extract information and make predictions, and a discriminator to introduce adversarial loss into the training process (**Fig. 1A**). In the generator, we first used a series of convolution modules to featurize epigenomic information in a sliding window pattern. For one output vector, which covers a distance of 1Mb orthogonal to the diagonal, we used a window size of 1.2 Mb centered at the corresponding region as input (**Fig. 1B**). We then employed a Bi-LSTM layer to capture the dependencies between output vectors, and a total of 3.2 Mb input were processed in one pass for prediction of 200 output vectors. At the end, a fully connected layer was used to integrate signals and make the final prediction. We also introduced an adversarial loss and a discriminator, which consists of several convolution modules that are applied during training and pushes the generator to produce realistic samples (**Fig. 1C**).

CNN layers. The input epigenomic tracks are divided into overlapping windows, with a window length of $m = 12,000$ bins (1.2Mb) and a stride of 1,000 bins (100Kb). We refer to the windowed inputs as $X = \{x_1, ..., x_n\}$, where $x_i \in \mathbb{R}^{c \times m}$ corresponds to window i, n is the total number of windows, and c is the number of epigenetic tracks. A series of four convolution modules are used to featurize each window into a vector of dimension $d = 900$ (after flattening), where each convolution module consists of a convolutional layer with ReLU activation, max pooling, and dropout. We define $Z = \{z_1, ..., z_n\}$ as the flattened output of the final convolution module where $z_i \in \mathbb{R}^d$ is the representation for window x_i.

Bi-LSTM layers. The Bi-LSTM layers receive sequence $Z = \{z_1, ..., z_n\}$ as input and generate a new sequence $\tilde{Z} = \{\tilde{z}_1, ..., \tilde{z}_n\}$, where $\tilde{z}_i \in \mathbb{R}^{2d}$. To produce the final output, every element of $\tilde{Z}$ is passed through a fully connected layer yielding the output sequence $\hat{Y} = \{\hat{y}_1, ..., \hat{y}_n\}$. Each $\hat{y}_i \in \mathbb{R}^{d'}$ is a vector of dimension $d' = 100$ and corresponds to a *zig-zag* pole in the Hi-C matrix, similar to DeepC (shown in **Fig. 1**). Epiphany uses a total of three Bi-LSTM layers, with residual connections between successive layers.

Adversarial loss. Generative adversarial networks (GAN) consist of two networks, a generator $\mathcal{G}$ with parameters $\theta^{\mathcal{G}}$ and a discriminator $\mathcal{D}$ with parameters $\theta^{\mathcal{D}}$, that are adversarially trained in a zero-sum game (10; 11). During training, the generator learns to fool the discriminator by synthesizing realistic samples from a given input, while the discriminator learns to distinguish real samples from synthetic samples. To train Epiphany, we employed a convex combination of pixel-wise MSE and adversarial loss. Given a dataset D and a trade-off parameter λ, Epiphany solves the following optimization problem during training:

$$\min_{\theta^{\mathcal{G}}} \max_{\theta^{\mathcal{D}}} \lambda \mathcal{L}_{adv}(\theta^{\mathcal{G}}, \theta^{\mathcal{D}}) + (1 - \lambda)\mathcal{L}_{MSE}(\theta^{\mathcal{G}})$$

$$\mathcal{L}_{adv}(\theta^{\mathcal{G}}, \theta^{\mathcal{D}}) = \mathbb{E}_{(X,Y)\sim D}[\log(\mathcal{D}(Y)) + \log(1 - \mathcal{D}(\mathcal{G}(X)))]$$

$$\mathcal{L}_{MSE}(\theta^{\mathcal{G}}) = \mathbb{E}_{(X,Y)\sim D}\left[\sum_{i\in[n]}\sum_{j\in[d']}(Y_{ij} - [\mathcal{G}(X)]_{ij})^2\right]$$

In our framework, $\mathcal{G}$ is the CNN-LSTM architecture described in the previous sections while $\mathcal{D}$ is a simple four layer 2D CNN. In our experiments, we used $\lambda = 0.05$.

3. Results and Evaluation

3.1. Epiphany generates realistic Hi-C contact maps

Past approaches that predict the 3D genome structure from 1D inputs use pixel-wise MSE to quantify the similarity between predicted and ground truth Hi-C maps. However, pixel-wise losses for images have been shown to be overly sensitive to noise (12) and to yield blurry results when used as objectives for image synthesis (13; 10). These issues can be mitigated with an adversarial loss, which enables the model to generate highly realistic samples, while circumventing the need to explicitly define similarity metrics for complex modalities of data.

We benchmarked the model between two loss functions, MSE only loss vs. convex combination of MSE and adversarial loss. The Pearson and Spearman correlations of both models are shown in **Table 1**.

Loss Function	Pearson (all)	Pearson (train)	Pearson (test)
MSE only	0.7833	0.8045	0.6494
MSE+GAN	0.7408	0.7687	0.5636
Loss Function	Spearman (all)	Spearman (train)	Spearman (test)
MSE only	0.7381	0.7605	0.5963
MSE+GAN	0.6899	0.7191	0.5048

Table 1. Mean Pearson and Spearman correlation for two loss functions

Epiphany demonstrates better performance for both correlation metrics with the MSE loss than the convex combination of MSE and adversarial loss. However, we observed that the high correlations from MSE trained models were associated with blurriness in the predicted contact maps, and while the correlations produced by the combined loss models may have been slightly diminished due to small deviations in the sharper predictions.

Therefore, we reasoned that correlation may not be an appropriate evaluation metric and decided to use results from the combined loss (MSE+Adversarial loss) for downstream analysis. A visual comparison of the blurry prediction made by MSE trained models vs. the more realistic prediction made by the combined loss is shown in **Fig. 2**.

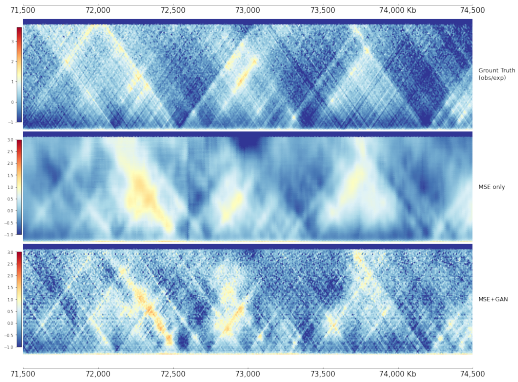

Figure 2. Prediction comparison between two loss functions on region (chr17:70670000-73880000). (Top) Normalized Obs/Exp ground truth. (Middle) Prediction from MSE-only model. (Bottom) Prediction from combined loss model.

3.2. Bi-LSTM layer captures potential contribution of distal elements

Given the sequential nature of Hi-C contact maps, interactions on consecutive output vectors are unlikely to be independent from one another. We found that Bi-LSTM layers introduce stronger dependencies between the output vectors, which better equips Epiphany to leverage structures that span multiple genomic positions in Hi-C maps (such as edges of TADs). Furthermore, Bi-LSTM layers overcome the limitation of CNNs by enabling each output vector to make use of important signals beyond the input window. To explore the contribution of input signals, we calculated the saliency score and SHAP value for attributions. **Fig. 3A** shows an example region (chr17:55945000-58945000) where a distal regulatory element may contribute to the 3D genome structure prediction through the Bi-LSTM layers.

3.3. Epiphany predicts cell-type specific 3D structure

Since Epiphany uses epigenomic marks as input, it can potentially generalize to a new cell type and predict cell-type specific 3D structures. Here we first trained Epiphany on a single cell type (GM12878) and tested on another (K562) to check the generalization performance. Chromosome 3, 11, 17 are completely held-out during training, and chr17 is used for testing on K562. We obtained an average Pearson correlation of 0.4732 across all genomic distances (Spearman correlation of 0.4268).

Fig. 3B shows a differential region (chr17:70670000-73880000) between GM12878 and K562, providing contact maps on the top, followed by input tracks on the bottom. For GM12878, peaks in DNaseI, H3K27ac and H3K27me3 between 71800000 and 72700000bp region are important in the model and contribute to the specific interaction highlighted in the map. K562 does not have strong interactions in this region, and due to the lack of input peaks, Epiphany

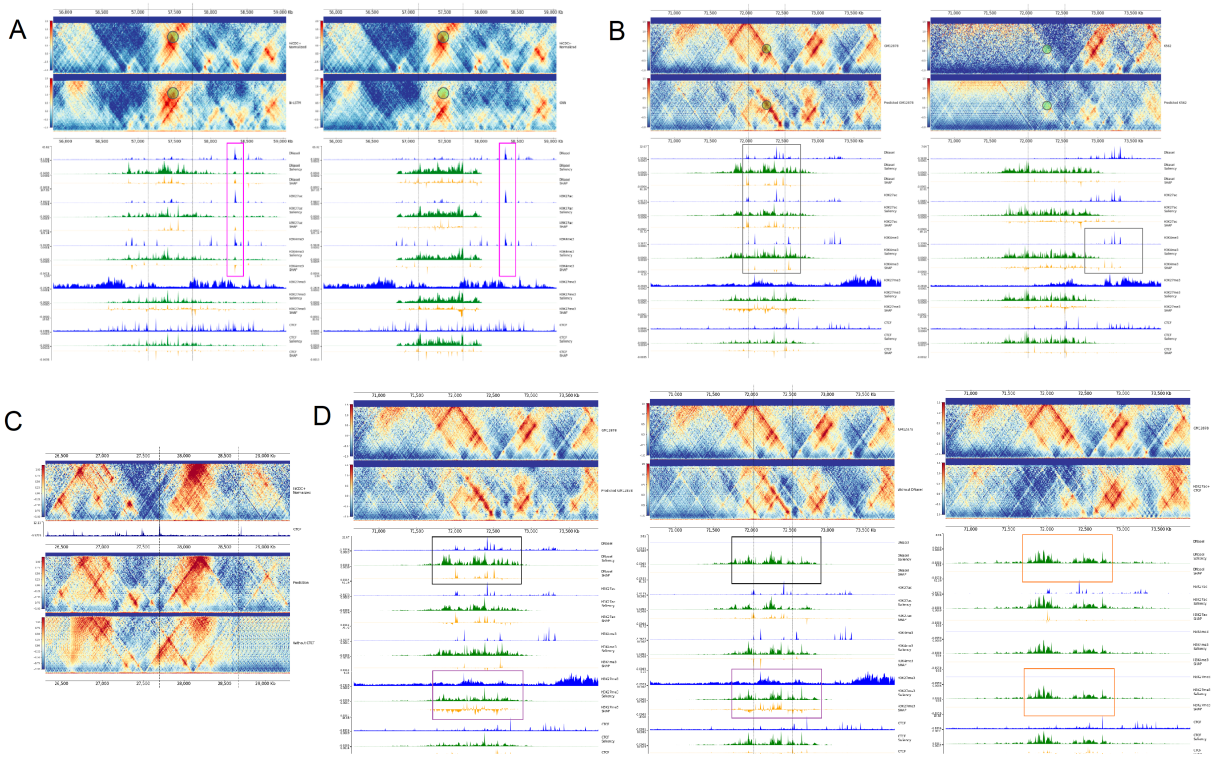

Figure 3. Results summary. **(A)** Example region (chr17:55945000-58945000) comparison between introduction of Bi-LSTM layer vs. 1d CNN layer. (Left) Ground truth Hi-C contact map and Bi-LSTM prediction, followed by input epigenomic signals (blue), saliency score (green) and SHAP values (yellow). Epigenomic signals from top to bottom: DNaseI, H3K27ac, H3K4me3, H3K27me3, CTCF. (Right) Prediction and attribution with 1D CNN layer. **(B)** Cell-type specific prediction comparing GM12878 (left) vs. K562 (right) on example region (chr17:70670000-73880000). **(C)** Prediction comparing full model (middle) vs. CTCF-ablated model (bottom) on region (chr11:26300000-2930000). **(D)** Prediction comparing full model (left), DNaseI-ablated model (middle) and CTCF+H3K27ac model (right) on region (chr17:70670000-73880000).

correctly predicts a weak interaction.

3.4. Epiphany learns the role of CTCF in genome folding

To confirm the importance of CTCF for predicting 3D interactions, we retrained the model with the entire CTCF track masked as zero and compared with the original model using all input tracks. **Fig. 3C** shows an example region (chr11:26300000-29300000) where the prediction of the full model and the CTCF-ablated model diverge. By exploiting the CTCF signal, the full model can accurately predict the entire TAD structure in this region (**Fig. 3C middle**), while the CTCF-ablated model (**Fig. 3C bottom**) only captured some of the interactions and "hallucinates" a small TAD structure between region chr11:27530000-28100000 that is not present in the true map (**Fig. 3C top**).

3.5. Epiphany identifies the contribution of epigenomic marks to 3D structure

In the comparison between GM12878 vs. K562, we found that some H3K4me3 peaks in distal regions gain importance in the model (**Fig. 3B**). We then wondered whether features from different epigenomic tracks could compensate for each other and more generally what redundancies were present among the tracks.

This idea matched our observations from the ablation analysis. We compared the prediction for this region using a model trained on all inputs, without DNaseI, or with CTCF+H3K27ac only. Epiphany was still able to accurate predictions in this region after ablating DNaseI; feature attribution indicated that in place of DNaseI signal (**Fig. 3D**, grey box), this model gave higher importance to H3K27me3 peaks (purple box) in order to predict the interaction. However, after ablating all signals except for CTCF and H3K27ac, the model failed to find alternative predictive signals and missed the boundary.

4. Conclusion

By training on epigenomic data, Epiphany is able to predict cell-specific 3D chromatin architecture, generalize across cell types, and attribute individual 3D interactions to epigenomic events. Future work will focus on integrating genomic DNA sequence with epigenomic data as inputs to the model.

References

[1] Sahin, M., Wong, W., Zhan, Y., Van Deynze, K., Koche, R., Leslie, C.S.: HiC-DC+: systematic 3D interaction calls and differential analysis for Hi-C and HiChIP. bioRxiv (2020)

[2] Dekker, J., Belmont, A.S., Guttman, M., Leshyk, V.O., Lis, J.T., Lomvardas, S., Mirny, L.A., O'shea, C.C., Park, P.J., Ren, B., *et al.*: The 4D nucleome project. Nature **549**(7671), 219–226 (2017)

[3] Zheng, H., Xie, W.: The role of 3D genome organization in development and cell differentiation. Nature Reviews Molecular Cell Biology **20**(9), 535–550 (2019)

[4] 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.: Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science **326**(5950), 289–293 (2009)

[5] Hsieh, T.-H.S., Weiner, A., Lajoie, B., Dekker, J., Friedman, N., Rando, O.J.: Mapping nucleosome resolution chromosome folding in yeast by Micro-C. Cell **162**(1), 108–119 (2015)

[6] Mumbach, M.R., Rubin, A.J., Flynn, R.A., Dai, C., Khavari, P.A., Greenleaf, W.J., Chang, H.Y.: HiChIP: efficient and sensitive analysis of protein-directed genome architecture. Nature Methods **13**(11), 919–922 (2016)

[7] Fullwood, M.J., Liu, M.H., Pan, Y.F., Liu, J., Xu, H., Mohamed, Y.B., Orlov, Y.L., Velkov, S., Ho, A., Mei, P.H., Chew, E.G., Huang, P.Y., Welboren, W.J., Han, Y., Ooi, H.S., Ariyaratne, P.N., Vega, V.B., Luo, Y., Tan, P.Y., Choy, P.Y., Wansa, K.D., Zhao, B., Lim, K.S., Leow, S.C., Yow, J.S., Joseph, R., Li, H., Desai, K., Thomsen, J., Lee, Y., Karuturi, R., Herve, T., Bourque, G., Stunnenberg, H.G., Ruan, X., Cacheux-Rataboul, V., Sung, W.K., Liu, E.T., Wei, C.L., Cheung, E., Ruan, Y.: An oestrogen-receptor-α-bound human chromatin interactome. Nature **462**(7269), 58–64 (2009)

[8] Schwessinger, R., Gosden, M., Downes, D., Brown, R.C., Oudelaar, A.M., Telenius, J., Teh, Y.W., Lunter, G., Hughes, J.R.: DeepC: predicting 3D genome folding using megabase-scale transfer learning. Nature Methods **17**(11), 1118–1124 (2020)

[9] Fudenberg, G., Kelley, D.R., Pollard, K.S.: Predicting 3D genome folding from dna sequence with akita. Nature Methods **17**(11), 1111–1117 (2020)

[10] Goodfellow, I.: NIPS 2016 tutorial: Generative adversarial networks. arXiv preprint arXiv:1701.00160 (2016)

[11] Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A.C., Bengio, Y.: Generative adversarial nets. In: Advances in Neural Information Processing Systems (2014)

[12] Zhao, H., Gallo, O., Frosio, I., Kautz, J.: Loss functions for image restoration with neural networks. IEEE Transactions on Computational Imaging **3**(1), 47–57 (2017). doi:10.1109/TCI.2016.2644865

[13] Xu, X., Sun, D., Pan, J., Zhang, Y., Pfister, H., Yang, M.-H.: Learning to super-resolve blurry face and text images. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 251–260 (2017)