

Epiphany: predicting Hi-C contact maps from 1D epigenomic signals

Arnav Das^{2†}, Rui Yang^{1†}, Vianne R. Gao¹, Alireza Karbalayghareh¹, William S. Noble², Jeffrey A. Bilmes^{2*}, Christina S. Leslie^{1*}

¹Memorial Sloan Kettering Cancer Center, New York, NY, USA

²University of Washington, Seattle, WA, USA

†Equal contribution, *Correspondence: cleslie@cbio.mskcc.org, bilmes@uw.edu



Overview

We propose Epiphany, a neural network to predict cell-type-specific Hi-C contact maps from five epigenomic tracks that are already available in hundreds of cell types and tissues: DNase I hypersensitive sites and ChIP-seq for CTCF, H3K27ac, H3K27me3, and H3K4me3. Epiphany uses 1D convolutional layers to learn local representations from the input tracks, a bidirectional long short-term memory (Bi-LSTM) layers to capture long term dependencies along the epigenome, as well as a generative adversarial network (GAN) architecture to encourage contact map realism. Epiphany is trained with a combination of MSE and adversarial (i.e., a GAN) loss to enhance its ability to produce realistic Hi-C contact maps for downstream analysis. At inference time, Epiphany can be used to study the contribution of specific epigenomic peaks to 3D architecture and to predict the structural changes caused by perturbations of epigenomic signals.

Materials and Strategy

Materials

Hi-C contact map: GM12878, K562, mES cell lines binned into 5 kb and 10 kb resolution, with ICE, KR, HiC-DC+ Z-score and observed-over-expected count ratio normalizations.

Epigenomic tracks: 5 epigenomic tracks including DNaseI, CTCF, H3K27ac, H3K27me3, H3K4me3

Prediction scheme

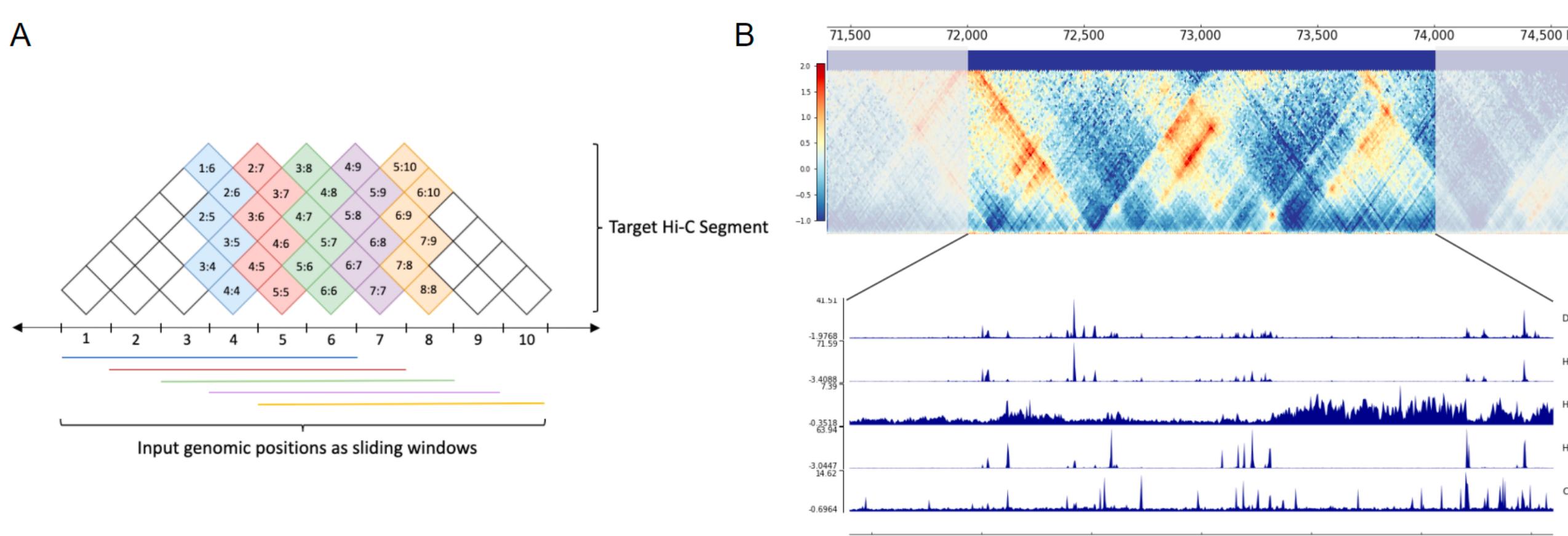


Fig 1. Prediction scheme

Key ideas:

- **Vertical stripe prediction:** predict 1Mb orthogonal-to-diagonal vectors on the contact map (A).
- **Receptive field:** for each 1Mb predicted vector, a window size of 1.4 Mb centered at the corresponding region is used on epigenomic input tracks.
- **Predict multiple output vectors:** Epiphany captures dependencies between adjacent vectors and predicts a consecutive of output vectors simultaneously.

Model Architecture

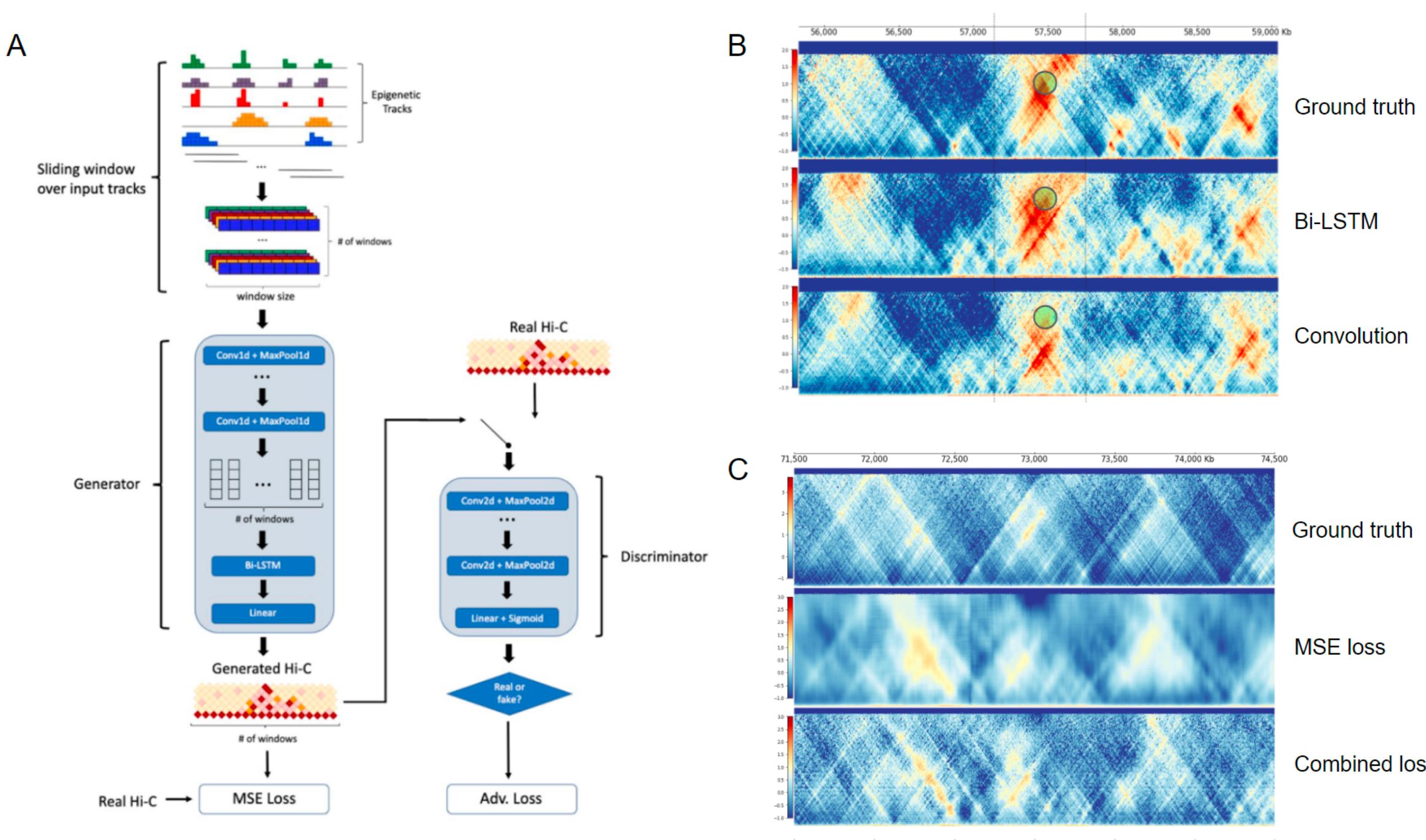


Fig 2. Model structure

Key structures

- **Bi-LSTM layers:** the introduction of Bi-LSTM layers instead of convolutional layers captures dependencies between adjacent predictions and increases the receptive field to capture distal elements.
- **Discriminator component:** introducing discriminative loss pushes Epiphany to predict realistic Hi-C contact maps.

Loss function

Convex combination of pixel-wise MSE and adversarial loss to enhance its ability to produce realistic Hi-C contact maps for downstream analysis. Given a dataset D and a trade-off parameter λ , Epiphany solves the following optimization problem during training:

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

Model Performance

Prediction performance

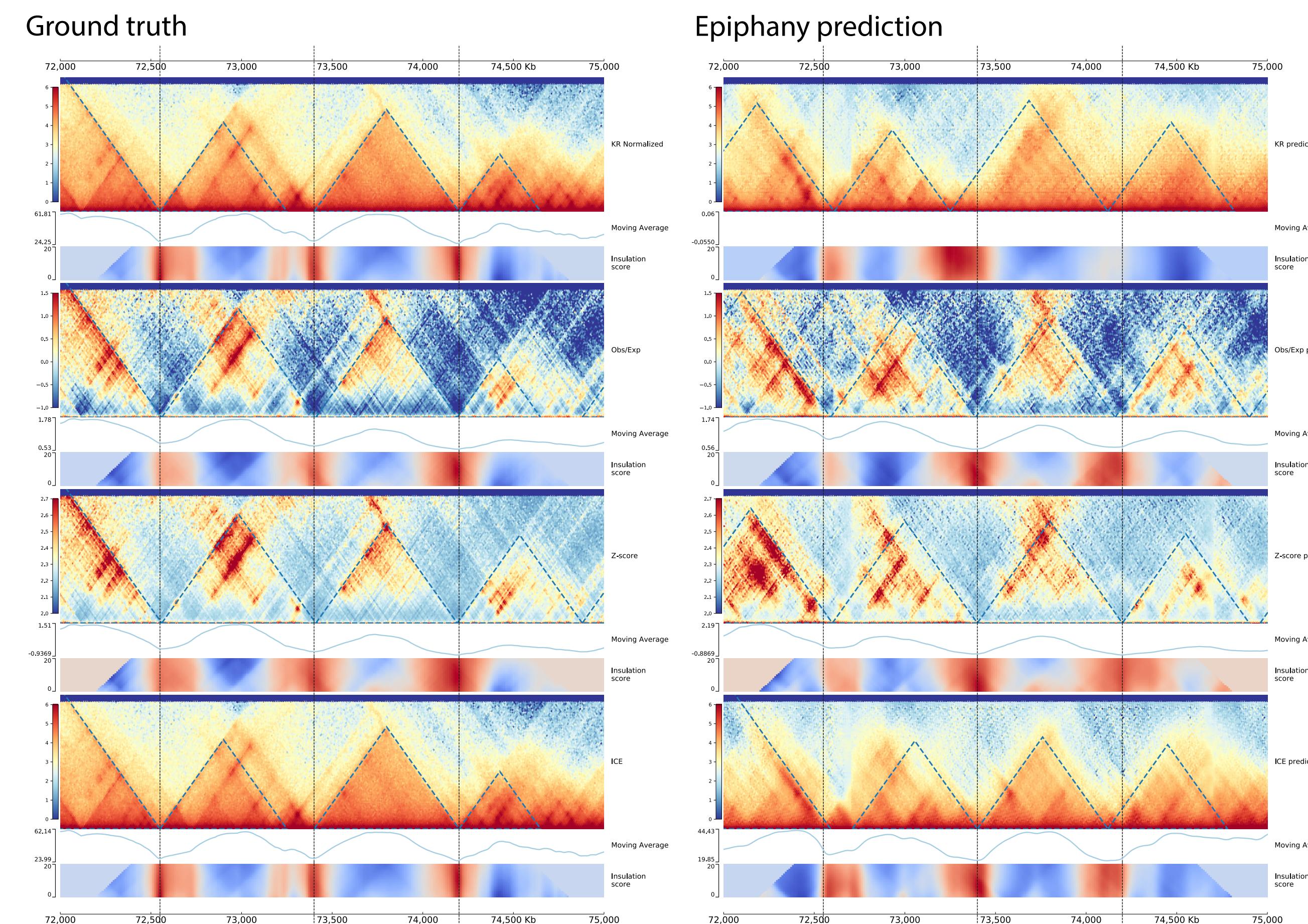


Fig 3. Prediction results. Left: ground truth with difference normalization methods. Right: Epiphany prediction.

Applications

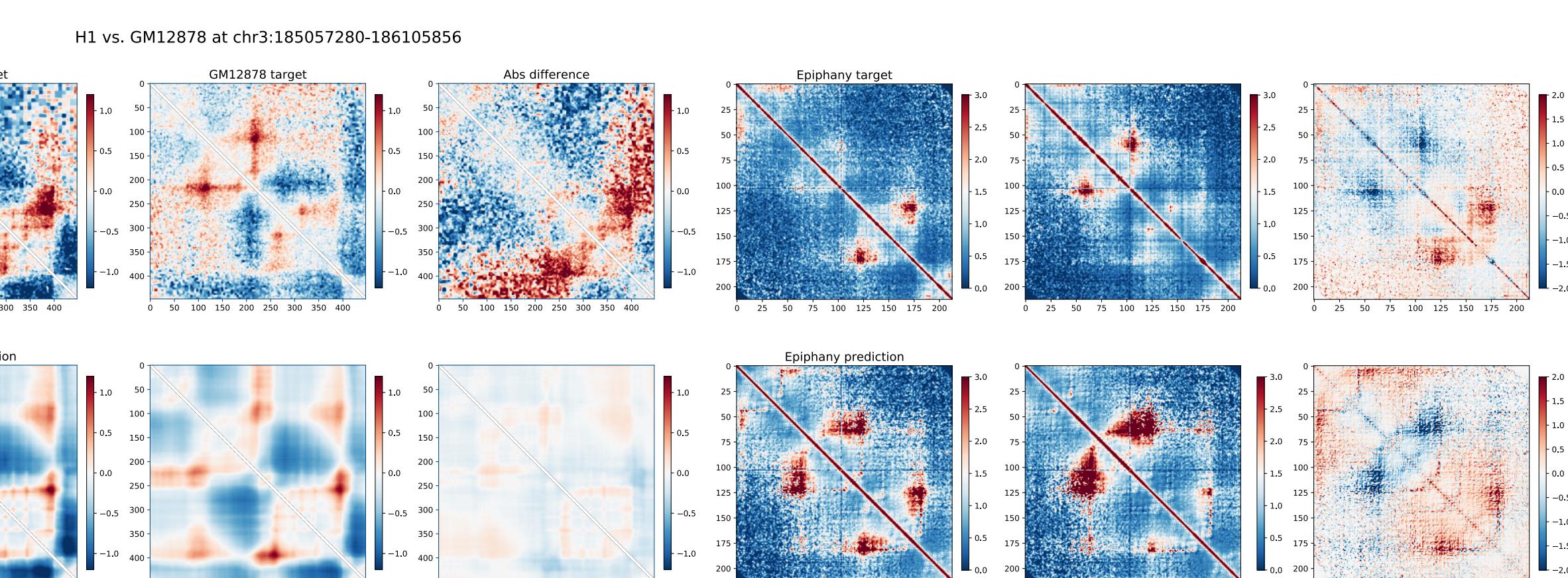


Fig 4. Cell-type specific prediction. Left: Akita prediction of H1ESC vs. GM12878. Right: Epiphany prediction of H1ESC vs. GM12878.

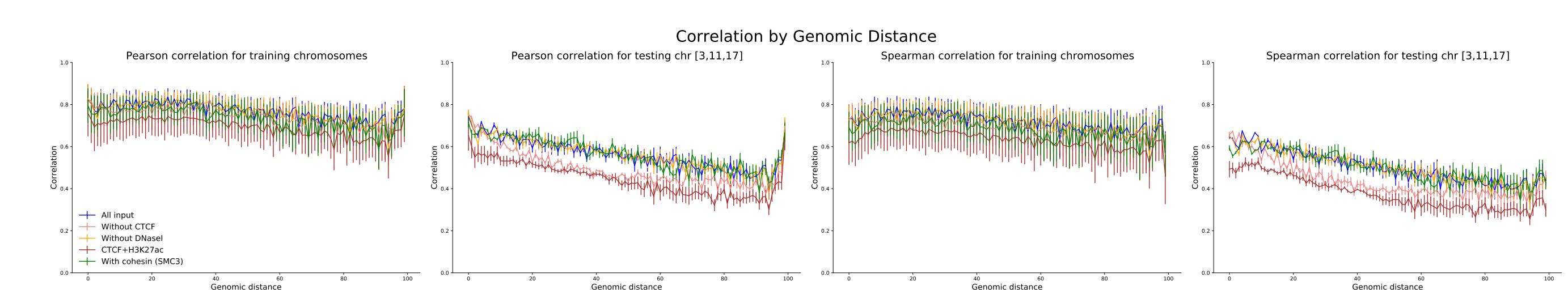


Fig 5. Pearson and Spearman correlation of feature ablation analysis.

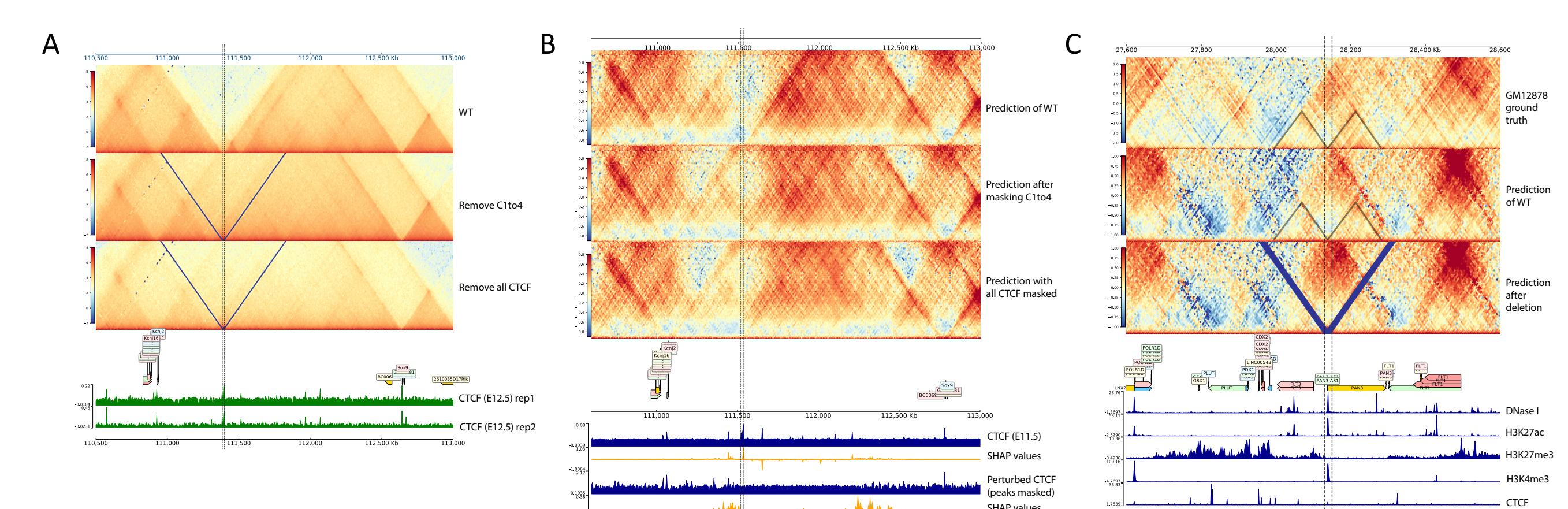


Fig 6. Biological applications.

Epiphany Highlights

Epiphany: a neural network to predict chromatin 3D structure from epigenomic tracks.

- A framework connecting 1D–3D
- A model to generate cell-type specific Hi-C contact maps for rare or new cell types.